CARACTERIZAÇÃO DA DISTRIBUIÇÃO DE CARGA EM REDES COMPLEXAS SUBMETIDAS A UM TRÁFEGO UNIFORME

Elias Bareinboim

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIAS DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. Celina Miraglia Herrera de Figueiredo, D.Sc.

Prof. Raul Donangelo, Ph.D.

Prof. Virgilio Augusto Fernandes Almeida, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2007

BAREINBOIM, ELIAS

Caracterização da Distribuição de Carga em Redes Complexas submetidas a um tráfego uniforme [Rio de Janeiro] 2007

VIII, 78 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2007)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1 - Distribuição de Carga

2 - Redes Complexas

3 - Tráfego Uniforme

I. COPPE/UFRJ II. Título (série)

B"H

A meu querido pai,
meu maior exemplo de vida,
quem sempre me orientou, me ensinou,
me apoiou, me incentivou, e me motivou,
e que, fisicamente, não mais está presente.

Agradecimentos

A minha mãe, meu irmão, meu primo e minha família, por tudo que representam para mim.

Ao Professor Valmir, que mesmo antes de ser meu orientador, conversou e me orientou a respeito das escolhas de disciplinas e temas em que me envolvesse durante o curso. Depois, já como orientador, por todo o suporte logístico, técnico e pessoal, e também por servir como um exemplo. Talvez tão importante quanto, ou até mais que os resultados específicos deste trabalho em si, foi o aprendizado quanto ao método de como desenvolver uma pesquisa; como olhar um problema; como se interessar e conviver com o mesmo; como manter uma linha de estudos coerente; quão indispensável é se manter focado; quanto tempo vale a pena insistir; quando vale a pena abrir outras alternativas; onde pesquisar determinado assunto; como conseguir manter uma rotina alinhada com seus objetivos. Realmente é um aprendizado constante, complexo, fundamental e indispensável. Muito obrigado.

Aos Profs. Celina, Raul e Virgílio, por aceitarem fazer parte desta banca, agregando valor inestimável a este trabalho.

Aos Profs. Fábio Protti, João Carlos P. Silva e Vitor Santos Costa, que me ajudaram desde a graduação, sempre solícitos, dispostos e amigáveis.

Aos colegas que colaboraram durante o meu mestrado, nos momentos mais variados e através de diversas formas: Alexandre Stauffer, Carina Lopes, Daniel Levitan, Daniel Sadoc, Paulo Carvalho, Rodrigo Hausen.

À CAPES, por viabilizar financeiramente este projeto, através da bolsa a mim concedida.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CARACTERIZAÇÃO DA DISTRIBUIÇÃO DE CARGA EM REDES COMPLEXAS SUBMETIDAS A UM TRÁFEGO UNIFORME

Elias Bareinboim

Setembro/2007

Orientador: Valmir Carneiro Barbosa
Programa: Engenharia de Sistemas e Computação

Uma rede é dita complexa quando possui um número grande de elementos e sua topologia geralmente não é conhecida em detalhes. A distribuição para o grau de seus vértices segue uma lei de potências. Exemplos de rede deste tipo, e pela qual temos interesse, são a Internet e a Web. Nosso trabalho tem como objetivo a caracterização de uma propriedade de tais redes conhecida como Carga, que representa o número de rotas que passam por cada nó, em nosso caso, quando submetidos a um tráfego uniforme (todos os nós enviam mensagens entre si). Essas rotas estão diretamente relacionadas com as árvores de caminhos mais curtos, e então desenvolvemos uma caracterização para a distribuição de graus em árvores desse tipo. Obtemos uma expressão analítica e efetuamos simulações sobre a mesma. O resultado obtido possibilitou caracterizarmos a distribuição da descendência de um vértice em árvores desse tipo, com uma expressão analítica que avaliamos por simulações. Contudo, chegamos a algumas restrições quanto a direta aplicabilidade da descendência para obtenção da Carga. Então, através de uma nova análise baseada em resultados experimentais propomos uma nova explicação para a Carga, utilizando inclusive a própria distribuição da descendência. Finalmente, discutimos e apontamos inconsistências existentes na literatura.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

CHARACTERIZATION OF THE LOAD DISTRIBUTION IN COMPLEX NETWORKS SUBMITTED TO A UNIFORM TRAFFIC

Elias Bareinboim

September/2007

Advisor: Valmir Carneiro Barbosa

Department: Systems Engineering and Computer Science

A network is regarded as a complex network when it has a high number of elements and the topology is not known in details. The distribution of the degrees of the vertices follows a power law. Examples of this kind of network, in which we are interested in, are the Internet and the Web. Our work aims to characterize the property of these networks known as Load which represents the number of routes passing through each node, and in our case, when submitted to an uniform traffic (all nodes sending messages to each other). These routes are directly related to the shortest path trees where we characterized the distribution of the degrees in this same type of trees, obtaining an analytical expression and evaluating some simulation on it. Based on these results, we characterized the distribution of the vertex's descendants in the shortest path trees. This was also verified by simulations. However, there are several restrictions to its applicability for Load estimation. Thus, we developed a new explanation based on our experimental results to characterize the Load, taking advantage of the vertex descendant distribution. Finally, we review and point out some inconsistencies available in the literature.

Conteúdo

Lista de Figuras

Capítulo 1

Introdução

1.1 Preliminares

Uma rede é dita complexa quando possui um número grande de elementos e sua topologia geralmente não é conhecida em detalhes. Muitas dessas redes, como a Internet, a WWW e diversas outras que retratam interações físicas, químicas, biológicas ou sociais (tabela 1.1) podem ser caracterizadas pelo fato de que o número de vizinhos de seus nós [1] (seus graus) são distribuídos segundo uma lei de potências. Tais redes têm, então, uma importante característica de independência de escala, conforme atestado em [18] para o caso da Internet e nos trabalhos coletados em [11, 27] para diversos outros casos.

Os modelos de redes complexas contemporâneos se contrapõem aos grafos randômicos [2] ditos clássicos propostos por Paul Erdõs e Alfréd Renyi em [15, 16, 17] nas décadas de 50 e 60, que consistem em agrupar, criando arestas, os n vértices do grafo através de uma probabilidade fixa p, considerando o número de vértices n grande (tendendo a infinito), resultando em uma distribuição de graus do tipo Poisson. Existem propriedades derivadas de tal modelo bastante interessantes,

[1] Iremos utilizar a palavra nós e vértices como sinônimos. Faremos o mesmo para arestas e conexões e também para rede e grafo. Chamaremos também de grafo mas nos referimos de fato a sua versão mais geral, ou multigrafo, em seu sentido estrito.

[2] Um grafo é dito randômico quando ele é gerado através de um processo aleatório.

como a transição de fase [3] bem definida e o então surgimento de uma componente gigante [4] através de determinado valor de grau médio z.

Rede	Vértices	Arestas
Internet	Roteadores	Conexão ótica ou física de outra natureza.
WWW	Páginas	Links entre as páginas através de URLs.
Colaboração científica	Cientistas	Co-autoria de um artigo.
Hollywood	Atores	Apareceram no mesmo filme.
Metabolismo Celular	Moléculas	Participação na mesma reação bioquímica.
Rede de **regulação proteíca**	**Proteínas**	Interação para **regulação celular entre** protcínas.

Tabela 1.1: Exemplo de algumas redes complexas.

Então, os grafos randômicos no sentido clássico diferem do comportamento encontrado em redes do mundo real de forma fundamental; iremos citar duas básicas para exemplificar. Primeiramente, as redes reais têm um forte coeficiente de agrupamento, o que não acontece com as randômicas. Uma rede é dita com alto agrupamento se a probabilidade de dois vértices estarem conectados por uma aresta for maior quando ambos tiverem um vizinho em comum. Tal propriedade pode ser medida através de, por exemplo, um coeficiente representando a probabilidade média de dois vizinhos de um dado vértice serem também vizinhos entre si.

Em redes reais, o coeficiente de agrupamento exibe valores como de alguns pontos percentuais até 50% ou mais, enquanto no modelo original de Erdõs e Rényi este valor é, por definição, independente dos vértices serem ou não vizinhos. Conforme [11], temos exemplos de variados tipos de redes e seus respectivos coeficientes de agrupamento, mostrando claramente tais evidências.

Outra característica fundamental diz respeito a sua distribuição de graus conforme enfatizado em [4, 5], onde, em redes randômicas clássicas a distribuição de graus é do tipo Poisson, como já dito, enquanto em redes reais isso não acontece, pois seguem uma lei de potência. Iremos estudar e aprofundar mais justamente sobre esta característica que é fundamental para o nosso trabalho.

[3] A expressão transição de fase é muito utilizada em termodinâmica para descrever uma transformação em um sistema de uma fase a outra, onde de forma abrupta uma mudança em alguma de suas propriedades ocorre.

[4] A transição de fase citada anteriormente se dá justamente no aparecimento de uma componente chamada gigante, que representa um subgrafo conexo que contém a maior parte dos vértices do grafo.

Em termos de explicação para essa diferenciação entre as redes clássicas e as reais, serão considerados os mecanismos que descrevem sua formação. Observa-se que as redes reais estudadas contêm um elemento fundamental derivado das características até aqui expostas, que é a existência dos vértices considerados concentradores (*hubs*), com seu grau de alta magnitude, conectando muitos nós e diminuindo o diâmetro da rede, o que não acontece no modelo de grafos randômicos clássico. Ao desenvolverem seu modelo, Erdõs e Rényi assumiram que tinham a lista completa dos nós antes dos mesmos terem arestas, *a priori*. Em contraste, em redes reais isso não acontece. Por exemplo, o número de documentos da Web vem crescendo, sendo que inicialmente, em 1990, havia somente uma página, e agora há mais de três bilhões. Outras redes reais se expandiram de forma similar, como, por exemplo, os atores em Hollywood que eram alguns poucos em 1890 e agora formam um contingente com mais de meio milhão de membros. Outro exemplo, a Internet, considerando-se seus roteadores, e que tinha alguns poucos dos mesmos a menos de três décadas atrás e agora somam milhões, sendo que os novos se ligam aos já existentes.

Além disso, nem todos os nós são iguais e possuem ligações independentes, conforme acontece com o modelo clássico. Por exemplo, ao se criar uma página nova e ligá-la, teoricamente pode-se escolher entre bilhões de opções, só que na prática se conhece somente uma fração mínima deste universo. Em decorrência de tal fato, ao se criar uma nova página, a mesma terá ligações mais prováveis aos sites mais conectados (com maior grau), pois devido a isto são também os mais conhecidos, fáceis de encontrar e então, constituem os alvos realmente mais viáveis de tais ligações. O mesmo vale para artigos, atores, e demais exemplos. Tal mecanismo descrito é conhecido na literatura como o fenômeno de "anexação preferencial" (*preferential attachment*).

Ambos os mecanismos citados, o processo de crescimento contínuo a que a rede é submetida e a não simetria dos vértices ocorrida através da "anexação preferencial" ajudam a explicar os concentradores, segundo os quais os novos nós

tendem a se conectarem aos vértices com maiores graus (fenômeno conhecido como *"rich-get-richer"*), e também a conseqüente distribuição de graus resultante.

Os artigos coletados em [11, 27] demonstram que, até muito recentemente, praticamente todo esforço despendido no estudo de redes complexas voltou-se para a caracterização de mecanismos que expliquem o aparecimento de lei de potências e também de suas topologias, quase sempre relativamente a suas componentes conexas e estruturas semelhantes conforme [7, 8, 9, 10].

No entanto, no que diz respeito a redes de computadores como a Internet, uma vertente de pesquisa igualmente importante deveria dedicar-se ao estudo de novos algoritmos, especialmente distribuídos [6], que fizessem uso das características topológicas especiais das redes em busca de maior eficiência. Isto de fato vem ocorrendo, mas somente a partir dos últimos anos, como os trabalhos em [32, 33, 34].

Esses trabalhos dedicam-se quase exclusivamente ao estudo de heurísticas de natureza local que possam ser utilizadas para levar ao aparecimento global de estruturas importantes na rede. Algumas aplicações relevantes dessas estruturas são: a disseminação de informações, a imunização das redes no nível da camada de rede na pilha de protocolos, e também a busca em redes "peer-to-peer" desestruturadas.

Não obstante da importância desses avanços recentes, para que continue havendo progresso no sentido de buscarem algoritmos distribuídos apropriados à distribuição de graus segundo leis de potência, é necessário que se obtenham avanços na descrição mais detalhada da estrutura das redes. Além do conhecimento que já se tem em termos de componentes conexas, é agora necessário que busquemos conhecer com mais detalhes algumas particularidades daquelas componentes, fato motivador e objeto de nosso estudo.

Em particular, trabalharemos na caracterização de uma grandeza chamada Carga, que é sempre associada aos nós da rede. Tal propriedade aponta para a possibilidade de intervenção em algoritmos de roteamento de rede visando a uma distribuição mais eqüitativa não somente da Carga entre os nós, mas também do tráfego em cada canal quando se utilizam redes de "overlay" para aplicações "peer-to-peer".

Além disso, um conhecimento profundo de tal variável está intimamente relacionado à possibilidade de desenvolvimento de novos algoritmos para duas classes importantes de problemas distribuídos, a saber, o roteamento de mensagens e o estabelecimento de redes de "overlay" para aplicações "peer-to-peer". A propriedade Carga também está diretamente relacionada com a importância dos vértices no grafo, sendo uma métrica interessante, quanto maior a Carga, mais importante é o vértice.

1.2 Definição do problema

Seja G um grafo com n vértices. Iremos interpretar G como uma rede de computadores. Considerando que G é submetido a um tráfego uniforme, ou seja, todos os vértices enviam mensagens entre si. Estudaremos o número de rotas que passam por cada vértice, e que chamaremos de Carga L. Tal variável quando específica a um determinado vértice u é chamada de L_u.

Um dos elementos necessários para obtermos L_u será o conjunto de árvores de distâncias mais curtas dos vértices de G. Quando consideramos uma árvore específica, enraizada por um vértice v, por exemplo, chamaremos a mesma de T_v. As rotas, citadas acima, são equivalentes aos menores caminhos entre dois vértices, ou seja, são obtidas justamente através das árvores de distâncias mais curtas.

Outra variável que irá nos ajudar em tal estudo, é o número de descendentes de um vértice em uma árvore de distâncias mais curtas. Por exemplo, chamaremos a descendência do vértice u na árvore T_v de D_u^v.

O número de descendentes de u, ou D_u^v, é equivalente ao número de caminhos mais curtos (rotas) passando por ele quando v, a raiz da árvore, envia mensagem para todos os outros vértices. Uma outra forma de visualizar seria considerarmos a árvore de v, e então estarão passando o número de mensagens referente a quantidade de vértices em que u está no meio do caminho entre v e os outros vértices. Ou também, o número de vértices que estão "abaixo" de u na árvore, ou seja, dependem dele para receberem mensagens.

Para obtermos a Carga L_u devemos considerar todas as n árvores de distâncias mais curtas referentes aos vértices de G, e então considerarmos a soma das rotas (descendências) de u em cada uma dessas árvores, ou seja, teremos $L_u = \sum_{v=1}^{n} D_u^v$. Na verdade estamos fazendo a sobreposição de todas as árvores de caminhos mais curtos de G.

Tal fato vem justamente através da observação de que quando cada um dos vértices, específico, estiver enviando as mensagens para todos os outros, ele enraiza uma árvore de distâncias mais curtas, e claramente a descendência de todos os vértices em tal árvore contribuirá com uma parcela em suas respectivas Cargas (a raiz envia mensagens pelas arestas de sua árvore de caminhos mais curtos). A Carga é a soma do número de rotas que passam por um determinado vértice, que é equivalente ao número de vezes que o vértice está no meio do caminho entre outros dois vértices quaisquer, sua descendência.

Ao considerarmos o envio através de todos os vértices, e suas respectivas árvores, e as descendências dos outros vértices nelas, ao efetivarmos a sua soma, tem-se justamente a Carga total do vértice.

Para exemplificar os conceitos supracitados, consideremos a figura 1.1. Na primeira parte (letra a) podemos visualizar um grafo G, que interpretaremos como uma rede, e o vértice v enviando mensagens para todos os outros vértices. Mais especificamente, podemos visualizar claramente o menor caminho entre v e o vértice u.

Quando v envia mensagens para todos os outros vértices, conforme a definição de tráfego uniforme, devemos considerar a árvore de menores caminhos enraizada por ele, no qual trafegam suas respectivas mensagens, conforme ilustrado na mesma figura na segunda parte (letra b), onde as arestas do grafo que não pertencem a árvore estão pontilhadas. Temos em tal figura as rotas de v para todos os outros vértices.

Pela definição de Carga, considerando um outro vértice específico u, a parcela referente a L_u quando v envia mensagens é dada pela descendência de u na árvore de v, ou D_u^v, conforme exibido na terceira parte (letra c) da figura. Essa

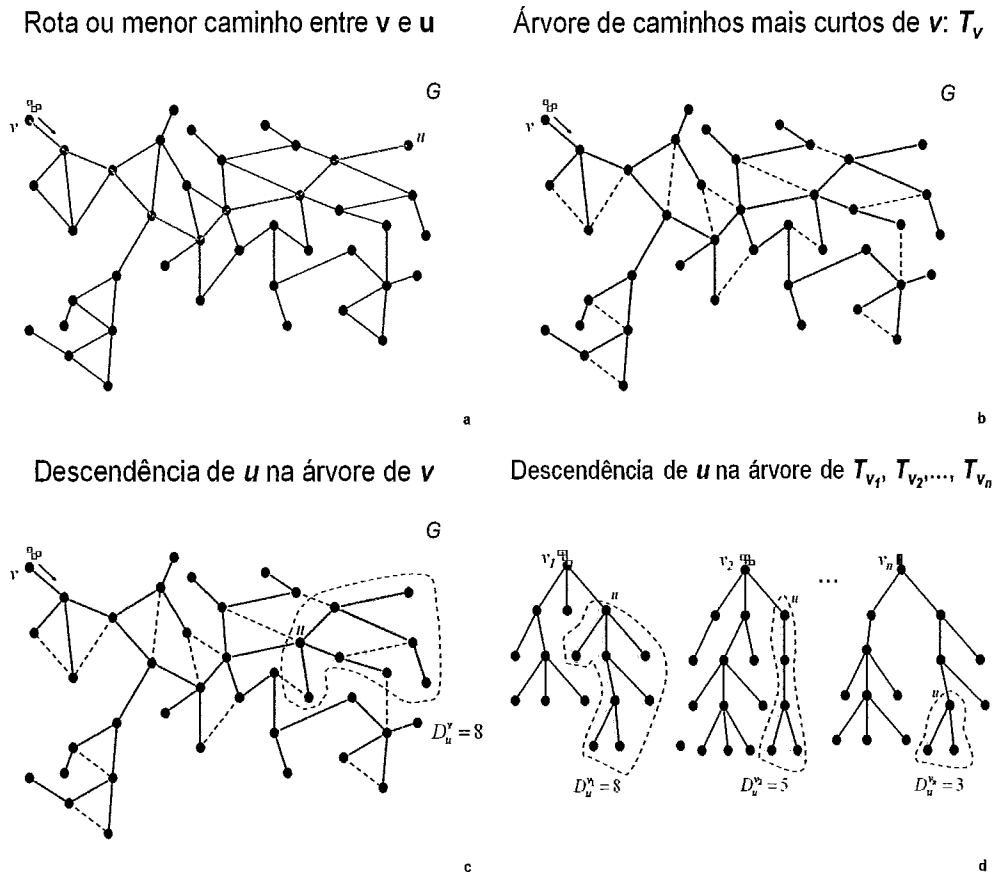

Figura 1.1: Exemplo de rota, árvore, descendência e carga.

descendência evidência o número de rotas que passam por u quando v envia mensagens (área em cinza na figura).

Considerando que desejamos obter a Carga de u, precisamos então levar em conta o envio de mensagens por todos os vértices, e então todas as suas árvores e suas respectivas descendências, mais especificamente as descendências de u em cada uma dessas árvores.

Para ilustrar tal conceito podemos ver na última parte (letra d) da figura as árvores de $v_1, v_2, ..., v_n$, onde trafegam as mensagens de tais vértices quando eles fazem suas disseminações, e então obtermos a descendência de u em cada uma dessas árvores, que caracteriza a Carga L_u propriamente.

No caso específico do nosso exemplo, as descendências de u nas árvores T_{v_1}, T_{v_2}, $..., T_{v_n}$ valem respectivamente $D_u^{v_1} = 8, D_u^{v_2} = 5, ..., D_u^{v_n} = 3$, e devem ser somadas conforme a definição de Carga.

Dado que está clara nossa problemática, queremos considerá-la quando o grafo G é um grafo aleatório, ou seja, é gerado através de um processo estocástico. Seus graus são amostrados de uma distribuição de probabilidades e a estrutura topológica do grafo é instanciada através de um processo aleatório, tipicamente um conjunto de sorteios.

Mais especificamente, a distribuição da qual serão amostradas os graus dos vértices de G é uma lei de potências, distribuição atualmente considerada como sendo uma boa aproximação para os graus dos vértices em redes como a Internet.

Iremos então considerar a distribuição α, como sendo uma lei de potências, e α_k como sendo a probabilidade de um vértice escolhido aleatoriamente ter grau k segundo esta distribuição. Tal distribuição tem importantes características, como por exemplo, a independência de escala [5].

Complementarmente, devemos considerar os graus dos vértices nas árvores de distâncias mais curtas, como por exemplo T_v, que baseados em um processo aleatório também seguirão uma distribuição de probabilidades, chamemos A, e sendo A_k como a probabilidade de um vértice aleatório ter grau k em T_v.

Para desenvolvermos nossas análises, utilizaremos de uma aproximação contínua através de um método baseado em uma fita, análogo a uma busca em largura em um grafo aleatório, e que iremos descrevê-la mais adiante.

1.3 Metodologia experimental

Em todas as simulações iremos precisar gerar os grafos para efetuarmos alguma computação sobre os mesmos. Mais especificamente, precisaremos gerar grafos com

[5] Sugerimos para maiores detalhes sobre a distribuição do tipo lei de potências e suas propriedades, a leitura do apêndice A.

uma seqüência de graus amostrada de uma distribuição pré-determinada, que em nosso caso é do tipo lei de potência.

Então, o procedimento básico para geração consistirá na amostragem de graus da distribuição, efetuando tal procedimento para cada um dos n vértices. Então, iremos conferir se a distribuição é factível de ser utilizada (se a seqüência de graus soma um número par).

Iremos utilizar para construção dos grafos nas simulações o método da urna, que é um dos mais populares e de simples manipulação.

Inicialmente, consideremos uma urna U. Seja um vértice qualquer v, por exemplo, que possui grau d_v. Deverá existir dentro da urna d_v bolas (ou cópias) rotuladas com v, referentes a tal vértice. Para construirmos G devemos sortear pares de bolas e inserirmos arestas entre os vértices relativos aos rótulos daquelas bolas.

Além disso, para as simulações consideraremos sempre as médias sobre um número específico de rodadas, dependendo da variável aleatória que desejamos estimar e o custo computacional para obtê-la. Sempre indicaremos quantas rodadas foram efetuadas para geração de determinado gráfico. Vale enfatizarmos que para mesmas variáveis, mas com parâmetros diferentes, como por exemplo, um parâmetro referente a distribuição, sempre utilizaremos um mesmo número de rodadas e simulações que sejam comparáveis.

1.4 Organização da tese

No capítulo 2, iremos desenvolver um estudo analítico sobre as árvores de distâncias mais curtas considerando, a priori, a distribuição dos graus do grafo G, ou seja, α como sendo do tipo lei de potência. Focaremos na obtenção da distribuição dos graus em tais árvores, ou A. Para confirmarmos nosso desenvolvimento, iremos também realizar simulações sobre nossa proposta, que de forma geral mostrarão concordar com a tendência do modelo analítico.

Então, no capítulo 3 abordaremos a distribuição da descendência D_u^v, onde desenvolveremos uma expressão analítica com base na distribuição de graus nas

árvores de distâncias mais curtas A, obtida anteriormente. Também efetuaremos simulações e mostraremos que elas se aproximam de forma bastante razoável de nosso desenvolvimento.

No capítulo 4 iremos estudar a variável Carga L propriamente dita, baseado no desenvolvido nos capítulos anteriores, e analisaremos a adequação da utilização da variável descendência D_u^v para sua composição. Além disso, efetuaremos simulações, e baseados nas mesmas proporemos uma nova explicação para L, incluindo algumas equações.

Então, no capítulo 5 iremos fazer considerações a respeito dos trabalhos relacionados existentes na literatura, comentando alguns problemas, sugerindo possíveis correções e comparando com nossos resultados.

Finalmente, no capítulo 6 faremos um resumo de nossas contribuições e proporemos algumas sugestões para desenvolvimento futuro.

Complementarmente, no apêndice A apresentamos as distribuições do tipo lei de potência e uma de suas principais características, a independência de escala.

No apêndice B, com intuito de simplificarmos a leitura deste trabalho, apresentamos uma tabela resumindo as principais variáveis utilizadas.

Capítulo 2

Árvores de caminhos mais curtos

2.1 Estudo analítico

Para o desenvolvimento de um estudo analítico, devemos inicialmente considerar o procedimento de busca em largura (BFS da sigla em inglês). Uma busca em largura é um algoritmo de busca não informada que se inicia através de um nó chamado raiz e explora todos os seus vizinhos. Depois de esgotar toda vizinhança, para cada um dos vizinhos, explora seus respectivos vizinhos, continuando assim sucessivamente até que o grafo seja esgotado (ou da componente conexa à qual pertence a raiz). Utiliza-se uma estrutura de dados conhecida como fila para auxiliar no processo de busca, armazenando os vizinhos pendentes que devem ser explorados em cada instante de tempo do algoritmo.

Tal procedimento é conhecido como uma busca em largura, pois expande a fronteira entre vértices "conhecidos" e "desconhecidos" uniformemente. De fato, o algoritmo "descobre" todos os vértices com distância k da raiz, antes de descobrir os vértices de distância $k+1$. Tal procedimento gera também nosso objeto de interesse, que são as árvores de caminhos mais curtos [1].

Segue o algoritmo básico em pseudocódigo:

(i) Inserir o nó raiz s na fila q.

[1]Para maiores detalhes sobre tal resultado, consultar [14]. Além disso, iremos utilizar o termo menores caminhos, rotas ou distâncias como sinônimos.

(ii) Retirar o nó v da cabeça da fila q e examiná-lo, explorando todos seus vizinhos u. Para cada u ainda não visitado, faça:

- Inserir aresta (v, u) na árvore T_s.
- Colocar u no final da fila q.

(iii) Se q vazio, todo vértice do grafo foi examinado, retornar a árvore T_s.

(iv) Caso contrário, voltar ao passo (ii).

Utilizaremos um procedimento análogo ao descrito anteriormente, conhecido como método da fita [2, 22, 23], que além de gerar uma árvore em largura possibilita a exploração de algumas propriedades analíticas interessantes, através de uma aproximação contínua na busca sobre tal árvore (iremos nos aprofundar mais adiante).

Tal algoritmo também utiliza a idéia do método da urna [29], e depois de amostrar o grau de cada um dos vértices, para um vértice v, por exemplo, gera-se $d(v)$ cópias rotuladas para o mesmo. Então, para cada uma dessas cópias sorteia-se um número uniformemente no intervalo unitário $[0, 1]$, e associa-se o mesmo a sua respectiva cópia, sempre cada uma delas sendo tratadas individualmente. Este procedimento é repetido para todos os vértices.

Então, devemos considerar uma fita ϕ, limitada no intervalo unitário, onde iremos inserir em ordem crescente todas as cópias dos vértices (análogo ao processo de inserção das bolas dentro da urna). Tal fita segue o padrão conforme a figura 2.1.

Podemos visualizar na figura as cópias inseridas na ordem crescente de seus rótulos, que aparecem abaixo do mesmo. Iremos percorrer a fita da direita para a esquerda, ou seja, em ordem decrescente. Por exemplo, o primeiro vértice a ser considerado será o que chamamos de A, pois existe uma cópia referente ao mesmo que tem o maior rótulo da fita, de valor $0,99$. Logo após, temos o vértice D, com sua cópia com rótulo $0,98$, e em seguida o vértice C com sua cópia de rótulo $0,97$, e assim sucessivamente.

Assim, percebemos que existirá na fita tantas cópias quanto forem os graus dos vértices. Por exemplo, um vértice B com grau 2, terá duas cópias relativas a ele na fita. Podemos visualizar suas cópias com os rótulos $0,92$ e $0,01$ na figura.

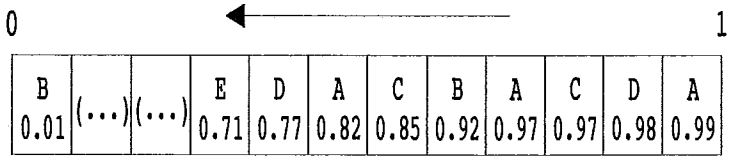

Figura 2.1: Exemplo com padrão de uma fita unitária.

Antes de apresentar o algoritmo que irá ser executado sobre a fita, devemos tecer algumas considerações básicas sobre o mesmo. O algoritmo tem como objetivo a execução de uma busca em largura sobre um grafo aleatório baseado numa seqüência de graus amostrada de uma distribuição.

Como o grafo é aleatório, ou seja, obviamente o mesmo não existe antes da fase de amostragem de seus graus, então necessariamente ele deverá ser construido em algum momento da execução do algoritmo, necessariamente após tal fase.

Devemos atentamente observar que iremos executar dois procedimentos simultaneamente, a construção do grafo baseado na sequência de graus amostrada de uma distribuição, e a construção da árvore de busca em largura sobre o grafo que está sendo gerado.

De fato, usaremos a ordem das cópias dos vértices na fita para formação da ordem dos vértices na fila da busca em largura. Tal fato será útil para futuramente efetuarmos nossa análise como uma aproximação contínua no tempo de execução da busca. Esperamos que tal fato fique mais claro quando efetuarmos a análise do algoritmo.

Somente como uma observação, uma questão pertinente que se coloca é porque utilizar tal artífício e não o método clássico da urna, dada as similaridades entre os mesmos?

Ao efetuarmos o procedimento de amostragem dos graus, inserimos cópias referentes aos seus graus na fita, e de forma análoga ao método da urna poderíamos diretamente construir um grafo baseado no mesmo.

Poderíamos fazer isso selecionando as cópias duas a duas, ou seja, devemos pegar todos os vizinhos de forma pareada e inserimos uma aresta entre os mesmos. Assim, teríamos através da ordem das cópias na fita um análogo ao sorteio das bolas da urna. Para ficar mais claro, no caso de nossa fita ϕ do exemplo da figura anterior, teríamos algumas arestas como $\{(A, D), (C, A), (B, C), (A, D)\}$.

Então, após o procedimento de amostragem e de construção do grafo através do pareamento, poderíamos executar o procedimento clássico de busca em largura para construção da árvore. Este seria um caminho perfeito para geração do grafo aleatório baseado na distribuição de graus e a construção de sua árvore da busca em largura, análogo ao efetuado quando utilizamos o método da urna.

Porém, devemos observar que utilizamos de forma diferente, pois temos como objetivo utilizarmos a ordem em que as cópias estão colocadas na fita como referência para a busca em largura. Novamente, fazemos isso pois iremos aproveitar tal fato para efetuarmos uma aproximação contínua, fazendo um paralelo dessa ordem das cópias com relação ao tempo de execução do algoritmo de busca. Poderemos fazer durante a análise um tratamento relativo ao tempo em que o vértice é anexado à árvore, utilizando uma integral, que será importante para o desenvolvimento.

A busca será feita sobre as cópias referentes aos vértices no grafo (cada vértice tem tantas cópias quanto for o seu grau), diferentemente da forma clássica onde são considerados os vértices propriamente.

Feita as considerações gerais a respeito do algoritmo, devemos considerar suas variáveis. Temos uma variável t, que representa o tempo referente a execução da busca, e que justamente possibilita a aproximação contínua do mesmo. O tempo vale inicialmente $t = 1$, e no final do procedimento $t = 0$, similar (paralelo) ao intervalo unitário em que a fita está limitada.

Cada vértice tem uma variável relativa ao seu grau em G, que foi amostrado da distribuição, e chamaremos, por exemplo, o grau em G de v por $d(v)$.

Além disso, cada vértice compartilha com suas cópias uma variável chamada *grau residual*, que armazena quanto falta para o vértice ser completamente saturado, obedecendo assim, à realização da distribuição de graus original da qual foi amostrada. Chamaremos o grau residual do vértice v, por exemplo, de $d_{res}(v)$. Em $t = 1$, temos $d_{res}(v) = d(v)$, ou seja, de forma lógica todo vértice inicia o procedimento com seu respectivo grau residual igual ao seu grau amostrado.

O algoritmo também contém uma estrutura de dados auxiliar $q()$, uma fila do tipo fifo (*first in first out*) que armazena as cópias correntes a serem exploradas em determinado instante de tempo da execução do algoritmo. Tal fila será preenchida de maneira análoga ao cômputo de uma busca em largura em um grafo arbitrário qualquer.

Apesar da fila ter como parâmetro o tempo, o mesmo pode ser ignorado quando o algoritmo é executado, devendo ser tratado somente como uma fila normal, onde manipulamos sua cabeça, examinando e retirando elementos da mesma. Esse parâmetro tempo é útil somente como marcação para o desenvolvimento analítico, e iremos observá-lo mais adiante.

Também teremos um ponteiro p, que marca o próximo vértice a ser explorado na fita. Iniciamos apontando para a cópia de maior valor da fita ϕ, que se encontra mais à direita, considerando a orientação conforme a ilustração anterior.

Temos a variável *comp* representando o índice da componente conexa que está sendo explorada de forma corrente (no momento). Além disso, consideraremos a versão da árvore referente a uma componente conexa específica como sendo $T(comp)$, por exemplo.

Então já podemos introduzir o algoritmo, e posteriormente um exemplo para ilustrar.

Núcleo básico do algoritmo de geração e busca da fita:

- Pré-processamento: $comp = 0, t = 1$. Efetuar amostragem da seqüência de graus da distribuição de graus e inserção de cópias rotuladas na fita em ordem crescente pelos rótulos. Ajuste do ponteiro p apontando para cópia mais a direita na fita;

- Passo 0: Selecionar cópia referente ao vértice v apontando por p e inserir $d(v)$ cópias na fila em $q(t)$. Avançar p. Atualizar t;

- Passo geral:

  (a) Seja o vértice v na cabeça da fila $q(t)$. Enquanto $(d_{res}(v) = 0)$: remover v da cabeça de $q(t)$, atualizar t e considerar nova cópia que está na cabeça como sendo o novo v;

  (b) Seja o vértice w apontado por p na fita. Enquanto $(d_{res}(w) = 0)$: avançar p;

  (c) Inserir aresta (v, w) em G. Avançar p.

  (d) Se $(d_{res}(w) = d(w))$: inserir aresta (v, w) em $T(comp)$, inserir cópias referentes a w em $q(t)$ (as cópias são equivalentes ao grau amostrado menos 1, pois já foi gasto uma unidade no processo de anexação que acabou de ocorrer).

  (e) Remover cópia de v da cabeça da fila $q(t)$. Atualizar graus residuais. Atualizar t.

- Reinicialização: Caso $q(t) = \emptyset$, então incrementar $comp$ e recomeçar procedimento no passo 0 para componente $comp$, ajustando ponteiro p para próxima vértice v tal que $d_{res}(v) > 0$; Senão, repita o passo geral.

- Finalização: Ao final do processo devemos checar o tamanho de cada componente $comp$, obter a de maior número de vértices e retorná-la (chamaremos a mesma de componente gigante de G, iremos comentar mais a respeito posteriormente).

Dado o descrito anteriormente, ao final do algoritmo sabemos qual maior componente, *comp* por exemplo, e então teremos $T(comp)$ como sendo a árvore de distâncias mais curtas de tal componente.

Vamos considerar um exemplo, conforme uma fita ϕ. Consideremos o conjunto de vértices $V = \{u, v, w, x, z, y, l\}$. Seja a sequência de graus nesta ordem, amostrada como $(3, 2, 2, 1, 4, 2, 2)$.

Consideremos então uma sequência na fita, já ordenada conforme os rótulos dos vértices, como por exemplo: $(u, v, w, x, z, z, y, l, l, u, v, z, w, u, y, z)$. Existem tantas cópias quanto os graus de cada vértice. Iremos através da execução do algoritmo obter a seguinte seqüência de passos, descrita na tabela 2.1. Cada linha nesta tabela representa a configuração ao final de cada passo do ponteiro p, da fila, de sua cabeça, qual aresta foi inserida no grafo, e qual aresta foi inserida na árvore.

Passo	**Ponteiro p**	**Fila**	**Cabeça**	**Aresta do grafo**	**Aresta da árvore**
pré	u	()	-	-	-
0	v	(u,u,u)	u	-	-
1	w	(u,u,v)	u	(u, v)	(u, v)
2	x	(u,v,w)	u	(u, w)	(u, w)
3	z	(v,w)	v	(u, x)	(u, x)
4	z	(w,z,z,z)	w	(v, z)	(v, z)
5	y	(z,z,z)	z	(w, z)	–
6	l	(z,z,y)	z	(z, y)	(z, y)
7	l	(z,y,l)	z	(z, l)	(z, l)
8	l	(y,l)	y	-	-
9	u	(l)	l	(y, l)	–
11	u	()	-	–	–

Tabela 2.1: Exemplo de execução do algoritmo sobre a fita.

Assim, construímos a árvore referente a busca em largura com as arestas $T(G) = \{(u, v), (u, w), (u, x), (v, z), (z, y), (z, l)\}$, e o grafo com arestas $E(G) = \{(u, v), (u, w), (u, x), (v, z), (w, z), (z, y), (z, l), (y, l)\}$.

Com objetivo de iniciarmos nossa análise em relação a distribuição de graus em árvores de caminhos mais curtos, devemos agora introduzir as seguintes variáveis aleatórias:

- α_i: probabilidade de um vértice v ter grau i no grafo G. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_G(v) = i)$.

- A_{m+1}: probabilidade de um vértice v ter grau $m+1$ na árvore de distâncias mais curtas T. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_T(v) = m+1)$.

- $\rho_{i,m}$: probabilidade de um vértice v ter m filhos em uma árvore em largura T, dado que o mesmo tem grau i em G. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_T(v) = m | d_G(v) = i)$.

- $\rho_{i,m}(t)$: probabilidade de um vértice v ter m filhos em uma árvore em largura T, dado que o mesmo tem grau i em G e foi anexado na árvore no tempo t, ou seja, sua cópia de maior índice na fita vale t. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_T(v) = m | d_G(v) = i, \max\{v\} = t)$.

- $\mu_{t,i}$: probabilidade para o valor do máximo entre cópias de um vértice v, sendo que o mesmo tem i cópias sorteadas uniformemente no intervalo unitário, ser igual a t. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(\max\{v\} = t \mid d_G(v) = i)$

Temos como objetivo a obtenção da distribuição de A_{m+1}, que justamente representa o número de filhos em uma árvore de caminhos mais curtos de um vértice[2]. Temos baseados no método da fita [2] a seguinte equação básica:

$$A_{m+1} \approx \sum_{i=m+1}^{\infty} \alpha_i \rho_{i,m} \tag{2.1}$$

[2]Devemos notar que o grau do vértice é $m+1$, sendo a primeira parcela m referente aos seus filhos na árvore, e a parcela unitária representando a aresta em que ele foi anexado à mesma.

que representa a multiplicação da distribuição marginal de α_i com a condicional $\rho_{i,m}$, originada através de trivial identidade [3].

Devemos notar que essa equação é aproximada, pois não consideramos a raiz da árvore, que não é anexada por outros vértices, ou seja, por definição não tem pai na mesma. A equação justamente leva em conta que os vértices são anexados a árvore de caminhos mais curtos, e com a raiz isso não ocorre. Apesar disso, a equação é uma boa aproximação, pois trabalharemos com grafos grandes onde um vértice dentro de todo o universo pode ser desconsiderado ($\lim_{n\to\infty} 1/n = 0$).

A distribuição a_i já é instanciada e não pode ser manipulada (conforme já definido na introdução como sendo uma lei de potências), porém a variável $\rho_{i,m}$ ainda não está completamente definida. Devemos então considerá-la através de simples interpretação, que podemos condicioná-la no tempo em que o vértice entra na árvore, e temos a equação:

$$\rho_{i,m} = \int_{t=0}^{1} \rho_{i,m}(t)\mu_{t,i}dt \tag{2.2}$$

similar a identidade anterior [4], porém agora considerando a distribuição do tempo que o vértice pode ser anexado na árvore, que representa o intervalo da fita unitário em que ele está inserido (através de suas cópias).

Devemos fazer uma pausa para considerarmos o desenvolvimento da distribuição de $\mu_{t,i}$, o máximo de um vértice v com i cópias uniformemente sorteadas no intervalo unitário [5]. Chamaremos os rótulos dessas i cópias de $t_1, t_2, \ldots, t_i$. E novamente, todos os rótulos são referentes ao intervalo unitário, são identicamente distribuídos conforme uma distribuição uniforme.

Seja $\mu_{t,i} = \max\{t_1, t_2, \ldots, t_i\}$. Seja então a distribuição do máximo, sendo $P(\mu_{t,i} \leq t) = P(\max\{t_1, t_2, \ldots, t_i\} \leq t) = P(t_1 \leq t, t_2 \leq t, \ldots, t_i \leq t) = P(t_1 \leq$

[3] De forma geral, tal regra diz que $P(X) = \sum_{\forall Y} P(X,Y) = \sum_{\forall Y} P(X|Y)P(Y)$.

[4] Podemos mais especificamente considerar o desenvolvimento similar $P(X|Z) = \frac{P(X,Z)}{P(Z)} = \frac{1}{P(Z)} \int_{\forall y} P(X,Y,Z) = \int_{\forall y} \frac{P(X,Y,Z)}{P(Z)} = \int_{\forall y} \frac{P(X,Y,Z)}{P(Y,Z)} \frac{P(Y,Z)}{P(Z)} = \int_{\forall y} P(X|Y,Z)P(Y|Z)$.

[5] O desenvolvimento da distribuição do máximo é amplamente conhecido, e está contido em qualquer livro de probabilidades como em [30].

$t)P(t_2 \leq t) \ldots P(t_i \leq t) = [P(t_j \leq t)]^i$, tal que $1 \leq j$, ou seja, a função de probabilidade acumulada. Só para ficar claro, o j nessa expressão é um valor no intervalo $[1, i]$, dado que todas as cópias são identicamente distribuídas.

Sendo que para cada vértice no intervalo unitário a probabilidade $P(t_j \leq t) = t$, para um j qualquer [6]. Logo, $P(\mu_{t,i} \leq t) = t^i$. Precisamos, na prática, da função de densidade de probabilidades, logo derivamos em relação a t a densidade acumulada e obtemos $P(\mu_{t,i} = t) = it^{i-1}$.

Considerando tal desenvolvimento aplicados a 2.2, temos então:

$$\rho_{i,m} = i \int_{t=0}^{1} \rho_{i,m}(t) t^{i-1} dt \tag{2.3}$$

Para avançarmos com a variável $\rho_{i,m}(t)$, devemos analisar mais profundamente como um vértice de grau i, anexado na árvore no tempo t, anexa outros vértices na árvore (ou seja, como v cria arestas partindo dele). Para tal, devemos considerar a figura 2.2. Consideremos uma configuração onde a fila vale $q(t_0) = \{a, b, b, \ldots, u, u\}$ no instante t_0, e $q(t_1) = \{u, u, \ldots\}$ no instante t_1.

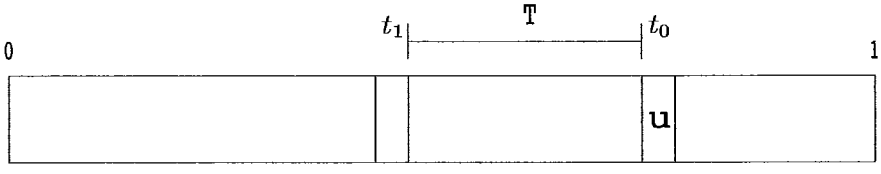

Figura 2.2: Exemplo de gap típico entre momentos de entrada em $q(t)$ e exploração do vértice.

Devemos notar que em t_0 o vértice u foi anexado na árvore, através de sua cópia de índice máximo, e conseqüentemente entrou no fim da fila q. Após o tempo $(t_0 - t_1)$, podemos observar que no topo da fila $q(t_1)$ temos u, logo devemos considerar quem serão seus vizinhos neste instante, ou de outra forma, quem ele irá anexar à árvore (também chamado na literatura de exploração do vértice, equivalente ao passo (ii) do algoritmo de busca em largura).

[6] Trivialmente obtido pois é proporcional a área na fita determinada por t_j. Em uma distribuição uniforme unitária U, temos $P(U \leq u) = u$. Em nosso caso t_j segue essa distribuição.

Podemos observar também que é possível u ser vizinho de outros vértices, e ter sido testado antes de t_1 e depois de t_0 (pois pode existir cópia de u localizada neste trecho da fita, entre t_0 e t_1), logo também devemos levar esta informação em consideração para obtenção de uma expressão mais precisa.

Então, baseados em tais observações, precisamos de uma forma para descrever: o *gap* existente entre o tempo de entrada na fila de um vértice e a respectiva exploração do mesmo; a existência do grau residual disponível em t_1, por exemplo, para que ele possa anexar outros vértices na árvore somente através do grau que ainda lhe resta, um grau residual positivo, e não o original, referente ao grau amostrado inicialmente, que não mais está disponível por completo e não seria razoável utilizá-lo. Além disso, devemos também contemplar o mecanismo de exploração do vértice, ou seja, uma expressão referente a anexação na árvore por parte de v de outros vértices (v criar arestas na árvore).

Assim, para tratar o *gap*, o grau residual disponível e a exploração do vértice, iremos introduzir as seguintes variáveis aleatórias:

- $Q(t)$: tamanho da fila no tempo t, ou seja, $P(Q(t) = q)$ é a probabilidade da fila ter tamanho q no tempo t.

- $\Delta_{i,t,q,r}$: probabilidade do grau residual de um vértice v ser r, dado que possui grau i no grafo G, entrou na fila no tempo t, sendo que o tamanho da fila em tal instante é igual a q. Utilizaremos a notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_{\text{residual}}(v) = r | d_G(v) = i, \max\{v\} = t, Q(t) = q)$.

- $p_{vis,q}(t)$: probabilidade referente à anexação na árvore, por parte de v, de um novo vértice, considerando-se que v entrou na fila em t e o tamanho dela é q.

- $\rho_{i,m,q}(t)$: probabilidade referente a um vértice v ter m filhos em uma árvore em largura T, dado que o mesmo tem grau i em G e foi anexado na árvore no tempo t, ou seja, sua cópia de maior índice na fita vale t, e o tama-

nho da fila em tal instante vale q. Utilizaremos à notação compacta proposta anteriormente, mas uma outra maneira de escrever a expressão seria $P(d_T(v) = m | d_G(v) = i, \max\{v\} = t, Q(t) = q)$.

Estamos em posição de avançarmos com o desenvolvimento de $\rho_{i,m}(t)$, e então iremos relaciona-lo com as novas variáveis, temos:

$$\rho_{i,m}(t) = \sum_{q \geq 0} \rho_{i,m,q}(t) P(Q(t) = q) \tag{2.4}$$

que representa identidade conforme a argumentação anterior, relacionando a distribuição marginal da fila $P(Q(t) = q)$ com a condicional $\rho_{i,m,q}(t)$. Na verdade contemplamos através desta expressão o *gap* entre a entrada do vértice na fila e a exploração do mesmo.

A distribuição de $Q(t)$ já é uma marginal e podemos considerá-la determinada (iremos comentar mais adiante), porém a variável $\rho_{i,m,q}(t)$ ainda não está completamente definida.

Devemos então considerá-la através de sua definição, ou seja, a distribuição de probabilidades de um vértice anexar na árvore m outros vértices, dado que possui grau i em G, foi anexado a árvore em t e a fila tem tamanho q. Temos a expressão

$$\rho_{i,m,q}(t) = \sum_{r=m}^{i-1} [\text{binomial}(r, p_{vis,q}(t)) = m] \Delta_{i,t,q,r} \tag{2.5}$$

que consiste em um condicionamento sobre os possíveis valores dos graus residuais.

Basicamente se fez uma marginalização sobre a distribuição de graus residuais $\Delta_{i,t,q,r}$, considerando-se uma distribuição binomial [7] parametrizada com r tentativas (referente ao total do grau residual) e a probabilidade de anexar um vértice

[7] A distribuição binomial é uma distribuição de probabilidades discreta do número de sucessos de uma seqüência de n experimentos que podem ser bem sucedidos ou não, com probabilidade de sucesso p. Falamos então que a binomial é parametrizada através de n e p, e desejamos contar o número de sucessos k, nos referindo a mesma como binomial$(n, p) = k$. Não iremos utilizar sua forma extensa, por fins de legibilidade, porém sabemos que Prob(binomial$(n, p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

como sendo $p_{vis,q}(t)$, considerando que se deseja obter m anexações de vértices (ou como chamado, sucessos).

Tal expressão é uma aproximação, dado que as tentativas não são exatamente iguais, pois existe um pequeno deslocamento na fita. Contudo, consideraremos aproximadamente as mesmas baseados na forma suave de Q, que iremos examinar mais adiante.

Para avançarmos devemos considerar a distribuição de graus residuais $\Delta_{i,t,q,r}$. Para tal, aplicamos raciocínio similar ao cálculo do máximo, ou seja, sobre a área na fita unitária, que é proporcional ao número de cópias que serão saturadas no intervalo da fita.

Observamos que o intervalo entre o tempo de anexação e exploração do vértice é $t - q * acc$, sendo acc uma constante normalização [8].

Se fixarmos o grau residual em r, utilizamos uma unidade do grau total i de v em G para anexá-lo a árvore, devemos então saturar $i - 1 - r$ cópias no intervalo descrito, e conseqüentemente deixaremos um excedente de r cópias para serem utilizadas após t, exatamente conforme o desejado. A expressão é facilmente calculada e ficaria como:

$$\Delta_{i,t,q,r} = (t - q * acc)^{i-1-r}(q * acc)^r \tag{2.6}$$

sendo acc uma constante, a se determinar.

O limite superior de $q*acc$ é determinado pelo tamanho da fita, que vale 1, logo teremos aproximadamente $q * acc = 1$. Assim, $acc = 1/q$, e em seu caso extremo, a fila q vale aproximadamente a soma dos graus dos vértices ($\sum_{\forall v} d_G(v)$), e logo temos aproximadamente $acc = \frac{1}{\sum_{\forall v} d_G(v)}$.

[8] Necessitamos de uma constante de normalização acc, pois apesar do tamanho da fila q representar o *gap* relativo a entrada do vértice na fila e a sua exploração propriamente dita, o mesmo é um número inteiro e ao trabalharmos em cima da fita estamos no intervalo $[0, 1]$. Logo precisamos nos posicionar dentro dela de maneira conveniente, proporcional a q, mas respeitando a restrição quanto ao seu intervalo unitário.

Este valor é aproximado [9], pois a soma dos graus dos vértices da expressão, considerando que os mesmos são distribuídos seguindo uma lei de potência tem variância infinita, por definição, e a soma de variáveis com variância infinita resulta em uma variável com variância também infinita, e então seria problemático obtê-la de forma simples. Porém, a princípio, tal variável pode ser aproximada através desta identidade.

Além dos graus residuais, devemos considerar a distribuição $p_{vis,q}(t)$. Para tal, podemos sugerir a aplicação da seguinte identidade trivial:

$$p_{vis,q}(t) = \overrightarrow{p_{vis}}(t - q * acc) \tag{2.7}$$

que simplesmente faz a troca da variável, chamando-a agora de versão "instantânea", já considerando no argumento da função o deslocamento referente ao tamanho da fila, conforme aplicado anteriormente.[10].

Agora temos a expressão $\overrightarrow{p_{vis}}(t)$, que representa a distribuição de probabilidades referente a anexação, por parte de v, de um novo vértice, considerando para tal a exploração exatamente no tempo t, independente da fila, do tempo de entrada na mesma, e de seu grau.

Para que seja efetuada uma anexação à árvore por parte do vértice v no instante de tempo t, devemos considerar o vértice w localizado em t na fita, e que irá ser testado e deve ser inexplorado, ou seja, nenhum vértice pode ter anexado o mesmo previamente, antes de t (senão ele já estaria na árvore, e então poderia ser vizinho de v no grafo G, porém não na árvore).

Para tal, o mesmo deve possuir todas as suas cópias localizadas abaixo de t na fita (menores que t). Repare que de forma similar a parte do desenvolvimento da distribuição do máximo, se desejamos saber a probabilidade de uma cópia ser

[9]De fato, existe uma modelagem para *acc* utilizando teoria de filas, porém devido a mesma não ter solução fechada preferimos omití-la.

[10]Vale sublinharmos que a seta em cima do símbolo não tem relação alguma com a notação vetorial.

menor que t, temos Prob($cópia \leq t$) $= t$, proporcional a área na fita, ou ao intervalo correspondente $[0, t]$.

Se desejamos saber a probabilidade de se ter k cópias na fita menores que k, e sabendo que as cópias são sorteadas uniformemente e independentes, temos [Prob(k cópias $\leq t | d_G(v) = k) = t^k$]. Baseado em tal desenvovimento, podemos então sugerir a expressão:

$$p_{vis}^{\rightarrow}(t) = \sum_k \alpha_k t^k \tag{2.8}$$

que justamente faz a marginalização por todos os graus possíveis de w, e considera a probabilidade de w com grau k ter as cópias menores ou iguais a t.

Estamos agora em posição para propor uma expressão da probabilidade de um vértice v ter grau $m+1$ na árvore de menores caminhos baseado no desenvolvido

de 2.1-2.8, fazendo as respectivas substituições temos:

$$A_{m+1} \overset{(2.1)}{\approx} \sum_{i=m+1}^{\infty} \alpha_i \rho_{i,m}$$

$$\overset{(2.2)}{=} \sum_{i=m+1}^{\infty} \alpha_i \int_{t=0}^{1} \rho_{i,m}(t)\mu_{t,i}dt$$

$$\overset{(2.3)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \rho_{i,m}(t)t^{i-1}dt$$

$$\overset{(2.4)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \rho_{i,m,q}(t)P(Q(t)=q)t^{i-1}dt$$

$$\overset{(2.5)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \sum_{r=m}^{i-1} [\text{binomial}(r, p_{vis,q}(t)) = m]\Delta_{i,t,q,r}P(Q(t)=q)t^{i-1}dt$$

$$\overset{(2.6)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \sum_{r=m}^{i-1} [\text{binomial}(r, p_{vis,q}(t)) = m]$$
$$(t - q * acc)^{i-1-r}(q * acc)^r P(Q(t)=q)t^{i-1}dt$$

$$\overset{(2.7)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \sum_{r=m}^{i-1} [\text{binomial}(r, \overrightarrow{p_{vis}}(t - q * acc)) = m]$$
$$(t - q * acc)^{i-1-r}(q * acc)^r P(Q(t)=q)t^{i-1}dt$$

$$\overset{(2.8)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \sum_{r=m}^{i-1} [\text{binomial}(r, \sum_k \alpha_k (t - q * acc)^k) = m]$$
$$(t - q * acc)^{i-1-r}(q * acc)^r P(Q(t)=q)t^{i-1}dt$$

$$\overset{(binomial)}{=} \sum_{i=m+1}^{\infty} \alpha_i i \int_{t=0}^{1} \sum_{\forall q} \sum_{r=m}^{i-1} \binom{i-1}{r} (\sum_k \alpha_k (t - q * acc)^k)^m (1 - \sum_k \alpha_k (t - q * acc$$
$$(t - q * acc)^{i-1-r}(q * acc)^r P(Q(t)=q)t^{i-1}dt$$

A expressão 2.9 gerada através da cadeia de substituições anteriores pode ser vista como uma função $A_{m+1}(\alpha, Q, acc)$, pois depende desses três argumentos. A relação de dependência com α, que é a distribuição original de graus dos vértices de G e segue uma lei de potências, é simples de ser obtida, plenamente determinada.

Ela também depende das variáveis acc e $Q(t)$, que são distribuições de difícil obtenção. Iremos utilizar a versão empírica das mesmas [11].

2.2 Estudo computacional

Obtemos as medidas desta seção através de 5000 rodadas, para cada um dos valores de τ [12] escolhidos para a simulação, e respeitando nosso intervalo de interesse $[2, 3]$, representativo para distribuição de graus do tipo Internet. Efetuamos as simulações com grafos com número de vértices $n = 5000$.

Trabalharemos inicialmente com o valor de $\tau = 2.50$, que está exatamente no meio do intervalo, referente a distribuição de graus do grafo G [13], que conforme sabemos segue uma lei de potências parametrizada através de tal variável.

Então, no gráfico da figura 2.4, temos variados instantes de tempo, amostrados dentro do intervalo unitário. Notamos claramente no início do processo (valores maiores do tempo, como por exemplo, e aproximadamente, $t > 0,70$) a fila crescendo, então se estabilizando (valores intermediários do tempo, como $t = [0.30, 0.70]$), e depois voltando a diminuir (valores menores do tempo, como $t < 0,30$).

Esse movimento é intuitivo, razoável e esperado, para uma função desse tipo, onde inicialmente muitos vértices não explorados entram na fila, depois se inicia uma fase na qual eles vão sendo explorados e outros novos vão sendo colocados em seus lugares, e em seguida os vértices vão simplesmente esvaziando a fila, pois não existem mais vértices, em quantidade, inexplorados e com alto grau, que são justamente os que aumentam o tamanho da fila, pois são suas cópias que contribuem para tal.

[11] Desenvolvemos algumas propostas de modelagem para essas variáveis, como com cadeias de Markov e funções do tipo multinomial, porém em todos os casos a obtenção de expressões analíticas fechadas não foi possível.

[12] O parâmetro da distribuição do tipo lei de potência chamaremos de τ. Caso o leitor não esteja familiarizado com tal distribuição, sugerimos a leitura do apêndice A.

[13] Equivalente a distribuição α, que segue uma lei de potências e foi introduzida no primeiro capítulo.

Para entendermos esta última afirmação, devemos considerar a fórmula referente a lei de potências, que implica em um decaimento acentuado, existindo muitos vértices de grau 1, muitíssimos menos de grau 2, menos ainda de grau 3, e assim sucessivamente. Além disso, devemos intuitivamente notar que os vértices de graus maiores, e que conseqüentemente possuem muitas cópias, tem maior probabilidade de serem anexados na árvore mais cedo (valores de tempo maiores), pois justamente tem um número maior de cópias espalhadas por toda a fita.

Assim, no início do processo esses vértices entram na fila, e começam a anexar outros vértices na árvore. Depois de determinado instante de tempo, praticamente não existem mais vértices desse tipo, ou seja, a fila só irá diminuir anexando na árvore vértices de grau 1, 2 ou graus baixos, que não contribuem para o crescimento da fila, e somente por algum tempo com sua manutenção, e posteriormente sua inexorável diminuição.

Temos na figura 2.3 o gráfico para as simulações da expressão proposta para A_{m+1}. Obtemos as medidas simuladas para cada um dos valores de τ (respectivamente $\tau = \{2.00, 2.25, 2.50, 2.75\}$) .

A nova expressão apresentou valores próximos aos simulados, com maiores desvios na cauda da distribuição. Podemos explicar tais desvios devido ao truncamento necessário na soma das parcelas de menor magnitude, em que deixamos de somar uma quantidade grande de parcelas devido ao grande tempo gasto para obtenção destes valores analiticamente. O impacto para as parcelas de maiores índices (maiores valores de m) em relação aos menores é considerável, visto que trabalhamos com muitas casas decimais, e elas são mais sensíveis a erros nas mesmas.

Mais especificamente, levamos em consideração as 100 primeiras parcelas (valores referentes ao i da fórmula 2.9, de seu somatório mais externo, que variavam de $[1, 100]$), apesar dos valores das probabilidades amostradas serem relativas a uma distribuição com limite superior como o n da simulação. De fato, para os valores de A_{m+1} altos, como i variava de $[1, 100]$, valores de graus mais próximos de 100 acabaram constituindo-se (utilizando-se) de menos parcelas, prejudicando ob-

viamente suas estimativas. Por exemplo, ao estimar A_{99}, somente efetuamos dois somatórios referentes a $i = 99$ e $i = 100$, ou seja, uma aproximação muito pior do que quando consideramos valores menores m em que mais parcelas contribuiam.

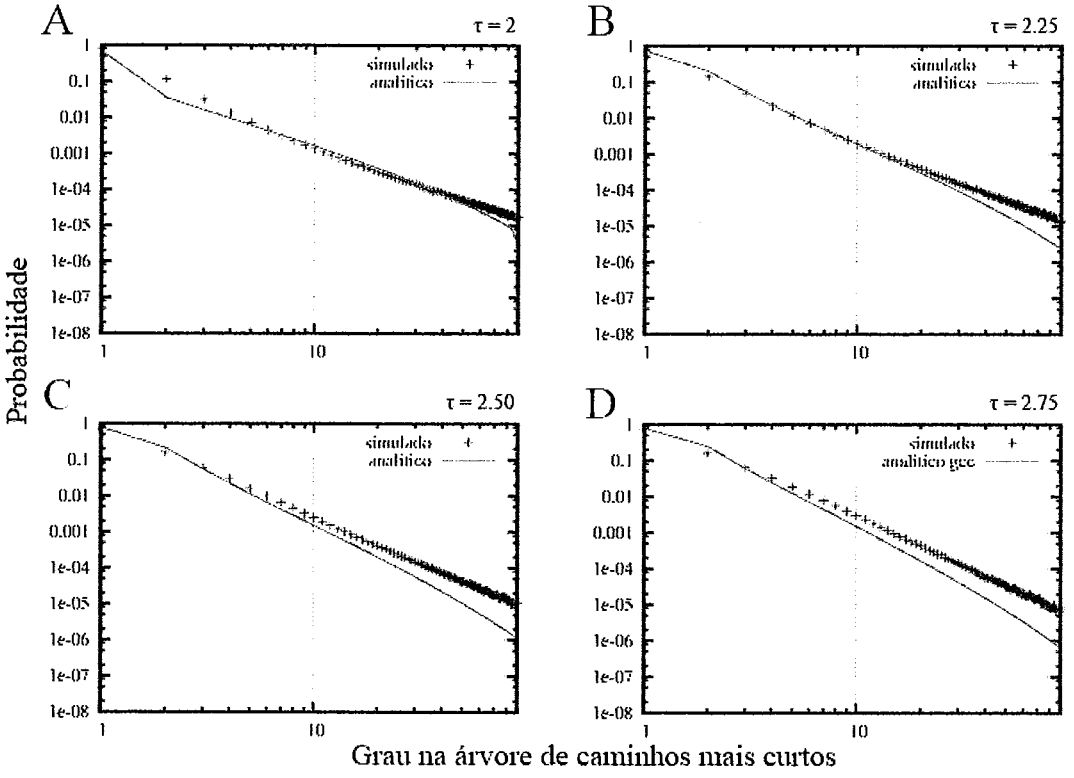

Figura 2.3: Distribuição de graus na árvore com o expoente variando, sendo $\tau = \{2.00, 2.25, 2.50, 2.75\}$ (grau x probabilidade).

Uma outra dificuldade é relativa a todos os somatórios, integrações e substituições decorrentes do desenvolvimento da expressão conforme 2.9, sendo necessário aplicar técnicas numéricas para sua resolução (por exemplo, para métodos de integração).

Uma outra questão bastante complexa é o cômputo das distribuições empíricas, e principalmente quanto a distribuição da fila $Q(t)$, auxiliar para o cômputo da distribuição de graus.

Vale adicionarmos uma nota referente ao cálculo de tal distribuição. Sua estimativa é muito complexa de ser calculada, pois o tempo t da expressão analítica,

que nos facilita em determinadas deduções, não existe do ponto de vista real da execução do algoritmo. Então, também tivemos que desenvolver uma aproximação para emular o comportamento de tal variável dentro do algoritmo, um tempo virtual, para que assim pudéssemos ter como amostrar as filas no intervalo unitário. Conforme já dito, desenvolvemos uma aproximação, que introduz erros e não é exatamente equivalente a fila $Q(t)$ teórica propriamente, ou seja, consideramos como uma das fontes de ruído.

Devemos observar também, que para valores de τ mais distantes de 2.00, os grafos são mais esparsos, e então possivelmente um maior número de rodadas poderia ajudar para obtenção de uma expressão mais precisa.

Apesar de todo o exposto, podemos considerar nosso desenvolvimento como uma boa aproximação do processo subjacente original.

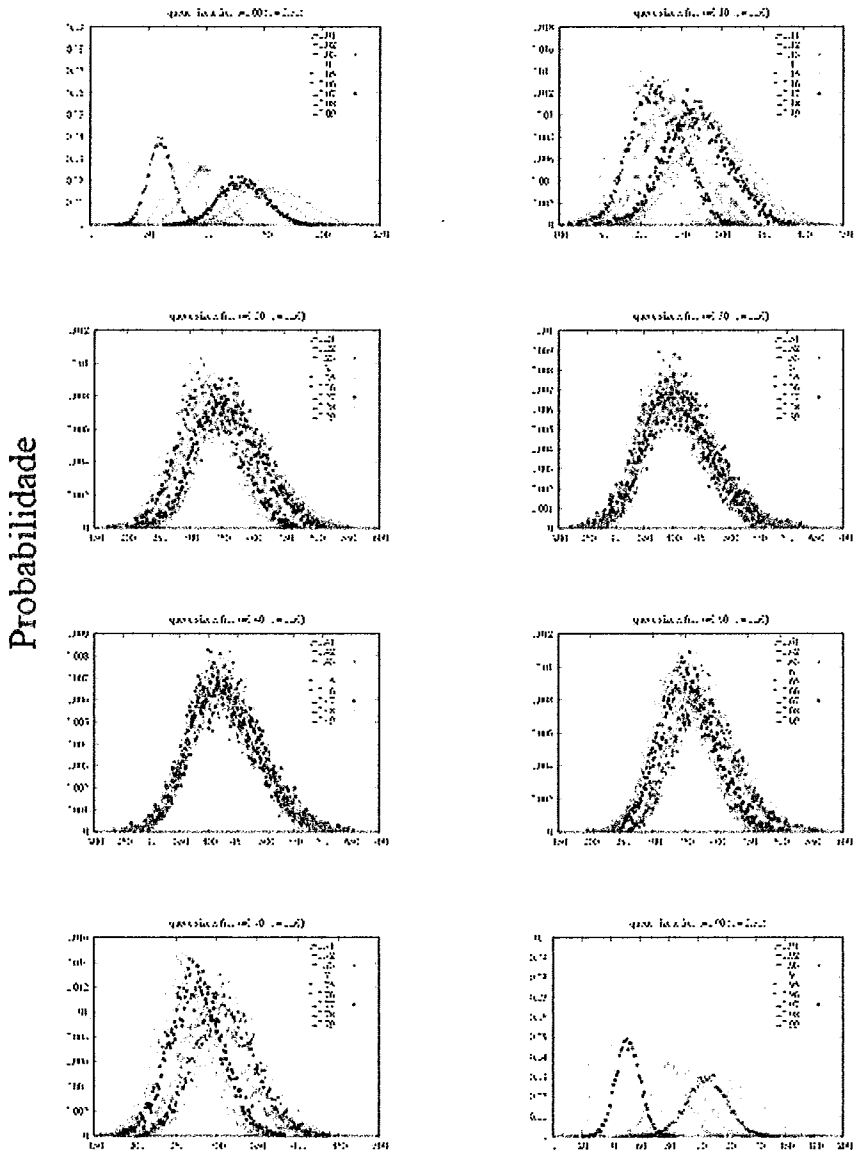

Figura 2.4: Estimativa da fila $Q(t)$ para t variando $[0, 1]$ e $\tau = 2.50$.

Capítulo 3

Distribuição da descendência

3.1 Estudo analítico

Neste capítulo iremos estudar a variável aleatória relativa ao número de descendentes de um vértice em uma árvore de distâncias mais curtas. Conforme nossa definição inicial, chamaremos a descendência do vértice x na árvore T_v de D_x^v.

Por exemplo, D_x^v considera o número de descendentes diretos (filhos) e indiretos (filhos dos filhos, recursivamente) de um vértice x em uma árvore de distâncias mais curtas T_v. Quando não precisarmos especificar o vértice e nem a árvore relacionada a tal variável, iremos simplesmente chamá-la de D.

O cálculo da distribuição da descendência é abordado em diversos formatos, entre eles utilizando *branching process* conforme [30]. Outra abordagem possível é uma versão modificada do problema da ruína do jogador com distribuições generalizadas, bastante utilizada com cadeias de Markov.

Existem também abordagens explícitas para o cálculo de D, utilizando expressões recursivas que decorrem diretamente da definição da mesma, porém não viáveis de se computar devido à explosão combinatória intrínseca à definição de tal distribuição. Uma forma mais adequada que iremos explorar é utilizando recorrência com funções geradoras.

Em [20] foi derivada a expressão da distribuição de Poisson Generalizada, da qual utilizaremos ferramental similar. Nosso objetivo não é considerarmos uma distribuição específica, conforme o desenvolvimento citado, mas uma forma generalizada, sem instanciação de uma distribuição pré-determinada.

Tal derivação então poderá ser utilizada para qualquer distribuição, principalmente quando se tem uma função geradora complicada, como em nosso caso, e então trabalharemos com o polinômio gerador em que seus coeficientes codificam (representam) as probabilidades.

Consideremos então uma árvore, um vértice i qualquer pertencente a mesma, e consideremos também que ele possui C_i filhos. A descendência do vértice i é uma variável chamada B_i. Cada um de seus C_i filhos tem descendentes conforme uma outra variável chamada B_{ij}. Então, podemos escrever uma expressão

$$B_i = 1 + \sum_{j=1}^{C_i} B_{ij} \tag{3.1}$$

que representa o número de descendentes de B_i como função dos descendentes de seus C_i filhos, que chamamos de B_{ij}, e adicionado ao número um, pois o próprio vértice está incluso em sua descendência. Ou seja, o número de descendentes de i vale 1, mais o número de descendentes de cada um dos C_i filhos do mesmo, ou seja, somatório das C_i parcelas de B_{ij}.

Agora podemos considerar uma recursão, onde B_i e B_{ij} são variáveis aleatórias baseadas na mesma distribuição, chamemos de B. Podemos também considerar que o número de filhos de i é uma variável aleatória chamada C_i.

Façamos uma pequena pausa para introduzirmos as funções geradoras de probabilidades.

Iremos considerar o caso simples, uma função geradora da variável aleatória X em relação a u, ou $G_X(u)$ [1], temos a definição:

$$G_X(u) = E[u^X] = \sum_{x=0}^{\infty} f(x)u^x \tag{3.2}$$

sendo $f(x)$ a função de densidade de probabilidades da variável X.

Considerada tal definição, desejamos obter a transformada da expressão de 3.1, conforme já exposto no início do capítulo, pois iremos trabalhar com a mesma para o cômputo da expressão final da descendência. Utilizaremos tal abordagem, pois na maioria dos casos as funções geradoras facilitam nas deduções, ou até mesmo viabilizam a mesma.

Podemos observar na expressão de B_i que constam duas parcelas bem definidas, um somatório bem delimitado, com C_i parcelas de variáveis B_{ij}, e a adição de tal parcela a uma constante, no caso o número um.

Mais especificamente, podemos identificar pelo descrito anteriormente, e em ordem de complexidade, três operações primitivas em relação a suas transformadas: a transformada de uma constante referente ao número um[2], a soma de duas variáveis referente a adição da constante ao somatório[3], e um somatório com parcelas bem definidas[4].

Assim, aplicando as três primitivas na expressão 3.1, considerando a transformada de B como sendo $G_B(u)$ e a transformada de C_i como sendo $G_C(u)$, obtemos a expressão:

$$G_B(u) = uG_C(G_B(u)) \tag{3.3}$$

[1]Poderia ser utilizada em relação à z, também conhecida como transformada z unilateral.

[2]Seja uma constante qualquer c. Para obtermos sua transformada devemos aplicar a definição da mesma, e temos $G_c(u) = E[u^c] = u^c$. Em nosso caso $c = 1$ e temos $G_1(u) = u$.

[3]Seja a soma de duas variáveis aleatórias independentes X e Y, e aplicando a definição de transformada temos $G_{X+Y}(u) = E[u^{X+Y}] = E[u^X] * E[u^Y] = G_X(u)G_Y(u)$, ou seja, a soma das variáveis aleatórias equivale a multiplicação de suas transformadas.

[4]Desejamos a transformada da soma de variáveis independentes e identicamente distribuídas W_i, que seguem W, limitadas por uma variável independente N, ou seja, $S_N = \sum_{i=1}^{N} W_i$. Aplicando a definição de transformada temos $G_{S_N}(u) = E[u^{S_N}] = E[u^{\sum_{i=1}^{N} W_i}] = E[E[u^{\sum_{i=1}^{N} W_i}]|N] = E[[G_W(u)]^N] = G_N(G_W(u))$.

que representa a transformada da variável aleatória B. Podemos observar claramente uma recorrência em tal expressão, onde a transformada de B aparece em ambos os lados.

De fato, desejamos obter as probabilidades da descendência na função $G_B(u)$, que são os coeficientes da mesma em sua representação como série de potências.

Para determinar tais coeficientes, devemos utilizar conforme [1, 35] a expansão de Lagrange, também utilizada por [20]. Tal resultado diz que, se $u = f(t)$, $f(0) = 0$, $f'(0) = 0$, e g uma função infinitamente diferenciável, então:

$$g(t) = g(0) + \sum_{k=1}^{\infty} \frac{u^k}{k!} [\frac{d^{k-1}}{dt^{k-1}} \frac{g'(t)t^k}{f(t)^k}]_{t=0} \tag{3.4}$$

Para utilizarmos tal identidade precisamos identificar quem são $f(t)$ e $g(t)$. Então façamos uma substituição conveniente, sendo $t = t(u) = G_B(u)$, aplicando a mesma em 3.3, então temos:

$$u = \frac{t}{G_C(t)} \tag{3.5}$$

Naturalmente sabemos que precisamos dos coeficientes de $G_B(u)$ (que agora vale t), logo é conveniente que façamos $g(t) = t$ (para obtermos através da expansão os coeficientes de sua representação em série de potências). Além disso, claramente u é uma função de t pela expressão 3.5, logo façamos $f(t) = u$.

Exatamente neste trecho devemos correlacionar o nosso problema original e a modelagem apresentada até aqui. Podemos reparar, claramente, a relação biunívoca de B com D, e de C_i com A_i, conforme definições anteriores.

Precisamos codificar as probabilidades oriundas do capítulo anterior para o número de filhos em uma árvore de caminhos mais curtos, em coeficientes para o polinômio gerador, então façamos $\beta_i = A_{i+1}$, para $i \geq 0$. Tal deslocamento aparece devido ao grau na árvore ser igual ao número de filhos (grau de saída) mais 1, que representa a aresta em que o vértice foi anexado a árvore.

Então, pela definição temos a função geradora (polinômio) do número de filhos de um vértice na árvore de distâncias mais curtas como $G_C(t) = \beta_0 t^0 + \beta_1 t^1 + \beta_2 t^2 + \ldots$.

Se observarmos que na expressão 3.4 existe um termo referente a $f(t)^k$, e observarmos também que $f(t)$ é uma função de uma série de potências, devemos ser capazes de efetuar tal cálculo, ou seja, fazer uma exponenciação de uma série de potências.

Temos uma identidade em [21] que sugere que a operação de exponenciação de uma série de potências com coeficientes constantes β_i, gera uma outra série de potências com coeficientes c_m, e que as mesmas se relacionam através da seguinte expressão:

$$c_m = \begin{cases} \beta_0^n & , \text{se } m = 0 \\ \frac{1}{m\beta_0} \sum_{k=1}^{m} (kn - m + k)\beta_k c_{m-k} & , \text{se } m \geq 1 \end{cases}$$

conseguindo assim eliminar a exponenciação de forma explícita. Poderemos visualizar tal fato ao efetuarmos a transformação $(\sum_{k=0}^{\infty} \beta_k t^k)^n = \sum_{k=0}^{\infty} c_k t^k$, constatando claramente que não existe mais exponenciação explícita na série resultante.

Então ao aplicarmos 3.5, com $f(t) = u = t/(a_0 t^0 + a_1 t^1 + a_2 t^2 + \ldots)$, em 3.4, temos:

$$g(t) = \sum_{k=1}^{\infty} \frac{u^k}{k!} \frac{d^{k-1}}{dt^{k-1}} \left[\frac{t^k}{(t/a_0 t^0 + a_1 t^1 + a_2 t^2 + \ldots)^k} \right]_{t=0} \tag{3.6}$$

Utilizando a identidade relativa a exponenciação, temos:

$$g(t) = \sum_{k=1}^{\infty} \frac{u^k}{k!} \frac{d^{k-1}}{dt^{k-1}} \left[\frac{t^k}{[t^k/(c_0 t^0 + c_1 t^1 + c_2 t^2 + \ldots)]} \right]_{t=0} \tag{3.7}$$

Simplificando t_k na expressão mais interna e arrumando a expressão, obtemos:

$$g(t) = \sum_{k=1}^{\infty} \frac{u^k}{k!} \frac{d^{k-1}}{dt^{k-1}} [c_0 t^0 + c_1 t^1 + c_2 t^2 + \ldots]_{t=0} \tag{3.8}$$

de onde precisamos extrair os coeficientes que acompanham u^k, que codificam as probabilidades da distribuição da descendência D.

Devemos agora considerar a parte mais interna da expressão gerada, $\frac{d^{k-1}}{dt^{k-1}} \sum_{\forall i} c_i t^i$. Para tal, devemos considerar a derivada de uma série de potências, que na verdade deve ser feita termo a termo, invertendo a ordem do somatório com o operador derivada. A idéia é fazermos:

$$\frac{d^{k-1}}{dt^{k-1}} \sum_{\forall i} c_i t^i = \sum_{\forall i} c_i \frac{d^{k-1}}{dt^{k-1}} t^i$$

uma versão mais simplificada, que nos permite facilmente desenvolvermos a expressão.

Então, desenvolvendo a mesma termo a termo, temos:

$$\sum_{\forall i} c_i \frac{d^{k-1}}{dt^{k-1}} t^i \mid_{t=0} = \begin{cases} 0 & , i < k-1 \\ c_i(i)(i-1)\ldots(i-(k-1)+1) & , i = k-1 \\ 0 & , i > k-1 \end{cases}$$

onde devemos observar que quando $i < k-1$, ao aplicarmos mais do que i vezes o operador derivada, iremos zerar a expressão. Quando $i > k-1$, a expressão será zerada ao substituirmos $t = 0$ da expressão mais externa de 3.8. A expressão somente não será nula quando $i = k-1$, e então teremos $c_i(i)(i-1)\ldots(i-(k-1)+1)$, e substituindo o valor de i temos $c_{k-1}(k-1)!$.

Vale notarmos por 3.8 que o valor de k no somatório mais externo varia, logo $\frac{d^{k-1}}{dt^{k-1}}$ também altera-se conforme k. Assim temos:

$$\begin{aligned} g(t) &= \sum_{k=1}^{\infty} \frac{u^k}{k!} c_{k-1}(k-1)! \\ &= \sum_{k=1}^{\infty} u^k c_{k-1} \frac{(k-1)!}{k!} \\ &= \sum_{k=1}^{\infty} u^k \frac{c_{k-1}}{k} \end{aligned} \tag{3.9}$$

onde notamos claramente que os coeficientes de cada termo u^k vale $\frac{c_{k-1}}{k}$.

Então, baseado no desenvolvimento de 3.9, podemos extrair seus coeficientes e escrever a distribuição da descendência D:

$$P[D = i] = \frac{c_{i-1}}{i}, i \geq 1. \tag{3.10}$$

Devemos observar que a expressão é uma função dos coeficientes c_i, que são baseados, conforme desenvolvimento anterior, na distribuição de graus na árvore em largura.

3.2 Estudo computacional

Dado o exposto anteriormente, executamos diversos experimentos para obtermos os valores simulados e analíticos da descendência. Considerando os custos computacionais envolvidos, efetuamos experimentos com 60000 rodadas e tamanho do grafo $n = 1000$.

Utilizamos os valores simulados dos graus de saída das árvores em largura, dada as restrições para o cálculo da versão analítica, conforme já visto, e temos então tal distribuição na figura 3.1.

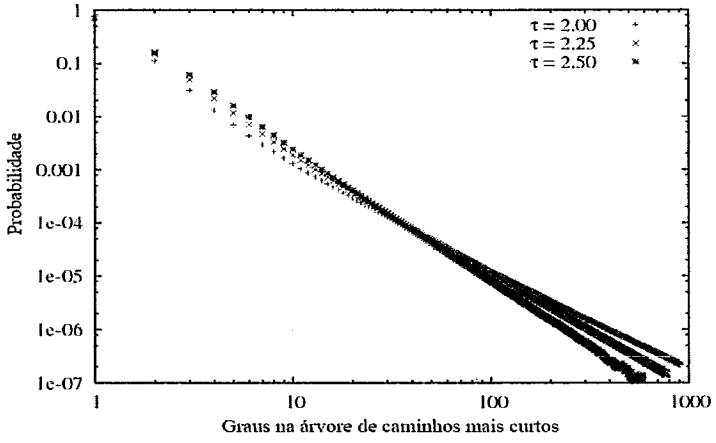

Figura 3.1: Distribuição simulada de A_{m+1} (grau x probabilidade).

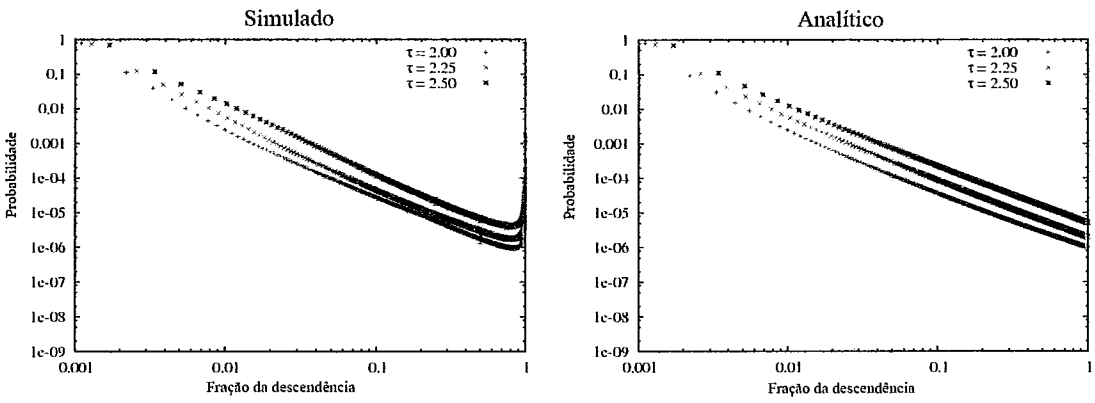

Figura 3.2: Tendências da descendência simulada (esquerda) e analítica (direita) (fração descendência/componente gigante x probabilidade).

Já na figura 3.2, exibimos as versões da descendência normalizada pela componente gigante [5], simuladas à esquerda e analíticas do lado direito, variando o expoente da distribuição de graus subjacente. Devemos observar que ambas seguem a mesma tendência, porém a versão analítica está ligeiramente acima da simulada.

Em 3.3 temos novamente a descendência, só que agora com a variância de sua medida simulada. Podemos notar que no final da mesma existe uma acentuada subida da função (na parte direita quase chegando a 1), o que não acontece na versão analítica. Tal fato pode ser justificado devido ao efeito da simulação finita, onde tal padrão, se estivéssemos trabalhando com um grafo infinito iria se diluir ao longo do tempo. A versão analítica já não possui tal indesejável característica em sua forma funcional.

Além disso, podemos notar claramente que elas seguem a mesma tendência, porém não estão exatamente sobrepostas. Logo, devemos observar que as estimativas estão dentro de um desvio padrão (a linha azul representa a versão analítica, a média da versão simulada esta realçada em verde, e os intervalos marcados pelas barras, em vermelho, delimitando um desvio padrão).

[5] O objetivo da normalização é minimizar o efeito de componentes gigantes diferentes, que podem ocorrer durante a versão simulada e não ocorrem na analítica.

Alguns motivos podem justificar a versão analítica se encontrar igual ou acima da média da versão simulada. Por exemplo, que o número de rodadas não foi suficiente para convergirmos para a sua média. Outra possibilidade é a variância ser alta demais, e então devemos reparar que os valores começam a divergir justamente quando a variância aumenta (barras vermelhas mais esticadas), sendo bastante acuradas quando a variância se mantém moderada.

Um outro motivo, bem mais delicado, diz respeito à dependência entre as variáveis responsáveis pela cascata, que podem atrapalhar tal desenvolvimento, dado o fato de ser uma das premissas mais fortes elas serem totalmente independentes. Devemos notar que talvez o grafo não seja grande o suficiente para uma aproximação desse tipo, porém, no caso ideal, ou seja, no limite, a aproximação poderia ser mais bem aplicada.

Além disso, do ponto de vista da versão analítica da descendência, conforme já dito, a mesma é obtida utilizando-se a distribuição A_{m+1} simulada, pois a versão analítica é difícil de ser obtida numericamente. Porém, a mesma também carrega erros, já que ela é só uma estimativa obtida através de um determinado número de rodadas.

Devemos observar também que, dado o formato de tais distribuições, elas não parecem ser, por inteira, exatamente uma lei de potências, porém sem dúvida são distribuições com caudas pesadas. Esse tipo de distribuição, independente de ser ou não uma lei de potências, tem muitos problemas quanto ao número necessário de amostras para que, de fato, uma estatística específica possa convergir.

Apesar de todos estes problemas, como já dissemos, a estimativa está dentro de um desvio padrão, o que é bastante razoável para uma aproximação, e consideramos satisfatório.

Adicionalmente, testamos a soma das variáveis descendência, a versão analítica, conforme sugerido ao final da introdução (soma das descendências de um vértice nas árvores de caminhos mais curtos, conforme a expressão do somatório de L_u) .

Claramente a variável relativa à soma das descendências convergiu para uma normal, conforme teorema central do limite [30], não concordando com seus valores

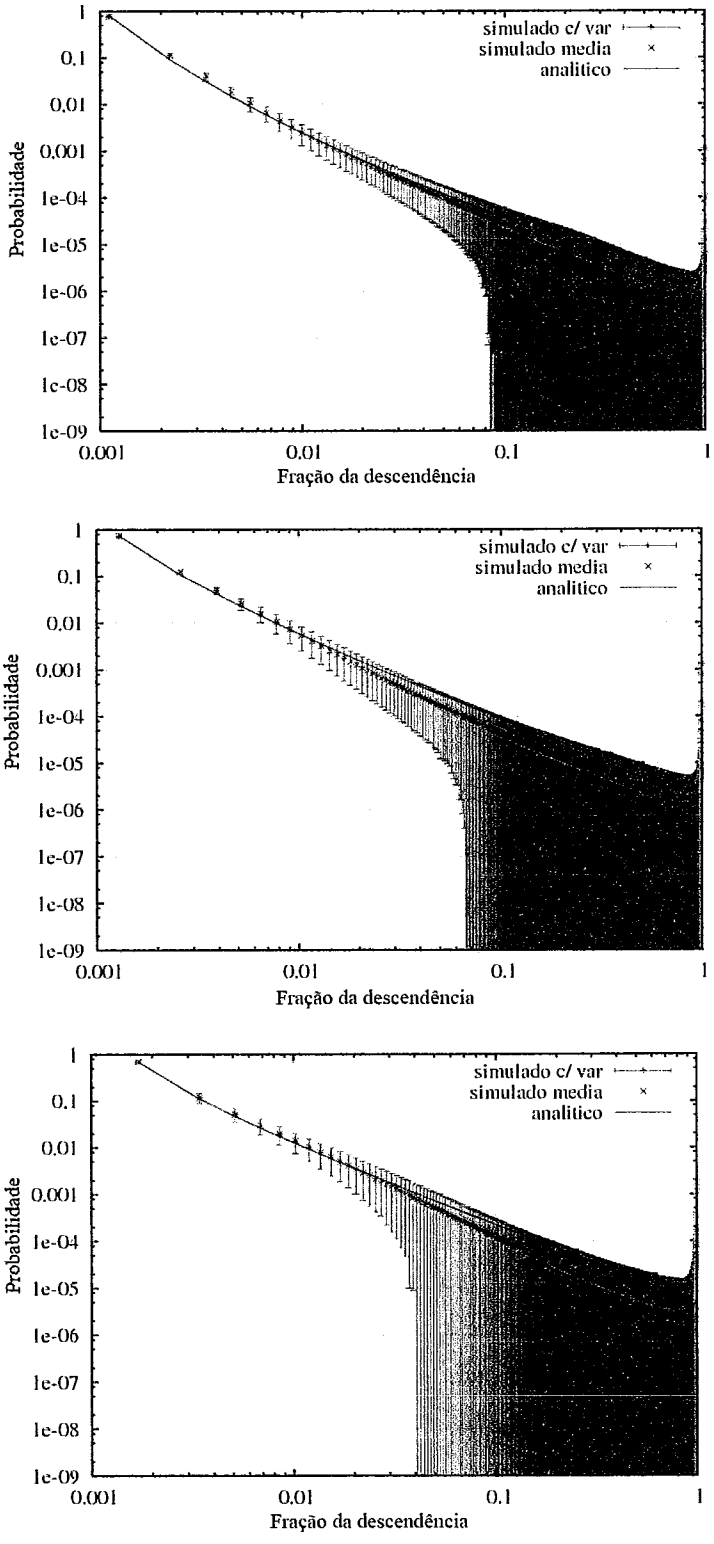

Figura 3.3: Descendência simulada x analítica com τ variável, sendo $\tau = \{2.00, 2.25, 2.50\}$, e desvio padrão (fração da componente gigante x probabilidade).

empíricos conforme desenvolvido no próximo capítulo, sendo um indício de que as parcelas possam realmente não serem independentes.

Vale notar que tal teorema, em sua versão mais conhecida, considera distribuições com variância finita, que não seria razoável de se aplicar para a distribuição de graus original (lei de potência), para a distribuição dos graus de saída, e também para a distribuição da descendência, todas elas possuindo nitidamente caudas pesadas.

Assim, talvez um estudo mais detalhado da estrutura de dependência entre as descendências seria necessário para avanços no cômputo da variável carga.

Capítulo 4

Distribuição da carga

4.1 Dificuldades do estudo analítico

O objetivo do presente estudo é caracterizar a distribuição da *Carga* em um vértice do grafo G que segue uma distribuição do tipo lei de potência para os graus de seus vértices.

Por hipótese, consideraremos a Carga quando a rede é submetida a um tráfego uniforme, ou seja, pacotes são enviados de um vértice i para um vértice j, para todos os possíveis pares (i, j). Para cada par específico, a mensagem é transmitida pelo menor caminho entre ambos e se houver mais de um menor caminho, os pacotes são divididos igualmente entre as possíveis rotas.

Com tal objetivo em vista, devemos considerar as árvores de caminhos mais curtos em G. Consideraremos também a distribuição de graus dessas árvores, ou seja, a distribuição A_{m+1} conforme obtida na seção 2.1.

Então, com as árvores de caminhos mais curtos e sua respectiva distribuição, valemo-nos do fato de que as descendências de um vértice em tais árvores representam o conjunto de rotas passando por ele quando a raiz da árvore estiver fazendo seu respectivo envio de mensagens para todos os demais vértices.

Na seção 3.1 consideramos a variável descendência, e mais especificamente temos a descendência de um vértice x na árvore T_v conhecida como D_x^v. Assim,

ao considerarmos o envio de mensagens por todos os vértices, devemos observar a descendência de x em todas as árvores, além da relativa ao vértice v. Cada descendência contribuirá com uma parcela para o cômputo da Carga L_x.

Mais especificamente, a parcela da Carga de x na árvore de T_v referente ao envio da raiz v vale D_x^v. Devemos considerar as descendências de x em todas as outras árvores, pois em última instância queremos $L_x = \lim_{n\to\infty} \sum_{i=1}^{n} D_x^i$.

Então, no passo subseqüente deveríamos considerar a composição clássica da distribuição da Carga através da soma das variáveis descendência. Porém, conforme observado de forma experimental no final do capítulo anterior, aparentemente tais variáveis não são independentes entre si, e provavelmente exista alguma estrutura de dependência entre elas.

Para analisarmos com mais atenção tal ponto, podemos escrever a descendência de diversas formas, explicitando vários enfoques sobre a mesma. Consideremos como exemplo a descendência de v_i[1] na árvore de v_k como D_i^k, ou de forma menos compacta $P(D_i^k \mid G)$, ou seja, a mesma variável porém deixando clara a dependência em relação a estrutura do grafo G.

Podemos também escrever tal variável como uma função $f(T_k, v_i)$, que explicita que a descendência é uma função de uma árvore e do vértice que se deseja observar, ou na mesma linha $f(A(G, v_k), v_i)$, explicitando assim que a árvore T_k é uma função (algoritmo) de G considerando-se sua raiz (ficando claro também que $T_k = A(G, v_k)$). Finalmente, podemos expressá-la como $h(G, v_k, v_i)$, explicitando todos seus argumentos. Assim, a descendência de v_i na árvore de v_k é, claramente, uma função $h(.)$ com argumentos G, v_i e v_k.

Podemos considerar valores fixos para o vértice que desejamos observar, v_i por exemplo, e em relação às duas árvores que temos interesse, como por exemplo as enraizadas por v_k e v_l. Podemos efetuar tal procedimento para estudarmos a relação existente entre as funções $h(.)$, fixando todos os pares possíveis de vértices, sempre dois a dois. Consideremos então nosso exemplo, com v_i e as árvores T_k e T_l,

[1]Consideraremos agora os vértices como sendo v_i ao invés de simplesmente i, para evitarmos confusão em relação a identificação de quais são os argumentos das funções. O mesmo valerá para todos os vértices até o fim desta seção.

e temos então $h(G, v_k, v_i)$ e $h(G, v_l, v_i)$ sendo que o único argumento indeterminado da função é G.

Quando G é fixado, a função $h(.)$ está plenamente determinada, e deixa de ser uma variável aleatória. Desejamos saber se $h(G, v_k, v_i)$ é independente de $h(G, v_l, v_i)$. Fica claro que toda aleatoriedade da função está contida em G, e por isso, e através de resultado intuitivo, porém que não iremos provar, a variável G não é independente dela mesma, e conseqüentemente $h(G, v_k, v_i)$ não é independente de $h(G, v_l, v_i)$. Baseado em tal argumentação, fica claro que não parece ser razoável esperarmos que a soma das variáveis D_i^k e D_i^l possa ser desenvolvida de forma simples, considerando as mesmas como sendo independentes.

Então, conforme resultado obtido no final da seção anterior e também por esta intuição, devemos somente considerar a soma das variáveis descendência quando for possível descrevermos de forma mais explícita a dependência entre elas, pois então poderemos utilizar tal estrutura para a obtenção de uma expressão específica para a soma, que contemple tal peculiaridade.

4.2 Estudo computacional

Dada a inviabilidade momentânea no avanço da expressão analítica para a variável Carga, efetuamos então um conjunto de simulações para observá-la e tentar propor uma explicação para a mesma, baseados em possíveis evidências experimentais.

Considerando os custos computacionais envolvidos, efetuamos experimentos com 60000 rodadas e tamanho do grafo $n = 10000$.

De fato, desejamos caracterizar a Carga através da descendência. Assim, dentro de cada simulação, geramos um grafo G baseado numa distribuição de graus do tipo lei de potências, e então construímos todas as árvores de distâncias mais curtas para cada vértice do grafo contido dentro da componente gigante[2] Assim, trivialmente determinamos suas descendências, e então a Carga L_x.

[2]Chamaremos a mesma muitas vezes somente de GCC.

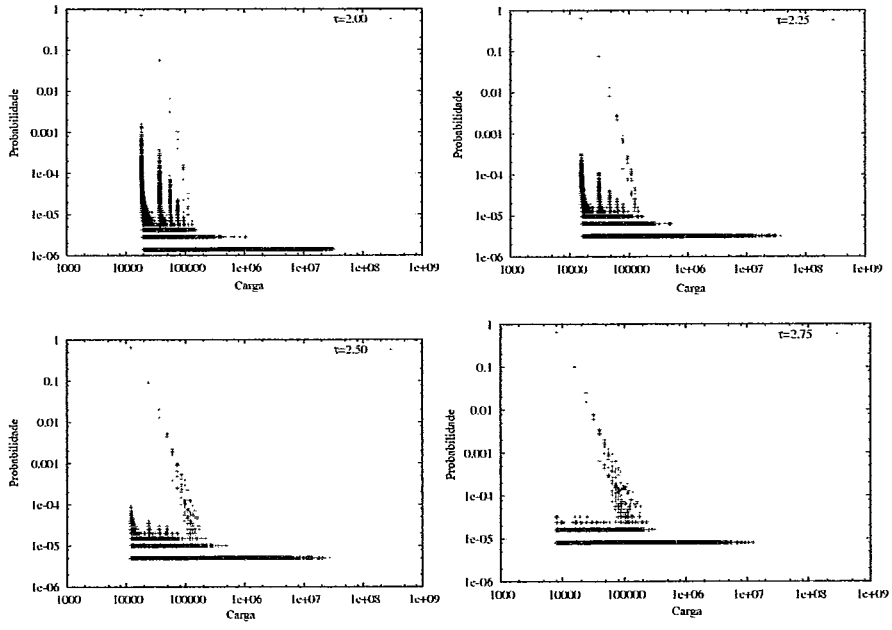

Figura 4.1: Distribuição da Carga com GCC de tamanho médio fixo (carga x probabilidade).

Inicialmente, e para facilitar nossa análise, trataremos a Carga de forma independente da variabilidade do tamanho da componente gigante [3], conforme o exposto na figura 4.1.

Neste gráfico, incluímos somente a seleção de grafos em que o valor da componente gigante foi igual à sua média, obtidos através de diversas realizações do procedimento aleatório de simulação[4]. Chamaremos o tamanho da componente gigante de n_{gcc}.

Então, podemos observar claramente picos bem definidos e um comportamento oscilatório entre os mesmos. No gráfico 4.2, agora com $n = 1000$, podemos ver mais claramente tal fato, dado que quando $n = 10000$ o mesmo é mais compacto devido

[3]Iremos comentar mais sobre a variabilidade da GCC mais adiante.

[4]Na prática, efetuamos simulações gerando grafos com componentes gigantes de todos os tamanhos possíveis, dado que não controlamos o mesmo, e então computamos a média experimental dos tamanhos das componentes baseado em todas as realizações, e finalmente selecionamos somente os grafos em que seu tamanho foi igual a média.

à escala, onde tal efeito é afetado visualmente. Tal comportamento oscilatório é bastante evidente, interessante e deve ser explicado.

Então, através da expressão do somatório de $L_z = \sum_{i=1}^{n_{gcc}} D_z^i$, notamos que a variável Carga é dependente do número de árvores que serão utilizadas para sua composição, indicado através da variável n_{gcc} do somatório (o caso extremo é no limite, conforme apresentado anteriormente na forma teórica ideal).

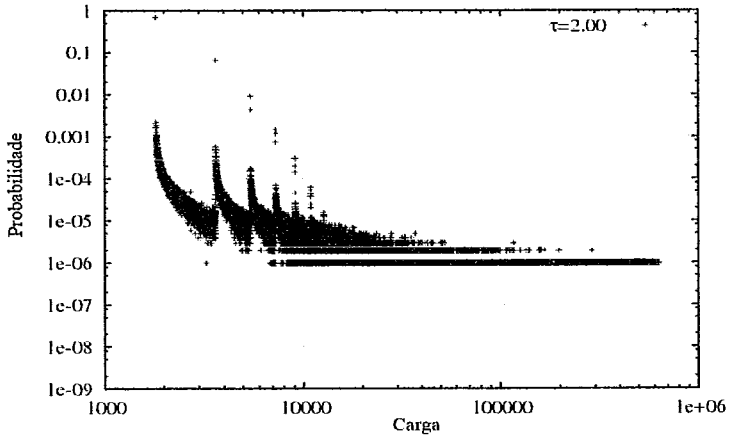

Figura 4.2: Carga selecionada para grafos com GCC iguais e tamanho $n = 1000$ (carga x probabilidade).

Seguiremos o raciocínio baseado na expressão de L_x, em que a Carga é composta pela soma das descendências. Podemos considerar, por exemplo, um vértice x que possui grau 1. Sabemos que com probabilidade 1 ele será raiz em uma das árvores (quando ele próprio estiver efetuando o envio de mensagens para todos os outros vértices), acumulando então uma parcela de n_{gcc} para contribuir com sua Carga (referente a sua descendência, que no caso é a componente gigante inteira), e, com probabilidade 1 ele será folha em $n_{gcc} - 1$ árvores, acumulando ambas as parcelas e totalizando exatamente $L_x = 2 * n_{gcc} - 1$.

Em relação a última parcela do parágrafo anterior, o vértice ter grau 1 implica diretamente em ele ter descendência 1 quando ele não é raiz, por definição. Caso isso não fosse verdade, ele estaria desrespeitando a restrição em relação a seu grau amostrado para G, que é 1. Através do valor unitário de seu grau ele é anexado a

árvore, e então, não mais restam possibilidades (grau residual) para anexar outros vértices à árvore.

O evento de um vértice de grau 1 e ter descendência 1, conforme o exposto acima, é o de mais alta probabilidade possível, pois o fato de ter grau 1 implica diretamente em ter descendência 1 (em todas as árvores menos a que ele é raiz), e através da forma da função da distribuição de graus em G (lei de potências), os vértices de grau 1 são os mais abundantes. Logo, a maioria dos vértices tem descendência 1 e conseqüentemente sua Carga vale $2 * n_{gcc} - 1$.

Para exemplificar, consideremos o caso concreto do gráfico 4.1, sendo $\tau = 2.00$ e de forma simulada $n_{gcc} = 9.252$ (média empírica).

Podemos observar que existe um pico na figura 4.1 no valor relativo à soma das parcelas descritas[5], que vale $2 * n_{gcc} - 1 = 18.503$. Podemos explicá-la como sendo 9.252 referentes aos descendentes na árvore em que o vértice é raiz, e as parcelas de 1 descendente, em 9.251 outras árvores em que ele é folha. Logo, é razoável considerar que o primeiro pico foi explicado.

Intuitivamente ao considerarmos agora o vértice ter grau dois, por exemplo, existe uma grande probabilidade de que ele tenha uma descendência menor do que um vértice de grau vinte. Podemos observar tal fato claramente, pois um vértice de grau alto tem mais cópias na fita do que um de grau baixo, assim, elas estarão mais espalhadas e então tal vértice deverá ser anexado na árvore antes do vértice de grau baixo. Logo, ele terá maior probabilidade de anexar outros vértices na árvore, dado que quanto mais cedo ele entrar na fila, mais cedo ele irá explorar outros vértices, e mais vértices estarão disponíveis para que ele possa anexar.

Essa explicação poderia ser substituída simplesmente pelo exame do gráfico da descendência (figura 3.2), que é uma função monotônica e decrescente. Pelo gráfico, mais fortemente do que a argumentação anterior, chegamos a essa conclusão sem termos que considerar os graus dos vértices, sabendo então que é mais fácil um vértice ter descendência dois do que três, três do que quatro, e assim sucessivamente.

[5]Fizemos uma verificação baseados na função de distribuição propriamente, que gerou a figura.

Consideremos agora o segundo pico da figura 4.1, que em nosso exemplo vale 37.003.

Agora, com o mesmo raciocínio aplicado as descendências de tamanho 1, podemos considerar as de tamanho 2, e temos as seguintes possibilidades:

$$\begin{cases} \text{Configuração } \mathbf{a}: 1 * 9.252 + 9.250 * 2 + 1 * 9.251 = 37.003 \\ 1 * (desc_y = 9.252) + \underline{9.250 * (desc_y = 2)} + 1 * (desc_y = 9.251) \\ \\ \text{Configuração } \mathbf{b}: 1 * 9.252 + 9.249 * 2 + 1 * 9.251 + 1 * 2 = 37.003 \\ 1 * (desc_y = 9.252) + \underline{9.249 * (desc_y = 2)} + 1 * (desc_y = 9.251) \\ \quad + 1 * (desc_y = 2) \end{cases}$$

Considerando a configuração **a**, temos um vértice que é raiz em uma das árvores, o que contribui com 9.252 para sua Carga, tem descendência 2 em 9.250 árvores e descendência 9.251 em 1 árvore.

Considerando a configuração **b**, temos um vértice que é raiz em uma das árvores, o que contribui com 9.252 para sua Carga, tem descendência 2 em 9.249 árvores, tem descendência 9.251 em 1 árvore e descendência 2 em 1 árvore também.

Dado que o desenvolvimento com os graus é complexo, e existe uma relação entre eles e as descendências, conforme argumentação anterior, iremos somente considerar as descendências. Então, conforme as parcelas das configurações **a** e **b** acima, notamos que o segundo pico tem relação preponderante com a segunda descendência mais popular, que é a de tamanho dois.

Na verdade, de forma mais geral para cada valor de Carga específico l_y, podemos considerar um sistema, sendo d_i o número de árvores em que o vértice y tem i descendentes e $desc_w$ como a cardinalidade w, ou seja, onde a descendência do vértice é igual a w.

Então, devemos determinar todos os coeficientes d_i em que o seguinte sistema se verifica:

$$\begin{cases} d_1 * desc_1 + d_2 * desc_2 + \ldots + d_{gcc} * desc_{gcc} = l_y & \text{(soma ponderada} = l_y\text{)} \\ d_1 + d_2 + \ldots + d_{gcc-1} + d_{gcc} = gcc & \text{(máximo de gcc árvores)} \\ d_{gcc} = 1 & \text{(quando é raiz)} \end{cases}$$

Podemos continuar com o mesmo raciocínio. Consideremos o terceiro pico através da figura 4.1, que vale 55.499. Temos então, por exemplo, os coeficientes $d_{9.252} = 1$, $d_3 = 9.249$, $d_{9.250} = 2$, e $d_y = 0, \forall y \mid y \neq \{1, 3, 9.250\}$. Podemos inferir relação similar entre as terceiras descendências mais populares, que são justamente as de tamanho 3.

Os picos 73.991, 92.485 e posteriores podem ser explicados de forma semelhante, considerando-se as descendências mais populares de tamanho 4, 5, e assim por diante.

Aparentemente ocorrem entre os valores dos picos variações sobre as configurações dos próprios. Por exemplo, o valor imediatamente após um pico tem coeficientes similares aos dele, porém as suas parcelas referentes as descendências mais populares vão transformando-se em outras menos populares. Ao se efetuar o cômputo da probabilidade de determinado evento ocorrer, teremos uma parcela de menor probabilidade multiplicando, compondo a probabilidade total referente aquela carga, e então seu valor será de menor magnitude.

Dado o exposto até aqui, consideramos a questão da Carga com valores fixos de componente gigante. Devemos observar através do exemplo da figura 4.3, que os tamanhos dessas componentes variam conforme o valor do expoente τ. Além disso, segundo a figura, mesmo com um valor τ fixado as componentes têm tamanhos variáveis.

Podemos constatar também que para valores de expoentes mais próximos ao valor crítico 2, quando o grafo é mais denso, há maior concentração em torno da média, enquanto para expoentes mais distantes a variância (dispersão) é maior, e a

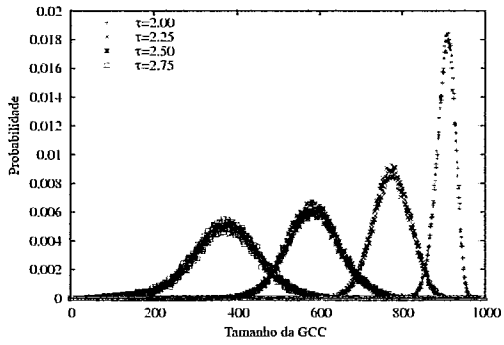

Figura 4.3: Exemplo de variação do tamanho da GCC em função do τ (tamanho da GCC x probabilidade).

componente gigante mais difusa. Mas o fato é que o valor de tal medida é variável, informação relevante e que devemos contemplar para a questão da Carga.

Baseado nessa variabilidade, devemos notar que ela altera a composição da Carga, introduzindo "ruído" na mesma. Podemos observar na figura 4.4, que para os maiores valores de τ, em que a componente gigante é mais difusa, os picos nitidamente tem maior variabilidade.

Essa variabilidade é expressa principalmente pela abertura dos picos. Devemos notar a diferença entre primeiro e último gráfico da figura, que são os casos mais extremos e onde ela se acentua. No primeiro gráfico não existe aparentemente uma abertura no pico, enquanto no último vemos ela de forma mais nítida.

Além disso, ao considerarmos a variação da n_{gcc}, existe a composição de algumas fontes geradoras para os picos, que sobrepõem-se, ou seja, acumulam-se. O efeito descrito anteriormente devido aos valores entre os picos, conforme quando o n_{gcc} está fixo, é também aplicado para quando o tamanho da GCC varia.

Por exemplo, as soluções para o sistema para l_y quando temos $n_{gcc} = 9.252$, também deverá ser resolvida para quando $n_{gcc} = 9.251$, só para citar uma, sendo que este último tamanho de GCC é um outro sistema, com outros valores para os coeficientes. Haverá uma superposição dos mesmos, e valores de n_{gcc} próximos darão coeficientes próximos. Tal sistema deverá ser considerado e resolvido para todos os tamanhos possíveis de componente gigante.

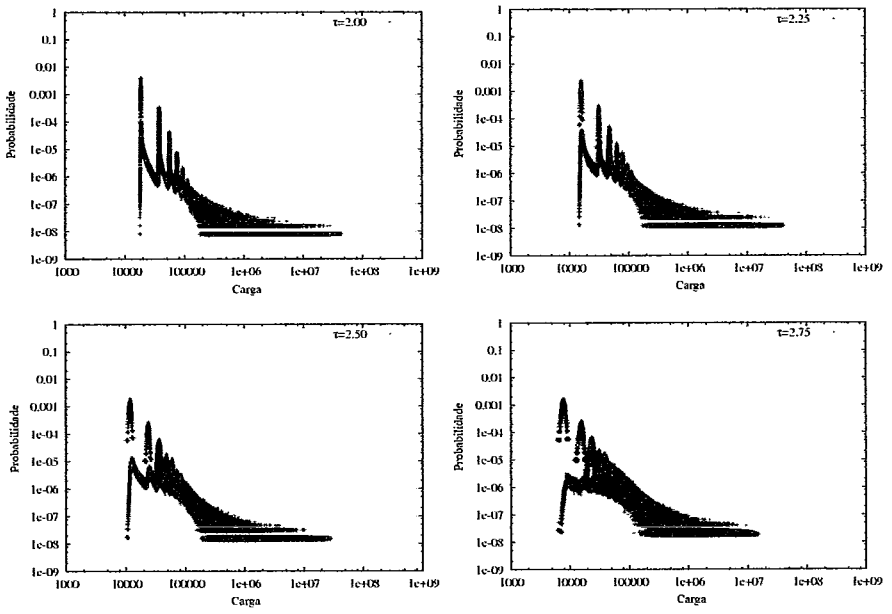

Figura 4.4: Distribuição da Carga com GCC variável (carga x probabilidade).

Assim, para contemplarmos essa variabilidade na expressão, devemos levar em consideração o valor da probabilidade de se ter a componente gigante com determinado tamanho específico, e somente depois as probabilidades de obtermos determinada configuração de Carga específica l_y, dado que a GCC está fixa naquele valor.

Então, baseados no exposto podemos sugerir uma expressão para a v.a. Carga L_y com a GCC fixa, como sendo a condicional:

$$P(L = l_y \mid GCC) \quad = \sum_{\forall d_i | \text{R se verifica}} f(d_1, \ldots, d_{gcc}) \tag{4.1}$$

sendo

$$R = \begin{cases} d_1 * desc_1 + d_2 * desc_2 + \ldots + d_{gcc} * desc_{gcc} = l_y & \text{(soma ponderada)} \\ d_1 + d_2 + \ldots + d_{gcc-1} + d_{gcc} = gcc & \text{(gcc árvores)} \\ d_{gcc} = 1 & \text{(quando é raiz)} \end{cases} \tag{4.2}$$

e d_i valendo o número de árvores em que o vértice y tem i descendentes.

Então, agora tratando a GCC conforme a sua distribuição e fazendo a sua marginalização, obtemos:

$$P(L_y = l_y) = \sum_{\forall gcc} \sum_{\forall d_i | R \text{ se verifica}} f(d_1, \ldots, d_n) P(GCC = gcc) \qquad (4.3)$$

Em [29] é discutida a obtenção da distribuição do tamanho da GCC, e mostra-se que não é possível obter uma fórmula fechada para a mesma. Dada esta impossibilidade, normalmente considera-se a utilização de seu tamanho médio, porém como em nosso caso é impreterível a distribuição e não sua esperança, tal desenvolvimento é inócuo. Contudo, ainda assim seria possível a obtenção de uma distribuição através de métodos numéricos.

Apesar de factível a obtenção de GCC, a solução do sistema de restrições conforme 4.2 é de difícil obtenção, como já discutido em [3, 37]. Assim, apesar de obtermos uma expressão em alto nível, para de fato podermos utilizá-la devemos fazê-lo de forma numérica e com elevado custo computacional.

Além disso, apesar da expressão ser função dos coeficientes que representam as descendências, a mesma também não pode ser computada facilmente. Conforme observado no início deste capítulo, essas descendências não são independentes entre si, ou seja, não poderíamos fatorá-la trivialmente.

Apesar do desenvolvimento diferenciado em relação ao nosso objetivo analítico inicial, e de toda explicação e entendimento que as simulações trouxeram em relação a Carga, houve um problema semelhante em relação a computação da função baseada na descendência. Logo, e novamente, um entendimento sobre a estrutura de dependência entre instâncias dessa variável seria necessário.

Ainda sim, consideramos interessantes os pontos levantados em tal seção, com intuito de melhor entendermos a variável Carga. Alguns desses pontos merecem destaque, como os problemas na obtenção da expressão analítica inicial, o comportamento oscilatório observado, sua análise e finalmente sua descrição, englobando

uma explicação através de algumas equações, em alto nível, também difíceis de serem obtidas em forma fechada.

Capítulo 5

Trabalhos relacionados

5.1 Distribuição de graus na árvore de caminhos mais curtos

A distribuição de graus em árvores de caminhos mais curtos foi inicialmente abordada em [2], que introduziu um formalismo bastante interessante baseado na utilização de uma fita unitária com interessantes propriedades analíticas. Seu desenvolvimento foi inspirado no trabalho de [23].

O formalismo proposto originalmente sugeria uma distribuição genérica para os graus dos vértices em G, podendo ser instanciada conforme o desejado. Em nosso caso específico desejamos aplicar uma distribuição do tipo lei de potências[1]. Assim, gostaríamos de aproveitar tal resultado para caracterização da distribuição da descendência nas árvores de distâncias mais curtas, e conseqüentemente na Carga de cada nó da rede.

Porém, ao efetuarmos um estudo mais detalhado de tal trabalho, encontramos algumas incongruências.

Por exemplo, a distribuição original para os graus dos vértices em G considerava somente vértices com um grau maior do que um determinado valor, no caso, $d(v) \geq 3$. Essa restrição quanto aos valores dos graus geram grafos subjacentes

[1]Tal distribuição também foi considerada no artigo original.

mais restritos, logo achamos mais pertinente removê-la, considerando redes com topologias mais gerais.

Ao tentarmos remover tal hipótese nos deparamos com uma variável que expressava uma probabilidade específica do modelo, e a mesma somava um valor maior do que um, que representava uma contradição. Então, notamos que não poderíamos simplesmente descartar tal restrição sem desenvolvermos algum outro tipo de solução para tal problema.

Resumindo o desenvolvimento para obtenção da contradição supracitada, a variável que representaria a probabilidade é chamada originalmente no artigo como $p_{unto}(t)$ [2], e pode ser reduzida, através de manipulações algébricas básicas, para $p_{unto}(t) = \frac{\alpha_1}{\delta t} + cte_1 + \sum_{k=3}^{\infty} \alpha_k k t^{k-2}$. Ao aplicarmos o limite quando t tende a zero em tal expressão, temos claramente a mesma indo para o infinito, um absurdo considerando que ela deve respeitar o limite unitário de toda probabilidade.

Ainda assim, não consideramos a restrição quanto aos graus dos vértices um erro, dado que existem algumas redes complexas que apresentam comportamento do tipo lei de potência somente a partir de determinados valores para o seu grau mínimo. Ainda assim, consideramos curioso tal fato aplicado ao caso da Internet, pois teríamos que desconsiderar os três graus mais populares, que concentram a quase totalidade dos vértices da rede, de forma real e prática[3].

Então, independente da restrição na distribuição de graus de G, e nossas considerações a respeito, executamos simulações para observarmos o uso da expressão original, obtendo assim o gráfico 5.1[4]. Podemos notar que os valores das previsões analíticas divergem em relação à distribuição simulada para os graus dos vértices

[2] O argumento da função é uma variável t, que é similar a utilizada na análise com fita dos capítulos anteriores.

[3] Apesar do deslocamento ser possível, do ponto de vista teórico, para todas as distribuições que seguem uma lei de potências, deveríamos somente considerá-lo quando pela definição real da rede subjacente tal deslocamento se impor, espelhando assim alguma propriedade real da mesma. Este claramente não é o caso da Internet.

[4] A cor vermelha nos três primeiros gráficos equivalem a versão simulada da distribuição, e a verde é relativa a versão analítica, conforme proposto originalmente. No último gráfico considera-se algumas variações para os valores do τ analítico.

na árvore em largura, ou seja, mesmo considerando os graus em que a expressão se propõe a funcionar a mesma não apresenta consistência.

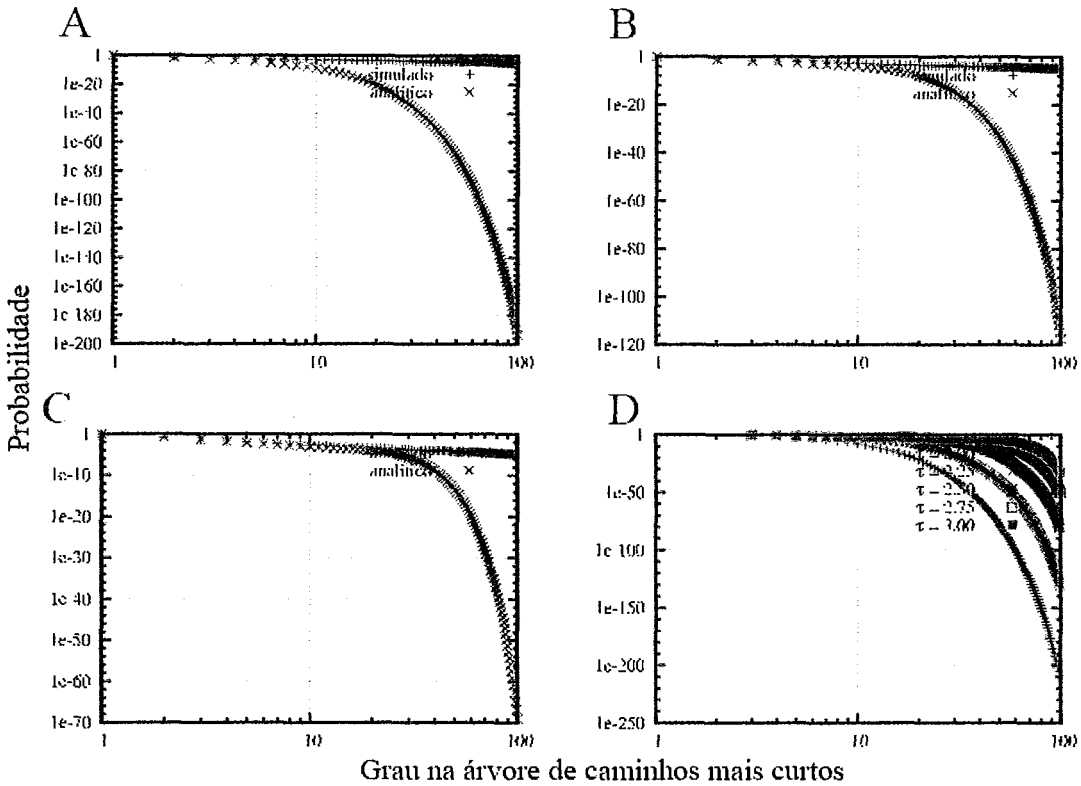

Figura 5.1: Distribuição de graus na árvore variando τ conforme expressão de [2] (grau x probabilidade).

Finalmente, ao examinarmos mais detalhadamente o artigo original, notamos uma outra incongruência em uma expressão chave do mesmo. Existe em tal trabalho uma variável com propósito relativamente similar a nossa chamada $p_{vis,q}(t)$, expressa em 2.7, que é anexar novos vértices individualmente na árvore através de um vértice que já está em tal árvore .

Contudo, em nosso trabalho contemplamos claramente o tempo de entrada do vértice na fila da busca em largura, e o momento em que o mesmo alcança sua cabeça e então deverá iniciar sua exploração. No artigo original não se considera tal fato, simplesmente não distinguindo entre esses dois momentos distintos. Porém, o mesmo não pode ser desconsiderado, pois pelo observado nas estimativas da fila

na figura 2.4), pode-se constatar que a mesma não é nula, não podendo então ser completamente negligenciada.

Complementarmente, e na mesma linha, ao negligenciar a fila e conseqüentemente o tempo de entrada e exploração relativo a cada vértice, também não se contempla o tratamento adequado para o grau residual dos vértices, requisito fundamental do processo, conforme discutido em nosso desenvolvimento e tratado na expressão 2.6.

Vale observarmos que o estudo original [2] não contemplou nenhum tipo de simulação, o que consideramos inadequado.

5.2 Distribuição da carga

Analisamos a proposta de [19], que foi o trabalho onde se introduziu o conceito de Carga aplicado as Redes Complexas. No mesmo, foi conjecturado resultado empírico bastante forte em que se sugere que a distribuição da Carga também segue uma lei de potência e seu expoente τ é universal. Esse expoente teria valor fixo sugerido $\tau = 2.2$, independendo do valor do expoente em relação à distribuição de graus original em que o grafo G foi instanciado, considerando-se sempre o intervalo referente aos valores encontrados para as redes de interesse como a Internet e a Web.

Podemos dividir nossa análise, classificando-a em três grandes grupos, e primeiramente abordaremos as questões referentes às interpretações conceituais, seguindo as relativas ao rigor estatístico e, finalmente, sobre as evidências empíricas.

Mais especificamente, quanto às questões conceituais, em tal artigo trabalha-se com um grau médio pré-fixado (chamado de m no mesmo), não levando em consideração a variação referente ao expoente da distribuição de graus. Sabe-se que o grau médio é função direta do expoente da distribuição de graus subjacente[5],

[5]Sabemos que o primeiro momento de uma distribuição do tipo lei de potência com expoente γ vale $\zeta(\gamma-1)/\zeta(\gamma)$.

apesar de não parecer ser contemplado adequadamente em tal trabalho. Supõe-se uma insensibilidade em relação ao mesmo que não se verifica.

Um outro ponto que é central, diz respeito à utilização equivocada da técnica de *binning*[6], uma vez que a mesma é aplicada com um propósito diferente do que quando utilizada corretamente, que é minimizar problemas relacionados a amostragem. Conforme observado no capítulo anterior, tal função tem uma forma funcional bem definida e não parece razoável aplicar tal técnica no caso em questão, inclusive mascarando a distribuição real obtida. Estes dois últimos pontos, graus médios e *binning*, são bastante fortes e devem ser levados em consideração.

Em relação às questões quanto ao rigor estatístico, também podemos dividi-las em duas partes. A primeira diz respeito aos fatores decorrentes de quando consideramos previamente que a distribuição obtida já é do tipo lei de potência. Baseando-se em tal premissa, cometeram-se erros como a não utilização de um estimador consistente e não-viciado, que existe para tal distribuição e poderia ser facilmente aplicado[7], e a não utilização da distribuição acumulada[8], que substituiria a técnica de *binning* para suavização da função em relação aos intervalos em que a amostragem pudesse ser problemática, introduzindo vícios (leia-se erros) no expoente quando os mesmos são obtidos através de um ajuste visual ou uma técnica de regressão em um gráfico duplamente logarítmico[9].

Além disso, e mais gravemente, a premissa de se utilizar uma lei de potências é equivocada, de considerá-la *a priori*, uma vez que esta seria justamente a nossa tarefa, e não a nossa hipótese de trabalho. Justamente devemos verificar se uma

[6]A técnica tem como objetivo resolver situações quando existe algum tipo de problema em relação à *amostragem*. Ela tem duas versões básicas: a fixamente espaçada e a logaritmicamente espaçada. Para maiores informações sobre este procedimento recomendamos [28].

[7]O estimador de β chamaremos de $\hat{\beta}$, e é obtido através da resolução da equação transcendental $\frac{\zeta'(\hat{\beta},x_{min})}{\zeta(\hat{\beta},x_{min})} = -\frac{1}{n}\sum_{i=1}^{n} lnx_i$, sendo x_{min} o limite inferior para o grau da distribuição de x.

[8]Conforme proposto em [13], se utilizarmos o complementar da distribuição acumulada de x como $P(x) = \int_x^\infty p(x')dx'$, e se sabemos que a mesma segue uma lei de potências, temos então $P(x) = C\int_x^\infty x'^{-\alpha}dx' = \frac{C}{\alpha-1}x^{-(\alpha-1)}$, que também é uma lei de potências, porém com o expoente deslocado.

[9]A aplicação de tal técnica deve ser evitada, conforme discussão do texto [13] e confirmado através de experimentos bastante contundentes, resumidos em sua tabela IV e figura 2.

distribuição de tal forma é ou não adequada em relação aos dados obtidos. Quando se propõe fazer uma estimativa para o expoente, já se está implicitamente considerando a distribuição como uma lei de potências, e, na verdade, somente se está buscando qual expoente melhor representa aqueles dados, um absurdo em nosso caso!

Complementarmente, vale sublinharmos, que a verificação de que em um gráfico duplamente logarítmico, uma reta representa bem os dados, é uma condição necessária[10], porém não suficiente para se conjecturar uma lei de potências, uma vez que existem outras distribuições que também podem ser aparentemente bem representadas de tal forma.

Não obstante, e conforme já descrito, existe ferramental estatístico que justamente se propõe a responder tal questão: é plausível considerar que um determinado conjunto de dados é originado de uma distribuição específica? Tais ferramentais são quantitativos, e conhecidos como testes de hipóteses[11], com utilização complementar do *p-valor*[12], testes do tipo *likelihood ratio*, entre outros.

Estes testes possibilitam aumentar nosso grau de confiança, com indícios e representatividade estatística, sendo condição necessária, porém não suficiente para a hipótese quanto à distribuição. E apesar de não poderem confirmar, somente corroborar, podem de forma definitiva descartar determinada possibilidade de forma funcional. Nenhuma dessas ferramentas, que são bastante populares, diga-se de passagem, foram utilizadas ou sequer cogitadas. Novamente, consideramos a não utilização das mesmas uma falha grave.

[10]Conforme estudo de [13], evidencia-se através da análise de 24 trabalhos bastante representativos da literatura, que seus autores utilizam de tal artifício sem qualquer rigor estatístico (verificação da reta no gráfico log-log). Ao se efetuarem alguns poucos testes estatísticos quanto a adequação da hipótese da distribuição ser do tipo lei de potências, em vários casos a mesma não se verifica.

[11]Existem algumas medidas possíveis para quantificar a distância entre duas distribuições de probabilidades para dados não distribuídos por uma normal, e se tais distribuições podem ser consideradas as mesmas, e um dos testes mais comuns para tal é devido a Kolmogorov-Smirnov, ou KS.

[12]Indica em conjunto com uma medida de distância entre distribuições, por exemplo a KS, se os dados foram obtidos daquela distribuição.

Então, aplicamos tais testes a um conjunto de simulações conforme algoritmo proposto em tal trabalho, e obtivemos os seguintes resultados expressos na tabela 5.1. Temos indicado na mesma os valores do expoente da distribuição dos graus (γ), o expoente da distribuição da Carga (δ) quando consideramos a mesma como uma lei de potência (obtidas através de seu estimador de máxima verossimilhança), o valor retornado pelo teste KS (que é utilizado para compor o *p-valor*), e o próprio *p-valor*.

Os resultados para o *p-valor* quando inferiores a 0.1 [13] foram verificados para todos os expoentes da distribuição de graus (sempre zerados), além da estimativa dos expoentes para a distribuição da Carga pelo estimador de máxima verossimilhança não ter sido universal para todos eles, muito menos com valor fixo 2.2.

Ou seja, a hipótese de que os dados são bem ajustados para uma distribuição do tipo lei de potência, e a suposta universalidade de seu parâmetro não são confirmadas pelos testes realizados. Ao observarmos a forma funcional do capítulo anterior (pela figura 4.4), verificamos também a total falta de evidências experimentais para suportar tais hipóteses.

γ	**Estimativa de δ**	KS	***p-valor***
2.00	1.8996	0.0178	0.00
2.25	2.0177	0.0179	0.00
2.50	2.1066	0.0185	0.00
2.75	2.2306	0.0187	0.00

Tabela 5.1: Estimação de máxima verossimilhança para expoente, valor do teste KS, e *p-valor*.

Adicionalmente ao elencado anteriormente, realizamos um conjunto de experimentos que evidenciaram através da figura 5.2 que a distribuição da Carga varia conforme o expoente e também em relação ao grau médio, ou seja, não podendo suportar a hipótese da universalidade do expoente.

Por todo o exposto, devemos ter cautela ao consideramos o resultado do artigo original [19].

[13] Conforme proposto em [13] .

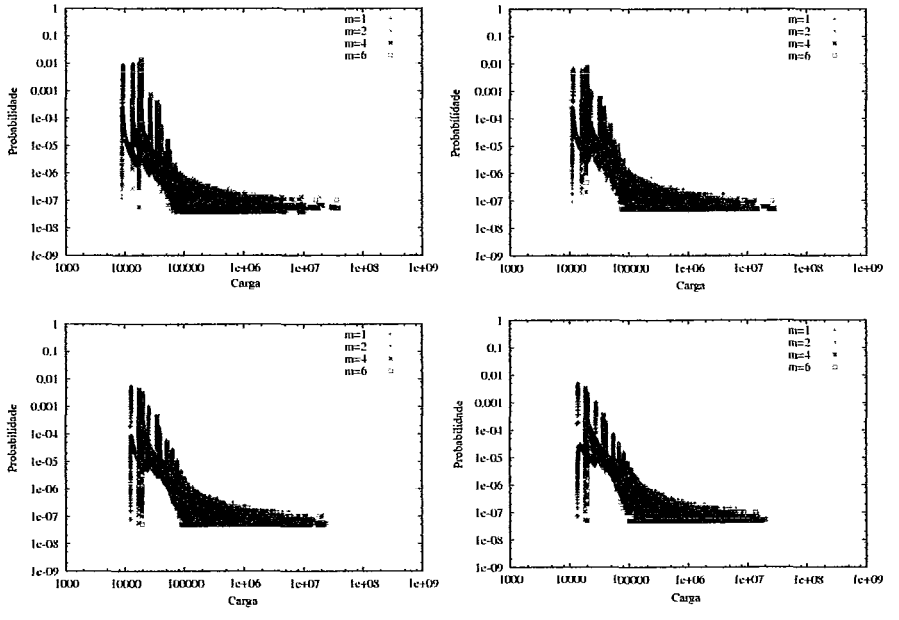

Figura 5.2: Distribuição da Carga com expoente variando, sendo $\tau = \{2.00, 2.25, 2.50, 2.75\}$, e com grau médio fixo (carga x probabilidade).

Capítulo 6

Conclusões e trabalhos futuros

Neste trabalho, abordamos um problema de Redes Complexas que é a caracterização da propriedade Carga, considerando-a quando a rede é submetida a um tráfego uniforme (isto é, o número de rotas que passam por cada vértice quando todos os pares dos mesmos enviam mensagens entre si). Para tal, desenvolvemos propostas e análises do ponto de vista teórico e simulado.

Nosso desenvolvimento se deu, inicialmente, através da idéia de utilizarmos o conjunto de árvores de caminhos mais curtos enraizadas nos nós da rede, para então manipulá-las para a caracterização da propriedade Carga.

No capítulo 2, introduzimos uma nova análise para descrever a distribuição de graus em tais árvores, e então sugerimos uma expressão analítica para a mesma.

Constatamos através de resultados simulados que nossa expressão está próxima, em termos de verossimilhança, ao processo subjacente que ela se propõe a descrever. Vale notarmos que uma parte específica da nossa expressão, referente a variável que descreve o tamanho da fila no processo de busca, não continha uma forma fechada e tivemos que estimá-la através de simulações.

Então, com a distribuição de graus na árvore caracterizada, no capítulo 3 valemo-nos do fato de que a descendência de um vértice na árvore em largura representava o conjunto de mensagens passando pelo vértice quando a raiz de tal árvore estiver fazendo o seu respectivo envio de mensagens.

Então, como nosso objetivo era obter tal variável quando todos os vértices fizessem seus envios, deveríamos considerar as árvores enraizadas por cada um dos vértices, e assim efetuarmos a sua soma (característica de quando o grafo esta sendo submetido a um tráfego uniforme).

Logo, para compor a Carga deveríamos caracterizar a distribuição da descendência de um vértice baseados na distribuição de graus em uma árvore em largura. Então, dada a natureza recorrente do padrão das descendências em que uma cascata de distribuições ocorriam, utilizamos do formalismo devido as funções geradoras, da própria recorrência identificada na definição da descendência, e de uma identidade devido a Lagrange [1, 35] para expansão de uma função em termos de série de potências, e desenvolvemos então uma expressão para a descendência fixada a distribuição de graus subjacente. Chegamos tanto à forma analítica como à simulada, e obtivemos resultados satisfatórios.

No capítulo 4, fizemos considerações a respeito da composição da variável descendência para utilização como parcela na expressão da Carga, que seria o nosso objetivo final. Porém, constatamos através de uma simples manipulação simbólica sobre a variável descendência, e também baseados em observações experimentais do final do capítulo anterior, que as variáveis descendência não são independentes entre si. Logo, não poderíamos utilizá-las de forma trivial sem contemplarmos a estrutura de dependência existente entre as mesmas.

Considerando-se esta impossibilidade momentânea, efetuamos então simulações para observarmos o comportamento da variável Carga, e chegamos a uma distribuição empírica, com uma forma bastante peculiar, e um comportamento oscilatório interessante.

Então, através de um estudo da mesma, propomos uma explicação para este comportamento, e sugerimos então outras equações, em alto nível. Tal expressão pode ser melhorada, pois em sua forma atual existem restrições quanto a computabilidade, dado o elevado custo computacional para obtê-la. Surgiram também problemas semelhantes aos relativos a descendência nesta nova expressão. Senti-

mos que essas equações poderiam ser mais bem estudadas, e possivelmente, serem melhoradas.

Finalmente, no capítulo 5 consideramos os trabalhos relacionados aos assuntos envolvidos neste estudo. Dividimos em duas partes este capítulo, sendo a primeira relacionada a distribuição de graus em árvores em largura conforme [2].

Ao efetuarmos um estudo mais detalhado sobre tal trabalho, encontramos algumas incongruências. Por exemplo, a distribuição original considerada era pouco realística para redes do tipo Internet, considerando somente vértices com um grau maior do que determinado valor.

Além disso, uma determinada parcela de uma expressão fundamental[1], e que deveria representar uma probabilidade continha inconsistências, tanto do ponto de vista técnico, onde seus valores tornavam-se maiores que um, como em termos conceituais, não se levando em consideração aspectos relevantes do processo que se propunha a descrever (por exemplo, o *gap* entre o tempo de entrada na fila e a exploração do vértice propriamente).

Então, executamos simulações que claramente demonstraram que nem em termos de tendência geral do processo a expressão original descrevia de fato o processo algorítmico subjacente.

Na segunda parte do capítulo, estudamos a proposta em [19], que foi o trabalho onde se introduziu o conceito de Carga aplicada as Redes Complexas. No mesmo, foi conjecturado resultado empírico bastante forte em que se sugere, primeiramente, que a distribuição da Carga também segue uma lei de potência, e em segundo lugar, que o parâmetro da distribuição, seu expoente τ, é universal. Tal expoente teria valor fixo sugerido $\tau = 2.2$, independente do valor do expoente da distribuição de graus original em que o grafo G foi instanciado, considerando-se sempre o intervalo referente aos valores encontrados para as redes de interesse como a Internet e a Web.

[1]Estamos nos referindo a expressão mais geral, que engloba $p_{unto}(t)$. Esta última foi a que mostramos a divergência quanto ao intervalo unitário, no capítulo anterior.

Então, dividimos nossa análise quanto a este trabalho em três grandes grupos, e nela primeiramente abordamos as questões referentes às interpretações conceituais do problema, seguindo as relativas ao rigor estatístico e, finalmente, sobre as evidências empíricas.

Baseados em todos os pontos levantados a partir de nossa análise, e em conjunto aos resultados obtidos no capítulo 4, esperamos ter deixado claro que o resultado de tal trabalho nos parece inadequado.

Como ponto interessante para desenvolvimento de trabalhos futuros, sugerimos o aperfeiçoamento de nossas expressões, seja do ponto de vista de acurácia, aumentando a verossimilhança em relação ao processo subjacente que ela descreve, como também na aplicação de resultados que possibilitem que as expressões possam ser computadas em um tempo mais razoável.

Para uma completa caracterização da variável do ponto de vista analítico, o desenvolvimento de uma expressão fechada para a fila se faria necessária. Além disso, também seria necessário contemplar a estrutura de dependência existente entre as variáveis descendência. Ambas as questões são desafiadoras, e interessantes para um aprofundamento.

Além disso, grafos com arestas ponderadas também poderiam ser consideradas, e aumentam a verossimilhança com as redes reais em que os canais tem configurações distintas. Porém, contemplar tal fato aumenta bastante a complexidade da modelagem (uma idéia inicial seria considerarmos ponderações sobre as diversas árvores de distâncias mais curtas de um vértice específico).

Um outro ponto interessante a se estudar seria a obtenção da variável Carga de forma distribuída pela rede, de modo que os vértices possam tirar proveito da mesma (nas redes que consideramos e que são de interesse geral, não existe uma entidade global, superior hierarquicamente, que possa computar tal varíavel).

Vale observarmos que o conhecimento da variável Carga dá informação, sobre duas outras grandezas: (i) o número de rotas passando por cada vértice; (ii) a largura de banda que cada canal da rede deve ser capaz de suster dadas as rotas que passam por ele.

Acreditamos que a melhor caracterização dessa grandeza poderá apontar para a possibilidade de intervir em algoritmos de roteamento de rede visando a uma distribuição mais eqüitativa, não somente da carga entre os nós, mas também do tráfego em cada canal quando se utilizam redes de "overlay" para aplicações "peer-to-peer".

Além disso, um conhecimento profundo de como se comportam essas árvores nas redes em questão está intimamente relacionado à possibilidade de se desenvolverem novos algoritmos para duas classes importantes de problemas distribuídos, a saber, o roteamento de mensagens e o estabelecimento de redes de "overlay" para aplicações "peer-to-peer".

Uma outra sugestão, e que poderia ser uma interessante aplicação do desenvolvido neste trabalho, seria considerarmos redes construídas baseadas na rede de trânsito de grandes metrópoles, e que parecem conter relação direta ao estudado até aqui.

Por exemplo, as ruas como sendo os vértices e havendo arestas entre eles caso exista uma ligação entre as respectivas ruas. Seria relevante medirmos justamente a Carga (tráfego) em cada rua, de tal sorte que abra possibilidade para criação de um estado de fluxo que seja o mais livre possível. Uma outra versão de tal problema, bastante diferente, seria considerarmos somente as ligações entre as ruas, que no caso de nossa modelagem inicial seria obtermos a Carga das arestas e não dos vértices.

Esperamos que os avanços teóricos e as idéias contidas nesta tese abram novas possibilidades para modelagem de propriedades tão interessantes como a Carga em Redes Complexas e conseqüentes avanços no entendimento e utilização de tais redes.

Apêndice A

Leis de potência e independência de escala

A distribuição do tipo Lei de Potência modela de forma mais realística redes do tipo Internet [18, 25] e WWW [24, 12], não possuindo decaimento exponencial em torno da média, conforme os modelos clássicos que utilizam a distribuição de Poisson. Em um grafo em que sua distribuição de graus segue uma lei de potências, um vértice v escolhido aleatoriamente possui grau $\alpha \geq 1$ com probabilidade proporcional a $\alpha^{-\tau}$, sendo que a distribuição é parametrizada através do τ. Iremos considerar os intervalos $2 \leq \tau \leq 3$ como representativos para a problemática em questão, conforme [19]. Consideremos uma constante de normalização C, a distribuição é:

$$P_G(\alpha) = C\alpha^{-\tau} \tag{A.1}$$

onde

$$\sum_{\alpha=1}^{n-1} C\alpha^{-\tau} = 1$$

Quando $n \to \infty$, a constante C pode então ser obtida por

$$C = \frac{1}{\sum_{\alpha \geq 1} \alpha^{-\tau}} = \frac{1}{\zeta(\tau)} \tag{A.2}$$

onde $\zeta(x) = \sum_{n=1}^{\infty} i^{-x}$ é a função zeta de Riemann [36]. Juntando A.1 e A.2 temos:

$$P_G(\alpha) = \frac{\alpha^{-\tau}}{\zeta(\tau)}. \tag{A.3}$$

A função $\zeta(x)$ converge para alguns valores de x, como $x > 1$, porém existem alguns desses valores positivos para os quais as funções não convergem, conforme demonstrado em [31].

Então, a distribuição com que iremos trabalhar primordialmente é a lei de potências, e possui a propriedade de ser livre de escala (*scale-free*). Existem muitos fenômenos naturais que dependem de uma escala, variando dentro de um intervalo bem definido em que a diferença entre valor mínimo e máximo encontra-se moderado, com uma mesma ordem de magnitude, e os valores encontram-se distribuídos em torno de uma média. Exemplos disso são a altura e o peso de pessoas, velocidade média de veículos, entre outros.

Existem fenômenos que não possuem essa característica de serem distribuídos em torno de sua média. Tal fato se verifica quando o intervalo de variação do domínio é de alta ordem de magnitude. Um exemplo clássico é a freqüência das populações das cidades, que foi verificado para o caso das cidades brasileiras em [26]. Conforme [28], na figura A.1, enquanto o intervalo entre menores e maiores alturas pode variar, por exemplo, entre 57cm e 272cm, o que nos dá uma variação de quase 5 vezes, e também a velocidade de um automóvel, que se encontra centrada em torno de uma média, e é bastante restrita no intervalo entre 40 e 100 quilômetros por hora, os dois últimos gráficos [1] mostram o tamanho das cidades, que podem variar de poucas dezenas de pessoas até quase 10 milhões delas, sendo

[1] Escala normal e logarítmica, respectivamente.

que a fração entre elas é cerca de centenas de milhares, e não encontra-se centrada em torno de uma média.

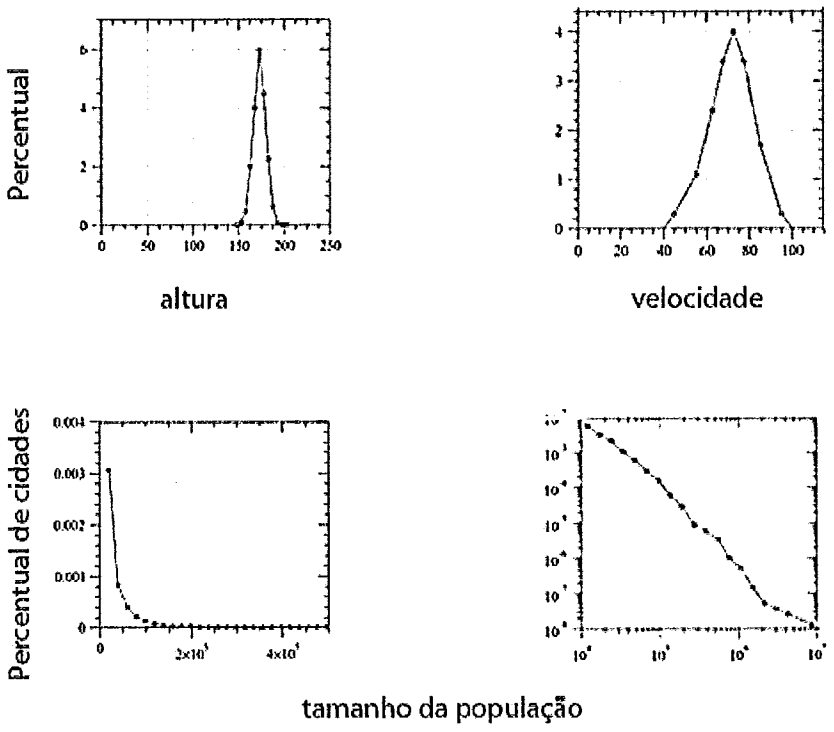

Figura A.1: Gráfico com frequências exemplificando propriedade de independência de escala.

Consideremos agora o gráfico da figura em escala logarítmica em ambos os eixos, conhecido também como escala logarítmica dupla, ou log-log. Podemos observar que uma linha reta em escala logarítmica representa as probabilidades.

Seja $p(x)dx$ uma fração de cidades com população entre x e $x + dx$. Sendo o histograma uma linha reta em escala log-log, então $\ln p(x) = \tau \ln x + C$, onde τ e C são constantes. Se observarmos a fórmula A.2 e aplicarmos log em ambos os lados da equação, temos o mesmo resultado. Ambas as formas representam a independência de escala (em escala logarítmica), ou seja, independente do tamanho do grafo, a dinâmica se mantém. Além disso, existe uma relação de proporcionalidade entre as probabilidades.

De forma mais específica, suponhamos que haja uma distribuição de probabilidades $p(x)$ para uma determinada variável x, e suponhamos que sabemos *a priori* que ela satisfaz a relação

$$p(bx) = g(b)p(x) \tag{A.4}$$

para algum b. Isto é, se aumentarmos a escala ou unidades em que o x esta sendo medido por um fator b, a forma da distribuição $p(x)$ não se alterará, a menos de uma constante multiplicativa.

De forma prática, podemos achar que arquivos de tamanho de 1kB são quatro vezes mais comuns dos que o de 2kB (note que estamos em escala de kilobytes, ou 2^10 bits. Considerando agora a escala de megabytes (MB e equivale a 2^20 bits), iremos achar que a relação dos arquivos de 1MB para os de 2MB também valem 4. O mesmo valeria para a proporção entre arquivos de gigabytes (equivalendo a 2^30).

Assim, a forma da distribuição não depende da escala em que os objetos são medidos. Essa propriedade não é verdadeira para a maioria das distribuições, como, por exemplo, a distribuição exponencial.

Considerando A.4, seja $x = 1$, o que nos daria $p(b) = g(b)p(1)$. Assim $g(b) = p(b)/p(1)$ e poderíamos reescrevê-la como:

$$p(bx) = \frac{p(b)p(x)}{p(1)} \tag{A.5}$$

Dado que tal equação deve ser verdadeira para qualquer b, podemos diferenciá-la em ambos os lados em relação ao b e temos:

$$xp'(bx) = \frac{p'(b)p(x)}{p(1)} \tag{A.6}$$

onde p' indica a derivada. Fazendo $b = 1$, temos:

$$x\frac{dp}{dx} = \frac{p'(1)}{p(1)}p(x) \tag{A.7}$$

que é uma equação diferencial de primeira ordem com solução:

$$\ln p(x) = \frac{p(1)}{p'(1)} \ln x + constante. \tag{A.8}$$

Fazendo $x = 1$ encontramos a constante simplesmente como $\ln p(1)$, e aplicando a exponencial em ambos os lados temos:

$$p(x) = p(1)x^{-\alpha} \tag{A.9}$$

onde $\alpha = -p(1)/p'(1)$. Assim, podemos afirmar que a distribuição do tipo lei de potência é a única distribuição que satisfaz o critério de ser livre de escala.

Tal definição captura a noção de auto-similaridade, que nos induz ao conceito relativo à simetria, quando parte do objeto mantém, exata ou aproximadamente, forma ou propriedade similar ao objeto como um todo. O exemplo mais conhecido é o dos fractais. Em nosso caso, a lei de formação dos graus dos vértices em redes como a Internet é independente de escala, e tem comportamento auto-similar independente do intervalo em que observamos. Tal propriedade evidencia o que já foi falado, que distribuições na família da Poisson são inadequadas para descrever redes desse tipo, dada a sua inaplicabilidade.

Apêndice B

Tabela de fórmulas

Devido a grande quantidade de variáveis envolvidas no trabalho, listamos as mesmas na tabela abaixo com intuito de facilitar a consulta e a respectiva leitura desta dissertação, segue:

Variável	Descrição
G	grafo ou rede corrente.
L_u	distribuição da varíavel Carga de um vértice u.
L	distribuição da varíavel Carga de um vértice qualquer.
T_v	árvore de caminhos mais curtos enraizada pelo vértice v.
T	árvore de caminhos mais curtos enraizada por um vértice qualquer.
D_u^v	distribuição da descendência do vértice u em T_v.
D	descendência de um vértice qualquer em T.
α	distribuição de graus dos vértices em G.
α_i	probabilidade de um vértice v ter grau i em G.
A	distribuição de graus dos vértices em T.
A_{m+1}	probabilidade de um vértice v ter grau $m+1$ em T.
$d(v)$	grau de v em G.
$d_{res}(v)$	grau residual de v em G.

Tabela B.1: Resumo das fórmulas (parte 1/2).

Variável	Descrição
$q(t)$	fila no tempo t.
$Q(t)$	distribuição do tamanho da fila no tempo t.
$\rho_{i,m}$	probabilidade de um vértice v ter m filhos em uma árvore T, dado que o mesmo tem grau i em G.
$\rho_{i,m}(t)$	probabilidade de um vértice v ter m filhos em uma árvore T, dado que o mesmo tem grau i em G e foi anexado em T no tempo t.
$\mu_{t,i}$	probabilidade para o valor do máximo entre cópias de um vértice v, sendo que o mesmo tem i cópias sorteadas uniformemente no intervalo unitário, ser igual a t.
$\Delta_{i,t,q,r}$	probabilidade do grau residual de um vértice v ser r, dado que possui grau i no grafo G, entrou na fila no tempo t, sendo que seu tamanho em tal instante de tempo é igual a q.
$p_{vis,q}(t)$	probabilidade referente à anexação na árvore, por parte de v, de um novo vértice, considerando-se que v entrou na fila em t e o tamanho dela neste instante de tempo é q.
$\rho_{i,m,q}(t)$	probabilidade referente a um vértice v ter m filhos em uma árvore T, dado que o mesmo tem grau i em G e foi anexado na árvore no tempo t, ou seja, sua cópia de maior índice na fita vale t, e o tamanho da fila em tal instante vale q.
C_i	número de filhos de um vértice i em T.
B_i	número de descendentes de um vértice i em T.
B_{ij}	número de descendentes de um vértice j em T, sendo que j é filho de i.
$G_X(u)$	transformada da variável aleatória X em relação a u.
$\zeta(\tau)$	função Zeta de Riemann com parâmetro τ.
GCC	variável referente a componente gigante.
τ	expoente típico de uma lei de potências.

Tabela B.2: Continuação do resumo das fórmulas (parte 2/2).

Bibliografia

[1] ABRAMOWITZ, M., STEGUN, I. A., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York, Dover, 1964.

[2] ACHLIOPTAS, D., CLAUSET, A., KEMPE, D., *et al.*, "On the bias of traceroute sampling: or, power-law degree distributions in regular graphs", In: *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 694 – 703, 2005.

[3] AIYANGAR, S. R., HARDY, G. H., AIYAR, P. V. S., *et al.*, *Collected Papers of Srinivasa Ramanujan.* Ams Chelsea Pub, 2000.

[4] ALBERT, R., JEONG, H., BARABáSI, A.-L., "The diameter of the world wide web", *Nature*, v. 401, pp. 130–131, 1999.

[5] BARABáSI, A.-L., ALBERT, R., "Emergence of Scaling in Random Networks", *Science*, v. 286, n. 5439, pp. 509–512, 1999.

[6] BARBOSA, V. C., *A Introduction to Distributed Algorithms.* The MIT Press, 1996.

[7] BARBOSA, V. C., DONANGELO, R., SOUZA, S. R., "Directed cycles and related structures in random graphs: I—Static properties", *Physica A*, v. 321, pp. 381–397, 2003.

[8] BARBOSA, V. C., DONANGELO, R., SOUZA, S. R., "Directed cycles and related structures in random graphs: II—Dynamic properties", *Physica A*, v. 334, pp. 566–582, 2004.

[9] BARBOSA, V. C., DONANGELO, R., SOUZA, S. R., "Emergence of scale-free networks from local connectivity and communication trade-offs", *Physica E*, v. 74, p. 016113, 2004.

[10] BARBOSA, V. C., DONANGELO, R., SOUZA, S. R., "Emergence of scale-free behavior in networks from limited-horizon linking and cost trade-offs", *Arxiv ?*, v. -, pp. ?–?, 2007.

[11] BORNHOLDT, S., (EDITORS), H. G. S., *Handbook of Graphs and Networks: From the Genome to the Internet.* Wiley VCH, 2003.

[12] BRODER, A., KUMAR, R., MAGHOUL, F., *et al.*, "Graph structure in the Web", *Computer Networks*, v. 33, pp. 309–320, 2006.

[13] CLAUSET, A., SHALIZI, C., NEWMAN, M., "Power law distributions in empirical data", *Arxiv ?*, v. -, pp. ?–?, 2007.

[14] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., *et al.*, *Introduction to Algorithms (Second Edition).* IT Press and McGraw-Hill, 2001.

[15] ERDőS, P., RéNYI, A., "On Random Graphs", *Publicationes Mathematicae*, v. 6, pp. 290–297, 1959.

[16] ERDőS, P., RéNYI, A., "On the evolution of random graphs", *Publications of the Mathematical Intstitue of the Hungarian Academy of Sciences*, v. 5, pp. 17–61, 1960.

[17] ERDőS, P., RéNYI, A., "On the strength of connectedness of a random graph", *Acta Mathematica Scientia Hungary*, v. 12, pp. 261–267, 1961.

[18] FALOUTSOS, M., FALOUTSOS, P., FALOUTSOS, C., "On power-law relationships of the internet topology", In: *SIGCOMM*, pp. 251–262, 1999.

[19] GOH, K.-I., KAHNG, B., KIM, D., "Universal Behavior of Load Distribution in Scale-Free Networks", *Physical Review Letters*, v. 87, n. 27, pp. 278701–278704, 2001.

[20] GOOVAERTS, M., KAAS, R., "Evaluating Compound Generalized Poisson Distributions Recursively", *Astin Bulletin*, v. 21, n. 2, pp. 193–198, 1991.

[21] GRADSHTEYN, R., *Table of Integrals, Series, and Products. Alan Jeffrey and Daniel Zwillinger (eds.)(Sixth edition).* Academic Press, 2000.

[22] KIM, J. H., "Poisson cloning model for random graphs", 2005.

[23] KIM, J. H., "Poisson cloning model for random graphs", In: *Proceedings of International Congress of Mathematicians*, v. 3, pp. 873 – 897, European Mathematical Society Publishing House, 2006.

[24] KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., *et al.*, "The Web as a graph", In: *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS*, pp. 1–10, ACM Press, 15–17 2000.

[25] MEDINA, A., MATTA, I., BYERS, J., "On the origin of power laws in Internet topologies", *ACM Computer Communication Review*, v. 30, n. 2, pp. 18–28, 2000.

[26] MOURA, N. J., RIBEIRO, M. B., "Zipf law for Brazilian cities", *Physica A*, v. 367, pp. 441–448, 2006.

[27] NEWMAN, M., BARABASI, A., WATTS, D., *The Structure and Dynamics of Networks.* Princeton, Princeton University Press, 2006.

[28] NEWMAN, M., "Power laws, Pareto distributions and Zipf's law", *Contemporary Physics*, v. 46, pp. 323–351, 2005.

[29] NEWMAN, M., STROGATZ, S., WATTS, D., "Random graphs with arbitrary degree distributions and their applications", *Physical Review E*, v. 64, pp. 026118–1,026118–17, 2001.

[30] ROSS, S. M., *Introduction to Probability Models (Eighth Edition).* Academic Press, 2002.

[31] STAUFFER, A. O., BARBOSA, V. C., *Novas heurísticas para a operação de Redes Complexas.* Dissertação de Mestrado, COPPE/Sistemas, Rio de Janeiro, Brasil, 2005.

[32] STAUFFER, A. O., BARBOSA, V. C., "Dissemination strategy for immunizing scale-free networks", *Physica E*, v. 74, p. 056105, 2006.

[33] STAUFFER, A. O., BARBOSA, V. C., "Local heuristics and the emergence of spanning subgraphs in complex networks", *Theoretical Computer Science*, v. 355, pp. 80–95, 2006.

[34] STAUFFER, A. O., BARBOSA, V. C., "Probabilistic heuristics for disseminating information in networks", *IEEE/ACM Transactions on Networking*, v. 15, pp. 425–435, 2007.

[35] WILF, H. S., *Generatingfunctionology (Second Edition).* Academic Press, 1994.

[36] YAN, S. Y., *Number Theory for Computing.* Springer, 2002.

[37] ZOGHBI, A., STOJMENOVIC, I., "Fast Algorithms for Generating Integer Partitions", *International Journal of Computer Mathematics*, v. 70, n. 2, pp. 319–332, 1999.