CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC	
( ) PRÉ-PROJETO (X) 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, surgiu nos últimos anos a definição do que seriam redes complexas (METZ *et al*., 2007). Originalmente identificadas na área da matemática discreta, essas redes permitem a representação de sistemas com uma complexidade mais proeminente com a qual outros modelos, como representações lineares ou em árvore falham em capturar as propriedades do mundo real. Redes complexas, como objeto 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 (DUAN *et al*., 2019; LARGERON *et al*., 2015; FORTUNATO, 2010).

As redes complexas, em específico a área de detecção de comunidades, tem ganhado mais atenção nos últimos anos (DUAN *et al*., 2019). A 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 com bilhões de vértices, 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.

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, mas são geralmente compreendidas como sub grafos conexos com alguma qualidade que o separa do restante do grafo. 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 têm, identificada no processo, uma solução determinada para a identificação das comunidades. Modelos distintos produzem grafos com diferentes propriedades de sistemas do mundo real~~., Como~~ como a propriedade de 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, como nos trabalhos de Largeron *et al*. (2015), Slota *et al*. (2019) e Duan *et al*. (2019). Mundo pequeno significando que o sistema representado como um grafo tem uma relação logarítmica entre o diâmetro e a quantidade de vértices. E ser livre de escala significando que o grau dos vértices tem uma relação logarítmica com a quantidade de vértices. 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, por tornarem 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 gerar redes complexas em que a topologia entre as comunidades possa ser definida com parâmetros demografias distintas.

Comentado [GJ1]: rever

A partir disso, esse trabalho visa a construção de um modelo para a geração de redes complexas que possibilite sua parametrização definindo propriedades topológicas de comunidades sobrepostas e hierárquicas. Esses parâmetros poderiam ser, respectivamente, uma árvore representando as comunidades e as relações hierárquicas entre elas, e uma lista tabela de quais itens da árvore (quais comunidades) teriam sobreposições.

1.1 OBJETIVOS

O objetivo do trabalho é disponibilizar um modelo para geração de redes complexas permitindo a parametrização da topologia extra comunitária (comunidades hierárquicas e sobrepostas), 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 trata de um algoritmo para a produção de redes complexas cujas comunidades é a propriedade e homofilia (LARGERON *et al*., 2015). Por fim, a seção 2.3 apresenta um modelo que mantém propriedades e adiciona dinamicidade em redes com comunidades (DUAN *et al*., 2019).

Comentado [GJ2]: rever

2.1 RTG: A RECURSIVE REALISTIC GRAPH GENERATOR USING RANDOM TYPING

Akoglu e Faloutsos (2009) propõe e demonstram as propriedades de um modelo algorítmico 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 (vértices de uma aresta) terem 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 de 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 a 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, no trabalho de Akoglu e Faloutsos (2009), se refere à onze diferentes proporções identificadas em grafos do mundo real;
b) modularidades especificas: É demonstrado pelos autores também que a alteração dos parâmetros tem efeitos determinados nas propriedades do grafo. Por exemplo, aumentando o valor de β é observado um crescimento da modularidade.

A simplicidade do modelo proposto traz 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, o consumo de tempo cresce linearmente. Outra característica que emerge dessa simplicidade, muito embora não apontado diretamente pelo autor 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 a identificação de quais sãos as comunidades e seus membros.

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 realizado promovendo-se a propriedade de homogeneidade das comunidades como observado em 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 essas 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;
c) $E^{max}{}_{btw}$: um número inteiro entre 0 e Emax wth que determina a quantidade máxima de arestas externas à comunidade por vértice;
d) MTE: um número inteiro determinando a quantidade minime de arestas no grafo;
e) A: um conjunto ordenado de desvios padrões para a inicialização dos parâmetros;
f) K: um número inteiro maior que zero que determina a quantidade de comunidades;
g) θ: um número real entre 0 e 1 que determina o limite de homogeneidade das comunidades;
h) $NbRep$: 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 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. Então 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 removidos vértices para que cada cluster tenha o mesmo tamanho (selecionando os mais próximos ao centro), e 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. Para 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 definido pelo parâmetro `MTE` seja atingido.

Largeron *et al*. (2015) fazem também a demonstração das propriedades do modelo proposto. A primeira propriedade demonstrada pelo trabalho é de que o grafo gerado é livre de escala. Isso é, nos grafos gerados, a frequência de vértices cai em função logarítmica do grau. 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 descreve 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. Quanto à variações do valor de $E^{max}{}_{btw}$, valores maiores degradam as funções de modularidade e média do coeficiente de aglomeração. 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. Ao utilizar valores maiores para esses dois parâmetros, o coeficiente de aglomeração 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 também é necessário aumentar o MTE para evitar uma queda agressiva no coeficiente de aglomeração. Os autores sugerem utilizar `MTE` = $10N$, resultando em uma queda bem mais gradual. Foi demonstrado também o tempo de execução do algoritmo em função dos parâmetros. 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. Por fim 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. 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 número 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 (DUAN *et al*., 2019). 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.

$$Q = \sum_s \frac{K_s^{in}}{2M} - \left(\frac{K_s}{2M}\right)^2 \quad (1)$$

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 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 por *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 significa que apesar de deixarem de ser os agrupamentos ótimos para definição de comunidades em alguns momentos, os agrupamentos se mantêm estruturalmente coerentes. Também é apontado a relação 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. 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 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, 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 se observou que apenas Largeron *et al*. (2015) e Akoglu e Faloutsos (2009) desenvolveram a geração de redes complexas incluindo o tratamento dos atributos homofilia.

Fortunato (2010) aponta a existência de comunidades sobrepostas em redes complexas. 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 em aplicações onde a topologia entre as comunidades é mais relevante que a intra comunitária. Como por exemplo, o trabalho desenvolvido por Stegehuis, Hofstad e Leeuwaarden (2016), em que ~~onde~~ se identificou que as relações entre membros de comunidades distintas têm grande impacto no espalhamento de uma doença.

Diante desse contexto, o desenvolvimento de um modelo para gerar redes complexas com uma verdade aproximada onde se façam presente as propriedades de comunidades hierárquicas e superpostas se torna bastante relevante. Com um modelo assim seria possível a sua utilização para aferir a precisão de 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 português. 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 compreessão 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.

Comentado [GJ3]: Os objetivos não contemplam a criação ou adaptação de um modelo.

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 (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 na validação das características topológicas das comunidades sobrepostas e hierárquicas;
c) elicitação de requisitos: baseando-se nas informações obtidas nas etapas anteriores, reavaliar, detalhar e, se necessário incluir novos requisitos;
d) proposição e implementação do modelo: implementar o modelo utilizando a linguagem de programação Python, utilizando a biblioteca Networkx;
e) 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;
f) validação de características de valor da comunidade: validar a homofilia comparando com outros algoritmos que geram grafos com comunidades homofílicas;
g) validação das propriedades de mundo pequeno e de liberdade de escala: comparando com outros modelos que geram redes complexas onde essas qualidades são mantidas;
h) 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 mantêm coerência interna;
i) validação da sensibilidade dos parâmetros na manipulação das propriedades;
j) validação da sensibilidade dos parâmetros no desempenho em função de tempo do modelo.

Comentado [GJ4]: Não está claro

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										
elicitação de requisitos										
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 está dividido em três seções. A seção 4.1 aborda redes complexas. Já a seção 4.2 discorre sobre propriedades de redes complexas. Por fim, a seção 4.3 descreve redes complexas com comunidades.

4.1 REDES COMPLEXAS

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 não está contida apenas nos vértices e arestas individualmente, mas em propriedades do conjunto de vértices e arestas.

De acordo com Girvan e Newman (2002), um dos sistemas do mundo real que se pode modelar em uma rede complexa é o conjunto de relações sociais. Esse sistema poderia ser modelado de forma que cada indivíduo seria representado por um vértice, e as relações sendo arestas ligando dois indivíduos. Num modelo assim, observa-se na topologia do grafo, a existência de um sub grafo completo, denominado clique (FORTUNATO, 2010), ao qual todos os vértices se conhecem, tal grupo pode apresentar características próprias e relevantes.

Girvan e Newman (2002) também apontam que outros sistemas, como cadeias alimentares, cadeias de metabolização, redes de transmissão elétrica e redes de computadores podem ser representadas por redes complexas. Em cada sistema do mundo real modelado através de redes complexas, diferentes propriedades emergem, e elas apresentam diferentes significados sobre o grafo observado. No estudo de redes complexas, é relevante o entendimento de algumas terminologias tais como clique, partição, comunidade, entre outras (VIEIRA; XAVIER; EVSUKOFF, 2020).

Fortunato (2010) se refere a clique como sendo um conjunto de vértices que formam um sub grafo completo. ~~Definições~~ Variações ligeiramente distintas podem ser utilizadas, dependendo do contexto. Por exemplo, o autor apresenta a definição de um n-clique, como sendo um sub grafo onde todos os vértices estão a n ou menos arestas de distância. Ou, ainda uma cadeia de *k-clique* (*k-clique chain*), sendo a união de dois cliques com k ou mais vértices que compartilham um ou mais vértices.

Outra terminologia relevante é o conceito de partição, definido por Fortunato (2010) como sendo a divisão dos vértices de um grafo em conjuntos distintos de forma que todos os vértices pertençam a exatamente uma partição. Essas duas definições são relevantes, na forma como elas interagem, para conceituar comunidade. Em alguns contextos uma comunidade pode ser definida como sendo necessariamente um clique (FORTUNATO, 2010), ou uma partição (VIEIRA; XAVIER; EVSUKOFF, 2020; FORTUNATO, 2010)

Fortunato (2010) define triângulo como sendo um conjunto de três vértices conectados entre si, por exemplo, um 3-clique. E, em contrapartida a essa definição, o coeficiente de aglomeração, definido como a proporção em que as triplas conexas de um grafo são triângulos. Isso é, para um coeficiente de aglomeração igual a meio, no qual quaisquer três vértices estejam conectados entre si, existe cinquenta por cento de chance desses vértices formarem um triângulo.

Comentado [GJ5]: seria 0,5?

Por fim, também existe a definição de um grafo aleatório. Um grafo aleatório é produzido a partir de um grafo qualquer, mantendo a quantidade e o grau dos vértices, mas considerando que os vértices estão conectados em uma probabilidade uniformemente distribuída. Um grafo aleatório ~~é~~ geralmente é utilizado para comparação, tendo como exemplo o grafo nulo, no sentido de que ele não apresenta características topológicas relevantes (FORTUNATO, 2010).

4.2 PROPRIEDADES DE REDES COMPLEXAS

Fortunato (2010) destaca que as redes complexas são ferramentas de modelagem que preservam algumas propriedades observadas no mundo real e que não poderiam ser modeladas em estruturas lineares ou através de árvores. Definições como a do coeficiente de aglomeração ganham relevância quando se observa que, conforme apontado por Girvan e Newman (2002), grafos do mundo real tem a tendência de possuírem um coeficiente de aglomeração entre dez e cinquenta por cento. Um grafo sintético, gerado artificialmente, que replique esses valores apresenta, portanto, a propriedade de ter um coeficiente de aglomeração que simula um grafo do mundo real.

Outra propriedade que se pode identificar dentro de uma rede complexa é o grafo ser livre de escala, tendo uma distribuição logarítmica entre o grau dos vértices e o quão comum é a ocorrência desse grau. ~~Ou, em~~ Em outras palavras, representa que o grafo possui mais vértices com poucas arestas (LARGERON *et al.*, 2015). Muitos sistemas do mundo real apresentam essa propriedade. A liberdade de escala está intimamente relacionada com outra propriedade, a de preferência de anexação local, isso é, a tendência de que vértices tenham mais conexões com vértices de grau maior que o seu próprio (LARGERON *et al.*, 2015).

Girvan e Newman (2002) e Largeron *et al.* (2015) também definem mundo pequeno, como sendo a qualidade de grafos possuírem diâmetros pequenos e que crescem de forma logarítmica. ~~Ou, em~~ Em outras palavras, para um sistema do mundo real ser observado em duas escalas distintas, a diferença do diâmetro em comparação com a quantidade de vértices, é logarítmica.

Homofilia, como definida por Largeron *et al.* (2015) se refere à característica de que, em muitos sistemas do mundo real, os vértices tendem a se conectar mais comumente com outros vértices semelhantes. No modelo de

Largeron *et al*. (2015) isso é representado como sendo a distância entre vértices que possuem uma representação em um espaço euclidiano. Akoglu e Faloutsos (2009) implementam isso como caracteres compartilhados entre vértices representados por uma *string*. Ao representar os atributos como uma sequência de valores discretos que estão dentro de um conjunto finito, os vértices seriam considerados mais próximos se possuíssem, para o mesmo índice na sequência, valor igual.

Por fim, uma propriedade identificada em muitos sistemas do mundo real é a presença de estruturas de comunidade (GIRVAN; NEWMAN, 2002; AKOGLU; FALOUTSOS, 2009; FORTUNATO, 2010; LARGERON *et al*., 2015; DUAN *et al*., 2019; SLOTA *et al*., 2019; STEGEHUIS; HOFSTAD; LEEUWAARDEN, 2016), que será discutida na próxima seção, mas que pode ser entendido como a presença, na topologia da rede, de sub grafos conexos com alguma característica de grupo que os separe dos demais. A definição de uma comunidade é usualmente feita em função da dicotomia de vértices e arestas internas à comunidade e externas a ela. A presença de comunidades estruturais, ou seja, definidas a partir das conexões internas e externas, possibilita também a propriedade de homogeneidade dessas comunidades. Uma comunidade pode ser considerada homogenia por possuir membros mais semelhantes entre si, do que o grafo do qual a comunidade faz parte (AKOGLU; FALOUTSOS, 2009; FORTUNATO, 2010; LARGERON *et al*., 2015).

Comentado [GJ6]: Aconselho acompanhar a estes conceitos, a apresentação figuras/diagramas que auxiliem no esclarecimento.

4.3 REDES COMPLEXAS COM COMUNIDADES

De acordo com Fortunato (2010), existem três definições principais de comunidades~~,~~: comunidades locais, comunidades globais, e comunidades baseadas em similaridade de vértices. A 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. 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 formam as comunidades, isso é, modificando a topologia da comunidade ao mudar arestas, 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.

Essas diferentes definições se relacionam ao que é apontado por Vieira, Xavier e Evsukoff (2020) como a diferença identificável entre cada modelo de detecção de comunidades. Os autores apontam que cada modelo busca otimizar alguma função que aponta qualidades da comunidade, mas inevitavelmente ignora outras funções, e por conseguinte outras qualidades estruturais, que poderiam ser utilizadas para definir a comunidade.

Conclui-se a partir das afirmações de Vieira, Xavier e Evsukoff (2020) e Fortunato (2010) que a definição de comunidade é necessariamente contextual. No entanto, existem algumas qualidades compartilhadas entre todas essas definições, por exemplo, como apontado por Fortunato (2010) toda definição de comunidade passa pela qualidade de conectividade do sub grafo produzido. Isso é, não se pode considerar uma comunidade qualquer conjunto de vértices que não represente um sub grafo conexo.

Para uma definição local de comunidades, o critério que qualifica algum conjunto de vértices como uma comunidade é relativo a alguma propriedade interna à comunidade. Um critério que se pode utilizar são as definições de clique a suas variações. Com um critério assim, a definição de comunidade garante uma coerência interna (FORTUNATO, 2010).

Para a definição de comunidade global, em oposição à uma definição local, o que se procura são características estruturais do grafo que separem os indivíduos dentro da comunidade dos fora da comunidade (FORTUNATO, 2010). Para isso são elencadas diversas métricas possíveis, por exemplo a comparação do coeficiente de aglomeração da comunidade e do grafo como um todo, onde a comunidade seria a partição ótima que maximiza esse coeficiente para as comunidades. A métrica que se tornou a mais aceita para a definição de comunidades foi a modularidade (VIEIRA; XAVIER; EVSUKOFF, 2020), estando definida conforme a Equação 1. Esse valor é tipicamente contrastado com o um grafo aleatório servindo de modelo nulo, isso é, um modelo onde essa função tende a se manter constante para todas as partições possíveis. Essa definição de comunidade é extensivamente explorada nos trabalhos de Luo *et al*. (2020) e de Duan *et al*. (2019). Por fim, a definição de comunidades por semelhança (FORTUNATO, 2010) está intrincadamente ligada às propriedades de homofilia e homogeneidade da comunidade, como pode se ver no trabalho de (LARGERON *et al*., 2015).

A complexidade em se definir o conceito de comunidade cresce consideravelmente quando se considera o que é definido como comunidades sobrepostas (FORTUNATO, 2010; VIEIRA; XAVIER; EVSUKOFF, 2020). Comunidades sobrepostas são comunidades que apresentam não apenas relações entre vértices, mas que compartilham vértices. Esse tipo de estrutura de comunidades sobrepostas é muito comum em sistemas do mundo real, nessas regiões de sobreposição não é raro observar-se uma densidade maior na região sobreposta (SENGUPTA; HAMANN; WAGNER, 2017). Para definições locais de comunidade, Fortunato (2010) apresenta exemplos onde as comunidades compartilham vértices se esses participarem em uma cadeia de *k-clique*.

No entanto definições globais precisam de adaptações, conforme apresentado por Vieira, Xavier e Evsukoff (2020). A modularidade é a função mais aceita para a qualidade estrutura de comunidades que não possuam áreas de sobreposição. Mas, ainda não existe um consenso de qual das adaptações dessa função é a mais adequada para a qualificação de comunidades com áreas de superposição. E, quando se utilizam definições de comunidades por semelhança, torna-se inviável a manutenção da característica de homogeneidade da comunidade e homofilia nos vértices que fazem parte da área sobreposta (VIEIRA; XAVIER; EVSUKOFF, 2020).

Esse desafio, é eminente pela característica de se definir comunidades como partições, como é apontado por (VIEIRA; XAVIER; EVSUKOFF, 2020). No entanto, ainda existe um outro fator relevante que dificulta tal definição~~, . conforme~~ Conforme apontado por Fortunato (2010), muitos sistemas do mundo real se organizam em comunidades hierárquicas, significando que esses sistemas são repartidos em comunidades que por sua vez são repartidas em sub comunidades. Segundo o autor, existem ferramentas para a identificação de comunidades hierárquicas em um sentido estrutural, e pelas propriedades de árvore contidas em uma organização assim não é difícil derivar uma definição recursiva. Para Fortunato (2010) é natural que as comunidades superiores, compostas de sub comunidades, possuam uma definição por semelhança menos presente, ou em outras palavras, os níveis externos de comunidades aninhadas apresentam uma diminuição na presença da propriedade de homogeneidade.

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.

GIRVAN, M.; NEWMAN, M. E. J. Community structure insocial and biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 99, n. 12, p. 7821–7826, 2002.

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.

SENGUPTA, N.; HAMANN, M.; WAGNER, D. Benchmark generator for dynamic overlapping communities in networks. In: IEEE. 2017 IEEE **International Conference on Data Mining (ICDM)**. [S.l.], 2017. p. 415–424.

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.

VIEIRA, V. da F.; XAVIER, C. R.; EVSUKOFF, A. G. A comparative study of overlapping community detection methods from the perspective of the structural properties. **Applied Network Science,** SpringerOpen, v. 5, n. 1, p. 1–42, 2020.

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: (X) 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.