Sistema de Recomendação em Grafos utilizando Influência Coletiva

PONCIANO, V. S. , MOTTA, C.L.R. , SCHNEIDER, D. S., QUEIRÓZ, A.B., OLIVEIRA, R. L.

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Rio de Janeiro (UFRJ)

Abstract. *This article presents a theoretical study on convexity parameters in graphs with an application in the Recommender System area. The objective is to model the propagation of a recommendation on a network so that a minimal set of users can recommend items to other users associated with maximum acceptance and less loss of information passed on.*

Resumo. *Este artigo apresenta um estudo teórico sobre parâmetros de convexidade em grafos com uma aplicação na área de Sistema de Recomendação. O objetivo é modelar a propagação de uma recomendação em uma rede de modo que um conjunto mínimo de usuários possa recomendar itens aos demais usuários associados com o máximo de aceitação e menor perda de informação repassada.*

1. Introdução

Os sistemas de recomendação estão entre as aplicações mais populares do uso de algoritmos, uma vez que diversas empresas estão interessadas em analisar dados para descobrir o perfil, comportamento e detalhes da navegabilidade virtual de cada usuário. O objetivo das empresas é oferecer melhores direcionamentos de serviços e produtos com intuito de transformar o usuário da rede em cliente consumidor. Esses sistemas são usados para prever a classificação ou preferência que um usuário sinaliza a um item pesquisado na rede e, assim, recomendar produtos e/ou serviços com maior probabilidade de aceitação pelo usuário. Atualmente, a maioria das grandes empresas de tecnologia usam as mais variadas técnicas computacionais para recomendar serviços e/ou produtos a usuários com base em perfis e dados de comportamento de busca na rede. No entanto, essa capacidade automatizada do sistema computacional para predição precisa ser realizada de maneira eficiente.

O objetivo neste artigo é propor uma ideia de um modelo de recomendação representando a rede de usuários por um grafo e usando a convexidade P_3 através de conceitos conhecidos de convexidade em grafos.

2. Definições básicas

Seja $G(V, E)$ um grafo. O conjunto $E(G)$ é chamado de conjunto das *arestas* do grafo, podemos entender uma aresta $uv \in E(G)$ como uma conexão de dois vértices no grafo, os vértices u e v são chamados de *extremos* da aresta uv. O *grau* de um vértice $v \in V(G)$ é o número de vizinhos deste vértice. Uma *clique* K é um conjunto de vértices onde qualquer par de vértices $u, v \in K$ são vizinhos. Um grafo G é dito *completo*, quando $V(G)$ for uma clique. Uma clique com p vértices é denotada por K_p. Duas arestas são

ditas *emparelhadas* quando não compartilharem um mesmo vértice extremo. Um emparelhamento perfeito é quando todas as arestas do grafo são emparelhadas. Chamamos de *iteração* do operador de influência I sobre um subconjunto de vértices S ao conjunto $I(I(...I(S))) = I^k(S)$. Esses e outros conceitos sobre grafos podem ser encontrados em [Bondy and Murty 2011].

Seja S um subconjunto de vértices de um grafo G, suponhamos que S seja um grupo de influenciadores do grafo G. Partindo do conjunto S, iremos infectar o grafo através de sucessivas iterações de I passo a passo até obtermos um conjunto que não infecta mais nenhum outro vértice e esse conjunto é o fecho convexo de S, denotado por $H(S)$.

$$I : I^0(S) \subset I^1(S) \subset I^2(S) \subset ... \subset I^{j+1}(S) = H(S)$$

Caso $H(S) = V(G)$, ou seja, o conjunto S infecta todos os vértices do grafo em um número finito de passos de infecção, dizemos que S é um conjunto de fecho do grafo (em inglês, P_3-*hull set*). Alguns estudos de convexidade podem ser encontrados em [Dourado et al. 2009] e [Nascimento 2020] .

3. O problema de maximização de influência e bibliografias relacionadas

Nas redes sociais em grande escala, o comportamento coletivo de grandes populações pode ser influenciado por um pequeno número de usuários. Tais usuários são denominados *better influencers*. A identificação dos *better influencers* ajuda a controlar uma rede inteira ou uma grande parte da rede. O problema definido como encontrar *better influencers* em uma rede é formalmente definido como a maximização da influência coletiva. De modo a sugerir meios eficientes de tratar o problema, iremos abordar alguns modelos encontrados na literatura assim como introduzir uma nova proposta de modelo.

Segundo o modelo apresentado em conhecido na literatura como *Linear Threshold Model*, cada vértice v do grafo é influenciado por cada um de seus vizinhos de acordo com peso $p_{v,w}$ de modo que a soma desses pesos seja igual a 1. Segundo o autor, a dinâmica de contaminação acontece da seguinte maneira: Cada vértice v escolhe um *threshold* (limite) θ_v de maneira uniformemente aleatória em um intervalo $[0, 1]$; isso representa o números de vizinhos de v que devem se tornar ativos para que v se torne ativo. Dada uma escolha aleatória de limites e um conjunto inicial de vértice ativos I_0, as interação $I(I(...I(S))) = I^t(S)$ são realizada em t etapas, onde todos os vértices que estavam ativos na tempo $t - 1$ permanecem ativos e será ativado um vértice v para o qual o peso total de seus vizinhos ativos seja pelo menos θ_v.

Outro modelo que podemos citar que é apresentado em [Kempe et al. 2003] é o modelo *Cascata Independente*. O processo no modelo descrito em [Kempe et al. 2003] se inicia com um conjunto inicial de vértices ativos I_0, quando um vértice v se torna ativo pela primeira vez na etapa t, é dada uma única chance de ativar cada vizinho dele que é inativo w. O vértice v terá sucesso com uma probabilidade $p_{v,w}$ que é o parâmetro do sistema. Se o vértice v for bem-sucedido, então w se tornará ativo na etapa $t + 1$; caso contrário, não poderá fazer outra tentativa de ativar o vértice w nas etapas posteriores.

4. Propondo uma nova modelagem do problema

Queremos encontrar um conjunto *better influencer* S que infecta $V(G)$ em tempo mínimo $t = k$. Assumimos neste problema que a informação se perde com o tempo, veja um exemplo na Figura 1. Neste problema, a recomendação tem um custo, denotado $C(R)$, que é proporcional ao número de vértices infectados no instante $t = 0$. É preciso que o tempo de infecção seja mínimo para que a perda de informações seja mínima. Esse problema pode ser resolvido de duas formas: A primeira forma é aumentar o custo $C(R)$ e, mais vértices são infectados em um tempo menor. Veja na Figura 1. A segunda forma é verificar o menor tempo possível dado conjunto S. Veja na Figura 2.

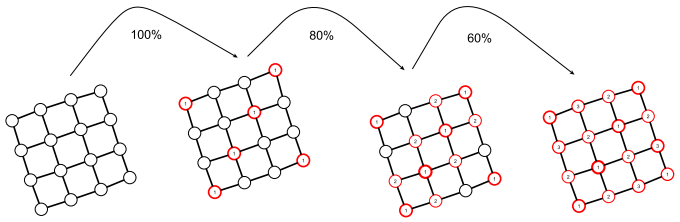

Figure 1. Ilustração da contaminação P_3 com tempo $t = 3$, $|S| = 6$ e as perdas de informação em porcentagem ao longo do tempo.

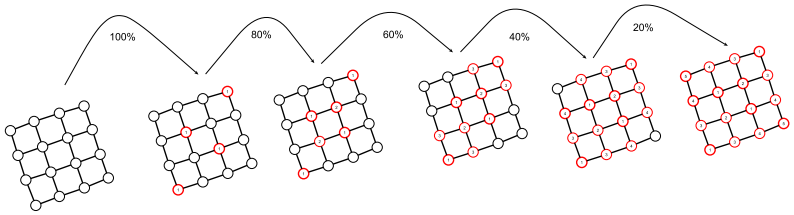

Figure 2. Ilustração da contaminação P_3 com tempo $t = 5$, $|S| = 4$ e as perdas de informação em porcentagem ao longo do tempo.

5. O conjunto de *better influencers* restrito a Grafos Inflados

Seja $G(V, E)$ um grafo. O *grafo inflado* G_I de G é obtido por G trocando todo vértice v de grau $d(v)$ por uma clique $K_{d(v)}$ e cada aresta uv por uma aresta entre dois vértices da clique correspondente $K_{d(u)}$ e $K_{d(v)}$ de G_I de tal forma que G_I que vem das arestas de G forma um emparelhamento de G_I. Os grafos inflados são introduzidos por [Dunbar and Haynes 1996]. O estudo de dominação e de outras variações do problema de dominação pode ser observado em [Favaron 1998, Kang et al. 2004]. Seja $G = (V, E)$ um grafo simples. O conjunto $S \subseteq V(G)$ é uma dominação se para cada vértice $v \in V - S$. O vértice v tem pelo menos vizinho em S.

O conjunto $S \subseteq V(G)$ é uma *dupla dominação* de G se para cada vértice $v \in V(G) - S$, $|S \cap N_G(v)| \geq 2$, ou seja, cada vértice v no grafo tem pelo menos dois vizinhos em S. O cálculo da dupla dominação mínima foi provado em [Centeno 2012] que para grafos bipartidos, cordais, bipartidos cordais, a complexidade do problema é NP-completo. Note que o conjunto de *better influencers* para $t = 2$ é também uma dupla dominação.

Teorema 1. *O conjunto de better influencers para $t = 2$ (uma dupla dominação) é NP-completo para grafos inflados.*

Prova: O conjunto de *better influencers* para $t = 2$ é uma vez que é verdade para grafos em geral [Centeno 2012]. Para a redução, usaremos o problema da dominação simples **DSP** para grafo planar G com o grau no máximo 3 e construiremos G'_I da seguinte maneira: [Garey and Johnson 1979].

- Primeiro, criamos G' como o resultado de uma inflação de cada vértice para K_3, para todo $u \in V(G)$. Em seguida, para todo vértice previamente criado, também criamos um K_3 ligando tal vértice cada $v \in V(G')$. Veja a Figura 3 como exemplo de G e seus respectivos G'.
- Agora, criamos G'_I como resultado da inflação de G', e desejamos verificar a existência de um conjunto duplamente dominante em G_I com no máximo $21\ell + 20(|V(G)| - \ell)$ vértices. Veja um exemplo do resultado G_I por G' na Figura 3.

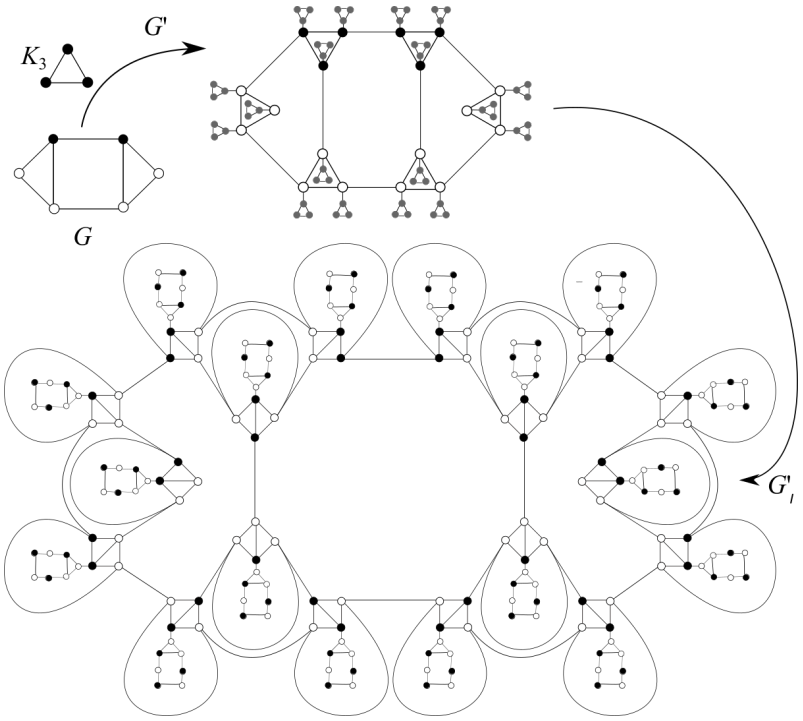

Figure 3. Grafo resultante da transformação polinomial.

Note que G_I ainda é planar e tem o grau no máximo 4. Agora, considere os gadgets da Figura 4. O Gadget 1 tem um conjunto duplamente dominado com 15 vértices. Além disso, os vértices do mesmo gadget que vieram de K_4 fornecem 5 vértices cada um para o conjunto o conjunto de *better influencers* para $t = 2$ e ainda menos do que é impossível. Como precisamos de 2 vértices de outras K_4, temos a seguinte afirmação:

Afirmação 1: O *gadget* contribui com pelo menos 15 vértices em um conjunto dominante.

No entanto, na Figura 4, temos o Gadget 2 com um vértice dominado por outro fora do gadget, o que diminui o tamanho do conjunto duplo dominante. Observe que não

há mais de um K_4 originado de K_3 contribuindo com apenas um vértice para o conjunto duplamente dominado já que os outros garantem que os vértices do mesmo K_4 sejam dominados duas vezes. Além disso, o K_4 com apenas um vértice em um conjunto dominante precisa de um vértice de algum Gadget 1. Portanto, temos a segunda afirmação:

Afrmação 2: O *gadget* 2 contribui com pelo menos 14 vértices se existe um vértice do gadget 1 adjacente a algum vértice nele.

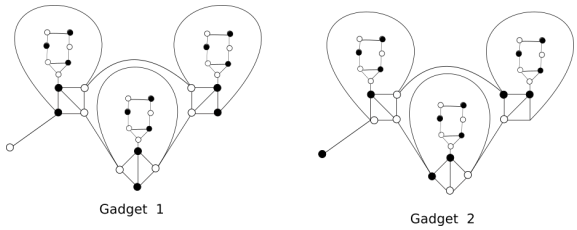

Figure 4. Gadgets *better influencers* possível.

Agora, vamos provar que $\langle G, \ell \rangle$ é uma instância `SIM` para `DSP` se e somente se $\langle G'_I,\ 15\ell + 14(|V(G)| - \ell)\rangle$ e ainda tem instância `SIM` para *better influencers*.

($\Rightarrow$) Seja D um conjunto em G com $|D| \leq \ell$. Para cada vértice de D, fazemos o conjunto P_3-interval, $D_{G'_I}$, de acordo como o gadget 1 para o gadget resultante da transformação. Agora, para os demais gadgets fazemos de acordo com gadget 2, onde o K_4 originado de K_3 que não tem dois vértices em uma dupla dominação deve ser dominado por um vizinho originário de um vértice em $D_{G'_I}$. Então, temos um conjunto duplamente dominado $D_{G'_I}$ para G'_I com $15|D| + 14(|V(G)| - |D|) \leq 15\ell + 14(|V(G)| - \ell)$ vértices.

($\Leftarrow$) Considere G'_I com um conjunto de *better influencers* de no máximo $15\ell + 14(|V(G)| - \ell)$ vértices. Pegue os vértices de G cujos gadgets têm exatamente 14 vértices no Tempo de infecção t=2 os coloque em $\bar{D}$. Lembre-se que na *Afirmação 2*, temos que este gadget tem um vizinho em algum gadget 1. Portanto, colocamos os vértices de G que correspondem a esses gadgets em D. Note que D é um conjunto dominante para G. Como $15|D| + 14\bar{D} \leq 15\ell + 14(|V(G)| - \ell)$, temos $|D| \leq \ell$.

Teorema 2. *Seja G_I um grafo inflado. O conjunto dos usuários considerados better influencers de G_I pode ser computado em tempo polinomial quando não temos as restrições para tempo t.*

Prova: Primeiro vamos provar que o conjunto de *better influencers* de G_I tem tamanho $|V(G)| = n$. Note que cada clique de G_I sempre terá pelo menos um vértice em qualquer P_3-hull set, pois de outra forma não poderíamos gerar os vértices dessa cliques.

Suponha que tenhamos um conjunto S de *better influencers* de tamanho $n + 1$, então existe uma clique $K \in G_I$ com dois vértices em S, digamos u e v. Isso implica que os vértices $K - \{u, v\}$ da clique K estão a uma distância 2 de um vértice gerado em $S - \{u, v\}$ e para o vizinho de u fora de K, digamos w, o vértice w não pertence a S e existe um único vértice w' em S na clique K_w e todos os vértices em $K_w - \{w, w'\}$ estejam a uma distância 2 de um vértice gerado $S - \{u, v\}$. De forma indutiva temos que $V(G_I)$ não está contido no fecho de S, que é um absurdo.

Vamos construir um conjunto $S \subseteq V(G_I)$, tal que S tem um vértice de cada clique, portanto $|S| = n$. Vamos provar por indução que S é um *better influencers* set de $V(G_I)$. Seja v um vértice isolado de G_I, obviamente $S = \{v\}$ e $n = 1$. Suponha por hipótese de indução um P_3-hull set S_k com k vértices, onde para todo $v \in S_k$, v pertence a uma clique distinta de G_I. Vamos mostrar que vale para S_{k+1}. Dado uma clique K que foi gerada por S_k, olhamos para as cliques vizinhas, se for uma clique com único vértice esse já havia sido gerado pois é uma clique de tamanho 1. Se for uma clique de tamanho maior que 1, coloque em S um vértice da clique vizinha $K' - w$ onde w é vizinho vértice u em K, assim a clique K' é gerada e S_{k+1}.

6. Conclusão e Trabalhos futuros

Pode-se observar que para os grafos inflados o problema pode ser computado em tempo polinomial para um tempo t sem restrições, apesar do fato observado de que o poder de influenciar cada vértice perde-se ao longo do tempo. Por outro lado, quando $t = 2$, existe uma chance maior de influenciar, porém o custo computacional é exponencial. Para trabalhos futuros pretende-se analisar uma rede real e analisar o problema em outras classes de grafos.

References

Bondy, A. and Murty, U. S. R. (2011). *Graph Theory*. Graduate Texts in Mathematics. Springer London.

Centeno, C. C. (2012). *A convexidade P_3 para grafos não direcionados*. PhD thesis, UFRJ.

Dourado, M. C., Gimbel, J. G., Kratochvíl, J., Protti, F., and Szwarcfiter, J. L. (2009). On the computation of the hull number of a graph. *Discrete Mathematics*, 309(18):5668–5674.

Dunbar, J. E. and Haynes, T. W. (1996). Domination in inflated graphs. *Congressus Numerantium*, pages 143–154.

Favaron, O. (1998). Irredundance in inflated graphs. *Journal of Graph Theory*, 28(2):97–104.

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.

Kang, L., Sohn, M. Y., and Cheng, T. E. (2004). Paired-domination in inflated graphs. *Theoretical computer science*, 320(2-3):485–494.

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.

Nascimento, J. R. ; Ferreira, D. J. . C. E. M. M. (2020). Número envoltório na convexidade p3: Resultados e aplicações. *Revista de Sistema e Computação - RSC*, 9(7):238–244.