

UNIVERSIDADE FEDERAL DO PARANÁ

GEORGEA DANIELEWICZ

ANÁLISE DE UMA MÉTRICA ALTERNATIVA PARA PREDIÇÃO DE LAÇOS
SOCIAIS EM GRAFOS LEI DE POTÊNCIA

CURITIBA PR

2016

GEORGEA DANIELEWICZ

ANÁLISE DE UMA MÉTRICA ALTERNATIVA PARA PREDIÇÃO DE LAÇOS
SOCIAIS EM GRAFOS LEI DE POTÊNCIA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: André Luís Vignatti.

CURITIBA PR

2016

Danielewicz, Georgea

Análise de uma métrica alternativa para predição de laços sociais em grafos lei de potência / Georgea Danielewicz. – Curitiba, 2016
47 f. : il.; tabs.

Dissertação (mestrado) – Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.
Orientador: André Luís Vignatti

1. Redes sociais on line. 2. Teoria dos grafos. 3. . I. Vignatti, André Luís. II. Título

CDD 004.65




MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor CIÊNCIAS EXATAS
Programa de Pós Graduação em INFORMÁTICA
Código CAPES: 40001016034P5


TERMO DE APROVAÇÃO

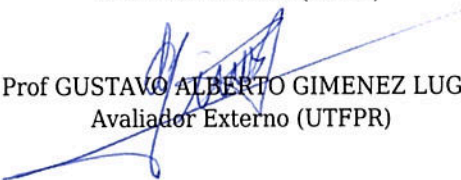
Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **GEORGEA DANIELEWICZ**, intitulada: "**Análise Alternativa de uma Métrica para Predição de Laços Sociais em Grafos Lei de Potência**", após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 28 de Julho de 2016.


Prof ANDRÉ LUIZ PIRES GUEDES
Presidente da Banca Examinadora (UFPR)


Prof ANDRÉ LUÍS VIGNATTI
Coorientador - Avaliador Interno (UFPR)


Prof EDUARDO JAQUES SPINOSA
Avaliador Interno (UFPR)


Prof GUSTAVO ALBERTO GIMENEZ LUGO
Avaliador Externo (UTFPR)



Agradecimentos

Agradeço de coração ao meu orientador Professor André Luís Vignatti, pela tamanha atenção e dedicação. Também agradeço à minha família e colegas pelo apoio e incentivos. Pois sem vocês esta conquista não teria sido possível.

Resumo

As redes sociais são uma maneira de descrever as interações sociais em um grupo ou comunidade. Podem ser modeladas por meio de grafos, em que um vértice corresponde a uma pessoa, e uma aresta representa alguma forma de associação entre duas pessoas [Hasan e Zaki, 2011]. As redes sociais são objetos altamente dinâmicos, elas crescem e mudam rapidamente ao longo do tempo devido à adição e exclusão de vértices e arestas. Compreender os mecanismos pelos quais estas estruturas sociais evoluem é uma questão fundamental, ainda não bem compreendida, e que constitui a motivação para este projeto. Mais especificamente, a pesquisa se dedica ao problema da predição de laços sociais: dado um *snapshot* de uma rede social em tempo t_0 , busca-se prever com precisão as arestas que serão adicionados à rede em um determinado momento t futuro, tal que $t > t_0$ [Liben-Nowell, 2005]. As soluções estudadas se dividem em dois grupos principais: predição de laço supervisionada e a predição de laço não supervisionada. A predição de laço supervisionada envolve técnicas de aprendizado de máquina como a extração de características e algoritmos de classificação [Zhang e Yu, 2011]. A predição não supervisionada busca de calcular métricas baseadas nas características topológicas do grafo [Hasan e Zaki, 2011]. Com base no segundo paradigma, e a partir do estudo de modelos de geração de grafos, é proposta como contribuição uma métrica para calcular a probabilidade de formação de arestas entre dois vértices específica para grafos com distribuição de grau Lei de Potência.

Palavras-chave: redes sociais, predição de laço, modelo de geração de grafo.

Abstract

Social networks are a popular way to model the interactions among people in a group or community. They can be visualized as graphs, where a vertex corresponds to a person in some group and an edge represents some form of association between the corresponding persons [Hasan e Zaki, 2011]. Social networks are very dynamic, since new edges and vertices are added to the graph over time. Understanding the dynamics that drive the evolution of social networks is a complex problem, yet to be fully understood, and which comprises the motivation of this project. A basic computational problem underlying social-network evolution is the *link-prediction problem*: given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t_0 , considering $t > t_0$ [Liben-Nowell, 2005]. Most works in this field branch into two main groups: supervised and unsupervised link prediction. Supervised link prediction models have two important components: feature extraction and classification [Zhang e Yu, 2011]. Unsupervised link prediction calculates metrics based on features which are extracted from the graph topology, some works develop a graph evolution model [Hasan e Zaki, 2011]. Based on unsupervised link prediction concepts, mainly on graph generation models, we propose a link prediction metric, which is specific to power-law graphs.

Keywords: social networks, link prediction, graph generation model.

Sumário

1	Introdução	1
1.1	Contexto	1
1.2	Motivação e Justificativas	2
1.3	Objetivos	2
1.4	Estrutura do Documento	2
2	Fundamentação Teórica	3
2.1	Grafos	3
2.2	Redes Sociais	4
2.2.1	Redes Sociais Homogêneas	5
2.2.2	Redes Sociais Heterogêneas	5
2.2.3	Fontes de Redes Sociais	6
2.2.4	Fechamento Triádico	7
2.2.5	Laços Fracos e Fortes	7
2.2.6	Pontes e Pontes Locais	8
2.2.7	Pontes Locais e Laços Fracos	9
2.3	Fenômenos de Popularidade em Grafos Sociais	9
2.3.1	Distribuição Lei de Potência	10
3	O Problema da Predição de Laços	13
3.1	Predição de Laços Não Supervisionada	14
3.1.1	Métricas Baseadas em Vizinhança Local	14
3.1.2	Métricas Baseadas em Caminho mais Curto	17
3.1.3	Passeio Aleatório	17
3.1.4	Métricas de Proximidade Baseadas em Passeio Aleatório	19
3.2	Predição de Laços Supervisionada	20
3.2.1	Extração de Características	20
3.2.2	Algoritmos de Classificação	20
3.3	Métricas de Avaliação	21
3.3.1	Tabela de Confusão	21
3.3.2	Taxa de Acerto	22
3.3.3	Precisão	22
3.3.4	Sensibilidade	22
3.3.5	Especificidade	22
3.3.6	Curva ROC (<i>Receiver Operator Characteristic</i>)	22

4	Análise de Modelos de Geração de Grafos e Proposta de Métrica	25
4.1	Análise de Modelos de Geração de Grafos	27
4.1.1	Modelo de Configuração	27
4.1.2	Modelo Grafo Aleatório Generalizado	28
4.1.3	Modelo Ligação Preferencial	29
4.1.4	Modelo de Bollobás e Riordan	31
4.1.5	Modelo ACL – Aiello, Chung e Lu	32
4.1.6	Modelo de Vignatti e da Silva	33
4.2	Proposta de Métrica	33
4.3	Análise de um Experimento Relacionado	35
4.3.1	Métodos Aplicados	35
4.3.2	Resultados Obtidos	36
5	Conclusão	39
	Referências Bibliográficas	41
A	Análise Avançada do Modelo Ligação Preferencial	43
B	Demonstração do Lema 4.1.1	47

Lista de Figuras

2.1	Grafo	3
2.2	Grafo direcionado	4
2.3	Rede social baseada em dados on-line	5
2.4	Representação de rede do tipo heterogênea	6
2.5	Grafos antes e após o fechamento triádico	8
2.6	Grafo exibindo ponte local entre vértices A e B	9
2.7	Distribuição Lei de Potência	11
3.1	Cálculo de coeficiente de <i>clustering</i> para o vértice A	16
3.2	Cálculo de coeficiente de <i>clustering</i> para o vértice A	16
3.3	Rede utilizada para exemplo de passeio aleatório	18
3.4	Curva ROC	23
4.1	Visão Geral das Pesquisas sobre Formação de Laços Sociais	26
4.2	Vértices com arestas incompletas	28
4.3	Exemplo de geração de grafo pelo modelo de Bollobás	31
4.4	Exemplo de aplicação da métrica sugerida sobre um grafo	34
4.5	Comparativo das métricas para predição de laços	36

Lista de Tabelas

3.1	Tabela de Confusão	21
-----	------------------------------	----

Lista de Acrônimos

ACL	Modelo de Geração de Grafo Aiello, Chung e Lu
DINF	Departamento de Informática
GRG	Modelo Generalized Random Graphs
PPGINF	Programa de Pós-Graduação em Informática
ROC	Receiver Operator Characteristic
UFPR	Universidade Federal do Paraná

Lista de Símbolos

α	logaritmo do tamanho de um dado grafo
β	taxa de crescimento de um grafo em escala log – log
E	conjunto de arestas de um grafo
G	grafo G
$\Gamma(i)$	conjunto de vizinhos do vértice i
w_i	grau do vértice i

Capítulo 1

Introdução

Neste capítulo são definidos o contexto, a motivação e as justificativas, os objetivos e a estrutura de organização do trabalho.

1.1 Contexto

As redes sociais são uma maneira bastante popular de modelar as interações sociais em um grupo ou comunidade. Podem ser modelados por meio de grafos, em que um vértice corresponde a uma pessoa, e uma aresta representa alguma forma de associação entre duas pessoas [Hasan e Zaki, 2011].

As redes sociais são objetos altamente dinâmicos: elas crescem e mudam rapidamente ao longo do tempo, por exemplo, por meio da adição de novas arestas, o que implica no aparecimento de novas interações na estrutura social [Liben-Nowell, 2005]. Compreender os mecanismos pelos quais elas evoluem é uma questão fundamental, ainda não bem compreendida, e que delinea o escopo deste projeto.

Um problema básico relativo à evolução de redes sociais é o problema da predição de laços sociais: dada uma rede social em um instante de tempo t_0 , busca-se prever com precisão as arestas que serão adicionados à rede em um determinado momento t futuro, sendo que $t > t_0$ [Liben-Nowell, 2005]. A pesquisa aqui apresentada pretendeu abordar o problema da predição de laços sociais, concentrando-se em grafos que apresentam um comportamento relativo à distribuição de graus de seus vértices denominado Lei de Potência.

A contribuição do trabalho consiste na proposta de uma métrica para predição de laços sociais específica para grafos com distribuição de grau Lei de Potência. Esta métrica foi derivada do modelo de geração de grafos de Vignatti e da Silva, específico para descrever grafos com esta distribuição de grau.

Cabe ressaltar que os modelos de geração de grafo não são soluções imediatas para o problema da predição de laços sociais. Os modelos de geração de grafo são algoritmos com a finalidade de gerar grafos com uma característica específica. São elaborados de modo a descrever matematicamente o comportamento de grafos que representam redes sociais reais, buscando capturar fenômenos que ocorrem no problema real, como o aparecimento de arestas. As soluções diretas para o problema analisado são as métricas. E estas são derivadas de modelos que descrevem a situação real.

1.2 Motivação e Justificativas

O estudo das redes reais de larga escala vem se tornando extremamente importante para a sociedade atual. Um exemplo disso são os sistemas econômicos e tecnológicos da atualidade, que se mostram cada vez mais dependentes de redes altamente complexas [Easley e Kleinberg, 2010].

A predição de laços sociais é um problema essencial, que busca compreender mecanismos que governam a evolução destas redes complexas. Contudo, há soluções já existentes que abrem espaço para aprimoramento. Como é o caso da métrica para predição de laços sociais conhecida como métrica Ligação Preferencial. Esta métrica foi derivada do modelo de geração de grafo Ligação Preferencial. Este modelo, no entanto, é baseado em uma análise inadequada, como será apresentado no Capítulo 4. Logo, buscou-se propor uma métrica derivada de um modelo construído com base em uma análise matemática adequada, que é o caso do modelo de Vignatti e da Silva.

A importância da temática e o fato de que as soluções já existentes ainda não esgotaram o problema delineiam a motivação para a realização deste trabalho, que consiste na proposta de uma métrica para predição de laços em grafos Lei de Potência.

1.3 Objetivos

O objetivo geral do presente trabalho consistiu em propor uma métrica para predição de laços em grafos Lei de Potência baseada em análise matemática adequada. Para cumprir o objetivo geral da pesquisa proposta, fez-se necessário atingir os seguintes objetivos específicos:

- Analisar a maneira como modelos de geração de grafos, especialmente grafos Lei de Potência, analisam a questão da formação de laços sociais;
- Obter uma métrica para predição de laços em grafos Lei de Potência derivada do modelo de geração de grafos de Vignatti e da Silva;

1.4 Estrutura do Documento

Este documento foi estruturado na forma de capítulos, e cada capítulo, por sua vez, é dividido em seções. O Capítulo 2 corresponde à fundamentação teórica, onde são apresentados e devidamente referenciados os princípios teóricos envolvidos no desenvolvimento do trabalho. Em seguida, o Capítulo 3 descreve os dois grupos principais em pesquisas acerca da predição de laços sociais: predição de laço supervisionada e a predição de laço não supervisionada.

O capítulo 4 apresenta uma análise de modelos de geração de grafos, com ênfase nos grafos Lei de Potência, e propõe a utilização de uma métrica para predição de laços em grafos Lei de Potência. Por fim, a conclusão a respeito das informações apresentadas neste documento corresponde ao Capítulo 5.

Capítulo 2

Fundamentação Teórica

Neste capítulo apresentamos os conceitos envolvidos no desenvolvimento do projeto. O conteúdo da Seção 2.1 foi baseado na obra de [Easley e Kleinberg, 2010].

2.1 Grafos

Grafos são uma maneira de representar relacionamentos entre um conjunto de itens. São extremamente úteis, pois podem servir como modelos matemáticos das estruturas de redes. Os grafos são constituídos de um conjunto de objetos, também chamados de *vértices*, que formam pares entre si conectados por ligações chamadas de *arestas*. A Figura 2.1 mostra os vértices A, B, C e D, sendo que B se conecta aos outros três vértices por meio de arestas. O conceito de *grau* é muito relevante neste estudo, e se refere à quantidade de arestas ligadas a um dado vértice [Newman, 2010]. Considerando a Figura 2.1, o *grau* do nó B é igual a três, pois está conectado a três arestas.

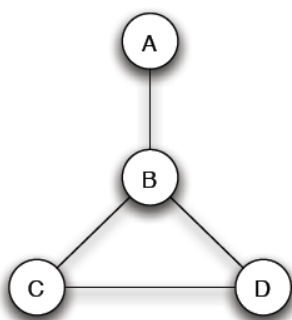


Figura 2.1: Grafo
Fonte: [Easley e Kleinberg, 2010]

Ainda na Figura 2.1 é possível observar que a ligação entre os pares de vértices são simétricas, ou seja, A está ligado com B assim como B está ligado com A. Em outros casos, podem acontecer ligações assimétricas, por exemplo, um vértice A está ligado com o vértice B mas não o oposto. Por esta razão, definem-se *grafos direcionados* como um conjunto de vértices cujas arestas “apontam” de um vértice a outro. Grafos deste tipo podem ser representados visualmente como na Figura 2.2.

Em um grafo, *caminhos* são definidos como sequências de vértices, com a propriedade de que cada par de vértices consecutivos está conectado por uma aresta. Ou seja, o caminho

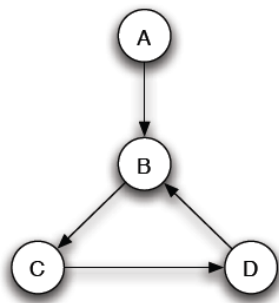


Figura 2.2: Grafo direcionado
 Fonte: [Easley e Kleinberg, 2010]

contém, além dos vértices, também as arestas que os ligam. Um grafo é dito *conexo*, se quaisquer dois vértices puderem ser ligados por pelo menos um caminho.

Além disso, um caminho pode conter vértices repetidos, caso isto não ocorra o caminho é dito simples. O *comprimento do caminho* é definido como a quantidade de arestas contidas na sequência, desde seu início até o fim. Já o conceito de *distância entre dois vértices* é definido como o comprimento do caminho mais curto entre eles. O *diâmetro* de um grafo consiste na maior distância entre dois vértices.

Adicionalmente a estes conceitos, define-se um *ciclo* como sendo um caminho com pelo menos três arestas, em que o mesmo vértice é ao mesmo tempo o primeiro e o último vértice, e o restante dos vértices são distintos.

Há estudos que se dedicam a analisar estas características topológicas de grafos que representam redes sociais reais, como se pode citar o trabalho de [Leskovec et al., 2005]. Ao analisar uma ampla variedade de grafos reais, o estudo observou alguns fenômenos surpreendentes. Primeiro, os grafos se densificam ao longo do tempo. Isso quer dizer que o número de arestas cresce mais rapidamente se comparado ao número de vértices. Além disso, o diâmetro efetivo entre os vértices frequentemente encolhe com o tempo, ao contrário da ideia convencional de que a distância deveria aumentar lentamente como função do número de vértices. Os grafos analisados apresentaram os seguintes padrões de crescimento:

- **Densificação:** as redes se tornaram densas com o passar do tempo, com o aumento do grau médio;
- **Encolhimento dos diâmetros:** o diâmetro efetivo ¹ da rede encolhe conforme a rede cresce, ao contrário do que se acreditava.

Estes tipo de estudo ilustra a maneira como a análise de grafos sociais reais permite que sejam observados e registrados comportamentos inerentes ao que ocorre nas situações reais. Em um segundo momento, estas características podem ser reproduzidas por meio de modelos matemáticos de geração de grafos.

2.2 Redes Sociais

Redes podem ser definidas como um conjunto de objetos, em que pares de objetos estão conectados por meio de *laços*, também chamados de *relacionamentos* ou *conexões*. Redes

¹O diâmetro efetivo corresponde ao percentual de 90% do diâmetro do grafo

podem ser encontradas em diversos contextos, dentre as quais podemos destacar as redes sociais. A Figura 2.3 mostra uma rede social formada por 436 empregados da *Hewlett Packard*. As pessoas são representadas pelos vértices, e as arestas conectam pares de pessoas que possuem um relacionamento, por exemplo, de amizade ou troca de *emails* [Easley e Kleinberg, 2010].

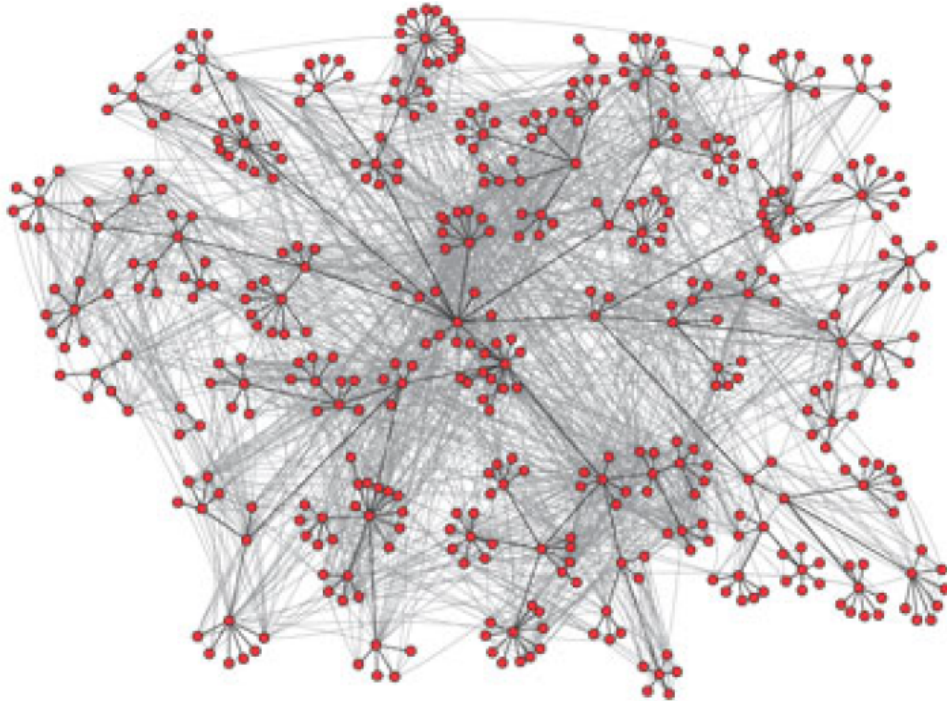


Figura 2.3: Rede social baseada em dados on-line
Fonte: [Easley e Kleinberg, 2010]

Observa-se que partes da rede podem apresentar maior ou menor densidade de conexões, algumas vezes com núcleos contendo a maioria das conexões. O modelo formal para representar redes sociais são grafos [Easley e Kleinberg, 2010].

As redes sociais são muito dinâmicas, uma vez que novas arestas e vértices são adicionados ao grafo ao longo do tempo. Compreender a dinâmica que impulsiona a evolução das redes sociais é um problema complexo devido ao grande número de variáveis envolvidas. Mas, um problema comparativamente mais fácil é entender a associação entre dois vértices específicos [Hasan e Zaki, 2011].

2.2.1 Redes Sociais Homogêneas

Uma rede $G = (V, E)$ é composta pelo o conjunto de vértices V e pelo conjunto de arestas ou laços E . Se todos os vértices em V são do mesmo tipo, assim como todos os laços em E , então a rede G é definida como uma rede social homogênea [Zhang e Yu, 2011]. O *Facebook*, considerando apenas os usuários e os laços de amizades entre eles, pode ser visto como uma rede social do tipo homogênea, pois os vértices existentes são apenas do tipo *usuário*, ao passo que os laços que ocorrem entre eles são sempre do tipo *amizade*.

2.2.2 Redes Sociais Heterogêneas

Redes sociais heterogêneas contêm tipos diferentes de vértices e de laços. Estas redes podem ser definidas como $G = (V, E)$, em que $V = \cup_i V_i$ é a união de diferentes tipos de vértices,

e $E = \cup_i E_i$ representa a união de diferentes conjuntos de laços [Zhang e Yu, 2011]. O *Facebook* pode ser visto como uma rede social do tipo heterogênea, considerando que além dos vértices do tipo usuários, também há vértices representando diferentes assuntos, como música, filmes e vídeos, e aos quais os usuários podem se associar, como ilustrado na Figura 2.4.

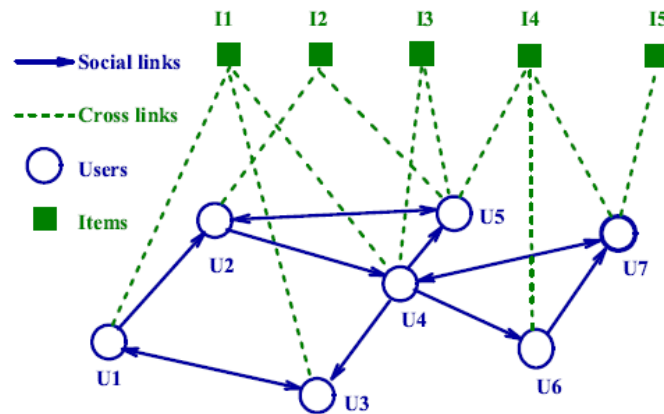


Figura 2.4: Representação de rede do tipo heterogênea
Fonte: [Li et al., 2013]

2.2.3 Fontes de Redes Sociais

A recente explosão de interesse na pesquisa de redes sociais levou à proliferação de dados em larga escala disponíveis no campo das interações sociais. Nesta subseção serão mencionados alguns dos tipos de redes derivados de várias interações sociais que atualmente se encontram disponíveis para a pesquisa [Liben-Nowell, 2005].

Redes Sociais do Mundo Real

Pesquisas sociológicas tradicionais sobre redes sociais se baseiam em entrevistas pessoais meticulosas com temas diversos, em que os pesquisadores perguntam aos indivíduos quem são seus amigos. Um exemplo famoso de um estudo deste tipo é o *Zachary Karate Club*, em que os dados sobre as interações entre trinta e quatro membros de um clube de karatê foram coletados por meio de entrevistas e observações durante um período de dois anos. Enfrenta-se uma dificuldade em relação aos dados recolhidos a partir de entrevistas e pesquisas: as pessoas têm definições extremamente diferentes do que constitui um amigo. Por exemplo, Milgram tratou este problema definindo explicitamente um amigo como uma pessoa com quem se comunica por meio do primeiro nome mutuamente [Liben-Nowell, 2005]. No caso de redes sociais virtuais, esta tarefa é facilitada, pois as pessoas atestam a amizade.

Redes de Colaboração

Outra fonte de dados sobre interações são as redes de colaboração, que conectam dois indivíduos se eles colaboraram em um projeto comum. As redes de colaboração mais estudadas são baseadas no grafo de atores de *Hollywood*, em que os atores que atuaram juntos em um filme são conectados por uma aresta, e no grafo acadêmico de colaboração, onde pesquisadores que participaram de um mesmo trabalho acadêmico são conectados por uma aresta. Redes de

citação, como a *arXiv*, em que os vértices são artigos ligados por arestas dirigidas a partir de cada artigo para os artigos que ele cita, não são as redes sociais, pois não representam realmente uma relação social entre seus vértices [Liben-Nowell, 2005].

O estudo desenvolvido por [Leskovec et al., 2005] teve como objetivo observar a evolução temporal de redes de colaboração, por meio da análise de *snapshots* das redes registrados em pontos regularmente espaçados no tempo. As bases utilizadas foram:

- *ArXiv citation graph* um grafo de citações de uma base de dados da *arXiv* com 29555 artigos e 352807 arestas;
- *Patents citation graph* uma base de dados que compreende patentes de utilidade registradas de 1963 a 1999, totalizando 3923922 patentes e 16522438 citações;
- *Autonomous system graph* um grafo dos roteadores existentes na Internet;
- *Affiliation graphs* grafos criados com base no grafo de citações da *arXiv*.

A Web e os Blogs

A própria *Web* apresenta características intrinsecamente sociais, como por exemplo a existência de páginas pessoais. *Blogs* são diários *online*, muitas vezes até atualizados diariamente, tipicamente contendo relatórios sobre a vida pessoal do usuário, reações a eventos mundiais, e comentários de leitores. *Links* existentes em um *blog* geralmente apontam para outros *blogs* lidos pelo autor, ou para *blogs* de amigos do blogueiro. Dessa forma, constata-se que a estrutura de *links* dentro das comunidades de *blogs* é uma relação essencialmente social. Assim, redes sociais podem ser derivadas da estrutura de *links* da comunidade de *blogs* [Liben-Nowell, 2005]. Além dos *blogs*, existem na *Web* as redes sociais de amizades como o *Facebook*.

2.2.4 Fechamento Triádico

Muito se questiona a respeito da evolução de uma rede social no decorrer do tempo, em particular sobre os mecanismos pelos quais arestas são formadas e depois desfeitas. Um dos princípios básicos é o de que se duas pessoas em uma rede social possuem um amigo em comum, então há uma chance elevada de que se tornem também amigos em algum ponto no futuro [Easley e Kleinberg, 2010].

Este princípio é chamado de *fechamento triádico*, e pode ser explicado por meio da Figura 2.5. Se os vértices B e C são amigos do vértice A, a formação de uma aresta entre B e C é altamente provável e produz a estrutura de um triângulo na rede. O termo fechamento triádico tem origem no fato de que a aresta B-C tem o efeito de “fechar” o terceiro lado do triângulo [Easley e Kleinberg, 2010].

Há diversas razões pelas quais B e C são propensos a interagirem, quando são amigos de A. Uma delas seria a oportunidade de B e C se encontrarem, pois ambos passam tempo com A. Além disso, a inexistência de uma amizade entre B e C poderia se tornar uma fonte de *stress* latente [Easley e Kleinberg, 2010].

A pesquisa de [Li et al., 2013] analisou bases de dados reais com foco nos novos laços sociais que se formavam, classificando-os em triádicos e não triádicos. Destas novas conexões geradas, a maioria foi decorrente do fechamento triádico. Mais de 80% dos laços sociais do *Flickr* e 70% dos laços sociais do *Epinions*, as bases de dados analisadas, nasceram de fechamentos triádicos.

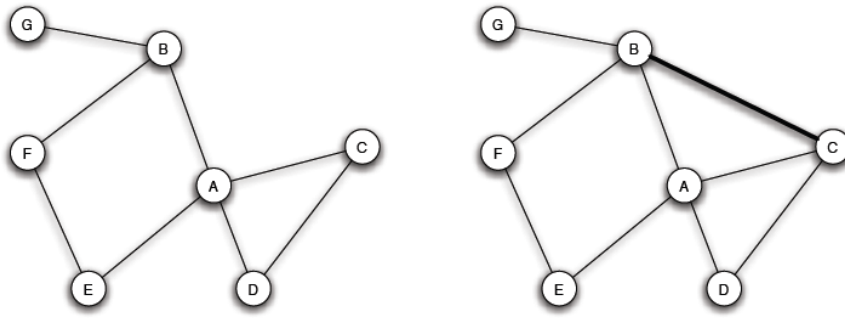


Figura 2.5: Grafos antes e após o fechamento triádico
 Fonte: [Easley e Kleinberg, 2010]

2.2.5 Laços Fracos e Fortes

As conexões nas redes sociais apresentam diferentes níveis de força. Conexões mais fortes representam amigos mais próximos e maior frequência de interação. Apesar dos diversos níveis de força, para efeitos de simplificação de um suposto modelo, muitos consideram apenas dois: laços fracos (que correspondem, por exemplo, a meros conhecidos) e fortes (que correspondem, por exemplo, à amizades). Mesmo sendo uma simplificação, isto acaba sendo uma generalização da definição mais comum de grafos, pois agora temos três possibilidades de relação entre dois vértices: sem aresta, aresta fraca e aresta forte. Sabe-se também que se um vértice A tem laços com B e C, o laço B-C apresenta chances elevadas de se formar, se os laços A-B e A-C forem fortes [Easley e Kleinberg, 2010]. A partir disso, pode-se chegar às Definições 2.2.1 e 2.2.2:

Definição 2.2.1. O vértice A viola o *Fechamento triádico forte* se possui laços fortes com vértices B e C, e não existe laço forte ou fraco entre B e C.

Definição 2.2.2. Uma rede satisfaz a propriedade do *Fechamento triádico forte*, logo, para todos os seus vértices: se um deles apresenta laços fortes com dois vértices vizinhos, estes dois vértices deverão obrigatoriamente apresentar pelo menos um laço fraco entre si.

2.2.6 Pontes e Pontes Locais

Como visto anteriormente, em alguns modelos de redes, os laços sociais podem ser caracterizados como fortes ou fracos. Além disso, podem também representar papéis importantes na estrutura da rede, como pontes e pontes locais. Na Figura 2.6, a pessoa A, por exemplo, possui cinco amigos, mas uma de suas amizades é qualitativamente diferente das demais: as arestas de A para C, D e E fazem ligação com um grupo de amigo em que todos se conhecem. No entanto, a ligação com B parece tornar uma parte diferente da rede mais acessível. Uma das consequências disso seria que o grupo de vértices A, C, D, E e F tenderão a ser expostos a informações similares. Já a amizade com B fornece acesso a novidades. Com base nisso, temos as seguintes Definições 2.2.3 e 2.2.4 [Easley e Kleinberg, 2010]:

Definição 2.2.3. A aresta entre A e B pode ser definida como uma *ponte* se ao ser removida, A e B pertencerem a componentes conexos diferentes.

Definição 2.2.4. A aresta entre A e B é dita uma *ponte local* se A e B não possuem amigos em comum, ou seja, se for removida, a distância entre A e B aumenta para um valor maior que dois.

Na Figura 2.6, a aresta A-B é uma ponte local com diâmetro quatro, pois a distância entre estes dois vértices sem considerar a ponte local A-B. Além desta, não há outras pontes locais na rede [Easley e Kleinberg, 2010].

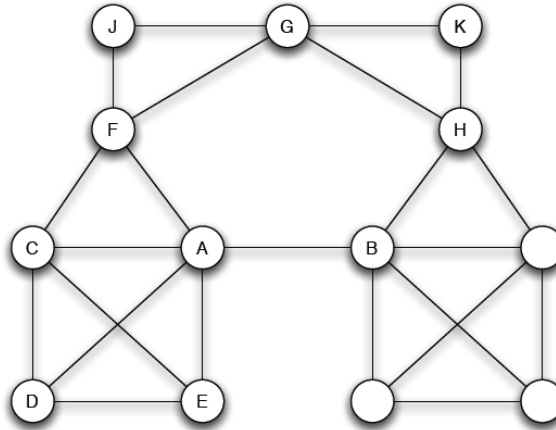


Figura 2.6: Grafo exibindo ponte local entre vértices A e B
Fonte: [Easley e Kleinberg, 2010]

2.2.7 Pontes Locais e Laços Fracos

Observa-se que pontes locais em uma rede que respeita a Definição 2.2.2 são necessariamente laços fracos, como mostra o Teorema 2.2.1:

Teorema 2.2.1. Se um vértice A satisfaz o fechamento triádico forte e está envolvido em pelo menos dois laços fortes, então qualquer ponte local na qual está envolvido deve ser um laço fraco.

Demonstração. Suponha que o vértice A de uma rede satisfaz ao fechamento triádico forte, e está envolvido em pelo menos dois laços fortes. Suponha também, por contradição, que A e B formam uma ponte local que é um laço forte. Como A possui pelo menos dois laços fortes, sendo um deles com B, o outro será com um vértice que chamaremos de C. Como a aresta entre A e B é uma ponte local, A e B não podem ter amigos em comum, logo, a aresta B-C não pode existir. Mas isto contradiz a propriedade do fechamento triádico forte, que afirma que se A-B e A-C são laços fortes, a aresta B-C existe. Esta contradição mostra que uma ponte local não pode ser um laço forte. □

2.3 Fenômenos de Popularidade em Grafos Sociais

A popularidade é um fenômeno que se caracteriza por desequilíbrios extremos: ao passo que alguns indivíduos são conhecidos apenas por seu círculo social imediato, outros alcançam larga visibilidade, e uma minoria muito pequena atinge o reconhecimento global. Este mesmo fenômeno ocorre com livros, filmes e qualquer item que demande uma audiência [Easley e Kleinberg, 2010]. Logo, questiona-se: é possível quantificar a popularidade?

Parece bastante difícil estimar o número de indivíduos do mundo que já ouviram falar de *Barack Obama* ou de *Bill Gates*. Contudo, é simples obter um registro do número de *links* direcionados para páginas famosas, como *Wikipedia* e *Amazon*. O número de *links* de diversas páginas que direcionam o usuário para outra página pode ser utilizado para medir a popularidade da página apontada. A quantificação da popularidade de uma página pode ser vista como uma função de k , representando o número de páginas com k *links* apontando para si. Valores elevados de k indicam grande popularidade [Easley e Kleinberg, 2010].

Compreender os mecanismos que governam o fenômeno da popularidade é importante para o prosseguimento da leitura, visto que o foco principal deste trabalho se concentra em grafos com distribuição de graus do tipo Lei de Potência, explicada com mais detalhes na Subseção 2.3.1.

2.3.1 Distribuição Lei de Potência

Ao medir a distribuição dos *links* na *Web*, no entanto, foi encontrado um comportamento conhecido com Lei de Potência, do inglês *Power Law*. Em diversos estudos da topologia da *Web*, registrados em diferentes momentos da história da *Internet*, a descoberta recorrente foi que o número de páginas apontadas por k *links* é aproximadamente proporcional a $1/k^2$ [Easley e Kleinberg, 2010].

O ponto crucial desta descoberta é que $1/k^2$ decresce de acordo como inverso de um polinômio, então as páginas com muitos *links* direcionados a elas são bem mais comuns do que se esperaria em uma distribuição normal. Por exemplo, $1/k^2$ é igual a $1/1000000$ para $k = 1000$, enquanto uma função decrescente como 2^{-k} apresenta um valor comparativamente menor considerando $k = 1000$. Uma função com este comportamento, ou seja, que decresça pelo inverso de um polinômio é chamada de Lei de Potência [Easley e Kleinberg, 2010].

A partir desta representação, podemos observar uma das colocações feitas anteriormente: a popularidade parece exibir desequilíbrios extremos. E está de acordo com a nossa intuição inicial sobre a *Web*, em que a fração de indivíduos com um determinado grau e popularidade diminui conforme a popularidade aumenta [Easley e Kleinberg, 2010].

Observa-se que outros comportamentos característicos da Lei de Potência, analisados na medida de popularidade também podem ser vistos em outros domínios, por exemplo, a quantidade de números telefônicos que recebe k chamadas por dia é aproximadamente proporcional a $1/k^2$; a quantidade de livros que são comprados por pessoas k é aproximadamente proporcional a $1/k^3$; a fração de artigos científicos que recebem k citações no total é aproximadamente proporcional a $1/k^3$ [Albert e Barabási, 1999], [Newman, 2003].

Na verdade, assim como a distribuição normal é amplamente difundida nas ciências naturais, a distribuição Lei de Potência parece dominar os casos em que a quantidade medida pode ser comparada a algum tipo de popularidade. Portanto, se os dados analisados em determinado estudo sejam por exemplo uma tabela com o número de *downloads* mensais para cada música em uma página famosa de música, é bastante provável que se comportem aproximadamente como uma distribuição Lei de Potência, ou seja, igual a $1/k^c$ para algum c constante. Em caso afirmativo, busca-se estimar o expoente c [Easley e Kleinberg, 2010].

Para descobrir se um conjunto de dados apresenta uma distribuição Lei de Potência existe um método bastante simples: seja $f(k)$ a quantidade de itens que apresentam valor k para uma determinada medida, supondo que se deseja saber se os dados refletem aproximadamente a equação $f(k) = a/k^c$, para algum expoente c e uma constante de proporcionalidade a [Easley e Kleinberg, 2010].

Então, ao escrever $f(k) = ak^{-c}$ e aplicar logaritmos em ambos os lados desta equação, tem-se que $\log(f(k)) = \log a - c \cdot \log k$. Isto representa uma relação Lei de Potência, e o termo $\log f(k)$ é plotado em função de $\log k$. O que se deve visualizar no gráfico é uma linha reta: $-c$ sendo o declive e $\log a$ como interceptação de y . Um gráfico como este fornece uma maneira rápida de verificar se dados apresentam uma distribuição Lei de Potência [Easley e Kleinberg, 2010].

Por exemplo, a Figura 2.7 apresenta a distribuição Lei de Potência de páginas da *Web* apontadas por k *links*. Nota-se que indivíduos mais populares tem menor prevalência dentre a população, o que gera o declive que se aproxima de uma linha reta por boa parte do gráfico. Considerando que fatores incontrolláveis participam da formação da estrutura de *links* da *Web*, questiona-se que processo explicaria esta formação que se aproxima de um declive em linha reta [Easley e Kleinberg, 2010]. O modelo Ligação Preferencial busca explicar este processo e é explicado com mais detalhes na Subseção 4.1.3.

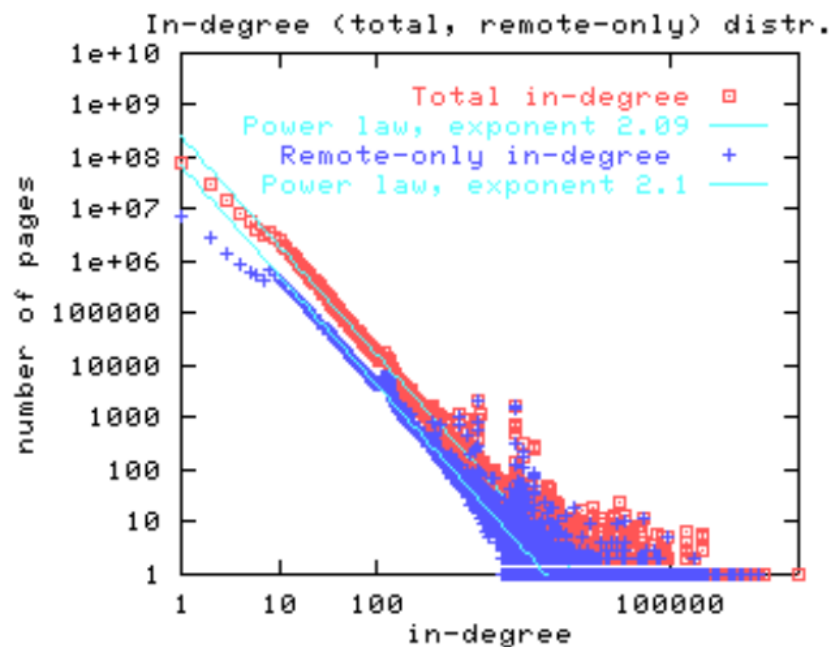


Figura 2.7: Distribuição Lei de Potência

Fonte: [Easley e Kleinberg, 2010]

Capítulo 3

O Problema da Predição de Laços

As redes sociais podem ser vistas como objetos altamente dinâmicos devido à adição e exclusão de arestas, o que implica em alterações em sua estrutura social. O problema da evolução de redes sociais busca compreender este processo [Liben-Nowell, 2005]. De modo mais específico, o problema da predição de laços sociais propõe prever as arestas que serão adicionadas uma rede social em um instante momento t futuro, com base em informações extraídas da mesma rede em um instante anterior t_0 . Logo, questiona-se: o que uma rede social consegue registrar sobre o seu próprio futuro? Em que medida a evolução de uma rede social pode ser modelada usando recursos puramente teóricos da própria rede, como por exemplo sua topologia [Liben-Nowell, 2005]?

Novas arestas de uma rede social se formam por uma variedade de razões, muitas das quais parecem ser completamente independentes da estrutura da rede. No entanto, também se percebe que um grande número de novas amizades são previstas com base na topologia da rede: duas pessoas que são próximas na rede terão amigos em comum, e participam de círculos semelhantes. Essa proximidade sugere que estes indivíduos são mais propensos a se encontrar e se tornam amigos em um futuro próximo [Liben-Nowell, 2005].

A predição da laço é aplicável a uma ampla variedade de áreas. Na *Internet*, pode ser usada em tarefas como a criação automática de *hiperlink*. No *e-commerce*, um dos usos mais importantes de previsão ligação é construir sistemas de recomendação. Também tem aplicações em outras disciplinas científicas. Por exemplo, na bioinformática, tem sido utilizada em interação entre proteínas [Hasan e Zaki, 2011].

Tradicionalmente, a predição de laços em redes sociais se define como a predição da existência de um relacionamento entre dois vértices por meio do cálculo da probabilidade de ocorrência dos laços. Por exemplo, para uma rede G , com um conjunto de laços E , pode ser previsto um conjunto de rótulos de relacionamento \mathcal{L} [Zhang e Yu, 2011].

Os métodos convencionais podem construir, por exemplo, um modelo M , com base nos laços em E , para então aplicar M na previsão do conjunto de suas probabilidades. Em outras palavras, neste exemplo, o modelo M irá mapear laços em \mathcal{L} a rótulos com valores do conjunto $\{-1, 1\}$. Se o laço l for existente, $f_M(l) = 1$, do contrário, $f_M(l) = -1$ [Zhang e Yu, 2011]. Também pode ser construído um modelo M capaz de determinar com probabilidade p a existência de um laço entre dois vértices.

A literatura costuma dividir os métodos de predição de laços sociais em dois grandes grupos: a predição não supervisionada, e a predição supervisionada. Cabe ressaltar que esta nomenclatura é utilizada de maneira diferente dos paradigmas dentro do segmento de aprendizado de máquina. Neste último campo de estudo, os termos *supervisionado* e não supervisionado se

referem ao fato de um algoritmo de classificação receber ou não um “treinamento”, para aprender a mapear os dados de acordo com rótulos já determinados.

3.1 Predição de Laços Não Supervisionada

A predição não supervisionada de laços sociais considera as técnicas que utilizam métricas baseadas na topologia da rede. Dentre estas destacam-se: as métricas baseadas em vizinhança local, métricas baseadas em caminho mais curto e métricas baseadas em passeio aleatório. [Zhang e Yu, 2011]. Estes indicadores também podem ser utilizados como blocos para construir sistemas de predição de laços baseados no paradigma supervisionado. Além das métricas, fazem parte deste grupo os modelos de geração de grafos.

Cabe ressaltar, que devido a razões práticas, utilizam-se grafos para modelar as redes sociais. Desta forma, deve-se entender que grafos sociais são a representação das redes sociais, pois esta estrutura permite análise matemática e computacional adequadas. Portanto, fica estabelecido que a problemática será tratada a partir da perspectiva de grafos, nas demais seções e capítulos deste trabalho.

3.1.1 Métricas Baseadas em Vizinhança Local

Há inúmeras abordagens que propõe a ideia de que se dois vértices u e v possuem vizinhos em comum dentro de seus conjuntos de vizinhos, $\Gamma(u)$ e $\Gamma(v)$, ambos tenderão a formar algum tipo de laço no futuro. Este conceito abstrato pode ser formalizado de acordo com cada indicador apresentado nesta seção.

Ligação Preferencial

A Ligação Preferencial, ou *Preferential Attachment*, parte do princípio que usuários com maior número de conexões sociais são mais propensos a formar novos laços. A Equação (3.1) calcula o produto dos conjuntos de vizinhos dos usuários u e v [Zhang e Yu, 2011].

$$PA(u, v) = |\Gamma(u)| |\Gamma(v)| \quad (3.1)$$

O conceito de ligação preferencial é análogo ao conceito do modelo “rico fica mais rico”, ou modelo ligação preferencial. Em suma, propõe que um vértice se conecta a outros vértices da rede com uma probabilidade baseada em seu conjunto de vizinhos [Hasan e Zaki, 2011].

A métrica Ligação Preferencial foi derivada do Modelo Ligação Preferencial [Hasan e Zaki, 2011], desenvolvido a partir de um argumento heurístico, em oposição a outros modelos de geração de grafos baseados em análises matemáticas rigorosas. Ademais, o valor fornecido pela Equação (3.1) consistirá em um número inteiro, e por si só não indica a probabilidade de formação de laço entre os vértices analisados. Estes fatores motivaram a realização deste trabalho, como veremos com mais detalhes no Capítulo 4.

Vizinhança em Comum

A medida Vizinhança em Comum, em inglês *Common Neighbours* $CN(u, v)$, mostra o número de vizinhos em comum entre os usuários u e v , conforme registra a Equação (3.2) [Zhang e Yu, 2011].

$$CN(u, v) = |\Gamma(u) \cap \Gamma(v)| \quad (3.2)$$

A ideia de usar o número de vizinhos comuns é apenas uma representação da transitividade na rede, ou do conceito de fechamento triádico, visto na Subseção 2.2.4. Em palavras mais simples, isso significa que nas redes sociais, se o vértice u é ligado ao vértice v , e o vértice w é ligado ao vértice v , existe uma probabilidade elevada de que o vértice u também seja ligado ao vértice w . Assim, como o número de vizinhos em comum cresce, a chance de u e w estarem ligados também aumenta [Hasan e Zaki, 2011].

Coeficiente de Jaccard

O coeficiente de Jaccard $JC(u, v)$ é a relação que mede a importância da vizinhança em comum CN . Se o denominador for um valor grande comparado ao numerador, então os dois usuários não apresentam uma relação muito forte, conforme a Equação (3.3) [Zhang e Yu, 2011].

$$JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \quad (3.3)$$

Em outras palavras, esta medida define a probabilidade de um vizinho comum de um par de vértices u e v ser selecionado de forma aleatória a partir da união dos conjuntos de vizinhos u e v . Assim, quanto maior o número de vizinhos comuns, maior o valor do Coeficiente de Jaccard [Hasan e Zaki, 2011].

Índice de Alocação de Recursos

O Índice de Alocação de Recursos, também chamado de *Resource Allocation*, confere a cada par de vértices u e v uma medida baseada no peso $\frac{1}{|\Gamma(w)|}$, que mede a importância de amigos em comum representados por w [Zhang e Yu, 2011], conforme a Equação (3.4).

$$RA(u, v) = \sum_{w \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{|\Gamma(w)|} \quad (3.4)$$

O peso $\frac{1}{|\Gamma(w)|}$ corresponde ao inverso do conjunto de vizinhos do vértice comum w . Isso significa que quanto mais “popular” for o vértice comum w , e em consequência maior o seu conjunto de vizinhos $\Gamma(w)$, menor sua influência no cálculo da métrica.

Medida de Adamic/Adar

Esta medida confere a cada par de vértices u e v uma medida baseada no termo $\frac{1}{\log|\Gamma(w)|}$, que similarmente ao Índice de Alocação de Recursos, mede a importância de amigos em comum representados por w [Zhang e Yu, 2011], conforme a Equação (3.5).

$$AA(u, v) = \sum_{w \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{\log |\Gamma(w)|} \quad (3.5)$$

Nesta métrica, entretanto, o peso do conjunto de vizinhos de w , $\Gamma(w)$, é amenizado pela aplicação do logaritmo.

Coeficiente de Clustering

O fechamento triádico nas redes sociais motivou a formulação de diversas medidas como objetivo de capturar tal conceito. Uma delas é o coeficiente de *clustering*, que pode ser vista como uma medida de vizinhança local. Esta medida é relativa a cada vértice. O coeficiente

de *clustering* de um vértice A é definido como a probabilidade de dois amigos de A, selecionados aleatoriamente, sejam amigos. Em outras palavras, refere-se a quantidade de pares conectados, em que os dois indivíduos são amigos de A [Tang et al., 2012].

Tomando com exemplo a Figura 3.1, o coeficiente de *clustering* do vértice A é $1/6$. Isso ocorre porque existe apenas uma aresta em que os dois vértices são amigos de A, ou seja C-D, dentre todos os pares possíveis formados por vértices ligados a A, B-C, B-D, B-E, C-D, C-E e D-E [Tang et al., 2012].

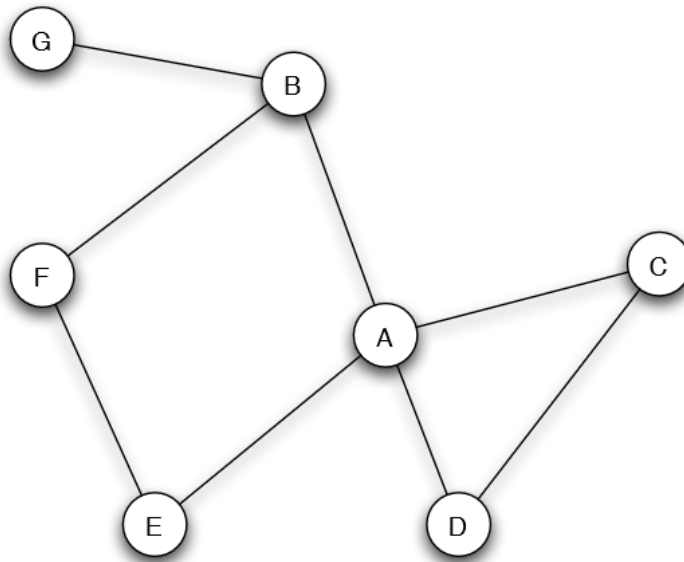


Figura 3.1: Cálculo de coeficiente de *clustering* para o vértice A
Fonte: [Tang et al., 2012]

Na Figura 3.2 o coeficiente de *clustering* do vértice A aumentou para $1/2$, porque nesta situação existem três arestas, B-C, C-D e D-E sobre os mesmos seis pares do caso anterior. Em geral o coeficiente de *clustering* de um vértice varia de zero, o que significa que não há conexão entre os amigos do vértice, a um, quando todos os vértices amigos também são amigos entre si, resultando em um forte fechamento triádico operando na vizinhança do vértice analisado [Tang et al., 2012].

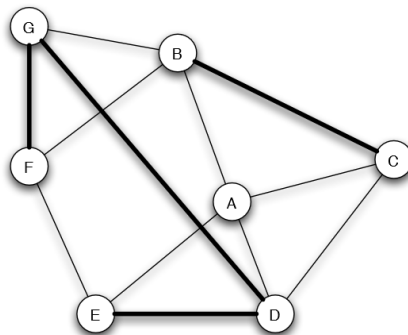


Figura 3.2: Cálculo de coeficiente de *clustering* para o vértice A
Fonte: [Tang et al., 2012]

A Equação (3.6) mostra o cálculo de uma medida de *clustering*.

$$cc(w) = \frac{|\{\{u, v\} \in E \mid u \in \Gamma(w), v \in \Gamma(w)\}|}{\binom{|\Gamma(w)|}{2}} \quad (3.6)$$

3.1.2 Métricas Baseadas em Caminho mais Curto

O fato de os amigos de um dado indivíduo poderem se tornar amigos sugere que a distância entre dois vértices em uma rede social pode influenciar a formação de um laço entre eles. Quanto menor for a distância, mais elevada a possibilidade de um laço se formar [Hasan e Zaki, 2011]. Para representar esta ideia podem ser aplicadas as métricas *Katz*, *Tempo de acerto*, *Tempo de Comutação* e *Rooted Pagerank*. Cabe ressaltar que as três últimas também podem ser consideradas medidas derivadas de passeio aleatório.

Katz

A medida Katz está relacionada ao conceito de caminho mais curto, e geralmente funciona na predição de laços. Este atributo soma diretamente todos os caminhos que existem entre um par de vértices u e v . Para penalizar a contribuição de caminhos mais longos no cálculo de semelhança, amortece exponencialmente a contribuição de um caminho por um factor de β^l , em que l representa o comprimento do percurso. A Equação exata para calcular o valor Katz é mostrada na Equação (3.7) [Hasan e Zaki, 2011].

$$katz(u, v) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{u,v}^{(l)}| \quad (3.7)$$

O termo $|paths_{u,v}^{(l)}|$ representa o conjunto de todos os caminhos de comprimento l de u para v . Katz constitui uma medida robusta, uma vez que é baseado no conjunto de todos os caminhos entre os vértices de u e v . O parâmetro β é usado para normalizar esta medida. Um pequeno valor de β considera somente os caminhos mais curtos, o que faz esta característica se comportar muito como características baseadas na vizinhança. Uma desvantagem desta característica é ser computacionalmente cara, ao considerar todos os vértices da rede [Hasan e Zaki, 2011].

3.1.3 Passeio Aleatório

Passeio aleatório, do inglês *random walk*, é uma ferramenta bastante útil para extração de informações de uma rede, e portanto, para a construção de métricas. Em uma rede é realizada uma trilha definida por passos aleatórios. O início se dá em um determinado vértice, e a cada passo do passeio é escolhida uma dentre as arestas conectadas ao vértice em andamento. Normalmente, é permitido visitar o mesmo vértice mais de uma vez [Newman, 2010].

O passeio aleatório funciona como um método de amostragem da probabilidade de que cada vértice seja visitado. A partir da obtenção destes valores, é possível calcular a probabilidade de determinada interação entre vértices, como por exemplo a formação de laços [Newman, 2010]. É intuitivo concluir que os vértices com maior chance de serem visitados também apresentam maior probabilidade de formar laços no futuro, devido ao seu grau elevado.

Considera-se que um passeio aleatório é iniciado em um vértice específico e é composto por t passos aleatórios. Seja $p_i(t)$ a probabilidade de que o passeio se encontre no vértice i no instante t . Se o passeio estiver no vértice j no tempo $t-1$, a probabilidade de seguir um passo

em direção de uma das k_j arestas ligadas a j é $1/k_j$ [Newman, 2010]. Logo, em um grafo não direcionado, a probabilidade $p_i(t)$ é dada pela Equação (3.8).

$$p_i(t) = \sum_j \frac{A_{ij}}{k_j} p_j(t-1) \quad (3.8)$$

Sabe-se que a matriz de adjacência da rede em questão é representada por \mathbf{A} , e o elemento A_{ij} é igual a um no caso de existir aresta entre os vértices i e j , e caso contrário igual a zero. O termo k_j representa o grau do vértice j . Dessa forma, tem-se que a probabilidade $p_i(t)$ é na verdade o somatório das probabilidades de que o passeio esteja em um vértice j ligado ao vértice i , multiplicada pela existência ou não desta aresta (A_{ij}), e inversamente proporcional ao grau do vértice j .

Cabe ressaltar que o valor da probabilidade em $p_i(t)$ é obtido por amostragem, após sucessivos passeios pela rede. É possível realizar o mesmo cálculo na forma matricial, isto é, para toda a rede [Newman, 2010]. Nesse caso, adota-se a Equação (3.9).

$$\mathbf{p}(t) = \mathbf{A}\mathbf{D}^{-1}\mathbf{p}(t-1) \quad (3.9)$$

em que \mathbf{p} é o vetor de elementos p_i , e \mathbf{D} a matriz contendo os graus dos vértices em sua diagonal:

$$D = \begin{bmatrix} k_1 & 0 & 0 & \dots \\ 0 & k_2 & 0 & \dots \\ 0 & 0 & k_3 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Tomando como exemplo a rede da Figura 3.3, obtém-se a matriz de adjacência \mathbf{A} e a matriz inversa com a diagonal de graus \mathbf{D}^{-1} :

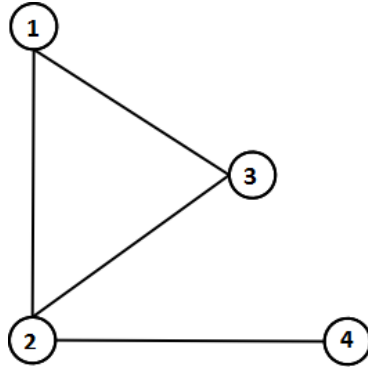


Figura 3.3: Rede utilizada para exemplo de passeio aleatório

Fonte: Autoria própria

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad D^{-1} = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Multiplicando os dois termos, é obtido o fator $\mathbf{A}\mathbf{D}^{-1}$, que será multiplicado pelo vetor de probabilidades $\mathbf{p}(t-1)$, obtido por amostragem. Este processo da obtenção de probabilidades por amostragem é conduzido diversas vezes, porque os valores obtidos a cada rodada podem apresentar variações, e não refletir a realidade.

Esta matriz de resultados pode ser interpretada da seguinte forma: a soma dos elementos de cada linha resulta em um valor indicativo de chance de o vértice ser visitado. Tomando como exemplo o vértice 2, tem-se que a soma dos elementos da segunda linha da matriz resulta no valor dois, soma mais elevada do que o valor $1/3$ correspondente ao vértice quatro.

$$AD^{-1} = \begin{bmatrix} 0 & 1/3 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 1 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \end{bmatrix}$$

3.1.4 Métricas de Proximidade Baseadas em Passeio Aleatório

Uma vez obtido o vetor \mathbf{p} de probabilidades, em que p_{ij} representa a probabilidade de uma visita ao vértice j ocorrer, partindo do vértice i , podem ser calculadas medidas de proximidade, como *tempo de acerto* e *tempo de comutação* [Zhang e Yu, 2013].

Tempo de Acerto

O tempo de acerto, ou *Hitting Time*, conta o número esperado de passos necessários para um passeio aleatório chegar do vértice u ao v . Sabendo que o termo $p_{u,v}$ foi definido na seção anterior, temos a Equação (3.10) [Zhang e Yu, 2011].

$$HT(u, v) = 1 + \sum_{w \in \Gamma(u)} p_{u,w} HT(w, v) \quad (3.10)$$

Um valor pequeno desta medida indica que os vértices são semelhantes uns aos outros, e tem uma chance maior de se ligarem no futuro. O benefício desta métrica é ser fácil de calcular por meio da realização de alguns passeios aleatórios [Hasan e Zaki, 2011].

Tempo de Comutação

O tempo de comutação, também conhecido com *Comute Time*, conta o número esperado de passos necessários para acessar o vértice u a partir de v , e no caminho de volta de v para u . Esta medida pode ser obtida com a Equação (3.11) [Zhang e Yu, 2011].

$$CT(u, v) = HT(u, v) + HT(v, u) \quad (3.11)$$

Uma vez que o tempo de acerto não é uma medida simétrica para grafos não direcionados, ou seja, os valores de $HT(u, v)$ e $HT(v, u)$ não são iguais, o tempo de comutação pode ser usado [Hasan e Zaki, 2011].

Rooted Pagerank

Esta medida é utilizada para o *ranking* das páginas da *Web*, e tem relação intrínseca com o tempo de acerto. Além disso, pode ser usada como um recurso para a predição de laços sociais. No entanto, uma vez que *Pagerank* [Page et al., 1999] é um atributo de um único vértice, necessita ser modificado de modo a representar uma similaridade entre um par de vértices x e y [Hasan e Zaki, 2011].

Considerando uma probabilidade α fixa, um usuário salta de uma página para outra página aleatória com probabilidade α e segue um *hiperlink* vinculado com probabilidade $1 - \alpha$. Para este passeio aleatório, a importância de uma página v é a soma esperada da importância de

todas as páginas u que apontam para v . Na terminologia do passeio aleatório, pode-se substituir o termo importância pelo termo distribuição estacionária [Hasan e Zaki, 2011].

Para a predição de laço, o passeio aleatório do *Pagerank* original pode ser alterado da seguinte forma: a métrica de semelhança entre dois vértices x e y pode ser representada como a probabilidade estacionária de y em um passeio aleatório que retorna a x com probabilidade $1 - \beta$ em cada passo, movendo-se a um vizinho aleatório com probabilidade β . Esta métrica é assimétrica e pode se tornar simétrica pela soma de um complemento, em que o papel de x e y são invertidos [Hasan e Zaki, 2011].

3.2 Predição de Laços Supervisionada

A predição supervisionada é o segundo paradigma de predição de laços sociais. Este modelo normalmente envolve técnicas de aprendizado de máquina como a extração de características e algoritmos de classificação [Zhang e Yu, 2011]. Um algoritmo de aprendizagem normalmente extrai uma métrica a partir de pares de vértices, e se baseia nestas pontuações de similaridade para prever a formação de uma aresta entre dois vértices [Hasan e Zaki, 2011].

3.2.1 Extração de Características

A extração de *características* consiste em uma etapa importante dentro do paradigma de aprendizado de máquina. No contexto de aprendizado de máquina, *características* se referem na verdade ao que se utiliza como entrada para os algoritmos de classificação. São medidas capazes de representar um determinado conjunto de dados [Guyon et al., 2008].

A extração de características é muito particular a cada domínio de conhecimento [Guyon et al., 2008]. No estudo de redes sociais há uma variedade muito ampla de informação que pode ser convertida em *características*. A partir de informação de uma rede social de usuários, é possível extrair diferentes tipos de *características*, como por exemplo as métricas como o coeficiente de Jaccard e a medida de Adamic [Zhang e Yu, 2011].

As características baseadas na topologia do grafo são as mais naturais para a predição de laço. De fato, muitos trabalhos se concentraram na predição de laço baseada nas características topológicas do grafo [Liben-Nowell e Kleinberg, 2007, Zhang e Yu, 2013]. Tipicamente, buscam calcular a semelhança com base na vizinhança do vértice ou com base nos conjuntos de caminhos entre um par de vértices. A vantagem destas características ocorre por serem genéricas e aplicáveis a grafos de qualquer domínio. Assim, pouco conhecimento da área específica é necessário para calcular os valores dessas características. No entanto, para grandes redes sociais, o cálculo de algumas destas características pode ser dispendiosa [Hasan e Zaki, 2011].

Devido à grande diversidade, não se pretende descrever todos os tipos de *características* possíveis neste trabalho.

3.2.2 Algoritmos de Classificação

Existe uma infinidade de modelos de classificação para a aprendizagem supervisionada, são bons exemplos máquina de vetores de suporte e redes neurais artificiais. Embora seus desempenhos sejam comparáveis, alguns métodos funcionam melhor do que outros para um determinado conjunto de dados ou domínio específico. No entanto, aplicar um sistema de classificação à tarefa de predição de laços apresenta alguns desafios, tornando alguns modelos mais atraentes do que outros [Hasan e Zaki, 2011].

Comumente, utiliza-se a palavra *classificação* para definir o ato de rotular itens diversos em categorias. Na área médica, por exemplo, quando se trata de uma determinada doença, os indivíduos são separados em grupos de doentes e saudáveis [Sammut e Webb, 2011].

Considerando o contexto de aprendizado de máquina, uma classe agrupa instâncias de um conjunto com características em comum. Um algoritmo de classificação pode ser alimentado com exemplos de uma ou mais classes, com seus respectivos rótulos. O algoritmo produz como saída um classificador que mapeia as propriedades dos exemplos fornecidos em rótulos [Sammut e Webb, 2011].

A tarefa de rotular laços sociais em redes sociais pode ser realizada de diversas maneiras, tais como positivos ou negativos, existentes ou não. Por exemplo, dado um laço social direcionado (u, v) , caso o vértice u não goste de v , então $y(u, v) = -1$, caso contrário, $y(u, v) = 1$ [Zhang e Yu, 2011].

3.3 Métricas de Avaliação

Uma vez que um sistema já foi implementado, normalmente segue-se a etapa de testes, que pode se basear em métricas de avaliação. Para avaliar a performance de sistemas de predição podem ser aplicadas métricas comumente utilizadas para medir o desempenho de sistemas classificadores. Esta seção inclui indicadores bastante utilizados para sistemas pertencentes aos grupos de predição supervisionada e não supervisionada.

3.3.1 Tabela de Confusão

A tabela de confusão, também chamada de tabela de contingência, é uma ferramenta originária da estatística médica capaz de representar, dentro de uma população, os indivíduos doentes, indivíduos sadios, indivíduos com teste positivo e indivíduos com teste negativo, conforme a Tabela 3.1 [Massad et al., 2004].

Tabela 3.1: Tabela de Confusão

	Positivo no exame	Negativo no exame
Doentes	VP	FN
Sadios	FP	VN

A partir da tabela de contingência, podem ser observados os seguintes índices:

1. Verdadeiro positivo (VP): doente com resultado positivo;
2. Falso positivo (FP): sadio com resultado positivo;
3. Falso negativo (FN): doente com resultado negativo;
4. Verdadeiro negativo (VN): sadio com resultado negativo.

Para realizar uma análise deste tipo, é preciso que os dados estejam validados. Isso quer dizer que se deve conhecer antecipadamente os rótulos de cada instância que o sistema se propuser a classificar.

Como este projeto pretende estudar sistemas de predição de laços sociais, os laços previstos corretamente que de fato existem são tratados como verdadeiros positivos. Laços

rotulados como existentes mas não verificados na realidade são falsos positivos. Quando um laço existente não é previsto pelo sistema, considera-se então como um falso negativo. Por fim, temos que os laços não existentes e não previstos são verdadeiros negativos.

Os índices da tabela de contingência são utilizados para o cálculo de outros índices, que incluem taxa de acerto, precisão, entre outros.

3.3.2 Taxa de Acerto

Taxa de Acerto corresponde a capacidade de um sistema identificar os verdadeiros positivos e negativos dentre todas as amostras [Zhang e Yu, 2011]. A fórmula para o cálculo da taxa de acerto é dada na Equação (3.12). O sistema apresenta mais acurácia quanto mais próximo o valor obtido se aproximar de um.

$$\text{Taxa de Acerto} = \frac{VP + VN}{VP + VN + FP + FN} \quad (3.12)$$

3.3.3 Precisão

Precisão corresponde ao número de positivos corretamente classificados dividido pelo número total de instâncias classificadas como positivo [Zhang e Yu, 2011]. É calculada pela Equação (3.13).

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (3.13)$$

3.3.4 Sensibilidade

O índice *sensibilidade* corresponde ao número de indivíduos corretamente classificados como verdadeiro positivo dividido pelo número total de positivos [Zhang e Yu, 2011]. Seu cálculo é obtido pela Equação (3.14).

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (3.14)$$

3.3.5 Especificidade

O índice *especificidade* cede a capacidade do sistema em identificar os indivíduos negativos entre os verdadeiramente não doentes, representada pela Equação (3.15). Assim, quanto menor o número de falsos positivos identificados pelo sistema, maior ser a sua *especificidade* [Massad et al., 2004].

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (3.15)$$

3.3.6 Curva ROC (*Receiver Operator Characteristic*)

A Curva ROC é útil para visualização da performance de classificadores. Seu uso vem crescendo em pesquisas nas áreas de *machine learning* e *data mining* [Fawcett, 2006].

Os valores dos índices estatísticos sensibilidade e especificidade são organizados em pares ordenados, e dispostos em um plano cartesiano. Com algum método de interpolação, são interligados desde a origem (sensibilidade = 0 e especificidade = 1), até o extremo oposto do

gráfico (sensibilidade = 1 e especificidade = 0), dando origem a curva ROC [Massad et al., 2004], conforme ilustra a Figura 3.4.

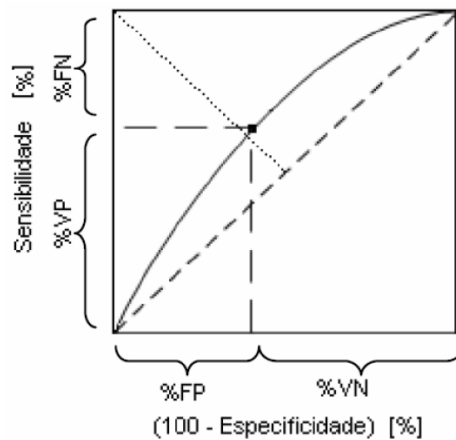


Figura 3.4: Curva ROC
Fonte: [Massad et al., 2004]

No eixo das ordenadas da curva ROC pode ser observada a fração dos verdadeiros positivos (sensibilidade). E no eixo das abcissas, a fração dos falsos positivos, correspondente ao complemento da especificidade (1 - especificidade) [Massad et al., 2004].

A linha tracejada representa um sistema classificador incapaz de discriminar, em outras palavras, um sistema que apresenta o mesmo percentual de verdadeiros positivos e de falsos positivos. Quanto mais afastada da linha tracejada estiver a curva ROC (de modo semelhante ao da Figura 3.4), melhor é o desempenho do sistema. O classificador ideal possui sensibilidade = 1 e especificidade = 1 [Brown e Davis, 2006].

Capítulo 4

Análise de Modelos de Geração de Grafos e Proposta de Métrica

Estudos recentes têm evidenciado o crescente interesse em redes reais de larga escala, como por exemplo, as várias referências apresentadas por [Liben-Nowell e Kleinberg, 2007], [Zhang e Yu, 2013] e [Leskovec et al., 2005]. Muitas pesquisas se concentram em modelar propriedades de redes reais por meio da geração de grafos aleatórios. Rapidamente se observou que os modelos clássicos não eram capazes de capturar a natureza dessas redes, e reproduzir, por exemplo, uma distribuição Lei de Potência para os graus, então novos modelos foram propostos. Os trabalhos realizados neste campo normalmente recaem nas seguintes categorias [Bollobás e Riordan, 2005]:

1. Estudos realizados diretamente nas redes, medindo propriedades como diâmetro, distribuição de grau, *clustering*, entre outras.
2. Criação de novos modelos de geração de grafos aleatórios motivada pelas características percebidas nos estudos do item 1.
3. Simulações computacionais dos modelos de geração de grafo, com análise das propriedades obtidas.
4. Análises heurísticas de novos modelos de modo a prever as propriedades dos grafos gerados.
5. Estudos matemáticos rigorosos de novos modelos, buscando provar os teoremas relativos a suas propriedades.

A contribuição deste trabalho consiste em propor uma métrica para predição de arestas para grafos sociais com distribuição Lei de Potência. Esta métrica foi derivada do modelo de geração de grafos de Vignatti e da Silva [Vignatti e da Silva, 2015], por sua vez, derivado do Modelo GRG (*Generalized Random Graphs*) ou Grafo Aleatório Generalizado [Hofstad, 2009]. Os modelos de geração de grafo não podem ser considerados soluções diretas para o problema da predição de laços. Seu objetivo é capturar comportamentos observados em grafos que representam redes sociais reais, e descrevê-los de maneira matemática, como descreve o item 2.

Devido ao fato de que a métrica proposta foi derivada da análise de um modelo de geração de grafos, faz-se necessária a análise do modelo em questão, e de outros modelos que contribuíram para sua elaboração. Para prosseguir com a leitura deste capítulo é sugerida a compreensão dos modelos analisados, por meio da Figura 4.1. A finalidade de todos os modelos

apresentados é a mesma: produzir grafos. No entanto há modelos que permitem que a distribuição de grau seja configurada, e modelos que são especificações da distribuição Lei de Potência, gerando apenas grafos com este comportamento.

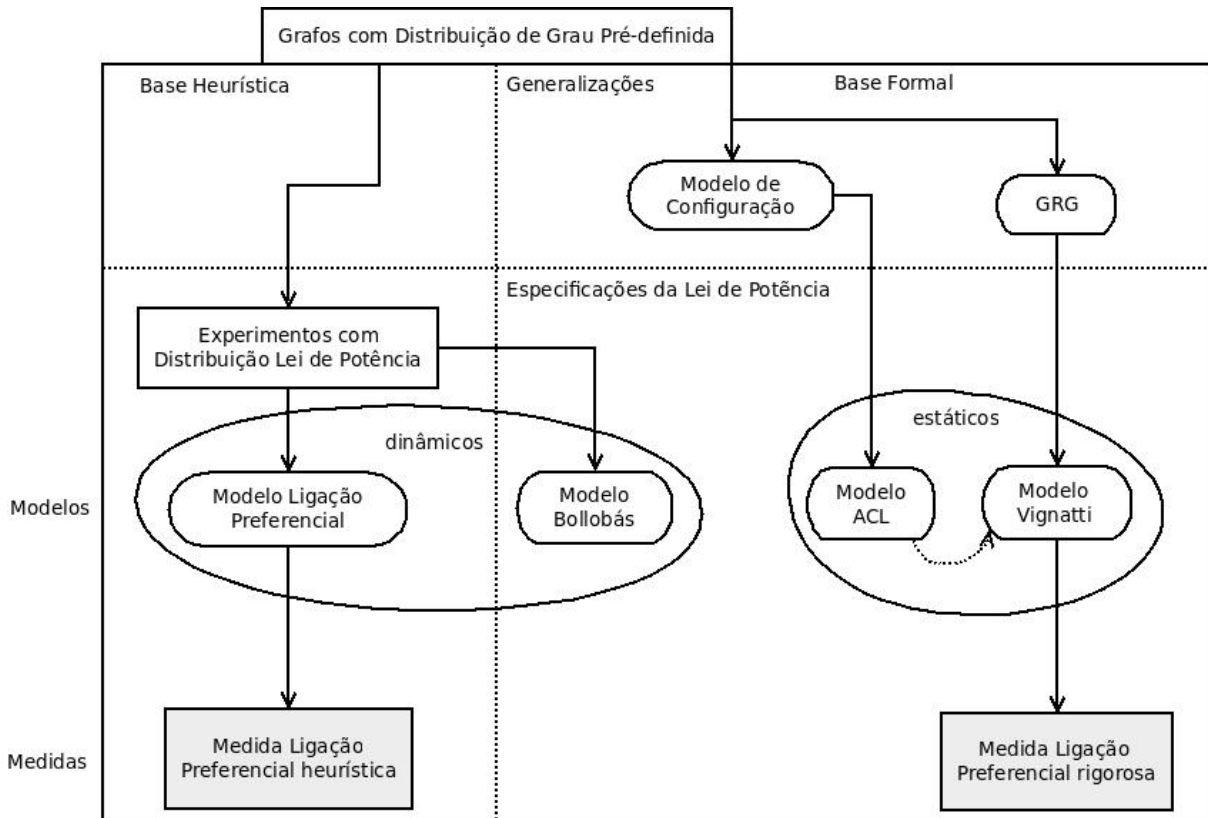


Figura 4.1: Visão Geral das Pesquisas sobre Formação de Laços Sociais

Fonte: Autoria Própria

A partir do estudo de grafos com distribuição de grau pré-definida, inicialmente cabe destacar os modelos de geração de grafo Modelo de Configuração [Newman, 2010] e o Modelo Grafo Aleatório Generalizado. Nestes modelos, a distribuição de grau de vértices que o grafo irá apresentar pode ser configurada, reproduzindo, por exemplo, a distribuição Lei de Potência, ainda que não sejam especificações para esta distribuição. Ambos os modelos foram desenvolvidos para permitir análises matemáticas adequadas.

Paralelamente, estudos foram desenvolvidos na busca pela compreensão do comportamento dos grafos [Albert e Barabási, 1999], concentrando-se na distribuição dos graus de vértices. Observou-se que diversas redes apresentaram um comportamento que foi denominado de Lei de Potência, ou mecanismo “rico fica mais rico” [Tang et al., 2012]. Este comportamento inspirou o desenvolvimento do Modelo Ligação Preferencial, por [Albert e Barabási, 1999]. Este modelo reproduz o comportamento “rico fica mais rico” nos grafos gerados. Contudo, por ser baseado em argumentos heurísticos, e desenvolvido por meio de uma análise contínua este modelo não apresenta uma base matemática adequada, por isso se encaixa no setor de “base heurística” da Figura 4.1.

O Modelo de Bollobás e Riordan foi o primeiro modelo baseado no conceito de ligação preferencial construído com rigor matemático [Bollobás e Riordan, 2005]. Similarmente, outros dois modelos foram desenvolvidos com base matemática adequada buscando capturar a distribuição Lei de Potência: o modelo ACL (*Aiello, Chung e Lu*) e o modelo de Vignatti e da Silva, inseridos no quadrante direito da Figura 4.1, assim como o modelo de Bollobás e Riordan.

O Modelo ACL [Aiello et al., 2001], cujo nome advém de seus autores, Aiello, Chung e Lu, consiste em uma especialização do modelo de Configuração aplicada à distribuição Lei de Potência. Outra especialização de um modelo não especializado corresponde ao Modelo de Vignatti e da Silva, derivado do modelo Grafo Aleatório Generalizado. O trabalho de [Vignatti e da Silva, 2015] também busca estabelecer a equivalência entre os modelos ACL e Grafo Aleatório Generalizado com distribuição Lei de Potência.

A análise de grafos que representam redes reais, bem como o desenvolvimento de modelos que possam recriá-los, trouxeram inspiração para elaboração das métricas de topologia. Os conceitos capturados nestes estudos estão diretamente relacionados ao processo de formação de arestas entre vértices. Como exemplo, pode-se citar a métrica Ligação Preferencial, utilizada na previsão de laços sociais e derivada do Modelo Ligação Preferencial [Hasan e Zaki, 2011].

A métrica proposta como contribuição foi derivada do modelo de Vignatti e da Silva, um modelo de geração de grafos construído com uma base matemática adequada. Este modelo consiste em objeto de análise e inspiração para a nova métrica capaz de prever com maior carga de informação a formação de laços entre vértices, como será visto em detalhes na Seção 4.2.

4.1 Análise de Modelos de Geração de Grafos

A análise de cada modelo faz parte da proposta, e será descrita nas subseções a seguir. Esta análise pode ser considerada parte essencial da contribuição deste trabalho. Pois a métrica proposta é derivada do modelo de Vignatti e da Silva, derivado do modelo GRG e de conceitos dos demais modelos apresentados, como se pode citar o mecanismo Lei de Potência.

4.1.1 Modelo de Configuração

O modelo de geração aleatória de grafos com distribuição determinada mais estudado é o modelo de configuração, em inglês *Configuration Model*. O modelo de configuração é, na verdade, um modelo de grafo aleatório com uma distribuição de grau configurável. Isto é, o grau exato de cada vértice na rede é fixo, em vez de simplesmente se utilizar uma distribuição de probabilidade para os graus [Newman, 2010].

Supondo que seja especificada a notação w_i que representa o grau que cada vértice $i = 1 \dots n$ na rede deve apresentar, pode-se criar uma rede aleatória com estes graus como descrito a seguir. Dá-se a cada vértice i um total de w_i “metades de arestas”, como representado na Figura 4.2. Existem $2m$ pontas de arestas, sendo que m é o número total de arestas. Em seguida, escolhe-se duas metades de aresta de maneira uniformemente aleatória, e é criada uma aresta completa, ligando-as uma ao outra, como indicado pela linha a tracejado na Figura 4.2. Então, escolhe-se um outro par do conjunto de $2m$ “metades de arestas” restantes que são conectadas, e assim por diante até que todas as metades de aresta sejam utilizadas. O resultado final é uma rede na qual cada vértice tem exatamente o grau desejado [Newman, 2010].

O processo acima gera cada par pela junção de “metades de aresta” com igual probabilidade. Tecnicamente, o modelo de configuração é definido como o conjunto em que cada par combinando com a distribuição de grau escolhida aparece com a mesma probabilidade. A distribuição uniforme dos emparelhamentos no modelo de configuração tem a importante consequência de que qualquer metade de aresta da rede de configuração é igualmente susceptível de ser ligada a qualquer outra [Newman, 2010].

Há algumas considerações acerca do processo descrito. Primeiro, deve haver um número par de metades de aresta no total, para não sobrar nenhuma metade de aresta. Isto significa que a soma dos graus deve ser feita até um número par. Uma segunda restrição é que a rede pode

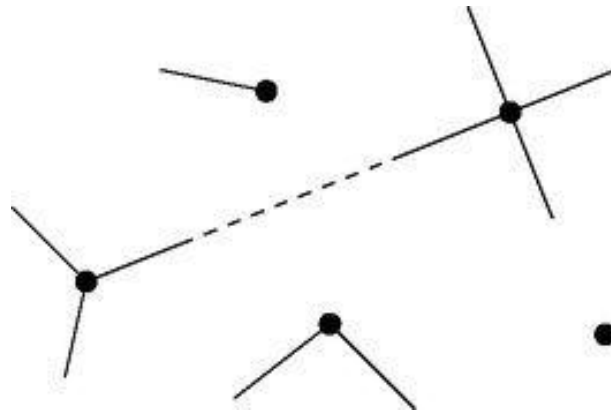


Figura 4.2: Vértices com arestas incompletas
Fonte: [Newman, 2010]

conter *loops* (arestas que saem e voltam para o mesmo vértice) ou arestas múltiplas, ou ambos. Questiona-se se isso poderia ser evitado, rejeitando a criação de quaisquer arestas durante o processo, mas verifica-se que esta não é uma boa ideia. Uma rede assim gerada não é projetada de maneira uniforme a partir do conjunto de emparelhamentos possíveis, o que significa que as propriedades do modelo já não podem ser calculadas analiticamente, pelo menos, por quaisquer meios atualmente conhecidos [Newman, 2010].

Na prática, portanto, faz mais sentido permitir a criação de arestas múltiplas e *loops*. Apesar de algumas redes do mundo real apresentarem *loops* e arestas múltiplas, a maioria não o faz. E isto, até certo ponto, torna o modelo de configuração menos satisfatório como um modelo de rede. No entanto, a média do número de arestas múltiplas e *loops* e no modelo de configuração é uma constante com o crescimento da rede, o que significa que a densidade de arestas múltiplas e *loops* tende para zero neste limite [Newman, 2010].

A principal conclusão é que o grafo aleatório resultante é uniforme dentro o conjunto de todos os grafos possíveis com a sequência de grau especificada. A construção pelo modelo de configuração também nos permite calcular quantos grafos podem existir, dentro das definições iniciais, com uma sequência grau específica, sendo, portanto, um belo exemplo do método probabilístico [Hofstad, 2009].

4.1.2 Modelo Grafo Aleatório Generalizado

Um modelo bastante conhecido é o modelo Grafo Aleatório Generalizado, do inglês *Generalized Random Graph* (GRG). Neste modelo, as arestas permanecem presentes de modo independente, sendo que o valor de suas probabilidades é variável. A motivação para a construção deste modelo nasceu da necessidade de construir um grafo aleatório com grande variabilidade nos graus de vértices, de modo que pudesse apresenta distribuições como a Lei de Potência, por exemplo [Hofstad, 2009].

Neste modelo, cada vértice $i \in [n]$ recebe um peso w_i . Este peso pode ser visto como a propensão do vértice a ter arestas, e tenderá a ser bastante próximo do grau esperado do vértice. Dados os pesos, as arestas estão presentes de forma independente, mas suas probabilidades de ocupação não são idênticas, e sim controladas pelos pesos dos vértices. Naturalmente, isto pode ser feito de várias maneiras diferentes [Hofstad, 2009].

No Grafo Aleatório Generalizado, a probabilidade de ocorrência uma aresta entre os vértices i e j , sendo que $i \neq j$, corresponde à Equação (4.1):

$$p_{ij} = \frac{w_i w_j}{l_n + w_i w_j} \quad (4.1)$$

em que $\mathbf{w} = (w_i)_{i \in [n]}$ são os pesos dos vértices, e l_n equivale à soma dos pesos de todos os vértices obtida por meio da Equação (4.2):

$$l_n = \sum_{i \in [n]} w_i \quad (4.2)$$

Assim sendo, a ausência de homogeneidade do modelo é projetada por meio de pesos dos vértices. A incidência de arestas ligadas a vértices com maior peso é maior, aumentando os graus de vértices com pesos elevados, em comparação ao grau de vértices com pesos menores. Quando os pesos são escolhidos de forma adequada, isto pode dar origem a grafos aleatórios com graus de vértice bastante variados [Hofstad, 2009].

Denota-se o grafo resultante por $GRG_n(\mathbf{w})$. Em muitos casos, os pesos realmente dependem de n , o que seria mais apropriado, mas também mais incômodo, escrever os pesos de $\mathbf{w}^n = (w_i^{(n)})_{i \in [n]}$. Para manter a notação simples, não será utilizada a dependência de n [Hofstad, 2009].

Sem perda de generalidade, assume-se que $w_i > 0$. Nota-se que, para um determinado $i \in [n]$, $w_i = 0$, o vértice i será isolado com probabilidade 1, de que i seja removido do grafo. Naturalmente, a topologia do grafo aleatório generalizado depende da escolha dos pesos dos vértices $\mathbf{w}^n = (w_i^{(n)})_{i \in [n]}$.

4.1.3 Modelo Ligação Preferencial

As ideias que fornecem a base para a ocorrência de uma distribuição Lei de Potência nascem a partir do modelo Ligação Preferencial, também chamado de modelo “rico fica mais rico”. É realmente um tema de pesquisa aberto e muito interessante obter um modelo satisfatório que descreva a distribuição Lei de Potência a partir de modelos simples de tomada de decisões individuais. Esta seção trata do modelo “rico fica mais rico” com base em [Easley e Kleinberg, 2010], que optou por explicá-lo por meio do conceito de páginas da *Web*. Contudo, este modelo pode ser aplicado na geração de grafos de outras naturezas.

O modelo é construído a partir das consequências observáveis da tomada de decisão na presença de grupos: supõe-se simplesmente que as pessoas têm uma tendência a copiar as decisões das pessoas de seu círculo. Com base nessa ideia, segue-se um modelo simples para a criação de *links* entre páginas da *Web*.

1. Páginas são criadas e chamadas de $1, 2, 3, \dots, N$.
2. Quando uma página j é criada, produz um *link* para outra página já existente de acordo com a seguinte regra probabilística (controlada por um único número p variável de zero e um):
 - (a) com probabilidade p , a página j escolhe uma página i aleatoriamente entre todas as páginas disponíveis, e cria um *link* para esta página i .
 - (b) com probabilidade $1 - p$, a página j escolhe uma página i uniformemente aleatoriamente entre todas as páginas disponíveis, e cria um *link* para a página que i aponta.

- (c) criação de um único *link* da página j . Este processo pode se repetir para criar múltiplos *links*, gerados de forma independente a partir da página j . No entanto, para manter as coisas simples, supõe-se que cada página cria apenas um link de saída.

A parte chave do processo consiste em após encontrar uma página disponível aleatória i na população, o autor da página j não cria um *link* para i . Em vez disso, copia a decisão tomada pelo autor da página i , e cria *link* apontando mesma página que o *link* de i .

O principal resultado deste modelo é que se executado em muitas páginas, a quantidade de páginas com k *links* apontados para si serão distribuídas aproximadamente de acordo com a Lei de Potência. Logo, o valor do expoente c depende da escolha de p . Esta dependência segue uma direção intuitivamente natural: como p fica menor, de modo que a cópia se torna mais frequente, o expoente c fica menor também, aumentando a probabilidade de se ver páginas extremamente populares.

Provar este resultado exige uma análise mais complexa, desenvolvida no Apêndice A, com base na ideia de números contínuos. Contudo, é útil trabalhar com algumas das ideias informais subjacentes a esta análise. Em primeiro lugar, o mecanismo de cópia é realmente uma implementação do processo “rico fica mais rico”: quando se copia a decisão de uma página anterior aleatória, a probabilidade de que seja criado um *link* para uma página l é diretamente proporcional ao número total de páginas atualmente ligados a l .

Por que este fenômeno é chamado de “rico fica mais rico”? Uma vez que a probabilidade da página l experimenta um aumento, este é diretamente proporcional à popularidade atual. Esta descrição corresponde ao modelo ligação preferencial, no sentido que as ligações feitas “preferencialmente” ocorrem para páginas que já tem alta popularidade. Da mesma forma, quanto mais popular um indivíduo, mais provável que se escute seu nome surgir em uma conversa, e, portanto, o mais provável que ele se torne popular para os participantes daquela conversa que não o conheciam.

A intuição por trás da análise é a seguinte: com o mecanismo “rico fica mais rico”, o modelo proposto prevê que a popularidade deve crescer de acordo com uma taxa proporcional a seu valor atual, e, portanto, exponencialmente com o tempo. Uma página que recebe uma pequena vantagem sobre outras tende a estender esta vantagem. A natureza deste mecanismo de copiar amplifica os efeitos de grandes valores, tornando-os ainda maior.

Como acontece com qualquer modelo simples, o objetivo não é capturar todas as razões por que as pessoas criam *links* na *Web*, ou em qualquer outra rede, mas mostrar que um princípio simples e muito natural por trás da criação *link* justificaria a distribuição Lei de Potência.

Por exemplo, as populações das cidades parecem seguir uma distribuição Lei de Potência: a quantidade de cidades com população k é aproximadamente $1/k^c$ para diversas constantes c . Considerando que as cidades são formadas em épocas diferentes, e que, uma vez formadas, crescem em proporção ao seu tamanho atual simplesmente como resultado de as pessoas terem filhos, então temos quase exatamente o mesmo modelo “rico fica mais rico”. Portanto, não devemos ficar surpresos ao ver a Lei de Potência esteja presente na realidade.

Encontrar leis que descrevam o mecanismo de popularidade das páginas *Web* é bastante intrigante. No entanto, enxergar este fenômeno como resultado do processo “rico fica mais rico” faz bastante sentido. Mais uma vez, deve-se salientar que este modelo apenas se aproxima da realidade e não é completamente aceito, visto que se trata de uma análise contínua, desenvolvida em detalhes no Apêndice A, quando a natureza real do problema exige uma análise discreta; e por ser baseado em um argumento heurístico. Além disso, no passo 2 do algoritmo fica claro que o modelo não é capaz de solucionar a situação em que não existe nenhuma outra página existente além da que está sendo criada. Outras linhas de pesquisa também buscam explicar o comportamento característico de Lei de Potência, como discutido na Subseção 4.1.4.

4.1.4 Modelo de Bollobás e Riordan

O modelo apresentado por Bollobás e Riordan foi o primeiro modelo baseado no conceito de ligação preferencial analisado rigorosamente [Bollobás e Riordan, 2005]. Considerando um conjunto de vértices v_1, v_2, \dots , tem-se que a notação w_s representa o grau de um vértice v_s no grafo G . Por indução, é definido então o processo de grafo aleatório $(G_1^t)_{t \geq 0}$, de modo que G_1^t seja um grafo em $\{v_i : 1 \leq i \leq t\}$ onde: inicialmente há um grafo vazio $G_1^{(0)}$ sem vértices, ou um grafo $G_1^{(1)}$ com apenas um vértice e uma aresta. Dado $G_1^{(t-1)}$, o grafo $G_1^{(t)}$ é formado por meio da adição do vértice v_t e de uma única aresta entre v_t e v_i , sendo i escolhido aleatoriamente de acordo com a Equação (4.3) [Bollobás e Riordan, 2005].

$$P(i = s) = \begin{cases} w_s / (2t - 1) & 1 \leq s \leq t - 1 \\ 1 / (2t - 1) & s = t \end{cases} \quad (4.3)$$

Esta equação é regida pelo princípio da Ligação Preferencial, pois uma aresta e nascente em v_t se liga ao vértice v_i , com probabilidade de escolha do vértice v_i proporcional ao seu grau naquele instante. A aresta e é contabilizada como se já estivesse contribuindo para o grau de v_t . Para $m > 1$, são adicionadas m arestas a partir de v_t , um de cada vez, contabilizando no cálculo dos graus as arestas já existentes e a metade solta da aresta sendo adicionada. Esta regra produz a seguinte definição equivalente: o processo $(G_m^t)_{t \geq 0}$ é obtido pela execução do processo (G_1^t) na sequência v'_1, v'_2, \dots , resultando no grafo G_m^t a partir de $G_1^{(mt)}$ por meio da identificação dos vértices v'_1, v'_2, \dots, v'_m , formando v_1 , e dos vértices $v'_{m+1}, v'_{m+2}, \dots, v'_{2m}$, formando v_2 , e assim por diante. [Bollobás e Riordan, 2005].

Considerando o vértice $v_i = i$, o grafo G_m^t apresenta $[t] = \{1, 2, \dots, t\}$. As arestas de G_m^t possuem uma orientação natural, que nasce nos vértices mais antigos e se ligam aos mais recentes, de modo que ij é orientado de i para j se $i > j$. Entretanto, no trabalho em questão, o grafo foi tratado como não orientado [Bollobás e Riordan, 2005].

Para ilustrar este modelo, consideremos o instante $t = 7$, em que o grafo já possui sete vértices conforme a Figura 4.3. Neste instante a distribuição das probabilidades dos vértices corresponde a $P(i = 1) = 3/13$, $P(i = 2) = 3/13$, $P(i = 3) = 3/13$, $P(i = 4) = 1/13$, $P(i = 5) = 1/13$, $P(i = 6) = 1/13$ e $P(i = 7) = 1/13$. Mais especificamente, para as arestas em que $i < 7$, o cálculo da probabilidade obedece à primeira linha da equação 4.3, $P(i = s) = w_s / (2t - 1)$. O vértice recém adicionado v_7 tem sua probabilidade calculada por meio da segunda linha da Equação (4.3), $P(i = s) = 1 / (2t - 1)$, que leva em consideração a metade de aresta incompleta.

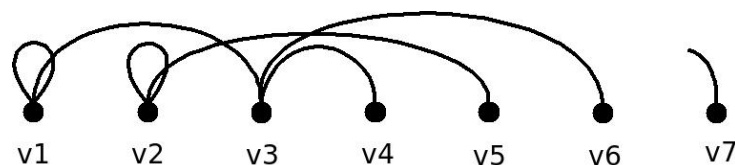


Figura 4.3: Exemplo de geração de grafo pelo modelo de Bollobás
Fonte: Autoria Própria

Além de satisfazer o critério básico matemático de ser especificado precisamente, o processo G_m^t apresenta diversas propriedades úteis. Uma delas é a definição de G_m^t em termos de $G_1^{(mt)}$, um termo muito mais simples, e dessa forma provar resultados por meio deste. Outra propriedade importante se deve ao fato de que enquanto G_1^t é dinâmico, a distribuição do grafo obtido $G_1^{(n)}$ em um dado instante $t = n$ apresenta pode apresentar um modelo estático, o modelo

de Bollobás [Bollobás e Riordan, 2005], cujo estudo detalhado não está entre os objetivos desta pesquisa.

4.1.5 Modelo ACL – Aiello, Chung e Lu

Este modelo de grafo aleatório foi baseado na distribuição de graus Lei de Potência e desenvolvido por [Aiello et al., 2001], apresentando dois parâmetros apenas. Os parâmetros utilizados apenas delinham genericamente o tamanho e a densidade do grafo analisado. Contudo, tais medidas são convenientes para descrever o comportamento característico da Lei de Potência em um grafo.

O modelo de grafo aleatório baseado na distribuição Lei de Potência $P(\alpha, \beta)$ considera $y(x)$ como sendo o número de vértices com grau x . O modelo $P(\alpha, \beta)$ atribui probabilidade uniforme a todos os grafos em que $y(x) = e^\alpha / x^\beta$. Portanto, o modelo depende de dois parâmetros de entrada α e β [Aiello et al., 2001]. Aplicando log em ambos os lados, detemos a Equação (4.4):

$$\log y(x) = \alpha - \beta \log x \quad (4.4)$$

A grosso modo, α corresponde ao logaritmo do tamanho do grafo e β pode ser descrito como a taxa de crescimento do grafo em escala *log-log*, parâmetro que se relaciona também com a densidade do grafo. Percebe-se que o número de arestas deve ser um número inteiro, mais precisamente, o termo da Equação 4.4 deve ser arredondada para $\lfloor \frac{e^\alpha}{x^\beta} \rfloor$. Se forem utilizados números reais, no lugar dos inteiros arredondados para baixo, poderão surgir termos de erro durante o processo de computação. No entanto, tais erros podem ser facilmente contornados, conforme explicado em [Aiello et al., 2001]. Por simplicidade e conveniência, serão utilizados números reais, considerando-os como partes de um número inteiro [Aiello et al., 2001].

Outra restrição é a de que a soma dos graus deve ser par. Isto pode ser garantido com a soma de um vértice com grau um, se a soma inicial for ímpar. Para simplificar ainda mais, também se assume que não existem vértices isolados. Ademais, podem ser deduzidos os seguintes fatos a respeito do grafo:

1. O grau máximo do grafo corresponde a $e^{\frac{\alpha}{\beta}}$, considerando que $0 \leq \log y = \alpha - \beta \log x$.
2. O número de vértices n pode ser calculado conforme a Equação (4.5):

$$n = \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{x^\beta} \approx \begin{cases} \zeta(\beta)e^\alpha & \text{se } \beta > 1 \\ \alpha e^\alpha & \text{se } \beta = 1 \end{cases} \quad (4.5)$$

em que $\zeta(t) = \sum_{n=1}^{\infty} \frac{1}{n^t}$ corresponde a função de Zeta de Riemann. Esta função converge para uma constante quando o valor de $\beta > 1$.

3. O número de arestas E é calculado a partir da Equação (4.6)

$$E = \frac{1}{2} \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x \frac{e^\alpha}{x^\beta} \approx \begin{cases} \frac{1}{2} \zeta(\beta - 1) e^\alpha & \text{se } \beta > 2 \\ \frac{1}{4} \alpha e^\alpha & \text{se } \beta = 2 \end{cases} \quad (4.6)$$

Para a imensa maioria dos casos, tem-se que $\beta > 2$, logo serão utilizadas as equações correspondentes a este valor de β . Na etapa de geração do grafo, utiliza-se o modelo de configuração.

4.1.6 Modelo de Vignatti e da Silva

O modelo de Vignatti e da Silva associa um peso w_i a cada vértice i pertencente a um conjunto de vértices $V = \{1, 2, \dots, |V|\}$. Os pesos dos vértices $w_1, \dots, w_{|V|}$ formarão um vetor w , que será diretamente responsável pela distribuição do grau. O modelo também leva em conta o conjunto de arestas E , para o qual cada aresta ij é criada independentemente com probabilidade $P_{ij} = \frac{w_i w_j}{\ell_n + w_i w_j}$, em que $\ell_n = \sum_{k \in V} w_k$ [Vignatti e da Silva, 2015].

Para avançarmos com a correspondência do modelo, faz-se necessário a prova do Lema 4.1.1 [Vignatti e da Silva, 2015]. A demonstração do 4.1.1 se encontra no Apêndice B.

Lema 4.1.1. Sejam $w_i, w_j \in \{1, \dots, e^{\frac{\alpha}{\beta}}\}$, então $e^\alpha \zeta(\beta - 1) + w_i w_j \approx e^\alpha \zeta(\beta - 1)$

Da definição de P_{ij} e do Lema 4.1.1, pode-se escrever que

$$P_{ij} = \frac{w_i w_j}{e^\alpha \zeta(\beta - 1) + w_i w_j} \approx \frac{w_i w_j}{e^\alpha \zeta(\beta - 1)} \quad (4.7)$$

Na literatura, a notação p_{ij} se refere à probabilidade de uma aresta surgir entre os vértices i e j . Para a correspondência do modelo apresentada nesta seção, esta probabilidade corresponde à Equação (4.8) [Vignatti e da Silva, 2015].

$$p_{ij} = \frac{w_i w_j}{e^\alpha \zeta(\beta - 1)} \quad (4.8)$$

Considerando que o termo $e^\alpha \zeta(\beta - 1)$ corresponde à soma dos graus do grafo, ou ℓ_n , também pode-se escrever a Equação (4.8) como a Equação (4.9) [Vignatti e da Silva, 2015].

$$p_{ij} = \frac{w_i w_j}{2 \cdot |E|} \quad (4.9)$$

4.2 Proposta de Métrica

A métrica *Ligação Preferencial* heurística foi derivada do modelo *Ligação Preferencial* [Hasan e Zaki, 2011]. Como foi visto na Subseção 4.1.3, o modelo *Ligação Preferencial* apresenta uma base matemática inadequada. Este modelo foi desenvolvido por meio de uma análise contínua, quando a natureza do problema exige uma análise baseada em números discretos. Além disso, o modelo foi construído partir de um argumento heurístico.

Em resposta a estas inconsistências da métrica *Ligação Preferencial* heurística, sugere-se a utilização da Equação (4.8) como uma métrica para predição de laços sociais. Este é o núcleo da contribuição do trabalho, resultado da etapa de análise dos modelos de geração de grafos.

A Equação (4.8) calcula a probabilidade de formação de aresta entre dois vértices i e j , e corresponde a multiplicação de seus graus, representados pela notação w_i e w_j sobre o termo $2 \cdot |E|$, que se refere à soma dos graus de todos os vértices.

Para ilustrar a utilização da métrica, considera-se o grafo da figura 4.4. Sabe-se que os graus dos vértices são $w_1 = 3$, $w_2 = 1$, $w_3 = 3$, $w_4 = 2$ e $w_5 = 3$, resultando em uma soma de graus igual a seis. Logo, calcula-se p_{13} , obtendo a pontuação $9/12$, ou $0,75$.

Em consequência de sua análise matemática adequada, as bases da métrica sugerida suplantam a inadequação da base matemática do modelo *Ligação Preferencial*. Outra vantagem da métrica sugerida é o fato de seu resultado fornecer um valor que reflete por si só a probabilidade de formação de laço entre o par de vértices analisados. Ao passo que a métrica ligação referencial fornece como resultado valores absolutos, tais como 50, 100 e 200. E exige, portanto, uma

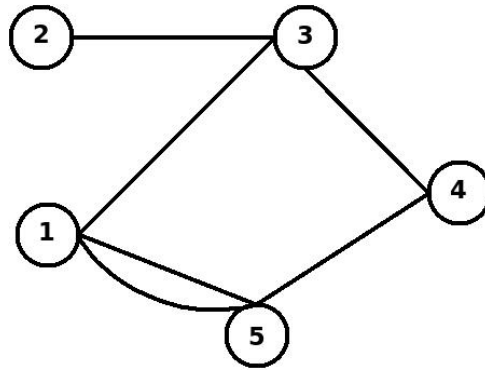


Figura 4.4: Exemplo de aplicação da métrica sugerida sobre um grafo

Fonte: Autoria Própria

comparação das pontuações obtidas por todos os pares de vértices do grafo analisado para que se possa categorizá-las como elevadas ou reduzidas.

O pseudocódigo do cálculo da métrica proposta é apresentado no Algoritmo 1. A métrica proposta também apresenta complexidade de tempo $O(m)$. Contudo, observa-se que o valor de m será equivalente a $|E|$, ou seja, ao número de arestas do grafo G .

Algoritmo 1 CÁLCULO DA MÉTRICA PROPOSTA

Entrada: i, j, G

Saída: Probabilidade de formação de aresta

início

$w_i \leftarrow$ número de arestas de i
 $w_j \leftarrow$ número de arestas de j
 $m \leftarrow$ número de arestas de do grafo G
 $p(i, j) = w_i \times w_j$
 $p(i, j) = p(i, j)/m$

fin

retorna $p(i, j)$

Por sua vez, algoritmo da métrica Ligação Preferencial heurística, descrito pela Equação (3.1), conforme pode ser observado no Algoritmo 2 apresenta complexidade de tempo $O(m)$.

Algoritmo 2 CÁLCULO DA MÉTRICA LIGAÇÃO PREFERENCIAL

Entrada: u, v

Saída: Métrica de formação de aresta

início

$\Gamma(u) = 0$
 $\Gamma(v) = 0$
 $PA(u, v) = 0$
 $\Gamma(u) \leftarrow$ número de vizinhos de u
 $\Gamma(v) \leftarrow$ número de vizinhos de v
 $PA(u, v) = \Gamma(u) \times \Gamma(v)$

fin

retorna $PA(u, v)$

Outra diferença evidente entre as duas métricas é que a métrica Ligação Preferencial heurística trabalha com os valores de $\Gamma(i)$ e $\Gamma(j)$, que representam o conjunto de vizinhos dos vértices i e j . Ao passo que a métrica proposta utiliza em seus cálculos os pesos w_i e w_j , que representam os graus dos vértices i e j . Em casos em que não existem arestas múltiplas ou *loops*, o valor de $\Gamma(i)$ e w_i serão os iguais para um mesmo vértice i . Do contrário, os valores serão diferentes, contudo, deverão ser próximos.

A métrica proposta exige a contabilização de todas as arestas de um grafo G para o qual se deseja obter a probabilidade de arestas i e j a ele pertencentes. Isto pode parecer um esforço adicional em relação à métrica ligação preferencial original. Contudo, visto que seu resultado por si só é capaz de refletir a probabilidade de formação de um laço, a métrica sugerida evita uma comparação das pontuações obtidas por todos os pares para obtenção de um referencial. Ademais, com a utilização de grafos de pequena escala, ou grafos em que já se conhece o valor de $|E|$, ou seja, o número total de arestas, o cálculo da métrica sugerida pode se tornar tão simples quanto o da métrica original.

4.3 Análise de um Experimento Relacionado

Para ilustrar a avaliação de desempenho de métricas de predição de laços, esta seção dedica-se a análise de uma pesquisa com esta temática. O estudo desenvolvido por [Liben-Nowell e Kleinberg, 2007] buscou formalizar o problema da predição de laços sociais e desenvolveu abordagens para o problema utilizando medidas referentes à topologia da rede.

Um grafo não direcionado $G = \langle V, E \rangle$ representa a estrutura topológica de uma rede social em que cada aresta $e = \langle v, e \rangle \in E$ apresenta uma interação entre u e v que ocorre em um dado instante $t(e)$. Para dois instantes t e $t' > t$, temos que $G = [t, t']$ denota um subgrafo de G , que consiste em todas as arestas dentro dos intervalos de tempo t e t' . Sejam t_0, t_0', t_1 e t_1' quatro instantes de tempo, e que $t_0 < t_0' \leq t_1 < t_1'$. Logo, a predição de laço funciona da seguinte forma: a partir de uma rede $G = [t_0, t_0']$ é obtida a lista de arestas ausentes em $G = [t_0, t_0']$ que estão previstas a aparecer na rede $G = [t_1, t_1']$. O intervalo $[t_0, t_0']$ é denominado intervalo de treinamento e $[t_1, t_1']$ intervalo de teste. Ao fim disso, avaliamos o quão acuradamente as novas arestas foram previstas.

O experimento utilizou cinco seções da base de dados acadêmica *arXiv*: **astro-ph**, **cond-mat**, **gr-qc**, **hep-ph** e **hep-th**. Considerando qualquer um destes cinco grafos, foram definidos os períodos de [1994,1996] e [1997,1999] como intervalos de treinamento e de teste.

4.3.1 Métodos Aplicados

Foram analisados uma série de métodos para predição de laços. Todos eles designam a cada par de vértices x e y do grafo G um peso $score(x,y)$. Os métodos podem ser classificados de três maneiras distintas: métodos baseados em vizinhança, métodos baseados em caminho e métodos de alto nível. Dada a importância de cada conjunto, serão trabalhados e separado.

1. *Métodos baseados em Vizinhança*: considerando um vértice x , seu grau é denotado por $\Gamma(x)$ e contabiliza o seu número de vizinhos. Existem métodos que se baseiam na ideia de que dois vértices x e y são mais propensos a formar laços no futuro se apresentarem muitos amigos em comum, dentre eles foram utilizados no experimento Vizinhança em Comum, Coeficiente de Jaccard, Adamic/Adar e Ligação Preferencial.

2. *Métodos baseados em Caminho*: determinados métodos buscam considerar o conjunto de todos os caminhos entre dois vértices no processo de predição de laços, como Katz, Tempo de Acerto, SimRank.
3. *Métodos de Alto Nível*: levam em consideração todo o grafo, e são exemplos deste tipo de método Aproximação de Nível Baixo, Bigramas Inéditos e *Clustering*.

4.3.2 Resultados Obtidos

A tabela mostra os resultados obtidos pelos diferentes métodos na predição de laços. A primeira linha indica os resultados obtidos por um preditor aleatório, que simplesmente seleciona um par de autores que não interagiram no intervalo de treinamento. A partir dos resultados da tabela, é possível observar que as medidas Katz e Adamic/Adar obtiveram acima da média para a maioria dos casos.

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-two pairs)		9.6	25.3	21.4	12.2	29.2
common neighbors		18.0	41.1	27.2	27.0	47.2
preferential attachment		4.7	6.1	7.6	15.2	7.5
Adamic/Adar		16.8	54.8	30.1	33.3	50.5
Jaccard		16.4	42.3	19.9	27.7	41.7
SimRank	$\gamma = 0.8$	14.6	39.3	22.8	26.1	41.7
hitting time		6.5	23.8	25.0	3.8	13.4
hitting time—normed by stationary distribution		5.3	23.8	11.0	11.3	21.3
commute time		5.2	15.5	33.1	17.1	23.4
commute time—normed by stationary distribution		5.3	16.1	11.0	11.3	16.3
rooted PageRank	$\alpha = 0.01$	10.8	28.0	33.1	18.7	29.2
	$\alpha = 0.05$	13.8	39.9	35.3	24.6	41.3
	$\alpha = 0.15$	16.6	41.1	27.2	27.6	42.6
	$\alpha = 0.30$	17.1	42.3	25.0	29.9	46.8
	$\alpha = 0.50$	16.8	41.1	24.3	30.7	46.8
Katz (weighted)	$\beta = 0.05$	3.0	21.4	19.9	2.4	12.9
	$\beta = 0.005$	13.4	54.8	30.1	24.0	52.2
	$\beta = 0.0005$	14.5	54.2	30.1	32.6	51.8
Katz (unweighted)	$\beta = 0.05$	10.9	41.7	37.5	18.7	48.0
	$\beta = 0.005$	16.8	41.7	37.5	24.2	49.7
	$\beta = 0.0005$	16.8	41.7	37.5	24.9	49.7

Figura 4.5: Comparativo das métricas para predição de laços

Fonte: [Liben-Nowell e Kleinberg, 2007]

A performance obtida puramente pelos indicadores foi relativamente baixa. Não houve uma técnica vencedora, mas os resultados obtidos sugerem que há de fato informação importante contida na topologia da rede.

Além da pesquisa de [Liben-Nowell e Kleinberg, 2007], o trabalho de [Zhang e Yu, 2013] também utiliza métricas de topologia como Vizinhança Comum, Coeficiente de *Jaccard* e Medida de *Adamic/Adar* para treinar um sistema classificador. Constrói medidas a partir do conteúdo textual das postagens de usuários nas redes, que podem revelar similaridade nos hábitos de vocabulário de dois usuários. Supondo que dois usuários sempre falam sobre comida no *Twitter*, há maior probabilidade de que estes dois indivíduos tenham *hobbies* em comum, e por conseguinte venham a se tornar amigos. As palavras utilizadas pelos

usuários u_i e u_j são então organizadas em dois vetores de conjuntos de palavras $w(u_i)$ e $w(u_j)$ com pesos atribuídos por TF-IDF ¹.

¹Frequência do termo inverso da frequência nos documentos, é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos.

Capítulo 5

Conclusão

A grande importância do estudo de redes sociais reside no fato de que ao esclarecer os mecanismos de funcionamento e evolução das grandes redes, nascem inspirações para outras aplicações, desde estratégias de *marketing* a previsão de parâmetros em redes sociais. A análise de grafos sociais confere descobertas importantes relativas ao comportamento humano, e da formação de laços entre indivíduos.

Apesar de muito já ter sido feito na pesquisa sobre formação de laços sociais, há ainda um imenso potencial a ser explorado. Pode ser observado em alguns trabalhos a elaboração de uma análise com base matemática inadequada, além de outras inconsistências. Estes fatores abrem espaço para que inúmeras outras pesquisas sejam elaboradas no sentido de aperfeiçoar estas falhas.

O estudo desenvolvido compreendeu uma visão geral do problema da predição de laços sociais, e uma revisão da bibliografia, com uma análise de modelos de grafos Lei de Potência, resultando na proposta de uma nova métrica para predição de laços em grafos Lei de Potência. O trabalho realizado se concentrou na temática da predição de laços, e apresentou como contribuição uma métrica específica para este evento. Contudo, conclui-se que muito pode ser extraído da topologia de um grafo social, o que inclui informações relacionadas a outros fenômenos.

A análise do tema trouxe como resultado diversas indagações, ainda não exploradas por pesquisadores, e ainda merecedoras de investigação. Seria possível criar um modelo de geração de grafos que incorpore, por exemplo, o conceito de pontes locais? Ou ainda, outros conceitos apresentados na fundamentação teórica, como o conceito de redes sociais heterogêneas? O fato de o modelo ser específico para uma rede heterogênea exigiria um tratamento matemático diferente dos modelos apresentados?

Estes questionamentos podem evoluir de maneira ainda mais profunda: seriam os modelos capazes de incorporar fenômenos mais abstratos, como por exemplo, a homofilia? A homofilia se refere à propensão de indivíduos se tornarem amigos considerando semelhanças da ordem de idade, profissão, etnia, etc. Considerando esta hipótese, de que maneira poderia ser mensurado um fenômeno como este?

Cabe ressaltar também que, embora os modelos de geração de grafo apresentem o mesmo resultado, alguns modelos são mais adequados em determinadas situações. O modelo Ligação Preferencial e o modelo de Bollobás e Riordan são incrementais. Estes modelos tem seu início a partir de apenas um vértice, e vão construindo o grafo por meio de iterações em que são adicionados novos vértices e arestas. Este cenário é ideal para situações em que não se conhece de antemão a estrutura desejada do grafo a ser produzido. Pois esta pode ser alterada de maneira dinâmica, ao longo da execução do algoritmo.

Os outros modelos, como o modelo de Configuração, o modelo GRG, e os modelos ACL e de Vignatti e da Silva são modelos estáticos. A distribuição de grau dos vértices deve ser especificada antes da execução do modelo. Desta forma, estes modelos são mais apropriados para um cenário estático, em que já se conhece a distribuição de grau dos grafos.

A análise teórica permitiu uma compreensão global a respeito do tema, o que se converteu na métrica proposta como contribuição. Porém, o principal resultado alcançado com a realização deste trabalho foi o conhecimento adquirido, inspirando inúmeras novas possibilidades a serem trilhadas.

Referências Bibliográficas

- [Aiello et al., 2001] Aiello, W., Chung, F. e Lu, L. (2001). A random graph model for power law graphs. *Experimental Math*, 10(1):53–66.
- [Albert e Barabási, 1999] Albert, R. e Barabási, A. L. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- [Albert e Barabási, 2002] Albert, R. e Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.
- [Bollobás e Riordan, 2005] Bollobás, B. e Riordan, O. M. (2005). Mathematical results on scale-free random graphs. páginas 1–37.
- [Brown e Davis, 2006] Brown, C. D. e Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80(1):24–38.
- [Easley e Kleinberg, 2010] Easley, D. e Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.
- [Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.
- [Guyon et al., 2008] Guyon, I., Gunn, S., Nikravesh, M. e Zadeh, L. (2008). *Feature extraction: foundations and applications*, volume 207. Springer.
- [Hasan e Zaki, 2011] Hasan, M. A. e Zaki, M. J. (2011). Chapter 1 link prediction in social networks link. <http://cs.iupui.edu/~alhasan/papers/SNDA11.pdf>. Acessado em 20/12/2015.
- [Hofstad, 2009] Hofstad, R. V. D. (2009). Random graphs and complex networks. <https://www.win.tue.nl/~rhofstad/NotesRGCN.pdf>. Acessado em 20/12/2015.
- [Leskovec et al., 2005] Leskovec, J., Kleinberg, J. e Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. páginas 177–187.
- [Li et al., 2013] Li, M., Zou, H., Guan, S., Gong, X., Li, K., Di, Z. e Lai, C. H. (2013). A coevolving model based on preferential triadic closure for social media networks. *CoRR*, 3(2512):24–38.
- [Liben-Nowell, 2005] Liben-Nowell, D. (2005). *An Algorithmic Approach to Social Networks*. Tese de doutorado, Massachusetts Institute of Technology, Massachusetts - US.
- [Liben-Nowell e Kleinberg, 2007] Liben-Nowell, D. e Kleinberg, J. (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):1019–1031.

- [Massad et al., 2004] Massad, E., Ortega, N. e Silveira, P. (2004). *Métodos quantitativos em medicina*. Manole.
- [Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.
- [Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.
- [Page et al., 1999] Page, L., Brin, S., Motwani, R. e Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>. Acessado em 20/12/2015.
- [Sammut e Webb, 2011] Sammut, C. e Webb, G. I. (2011). *Encyclopedia of Machine Learning*. Springer
- [Tang et al., 2012] Tang, J., Lou, T. e Kleinberg, J. (2012). Inferring social ties across heterogeneous networks. páginas 743–752.
- [Vignatti e da Silva, 2015] Vignatti, A. L. e da Silva, M. V. G. (2016). Minimum vertex cover in generalized random graphs with power law degree distribution. páginas 1–4.
- [Zhang e Yu, 2011] Zhang, J. e Yu, P. S. (2011). Link prediction across heterogeneous social networks: A survey. https://www.cs.uic.edu/~jzhang2/files/2014_survey_paper.pdf. Acessado em 20/12/2015.
- [Zhang e Yu, 2013] Zhang, J. e Yu, X. K. P. S. (2013). Predicting social links for new users across aligned heterogeneous social networks. páginas 7–10.

Apêndice A

Análise Avançada do Modelo Ligação Preferencial

Considera-se que uma quantidade de vértices com k *links* entrantes se comporta de acordo com a distribuição Lei de Potência. A distribuição Lei de Potência é descrita pela equação $1/k^c$, em que c depende de características dos vértices do modelo. Nesta subseção será construído um argumento heurístico capaz de analisar o comportamento do modelo e identificar as razões porque ocorre esta distribuição. Além disso, será analisada a maneira como expoente c se relaciona com características básicas do modelo. Esta análise é baseada em simples equações diferenciais que governam o crescimento exponencial presente nos cálculos introdutórios.

Esta análise aproximada foi desenvolvida por [Easley e Kleinberg, 2010], e fornece um argumento heurístico do comportamento característico de Lei de Potência, que foi posteriormente verificado por uma análise mais rigorosa do modelo probabilístico completo [Bollobás e Riordan, 2005].

A descrição do modelo ligação preferencial apresentada a seguir difere da descrição do Capítulo 4 por se tratar de uma análise baseada na ideia de números contínuos. Contudo, assim como na descrição do Capítulo 4, o exemplo para ilustração é composto pelas seguintes etapas:

1. Páginas são criadas ordenadamente e nomeadas como 1, 2, 3, ..., N .
2. Quando a página j é criada, é produzido um *link* para uma página *Web* já existente de acordo com uma regra probabilística, descrita por um número p entre 0 e 1.
 - (a) Com probabilidade p , a página j seleciona uma página i de maneira aleatória uniforme dentre as páginas já existentes, e cria um *link* para esta página.
 - (b) Com probabilidade $1 - p$, a página j seleciona uma página L com probabilidade proporcional ao número de *links* entrantes em L , e cria um *link* para esta página.
 - (c) Este processo descreve a criação de um único *link* a partir da página j . Pode ser repetido para criar múltiplos *links*, a partir da página j . Contudo, para fins de simplificação, será considerado que cada página gera apenas um *link* para outras páginas.

Há uma questão puramente probabilística: o processo especificado funciona para N passos, observando-se que as N páginas são criadas uma de cada vez, e pode ser determinado o número esperado de páginas com k *links* entrantes ao final do processo.

Aproximação Determinística do Modelo Ligação Preferencial

Inicialmente, temos que o número de *links* entrantes em um vértice j em um instante $t \geq j$, que é uma variável aleatória $X_j(t)$. Podemos então destacar duas propriedades de $X_j(t)$.

1. *Condição Inicial.* O vértice j inicia sem *links* entrantes ao ser criado no instante j , portanto $X_j(j) = 0$.
2. *A mudança esperada de X_j ao longo do tempo.* O vértice j recebe um *link* entrante no instante $t + 1$ somente se o *link* nascente do vértice recém criado $t + 1$ apontar para o vértice j .

Com probabilidade p , o vértice $t + 1$ cria *links* para vértices já existentes de maneira uniformemente aleatória. Enquanto que com a probabilidade complementar $1 - p$, o vértice cria *links* para vértice já existente com uma probabilidade proporcional ao número de *links* entrantes do vértice a ser apontado pelo novo *link*.

No caso anterior, o vértice $t + 1$ cria um *link* para o vértice j com probabilidade $1/t$. No momento em que o vértice $t + 1$ é criado, o número total de *links* na rede corresponde a t (cada um partindo do vértice anterior), e dentre estes, $X_j(t)$ apontam para o vértice j . Logo, o vértice $t + 1$ se liga ao vértice j com probabilidade $X_j(t)/t$. Portanto, a probabilidade total de que o vértice $t + 1$ se ligue ao vértice j corresponde à Equação (A.1).

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t} \quad (\text{A.1})$$

Faz-se necessário atentar para a natureza contínua desta descrição, o que faz dela uma aproximação, considerando que a natureza real exige uma análise discreta. O objetivo de tal aproximação consiste em analisar um processo mais simples do modelo ligação preferencial, no qual é mais fácil evidenciar a Lei de Potência. Novamente, apesar de os dois modelos não se comportarem exatamente da mesma forma, as similaridades entre eles oferecem evidências que podem ser confirmadas no modelo original.

A ideia central na formulação do modelo é construir um modelo determinístico, ou seja, um modelo sem probabilidades, em que todas as variáveis evoluem de maneira fixa ao longo do tempo, como um sistema físico ideal, descrito por equações de movimento a partir de determinadas condições iniciais. Para isso, considera-se o tempo contínuo de 0 a N , ao invés de passos discretos 1,2,3, . . . , e aproxima-se $X_j(t)$ número de *links* entrantes no vértice j de uma função contínua $x_j(t)$. A função x_j é caracterizada por duas propriedades que buscam se aproximar das condições iniciais e esperadas ao longo do tempo, com descritas em $X_j(t)$. As duas propriedades da função x_j são as seguintes:

1. *Condição Inicial.* Considerando $X_j(j) = 0$. Define-se $x_j(j) = 0$.
2. *Equação de Crescimento.* Considerando que quando o vértice $t + 1$ é criado, o número de *links* entrantes em j cresce com probabilidade descrita na Equação (A.2).

$$\frac{p}{t} + \frac{(1-p)x_j(t)}{t} \quad (\text{A.2})$$

Na aproximação determinística fornecida pela função x_j , modela-se a taxa de crescimento por meio da Equação diferencial (A.3).

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j(t)}{t} \quad (\text{A.3})$$

Utilizando equações diferenciais, foi descrito o comportamento de x_j , a aproximação determinística do número de *links* entrantes no vértice j ao longo do tempo. Em vez de trabalhar com variáveis aleatórias $X_j(t)$ que se movem em saltos probabilísticos em pontos discretos no tempo, trabalharemos com um valor de x_j que cresce de maneira suave ao longo do tempo, a uma taxa ajustada para corresponder as mudanças esperadas nas variáveis aleatórias correspondentes.

Resolução da Aproximação Determinística

Inicialmente é resolvida a Equação diferencial (A.3) que descreve x_j . Para fins de simplificação, considera-se $q = 1 - p$, transformando a Equação diferencial em (A.4).

$$\frac{dx_j}{dt} = \frac{p + qx_j(t)}{t} \quad (\text{A.4})$$

Dividindo os dois lados por $p + qx_j$, obtém-se (A.4)

$$\frac{1}{p + qx_j} \frac{dx_j}{dt} = \frac{1}{t} \quad (\text{A.5})$$

em seguida, integra-se ambos os lados

$$\int \frac{1}{p + qx_j} \frac{dx_j}{dt} dt = \int \frac{1}{t} dt \quad (\text{A.6})$$

e é encontrada a Equação (A.7).

$$\ln(p + qx_j) = q \ln(t) + c \quad (\text{A.7})$$

A partir da exponenciação e considerando $A = e^c$, obtém-se

$$p + qx_j = At^q \quad (\text{A.8})$$

logo, conclui-se (A.9).

$$x_j(t) = \frac{1}{q}(At^q - p) \quad (\text{A.9})$$

O valor da constante A pode ser determinado usando uma condição inicial $x_j(j) = 0$. Esta condição fornece a Equação (A.9).

$$0 = x_j(j) = \frac{1}{q}(Aj^q - p) \quad (\text{A.10})$$

A partir de conclui-se que $A = p/j^q$. Juntando este valor de A e a Equação (A.9), é obtida a Equação (A.11).

$$x_j(t) = \frac{1}{q} \left(\frac{p}{j^q} \cdot t^q - p \right) = \frac{p}{q} \left[\left(\frac{t}{j} \right)^q - 1 \right] \quad (\text{A.11})$$

Identificação da Distribuição Lei de Potência na Aproximação Determinística

A Equação (A.11) é uma etapa significativa da aproximação, considerando que obtém uma expressão descritiva do crescimento de x_j ao longo do tempo. Ou seja, descreve a quantidade de todos os vértices com pelo menos k *links* entrantes para um dado valor de k em um instante t . O valor de x_j se aproxima do número de *links* entrantes do vértice j . Analogamente a esta

expressão, o modelo simplificado se baseia na seguinte questão: para um dado valor de k , em um instante t , que fração de todas as funções x_j satisfazem $x_j(t) \geq k$?

Utilizando a Equação (A.11), encontra-se a desigualdade (A.12)

$$x_j(t) = \frac{p}{q} \left[\left(\frac{t}{j} \right)^q - 1 \right] \geq k \quad (\text{A.12})$$

reescrevendo em termos de j , tem-se que

$$j \leq t \left[\frac{q}{p} \cdot k + 1 \right]^{\frac{-1}{q}} \quad (\text{A.13})$$

Das funções x_1, x_2, \dots, x_t no instante t , a quantidade de valores j que satisfazem a desigualdade corresponde a

$$\frac{1}{t} \cdot t \left[\frac{q}{p} \cdot k + 1 \right]^{-1/q} = \left[\frac{q}{p} \cdot k + 1 \right]^{-1/q} \quad (\text{A.14})$$

Já é possível notar indícios da distribuição Lei de Potência. Considerando p e q como constantes, a expressão dentro dos parênteses do lado direito é proporcional a k , e a fração de x_j que seja no mínimo igual a k é proporcional a $k^{-1/q}$.

Nota-se que esta expressão é a quantidade de vértices $F(k)$ com pelo menos k links entrantes. É possível aproximar diretamente a quantidade de vértices $f(k)$ com exatamente k links entrantes por meio da derivação em outras palavras, aproximando $f(k)$ de $-dF/dk$. Diferenciando a expressão na Equação (A.14), obtém-se

$$\frac{1}{q} \frac{q}{p} \left[\frac{q}{p} \cdot k + 1 \right]^{-1-1/q} \quad (\text{A.15})$$

Em outras palavras, o modelo determinístico é capaz de prever que a quantidade de vértices com k entrantes é proporcional a $k^{-(1+1/q)}$, ou seja uma Lei de Potência com expoente

$$1 + \frac{1}{q} = 1 + \frac{1}{1-p} \quad (\text{A.16})$$

Com uma elevada probabilidade na formação aleatória de links, a quantidade de vértices com k links entrantes é de fato proporcional a $k^{-(1+1/(1-p))}$. O argumento heurístico obtido pela aproximação determinística fornece uma maneira simples de visualizar a origem do expoente da Lei de Potência $1 + 1/(1-p)$.

O comportamento deste expoente faz sentido intuitivamente a medida que p sofre variações. Quando p é próximo de 1, a formação de *link* é principalmente baseada em escolhas uniformemente aleatórias, e a dinâmica do modelo “rico fica mais rico” se modifica.

Correspondentemente, o expoente da Lei de Potência tende ao infinito, mostrando que vértices com muitos *links* entrantes vão se tornando mais raros. E que quando p é próximo de 0, o crescimento da rede é fortemente governado pelo comportamento ligação preferencial, e o expoente da Lei de Potência decresce para 2, permitindo muitos vértices com número elevado de *links* entrantes.

O fato de que 2 é um limite natural para o expoente, ao passo que a dinâmica “rico fica mais rico” se torna mais forte, também fornece uma maneira elegante de pensar que o fato de que muitos expoentes da Lei de Potência em redes reais (como o número de *links* entrantes de uma página *Web*) tendem a ser ligeiramente superiores.

Apêndice B

Demonstração do Lema 4.1.1

Demonstração. Basta mostrar que o Limite B.1

$$\lim_{\alpha \rightarrow \infty} \frac{e^{\alpha} \zeta(\beta - 1) + w_i w_j}{e^{\alpha} \zeta(\beta - 1)} = 1 \quad (\text{B.1})$$

Desta forma,

$$\lim_{\alpha \rightarrow \infty} 1 + \frac{w_i w_j}{e^{\alpha} \zeta(\beta - 1)} = 1 \quad \therefore \quad \lim_{\alpha \rightarrow \infty} \frac{w_i w_j}{e^{\alpha} \zeta(\beta - 1)} = 0 \quad \therefore \quad \lim_{\alpha \rightarrow \infty} \frac{w_i w_j}{e^{\alpha}} = 0 \quad (\text{B.2})$$

e também

$$\lim_{\alpha \rightarrow \infty} \frac{w_i w_j}{e^{\alpha}} \leq \lim_{\alpha \rightarrow \infty} \frac{e^{\frac{\alpha}{\beta}} e^{\frac{\alpha}{\beta}}}{e^{\alpha}} = \lim_{\alpha \rightarrow \infty} e^{\alpha(\frac{2}{\beta} - 1)} \quad (\text{B.3})$$

Se $\beta > 2$, o limite tenderá a zero.

□