CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC	
(X) PRÉ-PROJETO ( ) PROJETO	ANO/SEMESTRE: 2021.1

UTILIZAÇÃO DE REDES COMPLEXAS PARA GERAÇÃO DE COMUNIDADES CONFIGURÁVEIS

Gustavo Henrique Spiess

Prof. Aurélio Faustino Hoppe – Orientador

1 INTRODUÇÃO

Dentro das áreas da matemática, biologia, engenharia e outras, ~~surge~~ surgiu nos últimos anos a definição do que seriam redes complexas (METZ *et al*., 2007). ~~Essa área é um desenvolvimento primeiramente identificado n~~Na área da matemática discreta, as redes complexas são usadas para a representação de sistemas com uma complexidade mais proeminente. Também são utilizados ~~ou~~ quando o uso ~~em que~~ de outros modelos, tais como representações lineares ou em árvore, falham em capturar as propriedades do mundo real.

~~A área de~~As redes complexas, em específico na área de detecção de comunidades, tem ganhado mais atenção nos últimos anos (DUAN *et al*., 2019). Essa geração de redes complexas pode servir para treino de sistemas de aprendizado de máquina, bem como avaliação de diversas abordagens de detecção (LARGERON *et al*. 2015). Existem um conjunto de algoritmos para construção de redes em volume, como os utilizados nos trabalhos de Slota *et al*. (2019) e Largeron *et al*. (2015). Tais algoritmos possuem parametrizações que possibilitam a exploração de um espaço de geradores para grafos densos ou esparsos, variando parâmetros que propiciam mais ou menos arestas, entre outras configurações possíveis.

Redes complexas, como objetos de estudo, são definidas por Metz *et al*. (2007) como grafos com uma topologia não trivial. Elas podem servir de analogia para diversos sistemas mantendo os aspectos do mundo real (METZ *et al*., 2007; DUAN *et al*., 2019; LARGERON *et al*., 2015; FORTUNATO, 2010).

Segundo Fortunato (2010), a definição de comunidades em redes complexas passa necessariamente por alguns critérios qualitativos, e é fortemente dependente do contexto em que se está usando o conceito. Fortunato (2010) também descreve entre outras manifestações possíveis de comunidades em estruturas hierárquicas, e áreas onde as comunidades se sobrepõe. Também é apontado por Largeron *et al*. (2015), Slota *et al*. (2019) e Duan *et al*. (2019) a existência de uma coleção de algoritmos que são capazes de gerar redes complexas com comunidades. Muitos desses algoritmos tem, construído no processo, uma solução determinada para a identificação das comunidades. Modelos distintos incluem diferentes propriedades identificadas em sistemas do mundo real. Como homofilia no modelo proposto por Largeron *et al*. (2015) e Akoglu e Faloutsos (2009), coeficientes de aglomeração próximos aos identificados em grafos do mundo real (SLOTA *et al*., 2019) e dinamicidade (DUAN *et al*., 2019; LUO *et al*., 2020).

Na maioria dos modelos propostos para a geração de redes complexas com comunidades, algumas propriedades mais comuns são sempre implementadas, como mundo pequeno e ser livre de escala. Essas propriedades não estão intrincadamente ligadas à presença de comunidades, mas são identificadas, como apontado por Fortunato (2010), em sistema do mundo real onde comunidades tipicamente emergem. Essas características por si só comporiam uma rede complexa, tornando a topologia do grafo não trivial.

No entanto, a configurabilidade desses algoritmos não permite a construção de grafos que representem mais acuradamente um grupo com uma demografia conhecida. A geração de redes complexas com uma parametrização mais específica pode simplificar projetos de previsão de espalhamento de doenças em uma cidade (STEGEHUIS; HOFSTAD; LEEUWAARDEN, 2016), bem como servir como ferramenta para detecção de efeitos de diferentes políticas públicas.

Nesse contexto, a questão a ser respondida é como gerar redes complexas que representem mais realisticamente as comunidades de um grupo com uma demografia mais definida~~.~~ ? Esse problema pode ser descrito em como ~~de~~ gerar redes complexas ~~ao qual~~em que a topologia entre as comunidades possa ser definida com parâmetros demografias distintas.

A partir disso, esse trabalho visa a construção de um modelo para a geração de redes complexas que possibilite ~~a~~ sua parametrização definindo propriedades da topologia. I.e. informações de censo podem definir que em determinadas porcentagens das pessoas, o domicílio é habitada por uma quantidade de pessoas, ou que existem tantas pessoas participando do mercado de trabalho em uma ou outra área. Essas informações seriam passadas ao algoritmo por meio de parâmetros para que fossem representadas na topologia da rede. Esses parâmetros ~~para a representação~~ poderiam ser um conjunto de listas ordenadas destacando quais comunidades possuem áreas em superposição, e uma estrutura de árvore destacando as comunidades hierárquicas.

Comentado [GJ1]: O que são "redes em volume"?

Comentado [GJ2]: Não está clara esta frase.
O que seriam "comunidades" neste contexto?

Comentado [GJ3]: Citar referência que faz esta afirmação

Comentado [GJ4]: Conceituar "mundo pequeno" e "livre de escala".

Comentado [GJ5]: Estas são as propriedades que se pretende parametrizar? Não está claro no texto.

Comentado [GJ6]: "área" aqui seria o mesmo que região geográfica?

1.1 OBJETIVOS

O objetivo do trabalho é propor um modelo para geração de redes complexas permitindo a parametrização da topologia extra comunitária, demonstrando as suas propriedades globais.

Os objetivos específicos são:

a) disponibilizar um grafo não direcionado com comunidades;
b) etiquetar os vértices das comunidades de forma que seja possível utilizar essa informação como verdade construída;
c) demonstrar que o modelo é capaz de simular as propriedades de: comunidades se superpondo (compartilhando vértices), comunidades hierárquicas, mundo pequeno e o grafo livre de escala.

2 TRABALHOS CORRELATOS

Neste capítulo serão apresentados alguns modelos propostos até o momento que incorporam propriedades relevantes das redes complexas. Na seção 2.1 é apresentado um modelo basilar para a construção de grafos com comunidades (AKOGLU; FALOUTSOS, 2009). A seção 2.2 ~~se~~ trata de um algoritmo para a produção de redes complexas cujas comunidades apresentam a propriedade e homofilia e outras (LARGERON *et al.*, 2015). E a seção 2.3 apresenta ~~é~~ um modelo que mantém propriedades e adiciona dinamicidade em redes com comunidades (DUAN *et al.*, 2019).

2.1 RTG: A RECURSIVE REALISTIC GRAPH GENERATOR USING RANDOM TYPING

Akoglu e Faloutsos (2009) propõem e demonstram as propriedades de um modelo com o objetivo de recriar padrões de distribuição de vértices e arestas observados no mundo real. O modelo parte da representação de cada vértice como uma sequência de caracteres, e o modelo gera a lista de arestas como pares de sequências obtidos com a escolha aleatória de caracteres.

Os parâmetros para o algoritmo são:

a) k: A quantidade de caracteres possíveis;
b) q: A probabilidade de finalizar uma palavra;
c) W: A quantidade de arestas;
d) β: O reforço na probabilidade de origem e destino teres caracteres em comum.

O processo inicia criando uma matriz de K + 1 por K + 1, em que cada coluna representa um valor a ser acrescentado no final da origem, e cada linha representa um para o destino e a última coluna e linha representam a finalização da sequência respectivamente da origem e do destino. Nessa matriz cada célula tem um valor numérico que determina a probabilidade dessa ser a célula escolhida durante o processo de digitação. A diagonal principal, que representa a probabilidade de que o destino e a origem tenham valores em comum, tem a probabilidade aumentada subtraindo β das demais. Em seguida, é inicializada uma lista de arestas à qual são adicionadas W pares de vértices. Cada aresta é construída escolhendo um item da matriz, balanceando pela probabilidade na mesma, até que tanto a origem quanto o destino estejam finalizados. A partir dessa estrutura, uma série de propriedades emergem nos grafos produzidos:

a) grafo livre de escala / lei de potência: pela característica recursiva da construção da origem e do destino das arestas, os balanços da probabilidade, conforme apontado pelos autores, fazem com que as distribuições de arestas sigam a lei de potência. Essa lei de potência, segundo Akoglu e Faloutsos (2009), se refere à onze diferentes proporções identificadas em grafos do mundo real;
b) modularidades específicas: é demonstrado pelos autores também que a alteração dos parâmetros tem efeitos determinados nas propriedades do grafo~~.~~ ~~por~~ Por exemplo, aumentando o valor de β é observado um crescimento da modularidade.

A simplicidade do modelo proposto trás também outras características desejáveis, como a performance que nos testes que Akoglu e Faloutsos (2009) realizam num contexto de 1000 a 7000 para o valor do parâmetro W~~W~~, o consumo de tempo cresce linearmente. Outra característica que emerge dessa simplicidade, muito embora não apontado diretamente pelos autores, é a possibilidade de paralelizar a execução do algoritmo, distribuindo o processamento entre qualquer número de computadores sem incorrer em problemas. O trabalho representa um modelo que explora propriedades matemáticas triviais mas que consegue produzir grafos que mimetizam propriedades muito relevantes. No entanto, ele apresenta uma dificuldade por, apesar de garantir a presença de comunidades, não produzir junto ao grafo as etiquetas para as arestas.

Comentado [GJ7]: esta expressão não havia sido introduzida ainda. O que seria uma "topologia extra comunitária"?

Comentado [GJ8]: O que são "origem" e "destino" neste contexto?

Comentado [GJ9]: Usar letra minúscula caso se trate da quantidade de caracteres

Comentado [GJ10]: O que são as etiquetas?

2.2 GENERATING ATTRIBUTED NETWORKS WITH COMMUNITIES

Largeron *et al*. (2015) propõe um modelo para geração de redes complexas com comunidades baseadas em semelhanças por atributos. Isto é realizadoso promovendo-se a propriedade de homogeneidade das comunidades observada em outros sistemas do mundo real. O modelo proposto é composto por três partes: V, ε e A. As duas primeiras se referem, respectivamente, ao conjunto de vértices e o conjunto de pares ordenados de vértices que compões as arestas (sem direcionamento). A é um conjunto de comunidades, e esseas são conjuntos de vértices de forma que um vértice esteja em exatamente uma comunidade. O processo para geração desses dados é parametrizado com os seguintes valores:

a) N: um número inteiro maior que 0 que determina a quantidade de vértices;

b) $E^{max}{}_{wth}$: um número inteiro maior que 0 que determina a quantidade máxima de arestas internas à comunidade por vértice;

d) $E^{max}{}_{btw}$: um número inteiro entre 0 e $E^{max}{}_{wth}$ que determina a quantidade máxima de arestas externas à comunidade por vértice;

c) MTE: um número inteiro determinando a quantidade mínima de arestas no grafo;

d) A: um conjunto ordenado de desvios padrões para a inicialização dos parâmetros;

e) K: um número inteiro maior que zero que determina a quantidade de comunidades;

f) θ: um número real entre 0 e 1 que determina o limite de homogeneidade das comunidades;

g) $NbRep$: um número inteiro maior que 0 que determina a quantidade máxima de representantes por comunidade.

O modelo de Largeron *et al*. (2015) inicia criando uma nuvem de pontos n dimensionais servindo como conjunto de vértices., os Os valores para cada coordenada são obtidos como uma distribuição normal de A_1, A_2, ..., A_n onde A_n é a enésima componente do parâmetro A. São gerados N vértices. Depois é realizada a inicialização das comunidades, $K * NbRep$ vértices aleatórios são selecionados, com o uso do algoritmo *Kmedoids*. Os clusters gerados servem como sementes para a comunidade., . são São removidos vértices para que cada cluster tenha o mesmo tamanho (pegando selecionando os mais próximos ao centro) , entãoe são criadas ligações dos vértices de cada comunidade. Esses vértices agirão como representantes para a comunidade.

Depois disso, lotes de vértices sem comunidade são adicionados às comunidades das quais eles forem mais próximos dos representantes com uma chance θ de escolher uma comunidade aleatoriamente. A cada vértice adicionado à uma comunidade, são adicionados também um conjunto de arestas ligando-o interna e externamente à comunidade utilizando a lei de potência. A cada lote adicionado, novos representantes são escolhidos aleatoriamente. Por fim, são adicionadas arestas ligando vértices que compartilham pelo menos um vizinho até que o número mínimo de arestas (MTE) seja atingido.

Largeron *et al*. (2015) fazem também a demonstração das propriedades do modelo proposto: Sendo a primeira propriedade demonstrada pelo trabalho é de que a lei de potência é obedecida, gerando um grafo livre de escala. I.e., isso é, nos grafos gerados, a frequência de vértices cai em função logarítmica do grau, ou em outras palavras, para cada vértice com grau n maior que um, existem 10 com grau $n - 1$. Por conta desse modelo incorporar processos aleatórios em muitos momentos, esse valor não se expressa de forma tão exata, mas essa é uma tendência consistente nos grafos gerados.

Largeron *et al*. (2015) também descrevem em quais condições as estruturas que determinam a comunidade são degradadas, i.e. em quais condições a homogeneidade e a estrutura deixam de indicar que o que é produzido seria de fato uma comunidade. Com valores de θ maiores, as comunidades começam a perder a homogeneidade, de forma que uma execução onde theta é igual a um meio, as comunidades ocupam espaços muito semelhantes na nuvem de pontos. Com variações do valor de $E^{max}{}_{btw}$, é apresentado que valores maiores degradam as funções de modularidade e média do coeficiente de aglutinação. É demonstrado também que com esse valor igual a 0, as comunidades não possuem relações entre si, tornando o grafo desconexo.

Segundo Largeron *et al*. (2015), variando os parâmetros $E^{max}{}_{btw}$ e MTE é possível reforçar as características estruturais da comunidade gerando mais ligações internas. Aumentando esses dois valores, o coeficiente de clusterização e a modularidade aumentam, indicando uma comunidade mais densa.

Por fim, Largeron *et al*. (2015) descrevem os tempos e os problemas identificados aumentando a escala do grafo, nesse caso, foi identificado que ao aumentar o N é também é necessário também aumentar o MTE para evitar uma queda agressiva no coeficiente de clusterização. Os autores sugerem utilizar $MTE = 10N$ e os resultados indicam uma queda bem mais gradual. Foi demonstrado também que, dentro das faixas de parâmetro testadas pelo autor, variando os valores de N, $NbRep$ e K o tempo aumentou linearmente, as variações de $E^{max}{}_{wth}$ e $E^{max}{}_{btw}$

parecem não ter impacto no tempo de execução, e variando *MTE* o crescimento é por passo, isso é, até certo ponto ele cresse linearmente, depois disso o tempo se mantém constante.

Conclui-se que o modelo produz redes complexas com as propriedades determinadas: comunidades densamente conexas, homogêneas, com uma distribuição natural de graus, um mundo pequeno etc. Como extensão Largeron *et al*. (2015) indicam a adaptação do modelo para uso de atributos categóricos e não apenas numéricos. E, apesar de não ser apresentado pelos autores, uma limitação é que os atributos são preenchidos nos vértices por meio de distribuições normais.

2.3 DYNAMIC SOCIAL NETWORKS GENERATOR BASED ON MODULARITY: DSNGM

Duan *et al*. (2019) demonstram as propriedades de um modelo de geração de redes dinâmicas. O modelo proposto para a dinamização da rede com comunidades possui um conjunto de quatro parâmetros, e produz uma série de redes complexas com estruturas de comunidades. Os parâmetros são:

a) *G*: o grafo original;

b) *N*: o numero de instantes, isso é, quantas etapas serão geradas;

c) *count*: número de iterações máximas por instante;

d) *T*: temperatura, usada para determinar a probabilidade de aceitar uma mudança que não promova a modularidades esperada.

Duan *et al*. (2019) não aprofundam a definição de *T*, mas é apontado que o uso dessa temperatura é um valor para a manipulação dos componentes aleatórios do modelo proposto. Aumentando o valor de *T*, aumenta-se a probabilidade de ocorrência de alterações que não reforçam a modularidade esperada.

No trabalho, é utilizada a definição de modularidade como a soma da fração das arestas internas à comunidade menos o quadrado da fração das arestas que passam pela comunidade, conforme mostra a Equação 1. Isso é, para cada comunidade *s*, K^{in}_s é a quantidade de arestas com as duas pontas dentro da comunidade, K_s é a quantidade de arestas com uma ou mais pontas na comunidade, e *M* é a quantidade total de arestas no grafo.

Equação 1 – Modularidade do grafo

$$Q = \sum_s \frac{K_s^{in}}{2M} - \left(\frac{K_s}{2M}\right)^2$$

Fonte: Duan *et al*. (2019).

Inicialmente, o modelo define que o primeiro item da sequência é o próprio *G*, isso é $G_0 = G$. Depois é identificado o grupo de comunidades presentes no grafo, sendo calculada a modularidade para elas. Posteriormente, para cada item a ser gerado na sequência, é determinada uma modularidade esperada, valor randomizado entre 0,3 e 0,7, e é ~~realizado~~ realizada uma sucessão de trocas em pares de vértices. A troca consiste em remover a aresta se ela estava presente, ou adicionar se ela não estava. São realizadas tentativas de troca até que o número máximo seja atingido, ou até que a modularidade depois das trocas seja menor do que a modularidade esperada. Ao realizar cada troca verifica-se se essa diminuiu a modularidade, se sim, ela é efetuada. Caso a troca não promova a modularidade esperada existe uma probabilidade determinada pot *T* de que ela ainda assim ocorra.

Duan *et al*. (2019) descrevem os resultados experimentais do modelo, testando com uma base de dados previamente etiquetada em comunidades. Essa base de dados não se encontra mais disponível, mas descreve um grafo com doze comunidades, cento e quinze vértices e seiscentas e treze arestas. Com a execução é demonstrado que ao longo dos diferentes momentos do grafo a modularidade varia consideravelmente, bem como a quantidade de arestas. Os autores demonstram que a execução do algoritmo, apesar de produzir grafos vastamente distintos, os produz de forma que as estruturas de comunidade não são perdidas ao longo do processo, isso é, apesar de deixarem de ser os agrupamentos ótimos para definição de comunidades em alguns momentos, os agrupamentos se mantém estruturalmente coerentes. Também é apontado que as relações entre a mudança de valores nos parâmetros e o impacto na série produzida. Isso se manifesta em como é necessário um aumento coerente entre a temperatura e a quantidade de iterações por geração, para que a modularidade não seja perturbada demasiadamente.

Comentado [GJ11]: Revisar.

Duan *et al*. (2019) concluem apontando a necessidade de expansão na área, com a inclusão de outros processos além da adição e remoção de arestas, mas reforçando a relevância de um algoritmo para a geração de redes dinâmicas que considerem a modularidade.

3 PROPOSTA DO MODELO

Neste capítulo será descrita a proposta deste trabalho, justificando o desenvolvimento, definindo os requisitos funcionais e não funcionais, as metodologias abordadas e por fim o cronograma.

3.1 JUSTIFICATIVA

No quadro 1 é apresentado um comparativo entre os trabalhos correlatos. As linhas representam as características e as capacidades de cada modelo proposto, e as colunas representam os diferentes trabalhos.

Quadro 1 – Comparativo dos trabalhos correlatos

Características \ Trabalhos Correlatos	Akoglu e Faloutsos (2009)	Largeron *et al.* (2015)	Duan *et al.* (2019)
Gera grafos com comunidades	Sim	Sim	Não
Define os membros das comunidades	Não	Sim	Não se aplica
Produz um grafo dinâmico	Não	Não	Sim
Garante a homofilia das comunidades	Sim	Sim	Não se aplica
Garante propriedade de mundo pequeno	Sim	Sim	Sim
Gera grafos livre de escala	Sim	Sim	Sim
Possui vértices com atributos quantitativos	Não	Sim	Não se aplica
Possui vértices com atributos característicos	Sim	Não	Não se aplica
Gera comunidades hierárquicas / superpostas	Não	Não	Não se aplica

Fonte: elaborado pelo autor.

A partir do Quadro 1 é possível identificar que Duan *et al.* (2019) não geram propriamente uma rede complexa com comunidades. No entanto, eles a manipulam, tornando a rede em um grafo dinâmico, sem perder as propriedades relevantes. A manutenção das propriedades enquanto se manipula o grafo é, em si, o foco do trabalho. A relevância que ele traz é a capacidade de integrar o modelo proposto com outros trabalhos que fazem a geração mas, não necessariamente de forma dinâmica.

O trabalho desenvolvido por Akoglu e Faloutsos (2009) apresenta outras características que limitam o seu uso. Utilizando a terminologia determinada por Slota *et al.* (2019), o trabalho não apresenta uma verdade aproximada por engenharia. Apesar de gerar grafos garantindo a existência de comunidades homofílicas, ele falha em não etiquetar, durante o processo, a quais comunidades pertencem quais vértices. Isso significa, entre outras coisas, que ele não se torna relevante em processos como os de outros autores, onde os modelos propostos podem servir para aferir a capacidade de algoritmos que identificam as comunidades.

Por fim, ~~o~~ Largeron *et al.* (2015) apresentam um modelo similar ao que se pretende desenvolver nesse trabalho. Dadas as devidas proporções, o modelo proposto é bastante complexo, mas ele se vale dessa complexidade interna para gerar as redes complexas com as comunidades já estabelecendo quais vértices pertencem a quais comunidades. Essa verdade aproximada por engenharia, no caso do modelo de Largeron *et al.* (2015), pode ser efetivamente utilizada para a aferição dos resultados de processos que identifiquem essas comunidades. Além disso, este é um dos poucos trabalhos identificados pelo autor que promove a construção das comunidades por características topográficas e por características de homogeneidade. Também ~~O~~observou-se que apenas Largeron *et al.* (2015) e Akoglu e Faloutsos (2009) desenvolveram a geração de redes complexas incluindo o tratamento dos atributos homofilia.

Já Fortunato (2010) define o que pode-se encontrar em redes complexas comunidades sobrepostas. Isso tanto em comunidades hierárquicas quanto em comunidades que compartilham membros. Não foram identificados trabalhos que implementem a geração de redes onde essas propriedades estejam presentes, muito embora seja um fator de grande relevância, por exemplo para algumas aplicações onde a topologia entre as comunidades é mais relevante que a intra comunitária, como o trabalho desenvolvido por Stegehuis, Hofstad e Leeuwaarden (2016).

Diante desse contexto, o desenvolvimento de um modelo que ~~gere~~ gera redes complexas com uma verdade aproximada onde se façam presente as propriedades de comunidades hierárquicas e superpostas se torna bastante relevante. Sendo possível a utilização para aferir a usabilidade em processos que identificam comunidades com superposição. Além disso, também foi identificado que esse campo de estudos de redes complexas, e em especial as propriedades de comunidades, não possui publicações em língua portuguesa. Sendo academicamente relevante os processos de revisão bibliográfica e a disponibilização do conteúdo nessa língua para estimular o desenvolvimento de novos trabalhos limitando o efeito dessa barreira linguística em relação a compressão dos conceitos definidos na área de estudo.

Ressalta-se que o modelo proposto poderá ser utilizado em outras aplicações. Como por exemplo, gerando uma rede com uma topologia que mimetize uma população conhecida, é possível estudar os efeitos da disseminação de uma doença, como apresentado por Stegehuis, Hofstad e Leeuwaarden (2016). Também é possível utilizá-lo para estudar os efeitos de diferentes políticas públicas, tendências migratórias e outras mudanças na dinâmica social.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

O modelo proposto deve:

a) permitir ao usuário informar a profundidade de uma árvore de comunidades hierárquicas e a probabilidade de superposição de comunidades de um mesmo nível (Requisito Funcional – RF);

b) construir grafos não direcionados, sem pesos (RF);

c) gerar e identificar os membros de comunidades identificadas (etiquetando os vértices) (RF);

d) gerar as comunidades com a propriedade de homofilia (RF);

e) possibilitar a ocorrência de comunidades com a propriedade de superposição e hierarquia (RF);

f) produzir grafo sendo livres de escala e mundo pequeno (RF);

g) ter os processos computacionalmente mais intensos paralelizáveis (Requisito Não Funcional – RNF);

h) ser implementado utilizando a linguagem de programação ~~python~~ Python (RNF).

3.3 METODOLOGIA

a) revisão bibliográfica: identificar fontes bibliográficas com relação a redes complexas, detecção de comunidades e geração de redes com comunidades e trabalhos correlatos;

b) definição de ferramentas para modelagem e validação: identificar as ferramentas a serem utilizadas para validação de que as comunidades superpostas e hierárquicas mantém uma estrutura consistente;

c) proposição e implementação do modelo: implementar o modelo utilizando a linguagem de programação ~~python~~Python, utilizando a biblioteca Networkx;

d) validação de características estruturais da comunidade: validar com base no coeficiente de aglomeração e na modularidade, comparando com outros algoritmos que geram redes com comunidades;

e) validação de características de valor da comunidade: validar a homofilia comparando com outros algoritmos que geram grafos com comunidades homofílicas;

f) validação das propriedades de mundo pequeno e de liberdade de escala: comparando com outros modelos que geram redes complexas onde essas qualidades são mantidas;

g) validação de que as comunidades que se superpões e que tem estrutura hierárquica: com base nas ferramentas identificadas no item (b), verificar se essas comunidades mantém coerência interna;

h) validação da sensibilidade dos parâmetros na manipulação das propriedades;

i) validação da sensibilidade dos parâmetros no desempenho em função de tempo do modelo.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2 – Cronograma de atividades a serem realizadas

	2022									
	fev.		mar.		abr.		maio		jun.	
etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
revisão bibliográfica										
definição de ferramentas										
proposição e implementação do modelo										
validação de características estruturais da comunidade										
validação de características de valor da comunidade										
validação das propriedades de mundo pequeno e de liberdade de escala										
validação de que as comunidades que se superpões e que tem estrutura hierárquica										
validação da sensibilidade dos parâmetros na manipulação das propriedades										
validação da sensibilidade dos parâmetros no desempenho em função de tempo do modelo										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo explorar os principais assuntos que fundamentarão o estudo a ser realizado: redes complexas, detecção de comunidades e geração de redes com comunidades.

Redes complexas, como definido por Metz *et al*. (2007), são grafos com uma topologia não trivial. Elas são ferramentas que servem para a modelação de sistemas com a capacidade de manter propriedades que se observam no mundo real e que não podem ser observadas em outros modelos (FORTUNATO, 2010). A topologia não trivial de uma rede complexa se expressa como um conjunto de propriedade que podem ser observad~~a~~os nessa rede, mundo pequeno (LARGERON *et al*., 2015), grafos livres de escala (LARGERON *et al*., 2015; SLOTA *et al*., 2019) e comunidades.

De acordo com Fortunato (2010), existem três principais definições de comunidades~~,~~ : comunidades locais, comunidades globais~~,~~ e comunidades baseadas em similaridade de vértices. Uma definição local de comunidades passa pelo entendimento de que a topologia da mesma a difere do restante do grafo, de forma que não se poderia adicionar um vértice à comunidade sem perder alguma propriedade da estrutura da mesma. Uma definição global se aplica para cenários onde as pontas de cada aresta não são tão relevantes quando o isolamento do conjunto de vértices que foram as comunidades, isso é, trocando quais os vértices que cada aresta liga, mas mantendo o grau (interno e externo à comunidade) não se perde a definição de comunidade. Por fim, a definição de comunidade com base em similaridade dos vértices é a ideia de que os membros de uma determinada comunidade são mais semelhantes entre si do que são dos demais vértices.

Comentado [GJ12]: Não está claro

Essas definições são identificadas de uma perspectiva da detecção de comunidades, mas se aplicam claramente na geração de redes com comunidades. Duan *et al*. (2019) utilizam principalmente uma definição global de comunidades, enquanto Akoglu e Faloutsos (2009) e Largeron *et al*. (2015) utilizam uma definição local e por semelhança. Outros autores como Slota *et al*. (2019), apenas a definição local. Muito embora as definições não sejam mutuamente excludentes, a forma de se definir uma comunidade é muito dependente do sistema a ser representado.

Para o uso de uma definição tanto local como global, tipicamente são utilizadas métricas como a modularidade, definida por Fortunato (2010) e por Duan *et al*. (2019) como uma razão entre os vértices internos e externos da comunidade, como apontado na seção de trabalhos correlatos. Outra métrica que se pode usar é a diferença entre o coeficiente de aglomeração para o subgrafo que compõe uma comunidade e para o grafo como um todo. O coeficiente de aglomeração em si é definido como a proporção das triplas conexas que são grafos completos.

Além disso, quando se fala em uma definição por similaridade, uma métrica utilizada é a distância euclidiana quando se representa os vértices em uma nuvem de pontos. Dessa forma, pode-se medir a coerência de uma comunidade como a razão entre o desvio padrão dos vértices da comunidade, e o desvio padrão de todos os vértices do grafo. Uma comunidade mais coerente teria distâncias menores entre seus membros, diminuindo o desvio padrão. Essa qualidade de comunidades definidas por similaridade é expressa por Largeron *et al*. (2015) como homofilia. Mais precisamente, o autor declara que é esperado de uma rede complexa uma maior probabilidade de relacionamento entre vértices mais semelhantes.

No entanto, outras características tendem a estarem presentes em redes complexas observadas no mundo real. Akoglu e Faloutsos (2009), Slota *et al*. (2019) e Largeron *et al*. (2015) reproduzem com seus respectivos modelos, além de comunidades, grafos livres de escala. Essa propriedade é definida pelos autores como a proporção de ocorrência dos graus dos vértices. Isso é, espera-se que existem muitos vértices com grau baixo, e poucos vértices com grau elevado, seguindo uma proporção logarítmica. A propriedade de um grafo livre de escala também aponta que existem uma tendência de vértices com graus mais altos estarem mais conectados entre si do que vértices de graus mais baixos. Esses grafos comumente também apresentam a propriedade de mundo pequeno, definida por Largeron *et al*. (2015) como a tendência em grafos do mundo real de que o diâmetro do grafo (distância mínima entre os dois vértices mais distante) cresce de forma logarítmica com a quantidade de vértices do grafo. Outra definição relevante para as redes complexas é o conceito de clique. Um K − clique é definido Fortunato (2010) como um subgrafo completo contendo k vértices.

Uma definição estrutura de comunidades com superposição é oferecida por Fortunato (2010) onde os vértices que fazem partes de cliques em mais de uma comunidade efetivamente pertencem às duas. Outras definições de comunidades com superposição também podem ser derivadas conforme o autor descreveu, quando se considera comunidades como subgrafos que otimizam uma função arbitrária. Considerando o coeficiente de aglomeração por exemplo, podem ser descritos casos onde um vértice pertenceria a múltiplas comunidades. Por fim, outra possibilidade de comunidades se sobreporem é o conceito de comunidades hierárquicas. Fortunato (2010) descreve que em muitos sistemas do mundo real são identificadas comunidades que internamente são organizadas se dividindo em outras comunidades mais densas e homogênias.

Comentado [GJ13]: O que são "cliques"?

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.

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.

FORTUNATO, S. Community detection in graphs. **Physics reports**, Elsevier, v. 486, n. 3-5, p. 75–174, 2010.

LARGERON, C. *et al*. Generating attributed networks with communities. **PloS one, Public Library of Science**, v. 10, n. 4, p. e0122777, 2015.

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.

METZ, J. *et al*. **Redes complexas**: conceitos e aplicações. São Carlos, SP, Brasil., 2007.

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.

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.

ASSINATURAS

(Atenção: todas as folhas devem estar rubricadas)

Assinatura do(a) Aluno(a): _____

Assinatura do(a) Orientador(a): _____

Assinatura do(a) Coorientador(a) (se houver): _____

Observações do orientador em relação a itens não atendidos do pré-projeto (se houver):

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Acadêmico: Gustavo Henrique Spiess _____

Avaliador: Andreza Sartori _____

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?			
	O problema está claramente formulado?			
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?			
	Os objetivos específicos são coerentes com o objetivo principal?			
	3. JUSTIFICATIVA São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?			
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	4. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?			
	Os métodos, recursos e o cronograma estão devidamente apresentados?			
	5. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?			
ASPECTOS METODOLÓGICOS	6. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?			
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			
	7. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?			
	8. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?			
	9. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT?			
	As citações obedecem às normas da ABNT?			
	Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?			

PARECER – PROFESSOR DE TCC I OU COORDENADOR DE TCC
(PREENCHER APENAS NO PROJETO):

O projeto de TCC será reprovado se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos **4 (quatro)** itens dos **ASPECTOS TÉCNICOS** tiverem resposta ATENDE PARCIALMENTE; ou
- pelo menos **4 (quatro)** itens dos **ASPECTOS METODOLÓGICOS** tiverem resposta ATENDE PARCIALMENTE.

PARECER: ( ) APROVADO ( ) REPROVADO

Assinatura: _____ Data: _____

[1] Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR AVALIADOR

Acadêmico: Gustavo Henrique Spiess _____

Avaliador(a): Gilvan Justino

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	X		
	O problema está claramente formulado?		X	
	1. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	X		
	Os objetivos específicos são coerentes com o objetivo principal?	X		
	2. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?	X		
	3. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?	X		
	São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?	X		
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?	X		
	4. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?	X		
	5. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?	X		
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?	X		
	6. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?	X		
	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?	X		
ASPECTOS METODOLÓGICOS	7. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?	X		
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?		X	

PARECER – PROFESSOR AVALIADOR:
(PREENCHER APENAS NO PROJETO)

O projeto de TCC ser deverá ser revisado, isto é, necessita de complementação, se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos **5 (cinco)** tiverem resposta ATENDE PARCIALMENTE.

PARECER: ( ) APROVADO ( ) REPROVADO

Assinatura: _____ Data: _____

[1] Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.