Proposta do Trabalho Final PAA: Encontrando os Top-K Influenciadores em uma Rede Social

Vagner Clementino[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

vagnercs@dcc.ufmg.br

1. Introdução

Tradicionalmente as campanhas de marketing se baseiam em determinar um conjunto de consumidores, denominado público-alvo, e posteriormente focam suas ações naquela grupo [Hughes 1996]. Neste contexto, a mineração de dados desempenha um papel fundamental por permitir a construção de modelos que tentam predizer o comportamento de um cliente baseado em seu histórico de compras [Kumar et al. 1999]. Nos casos em que esta abordagem têm sucesso, foi possível perceber um aumento na lucratividade [Piatetsky-Shapiro and Masand 1999]. Contudo, este tipo de abordagem possui uma limitação básica: ela considera que a decisão de compra de uma pessoa é independente dos demais consumidores, desconsiderando o impacto que demais clientes, especialmente aqueles mais "próximos", por ventura possam exercer.

O efeito que os demais consumidores possuem sobre a decisão de compra de um cliente é conhecido em Economia como *externalidade da rede*. Com a expansão da Internet e do uso das redes sociais, este *"efeito da rede"* têm se mostrado de suma importância em diversos setores, especialmente naqueles ligados diretamente à informação (software, imprensa, telecomunicações e etc.) [Shapiro and Varian 2013].

Neste contexto, imagine que você trabalhe em uma empresa de marketing digital que pretende divulgar um novo produto *A* da marca *X* para o maior número possível de usuários em determinada rede social. Uma primeira estratégia seria divulgar o novo produto para cada usuário da rede, que é conhecida como marketing de massa. Tal opção é cara e baixa escalabilidade. Uma segunda alternativa seria apresentar o produto apenas ao usuários que *seguem* a marca *X*, utilizando, deste forma, um marketing direcionado. Esta estratégia peca pela sua abrangência, tendo em vista que não se tem a garantia que as informações do novo produto chegará aos demais usuários da rede social. Uma terceira via seria identificar um grupo de usuários tais que *a partir deles é possível alcançar qualquer outro usuário da rede*. Ao escolher este grupo de usuários, também denominados *"sementes"*, haverá uma maior probabilidade de que a informação chegue aos demais usuários. Este trabalho têm o foco nesta última abordagem.

A maneira que uma informação é propagada em uma rede é conhecida como *Modelo de Propagação*. Neste sentido, a "probabilidade"de propagação da informação dependerá naturalmente do modelo de difusão existente na rede. Contudo, neste trabalho, parte-se da premissa que quanto maior o número de usuários que a informação pode alcançar maior será a propagação independente do modelo de propagação da rede. Na subseção **??** iremos descrever o Modelo de Propagação adotado neste trabalho.

Uma rede social pode ser modelada como uma grafo não direcionado $G(V, E)$, onde o conjunto de vértices V representa os usuários e o conjunto de arestas E representa os relacionamentos entre os mesmos. Na área de *Teoria dos Grafos*, o problema de definir o conjunto $V' \subseteq V$, *de tamanho mínimo*, tal que para todo vértice em V' é possível alcançar qualquer vértice em V é conhecido como **Cobertura de Vértice**. A subseção **??** define o problema formalmente.

Em um primeiro momento o conjunto V' (Cobertura de Vértice) poderia ser definida como as sementes da ação de marketing. Contudo, na prática, o tamanho de V' pode ser grande o suficiente de modo a inviabilizar o orçamento da campanha. Desta forma, se faz necessário encontrar um subconjunto de $W \subset V'$ que alcance um maior número de pessoas sem, todavia, estourar o orçamento. Normalmente o tamanho deste subconjunto de W' deverá ser igual a k, onde k é o número máximo de usuários que a campanha conseguirá patrocinar. O problema de encontrar W tal que $|W| = k$ é conhecido na literatura como **Máxima Influência** e será definido formalmente na subseção **??**.

2. O Problema da Máxima Influência

Seja $G = (V, E)$ um grafo não direcionado que modele uma rede social. A *Máxima Influência* pode ser definido como o problema de encontrar u subconjunto $W \subset V'$, de modo que $|W| \leq K$ e $\cup_{j=1}^{k} w_j$ é *maximizado*, onde $w_j \in V'$ e $1 \leq j \leq k$. Este pequeno conjunto de vértices têm a característica de que sua influência agregada na rede é maximizada.

O conceito de "influência" é subjetivo e deve ser definido no contexto do problema estuado. Por exemplo, pode-se definir $w_j = 1$ caso o *j-ésimo* usuário da rede foi influenciado na compra de um produto ou *zero* caso contrário. Neste sentido, deve-se um buscar um conjunto mínimo de vértices que resulte em $\sum_{j=1}^{k} w_j$ *máximo*. Neste trabalho, um vértice u "influencia"um outro v se o fato de u comprar um um produto faça com que v também realize a mesma compra.

O processo de influenciação depende diretamente do *Modelo de Propagação* utilizado para descrever a forma que vértices interagem na rede. Em síntese, um Modelo de Propagação define para cada vértice uma probabilidade de ativação e a influência é propagada por meio da ativação de vértices vizinhos por meio da probabilidade de ativação destes últimos. Este trabalho utiliza o modelo denominado *Linear Threshold Model* [Granovetter 1978, Schelling 2006], cujos detalhes são descritos na Subseção 4.1.

Para modelo de propagação que foi considerado, o problema de determinar um valor de k que maximize a influência é *NP-Difícil* [Garey and Johnson 1979, Kempe et al. 2003]. Existe na literatura um algoritmo guloso que consegue um fator de aproximação da ordem de $1 - \frac{1}{e}$ [Hochbaum 1997], onde e é a base do logaritmo natural. Desta forma, no melhor caso existe uma garantia de desempenho um pouco acima de 63%. O algoritmo guloso que consegue este é baseado na abordagem apresentada em [Domingos and Richardson 2001]. Assim qualquer trabalho que proponha em resolver o problema da Máxima Influência pode considerar aquele valor como linha de base ("baseline"). Considerando que o Problema da Máxima Influência é *NP-Difícil* se faz necessário a definição de uma heurística ou algoritmo aproximado a fim de o problema seja resolvido em situações reais. A Seção 3 descreve a heurística proposta que é baseada na *Cobertura de Vértice* [Cormen et al. 2009] de um grafo.

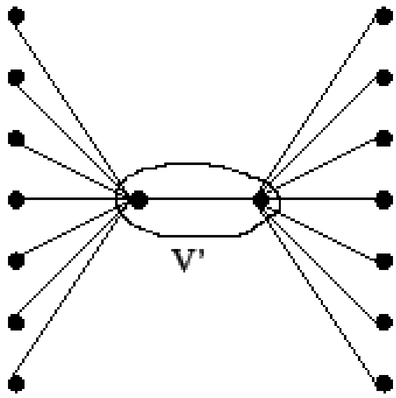


Figura 1. Cobertura de Vértice

3. Heurística Proposta

A heurística proposta tem como objetivo determinar o conjunto A_0, também conhecido como sementes, tal que $A_0 \in V$ e $|A_0| \leq k$ que maximize a influência em um grafo $G(V, E)$. Considerando que a partir de um vértice $v \in V$ é possível alcançar um grande número de outros vértices, pode-se inferir que v é um bom candidato para estar em A_0. Naturalmente os vértices pertencente à *Cobertura de Vértices* do grafo que modela a rede social seriam a primeira alternativa.

3.1. Heurística Baseada na Cobertura de Vértice

O problema da *Cobertura de Vértices* [Garey and Johnson 1979] pode ser definido formalmente como segue: dado um grafo $G = (V, E)$ e um número inteiro $K \leq |V|$ verificar se existe uma cobertura de vértice de tamanho K ou menos para G, isto é, se existe um subconjunto $V^{'} \subseteq V$ tal que $|V'| \leq K$ e, para cada vértice $\{u, v\} \in E$, pelo menos um, u ou v, pertence a $V^{'}$. A figura 5 ilustra um grafo com destaque para sua cobertura de vértice $V^{'}$ de tamanho 2.

Encontrar a Cobertura de Vértice mínima para um grafo $G = (V, E)$ qualquer é *NP-completo* [Garey and Johnson 1979, Cormen et al. 2009]. Diante da inexistência de um algoritmo que retorna a solução exata em tempo polinomial, vem sendo propostos na literatura diversos *algoritmos aproximativos*. Dentre eles, o mais utilizado, consegue executar em $O(V + E)$ em um nível de aproximação igual a *2* [Cormen et al. 2009], ou seja, seja C^{*} uma cobertura de vértice ótima, o algoritmo retornará uma cobertura de vértice C, tal que $|C^{*}| \leq |C| \leq 2 \times |C^{*}|$.

Na prática, caso utilizemos um algoritmo aproximativo de grau 2, o tamanho da Cobertura de Vértice retornada será muito maior do que k. Neste sentido, se faz necessário encontrar dentro da Cobertura encontrada os k vértices que maximizem a influência no grafo. Neste trabalho se realizou uma escolha gulosa dos vértices baseada na métrica denominada *Degree Acesss*. O *Degree Acesss* de um vértice v de um grafo $G(V, E)$ é definido como total de vértices que são passíveis de serem alcançados a partir de v. O *degree acesss* pode ser visto como o tamanho da contribuição exclusiva de um vértice para um possível cobertura de um grafo. A Figura mostra o valor do *Degree Acesss* como "label"de cada um dos vértices do grafo apresentado.

Com base na métrica proposta e na Cobertura de Vértice do grafo foi proposto a

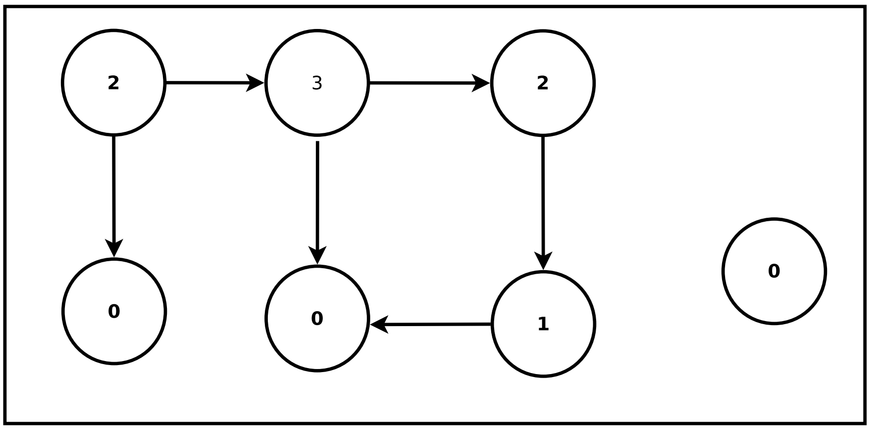


Figura 2. Degree Access para o vértice de um grafo

Algorithm 1: FIND-SEEDS retorna o conjunto semente A_0 com base na Cobertura de Vértice de um grafo.

Input: Um grafo não direcionado e não ponderado $G(V, E)$
um inteiro k correspondente a primeiro índice de A;
um inteiro r correspondente ao último índice de A
Output: Um conjunto $A_0 \in V$ tal que $1 \leq |A_0| \leq k$

$C \leftarrow \emptyset$
$A_0 \leftarrow \emptyset$
$Q \leftarrow \emptyset$ Q é uma fila
$C \leftarrow$ `FIND-VERTEX-COVER(`$G(V, E)$`)`
if $|C| = k$ **then**
  $A_0 \leftarrow C$
  return A_0
else
  `CALCULE-DEGREE-ACESSS(`$G(V, E), C$`)`
  `SORT(`C`)` Ordenando o conjunto C em ordem decrescente ao grau acessibilidade.
  $Q \leftarrow C$Atribuindo o conjunto C para uma fila
  while $|A| < k$ ***or*** Q ***is not*** $\emptyset$ **do**
    $v \leftarrow$ `DEQUEUE(`Q`)`
    if $v.degreeAcess > 0$ **then**
      $A_0 \leftarrow A_0 \cup v$
  return A_0
return A_0

heurística descrita no Algoritmo 1 para determinar o conjunto A_0. Seja $C \in V$ a cobertura de vértice de um $G(V, E)$ obtida utilizando a heurística proposta em [Cormen et al. 2009]. O tamanho de C é no máximo $2 \times C^*$, onde C^* é a Cobertura de Vértice ótima para o grafo $G(V, E)$. Caso $|C| = k$ podemos naturalmente definir $A_0 = C$. Do contrário, devemos encontrar em C os k vértices com o valor de *degree acesss*.

Conforme pode ser observado o Algoritmo 1 utilizada uma estratégia gulosa, com base na Cobertura de Vértices, para definir o conjunto A_0. A Cobertura de Vértice é definida pelo método FIND-VERTEX-COVER que é baseado em [Cormen et al. 2009]. O calculo do *degree acesss* é realizado pelo método CALCULE-DEGREE-ACESSS. Na subseção 3.2 iremos analisar a complexidade de tempo e espaço do algoritmo como um todo.

Naturalmente, por se tratar de um heurística, a escolha gulosa proposta no Algoritmo 1 não resultará em uma solução ótima global. Todavia, espera-se que o conjunto A_0 definido desta forma resultará em um maior número de vértices ativos (conjunto S).

3.2. Análise de Complexidade

3.2.1. Complexidade Temporal

A heurística descrita pelo Algoritmo 1 é baseada essencialmente nos métodos FIND-VERTEX-COVER e CALCULE-DEGREE-ACESSS. Conforme discutido anteriormente, o FIND-VERTEX-COVER têm como base o um algoritmo aproximativo bastante conhecido [Cormen et al. 2009] cuja complexidade temporal é dada por $O(V + A)$.

O método CALCULE-DEGREE-ACESSS é baseado no conhecido algoritmo de busca em largura em grafos *(Breadth-First Search - BFS)*. A ordem de complexidade temporal do BFS é dada por $O(V + A)$. Contudo, o método CALCULE-DEGREE-ACESSS é executado para vértice $c \in C$, onde C é a Cobertura de Vértice. Como no pior caso $C = V$ temos que a contribuição de CALCULE-DEGREE-ACESSS na complexidade temporal é dada por $O(V^2 + VA)$. Como não há outro método com complexidade maior do que a do CALCULE-DEGREE-ACESSS podemos concluir que a *complexidade temporal do Algoritmo 1* é $O(V^2 + VA)$.

3.2.2. Complexidade Espacial

A estrutura de dados básica utilizada neste trabalho foi um grafo a fim de representar uma rede social. A implementação neste trabalho utilizou um *lista de adjacência* o que implica em uma complexidade espacial da ordem de $O(V + A)$.

4. Avaliação da Heurística

4.1. Modelo de Propagação

Conforme exposto anteriormente, a fim de mensurar o valor $\cup_{j=1}^{k} w_j$ devemos definir um *Modelo de Propagação*. No trabalho proposto, será utilizado o modelo conhecido como *Linear Threshold Model* [Granovetter 1978, Schelling 2006]. Neste modelo, um vértice v é influenciado por cada um dos seus vizinhos w de acordo com um peso $b_{v,w}$ que respeita a Equação 1.

O problema da *Máxima Influência* pode ser modelado como um grafo não direcionado $G(V, E)$, onde o conjunto de vértices V representa os possíveis consumidores e as arestas E representam os relacionamentos entre os mesmos. O processo de propagação das informações utilizará *Linear Threshold Model*. Desta forma, para cada consumidor $v \in V$ será atribuído aleatoriamente um threshold θ_v no intervalo $[0, 1]$. As arestas em

E serão ponderadas com os valores $b_{v,w}$ respeitando à Equação 1. Cada vértice $v \in V$ terá um atributo determinando se ele está ativo, sendo que ante da execução do algoritmo todos estarão com o valor *não ativo*.

$$\sum_{\text{w vizinho de v}} b_{v,w} \leq 1 \tag{1}$$

Para cada vértice v é definido aleatoriamente um *threshold* θ_v no intervalo $[0, 1]$; este limiar representa a fração dos vizinhos que deve se tornar ativa para que o vértice v torne-se ativo. Dada uma escolha aleatória dos *threshold* e um conjunto inicial de nós ativos A_0 (sendo os demais nós inativos), o processo de difusão ocorre da seguinte forma: na etapa t, todos os nós que estavam ativos na etapa $t-1$ permanecem ativos, e um vértice v qualquer se torna *ativo* se o somatório dos pesos de seus vizinhos *ativos* sejam $\geq \theta_v$, conforme Equação 2. Cabe ressaltar que o conceito de *ativação* de um vértice depende do contexto que o modelo está sendo aplicado. Neste trabalho considera que um vértice v foi ativado se ele comprou um determinado produto anunciado.

$$\sum_{\text{w é um vizinho ativo de v}} b_{v,w} \geq \theta_v \tag{2}$$

Conforme pode ser observado, ao se aplicar *Linear Threshold Model* o problema se resume em encontrar o conjunto A_0 tal que $|A_0| = k$ e que maximize o total de vértice ativos, denominado conjunto S. É fácil verificar que $|S| = k$ antes da execução do algoritmo.

4.2. Análise Experimental

A fim de avaliar a heurística proposta se faz necessário utilização de um conjunto de dados que apresente as características estruturais de uma rede social de grande escala. Desta forma, de maneira análoga ao trabalho de [Kempe et al. 2003], será utilizado um grafo de colaboração obtido a partir de coautorias em publicações de física [J. Leskovec and Faloutsos 2015]. Têm sido mostrado que redes de coautoria são capazes de capturar as principais características das redes sociais de modo mais geral [Newman 2001].

O gráfico colaboração contém um vértice para cada pesquisador que tem pelo menos um artigo como coautor banco de dados da *arXiv*[1]. Caso um artigo tenha dois ou mais autores, o dataset possui uma aresta para cada par de autores. Observe que isso resulta em arestas paralelas quando dois pesquisadores estão como coautores de vários artigos. O gráfico resultante possui 9877 vértices e 25998 arestas e compreende de artigos publicados no período de janeiro de 1993 a abril de 2003 (124 meses). O data está disponível para download em `https://snap.stanford.edu/data/ca-HepTh.html`.

As baterias de teste consistirão da execução da heurística proposta por 10 mil vezes. Este valor se justifica por ser o número de execução realizada pelo trabalho utilizado como baseline do trabalho. A cada nova execução os valores dos *threshold* θ_v de cada vértice serão recalculados aleatoriamente com objetivo de remover qualquer viés que

[1] `http://arxiv.org/`

possa prejudicar os resultados do trabalho. Os dados que serão utilizados para comparação será o valor médio do conjunto S obtido de todas as execuções.

5. Resultados

Em [Kempe et al. 2003] foi proposto um algoritmo guloso que adiciona vértices ao conjunto A_0 que maximiza o conjunto S. Os autores não entraram em detalhes sobre como escolheram vértices que efetivamente aumentam o tamanho de S. Naquele trabalho, a heurística proposta foi comparada com três outras:

- Na primeira abordagem os vértices em A_0 forame escolhidas randomicamente;
- Na segunda heurística, foram escolhidos k vértices em ordem decrescente de seus graus d_v;
- A terceira abordagem utilizou o conceito de *Distance centrality* [Scott 2012], que é uma medida influência largamente utilizada na sociologia, e parte do pressuposto de que vértices mais próximos terão mais chances de influenciar uns aos outros.

A heurística proposta por [Kempe et al. 2003] teve um desempenho satisfatório comparada com as demais. Desta forma, este trabalho utilizará estas quatro heurísticas como baseline. Para tanto será utilizado o mesmo *dataset* (vide Seção **??**) e os mesmos valores para o atributo k.

—V—	k	—C—	—C— / —V—	Vértices Ativos	% Vértices Ativos
9877	5	8438	0,8543	2173	0,22
	10	8438	0,8543	2173	0,22
	15	8435	0,8540	2469	0,25
	20	8433	0,8538	2766	0,28
	25	8438	0,8543	2963	0,30
	30	8435	0,8540	2963	0,30

Tabela 1. Resultados para diversos valores de k

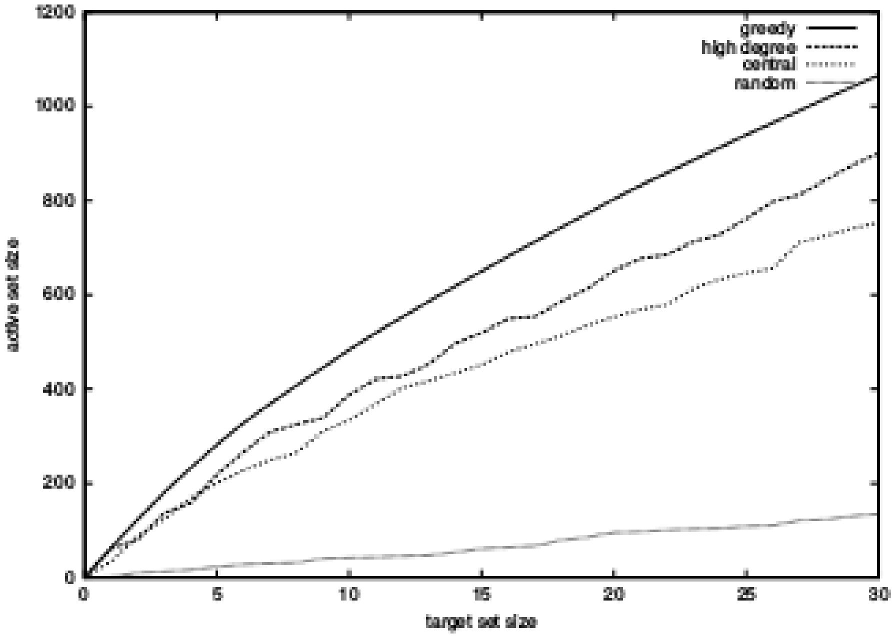
A partir de A_0 será executado o modelo de propagação *Linear Threshold Model* que ao final da execução resultará em um conjunto $S \in V$ que contêm os vértices que foram ativados durante o processo. O tamanho de S para diversos valores de k serão comparados com o baseline descrito na Seção **??**.

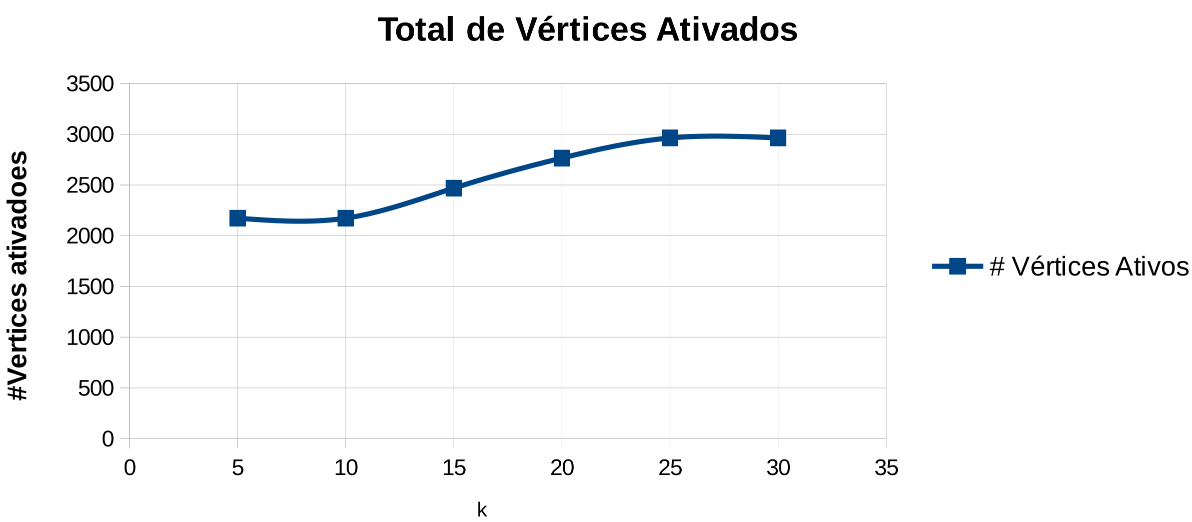
- Resultados da heurística melhor em média, mesmo para valor de k menores
- Variação do total de vértices ativados não acompanha o valor de k
- Valor de $k = 25$ se mostrou o ideal.
- Resultados bem abaixo da melhor solução [Hochbaum 1997]

6. Trabalhos Relacionados

7. Ameaças à Validade

- Modelo de Propagação é artificial
- A métrica DEGREE ACESSS não foi validada.
- Modelo aplicado a um único grafo.
- A escalabilidade da heurística não foi testada.





8. Conclusões e Trabalhos Futuros

- Heurística alcançou resultados satisfatórios.
- Heurística proposta possibilita refinamentos.
- Aplicação em grafos reais.
- Utilização de outros característica das redes sociais para a escolha gulosa.

Referências

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443.

Hochbaum, D. S. (1997). Approximation algorithms for np-hard problems. chapter Approximating Covering and Packing Problems: Set Cover, Vertex Cover, Independent Set, and Related Problems, pages 94–143. PWS Publishing Co., Boston, MA, USA.

Hughes, A. M. (1996). *The complete database marketer: second-generation strategies and techniques for tapping the power of your customer database*. McGraw-Hill.

J. Leskovec, J. K. and Faloutsos, C. (2015). SNAP Datasets: Stanford large network dataset collection. `https://snap.stanford.edu/data/ca-HepTh.html`.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Extracting large-scale knowledge bases from the web. In *VLDB*, volume 99, pages 639–650. Citeseer.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Piatetsky-Shapiro, G. and Masand, B. (1999). Estimating campaign benefits and modeling lift. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193. ACM.

Schelling, T. C. (2006). *Micromotives and macrobehavior*. WW Norton & Company.

Scott, J. (2012). *Social network analysis*. Sage.

Shapiro, C. and Varian, H. R. (2013). *Information rules: a strategic guide to the network economy*. Harvard Business Press.