Capítulo 5

Por dentro das redes complexas - detectando grupos e prevendo ligações

Ana Paula Appel[1], Estevam Rafael Hruschka Junior[2]

[1]Departamento de Engenharias e Computação – DECOM
Universidade Federal do Espírito Santo - UFES
Centro Universitário Norte do Espírito Santo - CEUNES
{anaappel@ceunes.ufes.br}
[2]Departamento de Computação - DC
Universidade Federal de São Carlos - UFSCar
{estevam@dc.ufscar.br}

Abstract

The growing volume of data modeled as complex networks, eg the World Wide Web, social networks like Orkut, Facebook, has raised a new area of research - mining complex networks. In this new multidisciplinary area some tasks can be highlighted: the extraction of statistical properties, community detection and link prediction. This short course aims to introduce not only the basics of mining complex networks, but also the techniques of community detection and link prediction. The mining of complex networks have not only been the focus of a large number of researchers but also big companies like Microsoft, Google, Yahoo and Facebook.

Resumo

O crescimento do volume de dados modelados como redes complexas, por exemplo a World Wide Web, redes sociais como Orkut, Facebook, fez surgir uma nova área de pesquisa - a mineração de redes complexas. Nesta nova área multidisciplinar destacam- se algumas tarefas: a extração de propriedades estatísticas, a detecção de comunidades, a predição de ligações arestas. Este minicurso tem como objetivo introduzir não só os conceitos básicos da mineração de redes complexas, mas também as técnicas de detecção

de comunidade e predição de ligação. A área de mineração de redes complexas têm sido o foco não só de um grande número de pesquisadores mas também de grandes empresas como Microsoft, Google, Facebook e Yahoo.

5.1. Introdução

De uma maneira simplista, uma rede é uma coleção de pontos conectados por pares de linhas. Dependendo da área de utilização os pontos recebem nomes específicos, como nós ou vértices e as linhas são referenciadas como arestas. Muitos objetos de interesses têm sido mapeados como redes, por exemplo, as redes sociais (Facebook, Flickr, Orkut, etc), redes biológicas, redes de computadores, ontologias, redes acadêmicas (DBLP, ARXIV) e a própria WWW - World Wide Web.

Tais redes, chamadas redes complexas são, normalmente, modelada como um grafo, ou seja, a rede complexa é representada por meio 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. Este tipo de representação é conveniente em muitas situações, já que ela abstrai o problema para uma representação onde apenas a conexão entre os objetos é levada em consideração eliminando muitas vezes a complexidade e informações desnecessárias. Informações adicionais podem ser inseridas, como nomes e pesos aos nós e arestas [Newman, 2010].

Nos últimos anos as redes complexas têm sido o foco de estudo de muitos cientistas, os quais vêm desenvolvendo uma ampla coleção de ferramentas para modelar, analisar e entender as redes complexas. Estas ferramentas são em sua maioria matemáticas, estatísticas e computacionais e muitas vezes partem de uma simples representação da rede com o conjunto de nós e arestas e depois de alguns cálculos produzem informações interessantes sobre ela, como por exemplo o caminho médio entre pares de nós, quantidade de triângulos, etc [Faloutsos et al., 1999, Leskovec et al., 2007, Tsourakakis, 2008].

A ciência moderna das redes complexas tem trazido um avanço significante no entendimento de sistemas complexos. Uma das mais relevantes características das redes que representam sistemas reais é a estrutura de comunidade, ou "clustering", isto é, a organização de vértices em grupos, com muitas arestas unindo vértices do mesmo grupo e comparativamente poucas arestas unindo vértices de grupos diferentes. Tais grupos, ou comunidades, podem ser consideradas partes independentes das redes, desempenhando um papel semelhante aos tecidos ou órgãos do corpo humano. Detectar comunidades é de grande importância em áreas como a sociologia, biologia e ciência da computação, disciplinas nas quais frequentemente os sistemas são representados como redes complexas. Contudo, encontrar comunidades em redes complexas é um problema de difícil solução e ainda não possui uma resposta satisfatória, apesar do grande esforço que tem sido feito pela comunidade científica nos últimos anos [Fortunato, 2010].

Outro problema bastante comum na área de redes complexas é a predição de ligação. A predição de ligação tem a sua fonte de inspiração as redes sociais, na qual quer se prever com uma acurácia cada vez maior quem são os seus possíveis novos amigos de cada um dos membros das redes sociais [Liu and Wong, 2008]. A tarefa de predição de ligações (*Link Prediction*) tem por objetivo prever quais arestas irão surgir em uma rede complexa em um futuro próximo [Liben-Nowell and Kleinberg, 2003]. De modo mais

formal, 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 $t+1$. Uma das dificuldades da predição de ligações é que as redes complexas tendem a ser esparsas.

Com o intuito de guiar o leitor se tentará uma exposição minuciosa do tema neste documento, a partir da definição dos principais elementos do problema, para a apresentação da maioria dos métodos desenvolvidos, com um foco especial em técnicas mais recentes, a partir da discussão de questões cruciais como a importância do métodos de agrupamento e de predição de ligação e as principais diferenças entre os diferentes métodos de cada técnica. Além disso, sempre que possível será dada uma descrição de aplicações para as redes complexas reais.

5.2. Conceitos

As redes complexas, ou simplesmente redes, são um conjunto de elementos discretos que são representados por vértices e arestas, sendo que as arestas 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.

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 detecção de comunidades e predição de ligação.

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 $e_k = \left\{(v_i, v_j) | v_i, v_j \in \mathscr{V}\right\}$. 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.

Seja S uma comunidade na qual nS é o número de nós em S e $nS = |S|$; mS o número de aresta em S e $ms = |(u,v) : u \in S, v \in S|$ e cS o número de arestas na fronteira de S, sendo $cS = |(u,v) : u \in S, v \notin S|$

Tabela 5.1. Símbolos utilizados neste trabalho.

Símbolo	Descrição
$\mathbf{A}$	Matriz de Adjacência
$\mathscr{G}$	grafo
$\mathscr{E}$	arestas
$\mathscr{V}$	nós
λ	autovalor do grafo
v_i	nó de um grafo
e_k	aresta de um grafo
Δ	triângulo
$d(v_i)$	grau do nó v_i
$\Gamma(v_i)$	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
$\mathrm{C}(v_i)$	coeficiente de clusterização do nó v_i
$\mathrm{C}(\mathscr{G})$	coeficiente de clusterização do grafo
$score(u;w)$	valor entre dois nós
κ_t	clique de tamanho t
k	número de nós em uma comunidade
S	comunidade
cS	número de arestas na fronteira de S
N	número de nós
nS	número de nós em S
M	número de arestas
mS	número de arestas em S

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 $\left\{(v_i,v_j) \in \mathscr{E} \Leftrightarrow (v_j,v_i) \in \mathscr{E}\right\}$, 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 $d(v_i)$ também representado por $\Gamma(v_i)$ 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 no 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 vértice.

- **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 $\mathscr{G}$.

5.3. 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 propriedades são distribuições que seguem leis de potência, como a distribuição do grau dos nós, número de triângulos e os autovalores da matriz de adjacência da rede complexa.

Grafos que modelam redes que representam sistemas reais não são regulares como os reticulados, na qual todos os vértices possuem o mesmo grau, nem randômicas, na qual são redes em que cada aresta possui uma probabilidade existir igual para todos os pares de vértices possíveis fazendo com que a distribuição de arestas entre os vértices seja altamente homogênea. A Figura 5.1 apresenta duas redes reais com distribuições diferentes. A primeira é uma randômica que representa algumas conexões entre as rodovias do EUA (à direita da Figura 5.1) e a segunda uma rede que segue uma distribuição de potência que representa as conexões entre os aeroportos dos EUA (à esquerda da Figura 5.1).

Como visto na Figura 5.1, as redes reais que seguem uma lei de potência exibem uma grandes inomogeneidade, a distribuição do grau dos nós é ampla, com uma cauda que na maior parte das vezes segue uma lei de potência: por isso, muitos vértices de baixo grau coexistem com alguns vértices de alto grau. Uma lei de potência, que é uma

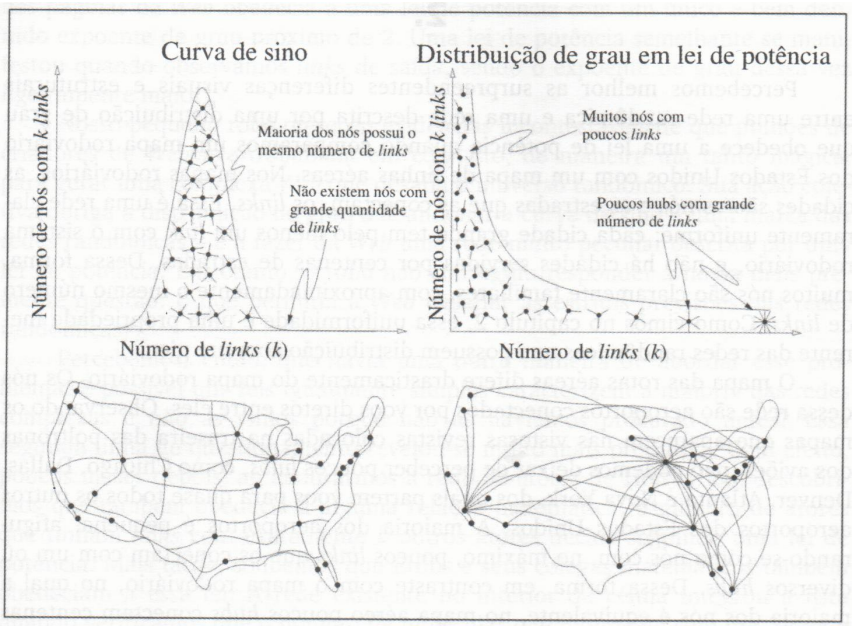

Figura 5.1. Comparação entre a distribuição dos graus dos nós das redes reais e randômicas [Barabasi, 2002]

distribuição na forma, $p(x) = a * x^{-\gamma}$, 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].

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.

Além disso, a distribuição das arestas não é só globalmente, mas também localmente heterogênea, com altas concentrações de bordas dentro de grupos especiais de vértices, e baixas concentrações de entre estes grupos. Esta característica das redes reais é chamado de estrutura da comunidade [Girvan and Newman, 2002], ou clustering. Comunidades, também chamadas de clusters ou módulos, são grupos de vértices com alta probabilidade de compartilharem propriedades ou desempenharem um papel semelhante dentro da rede.

A sociedade oferece uma grande variedade de possíveis organizações em grupos: as famílias, os colegas de trabalhos e os círculos de amizades. A difusão da Internet também levou à criação de grupos virtuais, que vivem na Web, como as comunidades onlines (Facebook, Orkut, Google+, MySpace, etc). De fato, as comunidades sociais têm sido estudadas por um longo tempo [Coleman, 1988, Freeman, 2004]. Comunidades também ocorrer em muitas redes que representam sistemas da biologia, ciência da computação, engenharia, economia, política, etc. Em redes de interação proteína-proteína, as comunidades são como grupos de proteínas que têm a mesma função específica dentro da célula [Newman, 2006], a rede da World Wide Web pode corresponder a grupos de páginas com os mesmos temas ou tópicos relacionados [Flake et al., 2002], em redes me-

tabólicas podem estar relacionadas aos módulos funcionais, tais como ciclos e caminhos [Palla et al., 2005], e assim por diante.

Comunidades têm aplicações concretas, por exemplo, agrupar clientes Web que têm interesses semelhantes e são geograficamente próximos uns dos outros pode melhorar o desempenho dos serviços prestados à World Wide Web, em que cada grupo de clientes pode ser servido por um servidor espelho dedicado [Krishnamurthy and Wang, 2000]. Identificar grupos de clientes com interesses semelhantes na rede de relações de compra entre clientes e produtos de varejistas online (como a www.amazon.com) permite a criação de sistemas de recomendação eficientes [Reddy et al., 2002], que melhor orientam os clientes por meio de uma lista de itens do varejista e aumentam as oportunidades de negócio entre outras aplicações.

A detecção de comunidade é importante não só pela grande quantidade de aplicações, mas por outras razões também. Identificar os grupos e suas fronteiras permite uma classificação dos vértices de acordo com sua posição estrutural nos grupos. Assim, vértices com uma posição central em seus clusters, compartilhando um grande número de arestas com outros membros do grupo, podem ter uma importante função de controle e estabilidade dentro do grupo; vértices encontrados nas fronteiras entre os grupos têm um papel importante de mediação e conduzem as relações e intercâmbios entre as diferentes comunidades, na qual os sinais são transmitidos em todo o grafo seguindo caminhos de comprimento mínimo.

Na literatura os termos *particionamento de grafo* e *detecção de comunidades* são utilizados para referenciar a divisão dos vértices em grupos. Contudo, os termos podem ser diferenciados pelo necessidade ou não de definir o número de grupos e objetos dentro do grupo.

5.3.1. Particionamento de Grafos

O particionamento de grafos é um problema clássico em ciência da computação, estudado desde a década de 60. O problema é a divisão dos vértices de uma rede em grupos não sobrepostos tal que o número de arestas entre os grupos seja minimizados. Neste caso o número de grupos e de vértices em cada grupo é fixo.

Um exemplo clássico é a alocação de tarefas em um processador para que a comunicação entre eles seja minimizada permitindo a alta performance em cálculos. Isto pode ser acoplado pela divisão de um cluster de computadores em grupos com aproximadamente o mesmo número de processadores, tal que o número de conexões físicas entre os processadores de diferentes grupos seja mínima.

Apesar de não ser o foco deste minicurso, o problema de particionamento de grafos é utilizado por alguns algoritmos de detecção de comunidades, assim será feita uma breve explanação sobre o assunto.

O problema mais simples de particionamento de grafos é a divisão de uma rede em duas, chamada bisseção do grafo. A maioria dos algoritmos de particionamento de grafos são de fato algoritmos de bisseção, já que após dividir uma rede em duas, esta pode ter cada uma das duas novas partes divididas em outras duas.

De modo mais formal, o problema de particionar um grafo consiste em dividir o

conjunto de vértices em k grupos de tamanho pré-definido, tal que, o número de arestas entre cada um dos grupos seja mínimo. O número de arestas removidas é chamado *tamanho de corte*. Especificar o número de grupos em que uma rede será particionada é necessário, pois se este número fosse deixado livre a resposta trivial seria uma única partição com todos os nós. A Figure 5.2 ilustra o particionamento de um grafo em dois grupos ($k = 2$) cada um com 7 nós.

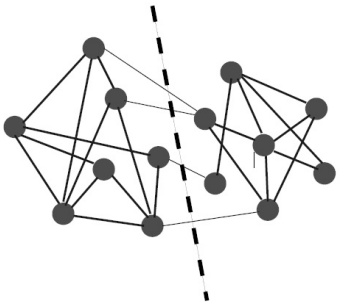

Figura 5.2. Um grafo sendo particionado em dois grupos iguais, cada um com 7 nós.

Apesar de ter uma definição simples, o problema em si não é tão simples de ser resolvido, já que o número de possibilidade de uma rede com N vértices ser dividida em dois grupos N_1 e N_2 é $N!/(N_1!N_2!)$.

O algoritmo mais 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 vértices 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 sucessivos da partição [Karypis and Kumar, 1998]. A Figura 5.3 ilustra o funcionamento do algoritmo METIS em uma rede. Note que apesar de na ilustração apresentar apenas o particionamento da rede em dois grupos, o METIS permite que o grafo seja particionado em k grupos, sendo k definido pelo usuário.

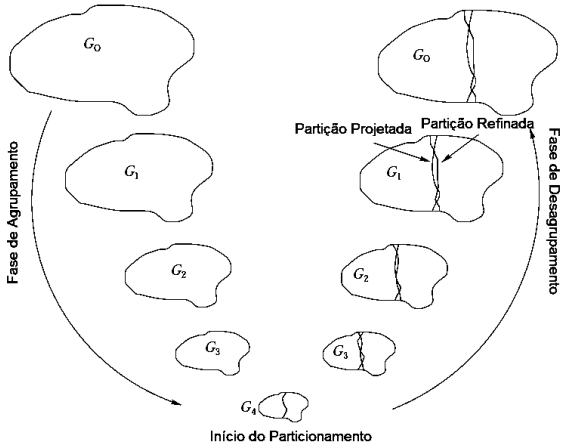

Figura 5.3. Exemplo de funcionamento do algoritmo METIS [Karypis and Kumar, 1998].

Outro método muito utilizado para o particionamento de grafos é o *Spectral Bi-*

section [Alon, 1998]. Este método é baseado no cálculo do segundo menor autovetor da matriz laplaciana da rede. 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. A matriz laplaciana é definida como B tem-se que $\mathbf{B}_{i,j} = d(v_i)$ se $i = j$, $\mathbf{B}_{i,j} = -1$ se $i \neq j \wedge (v_i, v_j) \in \mathscr{E}$ e 0 caso o contrário. Usando o segundo menor autovetor a matriz é reordenada e os grupos são identificados. Um exemplo deste método é apresentado na Figura 5.4. Primeiro a matriz de adjacência de um grafo $\mathscr{G}$, os pontos pretos indicam os uns na matriz e os espaço vazio os zeros, ao seu lado a matriz é reordenada pelo segundo menor autovetor da matriz laplaciana do grafo.

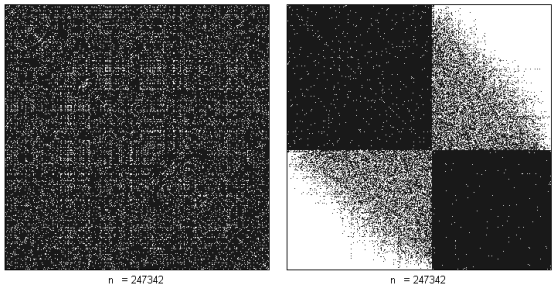

Figura 5.4. Matriz de adjacência de um grafo $\mathscr{G}$ e a matriz reordenada pelo segundo menor autovetor da matriz laplaciana do grafo $\mathscr{G}$

5.3.2. Detecção de Comunidades

O problema da detecção de comunidades se difere do particionamento de grafos, pois nela o número de grupos e o número de elementos dentro dos grupos não são especificados pelo usuário. Ao contrário, este número é determinado pela própria rede, isto é, o objetivo da detecção de comunidades é encontrar grupos formados naturalmente pela estrutura topológica da rede. Além disso, o tamanho de cada grupo existente na rede pode variar abruptamente. A Figura 5.5 apresenta a estrutura tradicional de comunidades em redes complexas. Contudo, o problema de detecção de comunidade é muito menos definido do que o particionamento de grafos. A começar pela definição mais aceita que diz: "Um grupo é um conjunto de vértices que tende a ser mais conectado entre si e menos conectado com outros grupos". O que exatamente significa “mais” e “menos” ? Para responder esta questão uma grande variedade de trabalhos têm sido propostos nos últimos anos.

Assim, a medida de “mais” aresta dentro da comunidade e “menos” aresta fora da comunidade é formalizada com a “densidade intra-cluster” ($\delta_{int}(S)$) e “densidade inter-cluster” ($\delta_{ext}(S)$). A densidade intra-cluster é o número de arestas que ligam dois vértices pertencentes a mesma comunidade dividido pelo número possível de arestas que possam existir dentro comunidade, $\delta_{int}(S) = \frac{arestas\ internas\ de\ S}{nS(nS-1)/2}$. A densidade inter-cluster é similarmente definida, sendo o número de arestas que conectam vértices da comunidade com o restante da rede dividido pelo número máximo de arestas inter-cluster que possa existir, $\delta_{ext}(S) = \frac{arestas\ inter-cluster\ S}{nS(n-nS)}$.

Para S ser uma comunidade é esperado que $\delta_{int}(S)$ ser um número maior que a densidade média da rede ($\delta(\mathscr{G})$), que é dada pelo número de arestas da rede dividido pelo

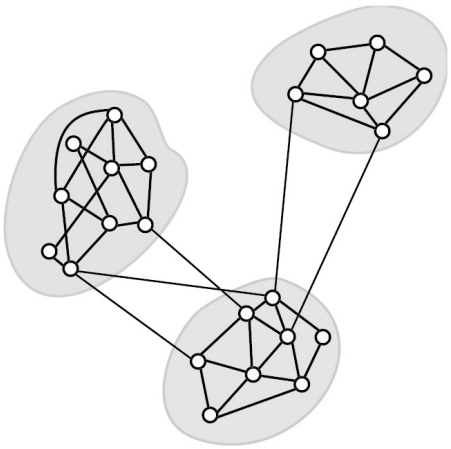

Figura 5.5. Grafo apresentando 3 comunidades do modo tradicional.

número máximo possível de arestas que possam existir na rede. Por outro lado, $\delta_{ext}(S)$ tem que ser muito menor que $\delta(\mathscr{G})$. A busca pela melhor combinação de um $\delta_{int}(S)$ maior e um $\delta_{ext}(S)$ menor está implícito ou explícito nos algoritmos de detecção de comunidades. Uma maneira simples de fazer isto é maximizar a soma das diferenças $\delta_{int}(S) - \delta_{ext}(S)$ sobre todos os grupos de uma partição [Mancoridis et al., 1998].

Uma propriedade indispensável de uma comunidade é a "*connectedness*". Espera-se que para S ser uma comunidade deve haver um caminho entre cada par de vértice da comunidade percorrendo apenas vértices de S. Nesta definição um rede com diversas componentes conexas teria cada uma de suas componentes analisadas independentemente.

Existem muitas outras medidas baseadas em densidade que têm sido usadas para particionar um gráfico em um conjunto de comunidades [Brandes and Erlebach, 2005, Schaeffer, 2007, von Luxburg, 2007]. Uma que merece menção especial é a modularidade [Newman, 2006, Newman and Girvan, 2003]. Para uma determinada partição de uma rede em um conjunto de comunidades, a modularidade mede o número de arestas dentro da comunidade em relação a um modelo (*null model*), que é geralmente considerado como sendo um grafo randômico com o mesmo grau de distribuição. Assim, a modularidade foi originalmente introduzida e normalmente usada para medir a força ou a qualidade de uma partição específica de uma rede.

Em mais detalhes, considere uma rede composta por N nós ou vértices conectados por M arestas e seja A_{ij} um elemento da matriz de adjacência da rede, o que dá o número de arestas entre os vértices i e j. E suponha que é dada uma divisão candidata dos vértices em um certo número de grupos. A modularidade desta divisão é definida como a fração das arestas que se enquadram dentro dos grupos menos a dada fração esperada, caso as arestas sejam distribuídas aleatoriamente. Na versão mais comum do conceito, a randomização das arestas é feita a fim de preservar o grau de cada vértice. Neste caso, o número esperado de arestas compreendida entre dois vértices i e j seguindo a randomização é $k_i * k_j/2M$ onde k_i é o grau do vértice i, e, portanto, o número real menos esperado de arestas entre os mesmos dois vértices é $A_{i,j} - k_i * k_j/2M$. Assim, Q apresentado na Equação 1 é a soma sobre todos os pares (i, j) que estão na mesma comunidade.

$$Q = \frac{1}{2M}\sum_{ij}[A_{ij} - \frac{k_i * k_j}{2m}](C_i, C_j) \quad (1)$$

Outra medida utilizada na detecção de comunidades é o *betwenness* introduzido por Girvan e Newman em [Girvan and Newman, 2002]. O *betwenness* é a medida que calcula a importância de um aresta baseado no número de caminhos mínimos entre pares de nós que passam por ela, ou seja, quanto mais caminhos mínimos passarem por uma aresta para conectar dois nós maior será a importância da aresta. No algoritmo proposto por Girvan e Newman, a detecção de comunidades é feita com sucessivas remoções de arestas que conduzem ao isolamento da comunidade, as arestas à serem retiradas são as que possuem maior *betwenness*. Um exemplo desse processo é apresentado na Figura 5.6

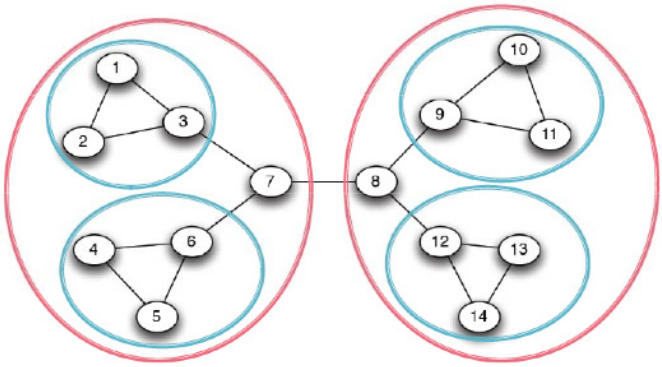

Figura 5.6. Cada um dos círculos representam a definição de uma comunidade. A aresta que conecta os nós 7 e 8 é a que possui a maior *betwenness*, 49, sendo a primeira aresta a ser retirada e dividindo a rede em duas comunidades. O processo de retirada de aresta continua até se encontrar as comunidades desejadas.

Além do *betwenness* e da modularização outra medida utilizada é a condutância. A condutância é uma das medidas mais simples já que leva em consideração o número de arestas dentro e fora da comunidade [Kannan et al., 2000, Leskovec et al., 2008]. Formalmente a condutância $\varphi(S)$ para um conjunto de nós S é $\varphi(S) = cS/min(Vol(S), Vol(V \backslash S))$, sendo cS o número de arestas na fronteira, $cS = |(u,v) : u \in S, v \in S|$ e $Vol(S) = \sum_{u \in S} d(u)$ sendo $d(u)$ é o grau do nó u. Assim, a condutância consegue capturar a noção de comunidades de nós que tenham uma melhor conectividade interna do que externa. A Figura 5.7 mostra as comunidades A, B, C, D e E. O grupo de nós B é mais propício a ser comunidade do que A, pois, $\varphi(A) = 2 > \varphi(B) = 1$.

Uma técnica que se baseia na condutância é o *network community profile* (NCP) que caracteriza a qualidade das comunidades encontradas na rede baseando-se em seu tamanho. Para cada k entre 1 e metade do número de nós da rede é definido $\Phi(k) = min_{|S|=k} f(S)$. O que significa que, para cada possível tamanho de comunidade k, $f(k)$ mede o valor da melhor comunidade de tamanho k e o NCP mede $\Phi(k)$ em função de k. A Figura 5.8 apresenta a melhor condutância para comunidade com $k = 4$ que é B, já que $\Phi(4) = \frac{1}{11}$. Similarmente, D e D+E apresentam a melhor condutância para 3 e 6 nós.

Para grandes redes reais foi observado também que o mínimo global é alcançado

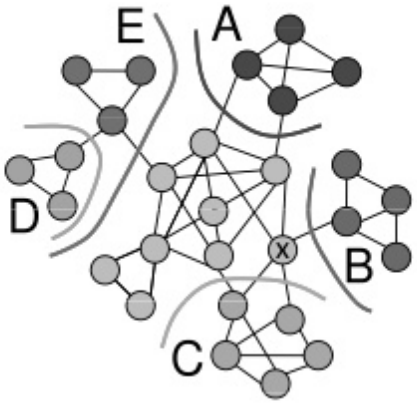

Figura 5.7. Uma rede com diversas comunidades: A, B, C, D e E. Todas foram encontrados utilizando a condutância como medida [Leskovec et al., 2008]

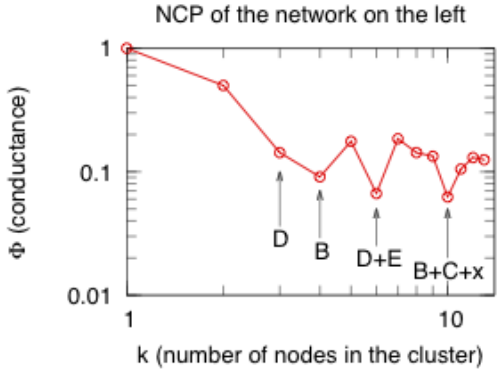

Figura 5.8. Gráfico NCP mostrando a condutância mínima para comunidade de tamanho de 3 a 6 nós. O gráfico se refere ao grafo apresentado na Figura 5.7, as comunidade A, B, C, D e E estão indicadas no gráfico [Leskovec et al., 2008]

por volta de $k = 100$, após esse k o valor da condutância tende a subir. O valor 100 é conhecido como o número de Dunbar, que é o número máximo de relacionamentos que uma pessoa consegue administrar [Dunbar, 1998]. Esta observação sugere que em geral as comunidades não possuem mais do que 100 nós.

Foi observado que o gráfico NCP tende a apresentar um formato em "V". Este formato pode ser explicado por um formato de comunidade diferente do formato apresentado pela Figura 5.5, observada em diversas redes reais, como [Newman, 2006, Ravasz et al., 2002, Girvan and Newman, 2002]. O formato do gráfico NCP indica que as grandes redes reais apresentam uma estrutra de "Centro-Periferia" aninhada [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 5.9. 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].

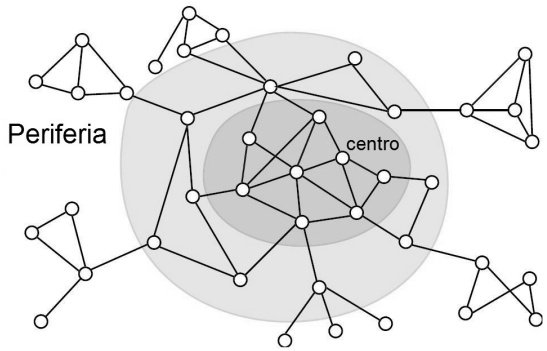

Figura 5.9. Grafo apresentando a topologia "Centro-Periferia"[Leskovec et al., 2008].

A técnica NCP pode ser usada com qualquer outra medida que defina a qualidade da comunidade, não só a condutância. Dentre essas medidas podemos destacar:

- **Espansão:** $f(S) = \frac{cS}{nS}$ - mede o número de arestas por nS que apontam para fora da comunidade [Radicchi et al., 2004].
- **Densidade Interna:** $f(S) = 1 - \frac{mS}{nS(nS-1)}$ - densidade interna das arestas da comunidade S [Radicchi et al., 2004].
- **Corte:** $f(S) = \frac{cS}{nS(n-nS)}$ - fração de todas as arestas possíveis deixando a comunidade [Fortunato, 2010].
- **Corte Normalizado:** $f(S) = \frac{cS}{2mS+cS} + \frac{cS}{2(m-mS)+cS}$ [Shi and Malik, 2000].
- **Maximo-ODF (*Out Degree Fraction*):** $max_{u \in S} \frac{|(u,v):v \notin S|}{d(u)}$ - é o fração máxima de arestas de um nó apontando para fora da comunidade [Flake et al., 2000].
- **Média-ODF:** $f(S) = \frac{1}{nS} \sum_{u \in S} \frac{|(u,v):v \notin S|}{d(u)}$ - é a fração de nós média apontando para fora da comunidade [Flake et al., 2000].
- **Flake-ODF:** $f(S) = \frac{|u:u \in S, |(u,v):v \notin S| < d(u)/2}{nS}$ - é a fração de nós em S que tenha menos arestas apontando para dentro do que para fora da comunidade [Flake et al., 2000].

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., 2004]. 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).

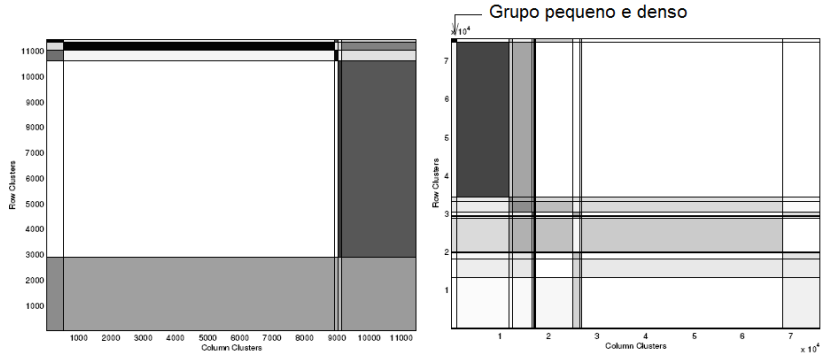

Figura 5.10. Duas redes reais, Epinions e Oregon, processadas pelo método Cross-Association [Chakrabarti et al., 2004].

5.3.3. Predição de Ligações

As redes sociais são objetos altamente dinâmicos, pois eles crescem e mudam rapidamente ao longo do tempo por meio da adição de novas arestas, o que significa o aparecimento de novas interações na estrutura social subjacente. Compreender os mecanismos pelos quais as redes evoluem é uma questão fundamental que ainda não é bem compreendida. 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 $t + 1$.

Com efeito, o problema de predição de ligação aponta para a seguinte pergunta: "até que ponto a evolução de uma rede social pode ser modelada utilizando características intrínsecas à rede em si?" Considere uma rede de co-autoria entre os pesquisadores, por exemplo. Há muitas razões, exógenas à rede, por que dois cientistas que nunca escreveram um artigo juntos vão escrevê-lo nos próximos anos: por exemplo, eles podem se tornarem geograficamente próximos se um deles muda de instituição. Tais colaborações são difíceis de prever. Mas também percebe-se que há um grande número de novas colaborações que são preditas pela topologia da rede: dois pesquisadores que estão "próximos" na rede e tenham inúmeros colegas em comum, e participam dos mesmo círculos sociais sugere que eles são propensos a colaborar em um futuro próximo. O objetivo da predição de ligação é encontrar tais pesquisadores, ou seja de uma maneira mais geral, prever tais ligações.

Uma das dificuldades da predição de ligações é que as redes complexas tendem a ser esparsas. Por exemplo no caso das redes sociais como o Facebook, um usuário tipico é conectado com cerca de 100 pessoas sobre um total de 500 milhos de usuários da rede. Para driblar esta dificuldade, alguns 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. Com isso, o desempenho dos algoritmos pode ser efetivamente melhorado, considerando algumas informações externas, como os atributos dos nós [Getoor and Diehl, 2005, Hasan et al., 2006, Taskar et al., 2004, Popescul et al., 2003]. No senso comum, duas pessoas compartilham mais gostos e interesses (e, portanto, há uma maior probabilidade delas estarem conectados em uma rede social) se elas têm mais características em comum, tais como idade, sexo, trabalho, e assim por diante.

As informações dos atributos podem ser usadas para prever as ligações sem considerar a estrutura de rede. Assim, quando os links existentes não são confiáveis, os métodos baseados em atributos são preferíveis, o que pode, de alguma forma, resolver o chamado problema do começo frio - um grande desafio de previsão link [Leroy et al., 2010]. Além disso, a estrutura de comunidade também pode ajudar a melhorar a precisão da previsão [Zheleva et al., 2010]. Em redes sociais, uma vez que uma pessoa pode desempenhar diferentes papéis em diferentes comunidades, a previsão em um domínio pode ser inspirada pelas informações de outros [Cao et al., 2010]. Por exemplo, quando as colaborações entre os autores são preditas, pode-se considerar suas filiações para melhorar a precisão. Entretanto, esta informação complementar dos nós e arestas nem sempre estão disponíveis, o que inviabiliza a aplicação desses algoritmos nesses casos.

Dentre as técnicas de predição de ligação destacam-se as baseadas em propriedades estruturais do grafo [Liben-Nowell and Kleinberg, 2003, Huang, 2006], especialmente as que trabalham de maneira local na rede. Essas medidas tem como técnica principal atribuir um valor de ligação, chamado $score(u;w)$, para pares de nós $\langle u,w\rangle$ baseado em um dado grafo $\mathscr{G}$. Os valores atribuídos são ordenados em ordem decrescente e então as predições são feitas de acordo com esta lista. Os valores computados podem ser vistos como medidas de proximidades entre nós u e w, relativos a topologia da rede. Contudo, qualquer tipo de medida que compute a semelhança entre dois nós pode ser utilizada. Uma boa revisão na área pode ser encontrada em [Lu and Zhou, 2010].

Para o nó u, seja $\Gamma(u)$ o número de vizinhos de u em $\mathscr{G}$. Um grande número de técnicas de predição de ligações são baseadas na ideia que dois nós u e w devem possuir uma aresta entre eles no futuro, se o conjunto dos seus vizinhos $\Gamma(u)$ e $\Gamma(w)$ tiverem uma grande sobreposição. Esta técnica segue a intuição natural que, em uma rede de co-autoria entre os pesquisadores supracitado, por exemplo, os nós u e w representam autores e $\Gamma(u)$ e $\Gamma(w)$ o conjunto de colegas com quem u e w tiveram pelo menos uma publicação em comum. Assim, quanto maior o número $\Gamma(u) \cap \Gamma(w)$ maior o número de colegas em comum u e w compartilha e consequentemente maior a probabilidade de u e W se tornarem colegas no futuro.

A implementação mais direta desta técnica de predição de ligação é chamada

"vizinhos-comum", sobre o qual para cada par de nó distante por duas arestas e não conectados diretamente é definido um valor que representa o número de vizinhos comuns que u e w compartilham. A Equação 2 apresenta esta medida.

$$score(u;w) := |\Gamma(u) \cap \Gamma(w)| \tag{2}$$

A técnica "vizinhos-comum" captura a noção de que dois estranhos que possuem um amigo em comum podem ser apresentados por este amigo. Isto introduz o efeito de "triângulo fechado" em um grafo e se assemelha a um mecanismo da vida real como apresentado na Figura 5.11.

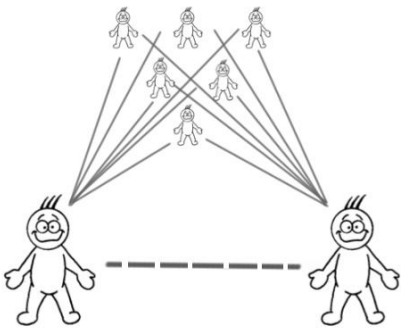

Figura 5.11. A conexão entre dois nós (pessoas) que compartilham muitos vizinhos.

Um exemplo clássico deste mecanismo é a indicação de possíveis amigos em redes sociais. 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.

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

Em Newman [Newman, 2001] foi computado esta quantidade ($\aleph(u,w)$) no contexto da rede de colaboração, verificando uma correlação positiva entre o número de vizinhos comuns de u e w no tempo t, e a probabilidade de u e w colaborarem em algum período depois de t.

O coeficiente de Jaccard [Salton and McGill, 1983] é uma similaridade métrica que é comumente utilizada na recuperação de informação. No contexto da predição de ligação ela é utilizada para medir a probabilidade de ambos u e w ter a característica f, para uma característica f selecionada randomicamente dentre as características em comum entre u e w. Se a característica f escolhida for o número de vizinhos em comum em G, então esta medida captura intuitivamente a noção de proporção de vizinhos de u que também são vizinhos de w (e vice-versa) o que é uma boa medida de similaridade de u e w. Formalmente, o coeficiente de Jaccard usa a Equação 5.

$$score(u;w) := \frac{|\Gamma(u) \cap \Gamma(w)|}{|\Gamma(u) \cup \Gamma(w)|} \tag{5}$$

Adamic e Adar é uma medida semelhante ao coeficiente de Jaccard, no contexto de decidir se dois nós são fortemente relacionados ou não. Para este cálculo, ele computa as características dos nós e defini a similaridade entre eles como apresenta a Equação 6.

$$score(u;w) := \sum_{v \in \Gamma(u) \cap \Gamma(w)} \frac{1}{log\Gamma(v)} \tag{6}$$

A medida Adamic/Adar [Adamic and Adar, 2003] avalia o grau dos vizinhos comum e dá ênfase aos nós em que os vizinhos em comum possuem um grau baixo. Isto por que nós com alto grau tem uma maior chance de ser vizinhos de muitos nós. Assim, se duas pessoas têm em comum um amigo que possui poucos amigos, estas duas pessoas tem uma probabilidade maior de serem amigas em um futuro do que duas pessoas que tem como amigo em comum uma pessoa muito popular que tem inúmeros amigos.

Em suma, se o fechamento de triângulo é um mecanismo pelo qual novas arestas são adicionadas em uma rede, então para u e w serem apresentados por um amigo em comum v, a pessoa v terá que escolher introduzir o par $\langle u;w \rangle$ de $\left(\frac{\Gamma(v)}{2}\right)$ pares de amigos, assim u e w têm mais chance de serem apresentados se v for uma pessoa não popular já que a quantidade de pares de amigos será menor.

Apesar da grande quantidade de medidas desenvolvidas na predição de ligação, os resultados apresentados em [Zhou et al., 2009] indicam que a medida mais simples, "vizinhos-comum", possui a melhor performance sobre as outras medidas. Além de ser uma medida de fácil cálculo e fácil adaptação a outros problemas especialmente quando se utiliza classes de nós.

Uma outra medida ainda local desenvolvida é a "Acoplamento Preferencial" (*Preferential attachment*). Esta medida recebeu bastante atenção com o desenvolvimento do modelo de crescimento de uma rede chamado *Preferential attachment*. A primícia básica desta medida é que a nova aresta é anexa a u proporcionalmente ao valor de $\Gamma(u)$. Newman e Barabasi propuseram, baseado em experimentos empíricos, que a probabilidade de u e w estarem correlacionados é corresponde a Equação 7.

$$score(u;w) := \Gamma(u).\Gamma(w) \tag{7}$$

Além dessas medidas baseadas em vizinhança também há medidas baseadas na contagem de caminhos. Um grande número de métodos refinam a noção de distância de caminho mínimo para a consideração implícita de utilizar todos os caminhos entre dois nós. Dentre esta abordagem se encontra a técnica *Katz* que define a medida como sendo a soma direta sobre toda a coleção de caminhos, exponencialmente amortecida para contabilizar caminhos mais curtos de maneira mais forte como apresenta a Equação 8.

$$score(u;w) := \sum_{l=1}^{\infty} \beta^l.|path_{u,w}^{\langle l \rangle}| \tag{8}$$

Sendo $path_{u,w}^{\langle l \rangle}$ todos os caminhos de tamanho l entre u e w. Usando um β muito pequeno esta medida se aproxima da medida "vizinhos-comum" já que caminhos de comprimento maior que 3 contabilizam muito pouco na soma. Para esta medida há variantes para grafos com peso e sem peso. Sem peso $path_{u,w}^{\langle 1 \rangle} = 1$ se há um caminho entre u e w, caso contrário é 0. Com peso o valor de $path_{u,w}^{\langle 1 \rangle}$ é o número de caminhos ou o peso da aresta entre u e w.

Outra medida comumente utilizada baseada em caminhos é a *random walk*. Uma *random walk* no grafo $\mathscr{G}$ começa no nó u e se move iterativamente para a vizinhança de u que é escolhida de maneira randômica e uniforme. O tempo de alcance $H_{u,w}$ de u para w é o número esperado de passos requerido para uma *random walk* começando em u alcançar w. Um exemplo eficiente dessa abordagem é feito em [Backstrom and Leskovec, 2011] na qual foi desenvolvido o método de *Random Walks* Supervisionada que naturalmente combina informações estruturais da rede com atributos de nós e arestas. Esses atributos são utilizados para guiar a *random walk*. A tarefa supervisionada é usada para aprender funções e atribuir peso às arestas da rede de maneira que nós que terão uma aresta criada no futuro tem mais probabilidades de ser visitados pela *random walk*.

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 detec-

tar 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].

Em muitos domínios identificar ligações anômalas pode ser mais útil que predizer ligações [Rattigan and Jensen, 2005]. A descoberta de ligações anômalas é uma tarefa dependente da predição de ligação, já que é baseada nas mesmas técnicas mas para encontrar arestas com comportamento suspeito ao invés de identificar arestas que aparecerão em um futuro próximo. Exemplos de aplicações estão descoberta de usuários fantasmas em redes sociais, cartões de créditos, aprendizado errôneo em ontologias, etc.

A significativa contribuição do estudo de redes complexas com a predição de ligação é o conhecimento profundo sobre os fatores estruturais que afetam o desempenho de algoritmos, que também pode ser considerado como a orientação da escolha dos algoritmos quando ambos a precisão e complexidade tem que ser levada em conta. Por exemplo, se a rede é altamente clusterizada, os algoritmos baseado em vizinhança comum podem ser boas escolhas, uma vez que pode dar uma previsão relativamente boa com uma complexidade muito baixa. No entanto, se a rede não é altamente clusterizada, ou a distribuição do número de vizinhos comuns decai muito rápido (como no nível de roteamento da Internet [56], 99,98% dos pares de nós não compartilham mais de dois vizinhos comum), algoritmos baseado em vizinhança comum são muito pobres, e devemos tentar abordagens baseada em caminho e *random walk* que fazem uso de mais informações.

Até agora a predição de ligação tem focado redes não direcionadas e sem pesos. Para redes direcionadas, encontrar triângulos é uma tarefa mais complexa fazendo com que mesmo a medida mais simples que é a de vizinhança comum necessite ser modificada para ser utilizada em tais redes. Nesta medida, mesmo que encontrar uma aresta seja uma modificação fácil, definir a direção da aresta pode ser uma tarefa complexa[Mantrach et al., 2009]. A maneira correta de explorar as informações de pesos para melhorar a precisão da predição ainda é um problema não resolvido. Um problema mais difícil é prever os pesos das ligações, o que é relevante para a previsão de tráfego para os sistemas de transporte urbano e transporte aéreo [Yin et al., 2002, Murata and Moriyasu, 2007, Lu and Zhou, 2009, Lu and Zhou, 2010]

Um grande desafio é a predição de ligação em redes multi-dimensionais, onde as ligações podem ter significados diferentes. Por exemplo, uma rede social pode consistir de links positivos e negativos, respectivamente, apontando para amigos e inimigos [Kunegis et al., 2009], ou de confiança e desconfiança [Guha et al., 2004]. Em [Leskovec et al., 2010] propôs-se um método para prever os sinais de links (positivo ou negativo), mas a previsão de tanto a existência de um link e seu sinal não foi bem estudada ainda. Desenvolvimento recente da teoria do equilíbrio social, podem oferecer sugestões úteis [Traag and Bruggeman, 2009, Marvel et al., 2009, Szell et al., 2010].

5.4. Conclusão

Este documento apresentou uma visão geral da área de detecção de comunidades e predição de ligação dentro da mineração de grafos e redes complexas. Esta área tem se

mostrado muito importante atualmente, principalmente pelo grande crescimento do domínio de aplicação, nos quais os dados podem ser modelados através de redes complexas com especial atenção às redes sociais.

Apesar das origens remotas e da grande popularidade dos últimos anos, a pesquisas em detecção de comunidades ainda não forneceu uma solução satisfatória ao problema e deixa com uma série de importantes questões em aberto. O campo não tem uma abordagem teórica que define precisamente o que algoritmos de detecção de comunidade deve fazer, porém, todos têm a sua própria ideia do que é uma comunidade, e mais, as ideias são consistentes entre si, mas, desde que ainda há discordância, continua a ser impossível decidir qual algoritmo faz o melhor trabalho e não há nenhum controle sobre a criação de novos métodos. Portanto, em primeiro lugar a comunidade científica que trabalha em detecção de comunidades deve definir um conjunto de grafos de referência os quais devem servir de base para novos algoritmos.

A predição de ligação apesar de estar melhor definida que a detecção de comunidade ainda possui também o longo caminho para alcançar a maturidade. Sendo a principal questão atender a diversos tipos de redes complexas, como as direcionadas e com pesos e também prever não só a falta de ligação mas também a ligação errônea.

O documento apresentou ainda 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. No campo na mineração de redes complexas, especialmente para a detecção de comunidades e a predição de ligações há muito trabalho a ser feito para estas tarefas atinjam a maturidade semelhante a tarefas na área de mineração tradicional como regras de associação e clusterização.

Agradecimentos

Os autores agradecem ao CNPq pelo financiamento.

Referências

[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.

[Adamic and Adar, 2003] Adamic, L. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

[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.

[Backstrom and Leskovec, 2011] Backstrom, L. and Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In King, I.,

Nejdl, W., and Li, H., editors, *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 635–644. ACM.

[Barabasi, 2002] Barabasi, A.-L. (2002). *Linked: The New Science of Networks*. Perseus Publishing, 1st edition.

[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.

[Brandes and Erlebach, 2005] Brandes, U. and Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Springer.

[Cao et al., 2010] Cao, B., Liu, N. N., and Yang, Q. (2010). Transfer learning for collective link prediction in multiple heterogenous domains. In *International Conference on Machine Learning*, pages 159–166.

[Chakrabarti et al., 2004] Chakrabarti, D., Papadimitriou, S., Modha, D. S., and Faloutsos, C. (2004). 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.

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

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

[Coleman, 1988] Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94.

[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.

[Flake et al., 2000] Flake, G. W., Lawrence, S., and Giles, C. L. (2000). Efficient identification of Web communities. In *KDD*.

[Flake et al., 2002] Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–71.

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

[Freeman, 2004] Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press.

[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.

[Guha et al., 2004] Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA. ACM.

[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)*.

[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.

[Kannan et al., 2000] Kannan, R., Vempala, S., and Vetta, A. (2000). On clusterings – good, bad and spectral. In *FOCS*.

[Karypis and Kumar, 1998] Karypis, G. and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392.

[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.

[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.

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

[Kunegis et al., 2009] Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 741–750, New York, NY, USA. ACM.

[Leroy et al., 2010] Leroy, V., Cambazoglu, B. B., and Bonchi, F. (2010). Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 393–402, New York, NY, USA. ACM.

[Leskovec et al., 2010] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 641–650, New York, NY, USA. ACM.

[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.

[Liu and Wong, 2008] Liu, G. and Wong, L. (2008). Effective pruning techniques for mining quasi-cliques. In *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 33–49, Berlin, Heidelberg. Springer-Verlag.

[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.

[Lu and Zhou, 2010] Lu, L. and Zhou, T. (2010). Link prediction in complex networks: A survey.

[Mancoridis et al., 1998] Mancoridis, S., Mitchell, B. S., and Rorres, C. (1998). Using automatic clustering to produce high-level system organizations of source code. In *In Proc. 6th Intl. Workshop on Program Comprehension*, pages 45–53.

[Mantrach et al., 2009] Mantrach, A., Yen, L., Callut, J., Francoisse, K., Shimbo, M., and Saerens, M. (2009). The sum-over-paths covariance kernel: A novel covariance measure between nodes of a directed graph. In *the IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Marvel et al., 2009] Marvel, S. A., Strogatz, S. H., and Kleinberg, J. M. (2009). The energy landscape of social balance.

[Murata and Moriyasu, 2007] Murata, T. and Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 85–88, Washington, DC, USA. IEEE Computer Society.

[Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

[Newman, 2001] Newman, M. E. J. (2001). Scientific collaboration networks: II. shortest paths, weighted networks, and centrality. *Physics Review E*, 69.

[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.

[Newman and Girvan, 2003] Newman, M. E. J. and Girvan, M. (2003). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+.

[Nicoletti, 2006] Nicoletti, Maria Do Carmo ; Hruschka Jr., E. (2006). *Fundamentos da Teoria dos Grafos.*, volume 1. EdUFSCar - Editora da Universidade Federal de Sï¿½o Carlos, 1. ed. revisada edition.

[Palla et al., 2005] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

[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.

[Rattigan and Jensen, 2005] Rattigan, M. J. and Jensen, D. (2005). The case for anomalous link detection. In *Proceedings of the 4th international workshop on Multi-relational mining*, MRDM '05, pages 69–74, New York, NY, USA. ACM.

[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.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

[Schaeffer, 2007] Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:888–905.

[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.

[Szell et al., 2010] Szell, M., Lambiotte, R., and Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world.

[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*.

[Traag and Bruggeman, 2009] Traag, V. A. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80(3).

[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.

[von Luxburg, 2007] von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

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

[Yin et al., 2002] Yin, H., Wong, S. C., Xu, J., and Wong, C. K. (2002). Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85 – 98.

[Zheleva et al., 2010] Zheleva, E., Getoor, L., Golbeck, J., and Kuter, U. (2010). Using friendship ties and family circles for link prediction. In *Proceedings of the Second international conference on Advances in social network mining and analysis*, SNAKDD'08, pages 97–113, Berlin, Heidelberg. Springer-Verlag.

[Zhou et al., 2009] Zhou, T., Lü, L., and Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(4):623–630.