# Encontrando os Top-K Influenciadores em uma Rede Social

Vagner Clementino

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais(UFMG)
Projeto e Análise de Algoritmos - 2015-1

# Agenda

Contexto

Problema

Objetivo

Modelo Proposto

Avaliação

Resultados

Ameaças à Validade

Conclusões e Trabalhos Futuros

# Contexto

- Tradicionalmente as campanhas de marketing se baseiam em determinar um conjunto de consumidores, denominado público-alvo [6].
- A *mineração de dados* permite a construção de modelos que tentam predizer o comportamento de um cliente baseado em seu histórico de compras [9].
- Nos casos em que esta abordagem têm sucesso, foi possível perceber um aumento na lucratividade [11]

# Contexto

- O efeito que os demais consumidores possuem sobre a decisão de compra de um cliente é conhecido em Economia como *externalidade da rede*.
- Este *“efeito da rede"* vêm crescendo em importância especialmente em setores ligados diretamente à informação (software, imprensa, telecomunicações e etc.) [14].

# Problema

- Suponha uma empresa de marketing digital que pretende divulgar um novo produto *A* da marca para o maior número possível de usuários em determinada rede social.
- Possíveis estratégias:
  - *Marketing de Massa*
  - *Marketing Direcionado*
  - **Divulgar para usuários que possam “influenciar"os demais**

# O Problema Máxima Influência

- Dado grafo não direcionado $G(V, E)$
  - $V$ representa os possíveis consumidores
  - $E$ representam os relacionamentos sociais entre consumidores
- Encontrar $W \subset V$, de modo que $|W| \leq K$ e $\cup_{j=1}^{k} w_j$ é *maximizado*
- Onde $w_j$ tal que $1 \leq j \leq k$ é alguma função sobre os vértices em $V$
- No contexto deste trabalho, maximizar $\cup_{j=1}^{k} w_j$ significa *“influenciar"a compra de um produto por um conjunto maior de usuários*

# O Problema Máxima Influência

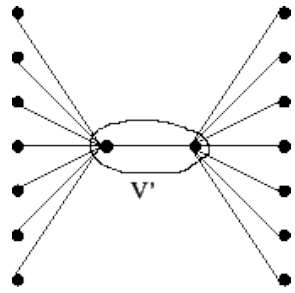
- 0 problema de determinar um valor de $k$ que maximize a influência é NP-Difícil [3]
- Existe na literatura um algoritmo guloso que consegue um fator de aproximação da ordem de $1 - \frac{1}{e}$ (aproximadamente 63%) [5]
- Algoritmo baseado na abordagem apresentada em [2].

# Objetivo

- Propor uma heurística que encontre $W \subset V$ tal que $|W| \leq K$ de modo a maximizar o número de usuários influenciados na rede.
- A Heurística é baseada na *Cobertura de Vértice* [1].

# Cobertura de Vértices

- Dado um grafo $G = (V, E)$ e um número inteiro $K \leq |V|$.
- Existe um subconjunto $V' \subseteq V$ tal que $|V'| \leq K$ tal que cada vértice $\{u, v\} \in E$, pelo menos um, $u$ ou $v$, pertence a $V'$.
- Problema *NP-completo* [3, 1].
- Existe um algoritmo aproximado com nível de aproximação igual a *2*.



# Heurística Proposta

- Seja $A_0 \in V$ e $|A_0| \leq k$ e $A_0$ maximize o número de usuários influenciados
- Seja $C \in V$ a cobertura de vértice de um $G(V, E)$ obtida utilizando a heurística proposta em [1].
- $C$ é no máximo $2 \times C^*$, onde $C^*$ é a Cobertura de Vértice ótima para o grafo $G(V, E)$.
- $v \in C$ é um bom candidato para estar em $A_0$.

# Heurística Proposta

- Na prática $C \gg k$, logo devemos escolher os “melhores" vértice em $C$
- Caso $|C| = k$ podemos naturalmente definir $A_0 = C$.
- Do contrário, devemos encontrar no máximo $k$ vértice em $C$ para fazer parte de $A_0$.
- Escolha gulosa baseada na métrica DEGREE ACESSS

# Algoritmo Proposto

**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 $1 \leq |A_0| \leq k$

```
1  C ← ∅
2  A_0 ← ∅
3  Q ← ∅ Q é uma fila
4  C ← FIND-VERTEX-COVER(G(V, E))
5  if |C| = k then
6      A_0 ← C
7      return A_0
8  else
9      CALCULE-DEGREE-ACESSS(G(V, E), C)
10     SORT(C) Ordenando o conjunto C em ordem decrescente ao grau acessibilidade.
11     Q ← CAtribuindo o conjunto C para uma fila
12     while |A| < k or Q is not ∅ do
13         v ← DEQUEUE(Q)
14         if v.degreeAcess > 0 then
15             A_0 ← A_0 ∪ v
16     return A_0
17 return A_0
```

# Análise do Algoritmo

- O método CALCULE-DEGREE-ACESSS calcula o número de vértices que podem ser alcançados a partir de um vértice $v$
  - Baseado no *BFS*.
  - *Complexidade de Tempo* $O(V + A)$.
  - Executado $|C|$ vezes.
- No pior caso $|C| = |V|$, complexidade do Algoritmo é dada por $O(V^2 + VA)$.
- *Complexidade de Espaço* $O(V + A)$ (lista de adjacência).

# Avaliação

- Baseada em *Modelo de Propagação*
- Utilizada o modelo conhecido como *Linear Threshold Model* [4, 12]
  - Cada vértice $v$ recebe um valor aleatório $\theta_v$
  - $v$ é influenciado por cada um dos seus vizinhos $w$ de acordo com um peso $b_{v,w}$ que respeita a Equação 1.

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

# Avaliação

- $v$ se torna *ativo* se o somatório dos pesos de seus vizinhos *ativos* sejam $\geq \theta_v$, conforme Equação 2
- Baseline [8]
  - Algoritmo guloso
  - Na primeira abordagem os vértices em $A_0$ foram escolhidas randomicamente;
  - Na segunda heurística, foram escolhidos $k$ vértices em ordem decrescente de seu grau $d_v$;
  - A terceira abordagem utilizou o conceito de *Distance centrality* [13]

# Dataset

- Grafo de colaboração obtido a partir de coautorias em publicações de física[7].
- Redes de coautoria são capazes de capturar as principais características das redes sociais de modo mais geral [10].
- O gráfico de colaboração contém um vértice para cada pesquisador com artigo em *arXiv*[1]
  - 9877 Vértices
  - 51971 Aresta

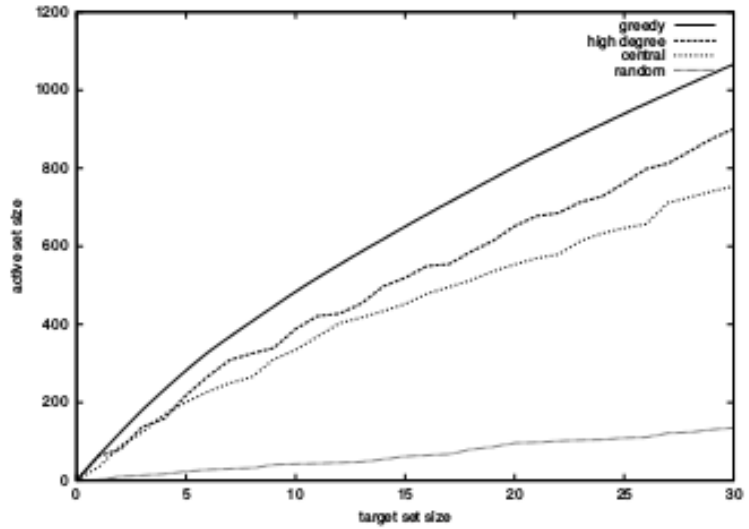
[1] `http://arxiv.org/`

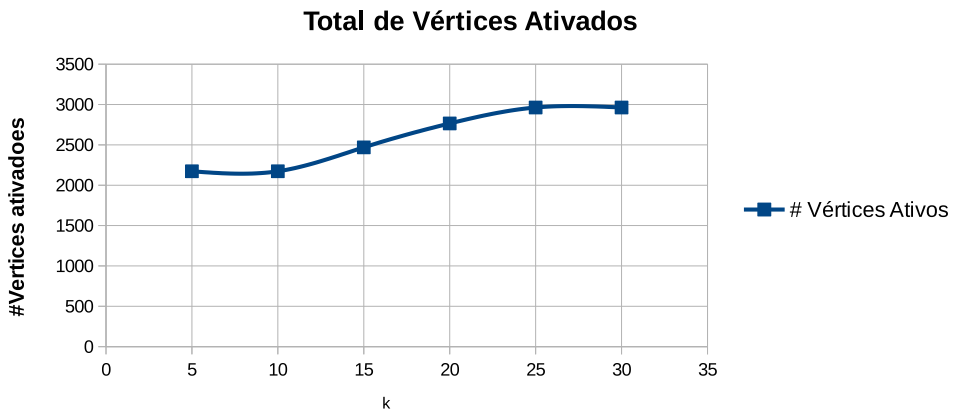
# Resultados

| \|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

# Resultados


greedy
high degree
central
random
active set size
target set size
0
200
400
600
800
1000
1200
0
5
10
15
20
25
30


Total de Vértices Ativados
#Vertices ativadoes
k
# Vértices Ativos
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
25
30
35

# Resultados

- 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 [5]

# 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.

# 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.

# References I


[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[2] P. Domingos and M. Richardson, “Mining the network value of customers,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

# References II

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

[4] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, pp. 1420–1443, 1978.

# References III

[5] D. S. Hochbaum, “Approximation algorithms for np-hard problems,” D. S. Hochbaum, Ed. Boston, MA, USA: PWS Publishing Co., 1997, ch. Approximating Covering and Packing Problems: Set Cover, Vertex Cover, Independent Set, and Related Problems, pp. 94–143. [Online]. Available: http://dl.acm.org/citation.cfm?id=241938.241941

# References IV


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

[7] J. K. J. Leskovec and C. Faloutsos, “SNAP Datasets: Stanford large network dataset collection,” https://snap.stanford.edu/data/ca-HepTh.html, Jun. 2015.

# References V


[8] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Extracting large-scale knowledge bases from the web," in *VLDB*, vol. 99. Citeseer, 1999, pp. 639–650.

# References VI

[10] M. E. Newman, “The structure of scientific collaboration networks,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[11] G. Piatetsky-Shapiro and B. Masand, “Estimating campaign benefits and modeling lift,” in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 185–193.

# References VII


[12] T. C. Schelling, *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[13] J. Scott, *Social network analysis*. Sage, 2012.

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