Estudo do impacto da seleção de sementes baseada em centralidade e em informações de comunidades sobrepostas para a maximização da influência

Gilma A. S. Campos[1,2], José M. Ribeiro[2], Vinícius F.Vieira[2], Carolina R. Xavier[2]

[1]Instituto Federal do Sudeste de Minas Gerais
São João del-Rei - MG - Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de São João del Rei (UFSJ) – São João del Rei, MG – Brasil

carolinaxavier@ufsj.edu.br

Abstract. *The problem of maximizing influence, proposed for social networks, involves identifying a set of influential nodes that initiate the diffusion process in a manner that maximizes the spread of influence. This study aims to compare the extent of diffusion in two different contexts. The first context involves selecting individuals based on centrality measures, while the second context involves selecting individuals using three criteria related to overlapping communities. A comprehensive comparison was conducted using the Independent Cascading Model as the diffusion model. The results revealed that in certain scenarios, the utilization of overlapping communities led to enhancements in the reach of diffusion.*

Resumo. *O problema de maximizar a influência, proposto para redes sociais, envolve identificar um conjunto de nós influentes que iniciem o processo de difusão de maneira a maximizar a propagação da influência. Este estudo tem como objetivo comparar a extensão da difusão em dois contextos diferentes. O primeiro contexto envolve a seleção de indivíduos com base em medidas de centralidade, enquanto o segundo contexto envolve a seleção de indivíduos usando três critérios relacionados a comunidades sobrepostas. Uma comparação abrangente foi realizada utilizando o Modelo de Cascata Independente como modelo de difusão. Os resultados revelaram que, em certos cenários, a utilização de comunidades sobrepostas resultou em melhorias no alcance da difusão.*

1. Introdução

O estudo da difusão envolve a análise da disseminação de um produto, ideia ou doença em uma rede. No contexto onde queremos maximizar essa difusão, surge o questionamento: quem seriam os indivíduos dessa rede que poderiam realizar este processo de forma mais eficiente? [Domingos and Richardson 2001] foram os primeiros a formular, para o contexto de marketing, a necessidade de encontrar uns poucos indivíduos que pudessem ampliar o efeito de adoção de um produto. [Kempe et al. 2003] trouxeram a questão para o contexto de redes sociais e encontraram uma resposta na criação de um algoritmo guloso para realizar a seleção destes indivíduos.

Na busca por melhores resultados em processos de difusão, surgem outros trabalhos, como [Leskovec et al. 2007] com o algoritmo CELF (*Cost-Effective Lazy Forward)*, que utiliza da estratégia *lazy evaluation* para diminuir significativamente o número de avaliações empreendidas pelo algoritmo de [Kempe et al. 2003]. Em [Goyal et al. 2011], os autores propõem uma melhoria para o algoritmo CELF de [Leskovec et al. 2007], o CELF++, cujos testes comprovam uma melhora significativa em relação ao tempo. Desde então, diversos trabalhos vem propondo heurísticas para a solução da maximização da influência, como [Chen et al. 2009],[Yang et al. 2020] e diversos outros que podem ser encontrados nas revisões feitas por [Sumith et al. 2018] e [Aghaee et al. 2021].

Outro ponto que deve ser considerado no entendimento dos processos de difusão, é a estrutura da rede. De acordo com Gulati [Gulati 1998], uma rede pode influenciar a ação de seus membros de duas formas: por meio do fluxo e compartilhamento da informação dentro da rede e pelas diferenças nas posições dos indivíduos nela contidos. Tal afirmação, traz a importância de conhecer a estrutura de comunidades inerente às redes sociais. Os autores, Granovetter [Granovetter 1978] e Bakshy *et al* [Bakshy et al. 2012], consideram que ligações entre comunidades funcionam diferentemente daquelas dentro das comunidades. Ou seja, amigos de uma mesma comunidade possuem laços fortes, mas os laços fracos entre amigos de diferentes comunidades são importantes na difusão de novas informações, pois possibilitam maior fluxo de informações entre eles. Encontrar esses indivíduos, que pertencem a mais de uma comunidade está relacionado à estudos de detecção de comunidades com sobreposição. Por exemplo, o trabalho de [Lancichinetti et al. 2009] .

Embasado na alegação de [Granovetter 1978] e [Bakshy et al. 2012], alocar estes indivíduos no um conjunto de sementes de difusão pode ser significativo para maximizar a disseminação da informação. Resta assim, buscar formas de encontrar estes indivíduos e avaliar a sua importância no processo. O uso de indivíduos com esta característica pode ser uma forma de contornar o problema apontado por Easley *et al* [Easley et al. 2010], que aborda como uma cascata de difusão pode ser interrompida quando encontra uma comunidade muito densa em uma rede. Dessa forma, a utilização de vértices alocados em comunidades sobrepostas pode ser uma boa estratégia para a difusão, uma vez que o fato de estar em mais de um grupo possibilita que a informação flua de um grupo para o outro.

Em [Shakarian et al. 2015] o processo de difusão para o *Independent Cascade Model* é definido da seguinte forma: a cada intervalo de tempo t onde χ_{t-1}^{new} é o conjunto de *nós* ativos iniciais no tempo $t-1$, cada $v \in \chi_{t-1}^{new}$ ativa o vizinho inativo $u \in \eta_{out} v$ com a probabilidade $p_{u,v}$, os nós ativados do tempo t, repetem este processo até que nenhum novo nó é ativado. No contexto da maximização da influência, quanto mais nós forem ativados ao final do processo, melhor é o conjunto inicial de sementes.

Este trabalho pretende comparar a qualidade da seleção de sementes com diferentes critérios. Como linha de base, foram utilizadas sementes obtidas através do ranking por medidas de centralidade em redes, essas foram comparadas com conjuntos de sementes extraídos das comunidades com sobreposição, com três diferentes propostas. O modelo utilizado nos experimentos será o *Independent Cascade Model*, que é adotado por ser amplamente utilizado na literatura para este fim.

2. Detecção de Comunidades

Comunidades podem ser entendidas como a partição dos vértices de uma rede em vários conjuntos, onde cada conjunto possui mais ligações entre seus vértice e menos ligações com os vértices de outras comunidades. No mundo real, indivíduos podem estar em mais de um conjunto, pois participam de mais de um grupo de forma coesa, por exemplo, grupo de amigos, grupo familiar, um grupo de praticante de esportes. Então surge a ideia de comunidades com sobreposição, onde os indivíduos são identificados em uma ou mais comunidades. Palla et al [Palla et al. 2005] foram os primeiros a levantar esta questão para a comunidade da ciência de redes. Existem diversos algoritmos para realizar essa tarefa, neste trabalho a detecção das comunidades com sobreposição é feita por dois algoritmos que são baseados no Label propagation algorithm(LPA).

O processo de detecção de comunidades do LPA é feito com a propagação de rótulos. Inicialmente, é feita uma atribuição de um rótulo para cada *nó* do grafo, representando a comunidade inicial. Em seguida, o algoritmo propaga esses rótulos por toda a rede, com base nas arestas entre os *nós*. Em cada iteração, cada *nó* recebe o rótulo com base na maioria dos rótulos de seus vizinhos. Se houver um empate, o rótulo é escolhido aleatoriamente. O algoritmo continua propagando os rótulos até que cada nó tenha o mesmo rótulo que a maioria de seus vizinhos.

Esse processo é repetido até que a rede seja dividida em um número desejado de comunidades ou até que um critério de convergência seja alcançado. Ao final desse processo, os nós são agrupados em comunidades conforme a identificação de seus rótulo. Este algoritmo detecta comunidades disjuntas, mas dá base para os algoritmos que seguem e fazem a detecção de comunidades com sobreposição.

Em [Coscia et al. 2012] é apresentado o algoritmo DEMON (*Democratic Estimate of the Modular Organization of a Network*). O algoritmo funciona da seguinte forma, inicialmente, cada nó na rede é considerado como uma comunidade. Em seguida, há o cálculo da ”democracia”dos nós. A ”democracia”de um nó é definida como a proporção de seus vizinhos pertencentes à comunidade atual desse nó. Essa medida de indica a afinidade de um nó com a comunidade à qual pertence em comparação com as outras comunidades em seu ambiente local. Cada nó é atribuído à comunidade com a maior medida de ”democracia”. Se a medida de ”democracia”do nó for menor que um determinado limiar, o nó permanece em sua comunidade atual. A cálculo da medida e a alocação dos nós nas comunidades são repetidos até que ocorra a convergência. Por fim, as comunidades locais formadas pelos nós são mescladas para formar uma coleção global de comunidades. A fusão é realizada combinando-se as comunidades que possuem nós em comum.

O algoritmo SLPA (*Spekers-Lisener Label Propagration*), proposto pelos autores [Xie et al. 2011], segue a ideia de [Gregory 2010] de *nós* possuindo vários rótulos de forma a permitir a sobreposição. A dinâmica do processo do algoritmo SLPA precisa determinar: (i) como a informação de um *nó* espalha para outros; (ii) como processar informações recebidas de outros. Ambas as questões dizem respeito em como prover a manutenção da informação, de maneira que, a solução encontrada está imitando o comportamento de comunicação humana, usando a propagação da informação pelo processo de (*speakers-listener*). No processo, cada nó pode ser ouvinte ou falante, dependendo do papel do nó no momento, fornecer ou consumir informação. Um *nó* pode armazenar

vários rótulos. Em vez de apagar o conhecimento de todos os rótulos observados, um *nó* os acumula repetidamente, apagando apenas um deles. Ademais, "quanto mais um nó observar um rótulo, maior a probabilidade de espalhar esse rótulo para outros *nós* (imitando a preferência das pessoas de espalhar opiniões com mais frequência)".

3. Materiais e Métodos

3.1. Dados

As redes selecionadas para investigação foram coletadas dos repositórios: SNAP [1], Projeto KONECT [2], Network Data [3] e Network Repository [4], e estão descritas na Tabela 1, que apresenta cada rede o seu identificador (REDE), o número de *nós* (N), o número de arestas (M), o grau médio (GRAU MÉDIO), a densidade da rede (DENS) e o coeficiente de *clustering* (CC). As informações apresentadas para cada rede, considera a componente gigante, como em [Salavati et al. 2018]. A escolha foi necessária, uma vez que, a falha na cascata de difusão pode ocorrer em consequência da possibilidade de que sementes alcancem um componente desconectado.

Tabela 1. Informações básicas das redes de estudo

	N	M	GRAU MÉDIO	DENS.	CC
GrQc	4158	13428	6.46	0.0015	0.556
Epinions1	75877	405739	5.35	0.000070	0.069
Sign-epinions	119130	704572	5.19	0.00005	0.070
Wiki-Vote	7066	100736	14.25	0.0020	0.071

3.2. Métodos de escolha de sementes

O primeiro passo foi a seleção de conjuntos de sementes utilizando *nós* da rede de acordo com as medidas de centralidade. Uma vez que as mesmas têm sido muito utilizadas para classificar a habilidade de propagação dos *nós*. Assim foram utilizadas as medidas de grau, *betweenness*, *closeness* e *PageRank*, além de sementes selecionadas aleatoriamente. Em um segundo momento, são propostos três critérios adicionais utilizando o conceito de comunidades sobrepostas. Para encontrar comunidades com sobreposição foram utilizados os algoritmos, SLPA [Xie et al. 2011] e DEMON [Coscia et al. 2012]. Os parâmetros de configuração dos algoritmos para execução, seguem os propostos pelos autores. Os algoritmos utilizados para a descoberta de comunidades foram executados a partir das implementações disponíveis na biblioteca CDlib (*Community Discovery Library*), versão 0.1.7) [Rossetti et al. 2019].

Após a obtenção das comunidades sobrepostas, sementes foram selecionadas de acordo com os seguintes critérios:

1. Critério A - para alcançar o valor de contribuição de cada comunidade foi calculado a relação do tamanho da comunidade com o tamanho da rede e a seguir o resultado encontrado em relação ao tamanho máximo do conjunto de sementes,

[1]https://snap.stanford.edu/data/
[2]http://konect.cc/networks/
[3]http://www-personal.umich.edu/ mejn/netdata/
[4]http://networkrepository.com/

ou seja 10% da rede. No caso das grandes redes, muitas comunidades pequenas tiveram seu valor de contribuição menor que 1, resultando na não contribuição destas comunidades para o conjunto de sementes. Dessa forma, para as comunidades que obtiveram este resultado foi especificado que as mesmas contribuiriam com 1 semente.

2. Critério B - foram selecionados *nós* que estivessem em qualquer sobreposição entre duas ou mais comunidades. Esses nós foram ordenados, em ordem decrescente, usando seu número de ocorrências em sobreposições, e a partir dessa ordem foram selecionadas como sementes aqueles com maior número de ocorrências.
3. Critério C - foram selecionados os *nós* que estivessem em qualquer sobreposição entre duas ou mais comunidades. Porém, neste caso, os *nós* foram ordenados de maneira decrescente, usando o grau de cada um.

3.3. Experimentos

Os experimentos foram realizados com 5 subconjuntos de sementes para cada rede, cada subconjunto contém propriamente os subconjuntos menores, e o tamanho deles variou de 2% a 10% do tamanho da rede, passo 2%. Esta estratégia considera a disparidade entre o tamanho das redes estudadas.

O processo de difusão foi realizado para cada subconjunto utilizando a implementação do modelo *Independent Cascade*, disponível na biblioteca NDlib [Rossetti et al. 2018] versão (5.0.0), e o parâmetro de configuração do modelo, segue o valor padrão da biblioteca, a probabilidade de v influenciar u é $p_{v,u} = 0,1$.

Cada experimento separado por rede, utilizando cada um dos subconjuntos, resultou em uma percentual de *nós* atingidos na rede. Dessa forma, a avaliação foi feita comparando os resultados obtidos pelas medidas de centralidade com os resultados obtidos pelos critérios baseados em comunidades sobrepostas.

Para avaliar o resultado dos experimentos, foi observado o valor quantitativo do alcance de cada conjunto de sementes e as interações entre os elementos desses conjuntos, ou seja, as arestas que eles compartilham, relacionando o conjunto de sementes com a topologia da subrede que eles formam. considera-se que o número de arestas de cada nó selecionados como semente é proporcional ao custo financeiro de um eventual contrato do interessado com um potencial semente.

4. Resultados e Discussões

Nesta seção serão apresentados os resultados obtidos durante os experimentos para cada uma das redes. A Tabela 2 apresenta o resumo dos resultados, cada linha da tabela traz o resultado de cada uma das redes e cada coluna se refere a um critério testado. Para cada uma das redes, o valor alcançado por cada critério foi normalizado de acordo com o valor do melhor deles para cada rede, assim, os valores variam de 0 a 1, sendo os critérios com nota 1 o melhor resultado para a rede em questão. As colunas A_S, B_S e C_S correspondem aos experimentos com o algoritmos de detecção de comunidades SLPA. Para a identificação dos resultados do algoritmo DEMON, os dados estão sob as colunas A_D, B_D e C_D. Para as demais colunas cujos experimentos são de acordo com as medidas de centralidade e o conjunto de sementes aleatórias: BC refere-se a *Betweenness Centrality*; CC ao *Closeness Centrality*, DC ao *Degree Centrality* (grau), Rand (aleatória) e PR ao *PageRank*.

Tabela 2. Resumo dos Resultados para o *threshold* = 0,1

	A_S	B_S	C_S	A_D	B_D	C_D	BC	CC	DC	Rand	PR
GrQc	0.39	0.62	0.85	0.57	0.80	0.85	**0.98**	0.72	0.76	0.73	**1.00**
Epinions1	0.94	0.67	0.75	0.94	0.93	**0.98**	**1.00**	0.96	**0.99**	0.68	0.70
Sign-epinions	0.96	0.53	0.59	0.97	0.96	**1.00**	**0.99**	**0.98**	**1.00**	0.66	0.72
Wiki-Vote	0.92	0.72	0.89	0.93	0.69	**0.99**	0.96	**0.99**	**1.00**	0.76	0.59

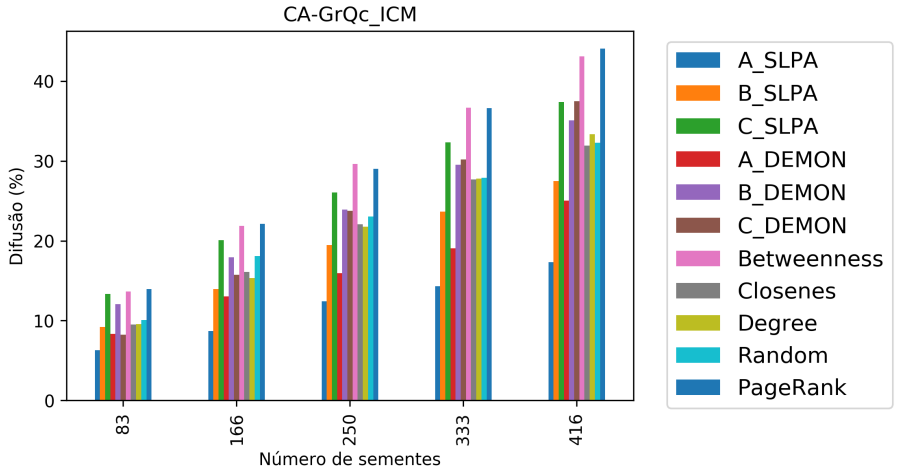

Figura 1. Resultado da difusão da rede GrQc

Os resultados da difusão para a rede GrQc estão presentes na Figura 1 para todos os conjuntos de sementes, no eixo x é variado o tamanho de subconjunto de semente, e cada coluna possui uma cor diferenciando o critério. O eixo y representa o percentual de elementos influenciados para cada subconjunto. A Figura 2 mostra as relações entre os subconjuntos de arestas, obtidos do subgrafo gerado a partir de cada subconjunto de sementes.

Observando-se a Tabela 2 e a Figura 1 é fácil identificar que os melhores resultados obtidos no processo são identificados para as medidas *PageRank* e *Betweenness*, seguidas pelos critérios C_D, C_S e B_D em comparação com as outras medidas. Na diagonal principal da matriz, na Figura 2, é mostrado a quantidade de arestas do subgrafo induzido por cada subconjunto de sementes. A paleta de cores indica a interseção entre os subconjuntos de sementes, quanto mais claro, maior é o número de arestas em comum entre os conjuntos. A partir desta observação, vê-se que o número de arestas para o subconjunto de grau é maior, seguido do subconjunto C_D.

Observa-se na Figura 2 que o subgrafo induzido pelas sementes que utilizaram o *PageRank* possui 2019 arestas ao passo que o subgrafo induzido pelas sementes que utilizaram o *Betweenness* possui 1280, no entanto, a segunda apresenta desempenho ligeiramente menor do que o desempenho da primeira (1), o que torna evidente que a medida de *Betweenness* possui sementes muito importantes para o processo de difusão, fato que pode ser observado pelo desempenho deste subconjunto com aproximadamente 60% das arestas utilizada pelo conjunto *PageRank*. Pode-se imaginar que o custo de uma semente

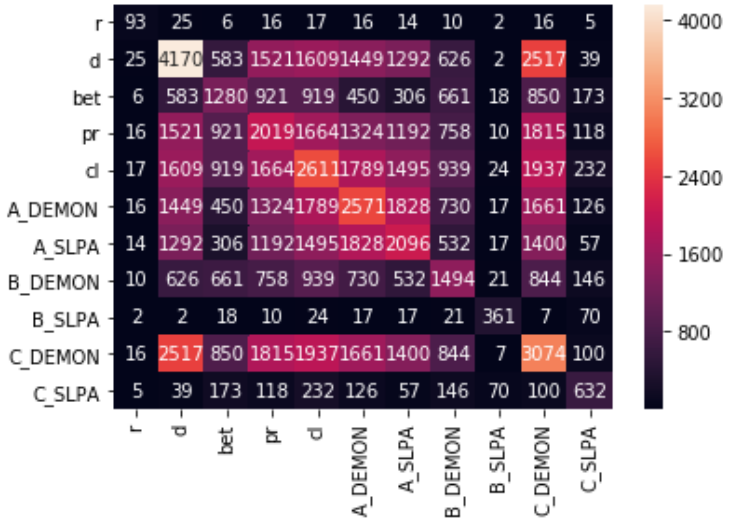

Figura 2. Relação entre Arestas da rede GrQc

que possui muitas conexões, principalmente no contexto de redes sociais, é maior do que o custo de sementes menos conectadas, então pode-se considerar que o benefício do uso do *Betweeneess* em relação ao *PageRank* é considerável. É importante observar que o critério baseado em comunidades C_S teve um resultado próximo (85%) do melhor valor, mas utilizou-se de sementes com menos conexões iniciais (30%) do melhor, o que poderia acarretar em um ganho ainda maior, o C_D também obteve desempenho similar, no entanto considerou uma quantidade maior de arestas em comum.

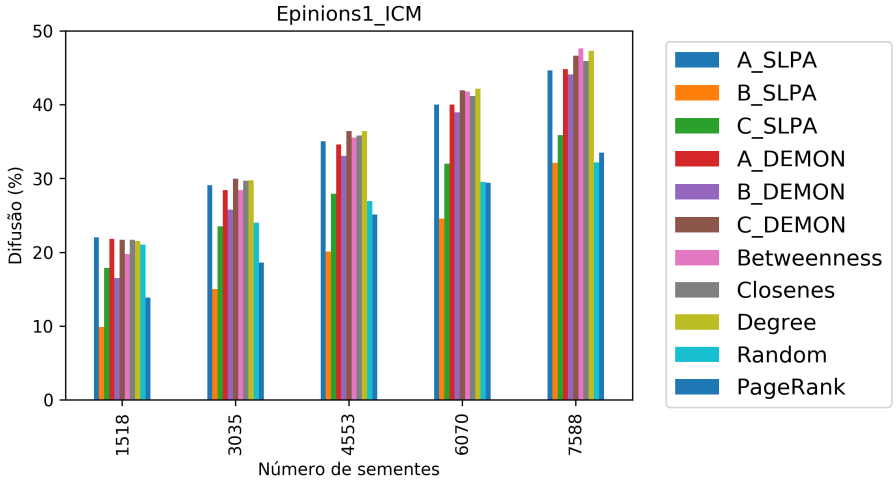

Figura 3. Resultado da difusão da rede epinions1

Os resultados para a rede Epinions1, como exibido na Figura 3, contabilizados o processo de difusão com o subconjunto com 10% da rede como sementes, alcança os melhores resultados com *Betweenness*, grau e C_D, respectivamente. No entanto, é importante notar que para conjuntos menores de sementes, o desempenho do *Betweenness* e do grau não se manteve, enquanto o critério baseado em comunidade se manteve entre os três melhores para qualquer tamanho de conjunto. Analisando as relações entre as

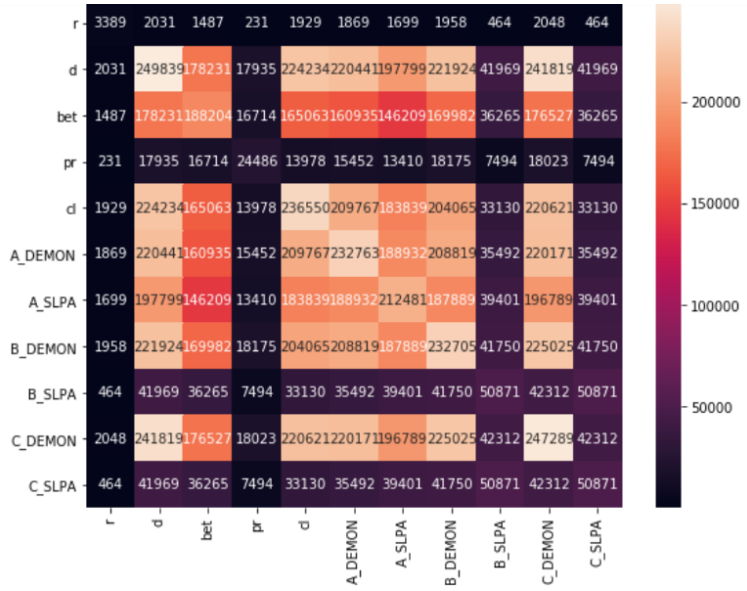

Figura 4. Relação entre Arestas da rede epinions1

arestas, pode-se identificar que entre os critérios que obtiveram os melhores alcances, o grau possui um número muito grande de arestas em comum, enquanto os que o seguem possuem menores números de arestas compartilhadas. Embora as diferenças entre a taxa de difusão entre eles no resultado geral normalizado seja de apenas um ponto percentual, seria mais vantajoso se utilizar dos conjuntos que compartilham menos arestas.

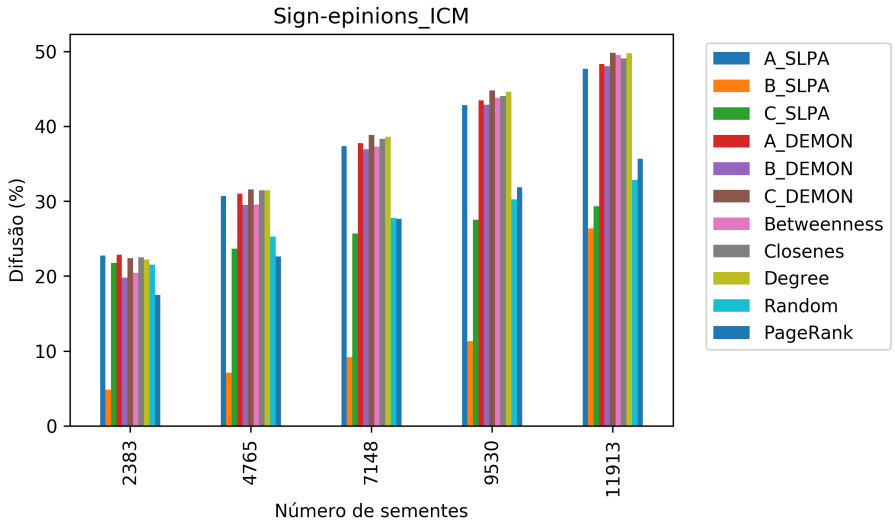

Figura 5. Resultado da difusão para a rede sign_epinions

O comportamento da rede sign-epinions é muito similar ao da rede Epinions1. Provavelmente porque ambas as redes estão inseridas no mesmo contexto, ambas são redes que mede a opinião de clientes sobre produtos em uma rede social. O grau e o critério C_D empatam em primeiro lugar, e para ambas possui uma quantidade significativamente maior de arestas do que os demais critérios, nos levando a concluir que seriam opções bastante custosas.

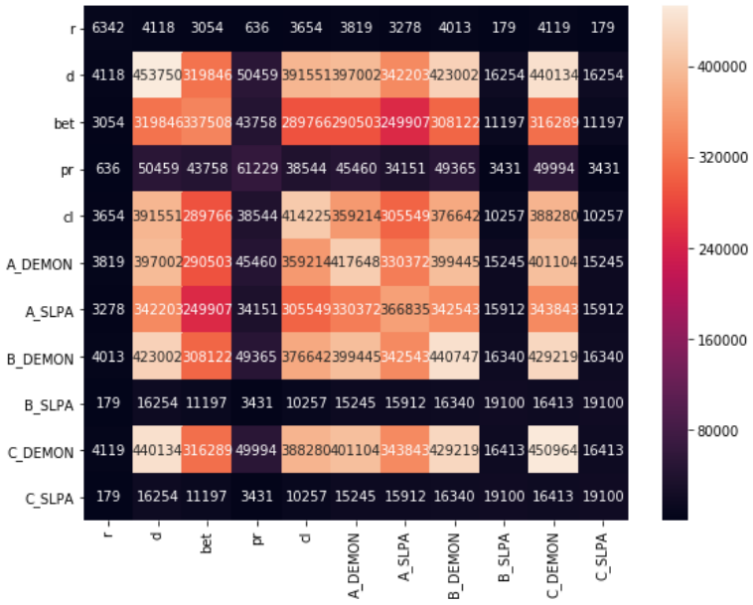

Figura 6. Relação entre Arestas da rede para a rede sign_epinions

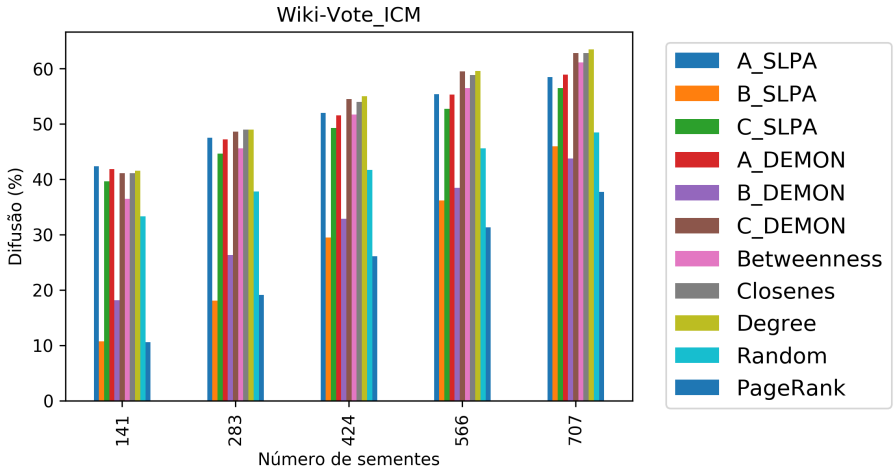

Figura 7. Resultado da difusão da rede Wiki-Vote

Na Figura 7 pode-se perceber uma similaridade entre o comportamento tanto das métricas de centralidade quanto dos critérios baseados em comunidades sobrepostas para as duas redes anteriores. Apesar de outro contexto, a Wiki-Vote também é uma rede de confiança entre os nós. O melhor desempenho neste caso, é da medida de grau, seguida respectivamente, pelo Closenees e o critério C_D. Vale ressaltar que das redes do estudo é aquela com o grau médio mais alto. Assim, o resultado é similar entre a métrica de centralidade de grau e o C_D era esperado. Pode-se observar que o número de arestas entre elas são similares. A métrica closenees é a que possui maior número de arestas, dessa forma, se pensarmos no custo, seria mais barato utilizar o critério C_D para difundir uma marca ou produto do que usar o closeness.

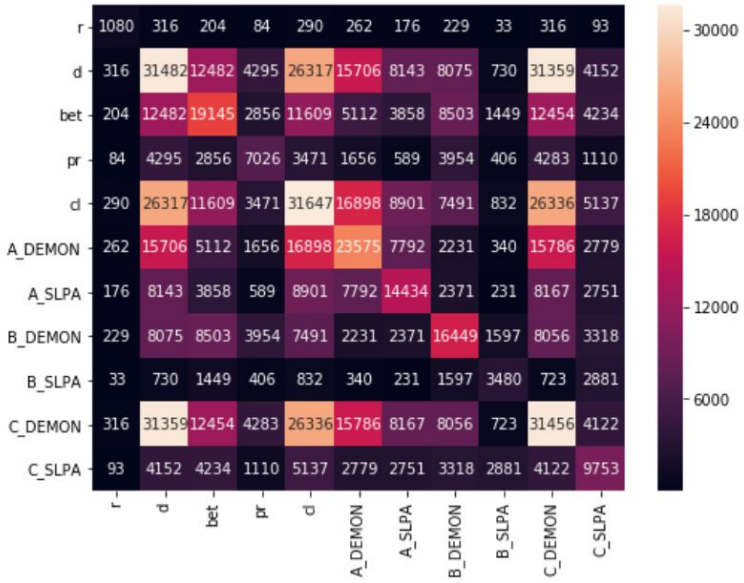

Figura 8. Relação entre Arestas da rede Wiki-Vote

5. Conclusões

Baseado na hipótese de que *nós* presentes na sobreposição das comunidades podem ser eficientes em propagar um comportamento em uma rede, haja vista que podem influenciar seus pares em mais de um grupo, e que as sobreposições existem no contexto real, este trabalho propôs três diferentes critérios baseados em comunidade sobrepostas para avaliar o desempenho desses nós no processo de difusão em redes utilizando o *Indepedent Cascade Model* para simular a difusão, tendo como linha de base sementes baseadas em centralidade.

Tendo em vista os resultados encontrados, elementos em sobreposições de grupos são promissores, mas não garantem o alcance máximo no processo de difusão. Nos experimentos realizados algumas métricas de centralidade obtiveram resultados melhores, mas por muitas vezes, trariam um maior custo para iniciar o processo.

Comparando-se o desempenho dos conjuntos com atributos topológicos da rede, observamos que o PageRank só teve bons resultados na rede GrQc, que possui um alto coeficiente de clustering, ao contrário do grau e closeness, que tiveram seus piores resultados justamente nessa rede. Entre os critérios baseados em comunidade, o algoritmo de detecção Demon obteve resultados muito próximos dos melhores resultados na maioria dos casos. Os critérios que utilizaram o algoritmo SLPA foram mais instáveis e não se mostrou melhor em nenhum cenário.

Em [Vieira et al. 2019] os autores apontam que a forma que os algoritmos detectam comunidades com sobreposição pode interferir mais nos resultados do que a estrutura da rede. Ou seja, os algoritmos podem não ser tão eficientes para encontrar a sobreposição, e cada um encontra um conjunto diferente, mostrando um grande viés dos métodos. No trabalho de [Zhou et al. 2017], embora se trate de agrupamento de arestas, o mesmo problema é colocado, ou seja, a dificuldade em detectar grupos reais, os autores apontam ainda a possibilidade de que seja encontrado um número excessivo de sobreposições.

Sendo assim, em trabalhos futuros pode-se agregar a heurística de [Chen et al. 2009] *Degree Discount* aos elementos encontrados nas sobreposições; realizar uma análise aprofundada das comunidades encontradas, de maneira que melhorias sejam propostas aos critérios; combinar critérios e variar o hiper-parâmetros dos algoritmos utilizados. Pretende-se ainda aplicar os critérios propostos em redes gerados por modelos de rede com comunidades sobrepostas para eliminar o viés trazido pelos algoritmos de detecção de comunidades. Deseja-se ainda, propor uma função de avaliação dos conjuntos que envolva, além da maximização do alcance, a minimização do custo de iniciar o processo.

Referências

Aghaee, Z., Ghasemi, M. M., Beni, H. A., Bouyer, A., and Fatemi, A. (2021). A survey on meta-heuristic algorithms for the influence maximization problem in the social networks. *Computing*, 103(11):2437–2477.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM.

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208.

Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D. (2012). Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623.

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.

Easley, D., Kleinberg, J., et al. (2010). *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge.

Goyal, A., Lu, W., and Lakshmanan, L. V. (2011). Celf++ optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New journal of Physics*, 12(10):103018.

Gulati, R. (1998). Alliances and networks. *Strategic management journal*, 19(4):293.

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.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New journal of physics*, 11(3):033015.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814.

Rossetti, G., Milli, L., and Cazabet, R. (2019). Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):52.

Rossetti, G., Milli, L., Rinzivillo, S., Sîrbu, A., Pedreschi, D., and Giannotti, F. (2018). Ndlib: a python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*, 5(1):61–79.

Salavati, C., Abdollahpouri, A., and Manbari, Z. (2018). Bridgerank: A novel fast centrality measure based on local structure of the network. *Physica A: Statistical Mechanics and its Applications*, 496:635–653.

Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E., and Guo, R. (2015). *Diffusion in Social Networks*.

Sumith, N., Annappa, B., and Bhattacharya, S. (2018). Influence maximization in large social networks: Heuristics, models and parameters. *Future Generation Computer Systems*, 89:777–790.

Vieira, V. d. F., Xavier, C. R., and Evsukoff, A. G. (2019). Comparing the community structure identified by overlapping methods. In *International Conference on Complex Networks and Their Applications*, pages 262–273. Springer.

Xie, J., Szymanski, B. K., and Liu, X. (2011). Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *2011 ieee 11th international conference on data mining workshops*, pages 344–349. IEEE.

Yang, P.-L., Xu, G.-Q., Yu, Q., and Guo, J.-W. (2020). An adaptive heuristic clustering algorithm for influence maximization in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(9):093106.

Zhou, X., Liu, Y., Wang, J., and Li, C. (2017). A density based link clustering algorithm for overlapping community detection in networks. *Physica A: Statistical Mechanics and its Applications*, 486:65–78.