Capítulo

5

Minerando a Web por meio de Grafos - da teoria às aplicações

Ana Paula Appel, Estevam Rafael Hruschka Junior

Abstract

The volume of data represented as graphs, such as complex networks, has been increased over the past years making possible the creation of a new research area called graph mining. The main source of graph data is the World Wide Web, being link page structure one of the most common targets in this area. In graph mining, some tasks as statistical properties, community detection and link prediction can be highlighted. This course aims to presenting a global view of graph mining area; focusing on graph from Web like social networks. Also, we will introduce some basic concepts and the main techniques of graph mining main areas.

Resumo

O crescimento do volume de dados modelados como grafos, como as redes complexas, motivou e impulsionou a criação de novas áreas de pesquisa como a mineração de grafos. Atualmente, tais dados são provenientes principalmente da Web e uma das principais fontes é a estrutura de links entre as páginas. Nesta nova área de pesquisa, algumas tarefas se destacam, a saber: a extração de propriedades estatísticas, a detecção de grupos (comunidades), a predição de ligações (arestas), entre outras. Este minicurso tem como objetivo apresentar uma visão geral da área de mineração de grafos, focando principalmente em grafos vindo da Web, como ocorre com as redes sociais, por exemplo. Além de introduzir os conceitos básicos da mineração de grafos, serão também apresentadas as principais áreas e técnicas de mineração de grafos.

5.1. Introdução

Atualmente a Web é uma das fontes de dados heterogêneas mais ricas e em crescente evolução. A Web abrange diversos tipos de dados como multimídia, texto (dados não estruturados), dados semi-estruturados, entre outros. Por essas características, a Web apresenta desafios e oportunidades para a mineração de dados e descoberta de informação

e conhecimento. Nos últimos anos a mineração de ligações em conjuntos de dados estruturados tem recebido uma grande atenção de pesquisadores da área, principalmente com relação às redes complexas.

Uma rede complexa é, normalmente, modelada como um grafo, ou seja, a rede complexa é representada através de um objeto matemático cujos nós, também chamados vértices, modelam elementos (que podem ser páginas web, pessoas, computadores) e as arestas modelam relacionamentos entre os nós.

Uma característica muito importante, nesta abordagem de mineração de dados, é que os grafos que representam problemas reais tendem a ser irregulares e por isso eles são chamados de redes complexas. A modelagem da Web através da abordagem de redes complexas tem sido foco de muitas pesquisas atualmente. Com o intuito de guiar o leitor interessado em mineração de grafos aplicada neste tipo de redes, este documento apresenta os conceitos básicos necessários para a compreensão e a execução de tarefas de mineração de dados estruturados vindos da Web e representados como redes complexas. Assim, vários domínios de aplicação podem ser abordados, tais como as redes sociais (Facebook, Flickr, Orkut, etc), ontologias, redes acadêmicas (DBLP, ARXIV) e a própria WWW - World Wide Web.

Outro aspecto bastante relevante e que teve papel de destaque no aparecimento desta nova área de pesquisa dentro da mineração de dados foi a limitação dos algoritmos tradicionais quando aplicados em dados modelados como redes complexas. Os algoritmos de mineração de dados tradicionais, tais como regras de associação, detecção de agrupamentos, classificação, entre outros, usualmente encontram padrões em relações únicas que armazenam uma coleção de instâncias independentes. Um desafio emergente para a descoberta de conhecimento é o problema de minerar coleções de dados que estão inter-relacionados. Tais inter-relações podem ser naturalmente representadas através de grafos e armazenadas em diversas relações. A Figura 5.1 apresenta o grafo que representa a rede Web do Google[1].

Grafos são representações convenientes para um conjunto numeroso de dados inter-relacionados, como ocorre nas redes complexas, representadas pelas redes sociais, redes de publicações científicas, autores vs. participação em conferencias, e muitos outros. Nas últimas décadas, o volume de dados representados como grafo (as redes sociais LinkedIn, Facebook e Flickr, por exemplo), tem aumentado exponencialmente fazendo com que houvesse uma mudança no paradigma de como as redes são analisas [Newman, 2003]. Assim, ao invés das redes serem analisadas de uma maneira centralizada, com nós e arestas sendo estudadas individualmente, as redes passaram ser analisadas de uma maneira mais geral por meio de propriedades estatísticas de larga escala.

O aumento no tamanho das redes complexas se deve ao fato, principalmente, destas redes serem construídas a partir de dados vindos da Web. Este dinamismo e a velocidade de crescimento destas massas de dados fazem com que a utilização de algoritmos tradicionais de mineração de dados não traga resultados satisfatórios nestes domínios. Com isso, a mineração de grafos tem se tornado essencial para extrair conhecimento de

[1]Disponível em:http://commons.wikimedia.org/wiki/Image:WorldWideWebAroundGoogle.png (Acesso em 09/08/2010)

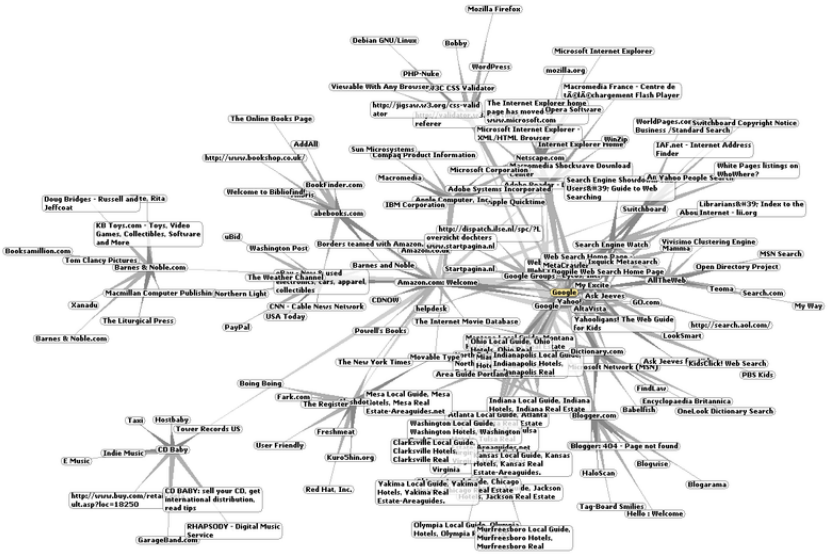


Figura 5.1. Rede Web da Google.

grandes redes complexas.

Muitas propriedades relevantes das redes complexas, algumas frequentemente não intuitivas, tais como diâmetro pequeno, distribuição do grau, triângulos, autovalores, presença de estrutura de comunidade, tem sido descobertas por meio da mineração das grandes redes complexas [Faloutsos et al., 1999, Leskovec et al., 2007, Clauset et al., 2007, Leskovec et al., 2008, Tsourakakis, 2008, McGlohon et al., 2008, Clauset et al., 2008]. Tais propriedades são importantes para o entendimento do comportamento e formação dessas redes. Além disso, essas propriedade provam que as redes complexas não são randômicas. Outro ponto é, que se a maioria dos nós de uma rede segue um padrão específico, o desvio de alguns nós desse padrão indica a presença de valores discrepantes ("outliers") que devem ser estudados.

5.2. Conceitos

As redes complexas, ou simplesmente redes, são um conjunto de elementos discretos que são representados pelos vértices e arestas, que são um conjunto de conexões entre os vértices. Os elementos e suas conexões podem representar, por exemplo, pessoas e ligações de amizade, computadores e linhas de comunicação [Faloutsos et al., 1999], componentes químicos e reações [Jeong et al., 2000], artigos e citações [Redner, 1998], entre outros. Assim, as redes complexas podem ser facilmente modeladas como um grafo.

Os grafos são capazes de abstrair os detalhes do problema ao descreverem características topológicas importantes com uma clareza que seria praticamente impossível se todos os detalhes fossem mantidos. Essa foi uma das razões por que a teoria dos grafos se espalhou, especialmente nos últimos anos, e tem sido utilizada por engenheiros, cientistas da computação e em especial por sociólogos.

Nesta seção serão apresentados os conceitos como os da teoria dos grafos, álgebra linear e outros que se fazem necessários para a o entendimento das tarefas de mineração

de grafos.

5.2.1. Teoria dos Grafos

Um grafo $\mathscr{G} = (\mathscr{V},\mathscr{E})$ é definido como um conjunto de nós $\mathscr{V}$ e um conjunto de arestas $\mathscr{E}$, sendo que $|\mathscr{V}| = N$ denota o número de nós e $|\mathscr{E}| = M$ denota o número de arestas, sendo $e_k \in \mathscr{E}$ e $e_k = \{(v_i, v_j)|v_i, v_j \in \mathscr{V}\}$. Os termos nó ou vértice são considerados sinônimos. Neste trabalho será usado o termo nó para referenciar os elementos do conjunto de vértices $\mathscr{V}$ e similarmente o termo aresta para referenciar os elementos do conjunto de arestas $\mathscr{E}$, que também é referenciado na literatura por meio dos seguintes sinônimos: *links*, *hops*, ligações ou conexões.

Uma maneira conveniente de representar um grafo $\mathscr{G}$ em um computador é usar uma matriz de adjacência, que é uma matriz $\mathbf{A}$ quadrada $N \times N$, sendo $N = |\mathscr{V}|$, em que $\mathbf{A}_{i,j} = 1$ se $(v_i, v_j) \in \mathscr{E}$ e 0 caso o contrário.

Tabela 5.1. Símbolos utilizados neste trabalho.

Símbolo	Descrição
$\mathbf{A}$	Matriz de Adjacência
$\mathscr{G}$	grafo
$\mathscr{G}_s$	subgrafo
$\mathscr{E}$	arestas
$\mathscr{E}_s$	arestas do subgrafo
$\mathscr{E}_{sp}$	arestas no `ShatterPoint`
$\mathscr{V}$	nós
$\mathscr{V}_s$	nós do subgrafo
$\mathscr{V}_{sp}$	nós no `ShatterPoint`
$\mathscr{D}$	diâmetro do grafo
$\mathscr{D}_e$	diâmetro efetivo do grafo
λ	autovalor do grafo
GCC	maior componente conexa do grafo
v_i	nó de um grafo
e_k	aresta de um grafo
Δ	triângulo
$d(v_i)$	grau do nó v_i
$d_{out}(v_i)$	grau de saída do nó v_i
$d_{in}(v_i)$	grau de entrada do nó v_i
d_{max}	maior grau do grafo
$P(v,u)$	caminho do nó v ao u
C(v_i)	coeficiente de clusterização do nó v_i
C($\mathscr{G}$)	coeficiente de clusterização do grafo
κ_t	clique de tamanho t
N	número de nós
M	número de arestas

A Tabela 5.1 apresenta os principais símbolos utilizados e a seguir apresentam-

se alguns conceitos básicos, extraídos de [Nicoletti, 2006, Bondy and Murty, 1979, Diestel, 2005], que serão usados neste minicurso:

- **Grafos Direcionados e Não Direcionados:** um grafo é *não direcionado* se $\{(v_i, v_j) \in \mathscr{E} \Leftrightarrow (v_j, v_i) \in \mathscr{E}\}$, isto é, as aresta são pares de nós sem ordem. Se um par de nós é ordenado, isto é, arestas tem direção, então o grafo é *direcionado*, também chamado de *dígrafo*.

- **Grau do Nó:** o nó v_i tem grau $d(v_i)$ se ele tem $|\mathscr{N}(v_i)|$ nós incidentes. Para grafos direcionados, o grau de um nó pode ser dividido em "grau de saída", $d_{out}(v_i)$ que é o número de arestas entram pelo nó v_i e "grau de entrada", $d_{in}(v_i)$ que é o número de arestas que saem para o nó v_i.

- **Triângulo:** em um grafo não direcionado um triângulo (Δ), também conhecido como fechamento transitivo, é uma tripla de nós conexos (u, v, w), tal que, $(u, v), (v, w), (w, u) \in \mathscr{E}$

- **Caminho:** é uma sequência de nós conectados entre si, $P(v_1, v_n) = (v_1, v_2, v_3, ..., v_n)$, tal que, entre cada par de nó existe uma aresta $(v_1, v_2), (v_2, v_3), ..., (v_{n-1}, v_n) \in \mathscr{E}$. Um caminho é **simples** se nenhum nó se repete. Dois caminhos são **independentes** se somente o primeiro e o último nó são comuns à eles.

- **Comprimento de um caminho:** é o número de arestas que o caminho contém. O **menor caminho** entre dois nós $P(v_i, v_j)$ é o caminho de menor número de arestas que ligam os dois nós.

- **Subgrafo:** um subgrafo $\mathscr{G}_s = (\mathscr{V}_s, \mathscr{E}_s)$ de um grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ é um subconjunto de arestas e todos os nós tal que $\mathscr{E}_s \subseteq \mathscr{E} \Rightarrow \mathscr{V}_s = \{v_i, v_j | (v_i, v_j) \in \mathscr{E}_s\}$.

- **Grafo Conexo:** é um grafo que possui pelo menos um caminho entre todos os pares de nós.

- **Componente Conexa:** é o maior subgrafo, na qual existe um caminho entre qualquer par de aresta.

- **Grafo Induzido:** um subgrafo induzido $\mathscr{G}_s = (\mathscr{V}_s, \mathscr{E}_s)$ de um grafo $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ é um subconjunto de nós e todas as arestas que ligam este subconjunto de nós no grafo original $\mathscr{G}$, tal que $\mathscr{V}_s \subseteq \mathscr{V}$ e $\mathscr{E}_s = \{(v_i, v_j) | (v_i, v_j) \in \mathscr{E}_s \, v_i, v_j \in \mathscr{V}_s\}$.

- **Clique (κ_t):** é um subgrafo completo que possui um subconjunto de nós $\mathscr{V}_s \subseteq \mathscr{V}$ e arestas conectando todos os pares de nós em $\mathscr{V}_s$. O tamanho t do clique é definido pelo número de nós, $|\mathscr{V}_s| = t$. Um triângulo é um clique de tamanho 3 - κ_3.

- **Diâmetro:** o diâmetro $\mathscr{D}$ de um grafo $\mathscr{G}$ é o maior caminho dentre todos os menores caminhos existentes entre todos os pares de nós do grafo $\mathscr{G}$.

5.2.2. Leis de Potência

Uma distribuição que segue uma lei de potência é uma distribuição na forma:

$$p(x) = a * x^{-\gamma} \tag{1}$$

na qual $p(x)$ é a probabilidade de x ocorrer, sendo a uma constante de proporcionalidade e γ o expoente da lei de potência [Newman, 2005, Clauset et al., 2009].

Distribuições que seguem uma lei de potência ocorrem em muitas situações de interesse científico e são importantes para o entendimento de fenômenos naturais e humanos. A população das cidades e as intensidades dos terremotos são exemplos de fenômenos que têm a distribuição seguindo uma lei de potência.

5.2.3. Autovalores e autovetores

Em geral, uma matriz opera em um vetor transformando tanto a sua magnitude quanto a sua direção. Contudo, uma matriz pode operar em certos vetores transformando apenas a sua magnitude, deixando assim, a sua direção a mesma ou então transformando-a para o inverso. Estes vetores são chamados autovetores da matriz. A matriz opera em um autovetor pela multiplicação da sua magnitude por um fator, que se positivo sua direção não é alterada e se negativo sua direção é invertida. Este fator é chamado autovalor e está associado a um autovetor.

Essa transformação é chamada "transformação linear" e é formalmente definida como $A * x = \lambda * x$, sendo A a matriz de transformação linear, x o autovetor não nulo e λ o escalar. O escalar λ é considerado o autovalor de A correspondente ao autovetor x.

Autovalores e autovetores são conceitos importantes de matemática, com aplicações práticas em áreas diversificadas como mecânica quântica, processamento de imagens, análise de vibrações, mecânica dos sólidos, estatística, etc. Na seção 5.3.3 serão apresentados como os autovetores e autovalores podem ajudar na mineração de grafos.

5.3. Mineração de Grafos

Diversos domínios de aplicações têm seus dados modelados como redes complexas, por exemplo, a Internet, a World Wide Web (WWW), as redes sociais, de colaboração, biológicas, entre outras. Os pesquisadores nos últimos anos têm identificado classes de propriedades que podem ser encontradas em muitas das redes reais de vários domínios, sendo que muitas dessas distribuições seguem leis de potência, como a distribuição do grau dos nós, número de triângulos e os autovalores da matriz de adjacência da rede complexa.

5.3.1. Distribuição do grau do nó

Em redes complexas, é comum que a distribuição do grau dos nós siga uma lei de potência. Assim, a distribuição do grau dos nós de uma rede é uma lei de potência se o número de nós N_ϕ que possui um grau ϕ é dado por $N_\phi \propto \phi^{-y}$ $(y > 1)$, sendo $\mathscr{V}_\phi = \{v_i \in \mathscr{V} | d(v_i) = \phi\}$, $\left|\mathscr{V}_\phi\right| = N_\phi$ e y é chamado de expoente da distribuição do grau. A grande maioria das redes reais apresentam uma distribuição do grau dos nós que segue uma lei de potência, por isso são chamadas redes livres de escala.

Uma distribuição livre de escala, significa intuitivamente que uma distribuição se parece com ela mesma independente da escala em que se esta olhando. A noção de auto similaridade é implícita no nome "livre de escala". A auto similaridade de elementos consiste no fato do elemento manter as mesmas propriedades seja qual for a escala utilizada [Schroeder, 1991].

Este tipo de distribuição tem sido encontrada em grafos de ligações telefônicas [Abello et al., 1998], na Internet [Faloutsos et al., 1999], na Web [Kleinberg et al., 1999, Broder et al., 2000, Flaxman et al., 2005, Huberman and Adamic, 1999, Kumar et al., 1999], em grafos de citações [Redner, 1998], *click-stream* [Bi et al., 2001], em redes sociais online [Chakrabarti et al., 2004b] e muitas outras.

Tipicamente, para muitos conjuntos de dados, o expoente da distribuição do grau tem valor $2 < y < 3$. Por exemplo, a distribuição do grau de entrada da Web é $y_{in} = 2,1$ e de saída $y_{out} = 2,4$ [Reka and Barabási, 2002], enquanto para as redes de computadores chamadas de Sistemas Autônomos $y = 2,4$ [Faloutsos et al., 1999]. Contudo, alguns desvios do padrão da lei de potência foram notados em [Pennock et al., 2002].

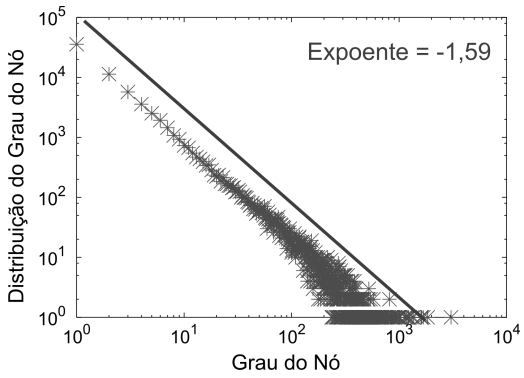


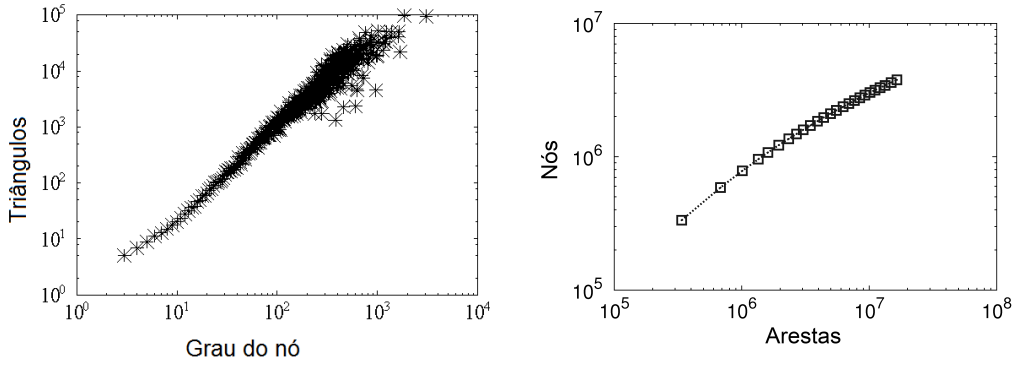
Figura 5.2. Gráfico da distribuição do grau dos nós da rede Epinions (quem confia em quem).

Além da distribuição do grau dos nós as seguintes distribuições tendem a seguir as leis de potência: número de triângulos em relação ao grau dos nós [Tsourakakis, 2008] e crescimento do número de nós e arestas na evolução das redes [Leskovec et al., 2007] entre outras. A Figura 5.3 apresenta os gráficos de algumas dessas distribuições. Um fato interessante sobre a distribuição dos triângulos em relação ao grau do nó é que a sua distribuição tem expoente oposto a distribuição do grau do nó.

5.3.2. Diâmetro efetivo

O diâmetro $\mathscr{D}$, como definido anteriormente, é o maior caminho dentre todos os menores caminhos existentes entre todos os pares de nós do grafo $\mathscr{G}$. O diâmetro também é referenciado como *diâmetro completo*. Para grafos com mais de uma componente conexa, o diâmetro usualmente é definido como infinito. Além disso, o diâmetro é suscetível aos efeitos degenerativos da estrutura do grafo como por exemplo, o surgimento de um caminho muito longo no grafo durante a sua evolução.

Calcular o diâmetro de um grafo grande é computacionalmente



(a) Δ vs. Grau do Nó (b) # Nós vs. # Arestas

Figura 5.3. Gráficos de outras distribuições que seguem lei de potência. Em (a) a distribuição dos triângulos versus o grau do nó da rede Epinions [Tsourakakis, 2008]. Em (b) a quantidade de nós versus a quantidade de arestas da rede Patente-US durante o seu crescimento [Leskovec et al., 2007].

caro(complexidade de tempo $O(N^3)$). Uma maneira mais eficiente de realizar este cálculo é pela amostragem de nós, isto é, uma quantidade de nós é amostrada e então o diâmetro é calculado entre os pares de nós amostrados [Albert et al., 1999]. Uma outra abordagem é usar o algoritmo de aproximação ANF [Palmer et al., 2002] que é baseado no cálculo aproximado chamado diâmetro efetivo. Define-se o diâmetro efetivo como sendo o menor número de "arestas" em que no mínimo 90% de todos os nós da maior componente conexa do grafo podem ser alcançados entre si.

O diâmetro efetivo é um valor mais robusto que o diâmetro, pois, nele somente os pares de nós conexos são considerados e a direção das arestas (no caso de grafos direcionados) são ignoradas. Além disso, muitos experimentos mostram que o diâmetro efetivo e o diâmetro exibem comportamento qualitativamente similar.

Em detalhes, seja $g_h(\mathscr{G})$ uma função que calcula para cada nó v_i o número de nós que tenham um caminho de máximo h arestas de distância partindo de cada nó v_i, sendo $i = 1, ..., N$. O valor de h inicia-se em h=1 até que $g_{(h-1)}(\mathscr{G})$-$g_h(\mathscr{G})$<$threshold$. Para h=1 tem-se $g_h(\mathscr{G})$= $\sum_{i=1}^{N} |d(v_i)|$ =$|\mathscr{E}|$. O gráfico que representa a função $g_h(\mathscr{G})$ é chamado "Hop-Plot" e é apresentado na Figura 5.4 [Leskovec et al., 2005]. A função $g_h(\mathscr{G})$ foi aplicada a uma rede complexa. O diâmetro efetivo desta rede é igual a $\mathscr{D}_e = 5$ sendo representado pela linha tracejada no gráfico apresentado na Figura 5.4.

Definindo mais formalmente o diâmetro efetivo tem-se: o diâmetro efetivo de $\mathscr{G}$ é definido como sendo $\mathscr{D}_e = h$ sendo que $g_h(\mathscr{G})$ representa 90% do número de pares de nós alcançáveis.

Muitos dos grafos reais exibem um diâmetro relativamente pequeno, conhecido como o fenômeno "Small-World"[Milgram, 1967]. Por exemplo, o diâmetro efetivo é pequeno para grandes redes reais, tais como a Internet, a Web, as redes sociais [Reka and Barabási, 2002, Barabasi and Albert, 1999, Broder et al., 2000, Bollobás and Riordan, 2004, Chung et al., 2002].

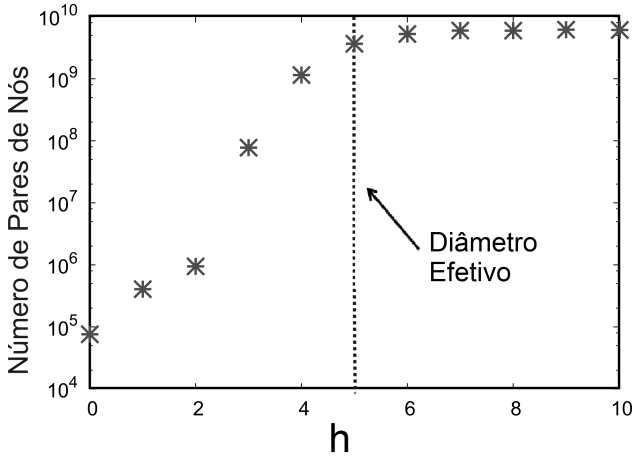


Figura 5.4. Gráfico Hop-Plot de uma rede complexa. A linha tracejada em 5 representa o Diâmetro efetivo da rede.

5.3.3. Autovalores e autovetores para grafos

Na teoria dos grafos, os autovalores λ_i de um grafo $\mathscr{G}$ são definidos como sendo os autovalores da matriz de adjacência A do grafo $\mathscr{G}$ ou a matriz Laplaciana B. Para A tem-se que $\mathbf{A}_{i,j} = 1$ se $(v_i, v_j) \in \mathscr{E}$ e 0 caso o contrário. Para B tem-se que $\mathbf{A}_{i,j} = d(v_i)$ se $i = j$, $\mathbf{A}_{i,j} = -1$ se $i \neq j \wedge (v_i, v_j) \in \mathscr{E}$ e 0 caso o contrário. Assim, como a matriz de adjacência, a matriz laplaciana pode ser usada para encontrar muitas propriedades dos grafos. As propriedades de ambas as matrizes fazem parta da teoria do *spectrum* do grafo. A seguir serão apresentadas algumas propriedades. Note-se que quando usado A será considerada a matriz de adjacência e quando usado B a matriz laplaciana.

Para todas as definições e teoremas que serão apresentados a seguir tem-se que o grafo $\mathscr{G}$ é não direcionado, com isso a matriz é simétrica. As provas dos teoremas e definições podem ser encontradas em [Mihail and Papadimitriou, 2002, Chung, 1994].

Seja A uma matriz de adjacência quadrada $N \times N$ com os seguintes autovalores λ_i, sendo $i = 1, 2, 3, ..., N$. Então:

Teorema 1 *Para um grafo $\mathscr{G}$ com $|\mathscr{V}| = N$ nós, tem-se que $\lambda_1(\mathscr{G}) \geq \lambda_2(\mathscr{G}) \geq \lambda_3(\mathscr{G}) \geq ... \ \lambda_N(\mathscr{G})$ são os autovalores da matriz de adjacência do grafo $\mathscr{G}$ em ordem decrescente.*

Assim, $\lambda_1(\mathscr{G})$ é o maior autovalor de $\mathscr{G}$, também chamado de principal autovalor.

Teorema 2 *Para um grafo $\mathscr{G}$ com N nós e $|\mathscr{E}| = N - 1$ tem se que $\lambda_1 \cong \sqrt{d_{max}}$.*

Teorema 3 *Um grafo $\mathscr{G}$ conexo é bipartido se $-\lambda_1$ também é um dos seus autovalores.*

Teorema 4 *Um grafo $\mathscr{G}$ a $\sum_{i=1}^{N} {\lambda_i}^2 = |\mathscr{E}|$.*

Seja B uma matriz de laplaciana quadrada $N \times N$ com os seguintes autovalores λ_i, sendo $i = 1, 2, 3, ..., N$. Então:

Teorema 5 *Se $\mathscr{G}$ é conexo então $\lambda_1 > 0$. Se $\lambda_i = 0$ e $\lambda_{i+1} \neq 0$, então $\mathscr{G}$ tem exatamente $i+1$ componentes conexas.*

O segundo autovetor da matriz laplaciana é largamente utilizado como método de bisseção de grafos para encontrar comunidades, como será apresentado na seção 5.3.5. Os autovetores e autovalores da matriz de adjacência de um grafo podem ser utilizados para diversas tarefas, como por exemplo na área de epidemiologia para indicar se a ocorrência de uma doença (modelada por uma rede complexa) está próxima ou não de uma epidemia, para isso basta comparar o número de morte do vírus versus o número de nascimento do vírus com $1/\lambda_1$ que é chamado de *Epidemic Threshold* [Chakrabarti et al., 2008]. Além disso, como será apresentado na seção 5.3.4 os autovalores também podem ser usados para calcular o número de triângulos aproximado de um grafo.

Um outro exemplo de aplicação de autovalores e autovetores é o algoritmo PageRank do Google [Page et al., 1998], que utiliza o primeiro autovetor da matriz de adjacência modificada de um grafo para fazer o *rank* de cada um dos nós. Considerando que o principal objetivo do PageRank é a ordenação de páginas Web para consultas, os nós do grafo são páginas e as aresta os *links* entre as páginas. Neste algoritmo também há um fator chamado *damping factor* - *df* - que é a probabilidade de um usuário encerrar a consulta e este valor é geralmente $0,85$. A Equação 2 representa o cálculo original do PageRank para todos os N nós de uma rede.

$$PR_{t+1} = \frac{1-df}{N} + df * A * PR_t \tag{2}$$

Neste algoritmo, a matriz de adjacência A é primeiro normalizada, isto é, cada elemento de uma linha da matriz é dividido pelo quantidade de "uns" existente na linha, sendo o somatório de cada linha igual a um. Na iteração inicial tem-se: $A(i,j) = \frac{1}{d_{out}(v_i)}$ se $\exists(v_i, v_j)$ senão $A(i,j) = 0$. O vetor PR é inicializado todo com 1.

5.3.4. Triângulos e coeficiente de clusterização

Em muitas redes, especialmente as redes sociais, é notado que se um nó u é conectado com um nó v que é conectado com w, então há uma grande probabilidade de u ser conectado com w. Esta relação é chamada de transitividade e é medida pelo coeficiente de clusterização [Watts and Strogatz, 1998]. A transitividade significa a presença de um alto número de triângulos ($\Delta(v_i)$) na rede. A contagem de triângulos é a principal parte do coeficiente de clusterização, que pode ser calculado para cada nó do grafo (Equação 3) ou para o grafo como um todo (Equação 4). Este coeficiente tem o objetivo de indicar quão próximo o grafo está de ser um grafo completo. O coeficiente de clusterização $C(v_i)$ de um nó v_i de grau $d(v_i)$ é definido pela Equação 3 a seguir.

$$C(v_i) = \frac{2 * \Delta(v_i)}{d(v_i) * (d(v_i) - 1)} \tag{3}$$

Seja v_i um nó com grau $|d(v_i)|$, então no máximo $d(v_i) * (d(v_i) - 1)/2$ arestas podem existir entre eles, sendo $\Delta(v_i)$ a fração de arestas que realmente existe, isto é o

número de triângulos. Isto significa que, o coeficiente de clusterização $C(v_i)$ de um nó v_i é a proporção de arestas entre os nós da sua adjacência dividido pelo número de arestas que podem existir entre eles. Equivalentemente, $C(v_i)$ é a fração de triângulos centrados no nó v_i entre $(d(v_i) * (d(v_i) - 1))/2$ triângulos que possam existir.

O coeficiente de clusterização global $C(\mathscr{G})$ é a média da soma de todos os $C(v_i)$ dos nós do grafo $\mathscr{G}$, dividido pelo número total de nós N. A equação do coeficiente de clusterização global é apresentada na Equação 4 a seguir.

$$C(\mathscr{G}) = \frac{1}{N} * \sum_{i=1}^{N} C(v_i) \tag{4}$$

Encontrar a quantidade de triângulos que cada nó possui, bem como a quantidade total de triângulos no grafo é um processo computacionalmente caro. A sua complexidade é de $O(N^2)$, sendo N o número total de nós do grafo.

Para reduzir essa complexidade, em alguns trabalhos como em [Latapy, 2008] os autores propõem algumas otimizações para a contagem e listagem dos triângulos em um grafo. Alguns trabalhos contam triângulos sem identificá-los [Bar-Yossef et al., 2002, Becchetti et al., 2008, Tsourakakis, 2008]. Sendo [Tsourakakis, 2008] o trabalho que apresenta uma melhor aproximação quanto ao número de triângulos e complexidade computacional. Neste trabalho é comprovado que o número total de triângulos é proporcional a soma dos autovalores da matriz de adjacência do grafo elevado ao cubo. A Equação 5 que representa esta proposição é apresentada a seguir:

$$\Delta(\mathscr{G}) = \frac{1}{6} * \sum_{i=1}^{N} \lambda^3{}_i \tag{5}$$

5.3.5. Detecção de Comunidades

Na sociedade há uma grande variedade de possíveis organizações em grupos: família, trabalho e circulo de amizade, cidades, nacionalidades. A difusão da internet tem direcionado a criação de grupos virtuais, que vivem na Web como as comunidades online Facebook, Orkut, LinkedIn. A detecção de comunidades pode ter diversas aplicações reais. Agrupar clientes web que tem interesses similares e são geograficamente próximos entre si pode aumentar a performance dos serviços oferecidos na WWW, em que cada grupo de cliente pode usar um servidor dedicado [Krishnamurthy and Wang, 2000]. Identificar grupos de clientes com interesses similares no relacionamento de compras entre produtos e consumidores de lojas onlines, como Amazon, permite o desenvolvimento de um sistema de recomendação eficiente, que ajuda a guiar consumidores nas lista de itens das lojas e permite uma oportunidade de negócio [Reddy et al., 2002].

Grupos de nós que tendem a ser mais conectados entre si que com o restante da rede são chamados de grupos ou comunidades. Pessoas tendem a formar comunidades, isto é, grupos pequenos no qual todo mundo conhece praticamente todo mundo. Com isto, grupos de nós pertencentes a uma mesma comunidade tendem a ter um número elevado de triângulos. Além disso, os membros das comunidades têm pouco relacionamento com

membros fora das comunidades que participam e cada um dos grupos tendem a estar organizados hierarquicamente, isto é comunidades dentro de comunidades.

Uma grande quantidade de algoritmos têm sido desenvolvidos para definir e identificar comunidades em redes sociais e de informações [Girvan and Newman, 2002, Radicchi et al., 2004]. Muitas vezes também é assumido que as comunidades obedecem a uma estrutura recursiva, em que grandes comunidades podem futuramente ser divididas em comunidades menores [Guimerà et al., 2007, Clauset et al., 2004].

A ideia do algoritmo de identificação de comunidades é particionar a rede em subgrafos menores por meio da remoção de um número mínimo de arestas. As comunidades encontradas podem ter seus elementos disjuntos, isto é, um nó que participa em um grupo não deve participar em outro, aceitando-se apenas algumas sobreposições [Fortunato, 2010]. Contudo, o contrário também pode ser encontrado, isto é, um mesmo nó faz parte de mais de uma comunidade. Esta situação pode ser vista na vida real, por exemplo: as pessoas nem sempre participam de apenas um grupo social. Boas revisões sobre detecção de comunidades podem ser encontradas em [Fortunato, 2010, Leskovec et al., 2008].

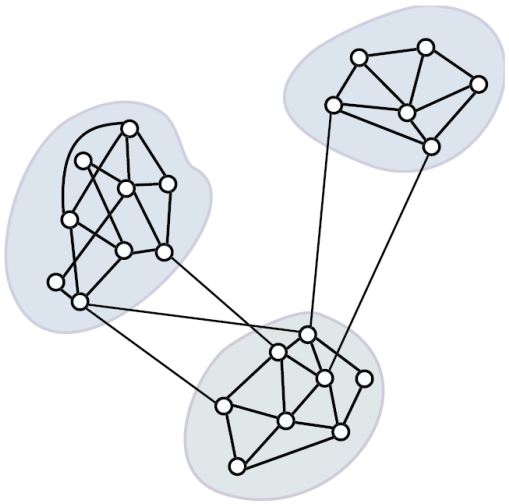


Figura 5.5. Grafo apresentando 3 comunidades do modo tradicional.

A estrutura de comunidades, como exemplificado pela Figura 5.5, foi observada também em diversas redes reais, como [Newman, 2006, Ravasz et al., 2002, Girvan and Newman, 2002]. Entretanto, esta forma de organização das redes complexas parece não valer para as redes complexas volumosas, com centenas de milhares de nós, como mostrado em [Leskovec et al., 2008], que demonstra que esta estrutura hierárquica está presente somente nas redes complexas pequenas, isto é redes com poucas centenas de nós.

As grandes redes complexas tendem a não possuir conjuntos de nós interligados formando comunidades muito bem definidas (Figura 5.5). Na verdade, as grandes redes tendem a apresentar uma estrutura chamada "Centro-Periferia" [Borgatti and Everett, 1999, Holme, 2005], que em ciência da computação é conhecida também pelo nome "*jellyfish*" [Tauro et al., 2001] ou "*octopus*" [Chung and Lu, 2006] e

que é exemplificada na Figura 5.6. Este conceito significa que uma rede é composta por um grande e denso conjunto de nós (core/centro) ligados entre si que basicamente não tem nenhuma estrutura de comunidade hierárquica, isto é, não podem ser quebrados em comunidades menores. Assim, a estrutura Centro-Periferia sugere o oposto da estrutura de comunidade hierárquica, e parece ser o mais encontrado em redes complexas de grande escala [Leskovec et al., 2008] e também em redes de computadores chamados Sistemas Autônomos [Siganos, 2006].

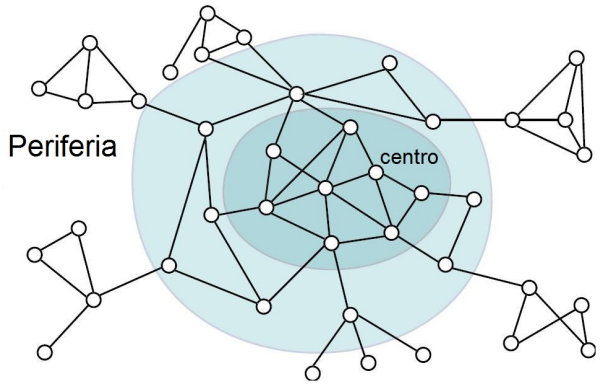


Figura 5.6. Grafo apresentando a topologia "Centro-Periferia"[Leskovec et al., 2008].

Também em [Leskovec et al., 2008] os autores mostram que as comunidades tendem a ser pequenas, com não mais do que 100 nós, e pouco conectadas com o restante da rede. O valor 100 é conhecido como o número de Dunbar, que é o número máximo de relacionamentos que uma pessoa consegue administrar [Dunbar, 1998].

A detecção de grupos (comunidades) não é importante apenas para redes sociais mas para inúmeras áreas da computação. Um exemplo, na computação paralela, é a determinação da melhor maneira de distribuir as tarefas para minimizar a comunicação entre os processadores, sendo as técnicas principais as baseadas no particionamento da rede. O problema de particionar um grafo consiste em dividir o conjunto de nós em k grupos de tamanho pré-definido, tal que, o número de arestas entre cada um dos grupos é mínimo. Cada grupo é chamado *cluster* ou comunidade e o número de arestas removidas é chamado *cut size*. Especificar o número de grupos em que o grafo será particionado é necessário, pois se este número fosse deixado livre a resposta trivial seria uma única partição com todos os nós. A Figure 5.7 ilustra o particionamento de um grafo em dois grupos ($k = 2$) cada um com 7 nós.

Um algoritmo muito tradicional para o particionamento de redes é o METIS, que permite o particionamento do grafo em k grupos. O algoritmo METIS funciona da seguinte maneira: dado um grafo $\mathscr{G}$, este é reduzido para um grafo com o agrupamento dos nós adjacentes em um mesmo nó. A bisseção deste grafo muito menor é computada e então o particionamento é projetado no grafo original por meio de refinamentos periódicos da partição [Karypis and Kumar, 1998]. A Figura 5.8 ilustra a aplicação do algoritmo METIS em um grafo. Note que apesar de na ilustração ser mostrado apenas o particionamento em dois, o METIS permite que o grafo seja particionado em k grupos, sendo k definido pelo usuário.

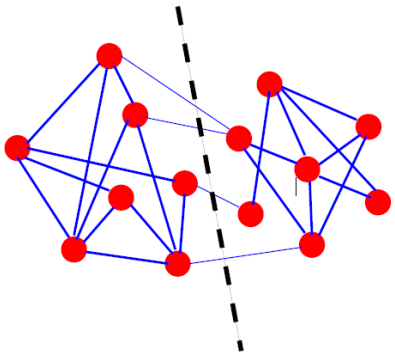


Figura 5.7. Um grafo sendo particionado em dois grupos iguais, cada um com 7 nós.

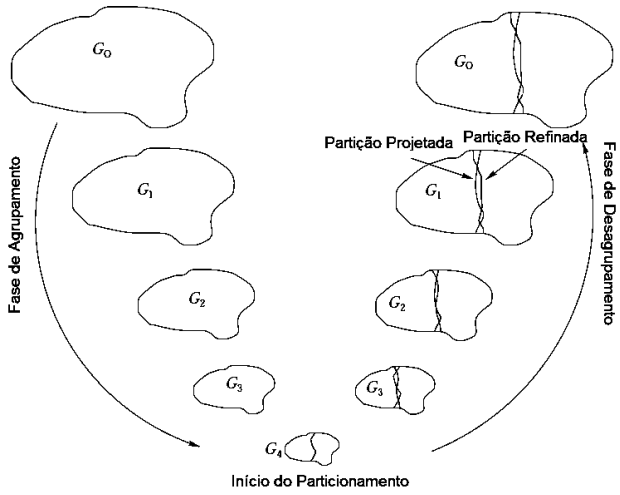


Figura 5.8. Exemplo de funcionamento do algoritmo METIS [Karypis and Kumar, 1998].

Outro método utilizados para o particionamento de grafos é o *Spectral Bisection* [Alon, 1998]. Este método é baseado no cálculo do segundo menor autovetor da matriz laplaciana da rede. Com este autovetor a matriz é reordenada e os grupos são identificados. Um exemplo deste método é apresentado na Figura 5.9. Primeiro a matriz de adjacência de um grafo $\mathscr{G}$ ao seu lado a matriz é reordenada pelo segundo menor autovetor da matriz laplaciana do grafo.

Um dos primeiros algoritmos realmente voltados para a identificação de comunidades em redes sociais foi o de Girvan e Newman [Girvan and Newman, 2002]. Neste algoritmo para cada aresta é calculado uma medida, chamada *betweenness*, que contabiliza o poder de "quebrar" de cada aresta. Esta medida calcula para cada aresta a quantidade de caminhos mínimos que utilizam esta aresta para interligar dois nós. Assim, quanto maior a quantidade de caminhos mínimos entre dois pares nós que contenham esta aresta, maior será o seu *betweenness*. Após o cálculo do *betweenness* para cada aresta, a aresta como o maior *betweenness* é removida da rede e então as arestas afetadas por esta remoção tem o seu *betweenness* recalculado. Este procedimento se repete até não haver mais arestas no grafo. O *betweenness* também pode ser calculado em relação a um nó da rede. O *betweenness* é uma medida muito cara computacionalmente já que é baseada no

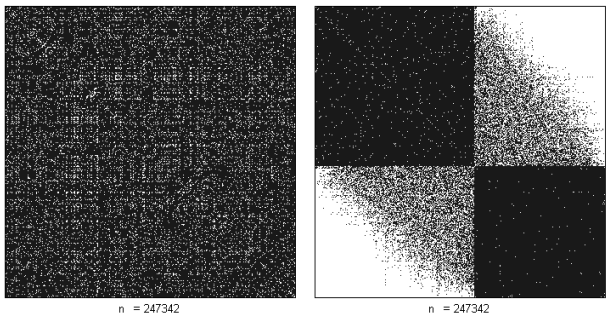


Figura 5.9. Matriz de adjacência de um grafo $\mathscr{G}$ e a matriz reordenada pelo segundo menor autovetor da matriz laplaciana do grafo $\mathscr{G}$

cálculo de caminhos mínimos o que a torna pouco aplicável a grande redes.

Os métodos supracitados não permitem a sobreposição de comunidades. A sobreposição é uma característica importante, principalmente nas redes sociais, já que, as pessoas naturalmente participam de mais de um grupo, como, escola, esportes, etc. Assim, um método bastante interessante que permite a sobreposição é o método Cross-Association [Chakrabarti et al., 2004a]. Este método faz uma decomposição conjunta da matriz de adjacência em grupos de linhas e colunas disjuntas, tal que intersecções retangulares são grupos homogêneos. O método é baseado em permutação de linhas e colunas utilizando-se do princípio MDL (Minimum Description Language). A ideia principal é que a matriz binária de uma rede representa a associação entre objetos (linhas e colunas) e quer se encontrar associações cruzadas entre esses objetos, isto é grupos homogêneos retangulares.

A Figura 5.10 apresenta a matriz de adjacência de duas redes reais Epinions e Oregon, processadas por este método. Quanto mais escura a área retangular mais denso é o grupo encontrado. Uma vantagem desse método é que o número de grupos é encontrado pelo método, além disso o número de grupos não precisa ser o mesmo para linhas (K) e colunas (l).

5.3.6. Resistência a ataques

Uma questão que tem sido considerada como foco de pesquisa recente é a definição de quão robusta é uma rede. Como visto anteriormente, uma rede pode representar o relacionamento interpessoal, as pessoas seriam os nós e o relacionamento as arestas. Imagine por exemplo que os nós da rede sejam marcados ou removidos conforme as pessoas contraiam uma doença ou tenham acesso a uma informação. Este tipo de análise é conhecido como resistência a ataques.

A Internet, por exemplo, é altamente robusta, já que há diferentes rotas ligando os computadores e/ou roteadores (nós), fazem com que a informação possua caminhos independentes para navegar de um nó para o outro. Assim, mesmo se um roteador falhar, o sistema é capaz de refazer a rota e fazer com que a informação chegue ao seu destino. De fato, na Internet sempre haverá um conjunto de roteadores que não estará funcionando em um dado momento. Entretanto, o importante é que a Internet como um todo continuará funcionando mesmo com uma grande quantidade de falhas [Wu et al., 2007].

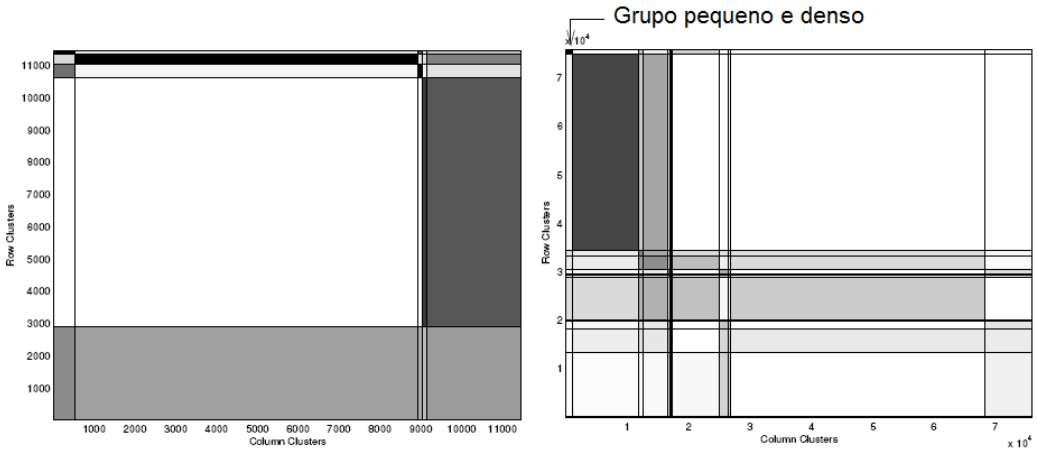
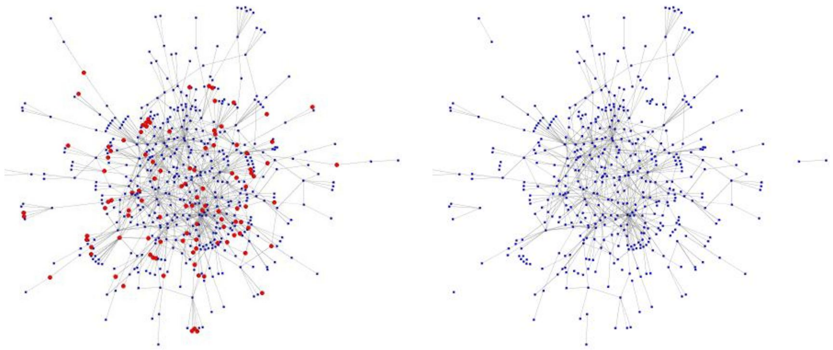


Figura 5.10. Duas redes reais, Epinions e Oregon, processadas pelo método Cross-Association [Chakrabarti et al., 2004a].



(a) (b)

Figura 5.11. Exemplo de uma rede complexa que tem 20% dos seus nós removidos aleatóriamente: (a) é a rede original e os nós marcados em vermelhos são os nós que serão removidos e (b) é a rede (a) já com os nós removidos.

Descobrir qual é o efeito da falha de uma certa fração de nós na conectividade do restante da rede tem sido o objetivo de inúmeros pesquisadores. Se ao invés da rede representar computadores ela representar pessoas, as falhas em uma rede podem representar a quantidade de pessoas que contraíram alguma doença, por exemplo, uma gripe.

Se os nós de uma rede são removidos de modo aleatório, como mostrado na Figura 5.11, o efeito é usualmente pequeno. Isto se deve ao fato de que as redes possuem um alto número de nós de grau ”um” e, em uma remoção aleatória, esses nós teriam uma alta probabilidade de serem removidos. A remoção de nós de grau um não altera a estrutura da rede já que eles se encontram na periferia da rede. A Figura 5.11 (b) apresenta a remoção aleatória de 20% dos nós de uma rede e como é notado, a rede continua robusta. Contudo, se os nós são removidos de um modo cuidadoso, isto é, segundo alguma ordenação que não seja aleatória, o dano causado na rede pode ser grande. Em [Albert et al., 2000, Cohen et al., 2000] a resistência das redes é analisada segundo a remoção aleatória, chamada de **falha** e a remoção baseada na ordenação dos nós pelo seu

grau, chamada de **ataque**.

As redes livres de escala são muito resistentes a falhas (remoção aleatória de nós), mas elas são substancialmente menos robusta quanto a um ataque. Isto acontece por que durante as falhas, uma grande quantidade de nós removidos são nós de grau um, já que estes são maioria nas redes reais. Já os ataques que removem os nós de alto grau, fazem com que a rede se torne desconexa muito rapidamente. Um exemplo do comportamento de uma rede real é ilustrado na Figura 5.12. Neste exemplo o diâmetro é avaliado quanto a quantidade de aresta após a deleção de nós. Note-se que, na remoção dos nós de alto grau o diâmetro cresce rapidamente e depois decai, mostrando que a rede torna-se desconexa muito rapidamente. Já na deleção de nós de baixo grau o diâmetro apresenta uma certa constância no seu valor.

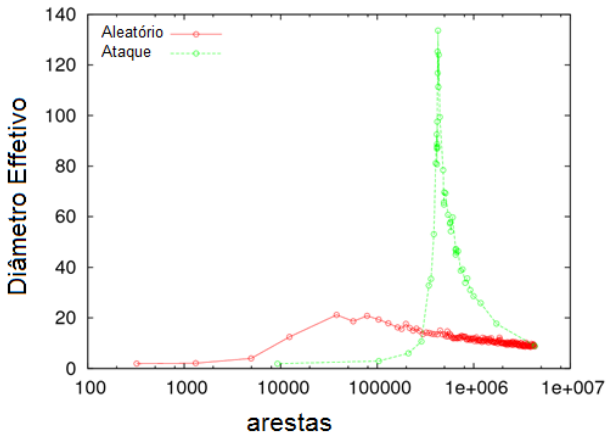


Figura 5.12. Gráfico mostrando o comportamento de uma rede real quanto a remoção de nós, ataque e falha.

As falhas e ataques às redes são geralmente analisadas quanto ao diâmetro e ao tamanho (número de nós) da maior componente conexa. O diâmetro aumenta conforme mais nós são removidos, mas esse aumento é mais lento nas redes livres de escala do que nas randômicas. Similarmente, o tamanho da maior componente conexa (GCC) decai mais devagar nas redes livres de escala do que nas randômicas. Este comportamento é ilustrado na Figura 5.13. As circunferências representam o tamanho da rede quanto as suas componentes conexas, as circunferências menores representam as componentes conexas menores.

Como ilustrado, as redes randômicas apresentam o mesmo comportamento quanto a falhas e ataques. Isso acontece pois as redes randômicas não tem a distribuição de grau seguindo uma lei de potência, na verdade, os nós têm o grau muito próximo a média do grau dos nós da rede.

O algoritmo `ShatterPlots` [Appel et al., 2009] encontra padrões em redes complexas, reais e sintéticas, por meio da remoção de arestas. Intuitivamente, dada uma rede $\mathscr{G}$ a cada passo t, Q_t arestas são escolhidas aleatoriamente para serem removidas da rede. Após cada remoção um conjunto de medidas, como diâmetro efetivo, autovalor da matriz de adjacência, número de nós e arestas são coletados da rede e o processo continua até todos os nós estarem isolados.

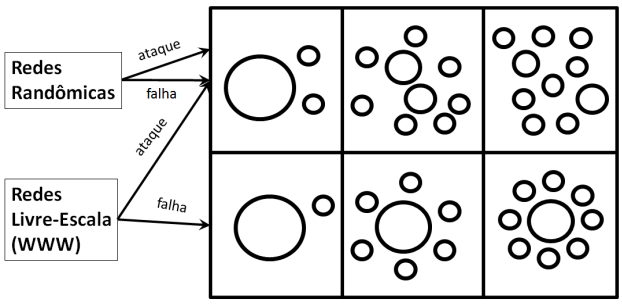


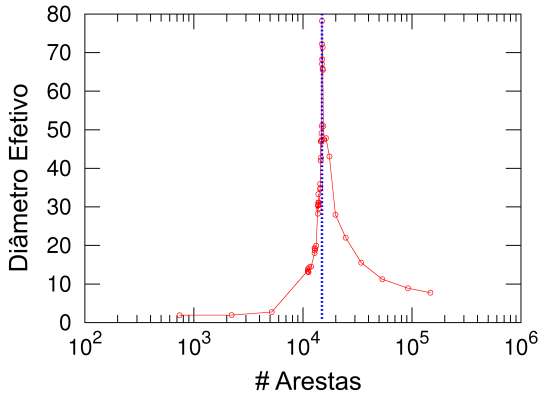
Figura 5.13. Comparação do comportamento do tamanho da rede quanto a remoção de nós aleatória e nós de alto grau. [Albert et al., 2000]

A Figura 5.14 apresenta medidas diâmetro efetivo, número de pares alcançáveis e número de nós pertencentes à maior componente conexa (GCC) na rede real Gnutella [Ripeanu et al., 2002] durante a remoção aleatória de arestas pelo algoritmo `ShatterPlots`. Todas as medidas são em relação a quantidade de arestas existente na rede. Como pode ser observado, apenas o diâmetro mostra um pico, chamado `ShatterPoint` e marcado por uma linha, nos gráficos apresentados na Figura 5.14. As outras medidas também possuem uma fase de transição, que é a mesma do diâmetro e que também esta marcada com uma linha azul, entretanto, elas não podem ser identificadas de maneira clara nos gráficos.

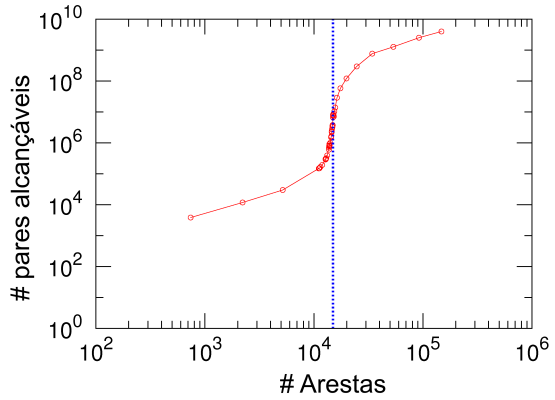
O algoritmo `ShatterPlots` é um método adaptativo, pois ele ajusta o número de arestas removidas na rede. Este ajuste faz com que o número de arestas removidas diminua se o diâmetro aumentar mais que um certo *threshold* e aumente se o diâmetro se manteve estável. Isto torna constante o número de iterações do algoritmo, fazendo com que ele seja escalável quanto ao número de arestas.

Dentre os padrões encontrados pelo `ShatterPlots` no `ShatterPoint` dois se destacam. O primeiro é o chamado *30-per-cent*. Este padrão revela que para todas as redes, sintética e reais, o `ShatterPoint` ocorre sempre na mesma proporção de nós e aresta, isto é, ele ocorre quanto o número de nós não isolados da rede é 30% maior que o número de arestas atual da rede ($\mathscr{V}_{sp} = 1.30 * \mathscr{E}_{sp}$). Este valor é conhecido na teoria dos grafos randômicos, *Erdös-Rényi*, como fase de transição ou *percolation*. Esta fase é na teoria conhecida como o ponto em que uma rede passa de desconexa para conexa, isto é, abaixo do ponto de transição a rede possui pequenas componentes conexas e acima deste ponto a rede possui uma grande componente conexa. Quanto mais acima deste ponto uma rede estiver mais bem conexa ela será. A Figura 5.15 apresenta o gráfico que mostra o padrão *30-per-cent*. Neste gráfico, os triângulos representam as redes sintéticas, entre elas *Preferential Attachment*, *Small-World* e *Erdös-Rényi*. Os demais símbolos representam as redes reais, dentre elas, Gnutell, Amazon, Oregon, etc.

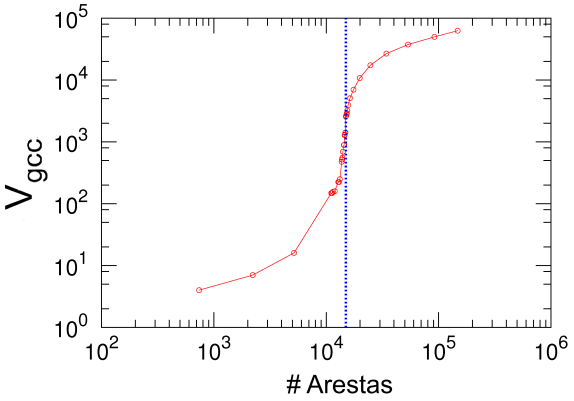
O segundo é o padrão chamado *NodeShatteringRatio* . Este padrão revela que em um gráfico com o número atual de nós da rede ($|V_{sp}|$) versus o número original de nós da rede ($|V|$) é possível traçar uma linha ($\mathscr{V}_{sp} = 0.37 * \mathscr{V}$) em que, acima desta linha todas as redes são sintéticas e abaixo dela todas são reais. A Figura 5.16 apresenta o gráfico do padrão *NodeShatteringRatio* . Mais detalhes podem ser encontrados em [Appel et al., 2009].



(a) Diâmetro Efetivo



(b) # pares alcançáveis



(c) # nós da GCC

Figura 5.14. Medidas estruturais da rede feitas durante a remoção aleatória de arestas. A rede apresenta `ShatterPoint` em todas as medidas mas somente o diâmetro apresenta um pico.

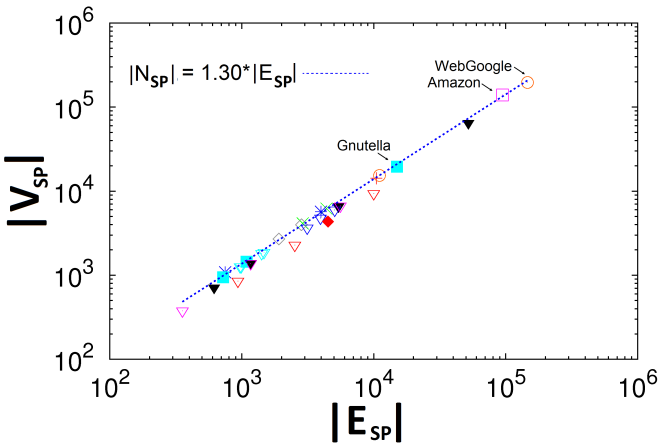


Figura 5.15. O padrão *30-per-cent* para redes reais e sintéticas (triângulos).

5.3.7. Predição de Ligações

A predição de ligações pode ser definida como, dado um "*snapshot*" de uma rede complexa em um tempo t, quer se prever com uma certa acurácia as arestas que irão surgir na rede complexa no tempo futuro $t+1$.

Dentre as técnicas de predição de ligação destacam-se as baseadas em propriedades estruturais do grafo [Liben-Nowell and Kleinberg, 2003, Huang, 2006]. Um exemplo interessante é apresentado em [Clauset et al., 2008], em que a descoberta de grupos, isto é comunidades, em redes complexas é usado para o auxílio na identificação de ligações faltantes, já que pares de nós pertencentes a uma mesma comunidade têm mais chance de serem conexos entre si do que pares de nós pertencentes a comunidades diferentes. Este método se diferencia da predição de ligação tradicional, pois normalmente esta tarefa visa descobrir arestas que virão a existir na rede complexa quando esta evoluir (crescimento do número de nós e arestas com o passar do tempo) e não uma aresta perdida na construção da rede complexa. Entretanto, este método não funciona para todos os tipos de redes complexas, já que o método não consegue detectar comunidades em redes complexas que não possuam grupos bem definidos. Há uma coleção de trabalhos que vem usando e desenvolvendo algoritmos na área de predição de ligações [Kashima et al., 2009, Hasan et al., 2006, Kunegis and Lommatzsch, 2009, Lu and Zhou, 2009, Acar et al., 2009].

Uma das dificuldades da predição de ligações é que as redes complexas tendem a ser esparsas. Para driblar esta dificuldade, outros modelos fazem uso não só de propriedades estruturais do grafo mas também de características relacionais baseadas nos atributos dos nós do grafo. Esta abordagem é mais conhecida na área de aprendizado relacional ou aprendizado multi-relacional, que tem por objetivo não só o uso da estrutura dos grafos, mas também a descrição dos mesmos por meio de uma base de dados relacional ou lógica relacional ou de primeira ordem. Assim, além das ligações entre as tuplas formando um grafo, também há características, isto é, informações, relacionadas aos nós do grafo [Getoor and Diehl, 2005, Hasan et al., 2006, Taskar et al., 2004, Popescul et al., 2003]. Entretanto, esta informação complementar dos

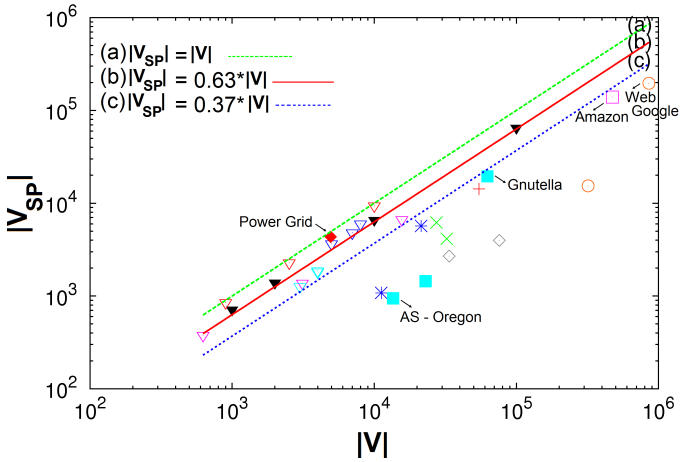


Figura 5.16. O padrão *NodeShatteringRatio* para redes reais e sintéticas (triângulos). A linha (a) ($\mathscr{V}_{sp} = \mathscr{V}$) mostra o valor de nós atual da rede igual ao valor original. A linha (b) mostra o valor de ($\mathscr{V}_{sp} = 0.37 * \mathscr{V}$) nós atual igual a 63% do valor de nós original, esta é conhecida como a quantidade de nós na fase de transição dos grafos *Erdös-Rényi* representados pelos triângulos pretos. A linha (c) ($\mathscr{V}_{sp} = 0.37 * \mathscr{V}$) o número de nós atual é 37% do número de nós total da rede. Isto comprova que as redes reais são bem resistentes e que a fase de transição esta bem distante.

nós e arestas nem sempre estão disponíveis, o que inviabiliza a aplicação desses algoritmos nesses casos.

5.4. Conclusão

Este documento apresentou uma visão geral da área de mineração de grafos e redes complexas. Esta área tem se mostrado muito importante atualmente, principalmente pelo grande crescimento de domínios de aplicação, nos quais os dados podem ser modelados através de redes complexas. Nestas aplicações, quando a modelagem é feita através técnicas tradicionais de representação e algoritmos tradicionais de mineração de dados (tais como regras de associação, detecção de agrupamentos, classificação, entre outros) são aplicados, os resultados tendem a não capturar todos os padrões relevantes (e presentes nos dados). Já as técnicas de mineração de grafos aplicadas a redes complexas podem trazer ganho substancial nos resultados da mineração. O documento apresentou ainda os principais algoritmos de mineração de grafos e propriedades estatísticas utilizados na área de redes complexas. Definições da teoria dos grafos necessárias para o entendimento desses algoritmos e propriedades básicas também foram abordadas. As redes complexas estão cada vez mais presentes nos sistemas computacionais e com isso o seu entendimento torna-se cada vez mais importante e relevante tanto para pesquisadores da área da computação quanto para de áreas em que problemas reais podem ser representadas através destes modelos.

Referências

[Abello et al., 1998] Abello, J., Buchsbaum, A. L., and Westbrook, J. R. (1998). A functional approach to external graph algorithms. In *Algorithmica*, pages 332–343.

Springer-Verlag.

[Acar et al., 2009] Acar, E., Dunlavy, D. M., and Kolda, T. G. (2009). Link prediction on evolving data using matrix and tensor factorizations. In Saygin, Y., Yu, J. X., Kargupta, H., Wang, W., Ranka, S., Yu, P. S., and Wu, X., editors, *ICDM Workshops*, pages 262–269. IEEE Computer Society.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabasi, A.-L. (1999). The diameter of the world wide web.

[Albert et al., 2000] Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378–381.

[Alon, 1998] Alon, N. (1998). Spectral techniques in graph algorithms. In Lucchesi, C. L. and Moura, A. V., editors, *Lecture Notes in Computer Science 1380*, pages 206–215. Springer-Verlag, Berlin.

[Appel et al., 2009] Appel, A. P., Chakrabarti, D., Faloutsos, C., Kumar, R., Leskove, J., and Tomkins, A. (2009). Shatterplots: a fast tool for mining large graphs. In *SIAM SDM*, pages 802–813. SIAM.

[Bar-Yossef et al., 2002] Bar-Yossef, Z., Kumar, R., and Sivakumar, D. (2002). Reductions in streaming algorithms, with an application to counting triangles in graphs. In *SODA*.

[Barabasi and Albert, 1999] Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Becchetti et al., 2008] Becchetti, L., Boldi, P., Castillo, C., and Gionis, A. (2008). Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *KDD*, pages 16–24.

[Bi et al., 2001] Bi, Z., Faloutsos, C., and Korn, F. (2001). The "DGX"distribution for mining massive, skewed data. *KDD*.

[Bollobás and Riordan, 2004] Bollobás, B. and Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.

[Bondy and Murty, 1979] Bondy, J. A. and Murty, U. S. R. (1979). *Graph Theory with applications*. Elsevier Science Publishing Co., Inc.

[Borgatti and Everett, 1999] Borgatti, S. P. and Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, 21.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web: experiments and models. In *Proceedings of the Ninth International World-Wide Web Conference (WWW9, Amsterdam, May 15 - 19, 2000 - Best Paper)*. Foretec Seminars, Inc. (of CD-ROM), Reston, VA.

[Chakrabarti et al., 2004a] Chakrabarti, D., Papadimitriou, S., Modha, D. S., and Faloutsos, C. (2004a). Fully automatic cross-associations. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 79–88. ACM Press.

[Chakrabarti et al., 2008] Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., and Faloutsos, C. (2008). Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1–26.

[Chakrabarti et al., 2004b] Chakrabarti, D., Zhan, Y., and Faloutsos, C. (2004b). R-mat: A recursive model for graph mining. In *Fourth SIAM International Conference on Data Mining*.

[Chung et al., 2002] Chung, F., Chung, F., Chung, F., Lu, L., and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Internet Mathematics*, 1:15879–15882.

[Chung and Lu, 2006] Chung, F. and Lu, L. (2006). *Complex Graphs and Networks*. American Mathematical Society.

[Chung, 1994] Chung, F. R. K. (1994). *Spectral Graph Theory*.

[Clauset et al., 2008] Clauset, A., Moore, C., and Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

[Clauset et al., 2004] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70:066111.

[Clauset et al., 2007] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2007). Power-law distributions in empirical data.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–704.

[Cohen et al., 2000] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85(21):4626–4628.

[Diestel, 2005] Diestel, R. (2005). *Graph Theory*. Springer-Verlag Heidelbergh.

[Dunbar, 1998] Dunbar, R. (1998). *Grooming, Gossip, and the Evolution of Language*. Harvard Univ Press.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM 1999*, volume 1, pages 251–262, Cambridge, Massachusetts. ACM Press.

[Flaxman et al., 2005] Flaxman, A., Frieze, A., and Fenner, T. (2005). High degree vertices and eigenvalues in the preferential attachment graph. *Internet Math.*, 2(1):1–19.

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[Getoor and Diehl, 2005] Getoor, L. and Diehl, C. P. (2005). Introduction to the special issue on link mining. *SIGKDD Explor. Newsl.*, 7(2):1–2.

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. In *Proc. Natl. Acad. Sci. USA*, volume 99.

[Guimerà et al., 2007] Guimerà, R., Sales-Pardo, M., and Amaral, L. A. (2007). Module identification in bipartite and directed networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76(3 Pt 2).

[Hasan et al., 2006] Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

[Holme, 2005] Holme, P. (2005). Core-periphery organization of complex networks. *Phys. Rev. E*, 72(4):046111.

[Huang, 2006] Huang, Z. (2006). Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In *Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (LinkKDD2006)*.

[Huberman and Adamic, 1999] Huberman, B. A. and Adamic, L. A. (1999). Internet: Growth dynamics of the world-wide web. *Nature*, 401(6749):131.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407.

[Karypis and Kumar, 1998] Karypis, G. and Kumar, V. (1998). Multilevel k-way partitioning scheme for irregular graphs. *J. Parallel Distrib. Comput.*, 48:96–129.

[Kashima et al., 2009] Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., and Tsuda, K. (2009). Link propagation: A fast semi-supervised learning algorithm for link prediction. In *SDM*, pages 1099–1110. SIAM.

[Kleinberg et al., 1999] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods.

[Krishnamurthy and Wang, 2000] Krishnamurthy, B. and Wang, J. (2000). On network-aware clustering of web clients. *SIGCOMM Comput. Commun. Rev.*, 30(4):97–110.

[Kumar et al., 1999] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. *Comput. Networks*, 31(11-16):1481–1493.

[Kunegis and Lommatzsch, 2009] Kunegis, J. and Lommatzsch, A. (2009). Learning spectral graph transformations for link prediction. In *Proc. Int. Conf. in Machine Learning*.

[Latapy, 2008] Latapy, M. (2008). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theor. Comput. Sci.*, 407(1-3):458–473.

[Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *eleventh ACM SIGKDD*, pages 177–187, New York, NY, USA. ACM Press.

[Leskovec et al., 2007] Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1):1 – 40.

[Leskovec et al., 2008] Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *CoRR*, abs/0810.1355.

[Liben-Nowell and Kleinberg, 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA. ACM.

[Lu and Zhou, 2009] Lu, L. and Zhou, T. (2009). Role of weak ties in link prediction of complex networks. In *Proceedings of the 1st ACM International Workshop on Complex Networks in Information and Knowledge Management (CNIKM)*, Hong Kong, China.

[McGlohon et al., 2008] McGlohon, M., Akoglu, L., and Faloutsos, C. (2008). Weighted graphs and disconnected components: patterns and a generator. In *KDD*, pages 524–532.

[Mihail and Papadimitriou, 2002] Mihail, M. and Papadimitriou, C. H. (2002). On the eigenvalue power law. In *RANDOM '02: Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, pages 254–262, London, UK. Springer-Verlag.

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

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

[Newman, 2005] Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323.

[Newman, 2006] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[Nicoletti, 2006] Nicoletti, Maria Do Carmo ; Hruschka Jr., E. (2006). *Fundamentos da Teoria dos Grafos.*, volume 1. EdUFSCar - Editora da Universidade Federal de São Carlos, 1. ed. revisada edition.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library.

[Palmer et al., 2002] Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002). Anf: A fast and scalable tool for data mining in massive graphs. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 1, pages 81–90, Edmonton, Alberta, Canada. ACM Press.

[Pennock et al., 2002] Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211.

[Popescul et al., 2003] Popescul, A., Popescul, R., and Ungar, L. H. (2003). Statistical relational learning for link prediction.

[Radicchi et al., 2004] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663.

[Ravasz et al., 2002] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

[Reddy et al., 2002] Reddy, P. K., Kitsuregawa, M., Sreekanth, P., and 0002, S. S. R. (2002). A graph based approach to extract a neighborhood customer community for collaborative filtering. In *DNIS*, pages 188–200.

[Redner, 1998] Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution.

[Reka and Barabási, 2002] Reka, A. and Barabási (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

[Ripeanu et al., 2002] Ripeanu, M., Foster, I., and Iamnitchi, A. (2002). Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1).

[Schroeder, 1991] Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman and Company, New York.

[Siganos, 2006] Siganos, Georgos; Sudhir L Tauro, M. F. (2006). Jellyfish: A conceptual model for the as internet topology. In *Journal of Communications and Networks*, volume 8, pages 339 – 350.

[Taskar et al., 2004] Taskar, B., Wong, M., Abbeel, P., and Koller, D. (2004). Link prediction in relational data.

[Tauro et al., 2001] Tauro, S. L., Palmer, C., Siganos, G., and Faloutsos, M. (2001). A simple conceptual model for the internet topology. In *Global Internet, San Antonio, Texas*.

[Tsourakakis, 2008] Tsourakakis, C. E. (2008). Fast counting of triangles in large real networks without counting: Algorithms and laws. In *ICDM '08*, pages 608–617, Washington, DC, USA. IEEE Computer Society.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

[Wu et al., 2007] Wu, J., Zhang, Y., Mao, Z. M., and Shin, K. G. (2007). Internet routing resilience to failures: analysis and implications. In *CoNEXT '07: Proceedings of the 2007 ACM CoNEXT conference*, pages 1–12, New York, NY, USA. ACM.