Capítulo 7

Introdução a Redes Complexas

Daniel Ratton Figueiredo

Abstract

In its most general definition, a network is an abstraction to encode some kind of relationship among pairs of objects. For example, in social networks objects are usually individuals and relationships are represented by some social tie, such as friendship or co-worker. Networks are everywhere and during the last decade a large number of empirical studies have been identifying peculiar properties in distinct and diverse networks, from the Internet and Online Social Networks (e.g., Facebook) to citation and neural networks. The topological structure of networks has a fundamental role in the functionality and processes exerted by the network. For example, the ranking of relevant web pages or the propagation of gossip through Twitter. Complex Networks *emerges as a multidisciplinary area of Science that aims at studying and understanding the following broad phenomenon: how "things" connect and what are its implications? In this chapter, we present the different aspects and challenges towards achieving this goal. We will first present empirical studies of several real networks, illustrating peculiarities found in their topological structure, such as power law distributions, which give rise to scale-free networks. We continue by presenting mathematical models to represent real networks, capturing their most important topological features. Finally, we present how functionality and processes on networks fundamentally depend on their topological structure.*

Resumo

Em sua definição mais geral, uma rede é uma abstração que permite codificar algum tipo de relacionamento entre pares de objetos. Por exemplo, em redes sociais objetos são geralmente indivíduos e relacionamentos representam algum tipo de relação social, como amizade ou trabalho em conjunto. Redes estão por todos os lados e durante a última década um grande número de estudos empíricos vem identificando propriedades peculiares em redes muito distintas, da Internet e Redes Sociais Online (ex. Facebook) até redes de citação e redes de neurônios. A estrutura topológica das redes possui um papel fundamental pois influenciam diretamente a funcionalidade e os processos que operam sobre a mesma. Por exemplo, o ranqueamento por relevância de páginas Web, ou a propagação de um boato pelo Twitter. Redes Complexas *surgiu como uma área multidisciplinar da Ciência que visa estudar e compreender este abrangente fenômeno: como as "coisas" se conectam e quais são as implicações disto?. Neste capítulo, apresentamos os diferentes aspectos e desafios na direção*

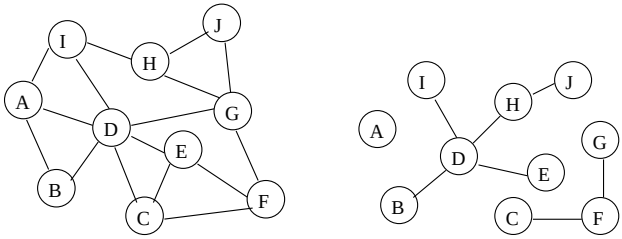
de atingir este objetivo. Iremos começar apresentando estudos empíricos da estrutura de diversas redes reais, ilustrando propriedades peculiares frequentemente encontradas nestas redes, tais como distribuições em lei de potência, que dão origem às redes livre de escala. Em seguida apresentaremos modelos matemáticos que são capazes de representar redes reais, capturando seus aspectos topológicos mais importantes. Por fim, iremos apresentar como algumas funcionalidades e processos que operam em redes dependem fundamentalmente de suas estruturas topológicas.

7.1. Introdução

A percepção de que estamos inseridos e cercados por diversas redes e que estas possuem um papel central em diversos aspectos de nossas vidas, vem crescendo vertiginosamente durante a última década, tanto no meio acadêmico quanto no público em geral. Não é raro encontrar uma notícia na mídia comum onde uma determinada rede possui grande importância à questão sendo abordada. O mesmo ocorre com artigos científicos em diferentes áreas do conhecimento, onde ao considerar um problema, pesquisadores formulam o mesmo de forma que uma determinada rede tenha um papel central e passam então a estudar e caracterizar esta rede. Estamos cada vez mais cientes de que o comportamento de muitas das coisas que nos cercam não pode ser estudado e caracterizado isoladamente, pois muitas destas coisas estão conectadas e a interação entre as partes influencia fundamentalmente o comportamento individual e coletivamente o comportamento global. Assim sendo, surge naturalmente a necessidade de estudar como as coisas se conectam e a importância desta conectividade para o problema em questão.

Mas o que são redes? Existem muitos tipos de redes em muitos domínios diferentes e as coisas que se conectam podem ter muitos significados. Desta forma, estamos interessados em uma definição que seja bem genérica para que possamos estudar todas as redes sobre a mesma perspectiva. De forma mais geral, uma rede é uma abstração que permite codificar algum tipo de relacionamento entre pares de objetos. Podemos assim considerar redes formadas por qualquer tipo de objeto, tais como um conjunto de indivíduos, de páginas Web, de neurônios, ou de computadores. Sobre este conjunto de objetos iremos *codificar* algum relacionamento que geralmente depende do conjunto de objetos. Por exemplo, considerando o conjunto de indivíduos, podemos codificar o relacionamento amizade. Assim, se dois indivíduos são amigos então dizemos que os dois estão relacionados, caso contrário, dizemos que os dois não estão relacionados. Desta forma, para cada par de indivíduos teremos ou não a existência do relacionamento sendo considerado, neste caso, amizade. A figura 7.1a ilustra uma rede de amizade, onde a ligação entre os indivíduos denota a existência do relacionamento. Repare que podemos codificar muitos relacionamentos diferentes, inclusive sobre o mesmo conjunto de objetos. Por exemplo, considerando um conjunto de pessoas, podemos codificar o relacionamento parentesco, trabalho em conjunto ou contato sexual. Diferentes relacionamentos sobre um mesmo conjunto de objetos em geral dão origem a diferentes redes, pois o relacionamento pode não existir exatamente entre os

mesmos pares de indivíduos. A figura 7.1b apresenta a rede de contato sexual sobre o mesmo conjunto de indivíduos, ilustrando que os diferentes relacionamentos deram origem a diferentes redes. Na seção 7.3 iremos apresentar e discutir diversos tipos de redes, tais como redes biológicas, redes de informação e redes tecnológicas.



(a) Relacionamento de amizade.

(b) Relacionamento de contato sexual.

Figura 7.1. Dois diferentes relacionamentos codificados sobre o mesmo conjunto de pessoas dando origem a duas redes distintas.

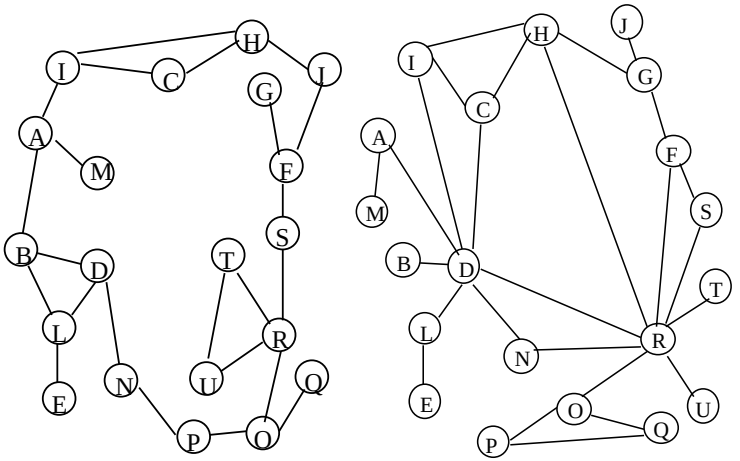
Um aspecto central ao estudo de redes está em descobrir, caracterizar e modelar a estrutura da rede. Informalmente, a estrutura da rede é dada pelo conjunto de relacionamentos existentes, ou seja, todas as "ligações" entre os objetos. Por exemplo, considere todos os alunos matriculados no curso de Engenharia de Computação da UFRJ. Qual é a estrutura da rede de amizade sobre este conjunto de pessoas? Uma primeira dificuldade seria descobrir esta rede, pois não há um repositório central onde podemos obter estas informações. Uma ideia é realizar uma pesquisa de campo com os alunos, perguntando a cada um deles para listar seus amigos no curso. Este método requer a colaboração dos alunos e bastante esforço da pessoa que irá realizar a pesquisa. Outra ideia é tentar inferir os relacionamentos de amizade, por exemplo, usando redes sociais online, como o Facebook, Orkut e Twitter. Neste caso, provavelmente não teríamos a rede de amizade completa, pois algumas pessoas ou alguns relacionamentos de amizade podem não estar nestas redes sociais online. Por outro lado, este método pode levar a uma rede bem próxima da real rede de amizade com bem menos esforço. Mas vamos supor que temos a estrutura desta rede de amizade.

Para que serve a estrutura de uma rede? Por que devemos caracterizar e modelar a estrutura de uma rede? Existem uma série de fenômenos que operam sobre as redes e que dependem fundamentalmente da estrutura das mesmas. Desta forma, para entender o comportamento de tais fenômenos é necessário antes entender a estrutura das redes. Por exemplo, considere um

sociólogo interessado em estudar como boatos se espalham por uma sociedade ou um conjunto de pessoas. Em particular, ele deseja estudar quanto tempo leva para um boato introduzido por um indivíduo chegar aos ouvidos de todos os outros indivíduos do conjunto. Além disso, ele quer saber se este tempo depende de quem introduz o boato. E mais ainda, ele quer identificar os indivíduos que são mais eficazes em espalhar boatos. O fenômeno de espalhamento de boatos é fundamentalmente influenciado pela estrutura da rede por onde o boato irá se espalhar. Vamos voltar a rede de amizade descoberta acima, entre os alunos do curso de graduação, e vamos supor que boatos se propaguem apenas por esta rede. Ou seja, indivíduos só contam boatos para seus amigos! Considere duas hipóteses para a estrutura da rede de amizade descoberta, ilustradas nas figuras 7.2a e 7.2b. Em qual delas o boato se espalha mais rapidamente? Em qual delas o espalhamento do boato depende de quem o inicia? Quem é a pessoa mais eficaz em espalhar os boatos? Para responder a estas perguntas com exatidão, precisaríamos definir como boatos são transmitidos entre os indivíduos. Por exemplo, uma pessoa irá contar o boato para todos os seus amigos ou somente para alguns deles, e neste caso quais? Apesar disto, podemos intuitivamente concluir que na rede A os boatos irão demorar mais para se espalhar e que não há muitas diferenças entre os indivíduos que começam o boato. Por outro lado, os boatos se espalharão bem mais rapidamente na rede B e os indivíduos neste caso são bem diferentes com relação ao tempo que leva para seus boatos se espalharem. Está claro que a estrutura da rede pode influenciar fundamentalmente fenômenos que operam sobre elas, como neste caso. Iremos apresentar alguns fenômenos importantes na seção 7.5 e discutir como estes dependem da estrutura da rede.

Um aspecto importante no estudo da estrutura das redes é a caracterização de sua estrutura. Caracterizar a estrutura de uma rede significa enumerar diversos aspectos que capturem e resumem a estrutura da rede. Apesar da estrutura codificar toda a informação de uma rede, em muitos casos esta informação é demasiada e não necessariamente ajuda a entender suas características. Além disso, uma propriedade estrutural pode ser fundamental, pois sua presença na rede leva a mesma a possuir alguma característica que independe do restante da sua estrutura. Isto é importante, pois podemos entender determinadas características apenas olhando para algumas propriedades estruturais. Desta forma, diferentes propriedades estruturais são usadas para nos dar uma ideia da estrutura da rede.

Vejamos uma analogia desta ideia com o nosso cotidiano. Imagine que você viu uma pessoa bem peculiar no ônibus a caminho da universidade. Ao chegar, você quer comentar desta esta pessoa para algum amigo. O que você faz? Muito provavelmente você irá descrever propriedades fundamentais da pessoa, tais como sexo, silhueta, altura, cor do cabelo, cor dos olhos, etc. A "estrutura" completa desta pessoa seria seu DNA, que apesar de conter todas estas informações, é pouco útil para construir uma imagem geral da pessoa.



(a) Hipótese de rede A. (b) Hipótese de rede B.

Figura 7.2. Duas possíveis redes de amizade.

Além disso, ao conhecer, por exemplo, o sexo da pessoa você pode concluir características que independem do resto da "estrutura" (DNA) da pessoa. Com redes acontece exatamente a mesma coisa, ainda mais quando estudamos redes muito grandes, quando a estrutura é gigantesca, do mesmo tamanho do DNA humano, com bilhões de relacionamentos codificados. Iremos usar diferentes propriedades estruturais para falar descrever as redes. Mas que propriedades são estas?

Da mesma forma como existem muitas maneiras de você descrever uma pessoa, existem muitas (na verdade infinitas) propriedades estruturais que podem ser usadas para caracterizar uma rede. Obviamente, algumas propriedades podem ser mais importantes do que outras, pois definem aspectos que são mais gerais e que possuem uma maior influência sobre a rede. A mais simples propriedade de uma rede é seu tamanho, ou seja, o número de objetos e o número de relacionamentos existentes. Na figura 7.2a, a rede possui $n = 20$ objetos e $m = 23$ instâncias do relacionamento (usaremos frequentemente n para denotar o número de objetos e m o número de relacionamentos presentes em uma rede). Com isto podemos definir a densidade da rede, que é dada pela fração de relacionamentos que existem na rede. Repare que a rede da figura 7.1a é mais densa do que a rede da figura 7.1b. Uma outra propriedade muito importante é a fração de relacionamentos que cada objeto possui. Repare que na rede ilustrada na figura 7.2b uma única pessoa possui uma grande

fração dos relacionamentos existentes. Estas e muitas outras propriedades serão apresentadas formalmente e discutidas na seção 7.2.

Dado as diferentes propriedades estruturais de uma rede, estamos interessados em entender como uma ou outra propriedade influencia o comportamento dos fenômenos que podem ocorrer na rede. Existem muitas vantagens em entender o comportamento de um fenômeno apenas conhecendo algumas propriedades estruturais da rede. Isto permite generalizar observações, pois basta que a rede possua tal propriedade para que tal fenômeno se comporte de tal maneira, sem que seja necessário conhecer maiores detalhes da estrutura da rede. Vejamos o estudo do sociólogo acima sobre espalhamento de boatos. Vimos, intuitivamente, que boatos se espalham mais rapidamente em B, mas para chegar a esta conclusão tivemos que inspecionar a estrutura da rede. Nosso sociólogo gostaria muito de poder generalizar suas observações. Ou seja, ele gostaria de poder dizer que se a rede de amizade possuir uma determinada propriedade estrutural, então boatos irão se espalhar muito rapidamente (ex. exponencialmente rápido). Desta forma, não precisamos conhecer os detalhes da estrutura da rede, e sim apenas se a mesma possui esta determinada propriedade estrutural para que boatos se espalhem muito rapidamente. Esta generalização é um dos mais importantes aspectos no estudo de redes, pois permite entender o comportamento de fenômenos apenas conhecendo algumas propriedades estruturais da rede. Na seção 7.5 iremos discutir a importância de algumas destas propriedades em alguns fenômenos em redes.

Mas como podemos generalizar as redes, uma vez que as redes são distintas? Por exemplo, a rede de amizade dos alunos no curso de Engenharia de Computação da UFRJ é diferente da rede de amizade dos alunos do mesmo curso na UFRN, que é diferente da rede de amizade de qualquer outro curso de Engenharia de Computação. Elas nem sequer tem o mesmo tamanho! Mas não surpreendentemente, estas redes apesar de serem distintas, possuem algumas propriedades estruturais que são praticamente idênticas. Isto ocorre pois estas redes são todas geradas a partir do mesmo *processo*. No caso, o processo que gerou estas redes é o processo de formação de amizades em nossa sociedade, que apesar das suas muitas sutilezas, de forma geral independe da Universidade. Assim sendo, para poder generalizar resultados teremos que trabalhar não com as redes específicas mas sim com os processos que dão origem a elas. Iremos representar os processos que dão origem a redes através de modelos matemáticos que simplesmente determinam matematicamente como a rede será construída. Redes construídas por um mesmo processo não necessariamente são idênticas, como no caso das redes de amizades acima. Desta forma, o modelo matemático para construir estas redes também deve ser capaz de dar origem a redes diferentes e para isto terá de ser aleatório.

Apesar de um modelo aleatório para redes dar origem a redes distintas, estas podem sempre possuir alguma determinada propriedade estrutural. E isto

então nos permite generalizar! Ou seja, redes reais bem representadas por um modelo aleatório que sempre induz uma determinada propriedade estrutural vão sempre ter um determinado comportamento funcional que depende desta propriedade. Nosso sociólogo tem que representar as redes de amizades com um modelo matemático e estabelecer que este sempre induz uma determinada propriedade estrutural. Ele também deve estabelecer que esta propriedade garante que boatos se espalham rapidamente. Logo, ele conclui que boatos se espalham muito rapidamente em *qualquer* rede de amizade. Ele vai ficar famoso! Na seção 7.4 iremos apresentar alguns modelos matemáticos aleatórios de redes muito utilizados e discutir diversas de suas propriedades estruturais.

7.2. Formalizando as redes

Antes de continuarmos, precisamos formalizar o conceito de redes para então definirmos matematicamente suas propriedades estruturais. Como vimos, uma rede é um abstração que permite codificar um relacionamento entre pares de objetos. O conjunto de objetos da rede será denotado por V e o número de objetos será dado por $n = |V|$. Iremos chamar estes objetos de vértices ou nós da rede. A existência de um relacionamento entre dois objetos será representado por um par não-ordenado. Desta forma, se $i, j \in V$ estão relacionados, iremos representar isto pelo par não-ordenado (i, j). Este par-ordenado, que representa a existência do relacionamento, será chamado de aresta. O conjunto com todos os relacionamentos existentes entre os objetos em V, ou seja, o conjunto de arestas da rede, será denotado por E. Desta forma, temos $E = \{(i, j)|i, j \in V, i$ está relacionado com $j\}$. O número de arestas da rede, m, é dado pelo número de pares não-ordenados no conjunto E, ou seja $m = |E|$. Finalmente, uma rede será definida por estes dois conjuntos, $R = (V, E)$. Na computação esta definição é equivalente a definição de um grafo, e na verdade, uma rede nada mais é do que um grafo!

Até o momento tratamos apenas com relacionamentos simétricos. Um relacionamento é simétrico se não existe distinção entre a ordem em que ele se apresenta. Ou seja, i está relacionado com j se e somente se j está relacionado com i. Exemplos de relacionamento simétricos são trabalhar em um mesmo projeto, contato sexual, vôo comercial entre cidades, entre muitos outros que veremos em breve. Entretanto, nem todo relacionamento é simétrico, sendo muito deles assimétricos. Neste caso, o fato de um objeto i estar relacionado com outro j não implica que j esteja relacionado com i. Exemplos de relacionamentos assimétricos são amizade, hiperlinks entre páginas web, troca de email, entre muitos outros. Entretanto, podemos facilmente estender a definição acima para lidar com relacionamentos assimétricos. Basta para isto representar o relacionamento como um par ordenado, onde a ordem dos vértices no par é importante. Ou seja, o par (i, j) é diferente do par (j, i), pois codificam dois possíveis e distintos relacionamentos.

Podemos representar uma rede, ou seja, representar sua estrutura, através

de uma matriz. Esta matriz codifica todas as arestas da rede e é conhecida como matriz de adjacência, denotada aqui por A. A matriz A é quadrada de o n e cada elemento $A(i,j)$ representa o par de vértices (i,j). Se o par estiver relacionado, então temos que $A(i,j) = 1$, caso contrário $A(i,j) = 0$. Podemos então definir a matriz A da seguinte forma:

$$A(i,j) = \begin{cases} 1 & \text{se } (i,j) \in E \\ 0 & \text{caso contrário} \end{cases} \quad (1)$$

Repare que se a rede estiver codificando um relacionamento simétrico, então a matriz de adjacência será simétrica com relação a sua diagonal principal. Pois neste caso, temos que $A(i,j) = A(j,i)$. Por outro lado, se o relacionamento for assimétrico, a matriz não será simétrica. A matriz de adjacência da rede de amizade ilustrada na figura 7.1a é dada por:

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix} \quad (2)$$

Onde as letras que identificam os vértices na figura foram mapeadas em números representando sua posição no alfabeto (A=1, B=2, etc).

Iremos chamar de *grau* o número de arestas que incidem sobre um determinado vértice da rede. Repare que o grau de um vértice representa o número de relacionamentos que existem com o vértice. Na figura 7.1a, o grau do vértice D é seis, pois existem seis relacionamentos de amizade. O grau de um vértice pode ser obtido somando os elementos da sua respectiva linha na matriz de adjacência. Repare que ao somar os elementos da linha 4 na matriz acima, obtemos o grau do vértice D, que é 6. Caso a matriz seja simétrica o grau também pode ser obtido somando a respectiva coluna, já que a linha i é idêntica a coluna i em uma matriz simétrica. Iremos denotar por d_i o grau do vértice $i \in V$.

É importante notar que a soma do grau de todos os vértices é igual ao dobro de número de arestas da rede. Ou seja,

$$\sum_{i \in V} d_i = 2m \quad (3)$$

Isto ocorre pois cada aresta possui duas pontas e desta forma é contada duas vezes quando somamos os graus de todos os vértices. Podemos ainda obter

o grau médio de uma rede, que é dada pela média aritmética do grau de todos os vértices. Ou seja,

$$\overline{d} = \frac{1}{n} \sum_{i \in V} d_i = \frac{2m}{n} \tag{4}$$

Repare que podemos obter o grau médio da rede simplesmente usando o resultado da equação (3), sendo necessário conhecer somente o número de arestas da rede. Além disso, sabemos que o grau de um vértice sempre está entre 0 e n-1, caso ele não possua relacionamentos ou se relacione com todos os outros vértices da rede.

Usando o número de vértices e arestas de uma rede, podemos definir a densidade da rede, ρ, que representa a fração de arestas que a rede possui. Considere uma rede com n vértices e m arestas. Qual é a sua densidade? Para calcularmos a densidade, precisamos determinar o maior número de arestas que a rede poderia ter. Isto ocorre quando todos os vértices estão relacionados entre si, ou seja, cada vértice possui o maior grau possível que é $n-1$. Como temos n vértices, o total de arestas será $n(n-1)/2$ e precisamos dividir por dois pois cada aresta na multiplicação $n(n-1)$ está sendo contada duas vezes. Podemos agora calcular ρ da seguinte forma:

$$\rho = \frac{m}{n(n-1)/2} = \frac{\overline{d}}{n-1} \tag{5}$$

Repare que usamos o resultado da equação (4) para simplificar a expressão acima, e desta forma temos a densidade da rede apenas em função do grau médio e do número de vértices.

7.2.1. Distribuição de grau

Uma importante propriedade estrutural de qualquer rede é sua distribuição de grau. A distribuição empírica[1] de grau é fração de vértices da rede que possui determinado grau. Considere uma rede $R = (V, E)$ e seja n_k o número de vértices com grau igual a k. A fração de vértices com grau k é simplesmente dada por:

$$f_k = \frac{n_k}{n} \tag{6}$$

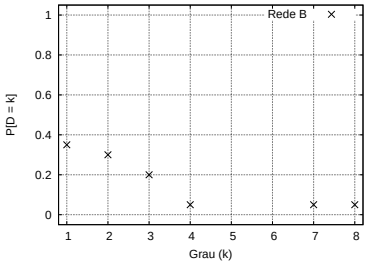
Em geral estaremos interessados na distribuição complementar cumulativa (CCDF) do grau. Ou seja, estamos interessados na fração de vértices que tem grau maior ou igual a k. Podemos obter este valor utilizando a equação (6), somando para todos os graus menores do que k e obtendo o complemento.

[1] Usaremos o termo empírica quando estivermos tratando de uma rede real onde a distribuição é obtida utilizando os graus da rede em questão. Omitiremos o termo quando estivermos tratando de uma função de distribuição de probabilidade, proveniente de um modelo matemático. Entretanto, às vezes omitiremos o termo quando o contexto estiver claro.

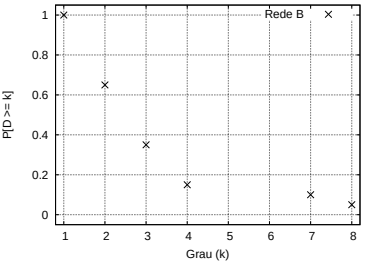
Ou seja,

$$F_k = 1 - \sum_{i=0}^{k-1} f_i \tag{7}$$

A razão para utilizarmos a CCDF está relacionado ao fato de que muitas redes reais possuem uma distribuição com cauda pesada ou até mesmo lei de potência (veremos o que significa isto na seção 7.4.1). Este tipo de distribuição é melhor visualizado utilizando o gráfico da CCDF.



(a) Distribuição de grau (convencional), equação (6).



(b) Distribuição complementar cumulativa (CCDF), equação (7).

Figura 7.3. Distribuição empírica de grau da rede ilustrada na 7.2b.

A figura 7.3 apresenta a distribuição de grau da rede ilustrada na figura 7.2b. Repare que a figura 7.3a apresenta a distribuição de grau, ou seja, a fração relativa de vértices com cada determinado grau, enquanto a figura 7.3b apresenta a distribuição complementar cumulativa. Podemos verificar que o grau mais frequente é 1, presente em 35% dos vértices. Também podemos verificar que o maior grau é 8 e que apenas um vértice, ou seja 5% dos vértices, possui este grau. Uma característica importante é que a CCDF é sempre uma função decrescente, conforme ilustrado no gráfico.

7.2.2. Caminhos e distância

Um outro aspecto importante de uma rede são seus caminhos. Um caminho é definido por uma sequência de vértices sem repetição onde existe uma aresta entre cada par de vértices adjacentes na sequência. Por exemplo, na figura 7.1a a sequência A,B,D,C,F é um caminho entre o vértice A e F, assim como a sequência A,I,H,J,G,F. Intuitivamente, um caminho é uma sequência de saltos pelas arestas da rede. O comprimento de um caminho é definido pelo número de arestas que o define, ou equivalentemente, pelo número de vértices da sequência menos um. Por exemplo, o comprimento dos caminhos A,B,D,C,F e A,I,H,J,G,F são 4 e 5, respectivamente.

Intuitivamente, existem muitos caminhos diferentes entre um determinado par de vértices em uma rede qualquer. Por exemplo, entre os vértices A e F da rede na figura 7.1a temos sete caminhos diferentes (você consegue enumerá-los?). Naturalmente, estes diferentes caminhos podem possuir diferentes comprimentos. Vamos definir a distância entre dois vértices como sendo o comprimento do menor caminho entre eles. No exemplo, a distância entre os vértices A e F é 3, dada pelo caminho A,D,E,F. Repare que entre quaisquer dois vértices, temos apenas uma distância, apesar de ser possível ter mais de um caminho com este comprimento. Por exemplo, a distância entre C e G na figura 7.1a é 2 que é obtida pelos caminhos C,D,G ou C,F,G. Veremos em breve que distâncias possuem um papel central no estudo da estrutura de redes, pois capturam o quão próximo os vértices da rede estão uns dos outros.

Uma importante propriedade estrutural de uma rede é sua distância média e seu diâmetro. A distância média é dada pela média aritmética das distâncias entre todos os pares de vértices da rede. Seja $l(i,j)$ a distância entre os vértices $i,j \in V$. A distância média $\overline{l}$ é definida por:

$$\overline{l} = \frac{\sum_{i,j \in V} l(i,j)}{\binom{n}{2}} \tag{8}$$

Repare que estamos considerando todos os pares não-ordenados, que ao todo são $\binom{n}{2}$. Uma outra propriedade relacionada é o diâmetro da rede, definida como sendo a maior distância entre qualquer par de vértices da rede. Desta forma, temos:

$$L = \max_{i,j \in V} l(i,j) \tag{9}$$

Repare que assumimos implicitamente nos cálculos acima que a distância entre qualquer dois vértices está definida. Entretanto, muitas redes não são *conexas*, ou seja, elas não possuem caminhos entre todos os pares de vértices. Logo, a distância entre estes pares de vértices não está definida. Neste caso, o cálculo da distância média e do diâmetro devem desconsiderar estes pares ou considerar apenas os pares que pertencem a maior componente conexa, ou seja, ao maior pedaço conectado.

A distância média da rede ilustrada na figura 7.1a é $82/\binom{10}{2} = 1.82$ e seu diâmetro é 3. Repare que esta é uma rede onde os vértices estão muito "próximos" uns dos outros, pois tanto a distância média quanto o diâmetro são relativamente baixos.

7.2.3. Coeficiente de clusterização

Uma característica importante de uma rede é sua redundância ou correlação das arestas ao redor de um vértice. Considere um vértice que está relacionado a dois outros. Quais as chances destes dois também estarem relacionados? Este propriedade possui importantes implicações, como veremos na seção 7.4. Uma métrica que captura esta ideia é o coeficiente de clusterização da rede. Iremos definir o coeficiente de clusterização de um vértice $i \in V$ como

sendo a fração de arestas que os vizinhos de i possuem entre si e o máximo de arestas que eles poderiam possuir entre si. Dado que o grau do vértice i é d_i, o maior número de arestas entre seus vizinhos é dado por $\binom{d_i}{2}$. Ou seja, todos os pares de vizinhos de i possuem aresta entre si. Seja E_i o número efetivo de arestas entre os vizinhos do vértice i. Podemos definir o coeficiente de clusterização do vértice i, c_i, como:

$$c_i = \frac{E_i}{\binom{d_i}{2}} = \frac{2E_i}{d_i(d_i-1)} \tag{10}$$

Repare que o coeficiente de clusterização não está definido para vértices com grau zero ou grau um. A definição acima se aplica somente a vértice com grau maior do que um. Utilizando o coeficiente de clusterização dos vértices, podemos agora definir o coeficiente de clusterização da rede como sendo a média aritmética destes. Ou seja,

$$\overline{c} = 1/n \sum_{i \in V} c_i \tag{11}$$

Repare que na média acima assumimos que todos os coeficientes estavam definidos, o que nem sempre é o caso. Nestes casos, a média deve ser apenas sobre os vértices que possuem coeficientes de clusterização definidos.

Considere a rede ilustrada na figura 7.1a. Os coeficientes de clusterização dos vértices A, B, C e D são 2/3, 1, 2/3 e 1/5, respectivamente. Repare que o vértice D tem baixa clusterização, pois existem poucas arestas entre seus muitos vizinhos. O coeficiente de clusterização da rede é $\overline{c} = 0.537$.

Por fim, é importante ressaltar que existem outras definições para o coeficiente de clusterização de uma rede. Entretanto, iremos trabalhar com a definição acima, as vezes chamada de coeficiente de clusterização local, apesar desta ter algumas desvantagens [Newman 2010].

7.2.4. Outras propriedades estruturais

Exite na literatura um bom número de propriedades estruturais que vem sendo utilizadas para caracterizar uma rede. Na verdade, existe um número infinito delas e você pode até mesmo definir a sua propriedade estrutural, e deve inclusive fazer isto mesmo quando for o caso. Podemos sempre enumerar e calcular mecanicamente muitas propriedades estruturais de uma rede. Entretanto, um aspecto bem mais importante é interpretar e compreender o significado das propriedades estruturais sendo observadas. O que significa que tal propriedade assume tais valores? Como isto está relacionado com a rede em questão? Outro aspecto fundamental é determinar quais propriedades estruturais são importantes. Mas o que é uma propriedade estrutural importante? Bom, uma propriedade estrutural é importante se ela revela algo interessante da rede em questão, ou melhor, de funcionalidades que a rede exerce. Repare que isto depende da rede e da funcionalidade que queremos estudar. Em todo o caso, você deve conhecer algumas outras propriedades estruturais que vem

sendo amplamente utilizadas, tais como reciprocidade, centralidade, *betweeness*, e *assortativity* [Newman 2010, Boccaletti et al. 2006].

Agora que definimos formalmente alguns conceitos em redes e apresentamos algumas propriedades estruturais, vamos voltar a falar sobre redes reais e conhecer algumas de suas importantes características estruturais.

7.3. Redes reais

Redes estão por todos os lados e em todos os domínios do conhecimento. O estudo empírico destas redes reais vem se tornando cada vez mais abrangente não apenas em relação ao que elas representam mas também com relação a seus tamanhos. Existem hoje estudos empíricos que vão desde a rede de neurônios de uma bactéria até a rede de contato do Twitter, com mais de 50 milhões de pessoas. No que segue iremos apresentar algumas importantes redes de diferentes domínios e descrever algumas das propriedades estruturais. Iremos ver que algumas destas propriedades aparecem em redes que aparentemente não possuem nenhuma relação.

7.3.1. Redes Sociais

Redes sociais são redes formadas por pessoas ou grupos de pessoas e por algum tipo de relacionamento. Por exemplo, a rede de amizade e a rede de contato sexual ilustradas na figura 7.1 são exemplos de redes sociais. Uma outra rede social bastante estudada são as redes de colaboração, cujo relacionamento é algum tipo de colaboração entre as pessoas. Na rede de colaboração científica, os vértices são pesquisadores e as arestas indicam algum tipo de colaboração científica, como por exemplo, publicação de artigos em conjunto. Neste caso, dois vértices estão relacionados se os dois pesquisadores são co-autores de ao menos um mesmo artigo científico. Na rede de colaboração artística, os vértices são artistas e as arestas algum tipo de colaboração artística. Por exemplo, os vértices podem ser atores de cinema e o relacionamento pode ser trabalhar (i.e., atuar) em um mesmo filme. Desta forma, dois atores estão relacionados se ambos atuaram em ao menos um filme em comum.

Assim como muitas redes, a rede de colaboração entre um conjunto grande de pessoas geralmente não está prontamente disponível e precisa ser inferida utilizando alguma outra informação. Por exemplo, a rede de colaboração científica pode ser inferida utilizando repositórios de dados ou bibliotecas digitais que cadastrem de forma estruturada artigos científicos publicados. Um repositório bem estruturado, de acesso público e relativamente grande é o DBLP, que hoje contém 1.6 milhão de artigos cadastrados principalmente da área de Computação [Ley 2011]. Entretanto, assim como muitos repositórios, o DBLP não é completo e muitos artigos publicados, principalmente em algumas subáreas da computação, não estão cadastrados. Em todo o caso, o DBLP é repositório importante e vem sendo usada em muitos estudos, inclusive em conjunto com outros repositórios [Newman 2001, Menezes et al. 2009, Freire e Figueiredo 2010].

A figura 7.4 ilustra a rede de colaboração dos professores do PESC (Pro-

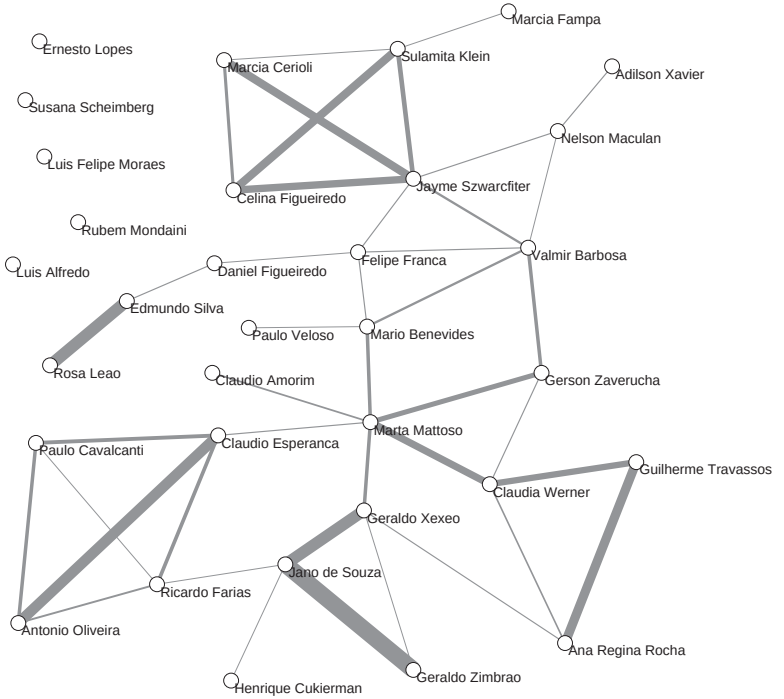


Figura 7.4. Parte da rede de colaboração do PESC-COPPE/UFRJ.

grama de Engenharia de Sistemas e Computação) na COPPE/UFRJ (desenhada utilizando a ferramenta Pajek). Esta rede foi inferida utilizando o DBLP em meados de 2009 considerando todos os professores do departamento na época[2]. Podemos observar diversos aspectos interessantes. Por exemplo, a rede está quase toda conectada, com 85% dos vértices pertencendo a maior componente conexa. O número de colaborações (grau) dos vértices é bem variado assim como a intensidade de colaboração, ilustrada pela espessura das arestas[3], mostrando que colaborar está longe de ser um processo uniforme. A

[2] É importante ressaltar que esta não é a real rede de colaboração do PESC, pois contempla apenas os artigos cadastrados no DBLP. Colaborações provenientes de outros artigos publicados por professores do PESC não foram considerados nesta figura.

[3] A métrica de intensidade de colaboração sendo usada foi definida por Newman e é proporcional ao número de artigos publicados em co-autoria normalizada pelo número

rede também nos ajuda a identificar diferentes grupos de pesquisa. Por exemplo, os professores Antonio, Claudio, Paulo e Ricardo formam uma subrede completamente conectada (i.e., um *clique*), e na verdade são os quatro professores do grupo de Computação Gráfica do PESC (na época em que os dados foram coletados). De forma similar, os professores Celina, Marcia, Sulamita e Jayme, representam o grupo de Algoritmos e Combinatória do departamento. Podemos observar ainda que algumas pessoas tem um papel central na rede, pois suas colaborações ajudam a manter a rede conectada e encurtam distâncias na rede. Ainda outras características podem ser observadas, e veremos algumas a seguir. Por fim, estas características não são particulares da rede de colaboração do PESC/COPPE, podendo ser encontradas em muitas redes de colaboração. Esta generalização é um indicativo que estas propriedades são consequências do processo de colaboração científico.

Um aspecto interessante e muito estudado em redes sociais é a distância entre as pessoas na rede. Podemos citar dois estudos populares nesta direção, conhecidos inclusive por leigos, que são o número de Bacon e o número de Erdős [Reynolds 2011, Grossman e Ion 2011]. Estes dois números medem a distância de um ator de cinema e de um pesquisador ao ator Kevin Bacon e ao matemático Paul Erdős, respectivamente, na rede de colaboração de atores e de pesquisadores. Um ator que tem número de Bacon igual a 1 atuou em ao menos um filme com Kevin Bacon. Similarmente, um pesquisador que tem um número de Erdős igual a 1 foi co-autor de Paul Erdős em ao menos um artigo científico. Um ator com número de Bacon igual a 2 atuou em um filme com algum outro ator que tem número de Bacon igual a 1 e nunca atuou com um ator com número de Bacon menor (que no caso seria o próprio Bacon). Analogamente para um pesquisador cujo número de Erdős é igual a 2.

A tabela 7.1 mostra a distribuição do número de Erdős para uma população de pesquisadores matemáticos [Grossman e Ion 2011]. Repare que apesar de Erdős ter apenas 504 colaboradores, as distâncias até os outros vértices é relativamente pequena. A mediana e a média da distribuição são 5 e 4.65, respectivamente. Ou seja, na média um pesquisador matemático está a 5 colaboradores de Erdős. Repare ainda que a cauda da distribuição é bem curta, pois uma fração muito pequena da população possui número de Erdős maior do que 7, sendo 13 a maior distância até Erdős, em comparação com os centenas de milhares de vértices da rede. É importante ressaltar que nem todos os pesquisadores possuem um caminho até Erdős pela rede de colaboração científica, apesar deles terem publicado com outros co-autores. Neste caso dizemos que estes pesquisadores possuem número de Erdős infinito. Isto ocorre por que a rede não está totalmente conectada e é formada por múltiplos componentes conexos.

Esta propriedade que observamos com o número de Erdős, que é ter distâncias muito pequenas entre os vértices da rede, não é uma característica

de autores em cada artigo [Newman 2004b].

Tabela 7.1. Distribuição do número de Erdős entre uma população de pesquisadores matemáticos (mediana 5, média 4.65, desvio padrão 1.21) [Grossman e Ion 2011]

# Erdős	1	2	3	4	5	6	7
# pessoas	504	6593	33605	83642	87760	40014	11591
# Erdős	8	9	10	11	12	13	∞
# pessoas	3146	819	244	68	23	5	50000

particular de Erdős nem da rede de colaboração científica matemática. Por exemplo, o mesmo fenômeno é observado com o número de Bacon e em outras redes de colaboração estudadas [Newman 2004a]. Além disso, de mudarmos o vértice inicial, considerando outro pesquisador, mesmo que bem menos prolífico do que Paul Erdős, observaremos novamente a mesma propriedade. Na verdade, distâncias curtas entre os vértices é uma propriedade estrutural inerente a muitas redes sociais, inclusive a rede de amizade.

Estudos empíricos realizados pelo influente professor de psicologia-social Stanley Milgram na década de 60 ajudou a corroborar esta ideia [Milgram 1967, Travers e Milgram 1969]. Ele solicitou indivíduos residentes no interior dos Estados Unidos que enviassem uma carta a uma determinada pessoa, informando aos indivíduos o nome completo, profissão e endereço postal do destinatário. Entretanto, os sujeitos do experimento só podiam enviar a carta para pessoas que elas conhecessem pessoalmente. No improvável caso de um sujeito conhecer o destinatário, ele poderia enviar a carta diretamente. Caso contrário, o sujeito deveria encaminhar a carta a algum conhecido. Este por sua vez repetia o procedimento, enviando a carta para o destinatário, caso ele o conhecesse pessoalmente, ou encaminhando a carta para outra pessoa. Milgram realizou o experimento com diversos sujeitos de diferentes cidades de origem e alguns diferentes destinatários. Em todos os seus experimentos ele observou algo surpreendente. As cartas que chegavam ao destinatário faziam um número médio de saltos entre cinco e seis. Ou seja, a distância pela rede social americana entre duas pessoas escolhidas relativamente ao acaso era em torno de seis. Este estudo, conhecidos por experimento de mundo pequeno, ajudou a popularizar o conceito de “seis graus de separação”, que se refere a ideia de que todas as pessoas do mundo estão em média a seis saltos de qualquer outra pessoa na rede de amizades.

Um outro aspecto interessante dos experimentos small world é que não apenas caminhos curtos existem pela rede social, mas que nós somos capazes de encontrá-los utilizando apenas informação local. Os sujeitos dos experimentos small world decidiam por conta própria para quem encaminhar a carta quando não conheciam o destinatário. Obviamente, eles não possuíam uma visão global da rede social e suas decisões eram puramente locais e in-

dependentes do restante do experimento. Por exemplo, um sujeito poderia encaminhar a carta para algum conhecido que morasse na mesma cidade do destinatário, ou para algum conhecido que tivesse a mesma profissão do destinatário. Como fazemos para navegar a rede social? Será que a estrutura da rede social facilita esta navegação? Esta observação assim como estas perguntas foram feitas e respondidas por Jon Kleinberg [Kleinberg 2000], e voltaremos ao assunto na seção 7.5.2.

Experimentos mais recentes utilizando email ao invés do correio postal, muito mais abrangentes e com muito mais participantes, foram realizados mais recentemente por pesquisadores da Universidade de Columbia [Dodds et al. 2003]. Apesar dos avanços tecnológicos na forma de comunicação nos últimos 50 anos, muitas das observações empíricas deste estudo corroboram os resultados de Milgram, ajudando a mostrar que realmente vivemos em um mundo pequeno.

O que você esperaria do coeficiente de clusterização de uma rede social? Quais as chances de dois amigos seus serem amigos entre si? Qual a chance de duas pessoas para as quais você enviou email no último ano também terem trocado email? Qual a chance de dois co-autores seus de artigos científicos também serem co-autores de um artigo? Qual a chance de duas pessoas que estão na sua lista de contatos também estarem em suas respectivas listas de contatos? Se sua intuição lhe diz que estas chances devem ser relativamente altas, então você acertou. Redes sociais estão cheias de triângulos! Esta é uma característica fundamental presente em muitas redes sociais que vem sendo estudadas [Newman 2010, Boccaletti et al. 2006]. Como exemplo, o coeficiente de clusterização da rede de colaboração científica apresentada acima é de 0.14.

De fato, não é raro encontrar redes sociais com coeficiente de clusterização maior do que 1/5. Mas 1/5 não parece ser um número muito alto. Em absoluto nenhuma quantidade é grande ou pequena, pois ficamos sem parâmetro para comparação! Desta forma, iremos comparar o coeficiente de clusterização de redes reais com o coeficiente induzido por um modelo matemático aleatório, a ser apresentado na seção 7.4.2. Comparado com este modelo, veremos que o valor 1/5 é extremamente alto.

Além disso, redes sociais possuem relativamente poucas arestas, quando comparado ao número de arestas que elas poderiam ter. Considere a rede de colaboração científica entre matemáticos, discutida acima. A maior componente conexa desta rede possui aproximadamente 268 mil vértice e 634 mil arestas, dando origem a um grau médio de 4.73 [Grossman e Ion 2011]. Repare que o maior número de arestas que a rede poderia ter é $\binom{268000}{2}$, que é aproximadamente 35 milhões. Logo, a densidade da rede, definia na equação (5) é de apenas $\rho = 1.7 \times 10^{-5}$.

Dado que redes sociais geralmente possuem tão poucas arestas, é de certa forma surpreendente que observamos ao mesmo tempo distância curtas e alta clusterização. Isto significa que as arestas da rede são “colocadas” de forma

criteriosa, fechando muitos triângulos para permitir uma alta clusterização e ao mesmo tempo cobrindo a rede rapidamente para permitir distâncias curtas. Voltaremos a discutir este assunto na seção 7.4.4 quando falarmos sobre o modelo de redes proposto por Watts e Strogatz.

7.3.2. Redes de informação

Em um mundo onde a informação tem um papel central cada vez maior, não é surpreendente que praticamente todos os dias nós interagimos com diferentes redes de informação. Em uma rede de informação os vértices representam alguma informação e as arestas representam algum tipo de relacionamento. Uma das maiores e mais populares redes de informação é a Web. Na rede da Web os vértices são páginas da Web identificadas por uma URL e as arestas representam os hiperlinks entre as páginas (entre URLs). Repare que o relacionamento neste caso é assimétrico, pois os hiperlinks da Web são direcionados: um hiperlink da página i para a página j não implica em um hiperlink no sentido contrário.

Um aspecto interessante é que não conhecemos a real estrutura da rede da Web. A Web é um sistema distribuído formado por milhões de servidores Web operados por milhares de empresas onde qualquer pessoa pode adicionar, modificar e remover páginas web e hiperlinks de forma independente. Não existe neste sistema distribuído um repositório central com todas as informações que estão na Web. Desta forma, a rede da Web, assim como muitas outras redes, precisam ser inferidas através de medições. Uma forma de medir e obter a rede da Web é através da navegação pela própria Web. Iniciando com um conjunto de páginas, podemos seguir os hiperlinks destas para descobrir outras, e destas outras seguimos os hiperlinks para mais outras, e assim por diante. Esta técnica é conhecida como *crawling* e é muito usada para medir redes, e não apenas a rede da Web.

Mas quais propriedades estruturais interessantes tem a rede da Web? Esta rede é enorme com dezenas de bilhões de páginas e trilhões de hiperlinks [de Kunder 2011]. Apesar disto, e de certa forma surpreendentemente, as distâncias entre duas páginas Web são muito pequenas, exponencialmente menores do que o tamanho da Web. Ou seja, se a Web tem n páginas, então a distância média entre as páginas é da ordem de $\log n$ [Albert et al. 1999, Broder et al. 2000]. Isto significa que em média precisamos de poucos clicks em hiperlinks para navegar de uma página a qualquer outra! Este fato se torna mais interessante quando pensamos em como a Web é construída. Não existe uma recomendação ou norma que estabelece como uma pessoa ou empresa deve criar suas páginas e seus hiperlinks. Apesar deste processo ser totalmente distribuído, sem nenhum tipo de coordenação, chegamos a uma rede enorme que tem distâncias muito curtas. Como é que fazemos isto? Na seção 7.4 veremos um modelo matemático que sugere um processo de formação da Web, dando origem a diferentes propriedades estruturais, inclusive esta.

Vamos considerar a rede da Wikipedia para ilustrar algumas propriedades estruturais importantes. A rede da Wikipedia é a rede formada pelas páginas da Wikipedia (documentos) e pelos hiperlinks entre as páginas. Repare que esta rede é um subconjunto pequeno e específico da rede da Web. Entretanto, a rede da Wikipedia é bastante interessante, pois é uma rede de informação com objetivos específicos (i.e., enciclopédia) e construída com mais controle.

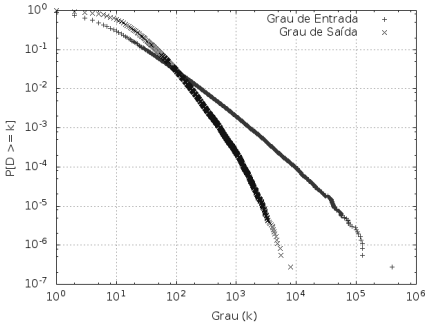


Figura 7.5. Distribuição (CCDF) do grau de entrada e saída da Wikipedia ($\overline{d} = 22.1$, desvio padrão 386.1 e 44.9 para o grau de entrada e saída, respectivamente).

Iremos analisar a rede da Wikipedia[4] em inglês obtida em janeiro de 2011. Esta rede possui 3.651.512 documentos (vértices) e 80.737.121 hiperlinks entre eles (arestas direcionadas). A figura 7.5 ilustra a distribuição complementar cumulativa (CCDF) do grau de entrada e de saída dos vértices da Wikipedia. Repare que os eixos do gráfico estão em escala logarítmica. Podemos observar diversos aspectos interessantes nesta distribuição. Um dos mais importantes é a diferença entre os graus dos vértices da rede, tanto de entrada quanto de saída. Podemos observar que existem páginas com desde poucos hiperlinks de saída e de entrada (1 ou 2) até páginas com dezenas de milhares, passando por diversas ordens de magnitude! Além disso, observamos que o grau médio é relativamente baixo (22.1) quando comparado com esta extensa faixa de valores. Logo, podemos concluir e observar que a grande maioria das páginas possui um grau baixo, enquanto uma pequena minoria possui um grau muito alto.

Podemos observamos ainda que a distribuição do grau de entrada e de saída são bem distintas, sendo que a primeira atinge valores bem maiores do que a segunda, apesar das duas terem a mesma média[5]. Dizemos que a cauda

[4] Todo o conteúdo da Wikipedia está disponível publicamente para download em http://en.wikipedia.org/wiki/Wikipedia:Database_download .

da distribuição de entrada é mais pesada do que a de saída, pois seus valores possuem maior massa de probabilidade. Por exemplo, enquanto a fração de vértices com grau de saída maior do que 5000 é 10^{-6}, esta mesma fração para o grau de entrada é 5×10^{-3}, ordens de grandeza maior. Esta característica é comum a muitas outras redes que possuem relacionamentos assimétricos, e não apenas a redes de informação, mas também redes sociais e redes tecnológicas. Na seção 7.4 iremos apresentar o modelo matemático proposto por Barabási e Albert que oferece uma explicação para o surgimento desta característstica. Outra observação interessante é que a distribuição de grau é bem comportada, podendo ser bem representada por um modelo matemático. De fato, esta distribuição empírica pode ser modelada por uma *lei de potência*, que é uma distribuição com forma simples mas com propriedades peculiares. Iremos apresentar e formalizar distribuições de lei de potência na seção 7.4.1.

A estrutura da Web não é estudada apenas por curiosidade, pois alguns bons usos desta estrutura podem valer muito dinheiro! A estrutura da Web pode ser utilizada para realizar diversas funções importantes sobre os documentos da Web. Talvez a função mais importante seja ordenar os documentos de acordo com sua relevância ou importância (*ranking*). No final da década de 90, Larry Page e Sergey Brin, dois alunos de pós-graduação em Stanford, observaram que a estrutura da Web poderia ser explorada para identificar documentos relevantes. Intuitivamente, páginas que recebem muito hiperlinks tendem a ser mais relevantes que páginas que recebem poucos hiperlinks. Além disso, a influência que um hiperlink possui é inversamente proporcional ao número de hiperlinks saem da mesma página. Estas ideias deram origem ao algoritmo PageRank para ordenar páginas Web [Brin e Page 1998]. Logo em seguida os dois fundaram a empresa Google, que viria revolucionar o mercado de máquinas de busca na Web (*search engines*). Além da ordenação de documentos a estrutura da Web pode ser utilizada na clusterização, categorização, tradução, entre outros. Você ainda pode pensar em como explorar a estrutura da Web para melhorar ou criar uma nova funcionalidade.

Uma outra rede de informação interessante é a rede de citação. Considere um conjunto de documentos que fazem referência a outros documentos. Por exemplo, artigos científicos em geral contém referências (citações) a outros artigos científicos; patentes em geral contém referências a outas patentes; decisões judiciais em geral contém referências a outras decisões judiciais. Existem muitas razões para um documento citar um outro, tais como informar ao leitor material relacionado, indicar algum documento influente, discordar de algum documento, etc. Em uma rede de citação, os vértices representam documentos e as arestas citações entre os documentos. Ou seja, se o documento i cita o documento j, então teremos uma aresta do vértice i para o vértice j. Repare que a relação de citação é assimétrica, pois um documento pode citar o outro

[5] O número total de arestas que saem dos vértices é igual ao número total de arestas que entram nos vértices, pois cada aresta possui duas pontas. Logo o grau médio de entrada é igual ao grau médio de saída.

e não vice-versa. Aliás, em uma rede de citação, é muito raro que dois documentos se citem mutuamente, pois em geral ao ser publicado, um documento não pode ser alterado. Exceções podem existir quando dois documentos relacionados são publicados no mesmo fórum (i.e., congresso ou revista) e então se citam mutuamente.

Redes de citação possuem propriedades estruturais bem interessantes e vem sendo estudadas por bastante tempo. Um dos estudos mais antigos foi realizado por Price na década de 60, inclusive abordando questões relacionadas a distribuição de grau [de Solla Price 1965]. Muitos outros estudos vem sendo realizados mais recentemente, com o aumento de informações sobre muitas redes de citações [Redner 1998, Börner et al. 2004]. Em todo o caso, uma propriedade estrutural muito frequentemente encontrada é uma distribuição empírica de grau que possui cauda pesada, se aproximando de uma lei de potência em alguns casos. Isto significa que a grande maioria dos documentos recebe poucas citações enquanto uma pequena minoria recebe um grande número de citações. De fato, no Science Citation Index (SCI), que é parte do repositório de artigos e citações Web of Science, em torno de 47% dos artigos cadastrados nunca receberam sequer uma citação e somente 1% dos artigos possuem mais do que 100 citações [Newman 2010]. Estas e outras observações sobre a estrutura das redes de citação são usadas para definir critérios de relevância dos documentos, em uma área conhecida por bibliometria (*bibliometrics*).

Por fim, é importante ressaltar que redes de informação e redes sociais muitas vezes se confundem. Por exemplo, considere alguns sistemas online como o Twitter, MySpace e o Blogger. Nestes sistemas temos uma mistura de redes sociais, induzida pelos relacionamento de amizade entre as pessoas, e redes de informação, induzida pela informação publicada e seus relacionamentos. No Twitter, uma pessoa segue a outra por diversos motivos, que incluem simplesmente amizade ou apenas interesse na informação. Recentes estudos inclusive vem tentando caracterizar o Twitter como sendo uma mistura de rede social com rede de informação [Kwak et al. 2010]. De fato, não há uma distinção muito clara dessas redes e a nomenclatura apresentada aqui serve mais para referência.

7.3.3. Redes Tecnológicas

Redes tecnológicas são redes construídas pelo homem em geral para transportar algo. Desta forma, redes tecnológicas são geralmente redes físicas, onde seus objetos e relacionamentos são concretos. Uma importante rede tecnológica é a Internet, formada por roteadores e canais de comunicação, conhecidos como enlaces. Um roteador é um computador especial dedicado a receber e enviar informação pelos seus enlaces. Um roteador está ligado a diversos enlaces, que podem ser fios de cobre, fio coaxial, fibra ótica, ou até mesmo canais de rádio e satélite. Na outra ponta de cada enlace há um outro roteador. Um roteador pode ter desde apenas dois canais de comunicação

até milhares. De forma simplificada, a Internet pode ser vista como uma infra-estrutura global para transportar bits.

Na abstração da Internet cada vértice corresponde a um roteador e uma aresta indica que existe um enlace entre os dois roteadores. Para facilitar a leitura, iremos chamar esta abstração também de Internet ao invés de rede da Internet. É importante não confundir a Internet com a Web. A rede da Web é uma rede virtual, pois representa documentos (URLs) e hiperlinks sem ligação com a Internet física. A Internet representa a infra-estrutura por onde a informação da Web irá trafegar. Entretanto, igualmente a rede da Web, a Internet é um sistema distribuído construído sem recomendações ou qualquer coordenação global. Qualquer um pode criar uma rede com roteadores e se ligar a outros roteadores da Internet (em geral, pagando por isto). Desta forma, a Internet precisa ser medida e inferida para que possamos conhecer sua estrutura. A Internet também é uma rede enorme com mais de um milhão de roteadores e obter uma representação fiel e sem tendência (*unbiased*) de sua estrutura ainda é um desafio. Em todo o caso, uma das maneiras para inferir a estrutura da Internet é através do programa *traceroute*, que informa o caminho (rota) tomada entre o computador onde o programa executa e qualquer outro roteador (ou computador) da Internet. Podemos então executar o *traceroute* de um alguns pontos da Internet para vários outros e fazer a união de todos os caminhos obtidos, conseguindo assim parte da estrutura da Internet. Esta técnica vem sendo usada para inferir a estrutura da Internet [CAIDA 2011], apesar dos seus potenciais problemas [Achlioptas et al. 2005].

Uma propriedade interessante na Internet é sua distribuição de grau. Novamente, observamos uma distribuição empírica de grau muito desigual, com a grande maioria dos vértices tendo grau baixo e uma pequena minoria com grau alto. Em muitos casos, esta distribuição empírica de grau se aproxima de uma distribuição de lei de potência. Além desta propriedade, observamos na Internet distâncias muito curtas entre os vértices. Em geral, a distância média entre os vértices é uma função logarítmica do tamanho da rede, ou seja, exponencialmente menores do que o número de vértices [Chen et al. 2002]. Esta propriedade está diretamente relacionada ao desempenho da rede e poderíamos dizer que seria esperada. Pois uma rede para transportar bits deve ter caminhos curtos entre seus vértices para oferecer um melhor desempenho (i.e., menor retardo na comunicação). Entretanto, a Internet não é construída de forma centralizada e é ao menos curioso que esta propriedade surja a partir da interação de milhares de entidades que aumentam e mantém a Internet, diariamente. Entender as razões que levam a observações empíricas como esta é um dos pontos mais fascinantes no estudo em Redes Complexas.

Existem muitas redes tecnológicas construídas pelo homem para transportar o próprio homem. Estamos falando da rede rodoviária de uma cidade ou um país, da rede ferroviária, e da malha aérea de um país ou do mundo. Estas redes são chamadas de redes de transporte e algumas delas possui características distintas da Internet, criadas para transportar bits. Por exemplo, considere

a rede rodoviária de um país, onde os vértices são cruzamentos (interconexão) de rodovias estaduais ou federais e as arestas representam a ligação física destes cruzamentos por estradas. O grau nesta rede representa o número de estradas que chegam ou saem do cruzamento. Podemos observar que a distribuição empírica de grau desta rede não possui uma cauda pesada, pois não há cruzamentos com um grande número de estradas (e.g., 10) e a grande maioria deles envolve um pequeno número de estradas (e.g., 2).

7.3.4. Redes Biológicas

Redes biológicas são geralmente formadas pela natureza e aparecem em muitos contextos. Algumas redes biológicas são construções físicas e outras representações mais abstratas que dependem do relacionamento. Uma das mais intrigantes redes biológicas é a rede neuronal humana, ou seja, a rede formada pelos neurônios de nosso cérebro. Nesta rede, vértices são neurônios e arestas representam ligações físicas entre os neurônios, conhecidas como sinapses. As ligações entre neurônios são direcionadas, com os dendritos recebendo sinais e axônios enviando sinais. Um neurônio humano geralmente possui poucos axônios e muitos dendritos, mas este número pode variar bastante entre os neurônios. Como não poderia deixar de ser, a rede neuronal humana é gigantesca e estima-se que tenha na ordem de 100 bilhões de neurônios e 100 trilhões de sinapses. Em comparação a rede neuronal do nematódeo *Caenorhabditis elegans*, que possui 1mm de comprimento, possui apenas 302 neurônios [Wikipedia 2011a].

A rede de neurônios humana e de outras espécies vem sendo estudada há décadas tanto por biólogos quanto por neuro-cientistas, e mais recentemente por físicos, matemáticos, e cientistas da computação. A razão para todos estes estudos está no fascinante papel que a rede neuronal desempenha. Sua estrutura é fundamental para que o homem desempenhe diversas funções cognitivas, tais como memória, consciência, linguagem e raciocínio. Seriam estas funções consequência direta da estrutura neuronal? Poderia outro mamífero exercer alguma destas funcionalidades caso sua rede neuronal tivesse a mesma estrutura que a do homem? Perguntas como estas ainda continuam a desafiar a ciência.

Por fim, existem ainda muitas outras redes biológicas, tais como a rede metabólica, a rede da cadeia alimentar e a rede de interação entre proteínas [Newman 2010].

7.3.5. Características em comum

O leitor mais atento deve ter reparado que apesar de termos discutido diversas redes de diferente domínios, algumas características estruturais são comum a muitas delas. Esta observação é surpreendente, pois a princípio não há razão alguma para a rede da Web compartilhar características estruturais com a rede de neurônios humana! De fato, algumas características são recorrentes em muitas redes reais. Vejamos algumas destas características:

- **Distribuição de grau com cauda pesada.** Muitas redes reais de diferentes domínios possuem uma distribuição de grau com cauda pesada, exibindo um espectro de grau que são ordens de grandeza maiores do que a média. Em alguns casos a distribuição de grau é bem representada por uma lei de potência, modelo matemático que será apresentado na seção 7.4.1. Redes como a Web, o Twitter, a rede de atores, a Internet, a rede metabólica, entre outras, exibem uma distribuição de grau bem representada por uma lei de potência.
- **Distância baixa.** Muitas redes reais possuem distância relativamente pequena entre seus vértices, inclusive entre os vértices mais distantes da rede, sendo ordens de grandeza menor do que o número de vértices. Em alguns casos a distância média da rede é bem representada por $\log n$, sendo exponencialmente menor do que seu tamanho. Redes como a Web, a rede de colaboração científica, a Internet, uma rede de neurônios, entre outras, exibem distâncias curtas.
- **Esparsas e conectadas.** Muitas redes reais são extremamente esparsas, exibindo uma densidade muito baixa, ordens de grandeza menor do que 1. Apesar disto, estas mesmas redes estão quase completamente conectadas, com quase todos os vértices pertencendo a mesma componente conexa. Este é o caso de redes como a Web, rede de atores, a Internet, uma rede de neurônios, a rede metabólica, entre muitas outras, são esparsas e conectadas.

Além destas, existem outras características estruturais que frequentemente aparecem em diversas redes reais. Na verdade, considerando as três características acima, é mais fácil encontrar uma rede real que exiba estas características do que uma rede que não exiba. Apesar disto, ainda não compreendemos muito bem se existe alguma razão fundamental que leva as redes reais a sempre (ou quase sempre) serem assim, independente do seu propósito ou domínio. Por fim, estas três características tem importantes implicações em muitas funcionalidades e processos que operam na rede e veremos algumas destas no decorrer do livro.

7.4. Modelos aleatórios de redes

Vimos na seção anterior que existem muitos tipos de redes reais em muitos domínios diferentes e algumas com propriedades muito particulares. Além disso, redes reais de um mesmo tipo também são diferentes. Considere a rede de colaboração científica entre matemáticos, apresentada na seção 7.3.1. Agora considere a rede de colaboração científica entre os cientistas da computação. Certamente estas redes não são iguais e possuem muitas propriedades diferentes. Por exemplo, seus tamanhos certamente serão diferentes, assim como o grau médio dos vértices e a densidade da rede. De fato sabemos que na Ciência da Computação, pesquisadores tendem a colaborar na produção de artigos científicos mais que os matemáticos, consequentemente apresentando

um grau médio maior na rede de colaboração. Mas apesar de diferenças, estas duas rede também possuem muitas similaridades estruturais. Talvez isto não seja surpreendente, uma vez que as duas são decorrências do processo de colaboração científica e interação humana, que visto de longe, não difere muito de uma área de ciência para outra. Mas como capturar este aspectos fundamentais destas duas redes? Para que possamos falar sobre redes de colaboração científica de forma mais abrangente será necessário generalizá-las. Para isto iremos precisar da ajuda da matemática, em particular, de modelos matemáticos aleatórios para as estruturas destas redes.

Um modelo aleatório de rede é um modelo matemático que determina como formar a estrutura de uma rede. Fazendo uma analogia com a Computação, ao executarmos o modelo ele irá gerar uma rede. Desta forma, as redes geradas pelos modelos irão ter determinada propriedades estruturais. Por ser um modelo matemático, podemos estabelecer rigorosamente algumas de suas características, por exemplo, que o modelo sempre irá gerar redes que tenham determinada propriedade estrutural. Isto é fundamental pois então podemos aplicar o modelo para representar redes que tenham estas mesmas propriedades estruturais.

É importante ressaltar que um modelo matemático não é bom ou nem ruim, pois trata-se simplesmente de um modelo. Entretanto, podemos dizer que um determinado modelo captura ou representa melhor uma determinada característica do que um outro qualquer, concluindo que o primeiro é melhor do que o segundo. Mas esta afirmação em geral depende das características que estamos querendo capturar, pois um modelo pode representar melhor determinada característica e pior outra.

No que segue apresentaremos três modelos muito utilizados para representar redes. Iremos discutir ainda algumas das propriedades estruturais destes modelos. Desta forma, podemos então determinar quais modelos podem ser empregados quando queremos capturar determinadas propriedades estruturais que aparecem em redes reais. Entretanto, é importante ressaltar que existem na literatura muitos diferentes modelos de redes, alguns bem mais avançados dos que discutiremos em breve, além de novos modelos estarem constantemente sendo propostos na literatura. Você não deve se ater aos modelos aqui apresentados e nem aos que existem na literatura! Se você achar adequado, proponha o seu modelo para capturar os aspectos que você considera importante na rede real que você está estudando.

7.4.1. Redes Livre de Escala

Antes de introduzirmos propriamente os modelos iremos apresentar um fenômeno muito interessante e igualmente importante, conhecido por de *lei de potência*. Apesar de se aplicar de forma mais geral, estamos interessados em distribuições de probabilidade que possuem este fenômeno, ou seja, em distribuições de lei de potência.

Uma distribuição é dita lei de potência quando sua função de probabilidade

possui as seguinte forma:

$$p_X(x) \sim cx^{-\alpha} \text{ para } x \text{ grande suficiente.} \tag{12}$$

onde $p_X(x)$ é a probabilidade da variável aleatória discreta X assumir o valor x, c é a constante de normalização e α é uma constante e parâmetro da distribuição, com $\alpha > 1$. O nome lei de potência vem do fato da distribuição ser caracterizada pela potência α, sendo este seu único parâmetro.

É importante reparar que a distribuição de lei de potência decresce muito mais devagar do que qualquer outra distribuição. Repare que $x^{-\alpha}$ é $o(\beta^{-x})$ para qualquer $\beta > 0$, ou seja, $\beta^{-x}/x^{-\alpha} \to 0$ quando $x \to \infty$. Consequentemente, dizemos que uma distribuição de lei de potência possui "cauda pesada", pois valores arbitrariamente grandes de x possuem muito mais massa de probabilidade não-desprezíveis. Ou seja, existe uma probabilidade relativamente alta da distribuição gerar números que são significativamente maiores do que sua média. A cauda da distribuição se recusa a ir para zero rapidamente!

Uma outra característica importante da distribuição de lei de potência é que seu gráfico traçado em escala $\log - \log$ é uma reta. Podemos observar isto diretamente da sua definição, na equação (12). Podemos aplicar o $\log$ aos dois lados da equação, obtendo assim a seguinte:

$$\log p_X(x) \sim -\alpha \log x + \log c \text{ para } x \text{ grande suficiente.} \tag{13}$$

Repare que esta é a equação da reta com inclinação $-\alpha$, quando traçado em escala $\log - \log$, para valores suficientemente grandes de x. Ou seja, relação linear deve surgir para valores grandes o suficiente de x, na cauda da distribuição.

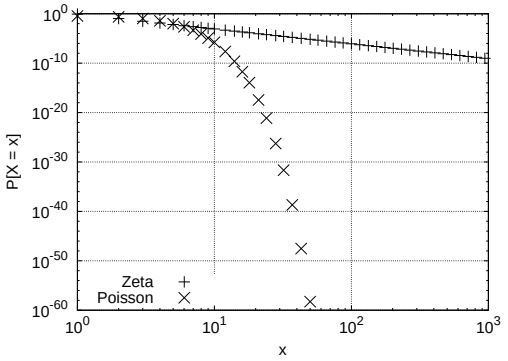


Figura 7.6. Comparação entre a distribuição zeta ($\alpha = 3$) e a distribuição de Poisson ($\lambda = \zeta(2)/\zeta(3)$) ambas com o mesmo valor esperado.

Vejamos o exemplo na figura 7.6, que mostra a distribuição de Poisson e zeta, ambas com o mesmo valor esperado, em um gráfico com escala log-log. Repare que a distribuição de zeta segue uma lei de potência o que não é o caso da distribuição de Poisson. Podemos observar que a distribuição zeta possui uma cauda muito mais pesada a Poisson: a probabilidade de termos um valor igual a 50 é 10^{-59} para Poisson e 10^{-6} para zeta, 53 ordens de grandeza maior! Note ainda que valores muito maiores do que a média podem ocorrer com probabilidade não desprezível na distribuição zeta, enquanto isto não ocorre com a Poisson.

A distribuição zeta possui exatamente a forma da equação (12) e no domínio $x > 0$. Desta forma, podemos calcular a constante de normalização somando todos os valores das probabilidades. Ou seja,

$$1 = \sum_{x=1}^{\infty} p_X(x) = c \sum_{x=1}^{\infty} x^{-\alpha} = c\zeta(\alpha) \tag{14}$$

O somatório ao final da equação é a famosa função zeta de Reimann. Logo, a constante de normalização c, é dada por:

$$c = \frac{1}{\zeta(\alpha)} \tag{15}$$

Pode-se demonstrar que a função zeta de Reimann converge apenas para valores maiores que 1 e diverge caso contrário. Desta forma, a constante de normalização só está definida para $\alpha > 1$.

Vamos obter os momentos da distribuição zeta de Reimann. Lembrando que o i-ésimo momento de uma distribuição é dado pelo valor esperado da variável aleatória elevada a i-ésima potência, ou seja, $E[X^i]$. Repare que o primeiro momento é, por definição, o valor esperado da distribuição, enquanto o segundo momento pode ser usado para obter sua variância[6]. Aplicando a definição de valor esperado, podemos obter o primeiro momento:

$$E[X] = \sum_{x=1}^{\infty} x p_X(x) = \sum_{x=1}^{\infty} x \frac{1}{\zeta(\alpha)} x^{-\alpha} = \frac{1}{\zeta(\alpha)} \sum_{x=1}^{\infty} x^{-(\alpha-1)} = \frac{\zeta(\alpha-1)}{\zeta(\alpha)} \tag{16}$$

Repare que na última passagem aplicamos novamente a definição da função zeta de Reimann, com argumento $\alpha - 1$. Repare que o valor esperado nem sempre estará definido, podendo divergir para infinito. Em particular, se $\alpha \leq 2$ temos que $\alpha - 1 \leq 1$ e consequentemente o valor esperado irá divergir, pois a função zeta no numerador da equação (16) irá divergir. Logo podemos concluir que o valor esperado da distribuição só está definido quando $\alpha > 2$.

É importante ressaltar este surpreendente aspecto da distribuição zeta, que na verdade é comum a outras distribuições de lei de potência. Imagine que temos uma distribuição zeta com $\alpha = 1.5$. Apesar da distribuição estar bem

[6] A variância de uma distribuição pode ser obtida como: $Var[X] = E[X^2] - E[X]^2$.

definida, seu valor esperado não existe, não está definido! Ele simplesmente diverge se tentarmos estimá-lo através da amostragem da distribuição. Considere a média amostral $\overline{X_n} = 1/n \sum_{i=1}^{n} X_i$, onde X_i são amostras iid (independentes e identicamente distribuídas) da distribuição zeta e n é o número de amostras. Ao estimar o valor esperado com a média amostral, veremos que neste caso $\overline{X_n} \to \infty$ quando $n \to \infty$. Pois mesmo depois de fazermos uma média com 1 milhão de amostras, ainda sim existe uma probabilidade não desprezível de obtermos amostras grandes o suficiente para influenciarem diretamente a média amostral. Leis de potência são objetos matemáticos fantásticos!

Podemos generalizar o desenvolvimento acima para calcular qualquer momento da distribuição zeta. Em particular, temos que $E[X^n] = \zeta(\alpha - n)/\zeta(\alpha)$ (mostre este resultado). Podemos concluir então que o n-ésimo momento da distribuição zeta só está definido quando $\alpha > n + 1$. Por exemplo, se $\alpha = 2.7$ então temos o primeiro momento (valor esperado) definido e o segundo momento (variância) não definido. A variância só está definida quando $\alpha > 3$.

Uma função distribuição de lei de potência também recebe o nome de "livre de escala". Intuitivamente, uma distribuição é dita livre de escala se a razão entre dois pontos de probabilidade depende apenas da razão entre os valores dos dois pontos, mas não dos valores em si. Ou seja, a proporcionalidade da distribuição se mantém em qualquer escala para os valores que ela pode assumir, sendo assim, livre de escala. Mais formalmente, seja b um valor inteiro positivo qualquer. Considere a razão entre um valor x qualquer para a distribuição e um valor b vezes maior, ou seja, bx. Temos então:

$$\frac{p_X(x)}{p_X(bx)} \sim \frac{cx^{-\alpha}}{c(bx)^{-\alpha}} = b^{\alpha} \quad \text{para } x \text{ grande suficiente.} \tag{17}$$

Repare que a razão entre as probabilidades é b^{α} independente do valor de x. Por exemplo, considere uma lei de potência com $\alpha = 2$. Então temos que um evento $b = 3$ vezes maior é 9 vezes menos provável, independente do tamanho do evento. Por fim, podemos mostrar que apenas distribuições que seguem lei de potência podem ser livres de escala.

Agora estamos prontos para definir redes livre de escala (*scale-free networks*). Uma rede é dita livre de escala se sua distribuição de grau é livre de escala. Ou seja, se sua distribuição de grau segue uma lei de potência. Consequentemente, em uma rede livre de escala, os graus dos vértices não são nada parecidos uns com os outros, pois podemos ter vértices com graus muito maiores do que média com probabilidade não desprezível. De fato, muitas redes reais possuem esta propriedade, como a distribuição de grau ilustrada na figura 7.5 (rede da Wikipedia), e por isto são chamadas de redes livre de escala.

Por fim, é importante ressaltar que existem outras distribuições que também seguem uma lei de potência. A conhecida distribuição de Zipf é uma variação da distribuição zeta onde os valores que a variável aleatória pode assumir são limitados superiormente. A distribuição de Pareto também segue uma lei

de potência para representar variáveis aleatórias contínuas. Esta distribuição também é bastante usada, inclusive por ser mais fácil de manipular algebricamente do que a distribuição zeta, que é discreta. O leitor interessado neste aspecto deve buscar mais informações sobre estas distribuições na literatura [Mitzenmacher 2004, Newman 2005].

7.4.2. Modelo de Erdős-Rényi, $G(n, p)$

Um dos mais antigos e mais estudados modelo aleatório para redes é conhecido como modelo de Erdős e Rényi, pois foi muito explorado no final da década 50 e início de 60, pelos influentes matemáticos Paul Erdős e Alfréd Rényi [Erdős e Rényi 1959, Erdős e Rényi 1960]. Em particular, Erdős e Rényi provaram diversas características fundamentais da estrutura da rede gerada por estes modelos e descobriram diversas propriedades muito interessantes, tudo sem a ajuda de um computador! Ao longo dos anos, muitas outras características foram estabelecidas por muitos outros pesquisadores, de forma que conhecemos muitos aspectos das redes geradas por este modelo. O modelo possui vários nomes e também é conhecido por $G(n, p)$, modelo binomial e às vezes até mesmo por "modelo de rede aleatória", dada sua grande influência.

O modelo $G(n, p)$ é bastante simples e possui apenas dois parâmetros, n e p. O parâmetro n determina o número de vértices da rede e estes são rotulados de 1 a n. O parâmetro p determina a probabilidade de uma determinada aresta ser incluída na rede. Além disso, cada possível aresta será incluída de forma independente das outras. Repare que o modelo não determina a estrutura da rede, e sim o processo aleatório que irá gerar a estrutura da rede. Desta forma, a rede gerada pelo modelo é uma variável aleatória.

Vamos exemplificar com o modelo $G(5, 0.25)$. Este modelo irá dar origem a redes com 5 vértices e cada uma das possíveis arestas desta rede irá existir com probabilidade 0.25. Ou seja, para cada aresta temos que decidir se a será incluída na rede. A figura 7.7 ilustra possíveis redes geradas pelo modelo $G(5, 0.25)$. Repare que na verdade o modelo pode gerar qualquer estrutura de rede com cinco vértices.

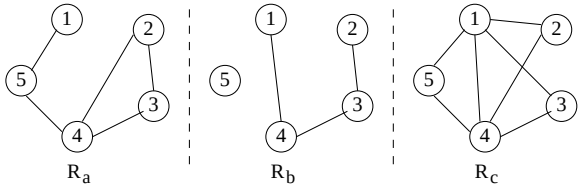


Figura 7.7. Exemplos de possíveis redes geradas pelo modelo $G(5, 0.25)$.

O espaço amostral S do $G(n, p)$ é dado pelo conjunto de redes que podem ser geradas pelo modelo. Podemos observar que qualquer rede com n vérti-

ces pode ser gerada, pois qualquer aresta pode ou não ser incluída na rede. Como cada aresta pode ou não estar na rede e temos um total de $\binom{n}{2}$ arestas possíveis em uma rede com n vértices, podemos obter o tamanho do espaço amostral:

$$|S| = 2^{\binom{n}{2}} = 2^{n(n-1)/2} \tag{18}$$

Repare que o espaço amostral cresce muito rapidamente. Para termos uma ideia, o número redes diferentes com 25 vértices é maior do que o número de átomos do universo[7]! Entretanto, é importante notar que a probabilidade do modelo gerar cada uma destas possíveis redes é muito diferente. Veremos que algumas redes são bem mais prováveis que outras e que algumas são simplesmente muito improváveis, com chances menores do que um em o número de átomos do universo!

Considere um conjunto de arestas $E = \{e_1, \ldots, e_m\}$, onde cada aresta é um par não ordenado de vértices, ou seja, $e_i = (u_i, v_i)$ com $u_i, v_i \in [1, \ldots n]$. Estamos interessados em determinar a probabilidade do modelo $G(n, p)$ gerar exatamente a rede representada pelo conjunto de arestas E. Esta probabilidade pode ser obtida se considerarmos que toda aresta do conjunto E deve estar na rede gerada pelo modelo e nenhuma outra aresta fora do conjunto E deve estar na rede. A probabilidade de uma aresta em E aparecer na rede gerada pelo modelo é p, independente da aresta. Além disso, a probabilidade de uma outra aresta não aparecer na rede é $1 - p$. Assim temos:

$$P[G(n, p) \rightarrow E] = p^m (1 - p)^{m - \binom{n}{2}} \tag{19}$$

onde m é o número de arestas em E. Repare que a probabilidade de gerar E depende apenas do número de arestas em E e não de suas arestas específicas.

Vamos voltar às redes ilustradas na figura 7.7 e calcular a probabilidade do $G(5, 0.25)$ gerar cada uma delas, pois cada uma delas é definida por um conjunto de arestas. Fazendo as contas, veremos que a probabilidade de gerar as redes R_a, R_b e R_c é aproximadamente 2×10^{-4}, 2×10^{-3}, e 3×10^{-5}, respectivamente. Podemos concluir que a rede R_b é bem mais provável de ser gerada pelo modelo $G(5, 0.25)$ do que as outras duas

Muitas vezes estamos interessados não em uma rede específica, mas em redes com características específicas. Por exemplo, podemos ter interesse em redes que tenham um certo número m de arestas, sem nos importarmos muito que arestas são estas. Podemos calcular a probabilidade do modelo $G(n, p)$ gerar uma rede com m arestas. Repare que estamos interessados em qualquer rede com m arestas. O número de redes com exatamente m arestas é dado pelo número de combinações diferentes de m arestas das $\binom{n}{2}$ arestas

[7] O número de átomos do universo observável está estimado em 10^{80} que é aproximadamente 2^{267} [Wikipedia 2011b]. O espaço amostral do $G(n, p)$ com $n = 25$ vértices tem tamanho 2^{300}.

possíveis da rede. Desta forma, obtemos:

$$P[G(n,p) \rightarrow |E| = m] = \binom{\binom{n}{2}}{m} p^m (1-p)^{m-\binom{n}{2}} \quad (20)$$

Repare que temos a distribuição binomial com parâmetros $\binom{n}{2}$ e p. Consequentemente, sabemos que seu valor esperado é dado por $\binom{n}{2}p$, que é o valor esperado do número de arestas nas redes geradas pelo modelo $G(n,p)$.

Vamos estudar agora o grau de um vértice no modelo $G(n,p)$. Repare que ao escolhermos um vértice ao acaso, seu grau será determinado pelo número de arestas incidentes ao mesmo. Logo, o grau é uma variável aleatória, denotada por D, que assume valores entre 0 (nenhuma aresta incidente) e $n-1$ (todas possíveis arestas incidentes). Mas qual sua distribuição? Cada uma destas arestas irá existir com probabilidade p independente das outras. Para que o grau seja igual a k, exatamente k arestas devem ser incidentes e todas as outras não podem ser incidente. Logo, a probabilidade do vértice ter grau exatamente k é dado por:

$$P[D = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k} \quad (21)$$

Temos então que o grau segue uma distribuição binomial com parâmetros $n-1$ e p. O valor esperado do grau neste caso é $E[D] = (n-1)p$.

Voltamos ao nosso pequeno exemplo, o modelo $G(5, 0.25)$. A probabilidade deste modelo gerar redes com exatamente 3 arestas é igual a 0.25. Repare que esta probabilidade é bem maior do que probabilidade de gerar a rede R_b, ilustrada na figura 7.7, que também possui três arestas. A diferença está no número diferente de redes com três arestas, que é relativamente grande (exatamente 120). Por fim, o grau médio das redes geradas por este modelo é $(5-1) \times 0.25 = 1$.

7.4.2.1. Propriedades estruturais em modelos de redes

Muitas vezes estamos interessados na probabilidade de um determinado modelo gerar redes que tenham uma determinada propriedade estrutural. Por exemplo, considere a propriedade "conexa". Uma rede é conexa quando existe ao menos um caminho entre qualquer dois pares de vértices. Qual é a probabilidade do modelo $G(n,p)$ gerar um grafo conexo? Teoricamente, podemos obter esta probabilidade considerando todas as redes do espaço amostral que são conexas e somando a probabilidade de cada uma delas ser gerada pelo modelo. Como estas são mutuamente exclusivas, esta soma nos dá exatamente a probabilidade do modelo gerar uma rede conexa.

Podemos generalizar esta ideia para qualquer propriedade. Seja X uma propriedade estrutural qualquer de forma que cada possível rede tenha ou não

esta propriedade. Seja $C_X \subset S$ o subconjunto de redes que possuem a propriedade X, ou seja, $C_X = \{R | R \in S,\ R \text{ possui propriedade } X\}$. A probabilidade do modelo gerar uma rede com propriedade X será dada por:

$$P[G(n,p) \rightarrow R \in C_X] = \sum_{E \in C_X} P[G(n,p) \rightarrow E] \tag{22}$$

Repare que esta probabilidade irá depender da propriedade X, assim como dos parâmetros n e p do modelo. Infelizmente, esta abordagem para calcular a probabilidade de termos uma rede com propriedade X em geral não é viável, pois sabemos que o tamanho do espaço amostral S é muito grande, assim como o subconjunto C_X de redes de interesse.

De forma a facilitar a análise, iremos considerar redes muito grandes, tão grandes que a influência de n na probabilidade de termos uma rede com uma propriedade X será desprezível. De forma ainda mais simplificada, estaremos interessados em estudar os casos onde esta probabilidade converge para um ou para zero, sem que seja necessário determinar de fato seu valor.

Iremos dizer que uma propriedade X é *asymptotically almost surely* (a.a.s.) se quase todas as redes geradas pelo modelo $G(n,p)$ possuem a propriedade X quando n tende ao infinito. Matematicamente, temos:

$$P[G(n,p) \rightarrow R \in C_X] \rightarrow 1 \text{ quando } n \rightarrow \infty. \tag{23}$$

Estamos interessados em caracterizar quais propriedades estruturais são aas no $G(n,p)$ e sobre quais condições de seus parâmetros. No que segue iremos apresentar algumas destas propriedades.

7.4.2.2. Propriedades estruturais do $G(n,p)$

Para estudarmos o comportamento assintótico do modelo $G(n,p)$ quando n tende ao infinito, precisamos caracterizar como p irá evoluir quando n cresce. Existem basicamente três opções: p pode crescer, se manter constante, ou diminuir. A primeira delas não é muito interessante, pois se p cresce com n, irá tender a um (seu limite superior) e neste caso teremos uma rede completa, com todas suas arestas. E se p cresce assintoticamente para um valor intermediário, então estamos praticamente no caso de p constante com o valor assintótico. O segundo caso também não será muito interessante, pois como veremos em breve, dá origem a grafos muito densos, quase completos. O caso de grande interesse é o terceiro, quando p diminui com n. Veremos que dependendo de como este decrescimento ocorre, teremos redes com propriedades muito diferentes.

Vamos começar com p constante, ou seja, $p > 0$ não varia enquanto $n \rightarrow \infty$. Neste caso, o grau médio de um vértice, dado por $(n-1)p$ diverge. Quanto maior é a rede maior é o grau médio dos vértices, e mesmo que p seja arbitrariamente pequeno, eventualmente teremos vértices com graus enormes.

Repare que a densidade média da rede (equação (5)) será dada por p, que é constante. Intuitivamente, estas arestas irão gerar uma rede muito conectada, conforme mostrado pelos seguintes teoremas:

Teorema 7.4.1 *Para $p > 0$ constante, o modelo $G(n, p)$ é conexo, a.a.s.*

Teorema 7.4.2 *Para $p > 0$ constante, o modelo $G(n, p)$ possui diâmetro 2, a.a.s.*

O teorema 7.4.1 nos diz que quase todas as redes geradas pelo $G(n, p)$ são conexas quando $p > 0$ é constante e n tende ao infinito. O teorema 7.4.2 estabelece que todos os vértices destas redes estão muito próximos uns aos outros, pois a rede tem diâmetro 2 quando $n \to \infty$. A prova destes dois teoremas não é difícil e se baseia em argumentos combinatoriais e a técnica de *union bound*. O leitor mais interessado deve procurar pelos detalhes na literatura [Bollobás 2001, Durrett 2006].

Conforme mencionamos, o caso mais interessante é quando p é uma função decrescente de n, que denotaremos por $p(n)$. Existem muitas maneiras de $p(n)$ decrescer com n e diferentes maneiras vão dar origem a diferentes propriedades estruturais assintóticas.

Vamos começar considerando $p(n) = c/n$, onde $c > 1$ é uma constante. Neste caso podemos observar que o grau médio dos vértices será $E[D] = (n-1)c/n \approx c$ para n grande. Ou seja, estamos em um regime onde o grau médio dos vértices é uma constante (c) maior do que 1 e isto independe do número de vértices da rede. Neste regime, temos o seguinte teorema:

Teorema 7.4.3 *Para $p(n) = c/n$ com $c > 1$, a maior componente conexa do modelo $G(n, p)$ possui ao menos αn vértices para alguma constante $\alpha > 0$, a.a.s.*

As componentes conexas da rede são os pedaços de redes que estão "conectados". Como vimos, podemos ter redes com múltiplas componentes conexas, e com o modelo $G(n, p)$ isto não é diferente. O tamanho de uma componente conexa é o número de vértices que fazem parte da componente. Desta forma, o teorema 7.4.3 nos diz que no regime em questão, o modelo $G(n, p)$ gera redes onde a maior componente conexa possui uma fração constante dos vértices. Apesar das redes geradas não serem conexas, teremos na rede uma componente conexa gigante, muito parecido com muitas redes reais. O próximo teorema estabelece que esta componente conexa gigante é única e que todas as outras componentes tem tamanho exponencialmente menor.

Teorema 7.4.4 *Para $p(n) = c/n$ com $c > 1$, a segunda maior componente conexa do modelo $G(n, p)$ possui no máximo $\alpha \log n$ vértices para alguma constante $\alpha > 0$, a.a.s.*

Você deve estar curioso para saber o que acontece com o $G(n,p)$ quando o grau médio é constante, mas menor do que 1. Neste caso não temos arestas suficientes para formar uma componente conexa gigante ao acaso, e a rede gerada pelo modelo está fragmentada em muitas componentes conexas pequenas. De fato, a maior delas tem tamanho logarítmico no número de vértices da rede. Esta observação está formalizada no seguinte teorema.

Teorema 7.4.5 *Para $p(n) = c/n$ com $c < 1$, a maior componente conexa do modelo $G(n,p)$ possui no máximo $\alpha \log n$ vértices para alguma constante $\alpha > 0$, a.a.s.*

Será possível o $G(n,p)$ gerar redes conexas mesmo quando $p(n)$ decresce? A resposta é afirmativa se o decrescimento não for rápido demais. Em particular, $p(n)$ pode decrescer e mesmo assim o grau médio crescer. Considere $p(n) = (c \log n)/n$ com $c > 1$. O grau médio neste caso é dado por $E[D] \approx c \log n$, que cresce devagar com n. Se o decrescimento de $p(n)$ for mais lento do que isto, então temos que o $G(n,p)$ dará origem a redes conexas, conforme estabelecido no seguinte teorema.

Teorema 7.4.6 *Para qualquer $p(n) > (c \log n)/n$ com $c = 1$, o modelo $G(n,p)$ é conexo, a.a.s.*

Por fim, vamos caracterizar as distâncias entre os vértices no $G(n,p)$. Intuitivamente, se temos uma componente conexa gigante ou uma rede conexa, esperamos que a mesma esteja relativamente bem conectada, pois as arestas são todas colocadas ao acaso. De fato, as distâncias nestes casos são relativamente curtas, e crescem de forma logarítmica com o tamanho da rede, como indica o seguinte teorema:

Teorema 7.4.7 *Para qualquer $p(n) > c/n$ com $c = 1$, a distância média entre os vértices do $G(n,p)$ é no máximo $\alpha \log n$, para alguma constante $\alpha > 0$, a.a.s.*

As provas destes teoremas está além dos objetivos deste livro, mas leitores interessados devem procurar por mais detalhes na literatura. Além disso, é importante notar que estas são apenas algumas das muitas propriedades que foram estabelecidas para o modelo $G(n,p)$. Exite uma extensa literatura bastante teórica sobre este assunto [Bollobás 2001, Durrett 2006].

7.4.2.3. O problema do modelo $G(n,p)$

Podemos utilizar o modelo $G(n,p)$ para representar redes reais. Neste caso, precisamos determinar os parâmetros do modelo que devem ser estimados a partir da rede real que desejamos modelar. Considere uma rede real com n_r vértices e m_r arestas. Nada mais natural do que determinar que o modelo

deve ter n_r vértices. Agora vejamos como estimar o parâmetro p. Podemos estipular que o modelo $G(n,p)$ deve gerar em média o mesmo número de arestas que a rede real. Sabemos que o valor esperado do número de arestas do $G(n,p)$ é $\binom{n}{2}p$ (ver equação (20)). Logo temos que:

$$\binom{n_r}{2}p = m_r \;\Rightarrow\; p = \frac{2m_r}{n_r(n_r-1)} = \rho \tag{24}$$

Onde ρ é a densidade da rede real, conforme definido na equação (5). Repare que o modelo terá em média não só o mesmo número de arestas que a rede real, mas também o mesmo grau médio que a rede real. Mas será que desta forma o modelo $G(n,p)$ irá capturar aspectos relevantes de redes reais?

Infelizmente o modelo $G(n,p)$ leva a redes que possuem estruturas muito diferentes de muitas redes reais que nos cercam. Por exemplo, sabemos que muitas redes reais possuem uma distribuição de grau com cauda pesada, o que não ocorre com o modelo $G(n,p)$, cuja distribuição de grau é binomial. De fato, estes dois tipos de distribuição são muito diferentes e por isto levam a redes com propriedades estruturais bem diferentes. Além disso, a distribuição de grau possui um papel central em diversas outras propriedade da rede.

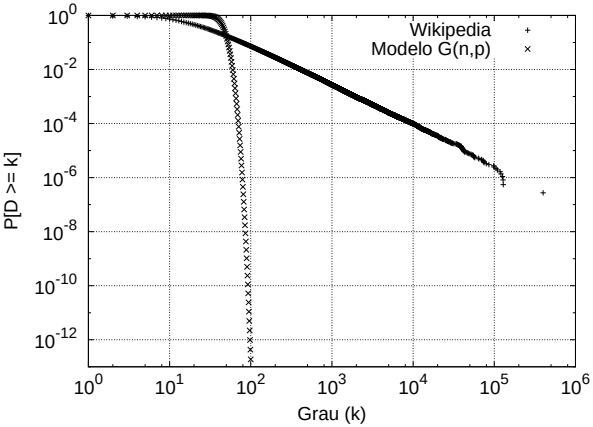


Figura 7.8. Distribuição (CCDF) de grau da rede da Wikipedia e do modelo $G(n,p)$ parametrizado (as duas distribuição tem a mesma média, $\overline{d} = 44.22$).

Vejamos um exemplo concreto desta diferença. A figura 7.8 mostra a distribuição de grau[8] (CCDF) da rede da Wikipedia, apresentada na seção 7.3.2.

[8] O grau de um vértice neste caso é dado pela soma do seu grau de entrada com seu

Esta rede possui $n = 3.651.512$ vértices e $80.737.121$ de arestas, e calculando o valor para p de acordo com a equação (24), temos que $p = 1.21 \times 10^{-5}$. A figura 7.8 também mostra a distribuição de grau do modelo $G(n, p)$ com estes parâmetros, dada pela equação (21). Repare que a distribuição empírica da Wikipedia e do modelo $G(n, p)$ parametrizado possuem a mesma média, mas a similaridade terminar por ai. Observe que no modelo $G(n, p)$ a probabilidade de um vértice ter grau maior do que 100 é de 10^{-12} enquanto na Wikipedia esta probabilidade é quase 10^{-1}, muitas ordens de grandeza maior!

Outro problema Do $G(n, p)$ está em seu coeficiente de clusterização. O coeficiente de clusterização médio no modelo $G(n, p)$ é dado por p, pois todas as arestas tem igual probabilidade de existirem (podemos mostrar este resultado formalmente). Sabemos que muitas rede reais são esparsas, apresentando uma densidade ρ muito baixa (ver equação (5)). Desta forma, ao parametrizarmos o modelo $G(n, p)$ de acordo com a equação (24) teremos um modelo $G(n, p)$ que irá gerar um coeficiente de clusterização muito baixo, pois $p = \rho$. Entretanto sabemos também que muitas redes reais possuem coeficiente de clusterização muito mais elevados do que suas densidades, ordens de grandeza maior. Logo o modelo $G(n, p)$ não é capaz de representar o coeficiente de clusterização de redes reais.

Vejamos um exemplo com a rede de colaboração científica entre matemáticos apresentada na seção 7.3.1. A densidade desta rede é de apenas $\rho = 1.7 \times 10^{-5}$ enquanto seu coeficiente de clusterização é de 0.14. Repare que ao parametrizarmos o modelo $G(n, p)$ para representar este rede, o modelo teria um coeficiente de clusterização médio de $p = 1.7 \times 10^{-5}$, algumas ordens de grandeza menor do que a rede real. Precisamos de novos modelos!

7.4.3. Modelo de Barabási-Albert, BA

Apesar de ser o modelo mais popular e mais estudado para representar estrutura de redes, vimos que o $G(n, p)$ não captura aspectos fundamentais das estrutura presente em muitas redes reais. Em particular, ele falha em capturar a grande variedade de grau das redes reais, como vimos na seção 7.4.2.3.

No final da década de 90, Albert-Lázló Barabási e Réka Albert estavam estudando a rede da Web e perceberam que esta possuía uma distribuição de grau muito particular. Eles observaram empiricamente que a distribuição de grau de entrada era muito bem representada por uma lei de potência, muito parecida inclusive com a distribuição que observamos na rede da Wikipedia, ilustrada na figura 7.5. A busca por uma possível explicação para este fenômeno levou a eles a propor um modelo matemático para construção da Web [Barabási e Albert 1999]. A ideia era mostrar que esta característica seria consequência do processo de formação da Web. Mas como podemos resumir o processo de construção da Web? Um mecanismo muito simples é considerar que páginas novas que entram na Web tendem a criar hiperlinks para páginas

grau de saída, uma vez que a rede da Wikipedia é direcionada.

mais populares na Web. Além de simples, este mecanismo de certa forma captura aspectos importantes do complexo processo de formação da Web. Por exemplo, é mais fácil encontrar uma página mais popular do que outra desconhecida sobre o mesmo assunto, induzindo assim a criação de hiperlinks para páginas mais populares.

Mas como definir popularidade das páginas? Para o modelo ser simples, o ideal seria usar somente a estrutura da rede para definir a popularidade das páginas. Neste caso, nada mais simples do que o grau de entrada da página para definir sua popularidade. Ou seja, páginas com grau de entrada maiores são mais populares que páginas com grau de entrada menores. Além disso, para simplificar o modelo, esta relação poderia ser simplesmente linear: a chance de uma página receber um novo hyperlink é proporcional ao seu grau de entrada. Barabási e Albert batizaram este mecanismo de *preferential attachment* ("anexação preferencial") e propuseram o seguinte modelo.

Considere o tempo discreto $t = 0, 1, \ldots$. No tempo $t = 0$ temos a rede da Web inicial. A cada instante de tempo $t > 0$ um novo vértice (página) é adicionada a rede da Web. Este vértice traz com ele m arestas (hiperlinks) que tem uma ponta no próprio vértice e a outra em algum outro vértice já na rede. A escolha dos vértices na outra ponta de cada um destas m arestas é feita de forma aleatória. Seja u um vértice presente na rede no instante t e seja $d_u(t)$ seu grau neste instante. A probabilidade de u receber uma ponta de aresta é dada por:

$$p_u(t) = \frac{d_u(t)}{\sum_{v \in V(t)} d_v(t)} \tag{25}$$

onde $V(t)$ é o conjunto de vértices que estão na rede no instante t. Repare que esta probabilidade é proporcional ao grau de u no instante em que a nova página entra na rede. Desta forma, a probabilidade de u receber uma ponta de aresta de uma nova página depende de t. Por fim, o vértice v_i que entra na rede no instante i, passa a fazer parte do conjunto de vértices nos instantes seguintes, ou seja, $v_i \in V_j$ para todo $j > i$.

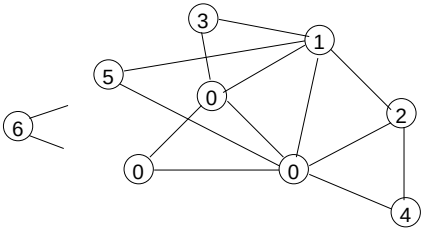


Figura 7.9. Modelo de Barabási-Albert para formação de redes com $m = 2$ no instante $t = 6$.

A figura 7.9 ilustra o instante $t = 6$ com a chegada do vértice 6 para o

modelo com $m = 2$. Os vértices com número 0 representam os vértices iniciais da rede, criados no tempo $t = 0$. Podemos calcular a probabilidade de cada um dos vértices receber uma ponta de aresta trazida pelo vértice 6. De acordo com a equação (25), por exemplo, o vértice 1 tem 0.38 de chance de receber uma das arestas, enquanto o vértice 4 tem 0.15. Vamos ver como obter estes números mais facilmente.

Podemos observar na figura que no instante de tempo $t = 6$ o número total de arestas é dado por $3 + 2 \times 5 = 13$, pois no inicio temos 3 arestas e cada vértice traz 2 arestas. Podemos generalizar este raciocínio e definir $m(t)$ o número de arestas no instante t, que é dado por:

$$m(t) = m_0 + mt \tag{26}$$

onde m_0 é o número de arestas na rede inicial, criada no instante $t = 0$. Como a soma dos graus dos vértices é igual a duas vezes o número de arestas (ver equação (3)), podemos simplificar a equação (25), de forma que:

$$p_u(t) = \frac{d_u(t)}{2(m_0 + mt)} \approx \frac{d_u(t)}{2mt} \text{ para } t \text{ grande o suficiente.} \tag{27}$$

Podemos também obter o número de vértices na rede no instante t, definido por $n(t)$. Como a cada instante um vértice entra na rede, o número total de vértices será dado por:

$$n(t) = n_0 + t \tag{28}$$

onde n_0 é o número de vértices na rede inicia, criada no instante $t = 0$. Podemos agora calcular o grau médio da rede, será dado por:

$$\overline{d}(t) = \frac{2m(t)}{n(t)} = \frac{2m_0 + 2mt}{n_0 + t} \approx 2m \text{ para } t \text{ grande o suficiente.} \tag{29}$$

É importante notar que a influência da rede inicial na probabilidade de anexação aos vértices e no grau médio vai a zero conforme a rede cresce. De fato, a rede inicial não possui muita influência sobre o processo de formação. Por fim, cada uma das m pontas de arestas trazidas pelo vértice v_i no instante i é escolhida de forma independente e com repetição de vértices de acordo com a equação (25). Este mecanismo pode dar origem a redes com múltiplas arestas entre dois vértices mas iremos permitir este tipo de aresta para simplificar o modelo e a análise. Entretanto, esta característica acontece com baixa probabilidade principalmente quando a rede cresce, não tendo influência significativa sobre a rede gerada.

Agora que entendemos a mecânica do modelo de Barabási-Albert (BA), vamos voltar a sua motivação original, que era explicar por que a rede da Web possui uma distribuição de grau de entrada que parece seguir uma lei de potência. Para isto precisamos determinar a distribuição de grau que será induzida pelo modelo BA. Em particular, depois de um tempo t grande o suficiente, qual será a distribuição de grau da rede gerada pelo modelo?

Podemos mostrar que a distribuição de grau do modelo BA será aproximadamente[9]:

$$P[d_u(t) = k] \approx \frac{2m^2}{k^3} \tag{30}$$

Observe que temos uma lei de potência com expoente 3 (média e variância finita) e que esta distribuição não depende do instante de tempo t em que estamos observando a rede, desde que t seja grande suficiente. Apesar da aproximação, este resultado foi inicialmente verificado por simulações numéricas e posteriormente formalizado.

O resultado obtido acima é bem interessante pois oferece uma possível explicação para a lei de potência observada empiricamente na rede da Web. Esta característica peculiar seria uma consequência do processo de formação da Web, que de certa forma é bem representado pelo mecanismo de anexação preferencial. Temos então que qualquer processo de formação baseado neste mecanismo irá dar origem a redes com lei de potência em seus graus. Na verdade, o mecanismo de anexação preferencial e suas consequências já haviam sido estudados por Simon na década de 60 no contexto de modelos econômicos para acúmulo de renda, e por Price na década de 70 no contexto de redes de citação de artigos científicos [Simon 1955, de Solla Price 1965]. Barabási e Albert redescobriram o mecanismo, demonstrando sua aplicação no contexto da rede da Web. Esta redescoberta chamou a atenção de muitos pesquisadores, que vieram a aplicar o mecanismo de anexação preferencial em diversos outros contextos.

Por fim, não é difícil encontrar aspectos fundamentais da Web que não são capturados pelo modelo BA. Por exemplo: a Web está em constante evolução, com páginas entrando e saindo da rede e hiperlinks sendo criados e removidos, diariamente; páginas mais antigas na Web não necessariamente possuem graus maiores; a distribuição empírica do grau de saída na Web também possui cauda pesada (no modelo é constante igual a m); o conteúdo das páginas é central no papel de criação dos hiperlinks; o expoente da lei de potência observada empiricamente não é exatamente 3. Desta forma, surgiram na última década diversos outros modelos e variações do BA para melhor capturar a evolução da Web e de outras redes [Krapivsky e Redner 2001, Leskovec et al. 2007]. O leitor mais interessado deve buscar pela literatura mais recente.

7.4.4. Modelo de Watts-Strogatz, WS

Apesar de capturar um aspecto central da estrutura de muitas redes reais, o modelo de Barabási-Albert não captura outros importantes aspectos encontrados em muitas redes. Em particular, o coeficiente de clusterização induzido pelo modelo BA é relativamente baixo, em contrapartida aos elevados coeficientes encontrados em redes reais, principalmente quando se tratando de redes sociais.

[9] A dedução deste resultado está disponível no suplemento deste texto [Figueiredo 2011].

Também no final da década de 90, Duncan Watts e Steven Strogatz estavam estudando a estrutura de diferente redes reais de diferentes domínios. Eles observaram que algumas destas redes possuíam as seguintes características estruturais em comum: alta clusterização, baixa distância, e muito esparsas. Eles batizaram estas redes de *small world network* ("rede de mundo pequeno") e propuseram um modelo matemático simples que fosse capaz de capturar estas três características [Watts e Strogatz 1998]. É importante ressaltar que o modelo *default* da época era o $G(n,p)$, que não é capaz de exibir alta clusterização e ser esparso ao mesmo tempo, conforme discutimos na seção 7.4.2.3.

O modelo de Watts-Strogatz é baseado em uma configuração inicial de rede regular, esparsa e com alta clusterização, seguida de mudanças aleatórias nesta estrutura para induzir baixas distâncias. O modelo WS não representa ou captura nenhum aspecto específico do processo de criação de redes reais, sendo neste sentido bastante estético, conforme descrito abaixo.

A configuração de rede inicial do modelo é um círculo com n vértices, todos com grau igual a 2. Considere um vértice u desta rede e seja $N_u(k)$ o conjunto de vértices que estão a distância menor ou igual a k do vértice u. Ou seja, $N_u(k) = \{v | l(u,v) \leq k\}$, onde $l(u,v)$ é a distância entre os vértices u e v. Para cada vértice u da rede, adicionamos arestas entre u e todos os vértices $v \in N_u(k)$. Este processo inicial dá origem a um látice regular onde cada vértice tem grau $2k$ e estão arrumados em um círculo. A figura 7.10 ilustra a configuração inicial da rede para $n = 18$ e $k = 2$.

Após a configuração inicial, o modelo determina que cada aresta da rede deve ser reposicionada com probabilidade p. Ao ser escolhida, uma aresta é reposicionada de forma aleatória e uniforme entre todos os possíveis pares de vértices. Ou seja, o par não-ordenado destino desta aresta é escolhido uniformemente entre todos os possíveis pares não-ordenados, $\binom{n}{2}$. O processo de formação da rede termina quando vez que todas as arestas tiverem sido consideradas para serem reposicionadas.

Repare que o modelo WS possui três parâmetros, n, k e p. O parâmetro p, que determina a probabilidade de reposicionamento das arestas é fundamental, pois controla que tipo de rede o modelo irá gerar. Por exemplo, quando $p = 0$ temos apenas o látice regular inicial, e quando $p = 1$ todas as arestas serão reposicionadas aleatoriamente, e temos uma rede parecida com o $G(n,p)$. O parâmetro p controla a aleatoriedade da rede, variando entre um látice regular e uma rede totalmente aleatória, sem estrutura. Iremos ver que o interessante acontece no meio do caminho! A figura 7.10 ilustra a configuração final de redes geradas pelo modelo, com valores de p diferentes.

Vamos caracterizar algumas propriedades estruturais do modelo WS. Seja $\overline{l}(p)$ e $\overline{c}(p)$ a distância média e o coeficiente de clusterização médio quando o modelo WS possui parâmetro p. Vamos assumir que n é grande suficiente e que $k \ll n$. Intuitivamente, quando p é muito pequeno, teremos uma rede com alta clusterização e distância alta, pois estamos quase com o látice regular

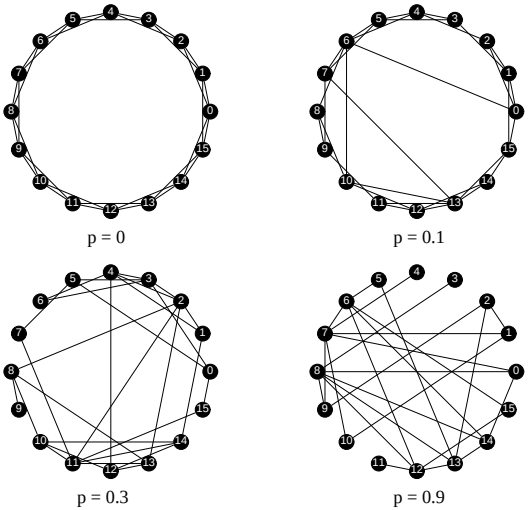


Figura 7.10. Exemplo de redes geradas pelo modelo WS para diferentes valores de p (com $n = 18$ e $k = 2$).

inicial. Quando p é muito grande, teremos uma rede com baixa clusterização, pois teremos distribuído o látice inicial, e distância baixa. Repare que, intuitivamente, ambos valores médios diminuem com p e talvez estas duas propriedades sejam antagônicas. Será que teremos alta clusterização e distância baixa simultaneamente, para algum valor de p intermediário?

Vamos começar calculando os valores exatos para o caso $p = 0$. Repare que neste caso, todos os vértices são idênticos, pois temos o látice inicial, de forma que a média da rede é igual ao valor da propriedade em qualquer vértice. Neste caso temos:

$$\overline{c}(0) = \frac{3(k-1)}{2(2k-1)} \approx \frac{3}{4} \tag{31}$$

$$\overline{l}(0) = \frac{n+2k-1}{4k} \approx \frac{n}{4k} \tag{32}$$

De forma esperada, podemos ver que quando $p = 0$ temos um coeficiente de clusterização alto e constante e distância proporcional a n/k.

No outro extremo $p = 1$, temos uma rede que se parece muito com uma rede gerada pelo modelo $G(n, p)$. Desta forma, suas características estruturais serão aproximadamente as mesmas. Usando os resultados da seção 7.4.2 e

observando que o grau médio da rede é dado $2k > 1$, temos:

$$\overline{c}(1) \approx \frac{2k}{n} \tag{33}$$

$$\overline{l}(1) \approx \frac{\log n}{\log k} \tag{34}$$

No extremo $p = 1$, temos uma clusterização muita baixa, inversamente proporcional a n, mas distâncias bem curtas, proporcionais a $\log n$. É importante ressaltar que estas duas aproximações podem ser obtidas formalmente [Newman 2000, Newman et al. 2000].

Vejamos o que ocorre com valores intermediários de p. Vamos considerar a relação $\overline{c}(p)/\overline{c}(0)$ e $\overline{l}(p)/\overline{l}(0)$ para diferentes valores de p. Esta relação determina a razão entre o valor da propriedade quando temos um valor qualquer para o parâmetro p e seu maior valor possível, obtido quando $p = 0$. Desta forma, podemos comparar em um mesmo gráfico o decrescimento dos valores médios relativos em função de p.

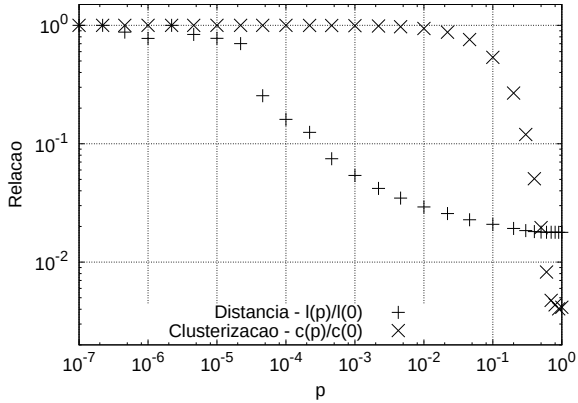


Figura 7.11. Relação entre distâncias médias ($\overline{l}(p)/\overline{l}(0)$) e coeficientes de clusterização médios $\overline{c}(p)/\overline{c}(0)$ no modelo WS em função de p ($n = 10^4$, $k = 15$).

A figura 7.11 ilustra exatamente as relações acima para o caso $n = 10^4$ e $k = 15$. Observamos que a distância começa a decrescer bem antes da clusterização. Desta forma, temos valores de p para os quais a rede exibe relativamente alta clusterização e baixas distâncias, por exemplo, entre $p = 10^{-2}$ e $p = 10^{-1}$. Conseguimos desta forma, obter redes de mundo pequeno com o modelo WS, que era seu objetivo. Por conseguir induzir redes de mundo pequeno o modelo WS também é conhecido como modelo *small world*.

Apesar do resultado ilustrado na figura 7.11 ter sido obtido numericamente, com simulações do modelo, é possível realizar um tratamento analítico aproximado das características induzidas pelo modelo [Newman 2000] [Newman et al. 2000]. Estes estudos mostram de forma aproximada que existem valores de p para os quais o modelo induz redes de mundo pequeno. Entretanto, podemos entender de forma intuitiva o que está ocorrendo. Quando p é bem pequeno, apenas algumas arestas são reposicionadas. Por serem escolhidas de forma aleatória e uniforme, estas arestas reposicionadas formam atalhos no látice, reduzindo drasticamente a distância entre muitos pares de vértices. Desta forma, são necessários poucos atalhos para reduzir a distância média da rede, o que pode ser obtida com p relativamente pequeno. Ao mesmo tempo, o reposicionamento de poucas arestas não influencia de forma significativa o látice inicial, que preserva sua estrutura e consequentemente deixa a rede com alta clusterização.

Assim como o modelo BA, o modelo WS também despertou a atenção de muitos pesquisadores e acabou recebendo muitas críticas. Por exemplo, o modelo deixa de capturar um aspecto fundamental de muitas redes reais, que é a distribuição de grau com cauda longa ou lei de potência. O modelo também é muito estético, não ajudando compreender de forma mais fundamental a razão para o surgimento destas características em redes reais. Desta forma, muitos outros modelos e variações do modelo WS foram propostos desde o seu surgimento [Newman 2000, Newman et al. 2002]. O leitor interessado neste aspecto deve buscar a literatura mais recente.

7.5. Funcionalidade e Processos

Um dos maiores objetivos ao estudarmos como as coisas se conectam é compreender como esta conectividade influencia a funcionalidade e os processos relacionados a estas coisas. Por exemplo, queremos entender a influencia da estrutura da rede da Web no processo de busca de informação. Será que sua atual estrutura facilita ou dificulta encontrar informação na Web? Voltando ao estudo do nosso sociólogo (ver seção 7.1, qual será a influência da estrutura da rede social de uma comunidade no espalhamento de boatos? Será que a percepção que boatos se espalham mais rapidamente por determinadas comunidades não pode ser justificada pela estrutura das suas respectivas redes sociais? Em um outro exemplo, queremos descobrir a influência da estrutura da rede de contato físico entre pessoas dentro de um hospital no espalhamento de infecção hospitalar. Como podemos modificar esta estrutura para reduzir a chance de alastramento de uma infecção? Este é um aspecto central no estudo de redes: caracterizar e entender como a estrutura influencia a funcionalidade.

Além disso, funcionalidade e estrutura muitas vezes se misturam não estando claro quem influencia quem. Uma rede pode possuir determinada estrutura justamente por estar almejando determinada funcionalidade. Por exemplo, talvez a estrutura da Internet seja simplesmente o resultado de um esforço de construir uma rede robusta e eficiente. Ou seja, sua estrutura é produto de sua

funcionalidade. Mas ninguém planejou ou gerenciou a evolução da Internet e sua estrutura é construída de forma distribuída, sem nenhuma entidade central. Então certamente as funcionalidades que a rede possui tem que operar sobre a estrutura que a Internet oferece. Por causa de sua estrutura, algumas funcionalidades podem não ser possíveis, ou extremamente ineficientes. Ou seja, a funcionalidade que a rede possui é produto da estrutura. Mas qual das hipóteses é verdadeira? O mais provável é que as duas sejam, dando origem a uma dependência mútua capturada por uma dinâmica evolucionária da rede: a estrutura influencia a funcionalidade, que por sua vez influencia a estrutura, que influencia a funcionalidade e assim por diante. Esta evolução contínua transforma tanto estrutura quanto funcionalidade ao longo to tempo. Entender esta dinâmica e a dependência mútua entre estrutura e funcionalidade é atualmente tema de pesquisa científica sendo estudada por muitos pesquisadores e um dos grandes desafios [Barrat et al. 2008, Newman 2010].

Veremos nesta seção dois aspectos funcionais fundamentais presentes em muitas redes. O primeiro está relacionado a capacidade das redes de funcionar quando temos falhas, que chamamos de robustez. Intuitivamente, os danos causados por falhas na rede irão depender da estrutura da rede e de como as falhas ocorrem. Veremos como as estruturas influenciam a robustez. O segundo aspecto funcional está relacionado a capacidade de "navegarmos" pelas redes com o objetivo de chegar a algum lugar específico, que chamamos de navigabilidade. Em particular, estamos interessados em entender como a estrutura influencia um processo de navegação que utiliza apenas informação local. Veremos que quando a estrutura é adequada, as redes podem ser navegadas com tremenda eficiência. Estas duas funcionalidades tem aplicações em diferentes contextos e diferentes rede, como por exemplo, a robustez da Internet a falhas de roteadores, ou a navigabilidade de uma rede social ao tentarmos encontrar uma pessoa que não conhecemos.

7.5.1. Falhas e robustez

Intuitivamente, a robustez de uma rede é a capacidade da mesma de operar na presença de falhas. Repare que esta definição é bastante abrangente, pois tanto "operar" quanto "falhas" podem ter inúmeros significados. Para estudar matematicamente a robustez de uma rede precisamos dar significado preciso a estas coisas. Uma saída elegante é não definir o que significa operar, pois este critério em geral depende do contexto da rede e de conhecimento das funcionalidades que a mesma exerce. Iremos apenas caracterizar a estrutura da rede após as falhas sem dizer se a mesma será capaz de operar ou não sobre esta nova estrutura. De forma similar, falhas nas redes podem ocorrer de muitas maneiras que em geral dependem do contexto da rede. Iremos definir alguns modelos de falhas simples que caracterizam de forma matemática como as falhas ocorrem. Obviamente, muitos outros modelos de falhas serão possíveis e você deve ter isto sempre em mente.

De forma geral, a capacidade de uma rede operar está relacionada com

as propriedades estruturais da rede. Neste sentido, um aspecto importante é o tamanho relativo da maior componente conexa, pois indica a maior fração de vértices que estão conectados entre si. Em muitas redes, ter uma componente conexa muito grande está diretamente relacionado com sua capacidade de operar. Por exemplo, considere a Internet. Enquanto a fração relativa da maior componente conexa é 1, todos os vértices podem se comunicar entre si; quando esta fração é 0.5 então apenas metade dos vértices, diminuindo a capacidade da rede de operar. Outra propriedade importante são as distâncias entre os vértices, por exemplo, o diâmetro da rede ou a distância média. Em geral, uma rede com distâncias menores opera melhor que uma rede com distâncias muito maiores. Novamente, a Internet é um bom exemplo, pois distância nesta rede está diretamente relacionado a atraso para entregar a informação (e.g., sua mensagem de chat). Apesar de não abordarmos, é importante notar que muitas propriedades podem ser utilizadas para caracterizar a capacidade de uma rede de operar, e que em alguns casos as propriedades mais adequadas dependem do contexto.

Falhas em redes podem ocorrer de muitas maneiras distintas e para representar estas falhas iremos trabalhar com modelo de falhas. Um modelo de falhas é uma abstração simplificada de algum processo de falha real. Existem muitos modelos de falhas e iremos estudar um modelo de falha bem simples, onde apenas os vértices podem falhar. Quando um vértice falha neste modelo, todas as arestas incidentes ao vértice também falham. Repare que a falha de um vértice é equivalente a remoção do vértice da rede, assim como das arestas incidentes ao mesmo. Desta forma, podemos pensar em falhas como sendo um processo de transformação estrutural na rede: a estrutura original da rede será transformada pelo modelo de falhas em uma nova estrutura.

O modelo de falha de vértice precisa também determinar *quais* vértices irão falhar. Existem apenas duas alternativas fundamentais para isto: falha aleatória e falha determinística. Na falha aleatória os vértices falham aleatoriamente e uma distribuição de probabilidade é necessária para determinar a probabilidade de um vértice específico falhar. Na falha determinística os vértices falham deterministicamente e é necessário então definir uma ordenação dos vértices, que determina a ordem em que os mesmos irão falhar. Repare que a escolha de falha aleatória ou determinística, assim como a probabilidade de falha e critério de ordenação dos vértices, devem ser feitas de forma que o modelo de falha capture a essência do real processo de falha que estamos representando. No que segue, veremos dois modelos de falhas específicos.

7.5.1.1. Falhas aleatórias

O mais simples modelo de falha aleatório assume que todos os vértices falham com igual probabilidade (distribuição uniforme), que os eventos de falha ocorrem de forma independente (i.e., a falha de um vértice não influencia na falha de nenhum outro) e que todos os vértices tem uma chance de falhar.

Repare que este modelo possui um único parâmetro que é a probabilidade de um vértice falhar, que chamaremos de q. Ou seja, cada vértice da rede permanece na rede pós-falha com probabilidade q.

O modelo acima é bastante representativo de falhas decorrentes de defeitos ou causas aleatórias. Por exemplo, a Internet tem milhões de roteadores e em qualquer período de tempo alguns falham por defeito em seu hardware (i.e., queimam). Estas falhas são bem representadas por este modelo aleatório.

Mas queremos entender como este modelo transforma a estrutura da rede. Em particular, como este modelo influencia as propriedades estruturais da rede. Iremos focar no tamanho relativo da maior componente conexa e determinar como este modelo afeta esta importante propriedade estrutural. Veremos que a influencia deste tipo de falha irá depender fortemente da estrutura da rede. Mas antes de iniciarmos uma dedução mais formal, vamos pensar neste processo de forma intuitiva.

Considere o modelo $G(n,p)$ com $p > c\log n/n$ e n grande o suficiente. Temos então uma rede conexa com distribuição de grau binomial com média proporcional a $\log n$. Ao aplicarmos o modelo de falha acima sobre esta rede iremos remover em média uma fração q dos vértices. Qual impacto que isto terá sobre o tamanho da maior componente conexa? Se q for pequeno o suficiente o impacto será desprezível e continuaremos com uma única componente conexa. Se q for grande o suficiente o impacto será devastador, desmembrando a rede em muitas componentes conexas, todas muito pequenas. Mas quando mudamos de um regime para outro? Se $q > c\log n/n$ então iremos remover uma fração dos vértices de forma que o grau médio dos vértice após a falha fique menor do que 1. Sabemos que neste regime o modelo $G(n,p)$ não possui uma componente conexa gigante e sua maior componente é da ordem de $\log n$ (ver seção 7.4.2). Intuitivamente, teremos uma rede fragmentada neste caso, com a maior componente conexa tendo um tamanho exponencialmente menor do que o número de vértices (mas repare que o número de vértices também diminuiu, o que deveria ser levado em consideração).

Considere agora o modelo BA, apresentado na seção 7.4.3. Temos uma rede conexa com distribuição de grau que se aproxima de uma lei de potência. Qual é o impacto que o modelo de falhas acima terá sobre o tamanho da maior componente conexa da rede? De forma intuitiva, o mesmo deve acontecer: para q pequeno o suficiente o impacto é desprezível; para q grande o suficiente o impacto é devastador. Mas onde ocorrerá esta transição? Qual estrutura de rede é mais susceptível a este modelo de falhas para um mesmo valor de q? Considere os vértices removidos pelo modelo de falhas aleatório na rede gerada pelo modelo BA. Intuitivamente, a grande maioria destes vértices terá um grau bem pequeno. Isto ocorre, pois apesar de todo vértice ter igual probabilidade de ser removido (q), o número de vértices com grau pequeno é muito grande, pois no modelo BA a distribuição de grau segue uma lei de potência. Ao serem removidos, estes vértices terão impacto desprezível na rede, pois são "folhas" e não estão no caminho mais curto entre um outro par

de vértices quaisquer. Além disso, vértices de grau muito alto mantém a rede conectada. Desta forma, intuitivamente, redes do modelo BA toleram maiores valores de q antes de se despedaçarem, sendo mais robustas que redes do modelo $G(n,p)$. A formalização desta intuição está disponível no suplemento deste texto [Figueiredo 2011].

7.5.1.2. Falhas determinísticas

Conforme mencionamos anteriormente, um modelo de falha alternativo ao aleatório é o modelo determinístico. Neste modelo, os vértices são ordenados de forma a definir a sequência em que irão falhar. É necessário, desta forma, usarmos um critério para ordenação dos vértices. Obviamente, este critério está relacionado ao processo real de falha que se deseja capturar com o modelo.

Um critério interessante e bastante utilizado é a ordenação decrescente dos graus dos vértices. Ou seja, o vértice de maior grau irá falhar primeiro, o de segundo maior grau irá falhar segundo, e assim por diante. Este critério é bastante representativo de falhas propositais, realizadas com o intuito de tentar desconectar a rede o mais rápido possível. Por exemplo, um hacker que deseja parar a Internet por algumas horas deve atacar o roteador de maior grau, na tentativa de causar um estrago maior. Certamente escolher um roteador ao acaso para atacar não trará muitas consequências, pois é grande a chance deste roteador não ter importância alguma para a rede como um todo (a Internet também tem distribuição de cauda pesada). Logo ataques propositais ou direcionados são bem representados por este modelo de falha.

Seja q o parâmetro que define a fração de vértices que irão falhar. Estamos interessados em entender como este modelo de falha transforma a estrutura das redes. Em particular, queremos caracterizar o tamanho ou existência da componente conexa gigante na rede pós-falha. Como era de se esperar o estrago que estas falhas irão causar depende fortemente da estrutura da rede original. Vamos discutir o que acontece de forma intuitiva, novamente para os modelos $G(n,p)$ e BA.

No modelo $G(n,p)$ com $p > c\log n/n$ temos uma rede conexa e vértices com grau que segue uma distribuição binomial. Repare que por ser binomial, os graus dos vértices não diferem muito pois a cauda da distribuição é exponencial. Como vimos na seção 7.4.2.3 é desprezível a probabilidade de termos um vértice com grau 10 vezes a média. Com isto, intuitivamente, podemos concluir que a remoção de uma fração q de vértices de maior grau terá um impacto na rede muito parecido com a remoção de uma fração q qualquer, escolhida de forma aleatória. De fato, isto ocorre e pode ser verificado experimentalmente e através de análise teórica aproximada [Albert et al. 2000, Callaway et al. 2000].

Vejamos agora o que ocorre com o modelo BA. Neste caso, temos uma rede conexa com distribuição de grau que segue uma lei de potência. Logo, a

grande maioria dos vértices tem grau pequeno e muito poucos tem grau muito grande, muitas vezes maior do que a média. Ao remover os vértices em ordem decrescente de grau, estaremos removendo os vértices de maior grau da rede. Estes vértices por sua vez são responsáveis por conectar muitos de outros vértices, pois tem um grau muito maior do que a média. Intuitivamente, o impacto será devastador mesmo quando removemos apenas uma pequena fração dos vértices. De fato, esta observação pode ser comprovada experimentalmente a análise teórica aproximada [Albert et al. 2000, Callaway et al. 2000].

É muito interessante que o modelo BA seja altamente robusto a falhas aleatórias e ao mesmo extremamente frágil a falhas determinísticas, direcionadas aos vértices de maior grau. Esta característica ajudou a popularizar a ideia de que algumas redes são robustas e frágeis ao mesmo tempo [Albert et al. 2000]. Este fato pode ter uma conotação positiva ou negativa, dependendo do contexto. No exemplo da Internet, esta observação tem conotação negativa pois implica que a rede pode ser mais facilmente atacada, interrompendo seu funcionamento. Entretanto, se tratando de uma rede de contato físico entre pessoas e espalhamento de um vírus ou uma doença, a observação tem conotação positiva, pois implica que podemos "remover" (por exemplo, vacinar) poucos indivíduos para conter eficientemente uma epidemia. Por fim, vale ressaltar que as observações acima não estão ligadas diretamente aos modelos $G(n, p)$ e BA, e sim a distribuição de grau que estes modelos induzem na rede. De forma geral, podemos estender os resultados a redes que possuem distribuição de grau com cauda exponencial e distribuição de grau com cauda pesada.

A influencia da estrutura da rede nos danos causados por falhas vem sendo muito estudado, inclusive sobre a perspectiva de outras propriedades estruturais, como distâncias, e outros modelos de falhas e de redes. O leitor interessado neste assunto deve procurar pela literatura mais recente.

7.5.2. Busca e navigabilidade

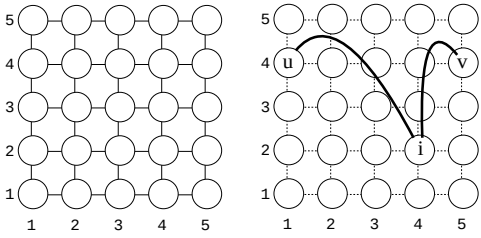
O problema de encontrar informação em redes é fundamental tendo diversas aplicações em diferentes contextos. Considere o problema de encontrar uma pessoa que você não conhece, mas sobre a qual você tem algumas informações, como nome, profissão, e endereço. Seria possível encontrar esta pessoa usando sua rede social? Por exemplo, você poderia pedir a um amigo para apresentar você a esta pessoa, e caso seu amigo não a conhecesse, ele poderia apresentar você a um amigo dele que talvez a conheça. Seria possível encontrar sequências de relacionamento de amizades que sejam curtas e que cheguem ao destino? Repare que não temos informação global sobre a rede social, ou seja, não sabemos quem conhece quem. O processo de decisão se baseia somente em informação local, que determina a quem devemos perguntar sobre a pessoa com a qual queremos encontrar.

Os resultados dos experimentos de Milgram, descritos na seção 7.3.1, mostram que não somente caminhos curtos existem em nossa redes social de amizade, mas somos capazes de buscar informação nesta rede de forma eficiente

utilizando apenas informação local. Mas o quanto desta capacidade de navegar a rede eficientemente está relacionada com a estrutura da mesma? Esta última observação e pergunta foram feitas por Kleinberg, também no final da década de 90, que propôs um modelo matemático para estudar este fenômeno e responder a esta pergunta.

Para formular o problema matematicamente, precisamos modelar o processo de busca que utiliza apenas informação local para navegar pela rede. Iremos modelar este processo de busca utilizando um algoritmo guloso, que funciona da seguinte maneira. Considere um vértice u da rede que possui uma mensagem que deve ser entregue ao vértice x, sobre o qual conhecemos algumas informações. Se x é vizinho de u, então u entrega a mensagem a x e o processo de busca terminar. Caso contrário, u repassar a mensagem ao seu vértice vizinho que está *mais próximo* do vértice x e o processo recomeça com este vizinho. Repare que precisamos definir uma função de distância entre um vizinho de u e o vértice destino x. Esta função distância pode utilizar apenas informação local ao vértice u, pois u não tem informação global sobre a rede.

Voltando ao exemplo acima com a rede social de amizade, a distância poderia ser a distância física entre as residências das pessoas. Por exemplo, conhecendo o endereço do vértice x, você poderia repassar a mensagem para seu amigo que mora mais próximo de x. Repare que isto dá origem a um algoritmo guloso, que sempre tenta dar o maior salto possível na direção do alvo.



(a) Rede original. (b) Atalho do vértice i.

Figura 7.12. Modelo de Kleinberg para estrutura de rede possivelmente navegável ($n = 5$).

O modelo de rede aleatória que Kleinberg se propôs a estudar é baseado em um látice regular adicionado de atalhos [Kleinberg 2000]. Considere um reticulado de duas dimensões com n^2 vértices (ou seja, o comprimento do lado do quadrado tem n vértices). Cada vértice está conectado aos seus vizinhos norte, sul, leste, e oeste, tendo todos grau igual a 4 (com exceção das bordas).

A figura 7.12a ilustra o reticulado. Todo vértice $i \in V$ da rede possui uma posição no reticulado, dada por (x_i, y_i) com $x_i, y_i \in [1, \ldots n]$. A distância entre dois vértices i e j é simplesmente dada pela distância de Manhattan entre os vértices, ou seja:

$$l(i,j) = |x_i - x_j| + |y_i - y_j| \tag{35}$$

Repare que desta forma temos uma função de distância entre os vértices que utiliza apenas a posição dos vértices no reticulado, e não utiliza nenhuma outra informação da rede. Iremos usar esta métrica de distância no algoritmo guloso para chegar a um vértice destino.

O reticulado acima não possui distâncias curtas na rede, pois para irmos de um vértice qualquer a outro, escolhidos de forma aleatória, seriam necessários na ordem de n saltos pela rede, em média. Vamos então criar atalhos no reticulado. Cada vértice do reticulado irá adicionar uma aresta que será incidente sobre o vértice e um outro escolhido ao acaso. A probabilidade de escolha será inversamente proporcional a distância entre os vértices. Mais precisamente, dado o vértice i, a probabilidade que o vértice j será escolhido para receber o atalho é dada por:

$$P[i \to j] = C \frac{1}{l(i,j)^\alpha} \tag{36}$$

onde $\alpha \geq 0$ é um parâmetro constante do modelo e $C < 1$ é uma constante de normalização que garante que $\sum_{j \in V} P[i \to j] = 1$. Repare que a probabilidade segue uma lei de potência na distância, com expoente α. A figura 7.12b ilustra a rede e possíveis atalhos do vértice i. Por exemplo, a probabilidade do vértice i criar um atalho para os vértices u e v é proporcional a $5^{-\alpha}$ e $3^{-\alpha}$, respectivamente.

Agora queremos entender o comportamento do algoritmo de busca guloso sobre esta estrutura em função dos seus únicos parâmetros, que são n e α. Antes vamos entender intuitivamente onde os atalhos nos levam. Repare que a probabilidade de um atalho ligar dois vértices que estão a distância d no reticulado é proporcional a $d^{-\alpha}$, com α controlando a cauda desta distribuição. Repare que a probabilidade de termos atalhos muito longos é inversamente proporcional a α; quanto menor for α maior é a chance de termos atalhos muito longos.

Considere agora o algoritmo guloso que utiliza a função de distância dada pela equação (35) para navegar, a cada passo saltando para o vizinho mais próximo do destino. Se α for grande o suficiente, então atalhos longos não irão existir, pois a probabilidade de existirem será desprezível. Desta forma, caminhos curtos não existem e consequentemente o algoritmo guloso não vai ser eficiente. Entretanto, se α for pequeno o suficiente teremos muitos atalhos longos e caminhos curtos irão existir. Em particular, quando $\alpha = 0$ os atalhos terão distribuição uniforme sobre o reticulado, não mais dependendo de suas distâncias. Mas o que ocorre com o algoritmo guloso neste caso?

Repare que o algoritmo guloso neste modelo sempre irá convergir para o destino, pois sempre é possível chegar mais próximo do destino usando um vizinho adjacente no reticulado, sem usar atalhos. Além disso, vamos caracterizar o desempenho do algoritmo guloso contando quantos passos o algoritmo faz para chegar ao destino. Com isto, temos o seguinte teorema [Kleinberg 2000].

Teorema 7.5.1 *Seja T o número médio de saltos que o algoritmo guloso realiza para sair de um vértice inicial qualquer e chegar a um vértice destino escolhido ao acaso. Então temos:*

$$\begin{array}{lccl} (i) & 0 \leq \alpha < 2 & \rightarrow & T \geq \beta n^{(2-\alpha)/3} \\ (ii) & \alpha = 2 & \rightarrow & T \leq \beta (\log n)^2 \\ (iii) & \alpha > 2 & \rightarrow & T \geq \beta n^{(\alpha-2)/(\alpha-1)} \end{array}$$

onde β é uma constante que não depende de n.

O resultado do teorema acima é fantástico! Vejamos o que ele nos diz. Se a probabilidade de termos atalhos muito longos for alta ou baixa, ou seja, $0 \leq \alpha < 2$ ou $\alpha > 2$, então o algoritmo guloso precisa de um número polinomial de saltos para chegar ao destino. Entretanto, quando a probabilidade de atalhos longos for justamente correta, ou seja, $\alpha = 2$, então o guloso é capaz de encontrar seu destino com um número logarítmico de passos, exponencialmente mais curtos do que nos outros casos. Ou seja, a estrutura da rede irá definir sua navigabilidade dela!

É interessante que o desempenho do algoritmo guloso seja tão influenciável pela estrutura da rede. No teorema acima, vimos que se a rede não tiver uma estrutura perfeita, ou seja α exatamente 2, o algoritmo terá um desempenho bem ruim. Entretanto, não parece que algoritmos de busca em redes reais sejam tão sensíveis a estrutura assim. Por exemplo, navegamos bem por redes sociais apesar destas aparentemente não exibirem um rigor científico ao serem construídas, como requerido pelo teorema acima. Por que então algumas redes são navegáveis? Este é uma questão que continua em aberto e muitos trabalhos vem sendo realizados nesta direção.

Por fim, é importante notar que outros trabalhos foram realizados para medir a navigabilidade de redes reais. Alguns trabalhos empíricos vem mostrando que diferentes redes reais são navegáveis, quando utilizamos funções de distância adequadas [Liben-Nowell et al. 2005, Boguna et al. 2009]. Modelos alternativos ao proposto por Kleinberg também foram estudados, indicando que algoritmos gulosos também podem levar a caminhos curtos, se a estrutura for correta [Watts et al. 2002]. Enfim, o leitor interessado neste aspecto deve buscar a literatura mais recente.

7.6. Considerações finais

Agora você já tem uma ideia do que vem a ser esta nova área da Ciência, conhecida por Redes Complexas, que visa entender como as coisas se

conectam e o impacto desta conectividade na funcionalidade. Apesar dos estudos de algumas redes específicas e de modelos aleatórios de redes mais antigos (e.g., experimento de Milgram, modelo $G(n, p)$), foi somente na última década que pesquisadores, em escala mundial e de diversos ramos da Ciência, começaram a estudar de forma mais sistemática redes, suas estruturas e seus impactos. Hoje já temos diversos centros de pesquisa nos EUA e na Europa dedicados exclusivamente a área de Redes Complexas, que são interdisciplinares e contam com pesquisadores renomados de diferentes áreas. É inegável que a área de Redes Complexas poderá ter um papel central para a Ciência moderna, revelando aspectos fundamentais e ainda desconhecidos sobre o mundo que nos cerca.

Neste livro fizemos apenas uma introdução ao tema, mas cobrimos todas as etapas do processo de estudar como as coisas se conectam. Começamos com a caracterização das redes reais, de suas estruturas, de onde surge a necessidade de definir propriedades estruturais, que resumem e nos dão uma ideia da estrutura da rede. Antes disto, precisamos obter, descobrir ou inferir redes reais, que é um passo fundamental neste processo, pois como podemos estudar e criar teorias sobre objetos (estruturas da redes) que não conhecemos? Aliás, grande parte do avanço e da atenção que a área recebeu é fruto das muitas redes reais dos mais variados domínios que vem sendo obtidas e disponibilizadas para estudo. Em seguida tentamos generalizar as redes reais, criando modelos matemáticos que possam representar alguns aspectos de suas estruturas. Um bom modelo, além de simples, deve capturar aspectos estruturais importante da rede real que ele representa, e também ser motivado por mecanismos que remetam às redes reais. Um modelo simples em geral é susceptível a tratamento analítico, estabelecendo uma base teórica sobre seu comportamento. Por fim, estamos interessados na funcionalidade e nos processos que operam sobre estas redes. Queremos entender o impacto da estrutura na funcionalidade e o impacto da funcionalidade na estrutura. Para isto, precisamos modelar a funcionalidade que temos interesse e estudá-la sobre modelos aleatórios de redes, possibilitando um entendimento mais geral. O tema estrutura *versus* funcionalidade é central, sobre o qual pouco conhecemos e um dos grande desafios da área [Barrat et al. 2008].

Além dos aspectos apresentados, há ainda outros importantes aspectos no estudo de como as coisas se conectam que não foram abordados. Por exemplo, redes reais não são estáticas no sentido de possuirem uma estrutura topológica que não varia ao longo do tempo. Redes reais estão em contante mudança estrutural, para alguma escala de tempo, pois até os continentes se movem! Apesar disto, quase toda a discussão neste livro girou em torno de redes estáticas e consequentemente de estruturas estáticas (com exceção do modelo BA). O estudo de redes estáticas é uma simplificação do mundo real, muitas vezes bem justificada pela enorme diferença nas escalas de tempo dos processos que operam na rede e de suas mudanças estruturais. Entretanto, em algumas redes reais, mudanças estruturais e processos na rede ocorrem em escalas

de tempo similar. O estudo de redes dinâmicas está em sua infância e até mesmo o significado das propriedades estruturais precisam ser repensados. Apesar disto, alguns trabalhos já começam a propor modelos aleatórios para redes dinâmicas e entender processos sobre redes dinâmicas [Avin et al. 2008, Clementi et al. 2008, Ribeiro et al. 2011, Grindrod e Higham 2010].

Um outro aspecto importante discutido apenas superficialmente no texto é a obtenção de redes reais. Como vimos, muitas redes reais não estão disponíveis prontamente, mesmo quando informação sobre a mesma é publica. Desta forma, é necessário utilizarmos algum mecanismo para descobrir a estrutura destas redes. Muitas vezes os mecanismos que podemos utilizar são restringidos pelo domínio e pelo tamanho da rede, que geralmente não permitem a descoberta da rede por completo. Temos então que projetar mecanismos capazes de inferir a estrutura da rede real. Mas como garantir que estes mecanismos descobrem redes representativas da rede real, sem tendências ou vícios? Este tema, conhecido por *network sampling*, é de extrema importância, pois as redes reais obtidas empiricamente são a base do processo de caracterização estrutural e modelagem matemática. Se estas estão fundamentalmente erradas por não refletir a realidade, todo o trabalho que se segue estará baseado em premissas falsas. Não por menos, muitos pesquisadores vem se dedicando a estudar este problema nos últimos anos [Achlioptas et al. 2005, Ribeiro e Towsley 2010, Kurant et al. 2011].

Concluímos aqui nossa introdução ao tema Redes Complexas e esperamos que o leitor tenha agora um bom entendimento de alguns dos avanços feitos pela área e dos seus muitos desafios. O leitor interessado em se aprofundar em algum tema específico deve começar sua busca na literatura pelas muitas referências apresentadas neste livro. Por fim, obrigado pela leitura!

Referências

[Achlioptas et al. 2005] Achlioptas, D., Clauset, A., Kempe, D. e Moore, C. (2005). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *ACM Symposium Theory of computing (STOC)*.

[Albert et al. 1999] Albert, R., Jeong, H. e Barabasi, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401(6749).

[Albert et al. 2000] Albert, R., Jeong, H. e Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378–82.

[Avin et al. 2008] Avin, C., Koucky, M. e Lotker, Z. (2008). How to explore a fast-changing world. (on the cover time of dynamic graphs). In *ACM International Conference on Automata, Languages and Programming (ICALP)*.

[Barabási e Albert 1999] Barabási, A. e Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286:509–512.

[Barrat et al. 2008] Barrat, A., Barthélemy, M. e Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge University Press.

[Boccaletti et al. 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. e

Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5).

[Boguna et al. 2009] Boguna, M., Krioukov, D. e Claffy, K. C. (2009). Navigability of complex networks. *Nature Physics*, 5(1).

[Bollobás 2001] Bollobás, B. (2001). *Random Graphs*. Cambridge University Press, 2nd edição.

[Brin e Page 1998] Brin, S. e Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *International World-Wide Web Conference (WWW)*.

[Broder et al. 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. e Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33(1-6):309 – 320.

[Börner et al. 2004] Börner, K., Maru, J. T. e Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proc. Natl. Acad. of Sciences of USA*, 101(Suppl 1):5266–5273.

[CAIDA 2011] CAIDA (2011). Archipelago measurement infrastructure. http://www.caida.org/projects/ark/.

[Callaway et al. 2000] Callaway, D. S., Newman, M. E. J., Strogatz, S. H. e Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85.

[Chen et al. 2002] Chen, Q., Chang, H., Govindan, R., Jamin, S., Willinger, W. e Shenker, S. (2002). The origin of power laws in internet topologies revisited. In *IEEE INFOCOM*.

[Clementi et al. 2008] Clementi, A. E., Macci, C., Monti, A., Pasquale, F. e Silvestri, R. (2008). Flooding time in edge-markovian dynamic graphs. In *ACM International Conference on Principles of Distributed Computing (PODC)*.

[de Kunder 2011] de Kunder, M. (2011). The size of the world wide web. http://www.worldwidewebsize.com/.

[de Solla Price 1965] de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

[Dodds et al. 2003] Dodds, P. S., Muhamad, R. e Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, 301.

[Durrett 2006] Durrett, R. (2006). *Random Graph Dynamics*. Cambridge University Press.

[Erdős e Rényi 1959] Erdős, P. e Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

[Erdős e Rényi 1960] Erdős, P. e Rényi, A. (1960). The evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

[Figueiredo 2011] Figueiredo, D. R. (2011). Introdução a redes complexas – suplemento. http://www.land.ufrj.br/∼daniel.

[Freire e Figueiredo 2010] Freire, V. e Figueiredo, D. R. (2010). Ranqueamento em redes de colaboração utilizando uma métrica baseada em intensidade do relacionamento. In *Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*.

[Grindrod e Higham 2010] Grindrod, P. e Higham, D. J. (2010). Evolving graphs: dynamical models, inverse problems and propagation. *Proc. of the Royal Society A: Mathematical, Physical and Engineering Science*.

[Grossman e Ion 2011] Grossman, J. e Ion, P. (2011). The Erdős number project. http://www.oakland.edu/enp/.

[Kleinberg 2000] Kleinberg, J. (2000). The small-world phenomenon. In *ACM Symposium on Theory of computing (STOC)*, pp. 163–170.

[Krapivsky e Redner 2001] Krapivsky, P. L. e Redner, S. (2001). Organization of growing random networks. *Phys. Rev. E*, 63(6):066123.

[Kurant et al. 2011] Kurant, M., Gjoka, M., Butts, C. T. e Markopoulou, A. (2011). Walking on a graph with a magnifying glass. In *ACM SIGMETRICS*.

[Kwak et al. 2010] Kwak, H., Lee, C., Park, H. e Moon, S. (2010). What is twitter, a social network or a news media? In *International World-Wide Web Conference (WWW)*.

[Leskovec et al. 2007] Leskovec, J., Kleinberg, J. e Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1.

[Ley 2011] Ley, M. (2011). The dblp computer science bibliography. http://www.informatik.uni-trier.de/l̃ey/db/.

[Liben-Nowell et al. 2005] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. e Tomkins, A. (2005). Geographic routing in social networks. *Proc. of the National Academy of Sciences*, 102:11623–11628.

[Menezes et al. 2009] Menezes, G. V., Ziviani, N., Laender, A. H. F. e Almeida, V. A. F. (2009). A geographical analysis of knowledge production in computer science. In *International Worl-Wide Web Conference (WWW)*.

[Milgram 1967] Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60–67.

[Mitzenmacher 2004] Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2).

[Newman 2000] Newman, M. E. J. (2000). Models of the small world. *J. Stat. Phys.*, 101.

[Newman 2001] Newman, M. E. J. (2001). The structure of scientific collaboration networks. In *Proc. Natl. Acad. Sci. USA*, volume 98, pp. 404–409.

[Newman 2004a] Newman, M. E. J. (2004a). Coauthorship networks and pat-

terns of scientific collaboration. *Proc. Natl. Acad. Sci. USA*, 101:5200–5205.

[Newman 2004b] Newman, M. E. J. (2004b). Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks. *Complex Networks*, pp. 337–370.

[Newman 2005] Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46.

[Newman 2010] Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

[Newman et al. 2000] Newman, M. E. J., Moore, C. e Watts, D. (2000). Mean-field solution of the small-world network model. *Phys. Rev. Lett.*, 84.

[Newman et al. 2002] Newman, M. E. J., Watts, D. J. e Strogatz, S. H. (2002). Random graph models of social networks. *Proc. of the National Academy of Sciences*, 99.

[Redner 1998] Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4:131–134.

[Reynolds 2011] Reynolds, P. (2011). The oracle of Bacon. http://oracleofbacon.org/.

[Ribeiro et al. 2011] Ribeiro, B., Figueiredo, D. R., de Souza e Silva, E. e Towsley, D. (2011). Characterizing continuous-time random walks on dynamic networks. In *ACM SIGMETRICS (extended abstract)*.

[Ribeiro e Towsley 2010] Ribeiro, B. e Towsley, D. (2010). Estimating and sampling graphs with multidimensional random walks. In *ACM Internet Measurement Conference (IMC)*.

[Simon 1955] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42:425–440.

[Travers e Milgram 1969] Travers, J. e Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32:425–443.

[Watts e Strogatz 1998] Watts, D. e Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.

[Watts et al. 2002] Watts, D. J., Dodds, P. S. e Newman, M. E. J. (2002). Identity and search in social networks. *Science*, 296(5571):1302–1305.

[Wikipedia 2011a] Wikipedia (2011a). Caenorhabditis elegans. http://en.wikipedia.org/wiki/C._elegans.

[Wikipedia 2011b] Wikipedia (2011b). Observable universe. http://en.wikipedia.org/wiki/Observable_universe.