UNIVERSIDADE REGIONAL DE BLUMENAU
CENTRO DE CIÊNCIAS EXATAS E NATURAIS
CURSO DE CIÊNCIA DA COMPUTAÇÃO – BACHARELADO

GERAÇÃO DE REDES COMPLEXAS COM COMUNIDADES SOBREPOSTAS E COMUNIDADES HIERÁRQUICAS

GUSTAVO HENRIQUE SPIESS

BLUMENAU
2022

GUSTAVO HENRIQUE SPIESS

GERAÇÃO DE REDES COMPLEXAS COM COMUNIDADES SOBREPOSTAS E COMUNIDADES HIERÁRQUICAS

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Ciências da Computação no Centro de de Ciências Exatas e Naturais da Universidade Regional de Blumenau como requisito parcial para a obtenção de grau de Bacharel em Ciências da Computação.

Professor Aurelio Faustino Hoppe, Mestre - Orientador

BLUMENAU
2022

FOLHA DE ASSINATURAS

Dedico esse trabalho a minha noiva, cuja paciência em me ouvir falar desse trabalho tornou-o possível.

AGRADECIMENTOS

A meu padrinho, Maiko Rafael Spiess, pelo sempre presente incentivo ao estudo.

Ao meu orientador, Aurélio Faustino Hoppe, por acreditar na conclusão desse trabalho.

A minha família, por todos os anos de apoio que foram necessários para chegar até aqui.

Aos amigos que fiz no percurso do bacharelado, pelo apoio recebido.

Aos professores do Departamento de Sistemas e Computação da Universidade Regional de Blumenau por suas contribuições durante os semestres letivos.

"Se eu vi mais longe, foi por estar sobre ombros de gigantes."

Isaac Newton

RESUMO

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Palavras-chave: Redes complexas. Geração de redes complexas. Comunidades. Comunidades sobrepostas. Comunidades hierárquicas.

ABSTRACT

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Keywords: Complex networks. Complex networks generation. Communities. Overlapping communities. Hierarchical communities

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IE – Internet Explorer

QE – Qternet Explorer

SUMÁRIO

1 INTRODUÇÃO

Redes complexas, como definido por Metz et al. (2007), são grafos com uma topologia não trivial. Isso é, são grafos onde parte ou toda a informação de interesse está contida não nos vértices e arestas individualmente, mas em propriedades do conjunto de vértices e arestas.

Como apontado por Girvan e Newman (2002), um dos sistemas do mundo real que se pode modelar em uma rede complexa é o conjunto de relações sociais. Uma modelagem simplista desse sistema apresenta é a representação de cada indivíduo como um vértice, e vértices adjacentes sendo pares de indivíduos que se conhecem. Nesse tipo de sistema um sub grafo completo, denominado clique (FORTUNATO, 2010), pode ser interpretada como uma propriedade relevante a indicação de que desse conjunto de indivíduos onde todos conhecem todos.

Girvan e Newman (2002) também aponta que outros sistemas, como cadeias alimentares, cadeias de metabolização, redes de transmissão elétrica e redes de computadores podem ser representadas como redes complexas. Muitas vezes propriedades que se observam em redes complexas de um domínio estão presentes também nas redes complexas de outros domínios, mas com interpretações distintas sobre o objeto modelado. O trabalho de Fortunato (2010) indica isso na discussão de múltiplas interpretações do que constitui uma comunidade em uma rede complexa, dividindo-se principalmente em características estruturais, e por semelhança de vértice.

No exemplo do trabalho desenvolvido por Largeron et al. (2015), as duas interpretações se encontram presentes, como é característica da literatura a despeito da geração de redes complexas. Largeron et al. (2015) descreve o que é chamado na literatura de um modelo de geração algorítmica de redes complexas onde os vértices do grafo estão dispostos em uma nuvem de ponto e a distribuição deles em diferentes comunidades leva em conta sua posição espacial, e as arestas são construídas em função desse pertencimento a uma comunidade. Akoglu e Faloutsos (2009) descreve um modelo mais primitivo, que não realiza a atribuição explicita de comunidades, mas que gera um grafo com essas comunidades ainda assim.

Indica-se, observando o trabalho de Fortunato (2010) de que ha uma vasta literatura a respeito dos processos de detecção dessas comunidades. Oberando-se a literatura da qual os trabalhos de Largeron et al. (2015), Akoglu e Faloutsos (2009) e Slota et al. (2019), é indicada a existência dos modelos necessários para a geração de redes complexas com comunidades. No entanto propriedades adjacentes a presença de comunidades para os quais

existe literatura a respeito da detecção, como comunidades hierárquicas e comunidades sobrepostas, parecem estar pouco presentes em modelos de geradores de redes complexas.

1.1 OBJETIVOS

Dado esse contexto, o objetivo do trabalho é a adaptação dos modelos presentes na literatura de geração de redes complexas para a incorporação de comunidades sobrepostas e comunidades hierárquicas.

Os objetivos específicos são:

a) A construção de um modelo algorítmico de geração de redes complexas que inclua a propriedade de comunidades.

b) A especificação, dentro desse modelo, de uma *ground truth* de quais vértices pertencem a quais comunidades.

c) A possibilidade, dentro desse modelo, de comunidades hierárquicas.

d) A possibilidade, dentro desse modelo, de comunidades sobrepostas.

e) A representação, dentro desse modelo, dos vértices como uma nuvem de pontos, para a definição de semelhança de vértices por distância.

1.2 ESTRUTURA

Esse trabalho se estrutura em quatro capítulos sendo o primeiro uma introdução aos temas abordados, bem como a apresentação dos objetivos do trabalho.

O segundo capítulo apresenta a fundamentação teórica da pesquisa, descrevendo o estado da arte do objeto de estudo.

O terceiro capítulo discute o desenvolvimento do modelo algorítmico proposto, incluindo ferramentas e técnicas utilizadas. Também são apresentados os blocos de pseudo código do modelo.

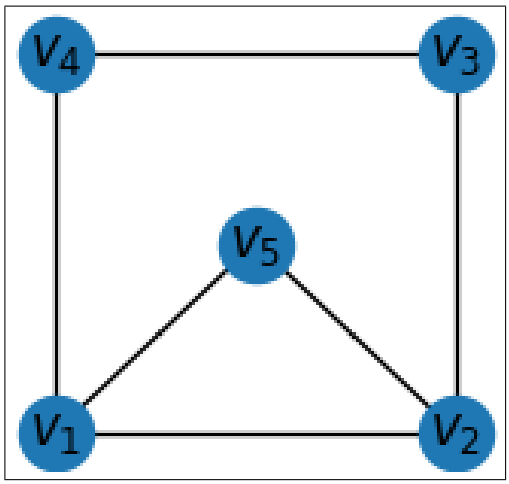
O quarto capítulo compõe os dados obtidos na avaliação dos resultados, bem como quaisquer discussões de implementações futuras ou outras formas de continuação.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 REDES COMPLEXAS E COMUNIDADES

Grafos podem trivialmente ser definidos como $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ onde $\mathcal{V}$ é um conjunto dos vértices de $\mathcal{G}$ e $\mathcal{E}$ é um conjunto de pares não ordenados de vértices ajacentes em $\mathcal{G}$, i.e. as arestas.

Figura 1 – Exemplo de grafo



Fonte: elaborado pelo autor

No exemplo da Figura 1, pode se representar o mesmo grafo com $\mathcal{V} = \{v_1, v_2, v_3, v_4, v_5\}$ e $\mathcal{E} = \{\{v_1, v_2\}, \{v_2, v_3\}, \{v_3, v_4\}, \{v_4, v_1\}, \{v_1, v_5\}, \{v_2, v_5\}\}$. Dentro desse grafo, o subgrafo formato pelos vértices v_1, v_2 e v_5 é também um grafo completo. A esse conjunto de vértices que forma um subgrafo completo dá-se o nome de clique (FORTUNATO, 2010). Essa descrição de clique é utilizada como uma definição inicial do conceito de comunidade, partindo de uma perspectiva local (FORTUNATO, 2010). Dentro desse conceito de comunidade, nomeia-se as arestas cujos dois vértices estão contidos em uma comunidade como sendo interna á comunidade.

É facilmente observável, no entanto, que essa é uma definição muito limitante de comunidade, é raro que comunidades de pessoas apresentem tanta homogeneidade a ponto de todos os membros conhecerem todos os outros membros. De fato, Fortunato (2010) indica que a definição precisa do que é uma comunidade varia de acordo também com o contexto de estudo, mas que algumas características são universais. Uma comunidade, dentro de qualquer definição, deve ser um sub grafo conexo, por exemplo (FORTUNATO, 2010).

Algumas das definições alternativas de o que é uma comunidade podem ser expressas.

Largeron et al. (2015) define comunidade como uma classe de estrutura topológica comum a redes complexas, essas comunidades são categorizadas por terem uma densidade de vértices elevada. O trabalho de Shen et al. (2009) implica que comunidades sejam estruturas que contenham múltiplos cliques dentro de si, e que essas comunidades se dispões em uma estrutura recursiva. Akoglu e Faloutsos (2009) descreve comunidades como estruturas modulares, onde nodos de um vértice formam grupos distintos entre si por que os membros do grupo tem maior chance de estarem conectados entre si do que estarem conectar com membros de outros grupos. Girvan e Newman (2002) define "Cluster" e comunidade como duas propriedades distintas, o primeiro sendo a probabilidade de dois nodos ambos adjacentes a um terceiro serem também ajacentes entre si, e a segunda como sendo condutos de vértices densamente conectados entre si, e esparsamente conectados para além de si.

Todas essas definições apresentam características específicas relevantes para a aplicação em que foram utilizadas. As definições são agrupadas em três classes distintas por Fortunato (2010):

a) Definição local

b) Definição global

c) Definição por similaridade de vértice

Essas definições não são mutuamente exclusivas, mas também não são ortogonais uma a outra. Segundo Fortunato (2010), a definição local parte das características topológicas internas á comunidade. Nominalmente, isso significa a existência de um conjunto considerável de arestas internas a comunidade e um conjunto limitado de arestas para além da comunidade.

A definição global de comunidades é aplicada aos casos onde a presença de clusters é uma característica inerente ao grafo que se está estudando Fortunato (2010). Essa propriedade inerente ao grafo pode ser definida como alguma propriedade dos vértices do objeto em questão e que partindo disso se atribui pertencimento á comunidades, ou ainda por comparação com um exemplo nulo Fortunato (2010). No caso de comparação com um exemplo nulo, define-se uma comunidade pela característica de uma não ser presente dentro de o que é chamado de "grafo aleatório" (FORTUNATO, 2010). Essa definição de um modelo nulo é crucial para o trabalho de Girvan e Newman (2002), o modelo nulo considerado é um onde o grafo original é alterado de forma aos graus de todos os vértices se manterem, mas a probabilidade de dois vértices estarem ligados é constante independente de quais os vértices.

A definição de comunidade por similaridade de vértice se baseia na tendencia de que em muitas aplicações, membros de comunidades são mais similares entre si do que seria esperado de um conjunto do mesmo tamanho escolhido aleatoriamente (FORTUNATO, 2010). Essa definição se faz visível no trabalhos de Akoglu e Faloutsos (2009) e de Largeron et al. (2015). Na observação desses dois trabalhos também é interessante o questionamento de como se define semelhança, Akoglu e Faloutsos (2009) representa os vértices como sequencias de caracteres de tamanhos variáveis em que a probabilidade de dois vértices estarem ligados é maior conforme mais caracteres eles compartilham; e Largeron et al. (2015) representa os vértices como pontos em um espaço n-dimensional e define que vértices são mais semelhantes quando a distância euclideana deles é menor.

Também independente de qual definição de comunidade que se esteja utilizando, existem os conceitos de partição e cobertura. Segundo Fortunato (2010), uma partição é uma divisão dos vértices de um grafo tal que cada vértice pertença a um e exatamente um cluster. O caso de um vértice "livre", não pertencendo a nenhuma comunidade, é trivialmente resolvido incluindo ele á comunidade com a qual ele mais tem jacências. Mas o caso de vértices que pertençam a mais de uma comunidade, i.e. comunidade que se sobreponham, é mais interessante. Fortunato (2010) define uma cobertura como uma divisão dos vértices em clusters onde cada vértice pertence a um ou mais clusters. Fortunato (2010) também descreve o conceito de comunidades hierárquicas, como sendo comunidades cuja estrutura interna também se organiza em clusters de escala menor do que o original.

Fortunato (2010) oferece também o conceito de "função de qualidade", sendo uma função que mapeia uma partição para um espaço de comparação, usualmente em números reais, onde partições que mapeiem para valores maiores são consideradas melhores. Segundo Fortunato (2010) função de qualidade mais comumente utilizada é a modularidade Q de Girvan e Newman (2002).

Quadro 1 – Função de modularidade Q

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{K_i K_j}{2m} \right) \delta(C_i, C_j)$$

Fonte: Girvan e Newman (2002)

Essa função no entanto não se aplica adequadamente ao caso de comunidades sobrepostas ou comunidades hierárquicas, para tanto, é necessário utilizar a função de

modularidade estendida, conforme desenvolvido por Shen et al. (2009).

Quadro 2 – Função de modularidade estendida EQ

$$EQ = \frac{1}{2m} \sum_{i} \sum_{v \in C_i, w \in C_i} \frac{1}{O_v O_w} \left[A_{vw} - \frac{K_v K_w}{2m} \right]$$

Fonte: Shen et al. (2009)

Tanto no caso da formula descrita no Quadro 1 quanto na do Quadro 2 a função definida é uma somatórias em que alguns termos se repetem. Primeiramente, é preciso descrever que a função $\delta(C_i C_j)$ retorna um se C_i for igual a C_j, e zero noutro caso (FORTUNATO, 2010). Considerando isso, no caso de uma partição (sem comunidades hierárquicas, e sem comunidades sobrepostas), as duas somatórias iteram sobre os mesmos valores, a primeira com os vértices i e j e a segunda com os vértices v e w.

Essa iteração olha para todos os pares de vértices que compartilham alguma comunidade, e soma o valor de A_{ij}, sendo A a tabela de adjacência do grafo em questão. Então é subtraído um valor $K_i K_j / 2m$, onde K_i é o grau do vértice i e m é a quantidade de arestas no grafo ($2m$ portanto é a soma dos graus de todos os vértices). Esse valor é o a probabilidade de uma aresta entre os vértices i e j no modelo nulo de Girvan e Newman (2002), considerando que os graus se mantém mas que a probabilidade da presença de uma aresta é uniforme.

A formula EQ de Shen et al. (2009) no entanto contém também o termo escalar $1/O_v O_w$, nesse caso o valor O_i é a quantidade de comunidades a qual pertence o vértice i. Isso permite a aplicação da modularidade estendida para os casos de grafos com comunidades sobrepostas. Vértices que estejam em duas comunidades contribuirão para a modularidade a partir das duas, mas tendo a magnitude da sua contribuição escalada à metade.

2.2 OUTRAS PROPRIEDADES DE REDES COMPLEXAS

Para além da presença de estruturas topológicas que podem ser denominas comunidades, redes complexas tem algumas propriedades topológicas bastante comuns e relevantes. São algumas delas:

2.2.1 Mundo pequeno, Anexação preferencial e Liberdade de escala

Largeron et al. (2015) descreve a propriedade de mundo pequeno como a característica de um sistema de ter um diâmetro loga ritmicamente proporcional a quantidade de vértices em um grafo. Isso é, a distancia entre os dois vértices que estão a mais arestas de distância, denominada diâmetro, cresce logaritmicamente conforme observamos exemplos maiores de grafos do sistema. Essa propriedade implica que em sistemas bastante grandes, é preciso uma quantidade relativamente pequena de saltos de nodo a nodo para se atingir qualquer membro do grafo.

Largeron et al. (2015) define a anexação preferencial como uma propriedade de um sistema em que vértices tendem a se ligar com outros vértices que sejam parecidos e que tenham grau elevado. Essa propriedade encontra-se presente também no trabalho Slota et al. (2019). Em ambos os casos, a implicação é que dado um sistema onde se vai adicionar um vértice, a maior parte das arestas desse novo vértice devem ligá-lo a outro com grau igual ou maior do que o próprio.

Para atingir essa distribuição característica o modelo de Slota et al. (2019) faz com que os vértices se dividam em diferentes escalas, de forma que os vértices de uma escala se liguem apenas entre si e com os membros das escalas imediatamente vizinhas. De grafos com essa distribuição onde o grau relativo de dois vértices adjacentes tende a não apresentar saltos demasiadamente grandes, se diz que são livres de escala (LARGERON et al., 2015).

Essas e uma série de outras proporcionalidades são comumente encontradas em redes complexas. Outros exemplos de proporcionalidades conhecidas na literatura envolvem os pesos das arestas de um vértice, a proporção de vértices e arestas do grafo e a já citada distribuição dos graus, todas são leis de potência (AKOGLU; FALOUTSOS, 2009).

2.2.2 Cluster e comunidades

Girvan e Newman (2002) diferencia explicitamente entre a definição de clusters e de comunidades. No trabalho seminal os autores apontam um cluster como sendo um triângulo, um em outras palavras um subgrafo completo com três vértices. Essa definição aparentemente arbitrária é relevante no entendimento do coeficiente de clusterização:

A conceitualização de um cluster é relevante dentro do estudo de redes complexas na medida em que a implicação é de dois vértices serem ligados por compartilharem uma relação com um terceiro. O coeficiente C sendo igual a 1 implica que o grafo é um grafo

Quadro 3 – Coeficiente de clusterização C

$$C = \frac{3 \times (\text{número de triângulos do grafo})}{(\text{número de triplas conexas do grafo})}$$

Fonte: Girvan e Newman (2002)

completo (GIRVAN; NEWMAN, 2002). Mais que isso, esse coeficiente, dado um vértice, é a probabilidade de quaisquer dois vértices adjacentes a ele serem adjacentes entre si.

2.2.3 Homofilia e Homogeneidade de comunidades

Conforme foi discutida quanto a definição de comunidades por semelhança de vértices, é de se esperar que vértices adjacentes compartilhem características. A essa preferência se dá o nome de homofilia (AKOGLU; FALOUTSOS, 2009). Essa definição de semelhança é deliberadamente vaga, pois dentro de sistemas distintos é trivial imaginar funções de compatibilidade distintas. Independentemente disso, observa-se que em grafos obtidos observando sistemas do mundo real, não raro as arestas tornam ajacentes vértices que otimizam alguma função de proximidade, (LARGERON et al., 2015).

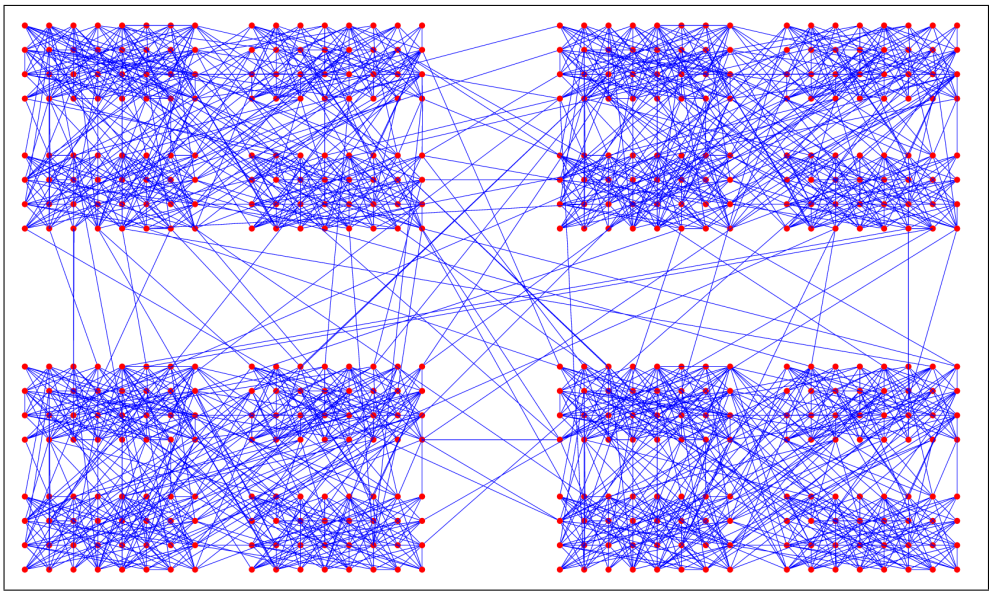
Homofilia como propriedade é intimamente ligada a uma outra característica que se observa de grafos do mundo real, em diferentes aplicações as comunidades tendem a ser mais homogêneas do que o grafo ao qual pertencem (LARGERON et al., 2015). Pode-se afirmar que um grafo onde isso ocorra tem a propriedade de “comunidades homogêneas”.

O exemplo da Figura 2 demonstra como a homofilia as vezes pode ter a presença visualmente verificada. Se considerarmos que a posição dos vértices na imagem corresponde a duas características ortogonais, a distância entre os vértices pode ser interpretada como uma função de similaridade de dois vértices. Nesse caso é intuitivamente entendido que vértices mais parecidos de conectam mais do que vértices mais dissemelhantes.

2.2.4 Agrupamentos hierárquicos e sobreposições

No exemplo de grafo da Figura 2 é possível demonstrar um entendimento intuitivo de como comunidades se organizam. A estrutura topológica de grupos densamente conectados fica visualmente identificável, onde cada quadrante contém uma comunidade coesa. Também visualmente acessível, cada comunidade desse exemplo tem uma estrutura interna auto semelhante.

Figura 2 – Exemplo de grafo com comunidades hierárquicas



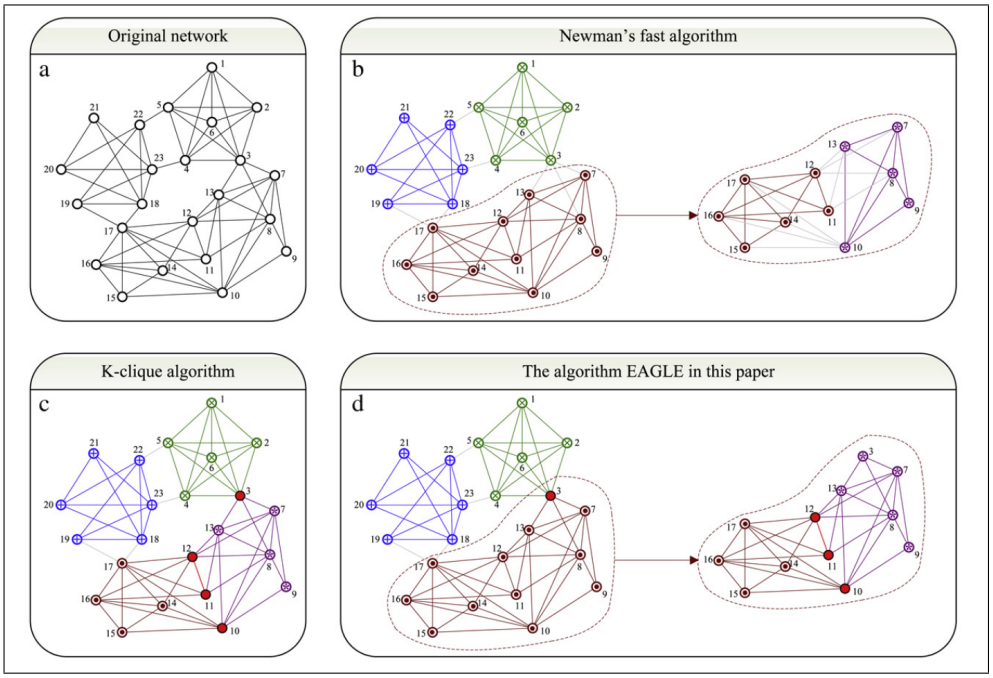
Fonte: Fortunato (2010)

Essa construção de estruturas topológicas recursivas é denominada por Girvan e Newman (2002) como “meta grupo”, onde as propriedade topológicas relativas a agrupamentos podem ser encontradas se repetindo em em escalas menores dentro das componentes de escalas maiores. Comunidades podem funcionalmente ser compostas por comunidades menores. Esse mesmo conceito recebe uma outra nomenclatura nos trabalhos de Largeron et al. (2015), Shen et al. (2009) e Fortunato (2010), onde são descritas como comunidades hierárquicas.

Essas estruturas seguem uma característica recursiva, como demonstrado pelo processo de detecção proposto por Shen et al. (2009), podendo ser concebidos exemplos de sistemas com qualquer sorte de diferentes níveis. E a elas também se aplica a compreensão de partição ou cobertura, na Figura 2 as comunidades de primeiro e de segundo nível caracteristicamente não compartilham vértices. No caso do que demonstra Shen et al. (2009) não só é possível que um vértice pertença a duas comunidades, é possível que ele pertença a duas comunidades de níveis distintos. Na Figura 3, os resultados de Shen et al. (2009) são demonstrados no quadro a respeito do algoritmo EAGLE, o vértice denotado como 3 é compartilhado entre duas comunidades de primeira ordem, mas em uma delas o vértice 3 encontra-se como membro de uma comunidade de segunda ordem.

Ressaltando que a exata definição de comunidade é altamente dependente do contexto (FORTUNATO, 2010), parece ser consenso na literatura que quando se consideram comunidades hierárquicas, todos os membros de uma comunidade de primeiro nível, devem

Figura 3 – Demonstração dos resultados de diferentes algoritmos de detecção em um grafo com comunidades hierárquicas e com sobreposição



Fonte: Shen et al. (2009)

fazer parte de uma das comunidades que compõe a primeira, como observado nos trabalhos de Fortunato (2010) e Shen et al. (2009). I.e.: nenhum vértice pertence exclusivamente a uma comunidade sem pertencer a alguma das sub comunidades. Alternativamente claro, o exemplo da Figura 3 mostra que a implementação de Girvan e Newman (2002) (quadrante superior direito) é capaz de produzir partições recursivas (note-se a distinção entre uma cobertura e uma partição).

Essa distinção entre cobertura e grafo implica também na definição de comunidades sobrepostas. Em sistemas do mundo real que produzem redes complexas, é bastante natural que comunidades compartilhem vértices, pois não raro alguma parte de um sistema é componente em dois grupos estruturalmente significantes (SHEN et al., 2009). Diz-se de duas comunidades que compartilham vértices que elas são comunidades sobrepostas sobrepostas.

O método de detecção de comunidades por K-cliques oferece alguma inspiração no entendimento das propriedades de comunidades sobrepostas. Fortunato (2010) descreve que a forma como esse método trabalha é pivotando subgrafos completos do grafo. Isso é, dado que um K-clique é um subgrafo completo de k vértices, de dois k-cliques compartilham $k-1$ vértices, eles devem de fazer parte da mesma comunidade. Armado desse conhecimento, é possível desenhar que uma cobertura ideal deveria priorizar comunidades com

grandes subgrafos completos internamente, mas de que os vértices da intersecção de duas comunidades deveriam de participar de k-cliques distintos, preferencialmente não estando anexos. A ideia por de trás disso é que se a intersecção de duas comunidades deveria ser parte da periferia das respectivas comunidades (FORTUNATO, 2010). Se a intersecção fosse tão densamente conexa quando o centro das duas comunidades, esses vértices não seriam mais valore intersectados entre duas comunidades distintas, e as comunidades seriam uma apenas.

2.3 O ESTADO DA ARTE EM GERAÇÃO DE REDES COMPLEXAS

Existe uma literatura muito prolífica de aplicações dos conceitos de redes complexas, como por exemplo o trabalho de Stegehuis, Hofstad e Leeuwaarden (2016), que faz uma análise do espalhamento de doenças em uma rede com comunidades, onde é demonstrado que a presença das comunidades tem um efeito significativo. Muitos métodos para a detecção de comunidades foram propostos, como fica evidente na ampla revisão feita por Fortunato (2010). É bastante mais escassa a literatura de modelos e metodologias capazes de gerar redes complexas onde se tenha a presença de redes complexas.

2.3.1 RTG: a recursive realistic graph generator using random typing

Akoglu e Faloutsos (2009) apresenta um trabalho seminal no que tange a geração de redes complexas. Como foi discutido anteriormente, o trabalho desenvolvido não apenas gera redes complexas com uma série de proporções que conhecidamente ocorrem em sistemas do mundo real, possibilitando a construção de grafos que se assemelhem aos produzidos pelos sistemas do mundo real, mas produz grafos com a presença de comunidades.

A implementação realizada por Akoglu e Faloutsos (2009) se baseia em um gerador de arestas que tem as probabilidades tendenciosas. Esse gerador é o que os autores chamam de um teclado recursivo, na realidade é uma matriz de possibilidades de escolha de uma característica discreta para a origem e o destino em simultâneo. Nominalmente, os vértices são uma sequência aleatória de caracteres de um conjunto finito de possibilidades, é repetidamente escolhido uma letra para o destino e uma para a origem em simultâneo. Com um parâmetro controlando um reforço que é feito para que a célula da matriz escolhida a cada interação seja da diagonal principal, existe uma tendencia de que vértices adjacentes tenham as mesmas letras nas mesmas posições.

Essas regras bastante simples são o bastante para que o modelo de Akoglu e Faloutsos (2009) tenha como emergentes algumas das propriedades desejáveis em um

modelo de geração de redes complexas para além das proporções, mundo pequeno, anexação preferencial e homofilia, o sistema de Akoglu e Faloutsos (2009) gera comunidades homogêneas.

Existe também um interesse relevante em quanto os algoritmos de geração de redes complexas precisam em tempo. Nesse quesito, a implementação de Akoglu e Faloutsos (2009) apresenta algumas das características mais desejáveis, ele é totalmente paralelizável, significando que para a construção de um grafo com o dobro de arestas, é possível dobrar a quantidade de recursos de processamento e assim dobrar a quantidade de arestas produzidas em um tempo constante.

2.3.2 Generating Attributed Networks with Communities

Largeron et al. (2015) apresenta um modelo algorítmico de geração de redes complexas com uma abordagem da anteriormente discutida. O processo realizado é bastante mais explícito e deliberado em quais condicionantes são utilizadas para afetar quais propriedades. Uma faze inicial gera uma nuvem de pontos com uma distribuição aleatória e uma amostra dessa população é usada para inicializar as comunidades. Essa amostragem é processada em um algoritmo k-means, gebrando os clusters iniciais, e as arestas iniciais são geradas. Uma segunda faze processa os demais vértices, escolhendo qual a comunidade ao qual serão inseridos se baseando na distância euclideana (homofilia), e gerando os vértices baseado na distribuição de graus. Numa última faze, opcional, é realizada a introdução de novas arestas, essas arestas são escolhidas de forma a fecharem triângulos, para aumentar o coeficiente de clusterização C.

Notadamente, a modelagem de grafo utilizada por Largeron et al. (2015) é $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ onde $\mathcal{A}$ é um conjunto de atributos dos vértices vértices, de forma que cada $v \in \mathcal{V}$ é um vetor de valores v_A. A modelagem tem acompanhando o grafo também um conjunto $\mathcal{P}$, composto por conjuntos de vértices, isolando assim as partições. Essa abordagem oferece uma funcionalidade bastante desejável, que é gerar um catálogo de qual vértice pertence a qual comunidade, dessa forma criando um "ground truth" contra o qual o desempenho de alguns algoritmos pode ser testado. Considerando os trabalhos que utilizam análise de redes complexas como Stegehuis, Hofstad e Leeuwaarden (2016), a possibilidade regerar um grafo com características topológicas conhecidas, podendo-se manipular o coeficiente de clusterização por exemplo, existem um conjunto de possíveis análises com relevância acadêmica. Para muitas dessas análises, as possibilidades de parametrização modelo de Largeron et al. (2015) parece ser interessante. É feita também uma discussão

de performance por parte de Largeron et al. (2015), mas é possível desenhar algumas críticas a forma como o modelo foi desenhado, que previnem a paralelização do processo de construção dos grafos.

2.3.3 Modelos dinâmicos de geração de redes complexas com comunidades

Para além do modelo trivial $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, é possível a concepção de um grafo dinâmico representado em T estados, $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ for $i \in \{1, \ldots, T\}$. Benyahia et al. (2016) apresenta uma implementação dinâmica do que foi apresentado por Largeron et al. (2015). Outras implementação de um gerador de grafos dinâmicos são disponibilizada por Duan et al. (2019) e Luo et al. (2020).

3 MODELO

O que este trabalho propõe é um modelo algorítmico de geração de redes complexas que produza grafos mais realistas. Realista nesse contexto é entendido como tendo a presença de um conjunto de propriedades que se observam em grafos do mundo real. A proposta desse trabalho é a extensão do modelo de Largeron et al. (2015) para a construção de grafos em que se apresentem também as propriedades de comunidades hierárquicas e comunidades sobrepostas. A modelagem proposta se baseia na construção de uma cobertura (em oposição á construção de uma partição) recursiva, de forma semelhante ao que é produzido pelo algoritmo de detecção proposto por Shen et al. (2009).

3.1 HIPÓTESE

3.1.1 A Representação do grafo

O model gera um grafo com atributos e uma cobertura de comunidades hierarquicamente dispostas.

Quadro 4 – Modelagem de grafo com comunidades hierárquicas e sobrepostas

$$
\begin{aligned}
&\mathcal{G} = (\mathcal{V}, \mathcal{E}, C_n) \\
&\mathcal{V} \subset \mathbb{Q}^p \\
&\mathcal{E} \subset \{\, v_1, v_2 \mid v_1 \in \mathcal{V}, v_2 \in \mathcal{V} \,\} \\
&C_0 \subset \{\, v \mid v \in \mathcal{V} \,\} \\
&C_n = \{\, C_{n-1_0}, C_{n-1_1}, \ldots, C_{n-1_m} \,\}
\end{aligned}
$$

Fonte: elaborado pelo autor

Os vértices em $\mathcal{V}$ são uma nuvem de pontos em um espaço de p dimensões, cada um sendo um vetor com p valores racionais (representados com ponto flutuante). As arestas em $\mathcal{E}$ são simples pares não ordenados de vértices. A cobertura C_n é uma estrutura recursiva de conjuntos com n níveis, onde o conjunto C_n é composto por m conjuntos C_{n-1}. No último níveo o conjunto C_0 é composto por vértices do grafo, esses conjuntos são as comunidades *folha*, significando que elas não são compostas por outras comunidades.

Essa estrutura recursiva é a representação das comunidades hierárquicas, onde os vértices que pertencem a uma comunidade C, denotado V_C são os membros do conjunto união dos membros de C, isso é, $\bigcup_{S \in c} V_S$. Como característica dessa modelagem, os vértices da comunidade representada por C_n são a totalidade dos vértices do grafo, portanto

$V_{C_n} = \mathcal{V}$ Isso implica que para qualquer vértice em $\mathcal{V}$, ele deve de estar presente pelo menos uma comunidade folha. Ao estar presente em uma comunidade folha, o vértice é considerando também parte de todas as comunidades compostas por esta comunidade folha.

A cardinalidade de cada um dos conjuntos que formam a cobertura é variável de acordo com o nível, isso é, dado um nível x todas as coberturas C_{n-x} possuem a mesma quantidade de elementos, mas conjuntos de níveis distintos podem possuir quantidades de elementos distintas. É considerada também a existência de um vetor, denotado K, em um espaço de $n-1$ dimensões, que denota a cardinalidade das coberturas compostas por outras coberturas. Todas as comunidades do grafo $\mathcal{G}$ contém pelo menos um membro.

Uma característica de notação é a função flat que mapeia uma comunidade para um conjunto de quais as comunidades que a compõe. Para efeito de notação flat(C_n) é um conjunto com todas as comunidades do grafo, incluindo a comunidade global. E o mapa L, que relaciona cada comunidade com a quantidade de ancestrais que a comunidade possui, ou seja, com a quantidade de comunidades que ela compõe. As notações utilizadas serão L_C, leia-se nível de C, ou L, leia-se níveis. A raiz da cobertura é o nível zero ($L_{C_n} = 0$) e a o nível de um nó folha é igual a n ($L_{C_0} = n$) As características que devem ser verdadeiras se um grafo for representado nesta modelagem encontram-se descritas na Tabela 1.

3.1.2 Propriedades desejáveis do modelo

Largeron et al. (2015) implementa um modelo algorítmico de geração de redes complexas que mantém uma série de propriedades desejáveis. Como a implementação proposta se baseia no modelo de Largeron et al. (2015), é desejável que as propriedades sejam mantidas. Nominalmente, são elas:

a) Mundo pequeno: O diâmetro das redes complexas geradas pelo modelo deve ter uma relação logarítmica com a quantidade de vértices no modelo.

b) Distribuição de graus em lei de potência: Os graus dos vértices devem estar distribuídos com uma lei de potência.

c) Homofilia: O grafo gerado deve apresentar uma tendência de priorização da adjacência com vértices semelhantes.

d) Estrutura de comunidades: O grafo gerado deve de ter comunidades, conforme etiquetadas na cobertura, de forma que todo vértice pertença a uma ou mais comunidades, e as comunidades se organizem em uma estrutura hierárquica.

Tabela 1 – Características da modelagem

Característica	Formalismo
Para toda a comunidade C, se ela não for folha, a função flat dela é a união de C com a função flat de seus componentes.	$\forall C(L_C < n \implies \text{flat}(C) = C \cup \bigcup_{S \in C} \text{flat}(S))$
Para toda a comunidade C, se ela for folha, a função flat dela é um conjunto consigo.	$\forall C(L_C = n \implies \text{flat}(C) = \{ C \})$
A comunidade raiz engloba todos os vértices do grafo.	$V_{C_n} = \mathcal{V}$
Para todas as comunidades C, se C não for folha, os vértices englobados em C são a união dos vértices englobados em seus componentes	$\forall C(L_C > 0 \implies V_C = \bigcup_{S \in C} V_S)$
Para todas as comunidades C, se C for folha, os vértices englobados em C seus componentes	$\forall C(L_C = 0 \implies V_{C_0} = C_0)$
Para todos os vetores do grafo, existe uma comunidade folha a qual ela pertence	$\forall v(\exists C(v \in C \wedge L_C = n))$
Pra todo vértice v, pra todo l, existe uma comunidade C que contenha o vértice e seja do nível l	$\forall v \forall l(\exists C(v \in V_{C_x} \wedge L_C = l))$
Todos as comunidades não folha tem a mesma quantidade de componentes se forem do mesmo nível, a cardinalidade de uma comunidade não folha é expressa num vetor K	$\exists K(K \in \mathbb{I}^{n-1} \wedge \forall C(L_C < n \implies K_{L_C} = \lvert C \rvert))$
Toda a comunidade tem pelo menos uma componente e engloba pelo menos um vértice	$\forall C(\lvert C \rvert \geq 1 \wedge \lvert V_C \rvert \geq 1)$

Fonte: elaborado pelo autor

e) Comunidades homogêneas: As comunidades devem ser coesas não apenas na perspectiva topológica, mas em similaridade.

Para tanto, a abordagem do modelo é a construção explicita das comunidades com base na similaridade dos vértices. Para isso, a similaridade dos vértices é definida com base na distância euclideana dos vetores de atributos dos vértices. As arestas do grafo são definidas com base nas comunidades das quais o vértice faz parte.

Essa implementação visa garantir a homogeneidade das comunidades e a homofilia

ao selecionar os membros das comunidades estocasticamente preferindo vértices com menor distância euclideana. A construção das arestas é feita priorizando a introdução de vértices a vértices com mais arestas dentro da comunidade, de forma a reforçar a distribuição de graus em lei de potência e a estrutura de comunidade, bem como a propriedade de mundo pequeno.

3.2 ALGORITMO

3.2.1 Parâmetros

As propriedades descritas podem ser controladas utilizando uma série de parâmetros. Os parâmetros seguem descritos na tabela Tabela 2. Eles são uma adaptação bastante direta dos parâmetros do modelo de Largeron et al. (2015), a difere-a mais significativa é no parâmetro K, que é um vetor multi dimensional de inteiros maiores que 1. Isso se deve à construção de uma árvore de comunidades hierarquicamente aninhadas.

Tabela 2 – Características da modelagem

Parâmetro	Descrição
$N \in \{\, n \in \mathbb{N} \mid n \geq 1 \,\}$	Quantidade de vértices.
$E_{\text{wth}}^{\max} \in \{\, i \in \mathbb{N} \mid i \geq 1 \,\}$	Número máximo de arestas (internas a comunidade) inseridas a um vértice ao introduzir ele a uma comunidade.
$E_{\text{btw}}^{\max} \in \{\, i \in \mathbb{N} \mid i \geq 1 \,\}$	Número máximo de arestas (externas a comunidade) inseridas a um vértice ao introduzir ele as comunidades.
$MTE \in \{\, m \in \mathbb{N} \mid m \geq 1 \,\}$	Número mínimo de arestas no grafo produzido.
$\mathcal{A} \in \{\, a \in \mathbb{Q} \mid a > 0 \,\}^{\lvert\mathcal{A}\rvert}$	Vetor de desvios padrão dos atributos dos vértices.
$K \in \{\, k \in \mathbb{N} \mid k \geq 2 \,\}^{\lvert K\rvert}$	Vetor de quantidade de comunidades por nível
$\theta \in \{\, t \in \mathbb{Q} \mid 0 \geq t \geq 1 \,\}$	Valor de interpolação entre homogeneidade por distância euclideana e distância por ortogonalidade de comunidade.
$\text{NbRep} \in \{\, n \in \mathbb{N} \mid n \geq 1 \,\}$	Número de representantes por comunidade.

Fonte: elaborado pelo autor

3.2.2 Inicialização

A primeira faze do algoritmo é a inicialização dos vértices e das comunidades. Conforme definido no Quadro 5.

O processo da inicialização se divide em gerar a nuvem de pontos e inicializar as comunidades. A linha quatro inicializa $\mathcal{V}$ com um conjunto vazio, e o laço de repetição

das linhas 6 á 10 insere vetores neste conjunto enquanto ele tiver menos de N membros. O vetor em si é definido como uma série de distribuições aleatórias com o centro em zero e o desvio padrão informado pelo parâmetro $\mathcal{A}$.

O processo de geração das estruturas de comunidade é mais complexo, exigindo uma função para possibilitar recursividade. A função cover tem como condicionante a característica de se a comunidade que se está processando é folha, isso é, se ela não possuirá subdivisões internas. Na linha 14 é feita essa ramificação, considerando que l, um parâmetro de controle que é incrementado a cada chamada recursiva. Se l for igual á cardinalidade de k, isso indica que estamos no último nível a ser gerado.

O processo para o último nível sendo gerado gera um conjunto de arestas entre os membros da comunidade, recebidos por parâmetro com p. Nas linhas 16 até 20 é iterado sobre os vértices, os vértices com quem é possível formar arestas, nomeado p', são definidos como os vértices em p diferentes de v com quem v não é ajacente. Nas linhas 19 e 20 uma quantidade aleatória das arestas possíveis são construídas.

As funções Rand_{Uni} e Sample são duas funções de escolha aleatória uniformes. $\text{Sample}(P, l)$ escolhe um sub conjunto de P com l elementos uniformemente distribuído, i.e., todos os membros de P tem a mesma chance de estar presente no conjunto construído. $\text{Rand}_{\text{Uni}(P)}$ funciona da mesma forma, mas retorna um único membro de P.

No caso de não ser uma comunidade folha, o processo de construção da comunidade encontra-se nas linhas 24 até 30. Para isso primeiramente é definido um tamanho de amostragem s. Esse tamanho é definido como um produto dos valores de K, filtrando para o nível atual em diante. Com isso, buscasse uma amostra p', com tamanho s ou o valor máximo possível se s for maior que a quantidade de membros em p.

Com essa amostra, é realizado um agrupamentos utilizando o algoritmo K Medoids. Nesses clusters iniciais é realizada a chamada recursiva da função cover, que faz a construção da comunidade composta pelos vértices do cluster. Com as comunidades definidas e agrupadas no conjunto c, que representa a comunidade que se está processando, é realizada a introdução de arestas para que a comunidade seja conexa. Assumindo que todas as comunidades geradas por meio da função cover sejam conexas, é construído um caminho que liga um membro de cada comunidade. A função se conclui retornando a comunidade criada.

A chamada original para a função $\text{cover}(l, p)$ é feita com l sendo zero e p sendo a nuvem de pontos. Por fim, o processo também realiza a atribuição dos representantes de cada comunidade folha como sendo a totalidade dos membros da comunidade, e mantendo

as demais comunidades sem representantes.

Quadro 5 – Faze 1 do modelo

Input: N, $\mathcal{A}$, K, NbRep
Output: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, C)$

$\mathcal{V} \leftarrow \emptyset$
$\mathcal{E} \leftarrow \emptyset$
while $|\mathcal{V}| < N$ **do**
begin
    $v \leftarrow (\mathcal{N}(0, \sigma_{\mathcal{A}_0}), \mathcal{N}(0, \sigma_{\mathcal{A}_1}), \ldots, \mathcal{N}(1, \sigma_{\mathcal{A}_{|\mathcal{A}|-1}}))$
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{ v \}$
end

Function cover(l, p)
begin
    if $l = |K|$ **then**
    begin
        for $v \in p$ **do**
        begin
            $p' \leftarrow \{ v' \in p \mid \{ v, v' \} \notin \mathcal{E} \wedge \{ v, v' \} \notin \mathcal{E} \wedge v' \neq v \}$
            $s \leftarrow \text{Rand}_{\text{Uni}}([1, |p'|])$
            for $v' \in \text{Sample}(p', s)$ **do** $\mathcal{E} \leftarrow \mathcal{E} \cup (v, v')$
        end
        return p
    end
    $s \leftarrow \text{NbRep} \times \prod_{i=l}^{|K|-1} K_i$
    $p' \leftarrow \text{Sample}(p, \min \{ s, \mid p \mid \})$
    $k \leftarrow \text{K Medoids}(p', K_l)$
    $c \leftarrow \{ \text{cover}(l+1, q) \mid q \in k \}$

    $p' \leftarrow \{ \text{Rand}_{\text{Uni}}(c') \mid c' \in V_c \}$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{ \{ p'_i, p'_{i+1} \} \mid i \in \{ 1, 2, \ldots \mid p' \mid -1 \} \}$

    return c
end

$C = \text{cover}(0, \mathcal{V})$
for $c \in \{ c' \in \text{flat}(C) \mid L_{c'} = |K| \}$ **do** $c.\text{rep} \leftarrow c$
for $c \in \{ c' \in \text{flat}(C) \mid L_{c'} \neq |K| \}$ **do** $c.\text{rep} \leftarrow \emptyset$

$\mathcal{G} \leftarrow (\mathcal{V}, \mathcal{E}, C)$
Return $\mathcal{G}$

Fonte: elaborado pelo autor

REFERÊNCIAS

AKOGLU, L.; FALOUTSOS, C. Rtg: A recursive realistic graph generator using random typing. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2009. p. 13–28. Citado 7 vezes nas páginas 14, 17, 18, 20, 21, 24 e 25.

BENYAHIA, O. et al. Dancer: Dynamic attributed network with community structure generator. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2016. p. 41–44. Citado na página 26.

DUAN, B. et al. Dynamic social networks generator based on modularity: Dsng-m. In: IEEE. *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*. [S.l.], 2019. p. 167–173. Citado na página 26.

FORTUNATO, S. Community detection in graphs. *Physics reports*, Elsevier, v. 486, n. 3-5, p. 75–174, 2010. Citado 8 vezes nas páginas 14, 16, 17, 18, 19, 22, 23 e 24.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citado 8 vezes nas páginas 14, 17, 18, 19, 20, 21, 22 e 23.

LARGERON, C. et al. Generating attributed networks with communities. *PloS one*, Public Library of Science, v. 10, n. 4, p. e0122777, 2015. Citado 11 vezes nas páginas 14, 17, 18, 20, 21, 22, 25, 26, 27, 28 e 30.

LUO, W. et al. Time-evolving social network generator based on modularity: Tesng-m. *IEEE Transactions on Computational Social Systems*, IEEE, v. 7, n. 3, p. 610–620, 2020. Citado na página 26.

METZ, J. et al. Redes complexas: conceitos e aplicações. São Carlos, SP, Brasil., 2007. Citado na página 14.

SHEN, H. et al. Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 388, n. 8, p. 1706–1712, 2009. Citado 5 vezes nas páginas 17, 19, 22, 23 e 27.

SLOTA, G. M. et al. Scalable generation of graphs for benchmarking hpc community-detection algorithms. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. [S.l.: s.n.], 2019. p. 1–14. Citado 2 vezes nas páginas 14 e 20.

STEGEHUIS, C.; HOFSTAD, R. V. D.; LEEUWAARDEN, J. S. V. Epidemic spreading on complex networks with community structures. *Scientific reports*, Nature Publishing Group, v. 6, n. 1, p. 1–7, 2016. Citado 2 vezes nas páginas 24 e 25.