sid.inpe.br/mtc-m21c/2020/05.29.15.18-TDI

DETECÇÃO DE PADRÕES EM DADOS ESPACIAIS E TEMPORAIS VIA REDES COMPLEXAS

Frank Moshe Cotacallapa Choque

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Marcos Gonçalves Quiles, Manoel Ferreira Cardoso e Elbert Einstein Nehrer Macau, aprovada em 29 de abril de 2020.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/42JBEF2>

INPE
São José dos Campos
2020

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GBDIR)
Serviço de Informação e Documentação (SESID)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

sid.inpe.br/mtc-m21c/2020/05.29.15.18-TDI

DETECÇÃO DE PADRÕES EM DADOS ESPACIAIS E TEMPORAIS VIA REDES COMPLEXAS

Frank Moshe Cotacallapa Choque

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Marcos Gonçalves Quiles, Manoel Ferreira Cardoso e Elbert Einstein Nehrer Macau, aprovada em 29 de abril de 2020.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/42JBEF2>

INPE
São José dos Campos
2020

Dados Internacionais de Catalogação na Publicação (CIP)

Cotacallapa Choque, Frank Moshe.

C456d Detecção de padrões em dados espaciais e temporais via redes complexas / Frank Moshe Cotacallapa Choque. – São José dos Campos : INPE, 2020.

xvi + 132 p. ; (sid.inpe.br/mtc-m21c/2020/05.29.15.18-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2020.

Orientadores : Drs. Marcos Gonçalves Quiles, Manoel Ferreira Cardoso e Elbert Einstein Nehrer Macau.

1. Sistemas complexos. 2. Redes complexas. 3. Redes temporais. 4. Redes dinâmicas. I.Título.

CDU 004.72

Aluno (a): ***Frank Moshe Cotacallapa Choque***

Título: "DETECÇÃO DE PADRÕES EM DADOS ESPACIAIS E TEMPORAIS VIA REDES COMPLEXAS"

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de ***Doutor(a)*** em

Computação Aplicada

Dr. **Helder Luciani Casa Grande**

Presidente / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Dr. **Marcos Gonçalves Quiles**

Orientador(a) / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Dr. **Manoel Ferreira Cardoso**

Orientador(a) / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Dr. **Elbert Einstein Nehrer Macau**

Orientador(a) / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Dra. **Karine Reis Ferreira**

Membro da Banca / INPE / São José dos Campos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Dr. **Solon Venâncio de Carvalho**

Membro da Banca / INPE / SJCampos - SP

(X) Participação por Video - Conferência

(X) Aprovado ***( ) Reprovado***

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 29 de abril de 2020

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Computação Aplicada

Dr. Márcio Porto Basgalupp

Convidado(a) / UNIFESP / SJCampos - SP

(X) Participação por Vídeo - Conferência

(X) Aprovado ( ) Reprovado

Dr. Fabrício Aparecido Breve

Convidado(a) / UNESP / Rio Claro - SP

(X) Participação por Vídeo - Conferência

(X) Aprovado ( ) Reprovado

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 29 de abril de 2020

AGRADECIMENTOS

Primeiramente, meu muito obrigado ao Brasil, por ter me acolhido e aberto diversas oportunidades nos últimos 8 anos. Fiz novas amizades e até formei uma nova família.

Agradeço também, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro ao longo de quase 4 anos de doutorado, e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa TT-5 no último ano do doutorado na Climatempo.

Sou grato também, aos meus orientadores e professores, pelo tempo, paciência, e comentários necessários para superar os desafios desta pesquisa. Por último, mas não menos importante, quero agradecer o apoio constante da minha família e companhia próxima de Deus ao longo dos 28 anos de vida.

RESUMO

A cada ano, novos desafios são encontrados na área de mineração de dados. O respeito pela privacidade dos usuários, o desenvolvimento de soluções em tempo real, armazenamento e processamento em grande escala são alguns dos temas que estão em destaque, atualmente, na computação. Além disso, sabemos que os dados gerados ao nosso redor estão relacionados a situações complexas, em diferentes níveis de detalhe, e mudando ao longo do tempo. Considerando esse contexto, o trabalho desenvolvido propõe métodos para obter informações sobre padrões em eventos temporais e espaciais, a partir do uso de redes complexas. Nesse sentido, sabendo que há uma demanda crescente por abordagens que lidem com os desafios modernos da mineração dos dados, o método proposto aqui, busca preencher essa necessidade com foco na simplicidade, flexibilidade e novos resultados. Em poucas palavras, este método identifica estágios ou estados pelos que passa um conjunto de eventos ao longo do tempo, usando como base a construção cronológica de uma rede complexa e aplicando algoritmos de detecção de comunidades. Adicionalmente, é feita uma detalhada comparação entre um algoritmo de agrupamento e o método proposto, a fim de obter as principais diferenças e mensurar como estes se comportam e conjuntos fictícios de dados temporais.

Palavras-chave: Sistemas complexos. Redes complexas. Redes temporais. Redes dinâmicas.

PATTERN DETECTION IN SPATIAL AND TEMPORAL EVENTS THROUGH COMPLEX NETWORKS

ABSTRACT

Every year, new challenges are faced in data mining. The respect for users' privacy, real-time solutions, data processing in large scale are some of the trending topics in this area. Besides this, huge amounts of generated data around us is related to complex situations, in different levels of details, and changing along time. Within this context, the work developed along these lines propose methods to get information related to significant changes in temporal datasets, based on the use of complex networks, with focus on temporal events and temporal sequences. In addition to this, considering the growing demand for new approaches to deal with modern challenges in data mining, the proposed algorithm intend to fill part of this gap by focusing on three characteristics: simplicity, flexibility and novelty results. In few words, this method is able to identify phases or stages in sets of spatio-temporal events, based on the use of community detection algorithms and cronological complex networks. Furthermore, a detailed comparison is made between a state-of-the-art clustering algorithm and the proposed method, in order to identify the main differences and how they behave in several toy data models.

Keywords: Complex systems. Complex networks. Temporal networks. Dynamic networks.

LISTA DE FIGURAS

Pág.

SUMÁRIO

Pág.

1 INTRODUÇÃO

É notório que há muitos anos a quantidade de dados gerados a cada dia é imensa, ao ponto de ser preciso armazenar boa parte destes dados em infraestruturas dedicadas e otimizadas para este fim, local também conhecido como *a nuvem* (ARMBRUST et al., 2010). Considerando que os dados, hoje em dia, são sinônimo de fonte de informação para aprimorar diversos processos e tomada de decisões, é natural a crescente demanda por ferramentas que consigam fornecer informações relevantes de forma automática (FAN et al., 2014; CHEN; ZHANG, 2014). Seguindo essa tendência, diversas pesquisas acadêmicas e iniciativas corporativas estão aprimorando e desenvolvendo tecnologias para auxiliar com os novos desafios da mineração de dados em grande escala (JIN et al., 2015).

Nesse contexto, emergiram outros desafios interessantes como o cuidado da privacidade, a implementação de políticas de segurança, o armazenamento em grande escala, a classificação correta, organização das fontes dos dados, execução de consultas rápidas, entre muitos outros (CHEN; ZHAO, 2012; XU et al., 2014). Além disso, apesar do grande poder computacional disponível hoje em dia, ela sozinha não é suficiente para ajudar na descoberta das informações, pois com o desenvolvimento de novas tecnologias, elas também adicionam novas fontes de dados, e portanto, maior complexidade no conjunto de dados (SHOBANA; KUMAR, 2015). Tudo isso é também acompanhado pela demanda de novas soluções em tempo real, que sejam executadas com um simples toque do dedo ou a partir de comandos de voz. Portanto, é esperado que novas formas de explorar os dados sejam procuradas, em face aos desafios modernos. (CHEN; ZHANG, 2014; JIN et al., 2015).

Uma variável que está presente nos conjuntos de dados que envolvem sistemas do mundo real é o tempo. É inevitável ignorá-lo, sobretudo nos dias atuais em que há diversas formas de coletas dados em incríveis velocidades e de formas diversas (ANTUNES; OLIVEIRA, 2001). O tempo é o que guia os ciclos, as estações e as temporadas; ajuda a se localizar no mar de dados que são gerados a cada nanosegundo. No intuito de otimizar ainda mais essa variável, este trabalho focase no desenvolvimento de novas técnicas que apresentam um uso potencial na extração de informações relevantes a partir do uso de redes complexas.

Mesmo que os grafos tenham suas origens na matemática há mais de 280 anos por Leonhard Euler, foi somente nos últimos 20 anos que a teoria das redes complexas emergiu e ganhou destaque, como uma nova forma de compreender a estrutura e relações que possuem os sistemas complexos (NEWMAN, 2010). Desse modo, as

redes complexas foram usadas em variados problemas e áreas, que vão desde os sistemas biológicos até a análise de fenômenos climáticos de grande escala (COSTA et al., 2007; COSTA et al., 2011).

Olhando para a mineração de dados e as redes complexas, é perceptível que, ambas as áreas, têm orientações ou focos diferentes. No entanto, estes também compartilham objetivos em comum: a partir de um conjunto de dados, desenvolver uma representação ou modelo que permita obter informações a partir dele. Algumas diferenças também são notórias: enquanto as redes complexas tem como principal tarefa a construção de relações entre elementos de um mesmo sistema, a mineração de dados procura a construção de padrões, mesmo que não pertençam ao mesmo sistema (ZANIN et al., 2016). Desse modo, a mineração de dados a partir das redes complexas adiciona informações que, de outra forma, seria muito difícil obtê-las (CHAU, 2012; HAN et al., 2006). Uma ilustração simples para comparar as vantagens seria o caso de uma escola com professores e alunos, que interagem em diversos momentos e lugares, e que além disso, possuem atributos pessoais como endereço, idade, ano de estudos, entre outros. A mineração de dados certamente ajudaria, por exemplo, na agrupação de alunos a partir de atributos específicos, como o rendimento escolar por da idade e bairro de origem. Por outro lado, as redes complexas ajudariam a enxergar as relações entre os alunos e avaliar quão engajada é a interação entre eles, segundo o bairro, rendimento escolar. Também seria possível identificar os grupos de amigos que se formam ao longo do tempo. Como visto, as possibilidades de aplicação são ilimitadas e em diversos sentidos.

Em particular, a mineração de dados temporais é uma área que explora, principalmente, variáveis na dimensão temporal (ANTUNES; OLIVEIRA, 2001). Nesse sentido, uma das tarefas é compreender as mudanças que acontecem ao longo do tempo, o que, em muitos casos, permitirá entender o comportamento de um sistema e até mesmo fazer um previsão do comportamento futuro (MITSA, 2010). Apesar da maioria dos algoritmos de mineração de dados considerar o tempo como uma variável entre muitas outras, em termos de ordem, ele é fundamental para uma sequência de eventos que, sem uma ordem temporal, pode tornar os dados sem utilidade (HOLME; SARAMÄKI, 2012). Isto é visível, por exemplo, quando analisamos o contágio de um vírus em uma população. Nele, a ordem do contato entre as pessoas tem impacto direto sobre a propagação do vírus. Portanto, neste exemplo, o tempo, o contato das pessoas e o estado do vírus precisam ser observados e analisados em conjunto.

Ao procurar modelar dados com atributos temporais, as redes complexas temporais emergem como uma alternativa para tratar de compreender como os sistemas evoluem no tempo (HOLME, 2015; HOLME, 2016). Como resultado disso, é cada vez mais frequente encontrar estudos sobre redes temporais e a dinâmica dos sistemas (DONNAT; HOLMES, 2018). No entanto, desde a perspectiva do desenvolvimento de métodos para auxiliar a mineração dos dados temporais no contexto atual, há muitos desafios que permanecem abertos (ZANIN et al., 2016; MARTÍNEZ et al., 2016). Nesse sentido, no intuito de mensurar as mudanças em dados com componente temporal, este trabalho propõe um conjunto de métodos baseados em redes complexas para encontrar padrões a partir do componente temporal.

1.1 Objetivos

Esta pesquisa trata sobre o desenvolvimento de novos métodos de mineração de dados temporais a partir da ciência das redes, que responde à pergunta: como encontrar padrões em dados com atributos espaciais e temporais usando redes complexas? Nesse sentido, também emergem outras perguntas que complementam este desafio:

- Quais são os métodos que existem na literatura capazes de transformar dados temporais e espaciais em redes?
- Como a rede, baseada em dados temporais e espaciais, muda ao longo do tempo?
- Quais são os desafios para encontrar padrões neste tipo de dados?

No intuito de responder estas perguntas, neste trabalho foram estudados os métodos tradicionais de mineração de dados temporais e espaciais, os baseados em redes complexas. Posteriormente, neste texto, são propostos novos métodos que aprimoram e complementam os métodos encontrados atualmente na literatura. Além disso, com este trabalho, procura-se preencher parcialmente as lacunas e os desafios modernos que a mineração de dados temporais e espaciais apresentam. Nesse sentido, duas frentes de trabalho foram desenvolvidas:

- A primeira procura analisar de forma conjunta dados que têm componentes temporais e espaciais, nesse sentido, o objetivo aqui é usar a localização e o registro temporal para para compreender como se relacionam as regiões onde os eventos acontecem, e desse modo, obter uma melhor

compreensão sobre os estágios ou estados que os eventos percorrem, considerando o agrupamento destes e seguindo uma sequencia cronológica;

- A segunda frente trata sobre a análise de dados temporais sem componente espacial. Aqui, o objetivo é obter novas informações a partir da sequência de dados temporais de diversos elementos de um sistema. Além disso, procura-se compreender, de modo global, o desempenho (em termos de interação sequencial) de cada elemento, ao longo do tempo, para encontrar grupos com o mesmo perfil de comportamento.

1.2 Motivação

Dados são o principal insumo para qualquer processamento que vise obter informações, seja para compreender o passado, ver o presente ou prever o futuro. A facilidade que hoje temos para gerar ou obter dados é impressionante, no entanto, ainda há muito espaço para explorar soluções que forneçam informações a partir delas, com o mesmo nível de facilidade que temos para gerá-las. Em muitas áreas existe o interesse pela extração e processamento de grandes quantidades de dados, de forma eficiente e rápida (FAN et al., 2014). Além disso, há muitas tarefas específicas que, em muitos casos é essencial que sejam realizadas de forma automática e em tempo real, como a detecção de anomalias na previsão do tempo ou a detecção de fraude em sistemas financeiros (SHOBANA; KUMAR, 2015).

É evidente que a relevância do tempo está atrelada à sua presença em diversos sistemas do mundo real, e, portanto, também presente nos sistemas complexos (DARST et al., 2016). Nesse sentido, devido ao crescimento da adoção das redes complexas para mapear as relações que existem nos sistemas complexos, muitos trabalhos usam esta variável (o tempo) como um dos componentes das regras para construir as redes e, desse modo, compreender a estrutura e dinâmica dos sistemas complexos que eles representam (ZANIN et al., 2016). O bem sucedido desenvolvimento delas em diversos casos motiva o estudo desta área em conjunto com a mineração de dados, desde que as redes complexas complementam e fornecem informações que, de outra forma, seria difícil encontrá-las (COSTA et al., 2011).

É conhecido que diversos trabalhos sobre ciência das redes abordam o tempo através das *redes temporais* (estruturas de rede que correspondem a momentos diferentes de um mesmo sistema), no entanto, a motivação deste trabalho não está baseada nelas, mas sim no uso da variável temporal para construir redes que não necessariamente são redes temporais (HOLME, 2015). Consequentemente, no in-

tuito de encontrar os padrões temporais e espaciais que existem no sistema, neste trabalho, desenvolveram-se métodos para inserir o tempo no processo de construção das redes e, dessa forma, explorar as propriedades da rede através de métodos próprios que permitem encontrar as variações.

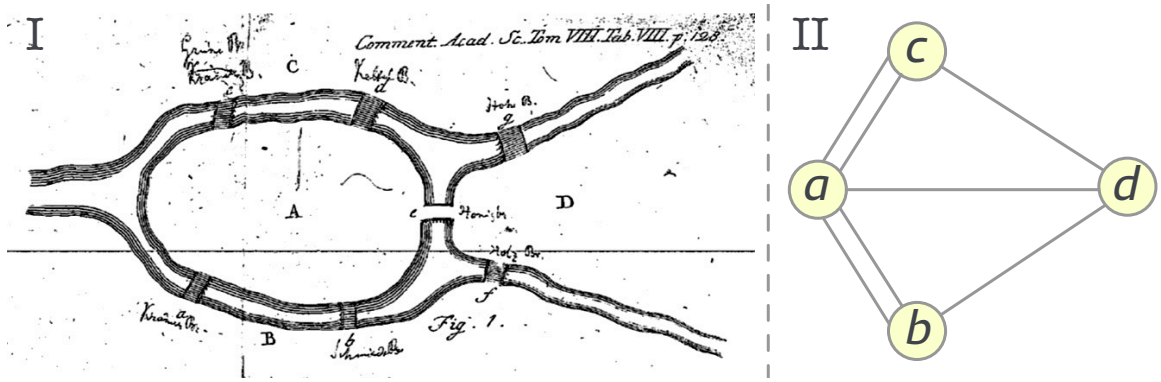
1.3 Organização do texto

O texto desta pesquisa é composta por oito capítulos. O primeiro deles é a Introdução, que coloca em contexto a relevância do assunto da tese e define as motivações e principais objetivos a serem alcançados ao longo do estudo. O segundo capítulo trata sobre as Redes Complexas de forma geral, colocando ênfase em algumas propriedades, resolução das redes, e tipos de redes, baseadas na literatura atual, que serão necessárias para a compreensão dos próximos capítulos. No terceiro capítulo, a ênfase está na revisão breve dos de métodos de mineração de dados em geral, e também das variações destes métodos para minerar dados temporais. No quarto capítulo, o foco é a sobre como transformar dados temporais em redes, fazendo uma revisão dos métodos atuais que são usados para esse fim. Também, explica-se alguns cuidados que devem ser considerados ao fazer a transição entre dados e redes. Depois, no quinto capítulo é apresentado a abordagem cronológica e três métodos desenvolvidos ao longo do doutorado, com maior ênfase na detecção de padrões em conjunto de dados que além de temporais, também possuem atributos espaciais ou de geolocalização. Também, é apresentado uma aplicação dos métodos sobre dados reais de queimadas na Bacia Amazônica. Já o sexto capítulo tem como foco a comparação entre os métodos propostos e o método ST-DBSCAN, de agrupação de dados temporais e espaciais, com destaque para as principais diferenças entre os resultados que ambos os métodos revelam. No sétimo capítulo são apresentados possíveis desdobramentos a partir dos métodos propostos, além disso, são mencionados os principais desafios e as aplicações onde estes seriam úteis. No oitavo capítulo são mencionadas as conclusões a partir dos resultados alcançados, e, finalmente, na última seção são listadas as referências da literatura usadas como suporte para o desenvolvimento desta pesquisa.

2 REDES COMPLEXAS

Parece inevitável discutir sobre a Ciência das Redes Complexas sem mencionar as origens dela, que tem seu fundamento na matemática, no ramo da Teoria de Grafos. O primeiro trabalho desenvolvido sobre este tema remonta-se ao ano 1736, quando o matemático suíço Leonhard Euler apresentou uma solução para o problema das Sete Pontes de Königsberg (EULER, 1953). Elas que fazem referência às pontes localizadas no rio Pregel (cidade de Königsberg, situada hoje na Rússia), que conectava diversas partes da cidade com uma ilha chamada de Kneiphof. O problema consistia em percorrer a cidade usando todas as pontes, porém atravessando apenas uma única vez cada uma delas (conforme Figura 2.1). Para provar que essa tarefa é impossível, respeitando as regras, foi preciso para ele desenvolver uma forma de análise que seja possível testá-la dentro do rigor matemático (RÄZ, 2018). Nesse sentido, ele reformulou o problema de forma abstrata, dessa forma dando origem à base da Teoria de Grafos. Baseado nesse teoria, as partes da cidade seriam representados por "vértices"e as pontes por "arestas", formando desse modo, um grafo.

Figura 2.1 - Ilustração esquemática das pontes de Königsberg segundo Euler



I) Cópia do desenho original em que Euler representou em letras em maiúsculo as quatro áreas da cidade. Também, usando as letras em minúsculo representou as sete pontes da cidade. II) Grafo representando as quatro partes (vértices) da cidade conectadas por sete pontes (arestas).

Fonte: Adaptado de Räz (2018).

Baseado no grafo, Euler percebeu que a possibilidade de "atravessar"as pontes uma única vez dependia da quantidade de arestas. O grafo precisaria ter, exatamente, nenhum ou dois vértices com um número ímpar de arestas para cumprir o desafio de percorre-lo sem repetir a mesma aresta. No caso das Sete Pontes de Königsberg, os quatro vértices possuem um número ímpar, portanto, o grafo não se cumpre as condições necessárias que Euler encontrou, e, portanto, não há forma de percorrer todas as pontes uma única vez (EULER, 1953; RÄZ, 2018).

A partir do momento que a teoria de grafos emergiu como uma nova área da matemática, ela desenvolveu algoritmos para encontrar caminhos e distancias entre vértices, assim com outras medidas. Fora da matemática, as aplicações dela são diversas. Na linguística, por exemplo, é usada para representar estruturas gramaticais ou de sintaxe; na química, é usada para o estudo de estruturas moleculares (BALABAN, 1985; CHEN; JI, 2010). Em outras áreas, como as ciências sociais, os grafos foram adotados como uma forma de explorar as relações entre indivíduos (BARNES; HARARY, 1983). Devido à capacidade de abstração que os grafos possuem, desenvolveram-se diversas áreas dentro da matemática para estudá-las com profundidade (HARARY, 2018).

Nas últimas décadas, a partir do estudo de certos sistemas complexos, surgiu o interesse pelo uso dos grafos de outra perspectiva, que considerasse a mecânica estatística, modelagem computacional e estatística, como parte de uma nova abordagem para compreender sistemas e fenômenos reais (STROGATZ, 2001). Esta nova área da ciência é chamada de *Ciência de Redes* ou *Teoria de Redes*, que tem como proposta a compreensão de sistemas onde seus componentes estão conectados, de forma complexa, e que ao mesmo tempo possuem padrões (NEWMAN, 2010). Um exemplo desse tipo de sistemas são as proteínas do genoma humano, que encontram-se relacionados em grande escala e de diversas formas, formando uma imensa orquestração de milhares de proteínas e genes, todas elas trabalhando de forma conjunta para determinar a variabilidade da vida humana (MICHAUT et al., 2011).

O início desta nova área de pesquisa tem seu começo em diversos estudos pioneiros, que usaram métodos baseados em grafos para resolver problemas que requeriam uma nova forma de compreender as relações entre diferentes elementos (NEWMAN, 2010). Nesse sentido, a partir da Teoria de Grafos e o estudo das propriedades que diversos sistemas complexos possuem em comum, construíram-se as bases para o desenvolvimento da Ciência de Redes (BARNES; HARARY, 1983). Toda-

via, apesar de existirem semelhanças entre ambas áreas (Teoria de grafos e Teoria de redes), algumas características particulares se destacam na Teoria de redes: a) ela aborda a complexidade de grandes conjuntos de arestas e vértices, analisando-as em diversas escalas de resolução; b) estuda a dinâmica do sistema, a partir de propriedades que emergem da topologia da rede; c) permite analisar fenômenos dinâmicos que surgem sobre uma topologia de rede, como a propagação de informações ou epidemias (STROGATZ, 2001; BARNES; HARARY, 1983; BRANDES; ERLEBACH, 2005).

A seguir, são descritos alguns conceitos importantes sobre redes baseado nos livros de Latora et al. (2017), Barabási e Pósfai (2016), Masuda e Lambiotte (2016), Newman (2010), Steen (2010) e Brandes e Erlebach (2005), que serão úteis para a compreensão deste trabalho. De modo formal, um grafo $G = (V, E)$ é formado pelo conjunto V, composto pelos vértices $\{v_1, v_2, ..., v_n\}$, e pelo conjunto E, que é composto pelas arestas $\{e_1, e_2, ..., e_m\}$, onde todos os elementos de ambos os conjuntos são diferentes entre si (LATORA et al., 2017). Desse modo, a quantidade de vértices e arestas é (usualmente) representado por N e M respectivamente. Também, a aresta que se encontra entre dois vértices pode seguir uma sequência ou ordem, e, desse modo, ser *direcionada*, e no caso contrário, *não direcionada* (BARABÁSI; PÓSFAI, 2016). Quando é direcionada, a aresta $e = (v_i, v_j)$ tem origem em v_i e destino em v_j. No caso de um grafo não direcionado, simplesmente afirmamos que tanto o vértice v_i quanto v_j são adjacentes à aresta e. Além disso, uma aresta é um *loop* se a origem e destino são os mesmos, isto é, $v_i = v_j$. Também, é possível que existam múltiplas arestas (entre dois vértices), e, neste caso, estas podem ser simplificadas numa aresta apenas (quando o grafo é não direcionado), ou em duas, uma entrada e outra de saída (quando a grafo é direcionado) (ESTRADA, 2010).

Há casos quando a aresta precisa salvar atributos, como o *peso*, que em representa a quantidade de arestas (entre dois vértices) que foram simplificadas. Nesse caso, o grafo $G = (V, E, W)$ é composto também pelo conjunto de pesos W, que correspondem aos pesos $\{w_1, w_2, ..., w_m\}$ de todas as arestas de E (LATORA et al., 2017). Portanto, uma aresta pode tomar diversas formas: ser direcionado ou não direcionado, com pesos ou sem pesos, e ser *loop*. Quando um grafo não possui *loops*, ela é chamada de grafo *simples*. Além disso, se todos os vértices do grafo possuem a mesma quantidade de arestas, o grafo é classificado como grafo *regular*, e quando todos os vértices estão conectados entre si, o grafo é nomeado como grafo *completo* (STEEN, 2010).

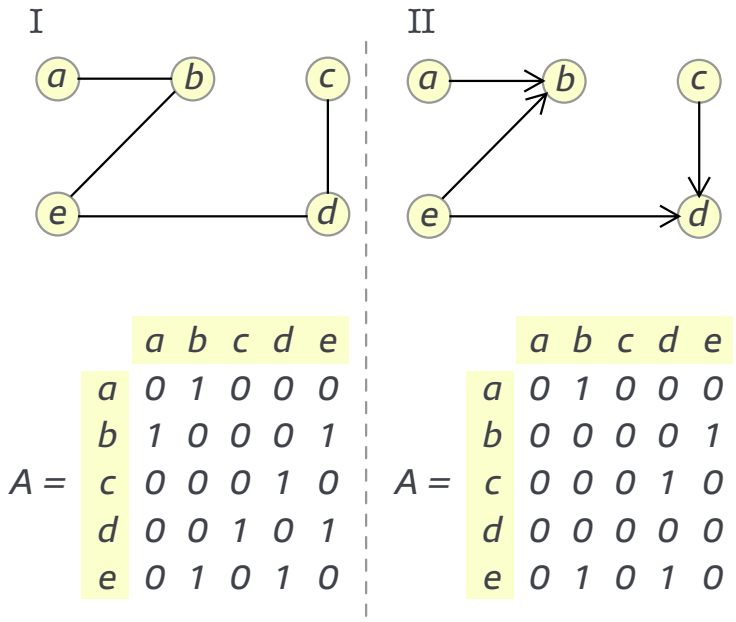
Outro conceito relevante na teoria de grafos que se aplica a vértices e arestas é a *adjacência*. Dois vértices são adjacentes se há uma aresta que une eles, também, duas arestas são adjacentes se compartilham um mesmo vértice (NEWMAN, 2010). Nesse sentido, o *grau* de um vértice é o número k de arestas adjacentes a este vértice. Se for um grafo direcionado, também é possível definir o grau de saída e entrada, de forma separada, correspondente à quantidade de arestas saindo e entrando, respectivamente (BARABÁSI; PÓSFAI, 2016).

Usando a adjacência, uma rede pode ser representada numa matriz $N \times N$, onde

$$A_{ij} = \begin{cases} 1; & \text{se o vértice } v_i \text{ é adjacente ao vértice } v_j, \\ 0; & \text{caso contrário.} \end{cases} \tag{2.1}$$

E no caso de grafos com pesos ou outros atributos, os valores podem ser diferentes de 1 e 0. Na Figura 2.2, mostra-se duas versões de grafos, uma direcionada e outra sem direção, e sua respectiva matriz de adjacências.

Figura 2.2 - Dois grafos (não dirigido e dirigido) e suas respectivas matrizes de adjacência.

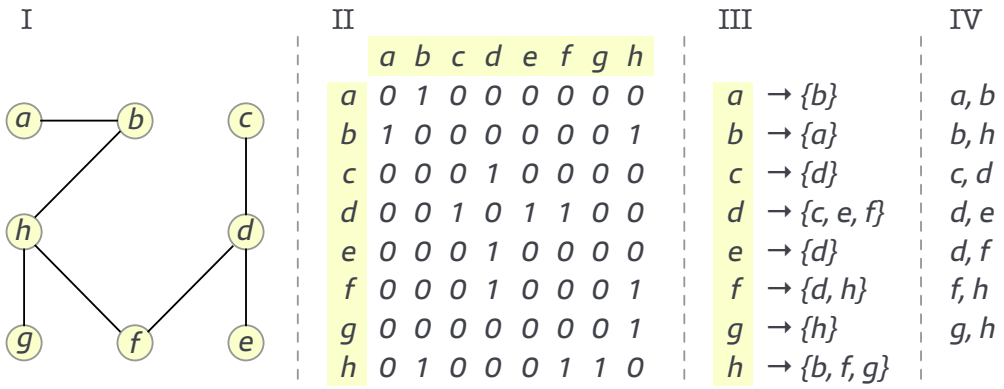


É interessante notar que a matriz da *adjacência* do grafo não dirigido (I) é simétrico, o que não necessariamente é verdadeiro para o grafo dirigido (II).

Fonte: Produção do autor.

Diversas análises e medidas de grafos estão baseadas na matriz da *adjacência*. No entanto, ao aplicar estes nas redes complexas, onde é possível ter milhares de nós de ligações, esse tipo de cálculo pode demandar alto custo computacional para o processamento de alguns algoritmos. Nesse sentido, outra forma de representar uma rede, é a partir da *lista de adjacências*. Nesta lista, por cada nó do grafo, são nomeados todos nós que estão ligados com ele. Adicionalmente, outra forma similar de descrever uma rede é a *lista de ligações*, que é o conjunto de todos os pares de nós que estão conectados entre si (MASUDA; LAMBIOTTE, 2016; ESTRADA, 2011). Nesse sentido, há algumas vantagens no uso desta última forma de representação. Algumas destas são: a simplicidade que possui para adicionar ou retirar ligações da lista, realizar simulações numéricas e a facilidade para exportar os dados da rede para softwares especializados para trabalhar com redes (MASUDA; LAMBIOTTE, 2016; ZINOVIEV, 2018). Na próxima Figura 2.3 observa-se algumas das principais formas de representar uma rede.

Figura 2.3 - Diversas representações de redes



I) Ilustração de uma rede com 8 nós e 7 ligações. II) Matriz de adjacência. III) Lista de adjacências. IV) Lista de ligações.

Fonte: Produção do autor.

Outra definição importante em grafos é o *caminho*. Esta é a composição de uma sequência de arestas $\{v_0 \rightarrow v_1, v_1 \rightarrow v_2, \ldots, v_{n-1} \rightarrow v_n\}$, que começa em v_0 e termina em v_n, onde cada uma delas são distintas entre si, e portanto, possui tamanho n (que corresponde à quantidade de arestas) (LATORA et al., 2017). E caso $v_0 = v_n$, o caminho é considerado um *ciclo*. Também, um caminho é chamado de

mais curto quando não há outro caminho menor entre dois vértices, e o tamanho deste caminho curto é chamado de *distância* (entre dois vértices). Desse modo, quando existe um caminho entre quaisquer dois vértices de um grafo, o grafo é considerado como *fortemente conectado* (MASUDA; LAMBIOTTE, 2016; STEEN, 2010).

Como mencionado anteriormente, os matemáticos começaram o estudo dos grafos a partir de um olhar teórico (STEEN, 2010). No entanto, com o passar do tempo foi necessário que os grafos também pudessem atender os desafios do mundo real (com milhares de interações entre diversos elementos de um sistema). Desse modo, com ajuda da física estatística, computação aplicada e matemática, a teoria das redes conseguiu estabelecer as bases de novos métodos e medidas para obter informações das redes (que representam sistemas complexos) (STROGATZ, 2001). Desse modo, como o foco deste trabalho encontra-se nas redes complexas, nas próximas seções, serão utilizados os termos *ligação* e *nó* para se referir a *aresta* e *vértice*, respectivamente.

2.1 Propriedades

A Teoria das Redes deu origem ao estudo das propriedades estatísticas e medidas das redes complexas, na procura por compreender as redes construídas com dados reais (STROGATZ, 2001). Devido ao amplo desenvolvimento delas, alguns conceitos surgiram na literatura acadêmica para estabelecer uma base para novas descobertas. Dessa forma, a seguir são apresentados alguns desses conceitos essenciais e que possuem relação com o escopo deste trabalho:

2.1.1 Grau de conectividade

De longe, é a medida mais conhecida e utilizada em muitos algoritmos sobre redes complexas. Em poucas palavras, esse conceito representa a quantidade de conexões com outros nós que um dado nó possui (BARABÁSI; PÓSFAI, 2016). Desse modo, o grau de um nó i, considerando a matriz de adjacência A da rede, é definido por

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{2.2}$$

e também, o grau médio é definido por

$$\langle k \rangle = \frac{\sum_{i=1}^{N} k_i}{N}. \tag{2.3}$$

Quando se trata de redes direcionadas, é claro que há diferença entre as ligações que entrem e saem, portanto, o grau é calculado de forma separada, para cada direção (CALDERELLI, 2007). Desse modo, o grau de entrada seria

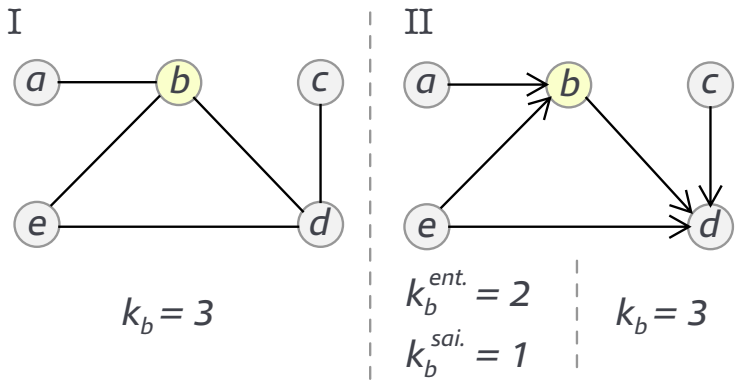
$$k_i^{\text{entrada}} = \sum_{j=1}^{N} A_{ji}, \tag{2.4}$$

e o grau de saída seria

$$k_i^{\text{saída}} = \sum_{j=1}^{N} A_{ij}. \tag{2.5}$$

Assim, o grau total de um nó i, numa rede direcionada, é a soma do grau de entrada e saída, isto é, $k_i = k_i^{\text{entrada}} + k_i^{\text{saída}}$ (SILVA; ZHAO, 2016). Na Figura 2.4 é mostrado um exemplo para ilustrar diferentes casos do grau de uma rede.

Figura 2.4 - Redes direcionada e não direcionada.



I) Grau (de conectividade) do nó *b*, considerando uma rede não direcionada. II) Grau de entrada e saída do nó *b*, considerando uma rede direcionada.

Fonte: Produção do autor.

2.1.2 Força

Quando existem diversas ligações entre dois nós, estas podem ser simplificadas em apenas uma ao adicionar um atributo w à ligação. Neste caso, o atributo será chamado de *peso*, (STEEN, 2010). É necessário destacar que o *peso* também pode ter valores não atrelados à simplificação e seguir outros mecanismos de cálculo. Por exemplo, se temos três ligações entre dois nós estes podem ser substituídos por apenas uma ligação com peso $w = 3$. No caso de redes em que as ligações têm pesos, a *força* (também conhecida como *strength*) de um nó é definida pela soma dos pesos das ligações que o nó está conectado (SILVA; ZHAO, 2016). Desse modo, a força s do nó i é

$$s_i = \sum_{j=1}^{N} W_{ij}, \tag{2.6}$$

onde W é a matriz dos pesos relativo à matriz de adjacência da rede. Em caso da rede ser direcionada, a força de um nó é definida separadamente, para cada direção das ligações em conexão com o nó. Para calcular a força de entrada a formula é

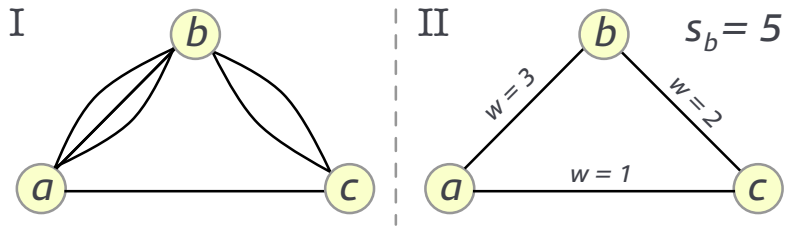
$$s_i^{\text{entrada}} = \sum_{j=1}^{N} W_{ji}, \tag{2.7}$$

e no caso da força de saída,

$$s_i^{\text{saida}} = \sum_{j=1}^{N} W_{ij}. \tag{2.8}$$

Na Figura 2.5 é mostrado um exemplo ilustrativo desta medida.

Figura 2.5 - Redes com e sem força.



I) Uma rede simples. II) Equivalente à rede I usando a *pesos* como atributo das ligações. Também, mostra-se a *força* do nó b, sendo este a soma dos pesos $w = 3$ e $w = 2$.

Fonte: Produção do autor.

2.1.3 Coeficiente de agrupamento

O *coeficiente de agrupamento* é também conhecido por alguns como *transitividade*. Esta medida quantifica quão próximo é a vizinhança de um nó, de formar um triângulo (três nós mutuamente conectados). Estes triângulos são comuns de aparecer em redes reais onde dois nós têm ligação entre si e também estão ligados com um terceiro nó (MASUDA; LAMBIOTTE, 2016).

Desse modo, quando a quantidade de triângulos numa rede é normalizada (dividido pela quantidade máxima de ligações na rede), é também chamado de *coeficiente de agrupamento local*, definido por

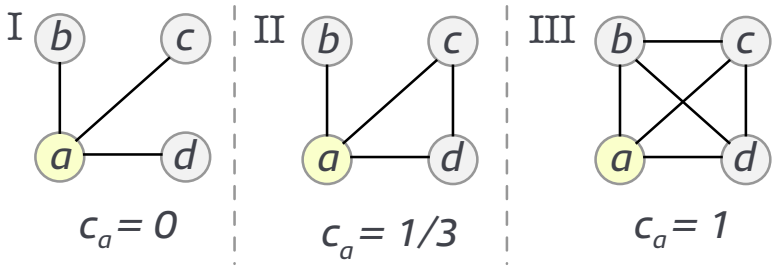
$$C_i = \frac{(\text{quantidade de triângulos com } v_i)}{k_i(k_i - 1)/2}, \tag{2.9}$$

também, note-se que $0 \leq C_i \leq 1$. Desse modo, o *coeficiente de agrupamento* (global) da rede é definido por

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{2.10}$$

Na Figura 2.6 mostra-se de forma ilustrativa o coeficiente de agrupamento local em três redes diferentes.

Figura 2.6 - Comparativo do coeficiente de agrupamento



I) Claramente, o coeficiente de agrupamento do nó a é zero devido à ausência de triângulos. II) A presença de um triângulo permite que o coeficiente de a seja $1/3$ III) Toda a rede está conectada, portanto, o coeficiente do nó a é 1.

Fonte: Produção do autor.

2.1.4 Centralidade

Para mensurar a centralidade de um nó há diversas abordagens possíveis, porém, cada uma segue o mesmo objetivo: quantificar a relevância dos nós de uma rede (MASUDA; LAMBIOTTE, 2016; KUNEGIS, 2014). Para alguns, o grau de conectividade é suficiente para conhecer os nós mais conectados (*hubs*). No entanto, em certas situações, a quantidade de ligações não é relevante, e, portanto, outras medidas devem ser consideradas.

- **Proximidade** (*closeness*): esta centralidade é baseada na distância entre um par de nós, definido por

$$\text{proximidade}_i = \frac{N-1}{\sum_{j=1;j\neq i}^{N} d(v_i, v_j)}, \tag{2.11}$$

  que representa quão próximo é um nó v_i do resto de nós v_j. Quanto mais próximo de 1 é o valor, mais central é o nó (MASUDA; LAMBIOTTE, 2016; STEEN, 2010).

- **Intermediação** (*betweenness*): Esta centralidade mensura o quão relevante é um nó segundo a quantidade de caminhos curtos, entre quaisquer vértices que passam por ele. Ela é definida como

$$\text{intermediação}_i = \sum_{r \neq i \neq s} \frac{\rho_{rs}(i)}{\rho_{rs}}, \tag{2.12}$$

onde ρ_{rs} é a quantidade de caminhos curtos do nó r para o nó s, e $\rho_{rs}(i)$, a quantidade de caminhos curtos que passam pelo nó i (SILVA; ZHAO, 2016).

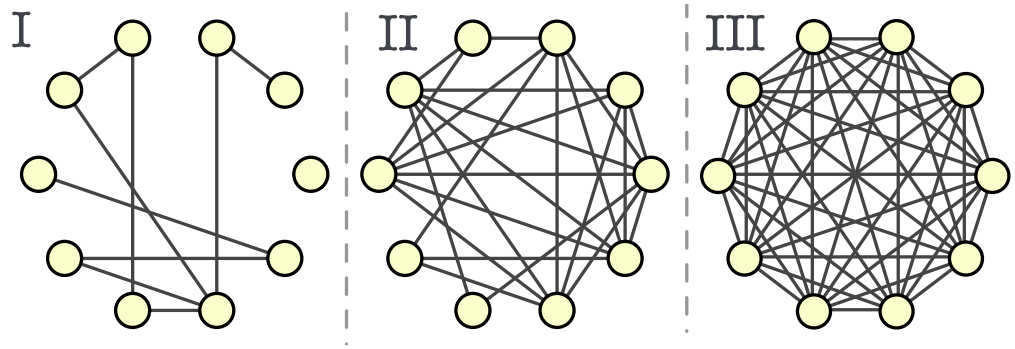
2.2 Modelos de rede

Os modelos de rede são estruturas construídas a partir de regras e parâmetros que permitem gerar redes com propriedades específicas. Desse modo é possível desenvolver redes que, mesmo tendo uma topologia e tamanho diferente, compartilham as mesmas propriedades.

2.2.1 Rede de Erdős e Rényi (ER)

Este foi um dos primeiros modelos de redes aleatórias a ser desenvolvido. A partir dele, muitas outras variações foram construídas. O modelo foi apresentado pelos matemáticos Erdős e Rényi (1960), e funciona baseado nas probabilidades. Uma forma de obter este modelo $G(n, m)$ está baseado na quantidade fixa de n nós e m ligações, desse modo, são geradas todas as combinações de redes possíveis que cumpram essa condição, e, posteriormente, é escolhido de forma aleatória uma rede entre este conjunto de possibilidades. Outra forma é construir a rede $G(n, p)$ a partir de uma quantidade fixa de n nós e uma probabilidade p (de ligar dois nós). Neste caso, a quantidade provável de ligações começa em 0 e vai até $\binom{n}{2}$. Desse modo, para cada possível ligação da rede é gerada uma probabilidade q, que se for menor ou igual ao valor de p, a ligação é construída entre os dois nós. Portanto, se $p = 1$, então a rede será completa (todos os nó conectados entre si), e no caso contrário ($p = 0$) a rede não terá ligações. Na Figura 2.7 são ilustradas diversas redes a partir desta última abordagem.

Figura 2.7 - Três redes geradas segundo o modelo de Erdős e Rényi



I) Rede gerada considerando $n = 10$ e $p = 0,1$ II) Outra rede, considerando $n = 10$ e $p = 0,5$. III) Por último, rede usando $n = 10$ e $p = 1$

Fonte: Produção do autor.

Nesta última abordagem, para termos $\langle m \rangle$ ligações é preciso que esta seja igual à probabilidade p vezes a quantidade máxima possível de ligações $\binom{n}{2}$,

$$\langle m \rangle = \binom{n}{2} p \tag{2.13}$$

desse modo, o coeficiente de grau (médio) é $\langle k \rangle = \frac{2\langle m \rangle}{n}$, que ao substituir a equação anterior, obtemos

$$\langle k \rangle = \frac{2\binom{n}{2} p}{n} = (n-1)p \tag{2.14}$$

portanto, a média de ligações que um nó possui é igual à probabilidade p vezes a quantidade de outros nós $n - 1$. Também sabemos que a probabilidade de um nó ter k ligações seria $p^k(1-p)^{n-1-k}$, desde que, se o nó tem k ligações, então ao longo da construção da rede é necessário ele ter ligado k vezes, e o restante das vezes $(n - 1 - k)$, não ligar. Levando esse cálculo para todos os nós da rede, obtemos a seguinte probabilidade binomial:

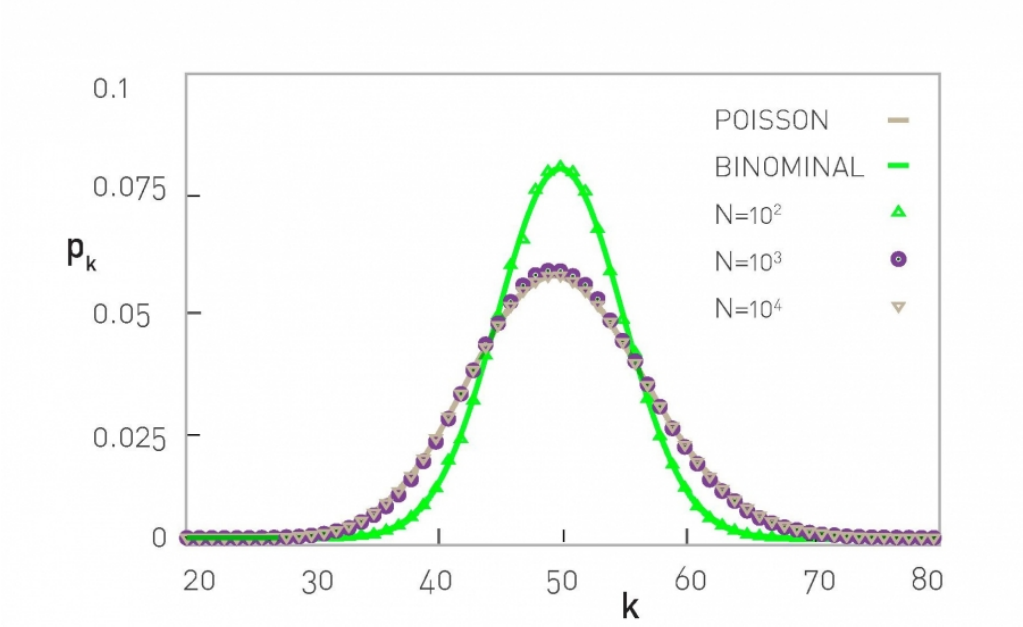
$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \tag{2.15}$$

após simplificar, chegamos na equação

$$P(k) = \frac{c^k e^{-c}}{k!}, \tag{2.16}$$

para $c = np$ e $n \to \infty$. Esta última equação é também conhecida como a distribuição de Poisson. Através dela é possível compreender que em redes aleatórias com muitos nós, a topologia comporta-se de forma similar à distribuição de Poisson. E no caso de redes com poucos nós, segue a distribuição Binomial (BARABÁSI; PÓSFAI, 2016). Na Figura 2.8 mostra-se de forma visual ambas as distribuições.

Figura 2.8 - Comparativo da distribuição do grau de conectividade para diversas redes seguindo o modelo aleatório.



É perceptível que as distribuições seguem regimes similares, porém, há uma diferença entre eles, especificamente entre a rede com $N = 10^2$ e as outras com quantidade maior de nós, conforme mostrado nas equações.

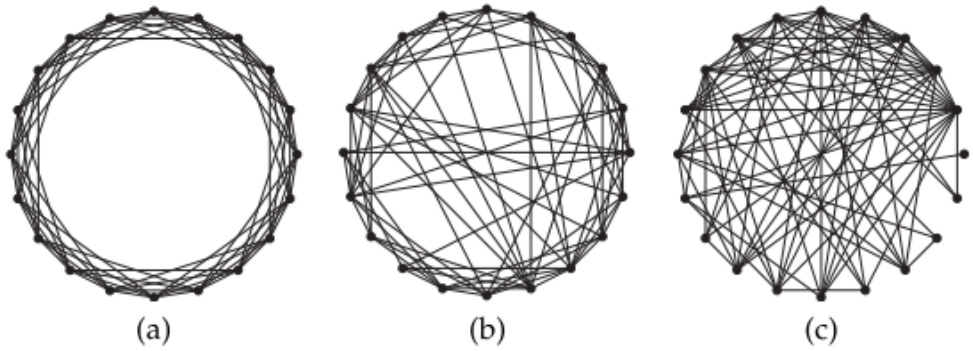
Fonte: Barabási e Pósfai (2016).

2.2.2 Watts e Strogatz (WS)

Apesar do trabalho de ER fornecer informações relevantes a partir de um modelo simples de rede aleatória, algumas propriedades observadas em sistemas reais não são encontradas nesse modelo (BARABÁSI; PÓSFAI, 2016). As principais diferenças estão no fato de que redes que se formam no mundo real possuem concentrações (*agrupamentos*) densas de nós em diversas partes da rede, algo não visto no modelo de ER. Também, num sistema real há nós com alta conectividade (*grau de conectividade*), muito acima do normal, enquanto o resto dos nós carecem dela. Devido a essas características do mundo real – que não é completamente aleatória mas também não é totalmente regular – Watts e Strogatz propuseram o modelo que faz a transição entre uma rede regular (todos os nós possuem a mesma quantidade de ligações) e aleatória, contendo nela o conceito de "mundo pequeno"(*small world*) (STEEN, 2010; ESTRADA, 2010; WATTS; STROGATZ, 2011). Basicamente, ele descreve que entre quaisquer dois vértices numa rede há um menor caminho entre eles, mesmo que a rede seja grande.

Um experimento conduzido por Milgram e publicado em 1967, sugeriu que é possível entrar em contato com qualquer pessoa do mundo, tendo apenas cinco pessoas como intermediadores entre remetente e destinatário, em outras palavras, em seis passos (BARABÁSI; PÓSFAI, 2016). O experimento, conhecido também como "Seis graus de separação", de certa forma, foi uma prova experimental do conceito que posteriormente WS propuseram no modelo de rede. A construção dele está baseada numa rede regular com n vértices, cada um deles ligado com k vértices. Depois, sequencialmente, é escolhido um vértice e uma aresta que está ligada ao vizinho mais próximo no sentido horário. Considerando a probabilidade p, esta ligação é desligada do vizinho e ligada em outro vértice, escolhido aleatoriamente. Após ter feito esse processo de forma sequencial para todos os nós e vértices, obtemos uma nova rede (SILVA; ZHAO, 2016). Na Figura 2.9 mostra-se um exemplo desta rede considerando três variações da probabilidade p.

Figura 2.9 - Três redes baseadas no modelo de Watts e Strogatz

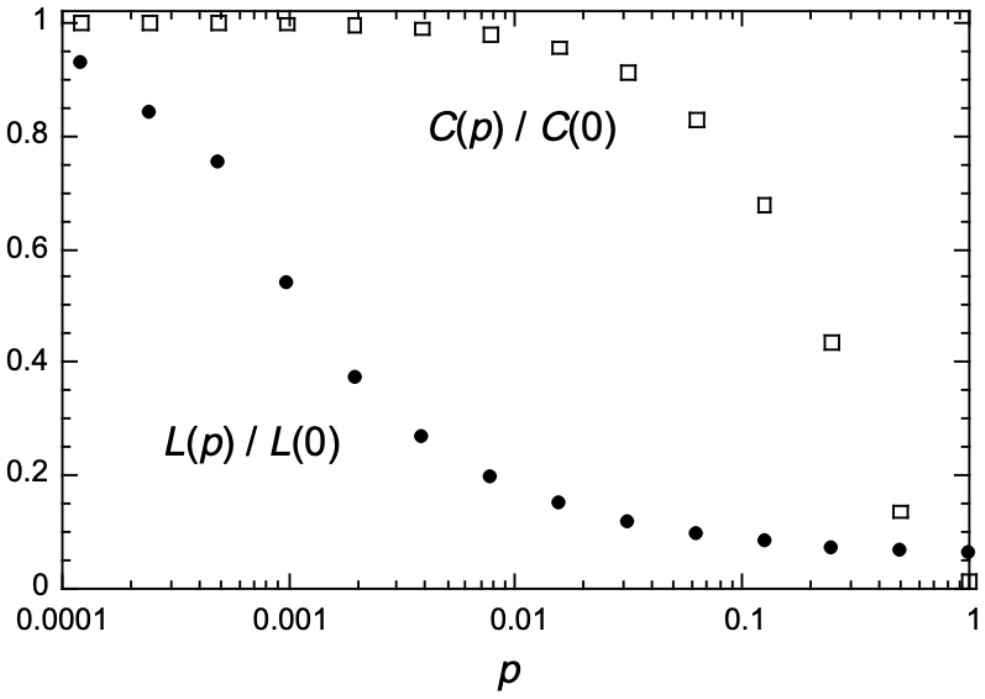


Redes com 20 nós e 8 ligações por cada nó, considerando a) $p = 0$, b) $p = 0.2$ e c) $p = 0.9$.

Fonte: Steen (2010).

É importante notar que, na rede WS, quando $p = 0$, a rede é regular e, portanto, bem agrupada; no entanto, quando $p = 1$ a rede é totalmente aleatória. Nesse sentido, quando o valor do p muda algumas propriedades se transformam gradualmente, entre essas propriedades estão o caminho médio e o coeficiente de agrupamento. Para valores de p próximos de 0 em direção a 1, há um efeito maior no tamanho do caminho $L(p)$ da rede do que no coeficiente de aglomeração $C(p)$. Isto deve-se ao fato de que ao ligar aleatoriamente algumas ligações são construídas "pontes"entre diversas seções da rede, permitindo alcançar em menos passos qualquer parte da rede, mesmo que sejam poucas as ligações que foram reconectadas (WATTS; STROGATZ, 2011). Na Figura 2.10 observamos como essa relação é construída a partir dos valores normalizados de $L(p)$ e $C(p)$.

Figura 2.10 - Valores de L e C em relação à probabilidade p em redes segundo o modelo de Watts-Strogatz.



Relação entre o Coeficiente de Agrupamento e o Caminho (normalizados) para diversos valores de p.

Fonte: Watts e Strogatz (2011).

2.2.3 Redes livres de escala

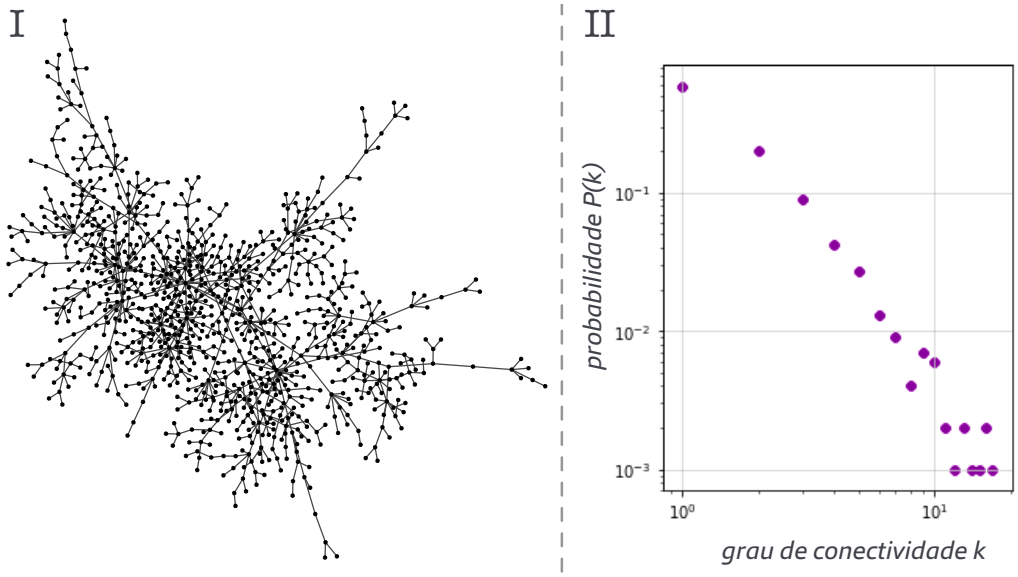
São redes que a distribuição do grau de conectividade segue o comportamento de uma distribuição baseada na lei de potência $P(k) \sim k^{-\gamma}$, onde $k > 0$ e $\gamma > 0$ (CALDERELLI, 2007). A simplicidade da equação e proximidade com redes do mundo real, fizeram que este tipo de redes sejam estudadas com profundidade em diversas áreas da ciência (FORTUNATO et al., 2006). Apesar de que nos últimos anos foram levantados questionamentos sobre a proximidade entre as redes empíricas e as redes livres de escala, é inegável que estas foram essenciais para fazer grandes avanços na exploração das redes complexas como uma forma de compreender os sistemas complexos (BROIDO; CLAUSET, 2019; BARABÁSI; PÓSFAI, 2016).

O modelo de Barabási e Albert (1999) conseguiu explicar de melhor forma as es-

truturas de sistemas reais que, até aquele momento, não eram tão bem representadas nos trabalhos já realizados. Era visto que existia uma concentração grande de ligações em poucos nós de um sistema, enquanto a maior parte dos nós possuía poucas ligações, além disso, elas seguiam uma certa ordem observada em diversos sistemas do mundo real, desde a Internet até a natureza (NEWMAN, 2010). O mecanismo proposto por BA explica este fenômeno com base no conceito da *ligação preferencial* e do *crescimento* da rede (BARABÁSI; ALBERT, 1999).

Apesar dos modelos de ER e BA seguirem um regime aleatório, há diferenças claras nas topologias que são geradas. A construção desta rede começa com uma quantidade pequena de nós n_0. Então, em cada passo de tempo é adicionado um novo vértice com uma quantidade $m < n_0$ de arestas, com probabilidade proporcional à quantidade de ligações que os nós já possuem (ESTRADA, 2011). Desse modo, os novos nós tendem a ligar com maior frequência com nós que são *hubs*. A adição de novas ligações e nós, assim como o mecanismo de ligação preferencial, são diferenças determinantes para a geração de uma topologia diferente se comparada com a rede de ER (BARABÁSI; PÓSFAI, 2016). A Figura 10 ilustra um exemplo do regime proposto por BA. Na Figura 2.11, mostra-se a distribuição de grau de uma rede livre de escala, composta por 10^3 nós e $m = 1$.

Figura 2.11 - Modelo Barabási-Albert



I) Ilustração de uma rede segundo o modelo Barabási-Albert considerando $n = 1000$ e $m = 1$. II) Distribuição de grau da rede mostrada, em escala logarítmica.

Fonte: Produção do autor.

2.3 Comunidades

Desde a estatística, passando pela computação e biologia, até a física, encontramos diversos problemas relacionados à detecção de conjuntos de elementos com atributos em comum (FAN et al., 2014; SHOBANA; KUMAR, 2015). Seja para obter a uma análise de sequencia do ADN (ácido desoxirribonucleico) ou encontrar grupos de amizades numa rede social, a tarefa de agrupar corretamente apresenta diversos desafios, tais como otimizar o desempenho computacional ou obter uma boa acurácia nos agrupamentos (MANYIKA et al., 2011).

Nas redes complexas, a procura de agrupamentos de nós ou arestas é nomeado como detecção de comunidades. As comunidades são identificadas quando há uma quantidade maior de ligações entre nós de um mesmo subgrupo da rede, e quantidade menor de ligações entre vértices que pertencem a subgrupos diferentes (BARABÁSI; PÓSFAI, 2016). Por isso, apesar de ser possível usar algoritmos de agrupamento tradicionais da mineração de dados nas redes complexas, devido ao fato de elas não considerarem a estrutura topológica da rede, os métodos tradicionais

não são apropriadas para a detecção de comunidades nas redes complexas (SILVA; ZHAO, 2016).

No intuito de mensurar o quão bem formados são os agrupamentos encontrados nas redes, foi proposto o conceito de *modularidade* por Newman e Girvan (2004), para estabelecer um indicador sobre as divisões das comunidades. Mais exatamente, ela é definida por

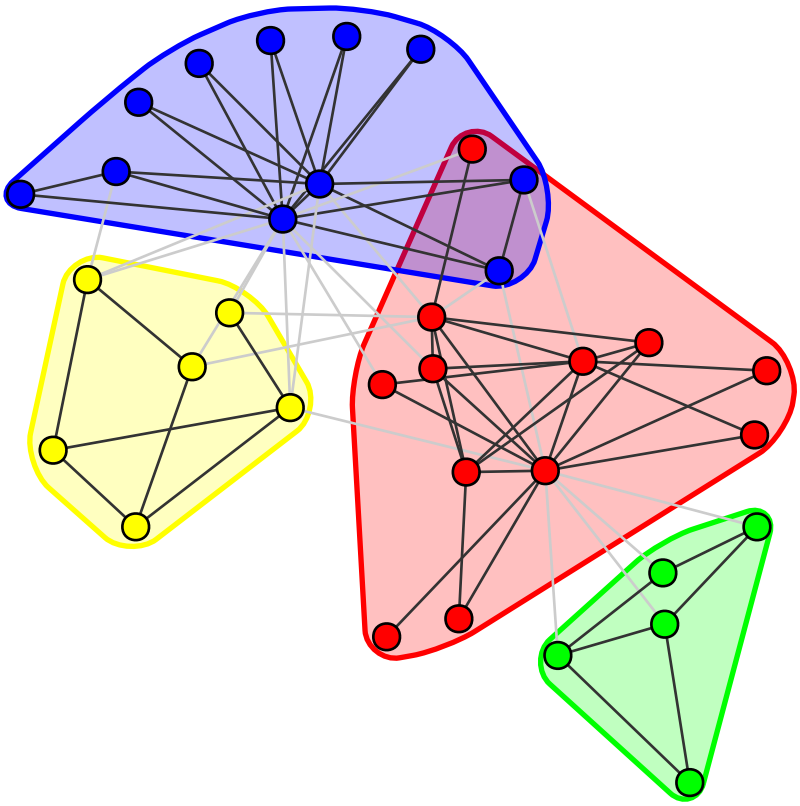
$$\text{modularidade} = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j), \tag{2.17}$$

onde c_i e c_j são comunidades com vértices i e j em cada comunidade, respectivamente. Se as duas comunidades de c_i e c_j são iguais, então $\delta = 1$, e no caso contrário, $\delta = 0$ (NEWMAN, 2010). Desse modo, quando a modularidade de uma rede é próxima de 0, significa que a rede é muito próxima de uma rede aleatória, onde não há presença de comunidades, e se a modularidade se aproxima de 1, isso significa que a rede muito provavelmente possui comunidades (BARABÁSI; PÓSFAI, 2016). A partir do trabalho onde a modularidade foi proposta, iniciou-se uma corrida para otimizar e desenvolver novos algoritmos para a detecção de comunidades. Essa nova área permitiu o desenvolvimento de outras linhas de pesquisa como a detecção de padrões em comunidades, interações entre comunidades, algoritmos eficientes para identificação de comunidades, entre outros (LATORA et al., 2017).

Na literatura encontramos diversos exemplos de comunidades em redes complexas. Por exemplo, conjuntos de paginas web que estão vinculadas seja tematicamente ou por fazerem parte da mesma organização (YANG; LIU, 2008). Também, grupos de amizades que se formam em redes sociais a partir do ambiente do trabalho ou estudo (HSU et al., 2007). Redes formadas por autores de artigos científicos também possuem comunidades, devido à proximidade da área de pesquisa ou por pertencer ao mesmo centro acadêmico (THELWALL; KOUSHA, 2014).

Para ilustrar este conceito, na Figura 2.12 mostra-se as comunidades da rede do famoso *Clube de Zachary*. Esta rede representa as interações que 34 membros de um determinado clube de karatê mantinha fora do clube, e que foi amplamente estudada por Wayne W. Zachary entre os anos 1970 e 1972 (GIRVAN; NEWMAN, 2002).

Figura 2.12 - Rede do Clube de Zachary



Quatro comunidades encontradas na rede *Clube de Zachary*, com modularidade $= 0.419$. O método usado foi o de *Louvain*, que é explicado posteriormente nesta seção.

Fonte: Produção do autor.

Apesar de atualmente termos dezenas de métodos disponíveis para detecção de comunidades, em termos de acurácia e custo computacional, ainda é difícil afirmar qual é o melhor algoritmo (YANG et al., 2016; KRAWCZYK, 2009; CHAKRABORTY et al., 2016). No entanto, entre todos os algoritmos propostos, há claramente um conjunto de algoritmos que se destacam pela sua flexibilidade (é possível usar em diversas estruturas de redes) ou velocidade (são eficientes em redes imensas) (WAGENSELLER et al., 2018; LATORA et al., 2017). Neste trabalho, considerando as avaliações e comparações feitas por diversas pesquisas que comparam os algoritmos de detecção de comunidades, será usado os método desenvolvidos por Blondel et al. (2008) (Multilevel). Outro algoritmo a destacar neste grupo é o desenvolvido por Rosvall e Bergstrom (2008) (Infomap), devido ao seu desempenho em diversos cenários, sejam redes imensas ou pequenas e pelo baixo custo computacional en-

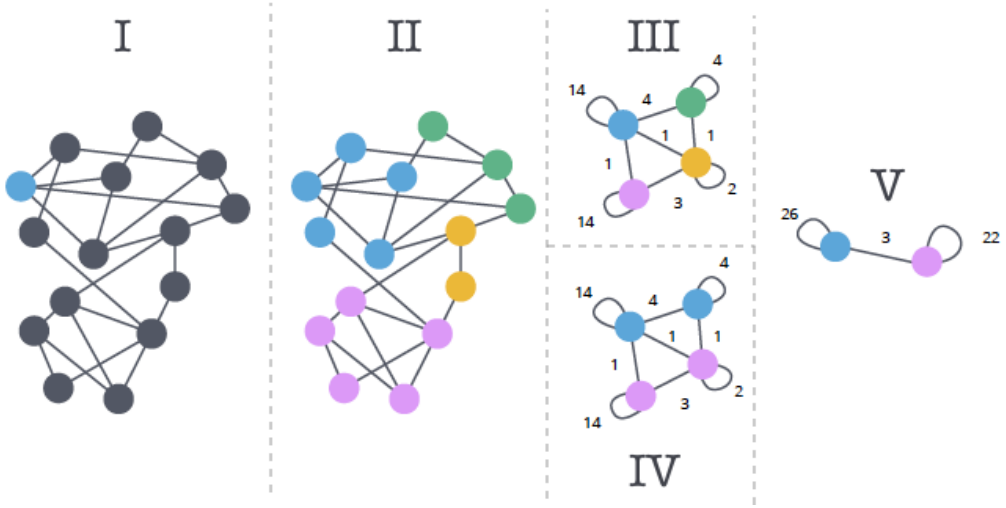
volvido. É importante ressaltar que o foco deste trabalho não é encontrar o melhor algoritmo de detecção de comunidades nem fazer um comparativo entre elas, mas usá-los em conjunto com os métodos propostos.

2.3.1 Método Louvain (Multilevel)

Ao ser desenvolvido pelos pesquisadores Blondel et al. (2008) da Universidade Católica de Louvain, o método acabou ganhando fama como Método Louvain, enquanto outros o nomeiam como Método *Multilevel*, devido à abordagem para construir as comunidades através de camadas de agrupamento de nós. O desempenho do método permite que seja usado tanto em pequenos grafos assim como em redes com milhões de nós e ligações (ZINOVIEV, 2018). Por exemplo, numa rede empírica (usando dados reais) da Internet com 70 000 nós e 351 000 ligações, o algoritmo conseguiu completar a tarefa de encontrar as comunidades em menos de um segundo (AYNAUD et al., 2013).

O algoritmo é dividido em duas partes: na primeira, cada nó é da rede é considerado uma comunidade, desse modo, numa rede com N nós, também teremos a mesma quantidade de comunidades. Depois, cada nó i é colocado na comunidade do vizinho j, e, então, é calculado o ganho na modularidade na adição ou remoção do nó i nas comunidades vizinhas. Assim, o intuito é que o nó i fique na comunidade que tiver a melhor modularidade, caso contrário, ele voltará à comunidade anterior que estava. Na segunda etapa, todas as comunidades são consideradas como nós, e as ligações entre elas agora possuem pesos correspondentes à quantidade de ligações que tinham inicialmente, antes de serem simplificadas. No caso das ligações dentro de uma mesma comunidade, elas serão convertidas em autoligações. Logo após este processo de transformação da rede, o processo (primeira e segunda etapa) é repetido até não ser possível melhorar a modularidade ou atingir um limiar desejado. Em poucas palavras, o algoritmo une recursivamente as comunidades em um único nó e calcula a modularidade destes, comparando os valores obtidos até chegar no melhor valor da modularidade possível (LATORA et al., 2017). Na Figura 2.13 observa-se uma ilustração do processo iterativo deste algoritmo.

Figura 2.13 - Método Louvain



Exemplo de quatro passos para encontrar as comunidades de uma rede seguindo o método Louvain. I) Rede original, e em azul, o nó escolhido para começar o cálculo da modularidade. II) Os nós são agrupados (em cores diferentes), segundo a maximização da modularidade. III) Transformação de todas as comunidades em nós. VI) Novamente, similar ao passo (II), os nós são agrupados segundo a modularidade. IV) E finalmente, a melhor agrupação possível, considerando a modularidade.

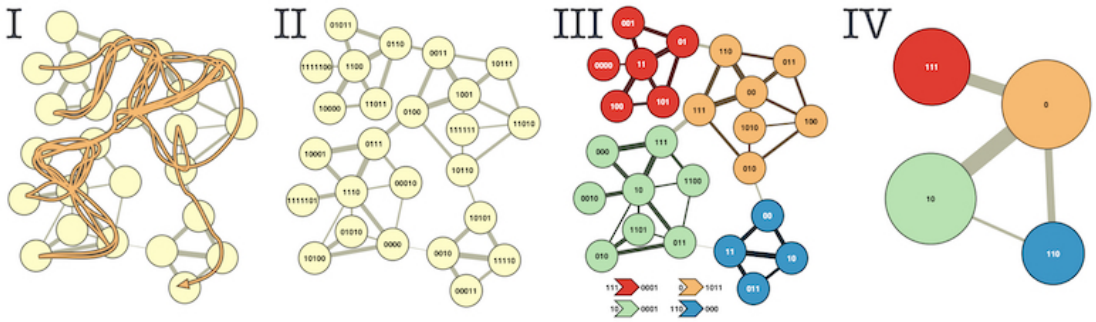
Fonte: Adaptado de Needham e Hodler (2019).

2.3.2 Método Infomap

O algoritmo é baseado na caminhada aleatória através do nós de uma rede, e representou um grande passo no uso dessa abordagem para detecção de comunidades. Desenvolvido por Rosvall e Bergstrom (2008), este método destacou-se por ser útil na procura de comunidades em redes imensas (LATORA et al., 2017). O algoritmo começa nomeando os nós através da codificação baseada no trabalho de Huffman (1952) (este método aloca nomes curtos para nós mais frequentes), e, desse modo, ao longo da caminhada aleatória sobre a rede obtemos uma série composta pelos nomes dos nós. Esta caminhada acontece sobre todos os nós. Depois disso, uma segunda interação é realizada, desta vez, agrupando os nós e usando diferentes nomes para cada grupo. Esta agrupação é baseada na ideia de que é muito provável que um caminhante aleatório permaneça mais tempo em nós que pertencem à mesma comunidade. Neste ponto, um processo de otimização é realizado a partir da *Equação de Mapeamento* (*Map Equation - MP*). Baseado

no fluxo da informação (nomes dos nós e a caminhada aleatória), o MP é usado para procurar as representações mais prováveis e curtas da caminhada, e assim, adicionar nós ou removê-los das agrupações segundo o valor do MP. O processo é repetido até que as comunidades se mantenham estáveis ou o MP alcance um limiar fixo (que não varia muito) (ROSVALL; BERGSTROM, 2008). Na Figura 2.14 ilustra-se o processo deste método de forma geral.

Figura 2.14 - Método Infomap



Processo de detecção de comunidades segundo o método Infomap. I) Um caminhante desloca-se pela rede de forma aleatória. II) Nomeia-se cada nó com códigos curtos usando o algoritmo de Huffman. III) Identifica-se os grupos de nós que pertencem à mesma comunidade. IV) Agrupa-se cada comunidade em nós, formando desse modo, uma rede menor. Após isso, repete-se o processo todo.

Fonte: Adaptado de Rosvall e Bergstrom (2008).

2.4 Resolução das redes

Todo o desenvolvimento realizado sobre redes complexas encontra-se num nível de resolução que é importante considerar em qualquer interpretação que seja feita (YANG et al., 2016; MARCHIORI; POSSAMAI, 2015). Por exemplo, o coeficiente de conectividade de um *nó* é diferente da média do coeficiente de conectividade da *rede*. Ambos valores tem utilidade em diferentes situações. O primeiro, serve para saber o quão conectado é um nó e desse modo saber se é um *hub* importante da rede, o segundo, pode ser utilizado para saber quão denso é uma rede. Nesse sentido, as medidas ou índices que encontramos na literatura, assim como os métodos ou algoritmos, precisam ser consideradas dentro de um nível de resolução ou escala

(TUNÇ; VERMA, 2015).

Na literatura encontramos três principais níveis:

- **Micro-escala**: neste nível o foco são as propriedades de uma ligação ou nó, de forma separada. Também, é possível desenvolver técnicas ou algoritmos que explorem propriedades de cada nó ou ligação. Algumas medidas que se encontram neste nível são: o grau de conectividade de um nó, a força dos nós, o peso de uma ligação, entre outros (MARCHIORI; POSSAMAI, 2015; TUNÇ; VERMA, 2015). Alguns exemplos de aplicação neste nível de resolução seriam: identificar a pessoa mais popular de um grupo de amizades na escola, ou saber qual avenida/ponte da rede de transporte é essencial para manter o bom fluxo de trânsito na cidade.

- **Meso-escala**: neste nível intermediário estuda-se as propriedades dos elementos de forma individual, mas considerando a rede como um sistema único. Em outras palavras, analisa-se as propriedades que afetam a conectividade de grupos de nós ou ligações da rede. Um exemplo nesta categoria é a detecção de comunidades ou o cálculo da modularidade de um conjunto de nós (GARGIULO et al., 2016; ROMBACH et al., 2014). Alguns exemplos do que é possível analisar nesta escala são: encontrar o grupo de estudantes mais engajados dentro da rede de alunos faculdade ou o conjunto de páginas web que são relevantes para divulgar ou lançar uma campanha de marketing.

- **Macro-escala**: aqui o foco é considerar a rede como um único elemento, analisando suas propriedades de forma geral ou global. Em alguns casos, as propriedades do nível micro são usadas para o cálculo no nível macro. No entanto, é possível que, dependendo do tamanho da rede ou do tipo de propriedade, qualquer mudança no nível micro sejam insignificante (TUNÇ; VERMA, 2015; MARCHIORI; POSSAMAI, 2015). Nesse sentido, neste nível de resolução, as medidas são utilizadas como indicadores globais, que permitem fazer comparações e simplificar informações a partir da rede toda. Algumas aplicações deste nível de análise são: o cálculo da densidade das amizades de uma rede social ou o índice de desigualdade existente na rede de parcerias econômicas entre cidades ou países.

2.5 Redes de sistemas complexos

O avanço das redes complexas seria impossível sem a necessidade constante de resolver desafios presentes em sistemas complexos (STROGATZ, 2001). Aliás, a principal motivação para o desenvolvimento de modelos, métodos e medidas nas redes complexas deve-se às limitações existentes nos métodos tradicionais para resolver problemas nos sistemas complexos, presentes no mundo real (LIU et al., 2011). Portanto, é preciso mencionar algumas destas áreas onde o desenvolvimento desta pesquisa pode ser aplicada, desde que, as redes complexas já fizeram contribuições significativas nelas. Estas são:

- **Redes biológicas**: na biologia vemos redes complexas em todos os níveis, seja nos seres humanos a nível microscópico – como a atividade no cérebro dos neurônios –, ou em dimensões maiores, como as interações entre especies de animais a nível intercontinental. Grandes avanços foram realizados, por exemplo, na compreensão da propagação de epidemias em estrutura de redes, a partir da construção de modelos de propagação sobre redes complexas (PASTOR-SATORRAS et al., 2015). Na neurociência, ajudou a mensurar, caracterizar e compreender a estrutura do cérebro, com base na análise das estrutura das comunidades em redes do cérebro (BETZEL; BASSETT, 2017). Também, na identificação de elementos que influenciam no desregulamento das funções biológicas que dão origem ao câncer de mama, usando redes complexas construídas a partir de dados de tecidos da mama saudáveis e com câncer (ANDA-JÁUREGUI et al., 2018).

- **Redes de infraestrutura**: é impraticável não fazer parte de redes de infraestrutura hoje em dia. O deslocamento dos caminhões que levam os alimentos e produtos para o comércio entre cidades, passam uma rede imensa de estradas e ruas (PORTA et al., 2006). Também, o transporte de energia é feito através de uma rede de distribuição e fornecimento a partir de diversas fontes, que são levadas de longínquos extremos de um ponto a outro num país ou até mesmo num dado continente, sobre uma rede complexa de cabos e estações de controle (GASTNER; NEWMAN, 2006). Nesse sentido, as redes de infraestrutura estão relacionadas a características espaciais ou de localização, e, portanto, não é suficiente considerar apenas a topologia, desde que há outros atributos envolvidos como a distância, a capacidade, etc (BARTHÉLEMY, 2011). Alguns dos estudos nesta área tratam, por exemplo, sobre como aprimorar a robustez

das redes de energia perante um eventual ciberataque ou até mesmo, de um ataque físico (HE; YAN, 2016). Também, noutro exemplo, uma pesquisa analisou a resiliência das redes de transporte aéreo na China e comparou o quanto ela é diferente da rede global de aviação (DU et al., 2016).

- **Redes do clima**: o uso das redes complexas no estudo do clima trouxe muito interesse sobre o seu potencial uso para a descoberta de teleconexões (fenômenos climático extremos e conectados) em diversos níveis, seja na superfície do mar ou a nível atmosférico, e entre áreas muito distantes (ZHOU et al., 2015). Também, devido à imensidade de dados e áreas a serem analisadas, novos métodos otimizados para este tipo de redes foram desenvolvidos a partir da similaridade (correlação) entre pontos do globo terrestre (DONGES et al., 2009). Por exemplo, a partir da análise de redes foi possível identificar áreas com maior impacto do fenômeno El Niño (FAN et al., 2017; GOZOLCHIANI et al., 2011). Também, foi possível encontrar padrões nas tele-conexões de eventos extremos de chuva ao redor do mundo (RHEINWALT et al., 2019).

- **Redes sociais**: ambientes sociais físicos e virtuais hoje atingem milhões de usuários a cada dia, além do tamanho, a velocidade com que elas mudam é impressionante (STARK; CASTELLS, 1997). Nesse sentido, devido à complexidade de dados necessário para analisar esse grande volume, as redes complexas apresentam-se como um caminho viável na descoberta de informações (GHOSH; GANGULY, 2014). Também, devido à necessidade de recomendar novas amizades, os algoritmos de redes complexas foram usadas para a identificação e recomendação de potenciais novas ligações entre os usuários (WANG et al., 2014). Noutro caso, por exemplo, as redes revelaram as mudanças que houve na comunicação entre diretivos e empregados da empresa Enron, a partir dos e-mails trocados entre eles, o que mostrou ser uma evidência que eles estavam prestes a falir após uma série de escândalos (DIESNER et al., 2005). De modo geral, as redes sociais são uma excelente fonte de dados para o estudo das relações sociais, e portanto é esperado que elas revelem fenômenos humanos como a propagação de epidemias, formação de protestos, entre outros (WEY et al., 2008; PIEDRAHITA et al., 2018).

- **Outras áreas**: é indiscutível que há muitas outras áreas que fazem uso das redes complexas em diversos cenários. Desde a luta contra a cor-

rupção, na identificação de coalizões políticas que, ao longo do tempo, colaboram entre si para se perpetuar em posições ou cargos de poder (RIBEIRO et al., 2018), até o uso de redes na arqueologia, para estabelecer relações entre lugares e tempos, a partir de dados incompletos (PRIGNANO et al., 2017). Sendo assim, é perceptível que as aplicações e usos das redes complexas são ilimitadas, e, nesse sentido, é esperado que novas linhas de pesquisa sejam desenvolvidas não apenas dentro das redes complexas, mas também em conjunto com outras áreas.

2.6 Linhas de pesquisa

Dentro das ciência das redes, algumas áreas incitaram o interesse a novas pesquisas. Entre elas, a descoberta do tipo de distribuição que redes reais possuem (KUNEGIS, 2014), e a discussão sobre se, de fato, as redes reais se assemelham às redes livres de escala (BROIDO; CLAUSET, 2019). Também, outra frente de pesquisa é o cálculo da resiliência da rede com aplicação direta na proteção de redes críticas, como a de energia ou transporte (GAO et al., 2016). Além disso, outra linha de pesquisa encontra-se na análise da topologia das redes de grande tamanho para obter características da rede em diversas camadas ou níveis (GÓMEZ et al., 2018). Por outro lado, algumas pesquisas são orientadas ao estudo da sincronização em redes, com utilidade na otimização de recursos distribuídos numa rede, seja de energia ou processamento computacional, por exemplo (ESTRADA, 2011). Também, a exploração de redes que são temporais é uma área com amplo crescimento na literatura, tendo ênfase na análise de dinâmicas temporais na rede, como o surgimento de comunidades, propa gação de informação, entre outros (HOLME, 2015).

3 MINERAÇÃO DE DADOS TEMPORAIS

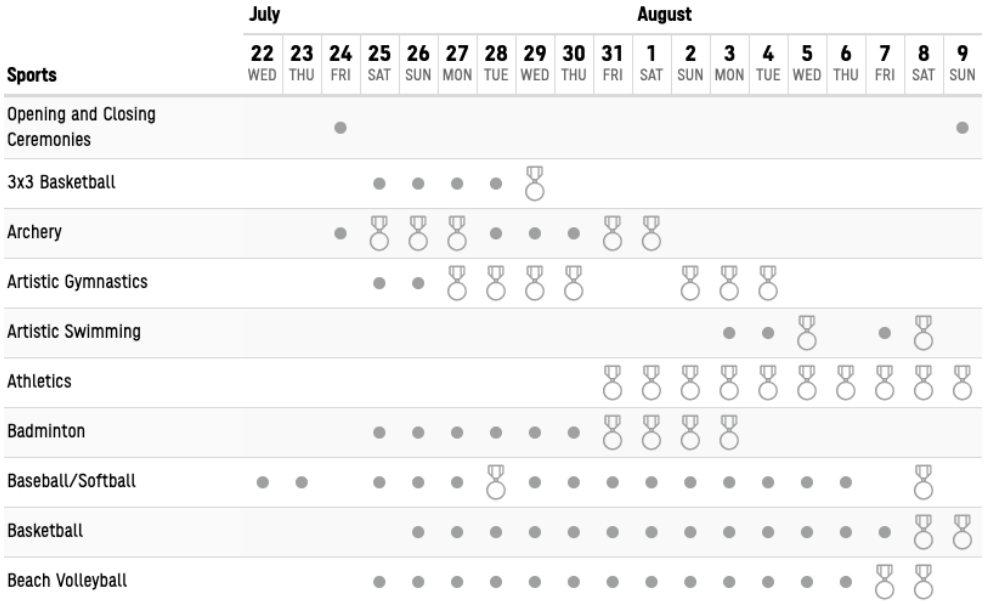
A Mineração de Dados Temporais (MDT) é uma área que, hoje em dia, inclui diversas linhas de pesquisa como: reconhecimento de padrões, banco de dados, visualização de dados, computação de alto desempenho, computação paralela, entre muitas mais. Portanto, neste capítulo explica-se de modo breve os conceitos e métodos mais estudados na literatura sobre mineração de dados temporais. Para esse fim, o conteúdo deste capítulo está principalmente baseado nos livros e artigos de (MITSA, 2010; LIU; ÖZSU, 2009; DUNHAM, 2002).

Em breves palavras, a mineração de dados temporais tem como foco encontrar padrões a partir dos desdobramentos das variáveis temporais (ANTUNES; OLIVEIRA, 2001). A complexidade desta tarefa é acompanhada de outros desafios quando, por exemplo, o volume de dados é imenso; os dados tem outros componentes importantes além do tempo; é necessária visualização dos dados, entre outros. Para compreender melhor os desafios atrelados à mineração, é preciso compreender as principais características dos dados temporais e como elas são estruturadas.

3.1 Características gerais dos dados temporais

Para que um conjunto de dados seja temporal, basta ter um atributo que faça referência ao tempo, e, nesse sentido, a maior parte dos bancos de dados relacionais e não relacionais já possuem tipos de dados pré-definidos que fazem referência a este tipo de atributo (SNODGRASS, 1992). Por exemplo, é comum encontrarmos o *Date* (ano, mês e dia), *Time* (hora, minuto, segundo), *Timestamp* (ano, mês, dia, hora, minuto, segundo), *Interval* (inicio e fim). Porém, os dados temporais não ficam restritos aos atributos pré-definidos nos bancos de dados, também é possível ter outros atributos temporais como a estação do ano, número da semana, do ano, entre outros (MITSA, 2010). Adicionalmente, em muitos casos é possível encontrar relações entre entidades temporais. Sobre isso, Chittaro e Montanari (2000) estabeleceu um conjunto de treze possíveis relações, como por exemplo: "antes de", "durante", "com inicio em", entre outros. De forma ilustrativa, na Tabela 3.1 vemos um exemplo de registros temporais.

Figura 3.1 - Calendário parcial dos Jogos Olímpicos - Japão 2020



Estrato da lista dos esportes em competição e as respectivas datas.
Fonte: Tokio2020 (2020).

3.2 Tipos de dados temporais

De modo geral, é possível concluir que são quatro os principais (ou mais comuns) tipos de dados temporais encontrados na literatura: *Series temporais*, *Sequências temporais*, *Eventos temporais* e *Dados semânticos temporais*, a partir dos trabalhos de (CHITTARO; MONTANARI, 2000; MITSA, 2010; ATLURI et al., 2018). No entanto, a depender da definição ou área de pesquisa, em alguns casos são estabelecidos novos tipos de dados temporais.

3.2.1 Eventos temporais

De modo geral, um *evento temporal* é caracterizado como o registro de uma *atividade* (ou categoria) com *tempo* (*timestamp*) específico (ATLURI et al., 2018). Em outras palavras, ele é composto pelo *quando* e *o quê*. Nesse sentido, um evento temporal corresponde a uma ação no tempo. Alguns exemplos são: uma compra online no domingo de manhã, o primeiro voo de São Paulo a Lisboa no dia 01 de janeiro, o dia do nascimento do filho(a), etc. Na Figura 3.2 ilustra-se um conjunto de eventos temporais.

Figura 3.2 - Eventos chaves da Segunda Guerra Mundial

Detalhes dos eventos que foram marcos importantes ao longo da Segunda Guerra Mundial.

Fonte: Folha de São Paulo (2009).

Também, quando a dimensão espacial está presente num evento temporal, ela é caracterizada como evento espacial e temporal (ST). Portanto, a partir da observação de um evento (do mundo real) é possível obter os três componentes básicos: tempo, espaço e ação (ATLURI et al., 2018). E a partir da combinação deles, compor outros tipos de dados. Alguns exemplos de eventos ST são: turnê que uma orquestra realiza ao longo do Brasil, paradas que um táxi realiza ao percorrer a cidade, etc.

3.2.2 Series temporais

Uma *serie temporal* é um conjunto de valores $X = \{x_1, x_2, \ldots, x_n\}$ a partir de uma medida, que correspondem a observações reais (eventos) em intervalos de tempo $t = t_1, t_2, \ldots t_n$ (MITSA, 2010). Uma serie temporal pode ser dividida em duas classes: *multivariada* e *univariada*. A primeira refere-se a série quando é composta por mais de uma variável, enquanto a segunda é baseada em apenas uma variável. Outra classificação também é feita em relação à variação da série. Ela é *estacionária* quando a média e a variância não mudam ao longo do tempo, e *não estacionária* quando caso contrário (as propriedades estatísticas mudando) (MITSA, 2010). São exemplos de series temporais: a temperatura média horária em graus Célsius de uma cidade no mês de agosto, a velocidade de um carro enquanto se desloca de uma cidade a outra, etc. Na Figura 3.3, por exemplo, observa-se o gráfico da evolução (da série temporal) do preço do dólar (dos Estados Unidos de América) em relação ao real (do Brasil), nos últimos 7 anos.

Figura 3.3 - Variação do preço em Reais (R$) do Dolar (USD$) - Série Temporal

Máximos e mínimos mensais do preço do dólar em relação ao real.
Fonte: Investing.com (2020).

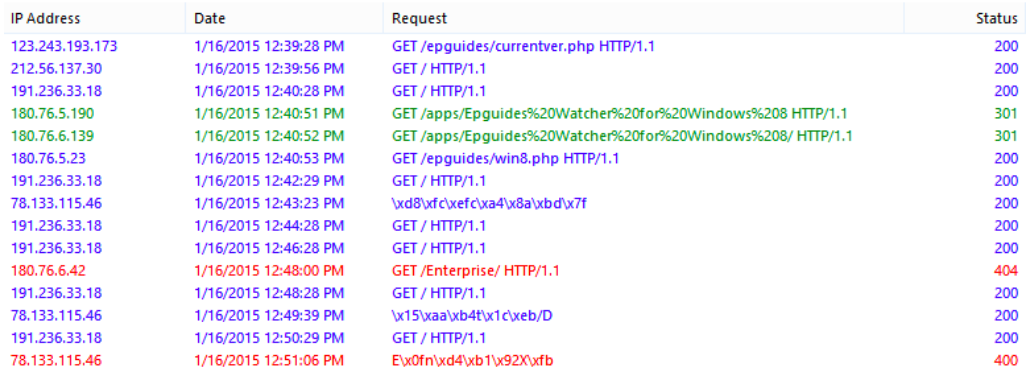
3.2.3 Sequências temporais

As *sequências temporais* são um conjunto de eventos temporais, que têm suas diferenças com as series temporais principalmente em relação à dispensa de estar baseada em alguma medida observada. Adicionalmente, nas sequências temporais é preciso ter um registro temporal relacionado a um acontecimento num período ou momento específico (MITSA, 2010). Desse modo, neste grupo encontram-se os dados temporais que possuem uma ordem temporal (com ou sem intervalos regulares) e a atividade realizada em cada registro temporal.

Por exemplo, a lista de entradas e saídas dos trabalhadores de uma empresa, a sequência de transações financeiras que uma pessoa faz numa loja, etc. Em ambos os casos, ao lado do registro temporal (dia e hora por exemplo) é definido

uma atividade realizada (entrou, saiu, comprou, etc). Para ilustrar, na Figura 3.4 observa-se os registros das requisições realizadas a um sítio web.

Figura 3.4 - Requisições de visitas web - Sequencia Temporal



IP Address	Date	Request	Status
123.243.193.173	1/16/2015 12:39:28 PM	GET /epguides/currentver.php HTTP/1.1	200
212.56.137.30	1/16/2015 12:39:56 PM	GET / HTTP/1.1	200
191.236.33.18	1/16/2015 12:40:28 PM	GET / HTTP/1.1	200
180.76.5.190	1/16/2015 12:40:51 PM	GET /apps/Epguides%20Watcher%20for%20Windows%208 HTTP/1.1	301
180.76.6.139	1/16/2015 12:40:52 PM	GET /apps/Epguides%20Watcher%20for%20Windows%208/ HTTP/1.1	301
180.76.5.23	1/16/2015 12:40:53 PM	GET /epguides/win8.php HTTP/1.1	200
191.236.33.18	1/16/2015 12:42:29 PM	GET / HTTP/1.1	200
78.133.115.46	1/16/2015 12:43:23 PM	\xd8\xfc\xefc\xa4\x8a\xbd\x7f	200
191.236.33.18	1/16/2015 12:44:28 PM	GET / HTTP/1.1	200
191.236.33.18	1/16/2015 12:46:28 PM	GET / HTTP/1.1	200
180.76.6.42	1/16/2015 12:48:00 PM	GET /Enterprise/ HTTP/1.1	404
191.236.33.18	1/16/2015 12:48:28 PM	GET / HTTP/1.1	200
78.133.115.46	1/16/2015 12:49:39 PM	\x15\xaa\xb4t\x1c\xeb/D	200
191.236.33.18	1/16/2015 12:50:29 PM	GET / HTTP/1.1	200
78.133.115.46	1/16/2015 12:51:06 PM	E\x0fn\xd4\xb1\x92X\xfb	400

Respostas do servidor web às requisições dos visitantes.
Fonte: ApacheViewer (2020).

Note-se que as *sequências* temporais também podem ser transformadas em *séries* temporais através de uma generalização ou agrupamento (ATLURI et al., 2018). No caso da generalização, será necessário adicionar alguma coluna que seja relativo a um valor mensurável (duração, intensidade, etc), e se não for em intervalos regulares, completar com algum método de interpolação dos dados. Por outro lado, também é possível usar o método do agrupamento, que quantifica a frequência de eventos em intervalos regulares, formando desse modo, uma série temporal (MITSA, 2010). Por exemplo, na Figura 3.4 seria possível adicionar uma coluna que contenha a quantidade de segundos que demorou a resposta do servidor, e, assim, obtermos a série temporal do tempo de resposta. Também, outra forma seria através da contagem de requisições em intervalos de 10 minutos. Em ambos os casos, obteríamos uma série temporal a partir de uma sequência temporal.

3.3 Medidas básicas

Alguma técnicas para reduzir a dimensão ou simplificar a complexidade de dados temporais está baseada em cálculos estatísticos (MITSA, 2010). Com eles é possível representar características globais de uma série temporal e, a partir deles, compor outros atributos. Desse modo, supondo que temos uma série temporal

$X = \{x_1, x_2, \ldots, x_n\}$ com n valores. Então, *média* de de X é dada por

$$\mu = \frac{\sum x_i}{n}. \tag{3.1}$$

Também, outra medida estatística importante é a *mediana*. Ela mostra o valor que encontra-se exatamente no meio de uma serie temporal ordenada em ordem crescente ou decrescente. Se houver dois valores no meio da lista ordenada (quando a quantidade é par), a mediana é a média destes dois valores (RATANAMAHATANA et al., 2009). Outra medida bastante usada é *variância*, que mensura a variação da série temporal em comparação com a média. Ela é definida por

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n - 1}. \tag{3.2}$$

3.4 Métodos de agrupamento

As duas maiores áreas ativas na mineração de dados são a Classificação e Agrupamento de dados (ANTUNES; OLIVEIRA, 2001). As diferenças destas duas área está no fato que a primeira procura alocar cada registro dos dados em grupos já definidos anteriormente, enquanto a segunda tem o desafio de agrupar os dados segundo sua similaridade ou distância (MARSLAND, 2014). Apesar de muitos métodos serem desenvolvidos em ambas áreas, o foco desta pesquisa estará concentrado na segunda área (agrupamento), desde que o intuito é encontrar padrões que não pertencem a alguma categoria. Desse modo, nas próximas linhas serão explanados brevemente os principais métodos de agrupamento (para dados temporais) que são encontrado na literatura. Antes disso, a seguir são mencionado alguns dos principais desafios na tarefa de agrupar dados, segundo Mitsa (2010):

- Apesar de existirem muitas abordagens para agrupar dados temporais, por muitas vezes não há uma forma interpretar cada agrupação gerada pelo algoritmo. Nesses casos o conhecimento de um especialista é necessário.

- Também, é necessário tratar as anomalias e dados errados que geram ruído e distorcem os resultados. Mesmo que algum algoritmo consiga agrupar corretamente, no caso de grupos grandes, algumas das anomalias acabam se inserindo em alguma agrupação.

- Ao longo do tempo, é possível que agrupações encontradas previamente, sejam extintas num tempo posterior, ou os elementos trocarem de grupo ou misturarem entre si.

- Em muitos casos, os atributos que permitem agrupar em diferentes grupos é desconhecido, ou seja, não há dados que indiquem quais atributos devem ser considerados para fazer a diferenciação entre os grupos.

Nesse sentido, segundo Steen (2010), os métodos de agrupamento foram desenvolvidos para serem usados em quaisquer tipos de dados, desde que possam ser transformados em ou possuam valores numéricos. Portanto, estes métodos também são aplicáveis a dados temporais.

Há uma ampla variedade destes métodos (de agrupamento) na literatura, que se dividem em três categorias principais de algoritmos:

3.4.1 Por partição

O objetivo nesta abordagem é separar os dados em K grupos, sendo que a quantidade grupos é definido por um especialista. O principal critério para agrupar cada registro de dados é baseado na "proximidade"entre eles. Pelo seu baixo custo computacional e facilidade de implementação, esta abordagem tornou-se popular. Porém, ela também possui algumas desvantagens importantes como a sensibilidade a ruído e anomalias, a limitação de pré-definir a quantidade de grupos (MITSA, 2010).

Alguns dos principais algoritmos que se encontram nesta categoria são: *K-Means, Nearest Neighbor Algorithm, Minimum Spanning Tree, Squared Error Clustering Algorithm*, entre outros. O algoritmo mais popular deste grupo é o *K-means* (GAN et al., 2007). De forma simplificada, este é o processo que segue para agrupar:

a) É definido a quantidade total N de elementos e a quantidade K de grupos.

b) Aleatoriamente, são escolhidos K elementos para serem os centroides iniciais dos grupos.

c) É calculado a distância entre os elementos e cada centroide. Cada elemento é alocado no grupo que têm a menor distância.

d) Calcula-se novamente o centroide a partir dos elementos de cada grupo seguindo a seguinte fórmula:

$$\text{centroide}_m = \frac{\sum_{i=1}^{n} s_{mi}}{n}, \tag{3.3}$$

onde m é um grupo com n elementos s.

e) Repete-se o processo a partir do passo (c) até alcançar algum critério de convergência: quantidade de repetições ou iterações, minimização do erro quadrático, etc.

Uma variação deste processo, é o algoritmo *nearest-neighbor*, que ao invés de adicionar cada elemento ao grupo com o centroide mais próximo, no passo 3, ele adiciona ao grupo do elemento vizinho mais próximo (e não o centroide).

3.4.2 Por hierarquia

Nesta categoria, a abordagem ordena os elementos de um conjunto de dados seguindo uma hierarquia, que pode ser de "cima para abaixo"(*top-down* ou *divisive*) ou de "abaixo para cima"(*bottom-up* ou *agglomerative*) (MITSA, 2010). Segundo Gan et al. (2007), um diferencial nesta abordagem é que não é preciso conhecer a quantidade de agrupações previamente, porém, o principal custo está na sua complexidade computacional. Sobre a hierarquia *agglomerative*, há três variações principais que são encontradas na literatura (MITSA, 2010). Elas explanam sobre como as agrupações são construídass:

- Única distância: dois grupos são unidos se a distância mínima entre quaisquer dois elementos de cada grupo é menor ou igual que um limiar.
- Distância média: dois grupos são unidos se a distância média entre quaisquer dois elementos de cada grupo é menor ou igual que um limiar.
- Distância completa: neste modo de unir dois grupos, a distância máxima precisa ser menor ou igual que um limiar.

Por outro lado, na hierarquia *divisive*, o processo começa considerando todo o conjunto de dados como uma agrupação única, e progressivamente, separando ou dividindo em grupos menores, segundo algum critério definido previamente (MITSA, 2010).

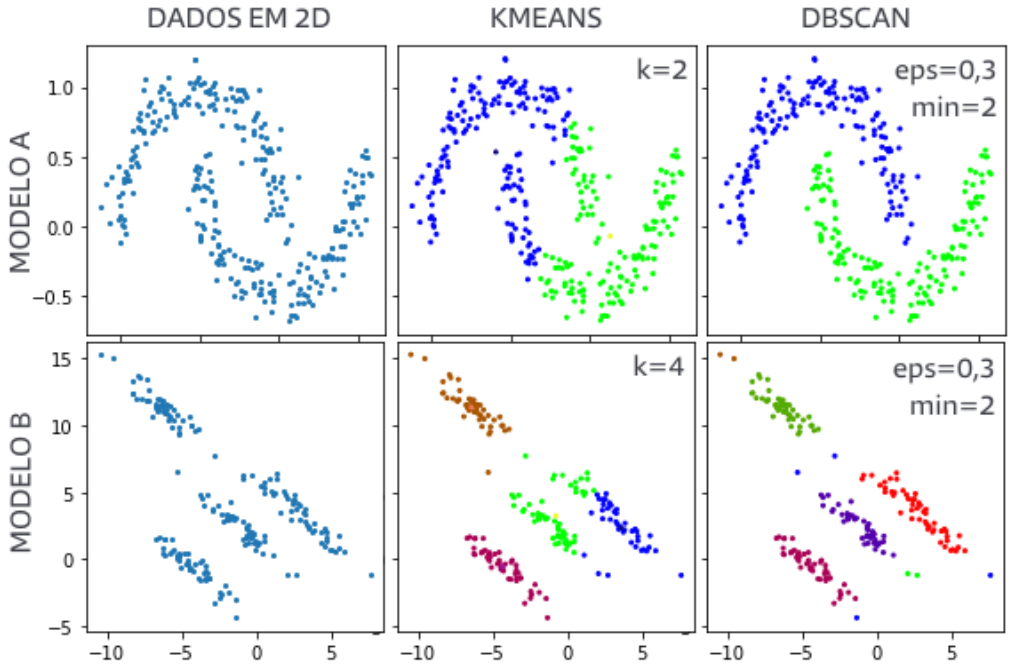
3.4.3 Por densidade

A principal vantagem nesta categoria, quando comparado com as anteriores, é a flexibilidade para encontrar formatos diversos de agrupações (MITSA, 2010). O algoritmos mais conhecido é provavelmente o *Density-based spatial clustering of applicacions whith noise* (DBSCAN), introduzido por Ester et al. (1996), que cria as agrupações com a maior densidade possível. Neste algoritmo, há dois parâmetros que precisam ser estabelecidos previamente pelo usuário: a) raio ou distância ϵ para considerar outros elemento como vizinhos e b) a quantidade mínima n_{min} de pontos para formar um região densa.

Além disso, o algoritmo verifica, para cada elemento, se a quantidade de elementos dentro da vizinhança ϵ é acima de n_{min}. Em caso afirmativo, considera o elemento avaliado como o centro de um novo grupo. Desse modo, o processo continua interativamente e, em alguns casos, pode unir dois grupos em um apenas (quando dois grupos estão muito próximos). Termina-se o processo de agrupamento quando nenhum elemento pode ser adicionado a algum grupo. Desse modo, elementos que não se encontram em algum agrupamento são considerados como ruído (ESTER et al., 1996).

Para ilustrar as diferenças nos resultados ao usar algumas das abordagens mencionadas, na Figura 3.5 mostra-se um comparativo visual do agrupamento de diversos conjuntos de dados em duas dimensões.

Figura 3.5 - Agrupamentos usando KMEANS e DBSCAN sobre dois conjuntos de dados



De modo geral, o algoritmo do KMEANS não consegue separar corretamente alguns grupos de dados devido à proximidade entre eles, mesmo colocando previamente a quantidade k correta de grupos. Por outro lado, o DBSCAN consegue capturar as agrupações e diferenciar os *outliers*. No entanto, a desvantagem desta abordagem é que é preciso calibrar o ϵ (*eps*) e o n_{min} (*min*) para encontrar as agrupações.

Fonte: Produção do autor.

3.5 O tempo e espaço nos métodos de agrupamento de eventos

Quando adicionamos a dimensão espacial aos dados temporais, também estamos adicionando maior complexidade, desde que a relação espaço-tempo têm diversos desafios dentro da mineração de dados (ATLURI et al., 2018).

Nesse sentido, com o crescimento da coleta de dados de forma massiva a partir de tecnologias com geolocalização (ex.: dispositivos móveis, relógios inteligentes, etc) aparece também o desafio de obter informações sobre padrões no espaço e tempo, de forma automática e em tempo real. Com esse fim, alguns métodos específicos para esta combinação (espaço-tempo) de dados foram desenvolvidos nos últimos anos.

Em síntese, os dados espaço-temporais podem ser divididos em cinco categorias, segundo Shi e Pun-Cheng (2019), Kisilevich et al. (2009): eventos, variáveis geo-referenciadas, series temporais geo-referenciadas, pontos em movimento e trajetórias. A seguir, uma breve explanação e respectivo exemplo de cada tipo de dado espacial e temporal:

- *Evento*: refere-se a elementos fixos no espaço e tempo. Por exemplo, a ligação telefônica atendida por uma pessoa (que tem local, data e hora).

- *Variável geo-referenciada*: são dados que estão fixados no espaço, porem que são atualizados ao longo do tempo. Por exemplo, a quantidade de pessoas trabalhando nos prédios da cidade. Cada prédio tem localização fixa, porém a quantidade de servidores vai sendo atualizada (não acumulada numa lista) ao longo do dia.

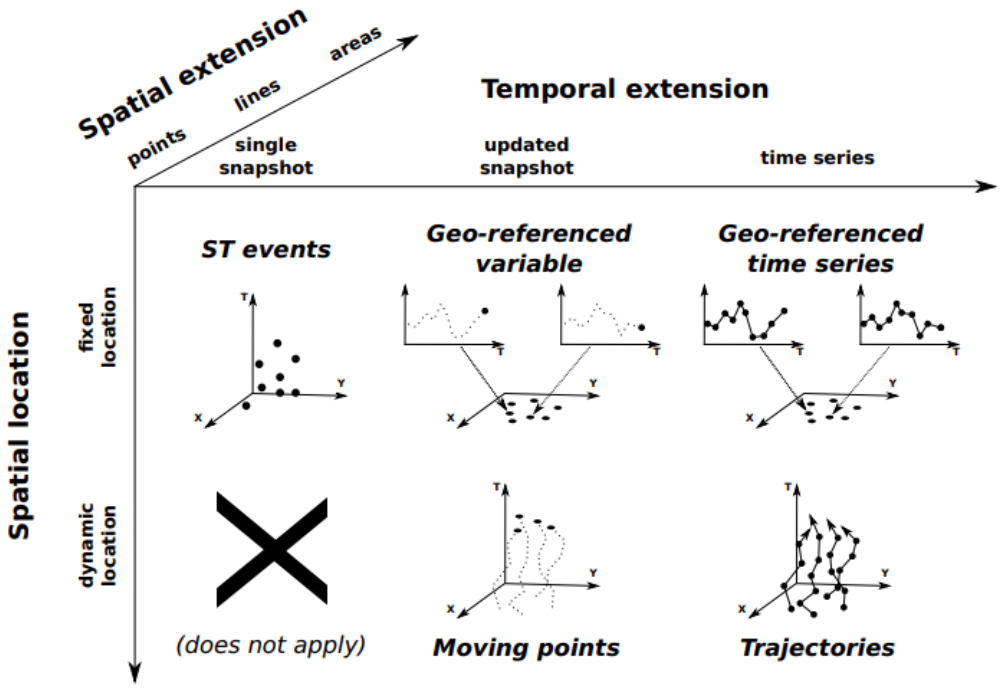
- *Series temporais geo-referenciadas*: correspondem a dados que salvam as mudanças ao longo do tempo, e que se encontram em locais fixos. Por exemplo, a variação ao longo do ano da quantidade de chuva detectada pelas estações meteorológicas da cidade. Neste caso, as estações são fixas, mas o histórico do volume é preservado. A diferença com a variável geo-referenciada encontra-se no de ela ser apenas uma captura dos valores no tempo, enquanto a serie temporal geo-referenciada corresponde a uma série de capturas dos valores.

- *Pontos em movimento*: refere-se aos dados que possuem localização dinâmica no espaço e tempo. Desse modo, sua localização é atualizada ao longo do tempo. Por exemplo, a posição atual dos satélites.

- *Trajetórias*: são dados que preservam as mudanças temporais e espaciais. Por exemplo: O percurso de um navio ao longo da viagem entre dois portos. A cada certo intervalo de tempo é possível estabelecer uma posição espacial, que é armazenada, formando uma série de pontos, que é a trajetória e ao mesmo tempo uma série temporal.

Na Figura 3.6 são mostrados as categorias a partir da combinação das dimensões espacial e temporal.

Figura 3.6 - Mapa dos tipos de dados espaciais e temporais.



Ao todo, cinco conjuntos possíveis de dados espaciais e temporais, a partir da combinação espaço-temporal.

Fonte: Kisilevich et al. (2009)

Na literatura, há dois algoritmos de agrupamento de eventos espaço-temporais que se destacam pela sua eficiência ao considerar ambas as dimensões dentro do mecanismo de agrupamento: *ST-GRID* e *ST-DBSCAN*, desenvolvido por Wang et al. (2006). Em ambos, as iniciais *ST* referem-se às dimensões espacial (latitude e longitude) e temporal. A dimensão espacial descreve a posição dos elementos, sejam fixos (ex.: a posição de uma torre de energia) ou dinâmicos (ex.: pontos do deslocamento de um carro). A dimensão temporal descreve o comprimento da evolução dos elementos, sejam em intervalos regulares ou capturas em momentos específicos (ANSARI et al., 2019; DUNHAM, 2002).

Basicamente, o conceito do *ST-GRID* é baseado na criação de celas (em três dimensões) em formato de cubos para toda a região dos dados, e desse modo, unir os cubos em grupos segundo a densidade de elementos em cada cubo e seus vizinhos. Para isso, é feito a contagem de elementos em cada cubo e seus vizinhos.

Quando a quantidade estiver acima de um limiar, os cubos são unidos para formar um grupo. Desse modo, o processo se repete até não ter mais cubos para unir a outras (WANG et al., 2006).

O *ST-DBSCAN* é um algoritmo baseado no *DBSCAN* com a dimensão temporal acrescentada, e, portanto, adiciona-se um raio ϵ_t para a nova dimensão. O principal diferença entre o ST-GRID e ST-DBSCAN está na forma que é feita a comparação com a vizinhança. O primeiro (ST-GRID), mensura a proximidade entre cubos, enquanto o segundo faz a comparação entre todos os pontos, desse modo, ele é mais eficiente que o segundo. No entanto, a desvantagem do ST-GRID encontra-se no fato de precisar armazenar a grade e seus respectivos pontos (em cada cubo). Apesar destas diferenças, em termos de qualidade dos resultados, ambos apresentam similar desempenho (WANG et al., 2006).

4 DADOS TEMPORAIS E REDES

Gerar redes complexas a partir de dados com atributos temporais é um processo que pode ser realizado de variadas formas. As vantagens e desvantagens de cada abordagem depende do tipo de dado temporal (evento temporal, serie temporal ou sequencia temporal), além disso, da resolução de tempo que deseja-se analisar e do tipo de informação que é procurado (HAN et al., 2006; ZANIN et al., 2016; HENDERSON et al., 2012). Por exemplo, se temos um banco de dados de conversas (mensagens) entre um grupo de amigos que contém os dados como: emissor, destinatário e a localização geográfica de cada pessoa, podemos construir diversas estruturas de redes. Uma abordagem possível pode estar baseada na troca de mensagens do último mês, onde os nós são as pessoas e as ligações as mensagens trocadas entre eles. Esta estrutura permitiria analisar os relacionamentos de um indivíduo e identificar os mais participativos. Além disso, seguindo outra abordagem, uma nova rede pode ser construída a partir da proximidade geográfica entre os membros do grupo. Desse modo, as pessoas são representadas como nós e a proximidade geográfica (definido por um limiar) como ligações entre dois nós. Esta última rede permitiria conhecer o grupo de indivíduos com melhor localização (em termos geográficos) em relação ao restante do grupo. Também, a partir das duas estruturas de redes seria possível conhecer os indivíduos que melhor interagem em termos sociais e geográficos, e, desse modo, obtermos uma rede multinível. Desse modo, a estrutura da rede pode ser diferente, dependendo do momento que ela foi construída.

4.1 Pontos a considerar

Antes de detalhar as abordagens mais conhecidas na literatura para transformar dados temporais em redes complexas, é preciso compreender as variáveis que mais influenciam neste processo:

- **Regras de ligação**: em alguns sistemas complexos é muito mais fácil de identificar os atributos de conjuntos de dados que serão representados como nós e ligações, desde que possuam essa característica de forma natural (ex: uma rede social ou uma rede de transmissão elétrica). Porém, definir o quê serão os nós e as ligações, na maioria dos conjuntos de dados, é uma das tarefas que ainda desperta dificuldade devido ao amplo leque de opções disponíveis (KUNEGIS, 2014).

  Para representar os nós, usa-se principalmente atributos que agrupam

um conjunto de dados ou que se encaixam numa categoria ou intervalo. Para representar as ligações, usa-se regras que ao serem válidas, permite construir uma ligação entre dois nós (SILVA; ZHAO, 2016). Nesse sentido, é possível ter diversas regras de ligações e, portanto, diversas estruturas de redes a partir do mesmo conjunto de dados. Por exemplo, se considerarmos os dados de uma partida de futebol é intuitivo imaginar os jogadores como os nós da rede. Porém, a construção das ligações podem ser diversas: cada ligação seria [A] o passe da bola entre quaisquer jogadores, [B] o passe da bola entre jogadores de um mesmo time, [C] o passe de bola que seja curto (menor que um limiar) [D] passe de bola aérea (acima de uma altura).

- **Janela de tempo**: conjuntos de dados do mundo real contém o tempo como atributo, na maioria dos casos. Em tal contexto, uma janela de tempo é o intervalo de tempo que será usado para construir a rede, a partir dos dados temporais. Desse modo, a partir do mesmo conjunto de dados, podemos construir diversas redes considerando intervalos de tempos diferentes. Estes intervalos podem ser regulares (intervalos constantes) ou irregulares (períodos de tempo diversos) (MASUDA; LAMBIOTTE, 2016; HOLME; SARAMÄKI, 2012).

  O intervalo regular é, com frequência, atrelado a períodos de tempo conhecidos ou determinados de forma natural (HOLME, 2015). Alguns exemplos são: as estações do ano, o período mensal, semestral, anual, mensal, semanal, diário, etc. Nesses casos, cada intervalo teria sua própria rede. Para intervalos irregulares, as janelas de tempo estão relacionadas a regras ou gatilhos para dar início e fim (HOLME, 2015; BATAGELJ; PRAPROTNIK, 2016). Por exemplo, os dias que houve um conflito social, as horas de uma operação cirúrgica ou os minutos de uma conversação entre colegas.

  A vantagem no uso de janelas de tempo está na facilidade para construir redes apenas sobre o período de interesse. Também, para analisar a evolução das redes ao longo do tempo (várias redes).

- **Direção das ligações**: em muitos conjuntos de dados é possível estabelecer de forma fácil uma relação entre origem e destino. Nesses casos, as ligações são direcionadas, por outro lado, em casos onde não é possível estabelecer este tipo de relação e as ligações não serão direcionadas (HOLME, 2015; STEEN, 2010). Por exemplo, os dados sobre voos de avião,

transporte, ou envio de e-mails podem ser facilmente transformados em redes direcionadas, desde que haja uma origem e um destino.

A direção das ligações também possui um papel importante na análise de fenômenos em redes, como a propagação de informação, doenças, etc. Nesses fenômenos, as direções nas ligações são de extrema importância para alcançar resultados válidos (HOLME, 2016)

- **Dimensão espacial**: este atributo tem um papel importante em qualquer conjunto de dados reais que pertencem a sistemas complexos, uma vez que muitos dos elementos destes sistemas possuem uma localização que, frequentemente, muda ao longo do tempo (SHI; PUN-CHENG, 2019). Nesse sentido, há dois grupos que abordam os métodos para representar esta dimensão nas redes complexas. O primeiro grupo é usado em dados de series temporais que estão atreladas a uma localização física fixa, como por exemplo, a velocidade do vento ao longo do dia mensurada por um grupo de barômetros (FENG; HE, 2017; DONGES et al., 2009; HLINKA et al., 2017). Nestes casos, seria possível definir regras e construir uma rede de barômetros a partir da similaridade (das séries) e a distância (espacial) entre eles. A abordagem do segundo grupo é baseada no uso da distância entre nós como um atributo (peso) nas ligações das redes (SILVA; ZHAO, 2016; MARWAN et al., 2009; JEBARA et al., 2009). Por exemplo, numa rede de passes de bola entre jogadores de futebol, a distância espacial entre os jogadores pode ser representada no peso das ligações, e, desse modo, fazer análises sobre o desempenho dos passes que cada jogador teve com seus colegas ao longo do jogo.

4.2 De dados temporais a redes

A transformação de dados temporais em redes temporais muda de abordagem dependendo do tipo de dado temporal (eventos, series, ou sequências) (SILVA; ZHAO, 2016). Devido à utilidade que o estudo de séries temporais tem para revelar padrões, identificar anomalias e fazer previsões (além de outras aplicações), na literatura são encontrados maiores avanços sobre a transformação em redes complexas, pois a partir delas seria possível obter medidas que ajudam a revelar informações sobre os sistemas complexos (AGHABOZORGI et al., 2015; ZHOU et al., 2015).

Como mencionado anteriormente, além das séries temporais também temos os

eventos temporais, que são caracterizados por terem atividade num momento específico (DUNHAM, 2002). Também é conhecido a possibilidade de transformar eventos em séries (temporais), e, portanto, obtermos redes complexas a partir dos eventos mesmo que indiretamente (DONGES et al., 2009).

A seguir, são apresentados os principais métodos que foram desenvolvidos especificamente para construir redes a partir de séries, sequências e eventos temporais.

4.2.1 Redes de series pseudo-periódicas

Zhang e Small (2006) desenvolveram um método para estudar séries temporais com base nos ciclos ou padrões que se repetem numa série temporal. Neste trabalho, os ciclos são transformados em nós e a ligação entre eles é construída se o coeficiente de correlação linear é superior a um limiar pré-estabelecido. Neste método, cada ciclo tem começo e fim nos mínimos locais da série temporal, desse modo, o que vemos como resultado final é uma rede de ciclos que são similares entre eles.

Segundo Zou et al. (2019), por ser um método que precisa da presença de padrões periódicos na série temporal, este fica restrito ou limitado ao estudo de séries temporais com essa característica. Também, ao depender de um limiar pré-estabelecido para construir as ligações, a rede final corre o risco de ser muito esparsa ou densa, e, assim, dificultar a análise correta da dinâmica da série temporal.

4.2.2 Redes de visibilidade

Desenvolvido por Lacasa et al. (2008), esta abordagem traz uma nova forma de olhar os dados de séries temporais para transformá-los em redes. O método é baseado no conceito de *visibilidade*, que, basicamente, consiste na conexão entre pontos da série temporal a partir da linha reta de visibilidade que existe entre eles sem que outro ponto intermédio corte a linha de visão.

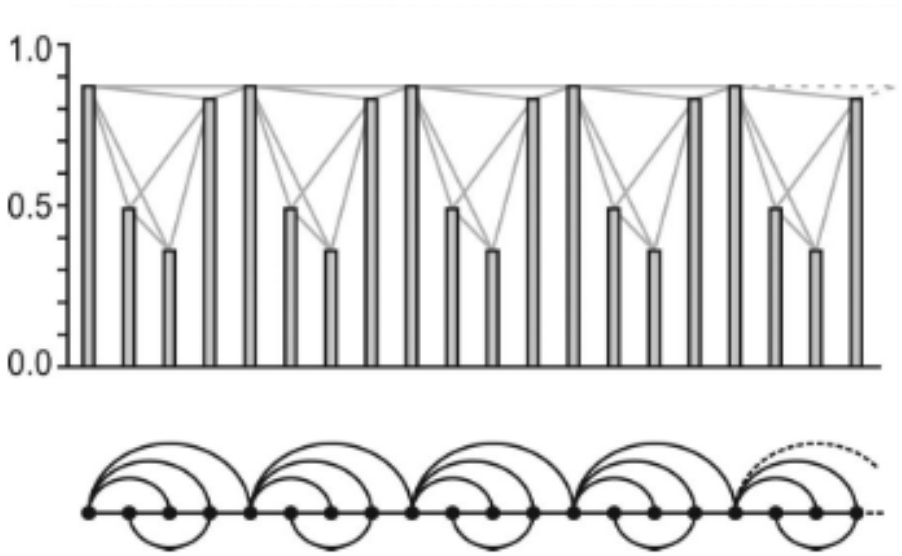
A visibilidade entre dois pontos (t_a, y_a) e (t_b, y_b) é criada se qualquer outro ponto (t_c, y_c), entre estes dois, cumprir a seguinte condição:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}. \tag{4.1}$$

Por exemplo, na Figura 4.1 mostra-se um exemplo ilustrativo de uma série e as

linhas de visibilidade, que posteriormente são transformadas em ligações.

Figura 4.1 - Rede de visibilidade



Exemplo de série temporal transformada em rede usando o método baseado na visibilidade.

Fonte: Lacasa et al. (2008).

As vantagens que possui este método são a flexibilidade para ser utilizado em qualquer série temporal e a garantia de formar uma rede conectada (um único componente) (ZOU et al., 2019). Também, ao não depender de um variável pré-estabelecida é gerado uma única estrutura de rede a partir de uma série temporal. Nesse sentido, diversas aplicações e desdobramentos deste método foram desenvolvidos para identificar dinâmicas ou padrões que existem nas séries temporais.

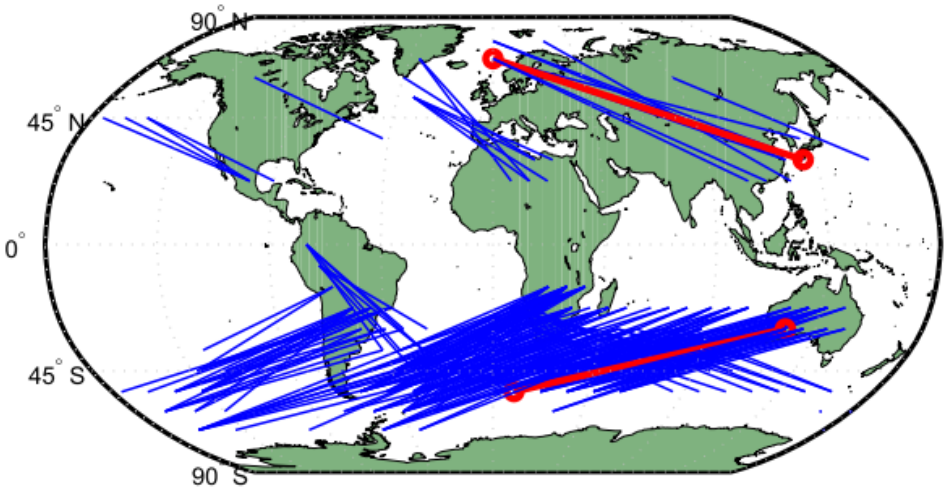
4.2.3 Redes de correlação

Esta abordagem, segundo Zou et al. (2019), é uma generalização das *redes de séries pseudo-periódicas*. Aqui também é usada a correlação para verificar se existe um conjunto de séries temporais que são similares no longo prazo, por exemplo, a velocidade do vento que marcam os anemômetros num parque eólico. Na literatura, uma área que tem adotado esta abordagem (ou suas variantes) em diversas aplicações é o estudo do *clima* e as relações entre diversos locais do planeta Terra

(ZHOU et al., 2015; YAMASAKI et al., 2008; TSONIS; SWANSON, 2008; RADEBACH et al., 2013; MENG et al., 2018; LUDESCHER et al., 2014).

A forma de construção deste método é simples. Primeiro, as séries temporais são consideradas como nós, estes, por sua vez, são ligados considerando o coeficiente de correlação entre eles. Desse modo, a ligação é determinada por um limiar fixo (acima de 0.5, por exemplo) ou uma porcentagem dos coeficientes (os 5% mais altos, por exemplo). Este método tem variantes, tanto na regra (limiar ou porcentagem) para gerar as ligações quanto na forma de mensurar a correlação (ZOU et al., 2019). Na Figura 4.2 mostra-se uma rede de correlação construída a partir dos dados da temperatura do oceano Pacífico numa região onde o fenômeno de El Niño está em formação.

Figura 4.2 - Rede de correlação



Exemplo de rede construída a partir do método baseado na correlação, e usando dados da temperatura sobre a superfície. Na imagem mostra-se 226 locais que estão relacionados direta ou indiretamente (segundo a rede) e com distância acima de 5.000 quilômetros. As linhas em azul representam as ligações mais relevantes da rede, considerando a correlação que existem entre estes pontos. As ligações em vermelho são apenas dois casos usados nessa pesquisa para explicar a correlação.

Fonte: Zhou et al. (2015).

Não há, até o momento, um consenso sobre qual variante deste método (dentro das opções que oferece esta abordagem) é o mais apropriado para o estudo de

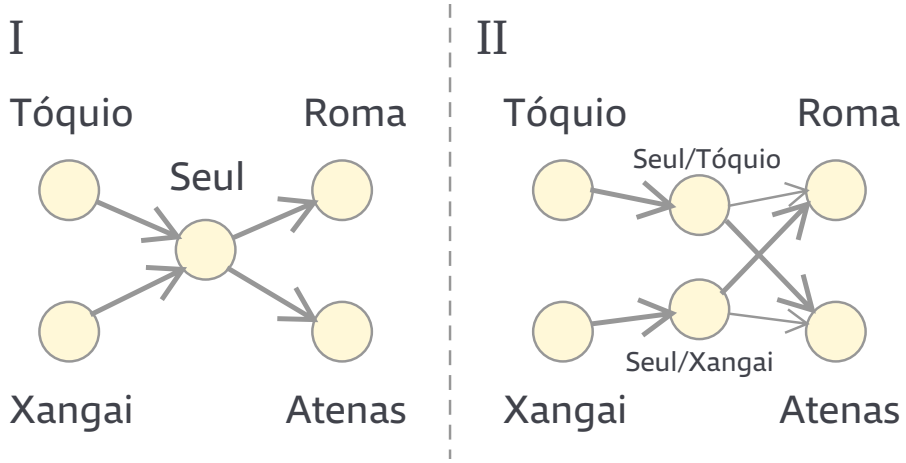
conjuntos de séries temporais, desde que uma pequena mudança nas variáveis pré-estabelecidas ou na regra de correlação pode gerar redes diferentes, e, portanto, análises diferentes. Além disso, a maior parte dos estudos foram aplicados em áreas ou problemas específicos, e, logo, não necessariamente replicáveis em outras áreas.

4.2.4 Redes de ordem superior

Este método é um dos mais recentes na literatura e desenvolvido por diversos grupos de pesquisa, em especial, Xu et al. (2016) e Benson et al. (2016). A diferença frente aos métodos mencionados anteriormente se destaca por utilizar dados que sejam sequências temporais (por exemplo, dados de transporte marítimo contendo origem, destino e hora de partida/chegada), e não séries temporais. Desse modo, sua relevância encontra-se no fato de poder capturar interações (dependências entre nós) não visíveis diretamente, a partir do mapeamento das interações anteriores.

Inicialmente, esse processo constrói uma rede de *primeira ordem*, que seria uma rede simples e direcionada. No último exemplo sobre rede de transporte marítimo, os nós seriam os locais de partida e chegada, e as ligações direcionadas seriam o sentido das viagens realizadas pelos navios entre os locais. Na etapa seguinte, a partir da rede de primeira ordem são gerados novos nós para representar dependências chaves entre origem e destino, e, desse modo, obter uma nova rede considerada como rede de segunda ordem. Assim, este processo pode ser utilizado de forma iterativa para obter redes de ordem superior. Na Figura 4.3, observa-se um exemplo visual da estrutura da rede em primeira e segunda ordem.

Figura 4.3 - Rede de ordem superior



Exemplo de rede de primeira (I) e segunda ordem (II). Na primeira imagem, mostra-se uma rede direcionada com Seul como nó chave no fluxo das ligações. Desse modo, na segunda imagem ele é decomposto em dois nós, Seul a partir de Tóquio e Seul a partir de Xangai.

Fonte: Adaptado de Xu et al. (2016).

O intuito das redes de ordem superior é similar ao da mineração de dados (na detecção de padrões, previsão de movimentos), porém, devido à possibilidade de usar redes, adiciona-se mais opções de estudo sobre dados temporais sequenciais (LAMBIOTTE et al., 2019).

4.2.5 Outras abordagens

Na literatura são encontradas outras abordagens para explorar dados temporais através das redes complexas que são, na maioria dos casos, variantes ou similares às abordagens apresentadas brevemente nesta seção. Nesse sentido, destacam-se alguns métodos baseados em redes e que são construídas sobre dados temporais. Estes trabalhos são de Marwan et al. (2009), pelas redes de recorrência; Campanharo et al. (2011) pelo modelo dual que permite transformar séries temporais em redes, e vice-versa. Também, pelo modelo baseado nas distâncias espaciais entre nós desenvolvido por Silva e Zhao (2016).

4.3 Oportunidades e limitações

É impossível um único método suprir ou solucionar todas as demandas relacionadas à exploração dos dados temporais. Por isso, os métodos que são desenvolvidos na literatura procuram atender problemas ou limitações específicos. Estes métodos, ao trabalhar em conjunto de modo complementar, conseguem auxiliar na descoberta de informações (ZOU et al., 2019). Assim, o propósito desta subseção é destacar algumas das principais limitações encontradas nos métodos apresentadas na subseção anterior.

Recapitulando, há principalmente três tipos de dados temporais: séries temporais, sequências temporais e eventos temporais. E como visto anteriormente, os métodos mencionados nesta seção são apropriados para dados de séries e sequências temporais, faltando desse modo, métodos para explorar eventos temporais. Apesar de ser possível transformar eventos temporais (furtos na cidade, por exemplo) em séries ou sequências temporais (quantidade de furtos diários ao longo do ano, por exemplo), nessa transição, perde-se detalhes sobre os dados, pois ao agrupá-los em séries a informação perde nível de resolução. Por exemplo, se olharmos nos dados sobre furtos que acontecem na cidade (contendo o local e a hora do evento), é possível transformá-la numa série temporal se contarmos os furtos diários que acontecem ao longo de um mês; cada ponto da série estará correspondido com cada dia do mês. Também, podemos gerar diversas séries se agruparmos a contagem dos furtos para cada bairro da cidade, desse modo, teremos uma série temporal que corresponde a cada bairro. Considerando dessa forma, a rede final, se usarmos a abordagem da correlação neste último caso, não permitiria analisar os furtos no nível semanal ou diário, resultando portanto, em limitação para explorar informações com maior nível de detalhe.

Além disso, outro desafio encontrado é que, por muitas vezes, a quantidade de pontos necessária para formar séries temporais é insuficiente, pois estas devem ter tamanho suficiente para realizar cálculos como a correlação. Desse modo, quando há poucos eventos temporais, a formação de séries temporais é muito limitado, e consequentemente, não é possível construir as rede. No exemplo anterior sobre os furtos na cidade, é possível que alguns bairros concentrem poucos furtos, e, portanto, as séries temporais que correspondem a estes bairros sejam inapropriados para formar redes. Desse modo, muitos locais não serão representadas na rede gerada, resultando em perda de informação.

Além da transformação de dados em redes, é preciso conhecer quais algoritmos

aplicar sobre elas, segundo as informações que estas podem revelar (KUNEGIS, 2014). Apesar da grande quantidade de medidas desenvolvidas na literatura acadêmica, ainda é difícil estabelecer uma regra de interpretação correta para muitas delas uma vez que a sua compreensão está atrelada à natureza dos dados originais (HOLME, 2015; MASUDA; LAMBIOTTE, 2016; YU et al., 2016). No caso de dados temporais a interpretação é mais evidente, considerando que é o tempo uma das peças fundamentais na construção da rede, portanto, é possível desenvolver medidas e interpretações específicas para a mineração de dados temporais.

Finalmente, considerando as diversas possibilidades que as redes complexas oferecem, é evidente que ainda existe espaço para explorar os dados com pouca atenção, como o são as sequências e eventos temporais.

5 A ABORDAGEM CRONOLÓGICA: MÉTODOS PARA DETECÇÃO DE PADRÕES EM EVENTOS ESPACIAIS E TEMPORAIS

Após a revisão dos métodos tradicionais de detecção de agrupamentos em dados temporais, também foram revisados os métodos de redes complexas que permitem transformar dados temporais em redes. Ao longo desse estudo, não foram encontrados métodos ou algoritmos que sejam específicos para a detecção de padrões em conjuntos de eventos temporais e espaciais usando as redes complexas como forma de explorar essas informações. Em contrapartida, diversos métodos foram desenvolvidos para a análise de series temporais a partir das redes complexas. Tendo em consideração esse contexto, neste capítulo é desenvolvido uma abordagem, baseada em redes complexas, para explorar eventos temporais e espaciais, para assim identificar os padrões que acontecem nesse tipo de dados.

A abordagem principal é nomeada como *cronológica*, por ter o tempo como base para a construção da rede. A partir desta abordagem, três métodos (ou variações) novos foram desenvolvidos, considerando seu uso em cenários diferentes. Em síntese, nesta abordagem a estrutura da rede é definida a partir dos valores espaciais e temporais dos eventos, depois, sobre essa rede, os métodos propostos são aplicados para obter padrões a partir da detecção de comunidades. Nesse sentido, o processo completo divide-se em duas partes: I) a abordagem cronológica, para construir a rede, II) e os métodos propostos, para modificar a rede e encontrar as mudanças nos dados.

A seguir, explana-se o funcionamento da abordagem, e as variações dela. Também, são mostrados diversos exemplos artificiais e a descrição das principais vantagens e desvantagens, assim como os métodos derivados desta. Além disso, no final é mostrado a aplicação do método em dados reais de queimadas na Bacia Amazônica.

5.1 A abordagem cronológica

Esta abordagem consiste na construção de uma rede a partir de eventos temporais e espaciais. Nesse sentido, a abordagem recebe como entrada os dados e entrega como saída uma rede complexa, ou seja, o trabalho da abordagem é transformar dados em rede.

Desse modo, a partir de um conjunto com n eventos $Z = z_1, z_2, z_3, ..., z_n$ que acontecem seguindo a ordem temporal $T = t_1, t_2, t_3, ..., t_n$, e que possuem localização

definida pelo conjunto $L = l_1, l_2, l_3, ..., l_n$, a geração da rede cronológica segue estes passos:

- A região espacial que contém os n eventos de Z é dividida por uma grade $g : \rho \times \omega$ formando q celas $c_1, c_2, \ldots c_q$, onde $\rho \times \omega = q$. Esta grade pode estar baseada em qualquer padrão (formato quadrado, hexagonal, entre outros), porém o recomendado é que todas as celas sejam iguais em tamanho para facilitar a divisão da região. Ao termos uma grade quadrada $g : \rho \times \rho$, recomenda-se que a quantidade total de celas seja próximo de $n/3$, para ter uma quantidade mínima de eventos em cada cela, ou seja, a quantidade de divisões em cada lado da grade seria $\rho \sim \sqrt{n/3}$.

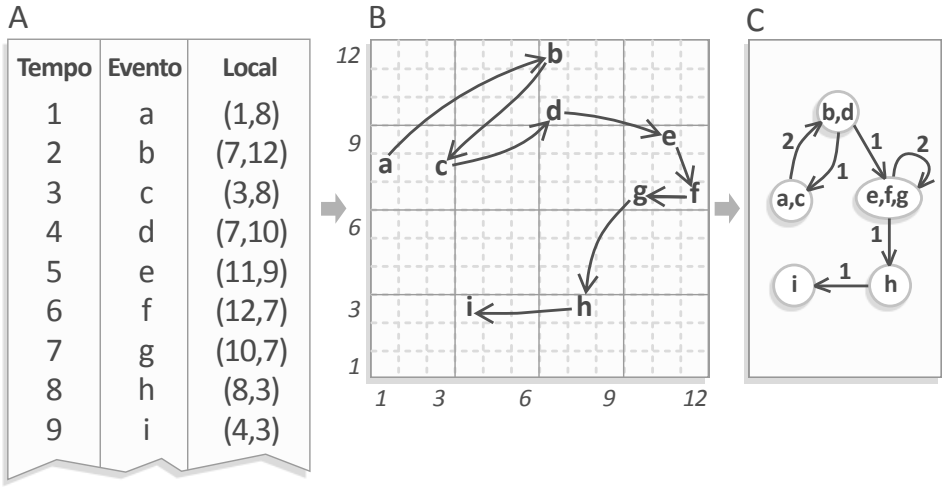
- Depois, para cada posição l_i, dos n eventos, é identificado a cela c_i em que o evento aconteceu. Desse modo, é possível que dois ou mais eventos localizados em pontos diferentes estejam na mesma cela.

- A seguir, cada cela c_i é transformado em nó n_i, e estes são ligados sequencialmente com o imediato próximo, considerando apenas uma ligação $\mu = 1$. A ordem sequencial é dada pela variável temporal em ordem crescente. Desse modo, a rede gerada será composta por no máximo q nós e $n - 1$ ligações.

- Finalmente, as ligações múltiplas entre dois nós são simplificadas em apenas uma ligação, e a ligação restante recebe o atributo *peso*, que representa a quantidade de ligações simplificadas. Quando dois eventos consecutivos acontecem na mesma cela é formado um *loop* no nó que corresponde a ambos eventos.

A Figura 5.1 mostra a ilustração da geração de uma rede cronológica a partir de uma lista de eventos com componente temporal e espacial.

Figura 5.1 - Construção da rede cronológica



A) Primeira etapa: Ordena-se a lista de eventos segundo o tempo. Também temos as informações sobre quando e onde aconteceu, e nome do evento. B) Segunda etapa: Divide-se a região em grade 4×4 para conectar as celas com ligações entre si, seguindo a ordem temporal. C) Terceira etapa: Rede formada e simplificada.

Fonte: Produção do autor.

É necessário destacar que a rede formada pode ser direcionada ou não, e se tiver múltiplas ligações entre dois nós, é possível simplificar através da adição do atributo *peso* (que corresponde à quantidade de ligações simplificadas) nas ligações.

5.2 Métodos de detecção de padrões de transição

Os métodos ou variações propostos nesta seção são baseados na rede construída a partir da abordagem cronológica, mostrada anteriormente. Desse modo, assumindo que a rede já existe, o funcionamento dos métodos propostos consiste em modificar a rede e aplicar algoritmos de detecção de comunidades para encontrar grupos de eventos que seguem um padrão no espaço e tempo. Estes grupos representam as transições que existe entre os eventos, ao analisar eles como um conjunto. Além disso, destaca-se que estes métodos, apesar de terem poucas mudanças entre si, o resultado delas pode ser diferente em muitos casos.

Neste contexto, a detecção de padrões de transição trata sobre encontrar câmbios espaciais e temporais significativos. Pois sabemos que eventos aparecem e desapa-

recem ao longo do tempo, e ao analisarmos estes de forma conjunta, em muitos casos há transições claras seja pela mudança de direção (por exemplo, a aparição de furtos em um novo bairro da cidade) ou também de intensidade (por exemplo, mais acidentes de carro ao longo de uma avenida). Desse modo, diversos padrões de transição podem acontecer sobre o mesmo grupo de eventos, ou seja, o intuito dos métodos não é encontrar, necessariamente, grupos de eventos, mas sim padrões de transição dentro desses grupos.

Considerando que as ligações da rede obtida correspondem a relações temporais próximas entre celas (mesmo que sejam distantes espacialmente), é possível formar de grupos de celas que possuem eventos muito próximos (temporalmente) entre si. Nesse sentido, a detecção desses grupos na rede é desenvolvido principalmente na área *comunidades*.

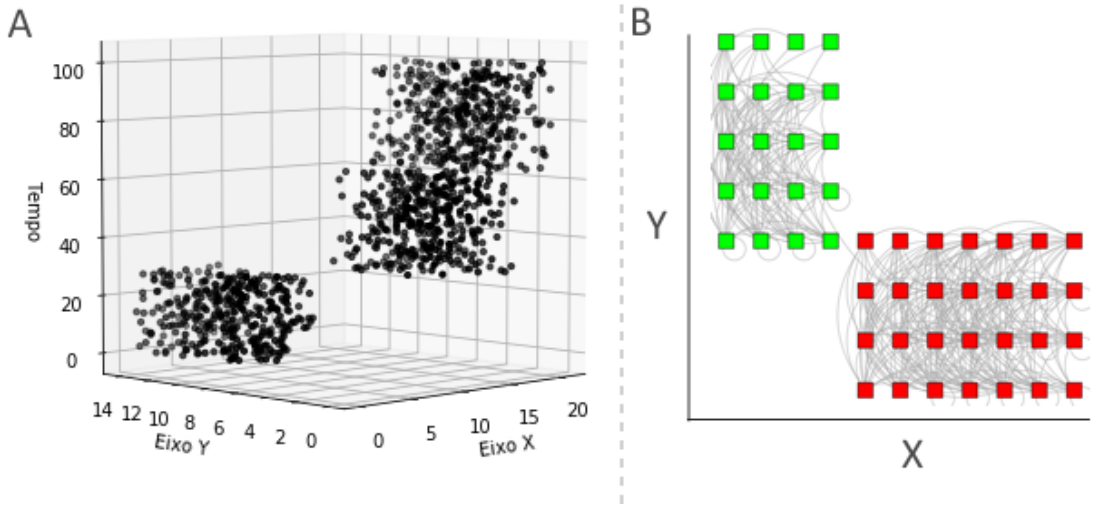
Sabendo disso, a seguir, detalham-se o procedimento *padrão* que é proposto para identificar as mudanças sobre a rede da abordagem cronológica:

- Aplicar o algoritmo (Infomap ou Louvain) de detecção de comunidades,

- Identificar as comunidades de celas q_i, que são representantes dos grupos de celas contendo padrões dos eventos ao longo do tempo. Desse modo, estes agrupamentos são os padrões ou estágios do conjunto de dados.

A Figura 5.2 ilustra a detecção dos grupos ou padrões num conjunto de eventos temporais após a construção da rede cronológica.

Figura 5.2 - Comunidades de celas



A) 1125 eventos espaciais e temporais gerados de forma aleatória em dois agrupamentos separados propositalmente ao longo do tempo e espaço. Para construir a rede, a região é dividida numa grade 12×12, que representa a quantidade de possíveis nós. B) Após construir a rede segundo o método proposto e usar a detecção de comunidades de celas (nós), são encontradas duas comunidades, que representam os dois grupos de eventos, separados pela transição ou deslocamento espacial que aconteceu entre ambos.

Fonte: Produção do autor.

5.3 As variações

O método padrão proposto é flexível na sua aplicação, por esse motivo nas seguintes linhas serão analisados algumas das principais variações e possíveis situações onde seu uso seria apropriado.

5.3.1 Método 1: Ligação única

Esta variação ou método corresponde ao procedimento padrão explanado anteriormente. O nome do método deve-se ao fato de não modificar a rede obtida a partir da abordagem cronológica. Desse modo, após construir a rede, procede-se com a identificação dos padrões ou transições dos eventos. A principal vantagem neste método encontra-se na simplicidade do método, pois apenas um parâmetro (tamanho da cela) será necessário em todo o processo. Destaca-se também que a complexidade da abordagem é $O(n)$, sendo n a quantidade de eventos, e sem considerar a complexidade do algoritmo de detecção de comunidades. É conhe-

cido que a complexidade do algoritmo de detecção de comunidades de Louvain ou Multilevel é aproximadamente $O(nlogn)$. No intuito de compreender melhor o desenvolvimento deste método, a seguir, mostra-se o pseudo-código.

Algoritmo 1: Método 1.

Entrada: lista de eventos, com local e tempo

Saída: grupos de nós

início

$eventos$: ordena a lista de eventos de forma crescente segundo o tempo;

n: quantidade de eventos;

divide a região numa grade quadrada com tamanho $\sim \sqrt{n/3}$ em cada lado;

repita

leia o evento α ;

identifique a cela que α pertence;

leia o evento $\alpha + 1$;

identifique a cela que $\alpha + 1$ pertence;

gere uma ligação entre as duas celas;

até *fim dos eventos*;

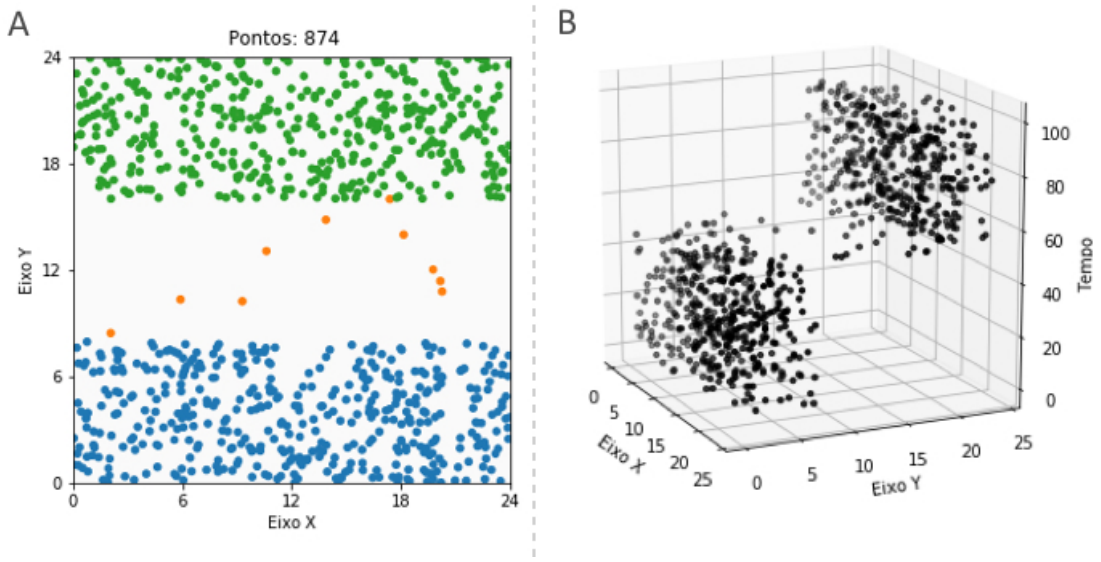
aplica o algoritmo de detecção de comunidades sobre a rede;

retorna os nós encontrados em cada comunidade;

fim

Para ilustrar o resultado deste método, na Figura 5.3A. é mostrado a geração artificial de eventos temporais e espaciais numa grade 24×24, dividido em três estágios claramente visíveis, sendo estes gerados de baixo pra cima, como é observado na Figura 5.3B. Propositalmente, o segundo estágio é composto por apenas 10 eventos, dentro do conjunto de 874 eventos gerados. Desse modo, o estágio intermediário pode ser considerado como ruido, desde que representa uma parte ínfima dos dados. O objetivo deste exemplo é ver o desempenho deste primeiro método e avaliar se conseguem identificar as transições entre padrões de eventos.

Figura 5.3 - Exemplo de eventos gerados

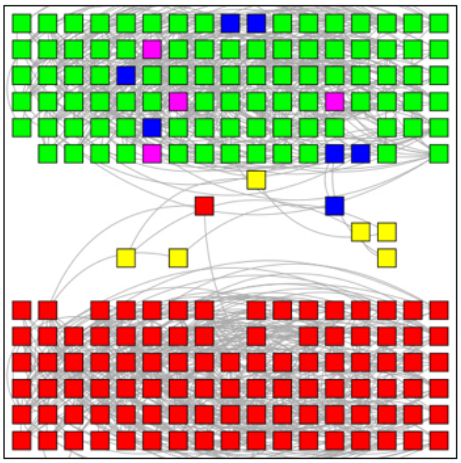


Eventos gerados artificialmente. A) Geração de três estágios usando 874 eventos espaciais e temporais. B) Os mesmos conjunto de eventos a partir de uma visão em três dimensões, considerando o tempo no eixo vertical.

Fonte: Produção do autor.

Seguindo com o procedimento estabelecido na abordagem cronológica, a região é dividida numa grade do tamanho 17×17, também, é utilizado uma ligação ($\mu = 1$) para cada par de eventos consecutivos, na construção da rede. Depois, ao aplicar a detecção de comunidades sobre a rede obtida, encontra-se diversos grupos de nós, como é visto na Figura 5.4. Ao comparar os resultados do Método 1 com os dados gerados na Figura 5.3, é preciso destacar dois pontos: primeiro, a identificação dos dois estágios principais no conjunto de eventos gerados é clara (celas de cor vermelha e verde), e em segundo lugar, o método tem dificuldade para agrupar os nós que representam o ruido (celas intermediárias).

Figura 5.4 - Método 1 aplicado sobre os dados gerados



Comunidades encontradas na rede de celas. Ao todo, cinco comunidades ou estágios identificados, sendo que dois deles correspondem aos dois principais conjuntos de eventos gerados (cor vermelha e verde). As comunidades restantes têm, principalmente, relação com o estágio intermediário ou ruido inserido no modelo de dados (cor amarela, rosa e azul).

Fonte: Produção do autor.

5.3.2 Método 2: Ligações simultâneas

Esta variação do método principal consiste em determinar uma quantidade fixa z de eventos subsequentes a cada evento, e assim, a partir de cada evento, construir uma ligação para cada uma das celas desses z eventos subsequentes. Desse modo, ao invés de ligar somente com o próximo imediato $\mu = 1$, nesta abordagem serão conectadas $\mu = z$ ligações a partir de cada evento. O ganho nesta variação é que, por permitir gerar mais ligações, a rede torna-se mais conectada e, portanto, permite formar comunidades mais coesas. Neste método, a complexidade da abordagem é $O(nm)$, sendo n a quantidade de eventos, e m a quantidade de ligações simultâneas, sem considerar a complexidade do algoritmo de detecção de comunidades.

No intuito de entender melhor o processo que segue este segundo método, nas seguintes linhas apresenta-se o respectivo pseudo-código. Note-se que há uma única

mudança em relação ao primeiro método, que está localizada na iteração.

Algoritmo 2: Método 2.

Entrada: lista de eventos, com local e tempo

Saída: grupos de nós

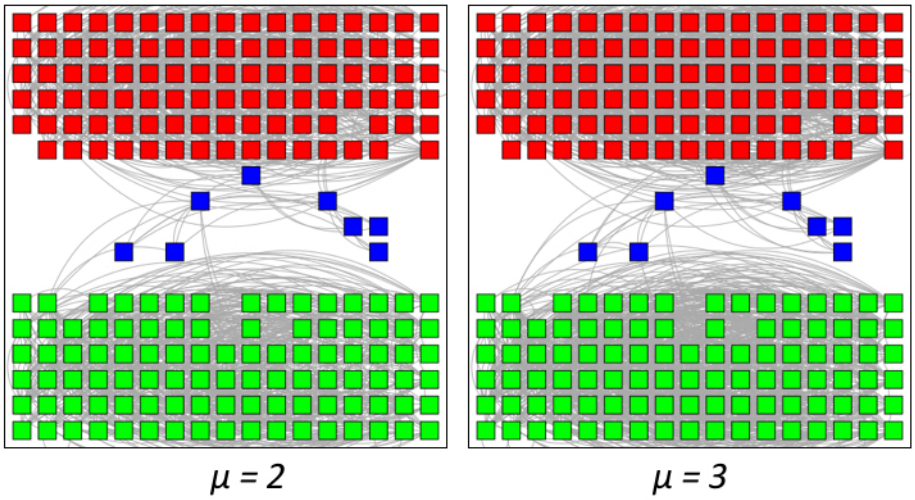
início

- $eventos$: ordena a lista de eventos de forma crescente segundo o tempo;
- n: quantidade de eventos;
- μ: quantidade de ligações simultâneas;
- divide a região numa grade quadrada com tamanho $\sim \sqrt{n/3}$ em cada lado;
- **repita**
  - leia o evento e ;
  - α: identifique a cela que e pertence;
  - **para** $i \leftarrow 1$ **até** μ **faça**
    - leia o evento $e + i$;
    - β: identifique a cela que $e + i$ pertence;
    - gere uma ligação entre α e β;
  - **fim**
- **até** *fim dos eventos*;
- aplica o algoritmo de detecção de comunidades sobre a rede;
- retorna os nós encontrados em cada comunidade;

fim

Desse modo, para ilustrar os resultados deste método, ao aplicá-lo sobre os dados gerados na Figura 5.3, observa-se na Figura 5.5 que tanto usando duas ou três ligações simultâneas, o método consegue diferenciar claramente os três estágios gerados no modelo original. Nesse sentido, no exemplo é evidente a principal vantagem do Método 2, que é a de fortalecer a conectividade da rede, resultando em comunidades ou estágios identificados com melhor definição. Além disso, outro ponto a destacar é sobre a similaridade dos resultados apesar de usar diferentes ligações simultâneas, portanto, mesmo que seja possível construir a rede com muitas ligações simultâneas, o ganho não é evidente.

Figura 5.5 - Método 2 aplicado sobre os dados gerados



Três comunidades em ambas as redes, construídas usando duas e três ligações respectivamente.

Fonte: Produção do autor.

5.3.3 Método 3: Ligações simultâneas e remoção de nós

Esta variação tem sua base na identificação de nós (ou celas) menos relevantes da rede, a partir da sua *força* (s ou *strength*). Portanto, este método torna-se útil quando há ruido no conjunto de dados, pois estes normalmente não acompanham o padrão do conjunto todo de eventos e portanto, é esperado que alguns destes encontrem-se isolados ou afastados dos grupos principais. Desse modo, ao remover os nós onde o suposto ruido se encontra, as comunidades (ou estágios) serão mais facilmente identificadas.

O processo de remoção consiste na identificação de nós com menor força (s), para isso é necessário saber quantas ligações simultâneas (μ) foram usadas na construção da rede, e duplicar este valor. Depois, os nós com força maior ou igual que este valor ($s > 2\mu$) serão mantidos, e o restante, removidos. Neste método, a complexidade da abordagem é $O(nm)$, sendo n a quantidade de eventos, e m a quantidade de ligações simultâneas e quantidade de de nós. Para melhor compreensão deste

processo, nas próximas linhas é mostrado o pseudo-código deste método.

Algoritmo 3: Método 3.

Entrada: lista de eventos, com local e tempo

Saída: grupos de nós

início

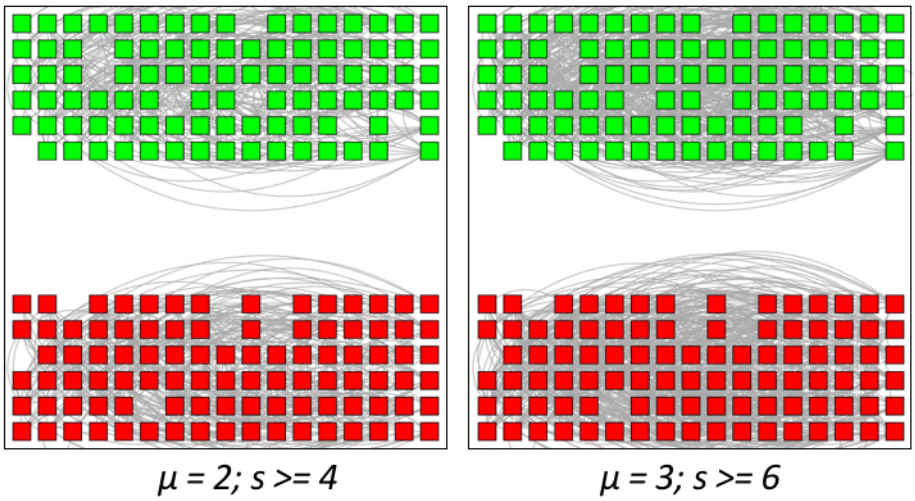
- $eventos$: ordena a lista de eventos de forma crescente segundo o tempo;
- n: quantidade de eventos;
- μ: quantidade de ligações simultâneas;
- $limiar$: $2 * \mu$
- divide a região numa grade quadrada com tamanho $\sim \sqrt{n/3}$ em cada lado;
- **repita**
  - leia o evento e ;
  - α: identifique a cela que e pertence;
  - **para** $i \leftarrow 1$ **até** μ **faça**
    - leia o evento $e + i$;
    - β: identifique a cela que $e + i$ pertence;
    - gere uma ligação entre α e β;
  - **fim**
- **até** *fim dos eventos*;
- **repita**
  - leia o par de nós p, q ;
  - w: some as ligações entre p e q ;
  - simplifique as ligações em apenas uma ligação e adicione o peso w ;
- **até** *fim dos pares de nós*;
- **repita**
  - leia o nó i ;
  - s: some os pesos w das ligações de i;
  - **se** $s <=$ *limiar* **então**
    - remova o nó i da rede;
  - **fim**
- **até** *fim dos nós*;
- aplica o algoritmo de detecção de comunidades sobre a rede;
- retorna os nós encontrados em cada comunidade;

fim

Para ilustrar o uso deste método, na Figura 5.6 mostra-se as comunidades encontradas após a remoção dos nós que não tinham a força suficiente. É notório

a diferença com os outros métodos, considerando que neste caso foi possível remover o estágio intermediário, que representava uma parte ínfima dos eventos. Desse modo, foram identificados satisfatoriamente os dois principais estágios que o conjunto de eventos teve ao longo do tempo. Além disso, é preciso destacar que, mesmo tendo diferentes quantidades de ligações simultâneas e limiares de força, o método conseguiu os mesmos resultados.

Figura 5.6 - Método 3 aplicado sobre os dados gerados



Duas comunidades em ambas as redes, construídas usando duas e três ligações respectivamente. Também, removeram-se os nós com força menor que a requerida pelo método.

Fonte: Produção do autor.

5.3.4 Outras possíveis variações

Ao usar outros limiares, relacionados à distância espacial ou temporal, é possível construir outros métodos. A seguir, descrevem-se algumas das variantes possíveis.

5.3.4.1 Janela de tempo

É possível que em alguns casos seja necessário considerar intervalos ou janelas de tempo para fixos ou dinâmicos para separar o conjunto de dados e construir diversas redes cronológicas. Esta abordagem é ideal quando os dados possuem

ciclos de repetição conhecidos e, portanto, é melhor construir a rede em intervalos de tempo que correspondam a esses ciclos. Por exemplo, se é desejado analisar diariamente os incidentes reportados no trânsito da cidade, então o recomendado seria dividir o conjunto de dados em intervalos diários para ter uma rede para cada intervalo e comparar as mudanças que acontecem neles. A partir dela, uma possível análise seria verificar se as mudanças em dias de semana são diferentes daquelas encontradas no final de semana.

Outro desdobramento ao ter diversas redes que correspondem a janelas de tempo é a possibilidade de construir séries temporais a partir das propriedades dos nós e ligações. Desse modo, ao invés de termos séries temporais sobre propriedades (quantidade, máximo, mínimo, etc) dos eventos, teremos séries temporais sobre características (centralidade, grau de conectividade) das celas onde estes eventos acontecem. Consequentemente, os novos dados gerados permitem que os métodos tradicionais de análise de séries temporais possam ser integrados com o método cronológico.

5.3.4.2 Limiar de distância temporal

Neste caso, um limiar temporal é inserido no momento de construir as ligações da rede cronológica. Desse modo, quando a distância temporal entre dois eventos é acima de um limiar estabelecido, os nós (celas) que representam estes dois eventos não serão ligados, mesmo que sejam próximos espacialmente ou consecutivos. Em alguns casos é natural considerar este tipo de regra quando os eventos são relevantes somente se estes são próximos temporalmente. Por exemplo, se temos um conjunto de movimentos sísmicos alocados numa região, estes precisam ser próximos temporalmente e espacialmente, uns dos outros, para serem estudados como um único fenômeno, caso contrário, eventos sísmicos isolados ou sem relação podem poluir o conjunto de dados.

5.3.4.3 Limiar de distância espacial

Neste caso, uma distância espacial é usada como o limiar para avaliar se é possível ligar dois nós (celas) onde dois eventos consecutivos acontecem. Desse modo, é possível construir uma estrutura de rede considerando a proximidade espacial dos dados. Por exemplo, quando analisamos eventos relacionados a crimes que acontecem numa cidade, é relevante considerarmos buscar se existem crimes similares acontecendo nas redondezas e não procurar no país todo, por exemplo. Portanto, nesse caso, é relevante inserir um limiar espacial.

5.4 Aplicação em dados reais: queimadas na Bacia Amazônica

Esta subseção é um breve resumo dos resultados publicados no trabalho "*From spatio-temporal data to chronological networks: An application to wildfire analysis*" e apresentado na *34th ACM/SIGAPP Symposium on Applied Computing*, em 2018 (VEGA-OLIVEROS et al., 2019). A menção desta publicação é relevante pois envolve o uso do método cronológico em combinação com outras técnicas de mineração de dados. Apesar de que o método usado nessa publicação considera somente a abordagem da construção da rede e não a parte da detecção de padrões, não deixa de ser relevante mostrá-lo aqui, pois desse modo também nos mostra a flexibilidade do método para ser usado de outras formas.

5.4.1 O desafio

Atualmente existe uma imensa quantidade de dados disponíveis sobre queimadas em todo o mundo, que podem ser usados para monitorar e mensurar diversos indicadores, assim como tomar decisões e estabelecer novas políticas. Nesse sentido, apesar de já ser conhecido pelos especialistas sobre o período que mais queimadas aparecem numa região com base nos dados históricos e a experiência ao longo dos anos, é difícil saber o comportamento que as queimadas possuem, uma vez que cada queimada é um evento temporal e espacial único. Portanto, o desafio é mensurar o comportamento das queimadas em conjunto para encontrar as possíveis mudanças ao longo dos anos sobre a região da Bacia Amazônica.

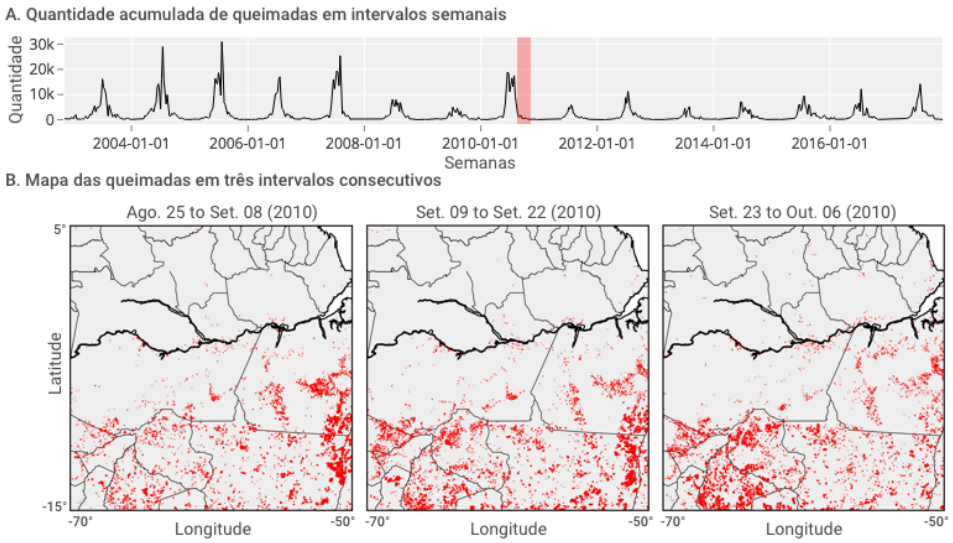
5.4.2 A abordagem

A região sob estudo encontra-se aproximadamente entre a longitude -70 e -50, e latitude 5 e -15, contendo $1,6$ milhão eventos de queimadas (com pelo menos 70% de acurácia), segundo informações coletadas pelos satélites Aqua e Terra da *National Aeronautics and Space Administration* (NASA), entre o primeiro de janeiro de 2003 e 31 de janeiro de 2018 (15 anos de dados). Os dados para este estudo estão disponíveis no site do projeto Active Fire Data, da NASA. Após baixar os dados, simplificaram-se os dados considerando a região de estudo, acurácia e período de análise mencionados anteriormente. Os dados são disponibilizados em formato .txt e além da localização da queimada, contém outras informações interessantes como brilho, acurácia, satélite que detectou a queimada, etc.

A modo de ilustração, na Figura 5.7 mostra-se três intervalos consecutivos, entre 25 de agosto e 6 de outubro de 2010, das queimadas na Bacia Amazônica. Nelas

é possível perceber visualmente a mudança da intensidade de queimadas nessa região, na direção leste - oeste, na parte inferior da região sob estudo.

Figura 5.7 - Mapa das queimadas



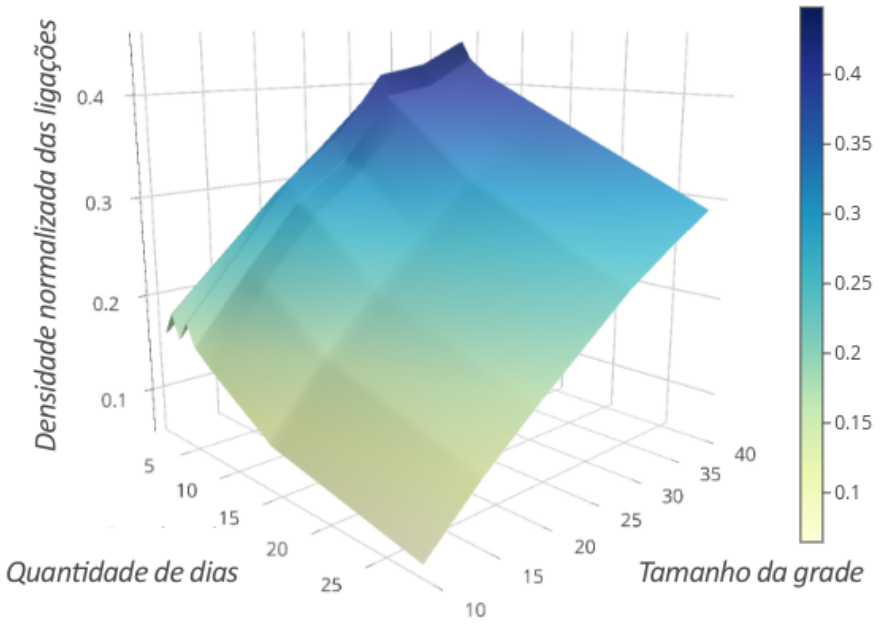
A) Serie temporal semanal da quantidade de queimadas na região da Bacia Amazônica. B) Mapa das queimadas em três intervalos de 14 dias cada.

Fonte: Produção do autor.

Apesar de que o método proposto na pesquisa de doutorado não considerou partições temporais, mas apenas partições espaciais (celas), no caso de grandes períodos de tempo, é necessário realizar este tipo de cortes ou divisões no tempo, pois caso contrário, teremos sobreposição de padrões que dificultarão a identificação correta dos padrões. Com esse intuito, dividiu-se região em 30×30 celas e em intervalos de 7 dias. O motivo para proceder com esta configuração deve-se à procura de mudanças em intervalos de tempo otimizados. Nesse sentido, para chegar nesta configuração foi realizado um teste de sensibilidade dos dados em função do tempo e espaço, assim como a porcentagem de ligações que persistem na rede após simplificar a rede (converter múltiplas ligações em apenas uma e removendo *loops*). Nesse sentido, a Figura 5.8 mostra que o melhor período encontra-se entre 5 e 10 dias, enquanto a melhor divisão de celas começa a partir de 30×30, por esse motivo, ao dividir o conjunto de dados em intervalos de 7 dias e transformá-los em redes segundo o método cronológico, foram obtidas 786 redes. O ganho ao

fazer esta análise de sensibilidade encontra-se no fato de encontrarmos um equilíbrio entre a quantidade de dados (informação disponível) e a resolução mínima (espacial e temporal).

Figura 5.8 - Otimização da configuração



Análise da sensibilidade das configurações para construir rede otimizando o equilíbrio entre a quantidade de dias (intervalo temporal no eixo x), tamanho da grade e mantendo o máximo de ligações na rede (ligações simplificadas / todas as ligações).

Fonte: Produção do autor.

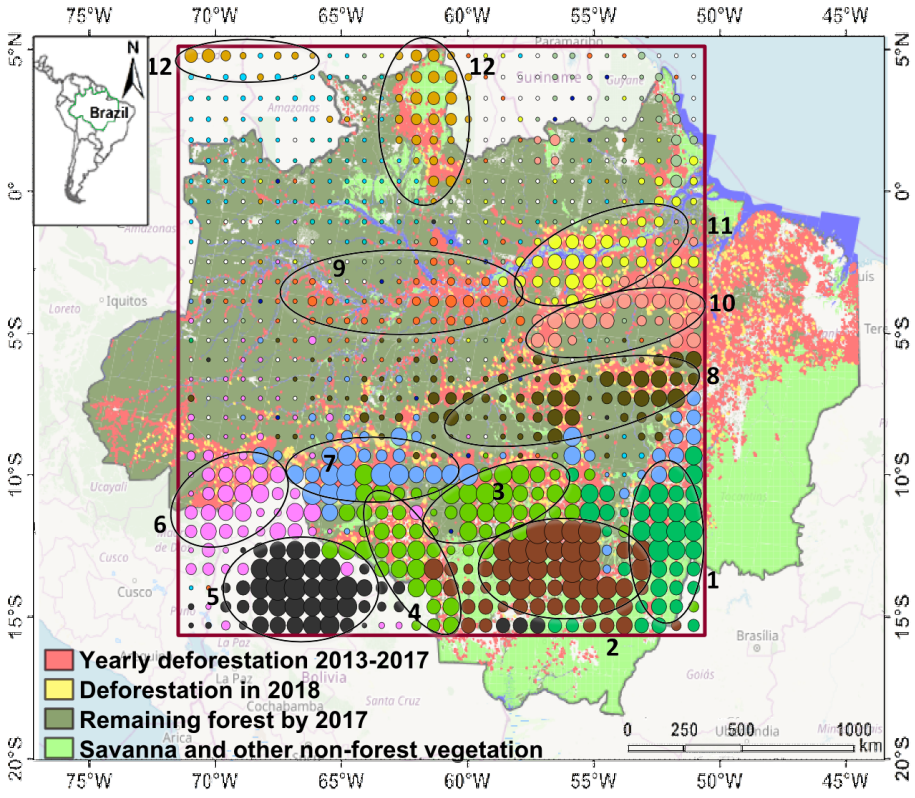
Até este ponto, a abordagem para construir as redes segue o método proposto nesta pesquisa, e a partir de este ponto, este caso de uso seguirá outras formas de explorar as redes construídas. Para isso, é elaborado uma série temporal para cada nó, considerando os valores semanais (cada rede representa uma semana) da centralidade baseada no método K-Core (DOROGOVTSEV et al., 2006). Desse modo, são obtidos 786 pontos em cada série temporal, que são utilizados para construir uma rede baseada no método da correlação, a partir do índice de correlação de Pearson, e limitando a quantidade de ligações para apenas os três vizinhos (nós) mais correlacionados. Desse modo, teremos uma nova rede, com no máximo 786

nós, que estarão conectados entre si a partir do quão correlacionados estão com outros pontos da rede.

5.4.3 Resultados

Tendo a rede construída a partir da similaridade entre nós, é aplicado a detecção de comunidades para identificar os nós que representam regiões com queimadas que possuem comportamento similar. Na Figura 5.9, mostram-se as 12 regiões encontradas que, de modo interessante, em muitas áreas coincidem com as regiões que o desmatamento tem avançado nos últimos anos, outros com regiões dedicadas à agricultura, assim como de pasto para animais.

Figura 5.9 - Comunidades e regiões com desmatamento



12 comunidades encontradas na rede de correlações construída a partir do método cronológico, apresentadas sobre o mapa de desmatamento e tipos de vegetação. Cada comunidade encontrada está relacionada ou coincide, por muitas vezes, com regiões que correspondem a áreas de desmatamento ou uso da terra diverso. De modo geral, a combinação de métodos usados conseguiu encontrar estes grupos mesmo sem termos considerado dados adicionais sobre o tipo de solo ou atividade realizada na região. Mais detalhes sobre as 12 regiões podem ser encontradas na publicação mencionada aqui.

Fonte: Vega-Oliveros et al. (2019).

Também, apesar de ser possível analisar as queimadas diretamente a partir do acumulado de queimadas (serie temporal), a vantagem do método cronológico neste caso se resume a dois pontos: primeiro, por permitir analisar intervalos curtos de tempo (intervalos semanais), o que não seria possível fazer diretamente com a correlação de séries temporais, e segundo, por oferecer flexibilidade para estudar, a partir do mesmo conjunto de redes, longos períodos de tempo e consolidar resultados de 15 anos em apenas um gráfico.

5.5 Considerações finais

A abordagem cronológica para construir a rede e os métodos propostos conseguem trazer informações relevantes a partir de conjuntos de eventos temporais e espaciais. E como visto, estas informações são padrões que correspondem aos estágios ou transições pelos que passa um conjunto de eventos, considerando o tempo e espaço. Também, foi mostrado que estes estágios, dependendo da sua relevância, podem ou não ser considerados, a depender do método utilizado. Desse modo, é claro que cada método avaliado neste capítulo têm aspectos que os diferenciam entre si, sendo seu uso apropriado em cenários diversos.

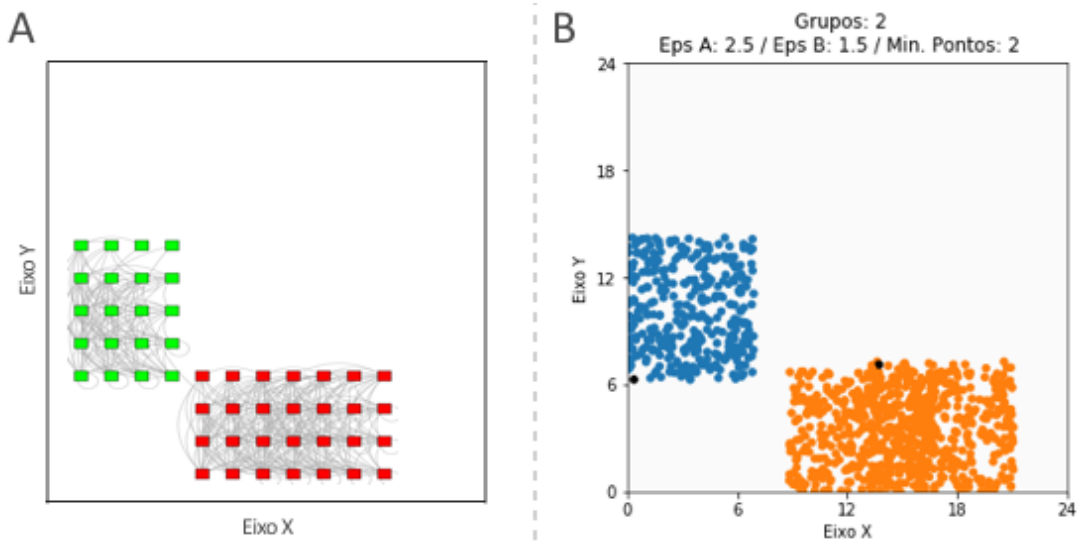
Em síntese, os métodos propostos são úteis para abstrair dados de eventos temporais e espaciais, o que representa um grande ganho desde que, na literatura há uma grande ausência de métodos para explorar este tipo de dados temporais. Em relação aos métodos propostos, o Método 1 (única ligação), por ser básico e o mais simples, apesar de conseguir identificar os estágios principais, têm dificuldades quando há presença de ruido. O Método 2 (ligações simultâneas) consegue aumentar a conectividade da rede, e portanto, evidenciar os estágios com melhor qualidade, mesmo que algum destes estágios sejam ínfimos. Por último, o Método 3 (remoção de nós) possibilita o destaque dos estágios principais ao remover nós fracos ou menos relevantes, desde que estes representam, em muitos casos, a presença de ruído no conjunto de dados.

Além disso, implementação da abordagem cronológica sobre dados reais, mostra a flexibilidade que este possui para se integrar com outros métodos de mineração de dados e redes complexas. Finalmente, mas não menos importante, é preciso destacar a simplicidade da abordagem cronológica, assim como facilidade da implementação de cada método sobre as redes obtidas.

6 COMPARAÇÃO COM O MÉTODO ST-DBSCAN

No capítulo anterior desenvolveu-se a abordagem cronológica e os métodos para identificar estágios ou fases pelos que passa um conjunto de eventos temporais. Também, em capítulos anteriores, mencionaram-se métodos de agrupamento de dados espaço-temporais que são amplamente usados na atualidade para minerar este tipo de dados. A pergunta que emerge naturalmente é, qual a diferença entre os métodos propostos e os métodos de agrupamento já conhecidos? Por exemplo, ao usar o método cronológico e o método ST-DBSCAN no exemplo da Figura 5.2, ambos os métodos obtém duas agrupações, como pode ser visto na Figura 6.1. No primeiro caso, obtemos duas comunidades de nós através do método cronológico simples (usando uma ligação), e, no segundo caso, obtemos dois agrupamentos através do método ST-DBSCAN. Portanto, mais uma vez, qual é o ganho ou diferença dos métodos propostos?

Figura 6.1 - Exemplo do método cronológico e o método ST-DBSCAN



A) Duas comunidades detectadas ao usar o método cronológico para identificar as fases dos eventos temporais. B) Dois agrupamentos encontrados pelo método ST-DBSCAN.

Fonte: Produção do autor.

Portanto, será que os métodos propostos acrescentam alguma informação nova além do que os métodos tradicionais de agrupamento já apresentam? Num simples

olhar, pela Figura 6.1, parece que o método ST-DBSCAN mostra o mesmo resultado do método cronológico.

A resposta a essa pergunta está no fato de que é possível os resultados coincidam quando as fronteiras das comunidades encontradas pelo método cronológico e as maiores distâncias entre os agrupamentos do método ST-DBSCAN sejam os mesmos. Porém, para entender melhor as diferenças é necessário compreender os padrões ou transições que os métodos propostos encontram. Como sabemos, transições acontecem constantemente nos sistemas complexos; portanto, identificar qualquer transição ou mudança de padrão não faz sentido neste contexto, mas aquelas que são significativas. Portanto é possível ter *muitas transições* significativas dentro de um *mesmo agrupamento de dados*. Desse modo, o desafio então, é encontrar as transições sem levar em consideração os possíveis agrupamentos.

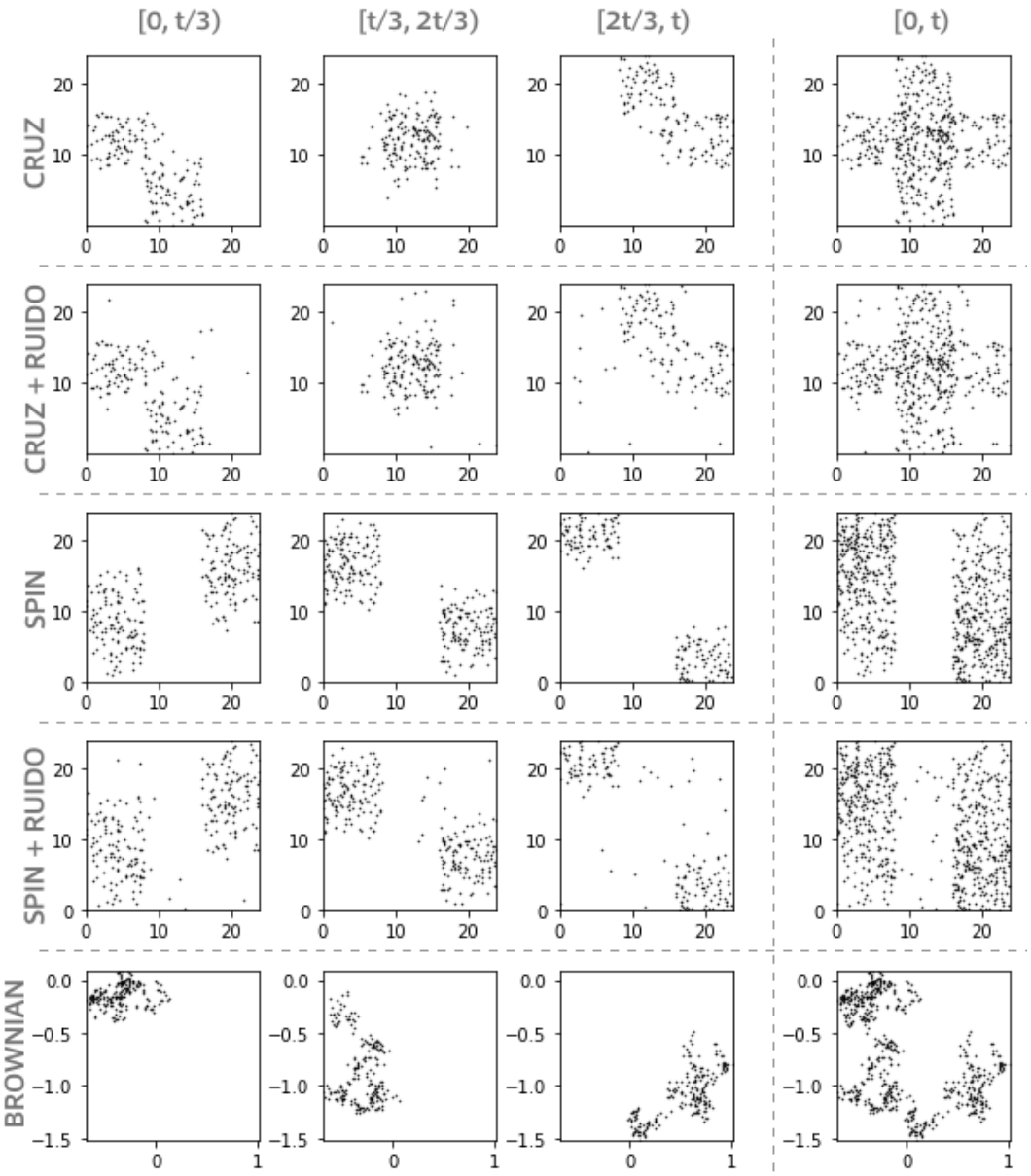
Para compreender melhor as diferenças entre ambas abordagens (detecção de agrupamento e mudanças), a continuação mostra-se alguns modelos artificiais ou sintéticos gerados para este fim. Destacando que, o intuito não é simular eventos reais mas avaliar o desempenho de ambas as abordagens em cenários onde padrões de transição foram gerados propositalmente.

Ao todo, são cinco modelos, como mostrados na Figura 6.2, sendo os dois primeiros baseados no formato cruzado ($\times$) de dados para analisar como é o comportamento do método perante a interseção de dois grupos de dados. Também, os dois modelos seguintes procuram avaliar o método proposto em condições de aumento gradual da quantidade de eventos em dois grupos que seguem direções opostas paralelas ($||$). E, por último, o quinto modelo segue o movimento *browniano* (similar ao movimento de partículas num fluído), para avaliar o desempenho dos métodos ao acompanhar o percurso de uma caminhada aleatória. Nesse sentido, na Figura 6.2, mostram-se os três momentos sequenciais de cada modelo nas primeiras três colunas e os eventos acumulados na quarta coluna. Posteriormente cada modelo artificial será analisado com maior detalhe.

Na primeira fila observa-se o modelo *cruz*, que mostra o *deslocamento* vertical e horizontal de dois grupos de eventos. Na segunda fila mostra-se o mesmo modelo, porém, inserindo ruído ao realocar 10% dos eventos de forma aleatória. Na terceira fila mostra-se o modelo paralelo (ou spin), que é formado por dois grupos de eventos indo em direções contrárias e de forma incremental (em quantidade de eventos). Na quarta fila apresenta-se o mesmo modelo, porém, realocando 10% dos eventos de forma aleatória (ruido). E, finalmente, na quinta fila temos um con-

junto de eventos que sinalizam os locais pelos que passa um caminhante aleatório seguindo a simulação do modelo browniano.

Figura 6.2 - Modelos de eventos artificiais



Cada fila representa um modelo artificial de eventos. As três primeiras colunas correspondem a três intervalos diferentes de tempo de cada modelo gerado. A última coluna corresponde ao acumulado dos eventos de cada modelo.

Fonte: Produção do autor.

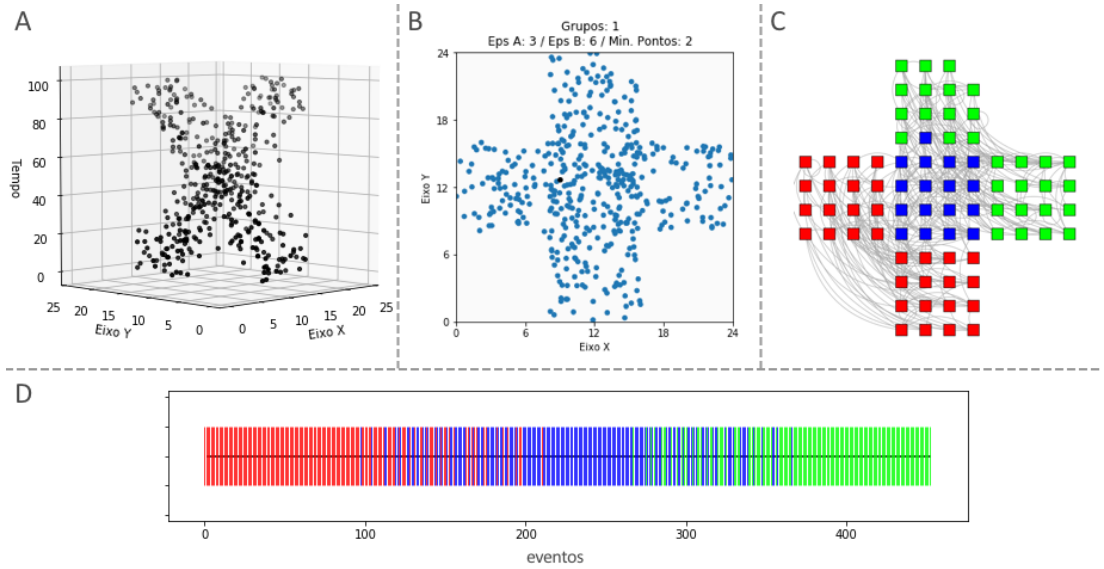
A seguir, uma análise comparativa de cada modelo de dados, considerando o método ST-DBSCAN e o método cronológico.

6.1 Modelo de cruzamento de eventos

O modelo: dois grupos de eventos são gerados de forma aleatória seguindo as direções norte-sul e oeste-leste. Ao todo são 465 eventos espaço-temporais que aparecem sequencialmente ao longo do tempo.

Ao usar o algoritmo ST-DBSCAN sobre o cruzamento ou interseção entre dois conjuntos de eventos espaço-temporais (gerados de forma aleatória mas seguindo um padrão) observa-se que o algoritmo encontra um único grupo, como mostrado na Figura 6.3B. Porém, se olharmos desde a perspectiva de padrões, com foco na detecção de transições, sabemos que são dois os grupos que se interceptaram, e, portanto, houve um antes, um durante e um depois dessa ação. De forma clara, essa é uma informação que o método ST-DBSCAN não consegue fornecer, mesmo mudando os parâmetros do algoritmo, como pode ser visto no Anexo A.1.

Figura 6.3 - Cruzamento de eventos



A) Modelo de cruzamento de eventos espaço-temporais. Ao todo, são 454 eventos alocados entre as coordenadas $(0, 0)$ e $(24, 24)$. B) Método ST-DBSCAN aplicado sobre o modelo A, usando os parâmetros $\epsilon_a = 3, \epsilon_b = 6, min = 2$, sendo estes as distâncias horizontal, vertical e a quantidade mínima de pontos para formar um grupo. C) Método cronológico aplicado sobre o modelo A, mostrando em evidência os três estágios pelos quais os dados temporais passaram. D) Linha do tempo de todos os eventos (linhas verticais) coloridos segundo a comunidade ou estágio que pertencem.

Fonte: Produção do autor.

Por outro lado, ao usar o método cronológico, na Figura 6.3C, conseguimos observar claramente três cores (vermelho, azul e verde), que representam as comunidades da rede e também os estágios ou fases pelas quais o conjunto de dados passou, que de forma interessante, correspondem com os estágios dos dados que foram gerados propositalmente: o antes da interseção, o durante, e o depois da interseção. Na Figura 6.3D é mostrado claramente a transição entre as diferentes fases do conjunto de eventos, tudo isso, sem precisar estabelecer a quantidade de mudanças que o algoritmo deve buscar. Neste modelo, ao construir a rede foi usado uma grade $g : 12 \times 12$, ou seja, a rede terá no máximo 144 nós. Este tamanho de grade segue a sugestão do método de que cada lado da região sob estudo deve ter $\sim \sqrt{n/3}$ divisões, onde n é a quantidade de eventos.

Além dos resultados mostrados aqui, também foram realizados outros testes com

variação de parâmetros no método ST-DBSCAN, assim como nos métodos propostos. Para ver os resultados dos testes, revise o Apêndice A.1 no final do documento. Neles é visível que o método ST-DBSCAN, mesmo usando diferentes parâmetros, não consegue encontrar alguma informação sobre a interseção dos dois conjuntos de eventos. Por outro lado, ao aplicar os métodos propostos, mesmo usando parâmetros não recomendados (tamanho de grade diferente do sugerido), ele consegue encontrar a interseção inserida propositalmente.

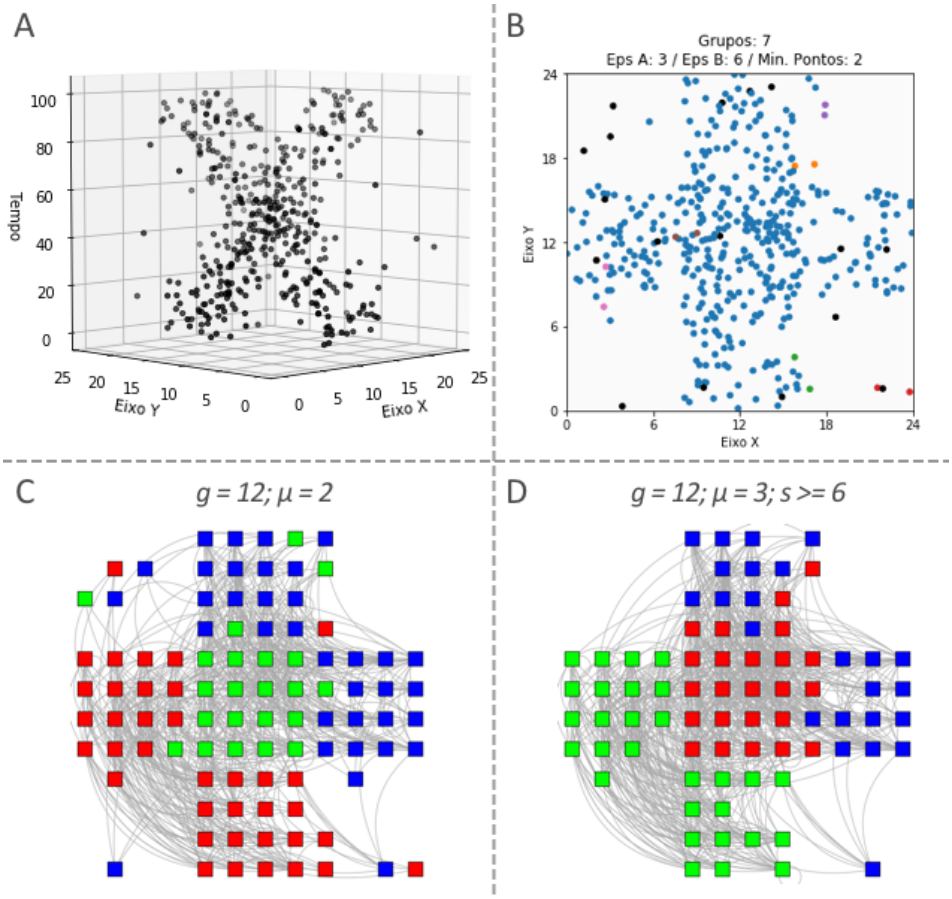
6.2 Modelo de cruzamento de eventos com ruído

O modelo: de modo similar ao anterior, dois grupos de eventos são gerados ao longo do tempo e de forma aleatória, seguindo as direções norte - sul e oeste - leste. A diferença, neste caso, é que 10% dos eventos foram removidos e adicionados de forma aleatória em locais diferentes, para representar o ruido do modelo.

Nesta análise, também usaram-se os método de agrupamento ST-DBSCAN e o método cronológico. O primeiro, encontrou apenas um grupo de eventos e alguns outros eventos isolados, sinalizados como ruído (que não pertencem a algum grupo), como é visível na Figura 6.4B. Desse modo, o método alcançou seu objetivo de agrupar os eventos, no entanto, de modo similar ao exemplo anterior, ele não mostrou alguma informação que revele alguma transição no conjunto de dados, mesmo mudando diversos parâmetros como mostrado no Apêndice A.3.

Por outro lado, aplicou-se o método cronológico numa grade $g : 12 \times 12$ considerando $\mu = 2$ ligações simultâneas (usando a variação das ligações simultâneas da seção anterior). Dessa forma, como explicado anteriormente, estaremos adicionando maior conectividade à rede, e mesmo com o ruído, seria possível identificar padrões que revelem as transições procuradas. Na Figura 6.4C mostra-se como o método cronológico consegue dividir os eventos nos três estágios esperados (mesmo com o ruído). Também, usando uma abordagem mais agressiva, ao remover nós que não tenham força suficiente, ou seja manter apenas os nós com $s \geq 6$, a rede diminui o ruido e deixa mais notório as fases que correspondem aos três estágios procurados, como visto na Figura 6.4D.

Figura 6.4 - Cruzamento de eventos com ruido



A) Modelo dos eventos espaço-temporais com cruzamento e com 10% de ruído. B) Método ST-DBSCAN aplicado sobre o modelo A. Nele foram usados os parâmetros $\epsilon_a = 3, \epsilon_b = 6, min = 2$, sendo estes as distâncias horizontal e vertical, respectivamente, e a quantidade mínima de pontos para formar um grupo. C) Método cronológico aplicado sobre o modelo A, mostrando em evidência os três estágios pelos que os dados temporais passaram, apesar do ruído inserido. A grade corresponde ao tamanho 12×12 e considerando $\mu = 2$ duas ligações simultâneas. D) Método cronológico aplicado sobre o modelo A, considerando a grade de tamanho 12×12, ligações simultâneas $\mu = 3$ e mostrando apenas os nós com força $s \geq 6$.

Fonte: Produção do autor.

Para ver outros testes realizados em ambos métodos e variando os parâmetros, revise o Anexo A.3. Neles é possível verificar que, quando há presença de ruído, o melhor é usar a variação do método considerando as ligações múltiplas, assim como estabelecer a quantidade de divisões da grade seguindo a regra de que cada lado da região sob estudo deve ter $\sim \sqrt{n/3}$ divisões, como sugerido no método.

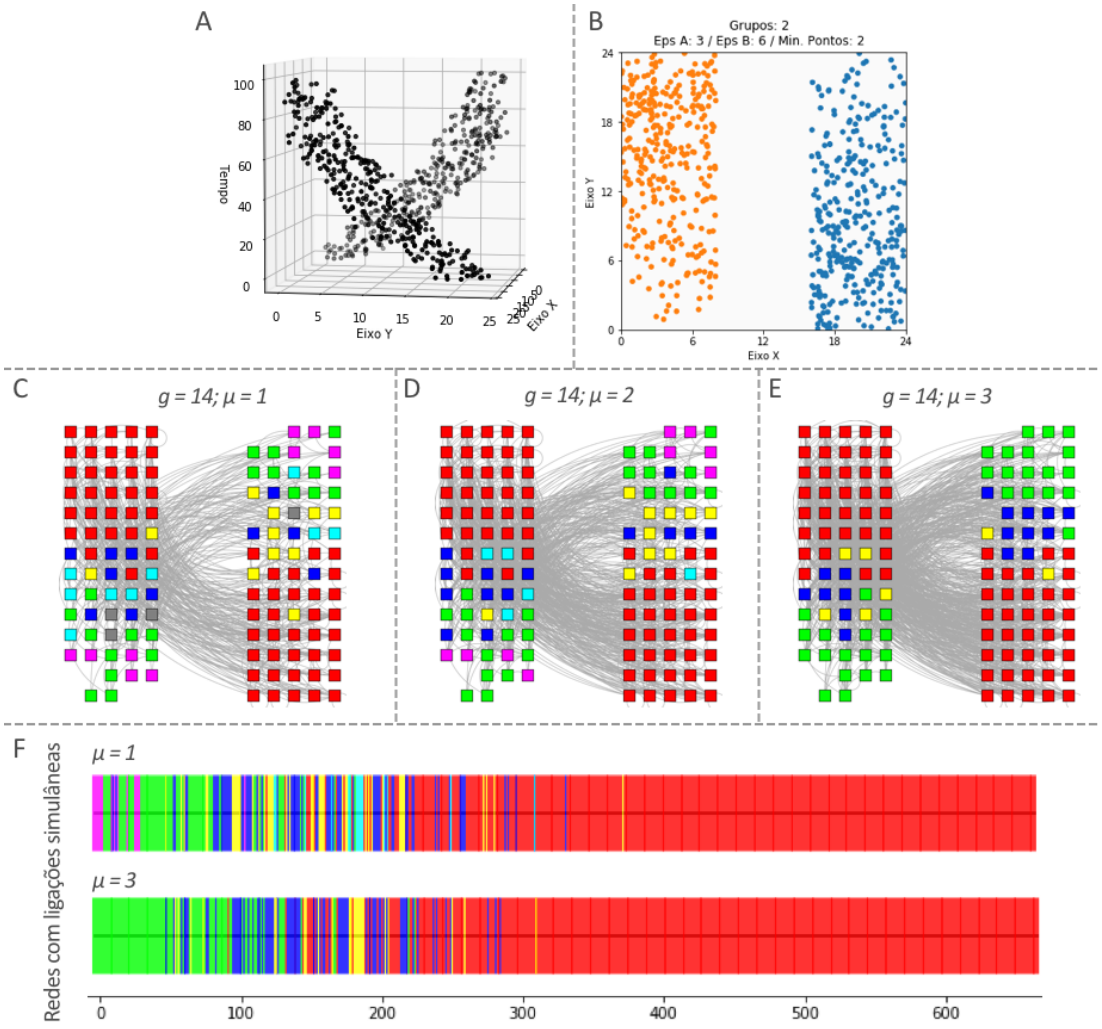
6.3 Modelo de eventos crescentes em paralelo

O modelo: dois grupos de eventos são gerados de forma aleatória, seguindo as direções norte- sul e norte- sul, de forma paralela e separada (sem interseção entre ambas). Também ao longo do tempo, a quantidade de eventos é incrementada gradualmente.

Neste modelo, será testado como os algoritmos reagem quando há variação na quantidade de eventos de forma incremental e em grupos separados acontecendo ao mesmo tempo. Em primeiro lugar, ao usar o método ST-DBSCAN, este encontra diversas quantidades de grupos, a depender dos parâmetros usados (Ver Figura 6.5B). Apesar disso, o algoritmo claramente consegue identificar os dois grupos principais que foram gerados pelo modelo. Testes com outros parâmetros podem ser revisados no Anexo A.2.

Por outro lado, ao usar o método cronológico, mesmo que a variação da intensidade tenha sido gradual e aos poucos, este método consegue encontrar diferentes estágios relacionados às mudanças graduais dos eventos. É claro que a quantidade de mudanças vai depender da variação das ligações múltiplas μ, pois ao aumentar a quantidade de ligações a rede torna-se mais conectada e, portanto, diminuirá a quantidade de comunidades na rede. Por exemplo, na Figura 6.5C, 6.5D e 6.5E, temos a rede com uma, duas e três ligações simultâneas, respectivamente. Nelas observamos que, ao aumentar a quantidade de ligações simultâneas, diminui a quantidade de mudanças (ou estágios) encontradas, porém, o método consegue manter o fluxo ou ordem das mudanças, marcando claramente a ordem e intensidade delas, como observado nas barras da Figura 6.5F. Nelas observa-se cada evento e a respectiva cor da comunidade a que pertence. Nota-se que quanto maior a quantidade de ligações simultâneas mais claro fica a transição entre os diversos estágios das mudanças ao longo do tempo.

Figura 6.5 - Eventos em paralelo



A) Modelo paralelo e incremental gradativo de dois grupos com 661 eventos espaço-temporais. B) Método de agrupamento ST-DBSCAN aplicado sobre o modelo, mostrando os dois grupos de eventos que foram gerados. C) Método cronológico aplicado sobre o modelo, considerando uma ligação simultânea e sobre uma grade 14×14 (seguindo a recomendação do método). Com essa configuração foram encontrados 6 estágios ou mudanças. D) Com configuração similar à imagem anterior, porém considerando duas ligações simultâneas. Ao usar o método cronológico, como resultado, foram encontradas 6 padrões, que correspondem às transições ao longo do tempo e espaço. E) Neste caso, usaram-se três ligações simultâneas. Ao aplicar o método cronológico, encontraram-se quatro mudanças. F) Duas barras de cores que representam os eventos de forma sequencial. Cada linha vertical é um evento e está colorido segundo a comunidade que representa. A barra superior corresponde à configuração com uma ligação simultânea, enquanto a barra inferior corresponde à configuração de três ligações simultâneas.

Fonte: Produção do autor.

Para comparar ambos métodos usando outros parâmetros, revise o Apêndice A.2 para maiores detalhes.

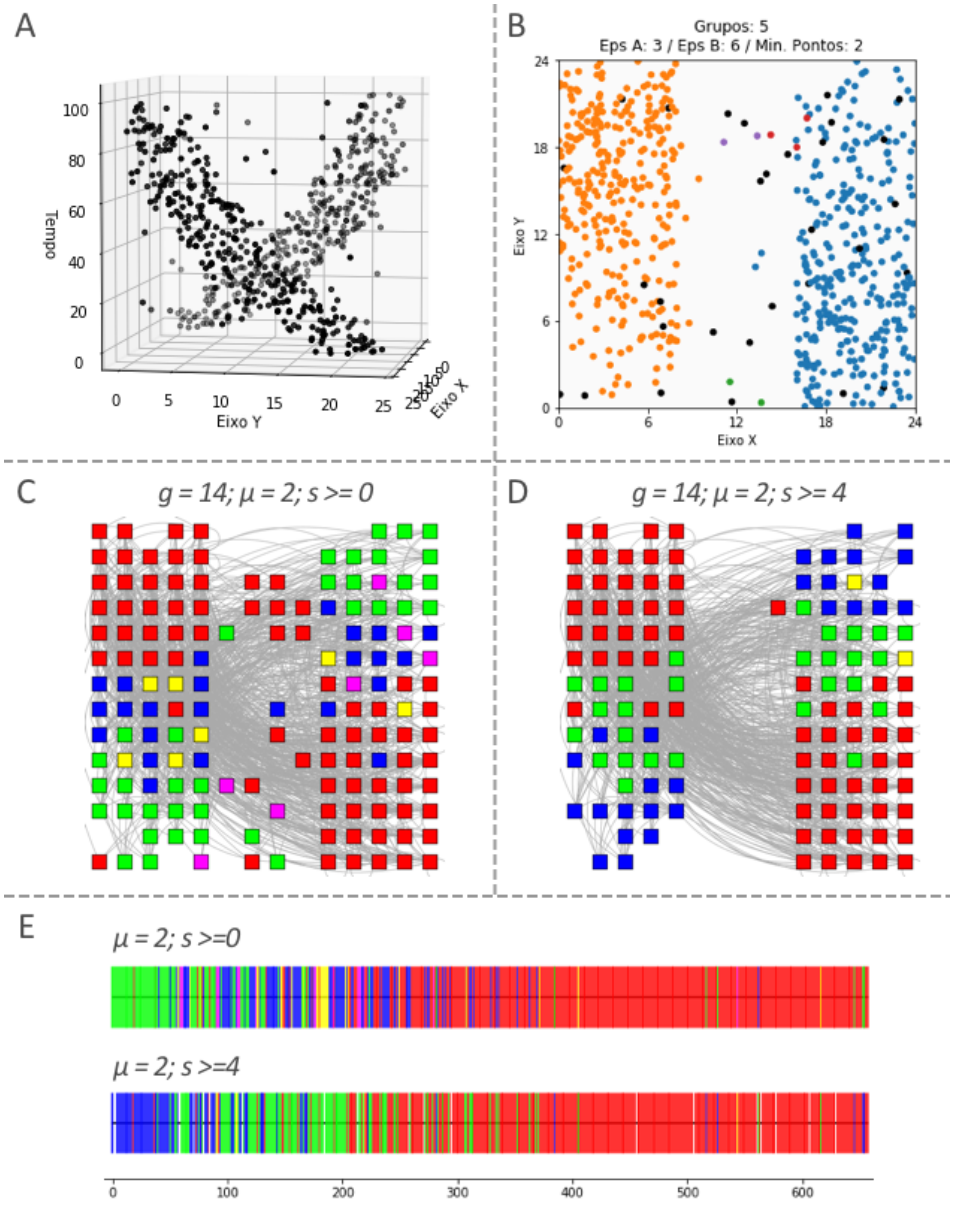
6.4 Modelo de eventos crescentes em paralelos e com ruído

O modelo: de modo similar ao modelo anterior, dois grupos de eventos são gerados de forma paralela e aparecem aumentando gradualmente a quantidade de eventos. No entanto, neste modelo é inserido ruído, que é gerado a partir da realocação espacial de 10% dos eventos.

O teste realizado é para avaliar o desempenho do método de agrupamento ST-DBSCAN e método cronológico sob as condições de mudança de quantidade de eventos e ruído. Desse modo, ao aplicar o ST-DBSCAN encontramos que ele consegue separar diversos grupos de eventos ao variar os parâmetros de entrada do algoritmo. Mesmo assim, é evidente que ele consegue encontrar os dois grupos principais de eventos paralelos, assim como alguns elementos que correspondem ao ruido (Ver Figura 6.6B).

Devido à presença de ruído, ao usar o método cronológico, construímos a rede usando duas ligações simultâneas para aumentar a conectividade da rede, e desse modo facilitar ao algoritmo a identificação das comunidades. Como resultado, o método consegue encontrar diversos padrões no conjunto de eventos, de forma similar aos resultados apresentados no modelo anterior, e mesmo com a presença do ruído. Na Figura 6.6C, por exemplo, encontramos cinco comunidades, representando os padrões de transição nos eventos. Além disso, é necessário destacar que apesar deste resultado, o ruído continua presente nos resultados, o que muitas vezes não é desejado. Por esse motivo, para diminuir a presença do ruído, na Figura 6.6D são removidos os nós com *força* menor ou igual a quatro, e como resultado, o ruído diminui consideravelmente, deixando assim, mais evidente os padrões de transição. No entanto, ao remover o ruído com essa abordagem, alguns nós ou celas que não são ruído acabam sendo removidos também. Mesmo assim, o objetivo não é afetado em ambas as situações, removendo ou não o ruído, o método cronológico consegue resultados consistentes, como visto nas barras da 6.6E. Em ambas barras é claro que o método encontra um estágio predominante (cor vermelha) por ser a fase em que se intensifica a quantidade de eventos, e os outros dois estágios (cor verde e azul) que representam o estágios iniciais onde os eventos são poucos e espaçados ou separados entre si.

Figura 6.6 - Eventos paralelos com ruido



A) Modelo paralelo e incremental gradativo de dois grupos de eventos com ruído. Ao todo são 661 eventos espaço-temporais. B) Método de agrupamento ST-DBSCAN aplicado sobre o modelo com ruído, mostrando principalmente os dois grupos de eventos que foram gerados. C) Estágios encontrados após usar o método cronológico considerando 2 ligações simultâneas. D) Transições encontradas após usar o método cronológico e removendo o ruído (deixando apenas os nós com força maior a 4). E) Comparativo da linha de tempo de ambas os resultados, com e sem remoção do ruído inserido no modelo original.

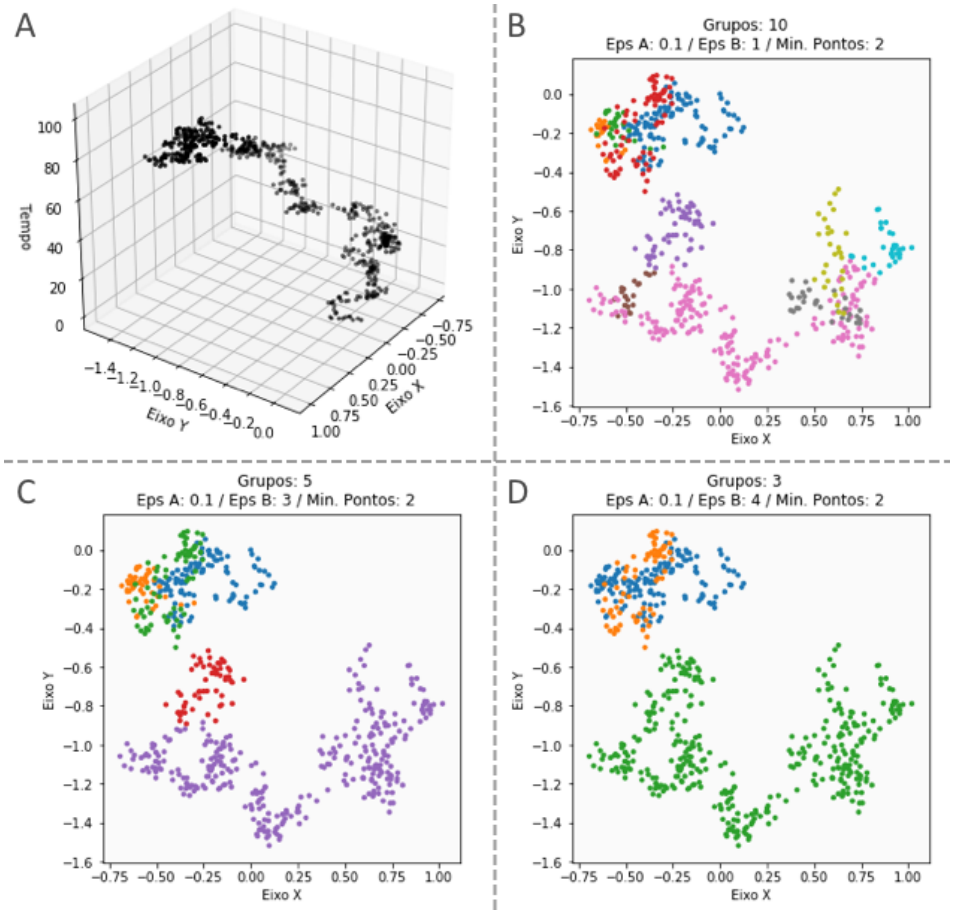
Fonte: Produção do autor.

Para comparações dos métodos usando outros parâmetros, ver no Apêndice A.4.

6.5 Modelo de eventos brownianos

O modelo: nesta simulação, os eventos espaço-temporais correspondem aos locais pelos aonde passa uma partícula que segue um comportamento segundo o modelo matemático de Wiener da teoria do Movimento Browniano (MALLIARIS, 2008). Este é, em poucas palavras, uma representação matemática da flutuação ou movimento de uma partícula num fluído, que pode ser líquido ou gasoso.

Figura 6.7 - Eventos relacionados ao movimento browniano - Método ST-DBSCAN

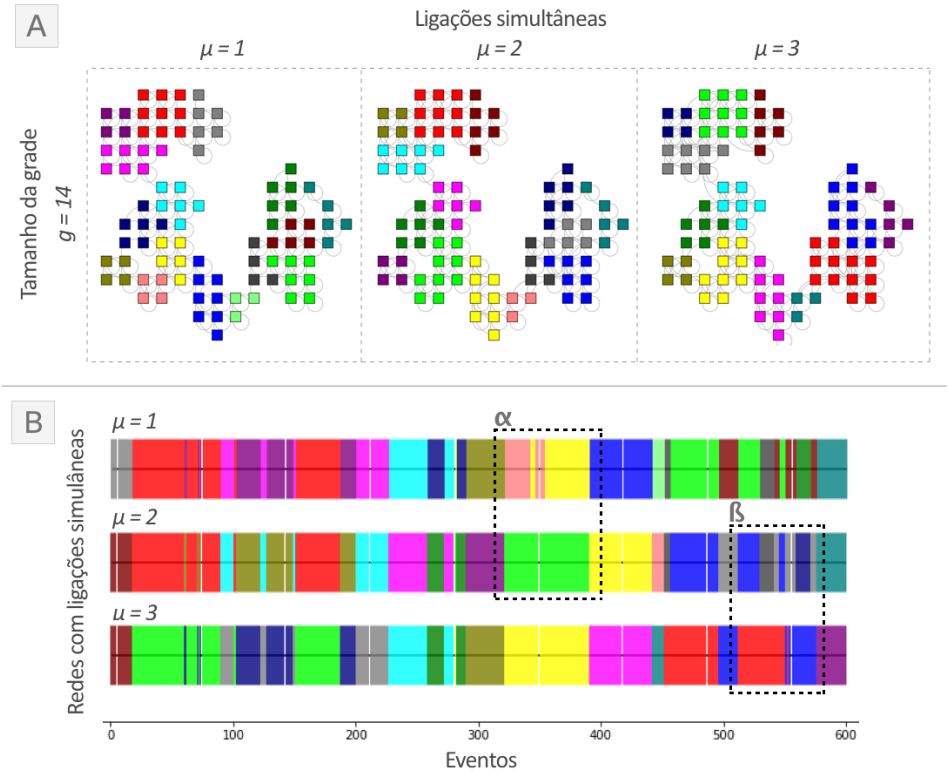


A) 601 eventos que marcam os locais que passa um caminhante aleatório, segundo o Movimento Browniano de Wiener. B) Algoritmo ST-DBSCAN aplicado sobre o modelo encontrou 10 grupos, usando os parâmetros $\epsilon_A = 0,1$ e $\epsilon_B = 1$. C) Ao usar os parâmetros $\epsilon_A = 0,1$ e $\epsilon_B = 3$, o algoritmo encontrou 5 grupos. D) Da mesma forma, ao usar os parâmetros $\epsilon_A = 0,1$ e $\epsilon_B = 4$ foram encontrados 3 grupos.

Fonte: Produção do autor.

Ao usar o algoritmo ST-DBSCAN sobre este modelo, observou-se que a quantidade de agrupações encontradas varia dependendo dos parâmetros iniciais que sejam configuradas no algoritmo. Portanto, não há como identificar uma quantidade predominante de grupos a partir dos pontos pelos que a partícula percorreu. Por exemplo, na Figura 6.7B, 6.7C e 6.7D foram identificados dez e cinco e três grupos, respectivamente. Nestas imagens, vemos que ao modificar apenas um parâmetro (ϵ_B) é gerado uma grande variação na quantidade de grupos de eventos encontrados pelo ST-DBSCAN.

Figura 6.8 - Eventos relacionados ao movimento browniano - Método cronológico



A) Três variações no método cronológico considerando uma, duas e três ligações simultâneas, sobre uma grade 14×14. Note-se a similaridade das divisões das comunidades entre as três redes. B) Linhas de tempo dos eventos coloridos com a cor da respectiva comunidade ou estágio que pertencem, das três redes. Também em α e β destaca-se como ao aumentar a quantidade de ligações múltiplas as comunidades se misturam ou unem entre si, formando comunidades ou estágios maiores.

Fonte: Produção do autor.

Por outro lado, ao usar o método cronológico para encontrar as mudanças nos eventos temporais, o algoritmo encontra claramente diversas transições pelos que o caminhante aleatório passou. Este resultado é consistente tanto para uma ligação simultânea como para duas ou três ligações simultâneas, como pode ser visto na Figura 6.8A. Também, a quantidade de transições ou mudanças encontradas se mantém muito próximos entre si, sendo 16, 15 e 13, para redes com uma, duas e três ligações simultâneas, respectivamente. Além disso, as mudanças de fases ou estágios acontecem de forma clara, com pouco ou quase nenhum ruido entre as transições, como apresentando na Figura 6.8B. Além disso, é preciso destacar que quanto mais ligações múltiplas (μ) são usadas na construção da rede, menor é a quantidade de estágios pelas que passa o conjunto de eventos, devido à fundição ou mistura de comunidades. Essa característica é evidente nas áreas marcadas como α e β da Figura 6.8B.

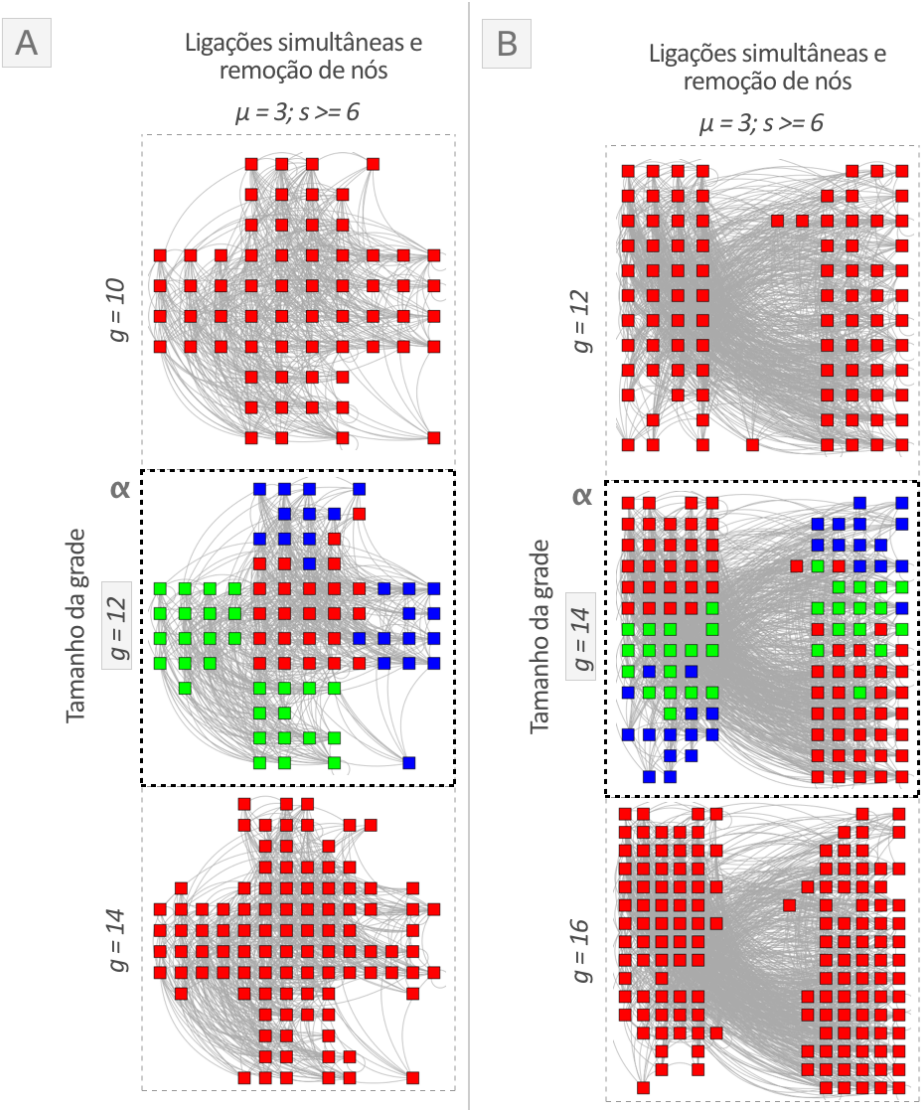
6.6 O tamanho da grade e o ruído

Na seção anterior apresentaram-se cinco modelos de eventos espaciais e temporais, sendo dois deles com inserção de ruído no conjunto de eventos. Nesse sentido, é preciso mencionar que, apesar da flexibilidade do método para mudar o tamanho da grade (da região sob estudo), a recomendação do método para calcular o tamanho da grade ideal ($\omega \sim \sqrt{n/3}$) se destaca nos cenários onde o conjunto de eventos têm ruído. Como dito anteriormente, esta regra procura garantir que a quantidade de eventos (em média) nas celas seja a suficiente para formar uma rede com suficiente capacidade de formar boas comunidades de nós. Isto deve-se ao fato de, se a rede tiver muitos nós e poucas interações, ela corre o risco de se tornar esparsa e portanto, será difícil obter comunidades a partir dela. Por outro lado, se houver poucos nós e muitas interações, a rede será densa, e portanto, há o risco de encontrarmos apenas uma comunidade.

Nos dois modelos com ruido apresentados anteriormente, a recomendação nesse caso era usar ligações múltiplas $\mu \geq 2$ para ajudar a aumentar a conectividade da rede e, mesmo com ruido, conseguir identificar os estágios pelos que passa o conjunto de eventos ao longo do tempo. Também, uma outra alternativa foi apresentada, que consiste na remoção de nós abaixo de um limiar. Este limiar está relacionado com a procura de nós com a menor *força* (s), pois há a possibilidade destes conterem eventos com ruido. Nesse sentido, ao usar esta abordagem nos dois modelos com ruido, considerando três ligações simultâneas $\mu = 3$ e mantendo apenas o nós com *força* $s \geq 6$, obtemos que o melhor resultado é mostrado ao

considerar a recomendação dada no método, sobre o tamanho da grade ideal.

Figura 6.9 - Tamanho da grade e a força dos nós



A) Modelo de cruzamento de eventos com ruido, construídos sobre diferentes tamanhos de grade, considerando três ligações múltiplas e mantendo apenas os nós com força maior ou igual a seis. B) Modelo paralelo e gradativo de eventos com ruido, considerando diversos tamanhos de grade na sua construção e três ligações múltiplas. Neste também foram mantidos apenas os nós com força maior ou igual a seis. As imagens marcadas com α representam as redes construídas considerando a configuração recomendada pelo método.

Fonte: Produção do autor.

Na Figura 6.9 mostra-se os testes realizados em ambos modelos com ruido, sendo que ambos têm diferente quantidade de eventos, e portanto, o tamanho da grade ideal também precisa ser diferente. O modelo cruzado (com ruído) têm 454 eventos e o modelo paralelo (com ruído) têm 661 eventos, desse modo, o tamanho ideal de grade seria 12×12 e 14×14 respectivamente. Na mesma imagem, em α, destaca-se dois detalhes: em primeiro lugar, ambos resultados conseguem identificar a transição entre os estágios procurados, e em segundo lugar, eles também diminuem o ruído quando comparado com os resultados ao usar uma ligação simultânea. É claro que, ao remover os nós "fracos", alguns nós que não contém eventos de ruido, também são removidos. Apesar disso, em ambos exemplos, o ganho obtido nos resultados supera a presença desta desvantagem ao remover alguns nós.

6.7 Principais diferenças entre métodos de agrupamento ST-DBSCAN e o Cronológico

As diferenças entre os métodos de agrupamento de dados espaço-temporais (Ex. ST-DBSCAN) e o método cronológico (explicado nesta seção) encontra-se, principalmente, em dois pontos:

6.7.1 Abordagem

Por um lado, os métodos de agrupamento estão baseados em métricas de similaridade espacial e temporal usando cálculos como a distância ou centroides, e tendo como base as comparações entre estes cálculos para otimizar o resultado final. Por outro lado, o método cronológico é baseado numa construção simples dos eventos, pois apenas precisam estar em ordem temporal e a identificação das celas a que cada evento pertence. Além disso, para fazer a detecção das comunidades, este é realizado sobre as celas apenas, e não sobre os eventos, o que permite reduzir consideravelmente o processamento computacional, desde que a quantidade de celas é muito menor que a quantidade de eventos.

Também, enquanto o ST-DBSCAN e similares precisam de limiares ou parâmetros (distância máxima, quantidade mínima de pontos) para identificar os grupos, no método proposto o único parâmetro necessário é o tamanho da cela, que é sugerido pelo próprio método, mesmo assim, essa variável pode mudar dependendo da resolução espacial procurada. Além disso, o método cronológico é flexível para adaptar o processo da detecção de padrões de transição. Neste ponto é necessário destacar que, como mencionado anteriormente, ambos os métodos tem objetivos

diferentes.

Em caso da presença de ruído, na abordagem cronológica é possível aumentar o número de ligações simultâneas para fortalecer a conectividade da rede, também, caso necessário, é possível remover o ruído de forma parcial ao deixar na rede apenas os nós que possuem *força* acima de um limiar (normalmente igual ao dobre do número de ligações simultâneas). Além disso, em relação à detecção das transições, como o método cronológico é suportado pelos algoritmos de detecção de comunidades em redes, é possível que outros algoritmos de detecção de comunidades sejam usados, oferecendo deste modo, flexibilidade para obter outros resultados a partir de novas abordagens.

6.7.2 Interpretação

Apesar de os métodos de agrupamento por similaridade serem úteis em diversos cenários, ao testar seu desempenho em situações como o *cruzamento* de grupos de eventos ou na *variação gradativa* da quantidade de eventos, conforme apresentado neste capítulo, é possível ver que eles não têm a capacidade de identificar estes padrões de transição através da detecção de grupos. Nesse sentido, o método cronológico ao ser testado nas situações mencionadas anteriormente, consegue resultados relevantes ao identificar estas transições (interseção de grupos de eventos, variação da intensidade de eventos, mesmo com ruído). Portanto, apesar de ambas abordagens encontrarem grupos de eventos, estes grupos possuem significados diferentes, i.e., nos métodos tradicionais de detecção de grupos, cada grupo representa o quão similar (baseadas na distância ou proximidade) é cada evento com outros eventos; por outro lado, no método cronológico, cada grupo representa uma transição ou mudança no conjunto de eventos.

6.8 Considerações finais

Foi mostrado que o método cronológico é uma nova abordagem para detecção de mudanças em dados de eventos espaciais e temporais; desse modo, este método não representa uma alternativa aos métodos de detecção de agrupamento de dados, mas um complemento para obter novas informações a partir do mesmo conjunto de dados.

Também, através de diversos modelos de dados foram comparados, lado a lado, os métodos baseados na abordagem cronológica e o método ST-DBSCAN. Os resultados deste comparativo mostraram que, em todos os casos, cada abordagem

cumpre sua função. No primeiro comparativo, onde existia cruzamento de dois grupos de eventos, o método ST-DBSCAN encontrou, principalmente, um grupo de eventos, por outro lado, os métodos propostos encontraram os três estágios das mudanças procuradas, mesmo em caso da presença de ruído. Neste último caso, é necessário usar o Método 2 ou 3, para diminuir a interferência do ruido nas transições encontradas. No segundo comparativo, ao termos variação gradativa da quantidade de eventos, e de forma simultânea em dois locais da região, o método ST-DBSCAN encontrou, em todos os casos, ambos os grupos de eventos, por outro lado, ao usar o método cronológico, encontraram-se três estágios nas mudanças gradativas na quantidade de eventos. Apesar do ruído, o método também conseguiu encontrar estes estágios. No último modelo, a partir dos eventos gerados pelo caminhante aleatório Browniano, o método ST-DBSCAN não teve resultado consistente, pois mostrava diversas quantidades de grupos a depender dos parâmetros usados. Por outro lado, o método proposto encontrou, praticamente, os mesmos estágios e mudanças ao longo do tempo.

Além disso, como visto nos experimentos realizados até este ponto, o método cronológico tem algumas limitações em relação ao ruído, que por vezes modifica os resultados desejados; contudo, o método consegue remover os nós que contem ruído, e, dessa forma, aprimorar a detecção de mudanças. Finalmente, a partir das mudanças é possível mensurar o tempo e tamanho que cada estágio tem, e, assim, construir o ciclo de mudanças.

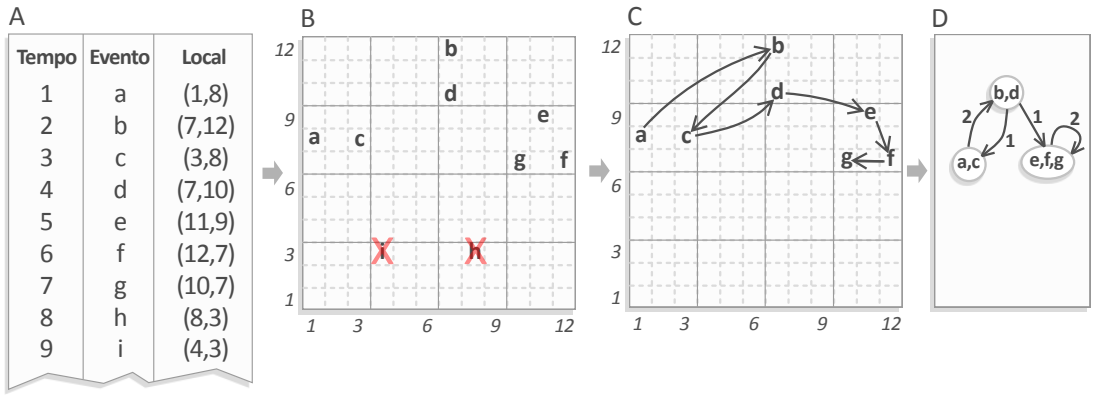
7 OUTROS DESDOBRAMENTOS DA ABORDAGEM CRONOLÓGICA

Além dos resultados mostrados até aqui, é esperado que novas abordagens sejam derivadas do método principal para situações diferentes das estudadas nesta pesquisa. Desse modo, nas próximas linhas são brevemente discutidas algumas considerações importantes assim como os desafios que devem emergir a partir das limitações atuais. Além disso, com esta discussão será possível avaliar o rumo de potenciais trabalhos futuros. A seguir, apresenta-se as possíveis três novas abordagens derivadas deste trabalho.

7.1 ST-DBSCAN + Cronológico

Esta nova abordagem consiste em unir o método de agrupamento por similaridade (ou densidade) ST-DBSCAN e o método cronológico, de detecção de mudanças em eventos. Até o momento, uma comparação exaustiva tem sido realizada entre ambas abordagens para mostrar as vantagens e desvantagens destas, porém, há a possibilidade de usar ambas em conjunto para aprimorar os resultados por três motivos: O primeiro motivo deve-se ao ST-DBSCAN conseguir detectar o ruido com maior acurácia, desde que ele mensura as distâncias a nível de eventos e não celas. O segundo motivo encontra-se no fato de o método cronológico lidar melhor em conjuntos de dados sem ruido, e por último, por ambos métodos não serem concorrentes entre si, como mostrado nas análises.

Figura 7.1 - Processo híbrido usando os métodos STDBSCAN e Cronológico



A) Lista de eventos espaciais e temporais. B) Identificação de eventos que fazem parte do ruído utilizando o método ST-DBSCAN. C) Construção da rede usando o método cronológico. D) Simplificação ou otimização da rede.

Fonte: Produção do autor.

A ideia por trás desta abordagem seria usar, num primeiro momento, o método ST-DBSCAN como forma de eliminar o ruído, e depois, prosseguir com o método cronológico para encontrar os padrões de transição nos conjunto de dados. Na Figura 7.1 ilustra-se este processo de forma resumida.

Apesar do potencial uso de ambas numa única abordagem, há certas considerações a fazer. Primeiro, a diversidade de combinações que o método ST-DBSCAN possui pode levar à remoção de eventos que não fazem parte do ruido. Como observados nas avaliações anteriores, este método é sensível aos parâmetros usados e tem influência na quantidade de grupos encontrados, no entanto, como o foco não é usar este método como detecção de agrupamentos e sim para remoção de ruídos, é possível que seja possível estabelecer alguma regra ou processo para encontrar parâmetros que permitam remover o ruído de forma otimizada. Nesse sentido, ainda será necessário fazer a análise de sensibilidade dos parâmetros na detecção correta do ruído, sua remoção e impacto na detecção das mudanças nos eventos.

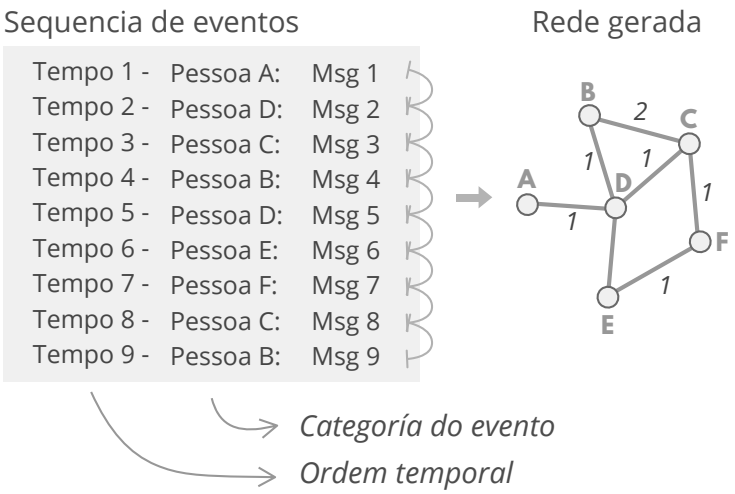
7.2 Eventos temporais sem atributo espacial

Nesta abordagem o desafio encontra-se em lidar com eventos temporais que não têm componente espacial, o que representa um grande desafio para o método,

desde que ele precisa desse atributo para gerar as celas e converti-las em nós. Por isso, é preciso adaptar o método para encontrar os padrões de transição neste tipo de dados.

Apenas para recapitular, como mencionado no capítulo sobre dados temporais (4), existem três tipos de dados temporais amplamente aceitos na literatura: eventos, sequências e series temporais. Nesse sentido, ao termos um conjunto de eventos temporais sem saber sua localização espacial, é possível que estes sejam transformados em sequências temporais. Por exemplo, no registro de entrada e saída dos funcionários de um prédio, cada entrada ou saída representa um evento, porém, como é possível agrupá-los por categorias (por funcionário ou hierarquia), conseguimos formar sequências temporais. Portanto, uma possibilidade de adaptação do método cronológico seria substituir as celas (espaciais) pelas categorias (das sequências), e transformar estas em nós. Desse modo, o restante do processo de construção da rede continuaria da mesma forma, ligando os nós de forma consecutiva, segundo a ordem temporal dos eventos. Na Figura 7.2 observa-se um exemplo de este processo de construção da rede.

Figura 7.2 - Processo de construção da rede



No lado esquerdo, a sequência de eventos ordenados segundo o tempo, e no lado direito, a rede gerada e simplificada a partir da sequência temporal dos eventos. Os nós representam as categorias e as ligações a respectiva sequência de eventos (mensagens) entre as categorias (pessoas). Também, os pesos das ligações são proporcionais à quantidade de ligações que foram simplificadas entre dois nós.

Fonte: Produção do autor.

A partir da rede, seria possível continuar com o método cronológico, aplicando a detecção de comunidades. Porém, neste ponto é necessário avaliar duas dificuldades: a primeira encontra-se em não ser possível comparar com o método ST-DBSCAN, considerando que não temos atributo espacial, em segundo lugar, mesmo que fosse identificado as comunidades, estas não representariam mudanças temporais, desde que, neste caso, os nós não tem uma posição espacial ou algum outro atributo além do tempo. Em outras palavras, não temos algum atributo ou característica que muda em relação ao tempo. Enquanto que, ao considerar o componente espacial, é possível mensurar as interseções de conjuntos de dados em relação ao tempo e espaço. Por isso, é necessário muito cuidado realizar cálculos sobre as redes sem termos identificado o seu verdadeiro significado.

Apesar de, por enquanto, não ser possível dar continuidade ao método cronológico, a rede construída tem utilidade, pois a partir das propriedades dela podemos inferir ou encontrar informações relacionadas às categorias (nós) das sequências temporais. Nesse sentido, ao procurar potenciais aplicações e uso desta rede cro-

nológica, escolheu-se abordar um desafio atual relacionado ao envio de mensagens em grupos, sem fazer uso do conteúdo delas, mas apenas sabendo *quem* (usuário) e o *quando* (data e hora) foram enviadas as mensagens. O que permitiria analisar grupos fechados ou privados, desde que apenas seria necessário identificar o remitente ou identificador do membro do grupo, e o registro do horário que foi enviado a mensagem, sem ler o conteúdo.

Nesse sentido, a partir dos dados de cinco grupos de WhatsApp (um aplicativo de mensagens popular em diversos países), obteve-se redes cronológicas considerando os membros dos grupos como nós e construindo as ligações entre estes com base nas mensagens consecutivas, seguindo um processo similar à Figura 7.2. A partir destas redes foi possível identificar mudanças no engajamento dos usuários em diversos períodos de tempo, como as eleições presidenciais no ano 2018. Também, foi gerado um índice de engajamento para cada usuário dos grupos de mensagens, usando apenas propriedades da rede. Com essa informação, preparou-se um ranking dos mais engajados e sua evolução ao longo do tempo. Os resultados desta pesquisa foram reunidas na publicação "*Measuring the engagement level in encrypted group conversations by using temporal networks*" que recebeu o aceite para apresentação no evento internacional "*IJCNN 2020: International Joint Conference on Neural Networks*" em Julho de 2020.

Espera-se que este desdobramento do método cronológico (para sequências temporais) seja ampliado para outra áreas com similar estrutura de dados e aplicações. Também, está em planejamento o desenvolvimento de uma versão aprimorada deste método, com ênfase na flexibilidade.

7.3 Considerações finais

A abordagem cronológica e os métodos propostos abre novas possibilidades a outras aplicações e cenários, a partir das limitações que os métodos propostos possuem atualmente. Como é mencionado anteriormente, o ruído representa um dos gargalos que faz o método cronológico recorrer a ter mais parâmetros (μes) para diminuir os impactos do ruído na correta detecção de transições ou mudanças. Nesse sentido, parece promissória a possibilidade de remover ou diminuir o ruído antes de construir a rede complexa.

No caso de eventos temporais sem componente espacial, o uso da abordagem cronológica para construir a rede e obter novas informações parece promissória, desde que ela permite o uso da bagagem científica das redes complexas na minera-

ção dos dados. Além disso, também é possível a construção de novas medidas que sejam, especificamente, para redes construídas a partir de eventos temporais. Finalmente, foi apresentado o desdobramento do método cronológico neste sentido, juntamente com a criação de uma nova medida para mensurar o engajamento dos usuários em conversas de grupo. A partir da nova medida também obteve-se o engajamento global, ou do grupo todo, e desse modo, mensurar o desempenho do engajamento grupal.

8 CONCLUSÕES

Neste trabalho reúne-se o desenvolvimento da abordagem cronológica e os métodos ou variações para encontrar padrões de transição em eventos temporais. Este conjunto de métodos estão baseados, principalmente, na sequência cronológica (temporal e espacial) dos eventos.

Em relação aos objetivos propostos neste trabalho, estes tratam sobre a identificação dos métodos desenvolvidos na atualidade para explorar dados, temporais e espaciais, e como transformar estes em redes complexas. Também, considerou-se o desenvolvimento de métodos para identificar transições que acontecem ao longo do tempo nestes conjuntos de dados, com ênfase no uso das redes complexas.

Nesse sentido, ao fazer a revisão sistemática dos métodos de mineração de dados para encontrar grupos ou padrões, encontrou-se principalmente, métodos de agrupamento e classificação. Destes dois, o de agrupamento é o que mais se aproxima para atender os objetivos desta pesquisa, por não precisar fazer uma preclassificação dos dados. Por outro lado, no conjunto de métodos para transformar dados temporais em redes complexas, verificou-se que estes focam, principalmente, em séries temporais. Portanto, a partir destes fatos, mostrou-se evidente a necessidade do desenvolvimento de métodos que identifiquem as transições entre grupos, em conjunto de evento temporais.

A seguir, é introduzido um breve resumo dos principais resultados obtidos ao longo da pesquisa de doutorado. Também, são mencionados as principais conclusões e potenciais trabalhos futuros.

8.1 Principais conclusões

Baseado no apresentado ao longo deste trabalho, as principais conclusões se resumem nestes pontos:

- a abordagem cronológica apresenta-se como um conjunto de métodos bem sucedidos para transformar eventos (temporais e espaciais) em redes complexas. Devido à simplicidade da abordagem, é fácil construir outros métodos a partir dela. Além disso, ela não está restrita a apenas eventos espaciais e temporais, mas também permite construir redes a partir de eventos sim componente espacial.
- o diferencial dos métodos propostos ao comparar com outros métodos

disponíveis na literatura encontra-se em dois pontos: 1) mostram informações sobre as transições ou mudanças significativas que existem num conjunto de eventos temporais, o que não é mostrado por nenhum outro método 2) precisam de poucos parâmetros para funcionar, o que permite que seja facilmente calibrado em diversas situações.

- verificou-se que os resultados dos três métodos mostram diferentes características dos dados, além de serem complementares. O primeiro método (ligação única) é simples e mostra as mudanças como esperado, no entanto, é altamente sensível ao ruído. O segundo método (ligações múltiplas), por aumentar a quantidade de ligações na rede, fortalece as comunidades deixando-as mais conectadas, e portanto, mostra de melhor forma os estágios dos eventos. O terceiro método (remoção de nós), destaca-se por diminuir consideravelmente a interferência do ruido e mostrar comunidades mais coesas.

- foi evidenciado a versatilidade da abordagem cronológica em problemas com dados reais. Em particular, ao usá-lo na detecção de regiões com queimadas, que possuem características similares, na Bacia Amazônica. Mesmo que apenas tivessem sido fornecidos dados sobre a localização e tempo de cada queimada, foi possível agrupar áreas que, em muitos casos, coincidiram com áreas em desmatamento, de agricultura, entre outros. Além disso, estes resultados demonstram a capacidade da abordagem por permitir condensar informações de 15 anos em apenas uma rede e suas comunidades.

- ao comparar o método ST-DBSCAN e os métodos desenvolvidos nesta pesquisa, mostrou-se que ambas as abordagem apresentam resultados complementares e com diferentes objetivos. O método ST-DBSCAN tem como foco encontrar eventos similares baseado na densidade espacial e temporal destas. Por outro lado, os métodos propostos trazem informações sobre as transições temporais e espaciais que existem nos eventos. Desse modo, foi mostrado a possibilidade de uma única agrupação de eventos ter diversas transições.

- ambas abordagens mostraram suas diferenças ao usá-las em modelos de eventos com características diferentes. No primeiro caso, ficou claro que o método ST-DBSCAN não consegue identificar, de forma clara, as agrupações que existem num cruzamento de dois grupos de eventos, encon-

trando apenas, um grupo principal. Por outro lado, a abordagem cronológica identificou corretamente as transições nesta dinâmica de eventos, mostrando o antes, durante e depois da transição, mesmo que esta tivesse ruido. No segundo caso, o método ST-DBSCAN identificou claramente dois grupos, como era previsto, no modelo onde dois grupos de eventos se deslocam e mudam gradativamente a quantidade de eventos. Neste caso, o método cronológico, identificou corretamente as mudanças gradativas, mesmo com a presença de ruido. Finalmente, no último modelo, que corresponde à caminhada aleatória de Brown, o método ST-DBSCAN não chegou a um resultado consistente, identificando diversas agrupações dependendo dos parâmetros usados. Por outro lado, o método cronológico encontrou as diversas regiões e transições pelas que o caminhante aleatório se deslocou ao longo do tempo. Mesmo mudando de parâmetros, as regiões encontradas são muito similares entre si.

- a partir das análises com modelos artificiais, foi identificado a relação que existe entre a quantidade de ligações múltiplas e a melhora na identificação das comunidades, porém, esta relação também está ligada ao tamanho da grade, em alguns casos. Nesse sentido, é recomendável estabelecer o tamanho da grade respeitando a regra de $w \sim \sqrt{n/3}$, onde w é a quantidade de divisões em cada lado da grade.

- o método cronológico, apesar de conseguir diminuir o ruído fazendo uso de alguns procedimentos, é sensível à influência deste em alguns casos, tendo dificuldade para limpar completamente a presença destes eventos ou limpando excessivamente ao ponto de remover nós sem ruido.

- em casos onde os eventos não tem componente espacial, a abordagem cronológica consegue fornecer uma nova forma de explorar os dados a partir das redes complexas. Desse modo, é possível desenvolver novas medidas com fins específicos para estes casos, como foi mostrado com o *índice* ou medida construído para mensurar o engajamento em grupos de bate-papo.

8.2 Trabalhos futuros

Apesar do modelo apresentado trazer muitas oportunidades e potenciais usos, há muito para ser explorado futuramente. Nesse sentido, alguns pontos se destacam para o desenvolvimento de futuros trabalho:

- o método cronológico, neste trabalho, apenas foi explorado desde a perspectiva de comunidades. Porém, na ciência de redes há muitos outros mecanismos que permitiriam aprofundar as descobertas de novas propriedades em este tipo de redes, assim como compreender no quê elas podem contribuir na procura de novas informações a partir dos dados temporais.

- a integração deste método com outros, seja da ciência de redes ou da mineração de dados, é uma área que não foi tratada neste trabalho, o que representa uma grande oportunidade considerando que, por estúdios preliminares apresentados aqui, parecem ser promissórios algumas destas abordagens. Como a integração do método cronológico com o ST-DBSCAN, ou do método cronológico com o método de redes por correlação.

- a criação de uma abordagem ainda mais abrangente, que permita a consideração de uma ou três dimensões espaciais. Também, adicionar ao método a flexibilidade para considerar ou não atributos espaciais, dessa forma, o método poderia ser aplicado em, praticamente, qualquer conjunto de eventos temporais.

REFERÊNCIAS BIBLIOGRÁFICAS

AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering - a decade review. **Information Systems**, v. 53, p. 16–38, 2015. ISSN 03064379. 51

ANDA-JÁUREGUI, G. de; ESPINAL-ENRÍQUEZ, J.; DRAGO-GARCÍA, D.; HERNÁNDEZ-LEMUS, E. Nonredundant, highly connected MicroRNAs control functionality in breast cancer networks. **International Journal of Genomics**, v. 2018, p. 1–10, maio 2018. ISSN 23144378. Disponível em: <https://doi.org/10.1155/2018/9585383>. 31

ANSARI, M. Y.; AHMAD, A.; KHAN, S. S.; BHUSHAN, G.; MAINUDDIN. Spatiotemporal clustering: a review. **Artificial Intelligence Review**, v. 53, n. 4, p. 2381–2423, jul. 2019. Disponível em: <https://doi.org/10.1007/s10462-019-09736-1>. 47

ANTUNES, C. M.; OLIVEIRA, A. L. Temporal data mining : an overview. **Lecture Notes in Computer Science**, p. 1–15, 2001. 1, 2, 35, 41

APACHEVIEWER. **Apache logs viewer**. 2020. Disponível em: <https://www.apacheviewer.com>. 40

ARMBRUST, M.; STOICA, I.; ZAHARIA, M.; FOX, A.; GRIFFITH, R.; JOSEPH, A. D.; KATZ, R.; KONWINSKI, A.; LEE, G.; PATTERSON, D.; RABKIN, A. A view of cloud computing. **Communications of the ACM**, v. 53, n. 4, p. 50, apr 2010. ISSN 00010782. Disponível em: <http://portal.acm.org/citation.cfm?doid=1721654.1721672>. 1

ATLURI, G.; KARPATNE, A.; KUMAR, V. Spatio-temporal data mining. **ACM Computing Surveys**, v. 51, n. 4, p. 1–41, set. 2018. Disponível em: <https://doi.org/10.1145/3161602>. 36, 37, 40, 45

AYNAUD, T.; BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R. Multilevel local optimization of modularity. In: **Graph Partitioning**. John Wiley & Sons, Inc., 2013. p. 315–345. Disponível em: <https://doi.org/10.1002/9781118601181.ch13>. 27

BALABAN, A. T. Applications of graph theory in chemistry. **Journal of Chemical Information and Modeling**, v. 25, n. 3, p. 334–343, ago. 1985. Disponível em: <https://doi.org/10.1021/ci00047a033>. 8

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, out. 1999. Disponível em: <https://doi.org/10.1126/science.286.5439.509>. 22, 23

BARABÁSI, A.-L.; PÓSFAI, M. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266 1107076269. Disponível em: <http://barabasi.com/networksciencebook/>. 9, 10, 12, 19, 20, 22, 23, 24, 25

BARNES, J.; HARARY, F. Graph theory in network analysis. **Social Networks**, v. 5, n. 2, p. 235–244, jun. 1983. Disponível em: <https://doi.org/10.1016/0378-8733(83)90026-6>. 8, 9

BARTHÉLEMY, M. Spatial networks. **Physics Reports**, v. 499, n. 1-3, p. 1–101, fev. 2011. Disponível em: <https://doi.org/10.1016/j.physrep.2010.11.002>. 31

BATAGELJ, V.; PRAPROTNIK, S. An algebraic approach to temporal network analysis based on temporal quantities. **Social Network Analysis and Mining**, v. 6, n. 1, maio 2016. Disponível em: <https://doi.org/10.1007/s13278-016-0330-4>. 50

BENSON, A. R.; GLEICH, D. F.; LESKOVEC, J. Higher-order organization of complex networks. **Science**, v. 353, n. 6295, p. 163–166, jul. 2016. Disponível em: <https://doi.org/10.1126/science.aad9029>. 55

BETZEL, R. F.; BASSETT, D. S. Multi-scale brain networks. **NeuroImage**, v. 160, p. 73–83, out. 2017. Disponível em: <https://doi.org/10.1016/j.neuroimage.2016.11.006>. 31

BLONDEL, V. D.; GUILLAUME, J. L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. 1–12, 2008. ISSN 17425468. 26, 27

BRANDES, U.; ERLEBACH, T. **Network analysis: methodological foundations**. [S.l.: s.n.], 2005. ISSN 03623319. ISBN 3540249796. 9

BROIDO, A. D.; CLAUSET, A. Scale-free networks are rare. **Nature Communications**, v. 10, n. 1, mar. 2019. Disponível em: <https://doi.org/10.1038/s41467-019-08746-5>. 22, 33

CALDERELLI, G. **Scale-free networks**. [S.l.]: Oxford University Press, 2007. ISBN 9780199211517. 13, 22

CAMPANHARO, A. S. L. O.; SIRER, M. I.; MALMGREN, R. D.; RAMOS, F. M.; AMARAL, L. A. N. Duality between time series and networks. **PloS one**, v. 6, n. 8, p. e23378, jan 2011. ISSN 1932-6203. Disponível em: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023378>. 56

CHAKRABORTY, T.; DALMIA, A.; MUKHERJEE, A.; GANGULY, N. Metrics for community analysis: a survey. **ACM Computing Surveys**, v. 50, n. 4, p. 1–37, 2016. ISSN 15577341. 26

CHAU, D. H. P. **Data mining meets HCI: making sense of large graphs**. 169 p. Tese (Doutorado) — Carnegie Mellon University, Pittsburgh, 2012. 2

CHEN, C. L. P.; ZHANG, C. Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data. **Information Sciences**, v. 275, p. 314–347, 2014. ISSN 00200255. 1

CHEN, D.; ZHAO, H. Data security and privacy protection issues in cloud computing. In: **INTERNACTIONAL CONFERENCE ON COMPUTER SCIENCE AND ELECTRONIS ENGINEERING, 2012. Procedings...** [S.l.: s.n.], 2012. ISBN 9780769546476. 1

CHEN, Z.; JI, H. Graph-based clustering for computational linguistics: a survey. In: **WORKSHOP ON GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010. Proceedings...** Uppsala, Sweden: Association for Computational Linguistics, 2010. p. 1–9. Disponível em: <https://www.aclweb.org/anthology/W10-2301>. 8

CHITTARO, L.; MONTANARI, A. Temporal representation and reasoning in artificial intelligence: issues and approaches. **Annals of Mathematics and Artificial Intelligence**, v. 28, n. 1/4, p. 47–106, 2000. ISSN 10122443. 35, 36

COSTA, L. d. F.; OLIVEIRA, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, jun 2011. ISSN 0001-8732. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/00018732.2011.572452>. 2, 4

COSTA, L. D. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. Characterization of complex networks: a survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, 2007. ISSN 00018732. 2

DARST, R. K.; GRANELL, C.; ARENAS, A.; GÓMEZ, S.; SARAMÄKI, J.; FORTUNATO, S. Detection of timescales in evolving complex systems. **Scientific Reports**, v. 6, p. 1–17, 2016. ISSN 20452322. 4

DIESNER, J.; FRANTZ, T. L.; CARLEY, K. M. Communication networks from the Enron email corpus it's always about the people. Enron is no different. **Computational and Mathematical Organization Theory**, v. 11, n. 3, p. 201–228, out. 2005. Disponível em: <https://doi.org/10.1007/s10588-005-5377-0>. 32

DONGES, J. F.; ZOU, Y.; MARWAN, N.; KURTHS, J. Complex networks in climate dynamics: comparing linear and nonlinear network construction methods. **European Physical Journal: Special Topics**, v. 174, n. 1, p. 157–179, jul 2009. ISSN 19516355. Disponível em: <http://www.springerlink.com/index/10.1140/epjst/e2009-01098-2>. 32, 51, 52

DONNAT, C.; HOLMES, S. Tracking network dynamics: a survey using graph distances. **Annals of Applied Statistics**, v. 12, n. 2, p. 971–1012, jun 2018. ISSN 19417330. 3

DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. k-core organization of complex networks. **Physical Review Letters**, v. 96, n. 4, fev. 2006. ISSN 10797114. Disponível em: <https://doi.org/10.1103/physrevlett.96.040601>. 74

DU, W. B.; ZHOU, X. L.; LORDAN, O.; WANG, Z.; ZHAO, C.; ZHU, Y. B. Analysis of the chinese airline network as multi-layer networks. **Transportation Research Part E: Logistics and Transportation Review**, v. 89, p. 108–116, maio 2016. ISSN 13665545. Disponível em: <https://doi.org/10.1016/j.tre.2016.03.009>. 32

DUNHAM, M. H. **Data mining: introductory and advanced topics**. USA: Prentice Hall PTR, 2002. ISBN 0130888923. 35, 47, 52

ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. **Mathematical Institute of the Hungarian Academy of Sciences**, v. 5, p. 17–61, 1960. 17

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2, 1996. Proceedings...** [S.l.: s.n.], 1996. p. 226–231. 44

ESTRADA, E. Quantifying network heterogeneity. **Physical Review E - Statistical, Nonlinear, and Soft Matter Physics**, v. 82, n. 6, p. 1–8, 2010. ISSN 15393755. 9, 20

_____. **The structure of complex networks: theory and applications**. USA: Oxford University Press, 2011. ISBN 019959175X. 11, 23, 33

EULER, L. Leonhard Euler and the Koenigsberg Bridges. **Scientific American**, v. 189, n. 1, p. 66–70, jul 1953. ISSN 0036-8733. Disponível em: <http://www.nature.com/doifinder/10.1038/scientificamerican0753-66>. 7, 8

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis. **National Science Review**, v. 1, n. 2, p. 293–314, jun 2014. ISSN 2053-714X. Disponível em: <https://academic.oup.com/nsr/article/1/2/293/1397586>. 1, 4, 24

FAN, J.; MENG, J.; ASHKENAZY, Y.; HAVLIN, S.; SCHELLNHUBER, H. J. Network analysis reveals strongly localized impacts of El Niño. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 29, p. 7543–7548, jul 2017. ISSN 1091-6490. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28674008http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5530664>. 32

FENG, C.; HE, B. Construction of complex networks from time series based on the cross correlation interval. **Open Physics**, v. 15, n. 1, p. 253–260, 2017. 51

Folha de São Paulo. **Veja cronologia dos fatos que marcaram a Segunda Guerra Mundial**. 2009. Disponível em: <https://m.folha.uol.com.br/mundo/2009/09/617745-veja-cronologia-dos-fatos-que-marcaram-a-segunda-guerra-mundial.shtml.>. 37

FORTUNATO, S.; FLAMMINI, A.; MENCZER, F. Scale-free network growth by ranking. **Physical Review Letters**, v. 96, n. 21, p. 1–4, 2006. ISSN 00319007. 22

GAN, G.; MA, C.; WU, J. **Data clustering: theory, algorithms, and applications**. [S.l.: s.n.], 2007. 42, 43

GAO, J.; BARZEL, B.; BARABÁSI, A. L. Universal resilience patterns in complex networks. **Nature**, v. 530, n. 7590, p. 307–312, fev. 2016. ISSN 14764687. Disponível em: <https://doi.org/10.1038/nature16948>. 33

GARGIULO, F.; CAEN, A.; LAMBIOTTE, R.; CARLETTI, T. The classical origin of modern mathematics. **EPJ Data Science**, v. 5, n. 1, p. 1–18, 2016. ISSN 21931127. 30

GASTNER, M. T.; NEWMAN, M. E. The spatial structure of networks. **The European Physical Journal B**, v. 49, n. 2, p. 247–252, jan. 2006. Disponível em: <https://doi.org/10.1140/epjb/e2006-00046-8>. 31

GHOSH, S.; GANGULY, N. Structure and evolution of online social networks. **Intelligent Systems Reference Library**, 2014. ISSN 18684408. 32

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, v. 99, n. 12, p. 7821–7826, jun. 2002. Disponível em: <https://doi.org/10.1073/pnas.122653799>. 25

GÓMEZ, S.; DOMENICO, M. D.; OMODEI, E.; ALBERT, S. R.; ARENAS, A. Multilayer networks. In: BATTISTON, S.; CALDARELLI, G.; GARAS, A. (Ed.). **Multiplex and multilevel networks**. Oxford University Press, 2018. p. 1–30. Disponível em: <https://doi.org/10.1093/oso/9780198809456.003.0001>. 33

GOZOLCHIANI, A.; HAVLIN, S.; YAMASAKI, K. Emergence of El Niño as an autonomous component in the climate network. **Physical Review Letters**, v. 107, n. 14, p. 148501, sep 2011. ISSN 0031-9007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22107243https://link.aps.org/doi/10.1103/PhysRevLett.107.148501>. 32

HAN, J.; KAMBER, M.; PEI, J. Graph mining, social network analysis, and multirelational data mining. In: KAUFMANN, M. (Ed.). **Data mining: concepts and techniques**. [S.l.: s.n.], 2006. cap. 9, p. 535–589. ISBN 1558609016. 2, 49

HARARY, F. **Graph theory**. [S.l.: s.n.], 2018. ISSN 2073-8994. ISBN 9780429962318. 8

HE, H.; YAN, J. Cyber-physical attacks and defences in the smart grid: a survey. **IET Cyber-Physical Systems: Theory & Applications**, v. 1, n. 1, p. 13–27, dez. 2016. ISSN 2398-3396. Disponível em: <https://doi.org/10.1049/iet-cps.2016.0019>. 32

HENDERSON, K.; GALLAGHER, B.; ELIASSI-RAD, T.; TONG, H.; BASU, S.; AKOGLU, L.; KOUTRA, D.; FALOUTSOS, C.; LI, L. RolX: structural role extraction

& mining in large graphs. In: **INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 18., 2012. Proceedings...** [S.l.: s.n.], 2012. p. 1231–1239. ISBN 9781450314626. 49

HLINKA, J.; HARTMAN, D.; JAJCAY, N.; TOMEČEK, D.; TINTĚRA, J.; PALUŠ, M. Small-world bias of correlation networks: from brain to climate. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, v. 27, n. 3, p. 035812, mar 2017. ISSN 1054-1500. Disponível em: <http://aip.scitation.org/doi/10.1063/1.4977951>. 51

HOLME, P. Modern temporal network theory: a colloquium. **European Physical Journal B**, v. 88, n. 9, 2015. ISSN 14346036. 3, 4, 33, 50, 58

_____. Temporal network structures controlling disease spreading. **Physical Review E**, v. 94, n. 2, p. 1–8, 2016. ISSN 24700053. 3, 51

HOLME, P.; SARAMÄKI, J. Temporal networks. **Physics Reports**, v. 519, n. 3, p. 97–125, 2012. ISSN 03701573. Disponível em: <http://dx.doi.org/10.1016/j.physrep.2012.03.001>. 2, 50

HSU, W. H.; LANCASTER, J.; PARADESI, M. S.; WENINGER, T. Structural link analysis from user profiles and friends networks: a feature construction approach. In: **INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2007. Proceedings...** [S.l.: s.n.], 2007. 25

HUFFMAN, D. A. A method for the construction of minimum-redundancy codes. **Proceedings of the IRE**, v. 40, n. 9, p. 1098–1101, set. 1952. ISSN 00968390. Disponível em: <https://doi.org/10.1109/jrproc.1952.273898>. 28

INVESTING.COM. **USD/BRL - Dólar Americano Real Brasileiro**. 2020. Disponível em: <https://br.investing.com/currencies/usd-brl-chart>. 39

JEBARA, T.; WANG, J.; CHANG, S. F. Graph construction and b-matching for semi-supervised learning. In: **INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 26., 2009.** Montreal, Canada: [s.n.], 2009. ISBN 9781605585161. 51

JIN, X.; WAH, B. W.; CHENG, X.; WANG, Y. Significance and challenges of big data research. **Big Data Research**, v. 2, n. 2, p. 59–64, 2015. ISSN 22145796. 1

KISILEVICH, S.; MANSMANN, F.; NANNI, M.; RINZIVILLO, S. Spatio-temporal clustering. In: MAIMON, O.; ROKACH, L. (Ed.). **Data mining and knowledge discovery handbook**. [S.l.]: Springer US, 2009. p. 855–874. 46, 47

KRAWCZYK, M. J. Communities in social networks. In: **INTERNATIONAL CONFERENCE ON BIOMETRICS AND KANSEI ENGINEERING, 2009. Proceedings...** [S.l.: s.n.], 2009. p. 111–116. ISBN 9780769536927. 26

KUNEGIS, J. **Handbook of network analysis**. [s.n.], 2014. ISBN 978-1-4503-2038-2. Disponível em: <http://arxiv.org/abs/1402.5500>. 16, 33, 49, 58

LACASA, L.; LUQUE, B.; BALLESTEROS, F.; LUQUE, J.; NUÑO, J. C. From time series to complex networks: the visibility graph. **Proceedings of the National Academy of Sciences**, v. 105, n. 13, p. 4972–4975, apr 2008. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/105/13/4972>. 52, 53

LAMBIOTTE, R.; ROSVALL, M.; SCHOLTES, I. From networks to optimal higher-order models of complex systems. **Nature Physics**, v. 15, n. 4, p. 313–320, 2019. ISSN 1745-2481. Disponível em: <https://doi.org/10.1038/s41567-019-0459-y>. 56

LATORA, V.; NICOSIA, V.; RUSSO, G. **Complex network: principles, methods and applications**. [S.l.]: Cambridge University Press, 2017. ISBN 9781107103184. 9, 11, 25, 26, 27, 28

LIU, L.; ÖZSU, T. **Encyclopedia of database systems**. [S.l.]: Springer US, 2009. ISBN 978-0-387-35544-3. 35

LIU, Y. Y.; SLOTINE, J. J.; BARABÁSI, A. L. Controllability of complex networks. **Nature**, v. 473, n. 7346, p. 167–173, 2011. ISSN 00280836. 31

LUDESCHER, J.; GOZOLCHIANI, A.; BOGACHEV, M. I.; BUNDE, A.; HAVLIN, S.; SCHELLNHUBER, H. J. Very early warning of next El Niño. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 6, p. 2064–6, feb 2014. ISSN 1091-6490. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24516172http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3926055>. 54

MALLIARIS, A. G. Wiener process. In: _____. **The new palgrave dictionary of economics**. London: Palgrave Macmillan UK, 2008. p. 7114–7115. ISBN 978-1-349-58802-2. Disponível em: <https://doi.org/10.1007/978-1-349-58802-2_1831>. 90

MANYIKA, J.; BROWN, M. C.; J., B. B.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. **Big data: the next frontier for innovation, competition and productivity**. [S.l.: s.n.], 2011. 24

MARCHIORI, M.; POSSAMAI, L. Micro-macro analysis of complex networks. **PLoS ONE**, v. 10, n. 1, p. 1–27, 2015. ISSN 19326203. 29, 30

MARSLAND, S. **Machine learning: an algorithmic perspective**. [S.l.: s.n.], 2014. ISBN 9781466583337. 41

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J. C. A survey of link prediction in complex networks. **ACM Computing Surveys**, v. 49, n. 4, p. 1–33, dec 2016. ISSN 03600300. Disponível em: <http://dx.doi.org/10.1145/3012704http://dl.acm.org/citation.cfm?doid=3022634.3012704>. 3

MARWAN, N.; DONGES, J. F.; ZOU, Y.; DONNER, R. V.; KURTHS, J. Complex network approach for recurrence analysis of time series. **Physics Letters A**, v. 373, n. 46, p. 4246–4254, 2009. ISSN 03759601. 51, 56

MASUDA, N.; LAMBIOTTE, R. **A guide to temporal networks**. World Scientific, 2016. Disponível em: <https://www.worldscientific.com/doi/abs/10.1142/q0033>. 9, 11, 12, 15, 16, 50, 58

MENG, J.; FAN, J.; ASHKENAZY, Y.; BUNDE, A.; HAVLIN, S. Forecasting the magnitude and onset of El Niño based on climate network. **New Journal of Physics**, v. 20, n. 4, p. 043036, apr 2018. ISSN 1367-2630. Disponível em: <http://stacks.iop.org/1367-2630/20/i=4/a=043036?key=crossref.9c9616cde36c0954b87878665a44b226>. 54

MICHAUT, M.; BARYSHNIKOVA, A.; COSTANZO, M.; MYERS, C. L.; ANDREWS, B. J.; BOONE, C.; BADER, G. D. Protein complexes are central in the yeast genetic landscape. **PLoS Computational Biology**, v. 7, n. 2, p. e1001092, 2011. ISSN 1553734X. 8

MITSA, T. **Temporal data mining**. Chapman and Hall/CRC, 2010. ISBN 9780429191855. Disponível em: <https://www.taylorfrancis.com/books/9781420089776>. 2, 35, 36, 38, 39, 40, 41, 42, 43, 44

NEEDHAM, M.; HODLER, A. E. **Graph algorithms in Neo4j: louvain modularity**. 2019. Disponível em:

<https://neo4j.com/blog/graph-algorithms-neo4j-louvain-modularity/>. 28

NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford University Press, 2010. 1–784 p. ISBN 9780191594175. 1, 8, 9, 10, 23, 25

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E - Statistical, Nonlinear, and Soft Matter Physics**, v. 69, n. 2, 2004. ISSN 1063651X. 25

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of Modern Physics**, v. 87, n. 3, p. 925–979, 2015. ISSN 15390756. 31

PIEDRAHITA, P.; BORGE-HOLTHOEFER, J.; MORENO, Y.; GONZáLEZ-BAILóN, S. The contagion effects of repeated activation in social networks. **Social Networks**, v. 54, p. 326–335, 2018. ISSN 03788733. 32

PORTA, S.; CRUCITTI, P.; LATORA, V. The network analysis of urban streets: a dual approach. **Physica A: Statistical Mechanics and its Applications**, v. 369, n. 2, p. 853–866, 2006. ISSN 03784371. 31

PRIGNANO, L.; MORER, I.; DIAZ-GUILERA, A. Wiring the past: a network science perspective on the challenge of archeological similarity networks. **Frontiers in Digital Humanities**, v. 4, 2017. ISSN 2297-2668. 33

RADEBACH, A.; DONNER, R. V.; RUNGE, J.; DONGES, J. F.; KURTHS, J.; RUNGE, J.; DONGES, J. F.; DONGES, J. F.; KURTHS, J.; KURTHS, J. Disentangling different types of El Niño episodes by evolving climate network analysis. **Physical Review E**, v. 88, n. 5, p. 052807, nov 2013. ISSN 1539-3755. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24329318https://link.aps.org/doi/10.1103/PhysRevE.88.052807>. 54

RATANAMAHATANA, C. A.; LIN, J.; GUNOPULOS, D.; KEOGH, E.; VLACHOS, M.; DAS, G. Mining time series data. In: MAIMON, O.; ROKACH, L. (Ed.). **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer US, 2009. p. 1049–1077. 41

RÄZ, T. Euler's Königsberg: the explanatory power of mathematics. **European Journal for Philosophy of Science**, v. 8, n. 3, p. 331–346, oct 2018. ISSN 18794920. 7, 8

RHEINWALT, A.; HOSKINS, B.; GOSWAMI, B.; BOOKHAGEN, B.; KURTHS, J.; BOERS, N. Complex networks reveal global pattern of extreme-rainfall teleconnections. **Nature**, v. 566, n. 7744, p. 373–377, 2019. ISSN 0028-0836. 32

RIBEIRO, H. V.; ALVES, L. G.; MARTINS, A. F.; LENZI, E. K.; PERC, M. The dynamical structure of political corruption networks. **Journal of Complex Networks**, v. 6, n. 6, p. 989–1003, 2018. ISSN 20511329. 33

ROMBACH, M. P.; PORTER, M. A.; FOWLER, J. H.; MUCHA, P. J. Core-periphery structure in networks. **SIAM Journal on Applied Mathematics**, v. 74, n. 1, p. 167–190, 2014. ISSN 00361399. 30

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences of the United States of America**, v. 105, n. 4, p. 1118–1123, 2008. ISSN 10916490. 26, 28, 29

SHI, Z.; PUN-CHENG, L. S. Spatiotemporal data clustering: a survey of methods. **ISPRS International Journal of Geo-Information**, v. 8, n. 3, p. 112, 2019. ISSN 22209964. 46, 51

SHOBANA, V.; KUMAR, N. Big data - a review. **International Journal of Applied Engineering Research**, v. 10, n. 55, p. 1294–1298, 2015. ISSN 09739769. 1, 4, 24

SILVA, T. C.; ZHAO, L. **Machine learning in complex networks**. [S.l.]: Springer, 2016. ISBN 9783319172897. 13, 14, 17, 20, 25, 50, 51, 56

SNODGRASS, R. T. Temporal databases. In: **INTERNATIONAL CONFERENCE ON SPATIAL AND TEMPORAL REASONING IN GEOGRAPHIC SPACE**. [S.l.: s.n.], 1992. p. 23–64. ISBN 9783540559665. ISSN 16113349. 35

STARK, D.; CASTELLS, M. **The rise of the network society**. [S.l.]: Wiley-Blackwell, 1997. 725 p. ISSN 00943061. 32

STEEN, M. van. **Graph theory and complex networks: an introduction**. [S.l.: s.n.], 2010. ISBN 9789081540612. 9, 12, 14, 16, 20, 21, 42, 50

STROGATZ, S. H. Exploring complex networks. **Nature**, v. 410, n. 6825, p. 268–276, mar. 2001. Disponível em: <https://doi.org/10.1038/35065725>. 8, 9, 12, 31

THELWALL, M.; KOUSHA, K. Academia.edu: social network or academic network. **Journal of the Association for Information Science and Technology**, v. 65, n. 4, p. 721–731, 2014. ISSN 23301643. 25

TOKIO2020. **Olympic competition schedule**. 2020. Disponível em: <https://tokyo2020.org/en/schedule>. 36

TSONIS, A. A.; SWANSON, K. L. Topology and predictability of El Niño and la Niña Networks. **Physical Review Letters**, v. 100, n. 22, p. 228502, jun 2008. ISSN 00319007. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.100.228502>. 54

TUNÇ, B.; VERMA, R. Unifying inference of meso-scale structures in networks. **PLoS ONE**, v. 10, n. 11, p. 1–14, 2015. ISSN 19326203. 30

VEGA-OLIVEROS, D. A.; QUILES, M. G.; COTACALLAPA, M.; ZHAO, L.; CARDOSO, M. F.; FERREIRA, L. N.; MACAU, E. E. From spatio-temporal data to chronological networks: an application to wildfire analysis. In: **ACM SYMPOSIUM ON APPLIED COMPUTING, 34., 2019**. [S.l.]: ACM, 2019. ISBN 9781450359337. 72, 76

WAGENSELLER, P.; WANG, F.; WU, W. Size matters: a comparative analysis of community detection algorithms. **IEEE Transactions on Computational Social Systems**, v. 5, n. 4, p. 951–960, 2018. ISSN 2329924X. 26

WANG, M.; WANG, A.; LI, A. Mining spatial-temporal clusters from geo-databases. In: LI, X.; ZAÏANE, O. R.; LI, Z. (Ed.). **Advanced data mining and applications**. Berlin: Springer, 2006. p. 263–270. ISBN 978-3-540-37026-0. 47, 48

WANG, P.; XU, B. W.; WU, Y. R.; ZHOU, X. Y. Link prediction in social networks: the state-of-the-art. **Science China Information Sciences**, v. 58, n. 1, p. 1–38, 2014. ISSN 1674733X. 32

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. In: NEWMAN M.; BARABASI, A. I. W. (Ed.). **The structure and dynamics of networks**. [S.l.: s.n.], 2011. v. 393, n. 6684, p. 440–442. ISBN 9781400841356. 20, 21, 22

WEY, T.; BLUMSTEIN, D. T.; SHEN, W.; JORDáN, F. Social network analysis of animal behaviour: a promising tool for the study of sociality. **Animal Behaviour**, v. 75, n. 2, p. 333–344, 2008. ISSN 00033472. 32

XU, J.; WICKRAMARATHNE, T. L.; CHAWLA, N. V. Representing higher-order dependencies in networks. **Science Advances**, v. 2, n. 5, p. e1600028, 2016. ISSN 23752548. 55, 56

XU, L.; JIANG, C.; WANG, J.; YUAN, J.; REN, Y. Information security in big data: privacy and data mining. **IEEE Access**, v. 2, p. 1149–1176, 2014. ISSN 21693536. 1

YAMASAKI, K.; GOZOLCHIANI, A.; HAVLIN, S. Climate networks around the globe are significantly affected by El Niño. **Physical Review Letters**, v. 100, n. 22, p. 228501, jun 2008. ISSN 0031-9007. 54

YANG, B.; LIU, J. Discovering global network communities based on local centralities. **ACM Transactions on the Web**, v. 2, n. 1, p. 1–32, 2008. ISSN 15591131. 25

YANG, Z.; ALGESHEIMER, R.; TESSONE, C. J. A comparative analysis of community detection algorithms on artificial networks. **Scientific Reports**, v. 6, n. e30750, 2016. ISSN 20452322. 26, 29

YU, F.; ZENG, A.; GILLARD, S.; MEDO, M. Network-based recommendation algorithms: a review. **Physica A: Statistical Mechanics and its Applications**, v. 452, p. 192–208, 2016. ISSN 03784371. 58

ZANIN, M.; PAPO, D.; SOUSA, P. A.; MENASALVAS, E.; NICCHI, A.; KUBIK, E.; BOCCALETTI, S. Combining complex networks and data mining: why and how. **Physics Reports**, v. 635, p. 1–44, 2016. ISSN 03701573. 2, 3, 4, 49

ZHANG, J.; SMALL, M. Complex network from pseudoperiodic time series: topology versus dynamics. **Physical Review Letters**, v. 96, p. 238701, Jun 2006. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.96.238701>. 52

ZHOU, D.; GOZOLCHIANI, A.; ASHKENAZY, Y.; HAVLIN, S. Teleconnection paths via climate network direct link detection. **Physical Review Letters**, v. 115, n. 26, p. 268501, dec 2015. ISSN 0031-9007. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.115.268501>. 32, 51, 54

ZINOVIEV, D. **Complex network analysis in python**. [S.l.]: The Pragmatic Bookshelf, 2018. ISBN 9781680502695. 11, 27

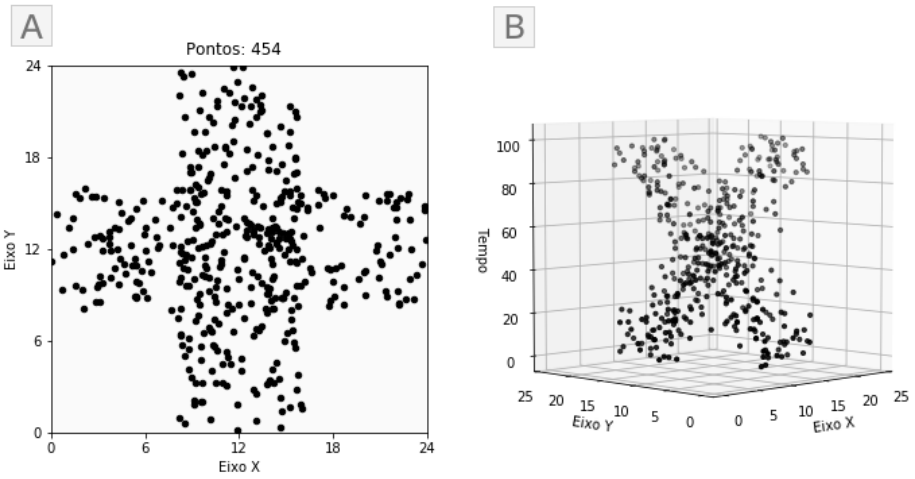
ZOU, Y.; DONNER, R. V.; MARWAN, N.; DONGES, J. F.; KURTHS, J. Complex network approaches to nonlinear time series analysis. **Physics Reports**, v. 787, p. 1–97, jan. 2019. Disponível em: <https://doi.org/10.1016/j.physrep.2018.10.005>. 52, 53, 54, 57

APÊNDICE A - COMPARAÇÃO DOS MÉTODOS ST-DBSCAN E CRONOLÓGICO SOBRE EVENTOS TEMPORAIS E ESPACIAIS

Neste apêndice detalham-se diversos resultados dos métodos ST-DBSCAN e Cronológico a partir do uso de diversos parâmetros.

A.1 Cruzamento de eventos

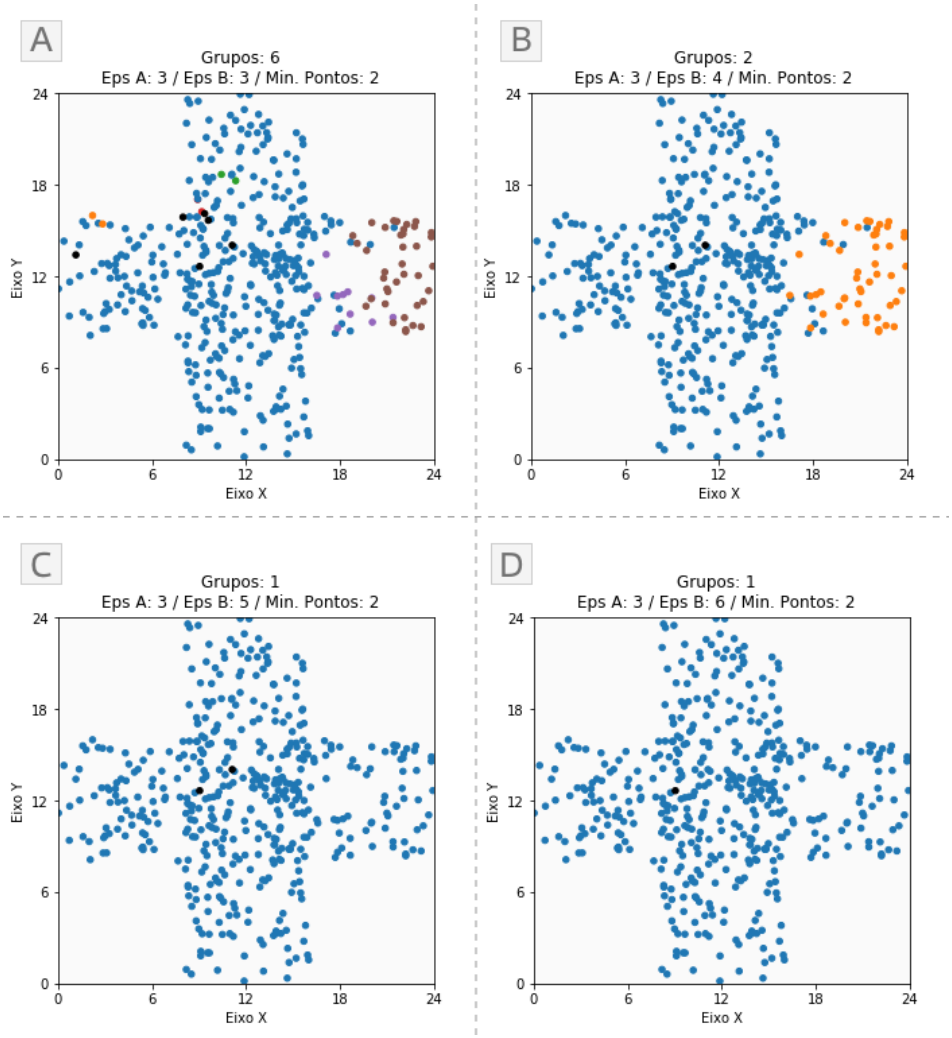
Figura A.1 - Modelo de cruzamento de eventos



A) 454 eventos gerados, composto por dois grupos se deslocando em formato cruzado. B) os mesmos eventos adicionando a dimensão temporal no eixo vertical.

Fonte: Produção do autor

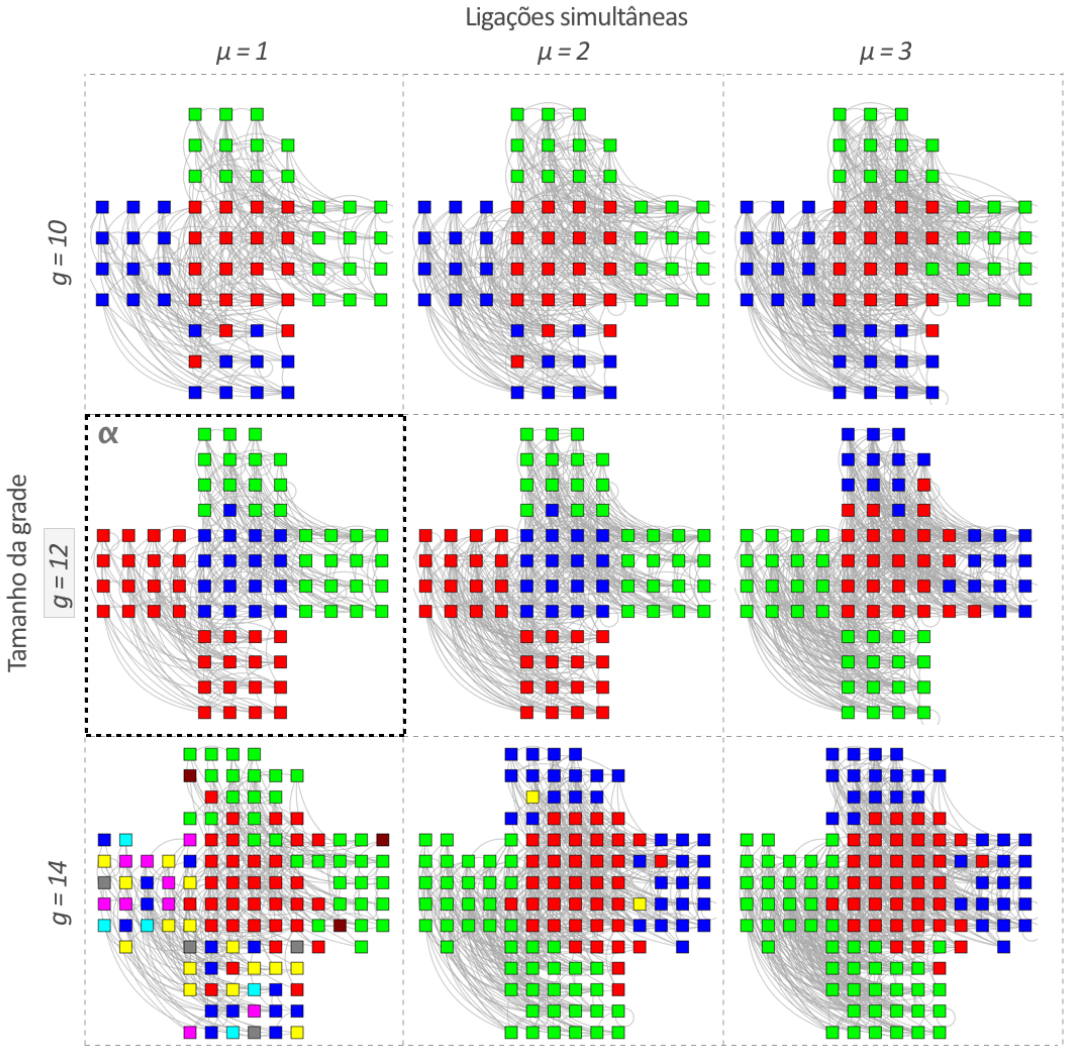
Figura A.2 - Cruzamento de eventos - Método ST-DBSCAN



Quatro imagens mostrando os grupos obtidos após aplicar o método ST-DBSCAN variando os parâmetros de entrada. Observe-se que, em todos os casos, um grupo é predominante (cor azul).

Fonte: Produção do autor

Figura A.3 - Cruzamento de eventos - Método cronológico

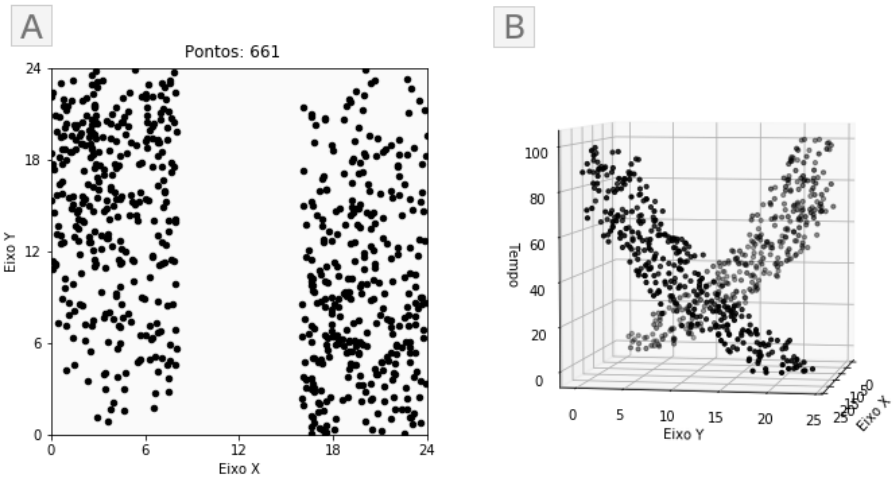


Nove imagens mostrando os resultados após aplicar o método cronológico, considerando entre 1 e 3 ligações simultâneas (μ), e tamanhos de grade (g) de 10, 12 e 14. α) Imagem mostrando o resultado com a configuração recomendado pelo método.

Fonte: Produção do autor

A.2 Eventos paralelos

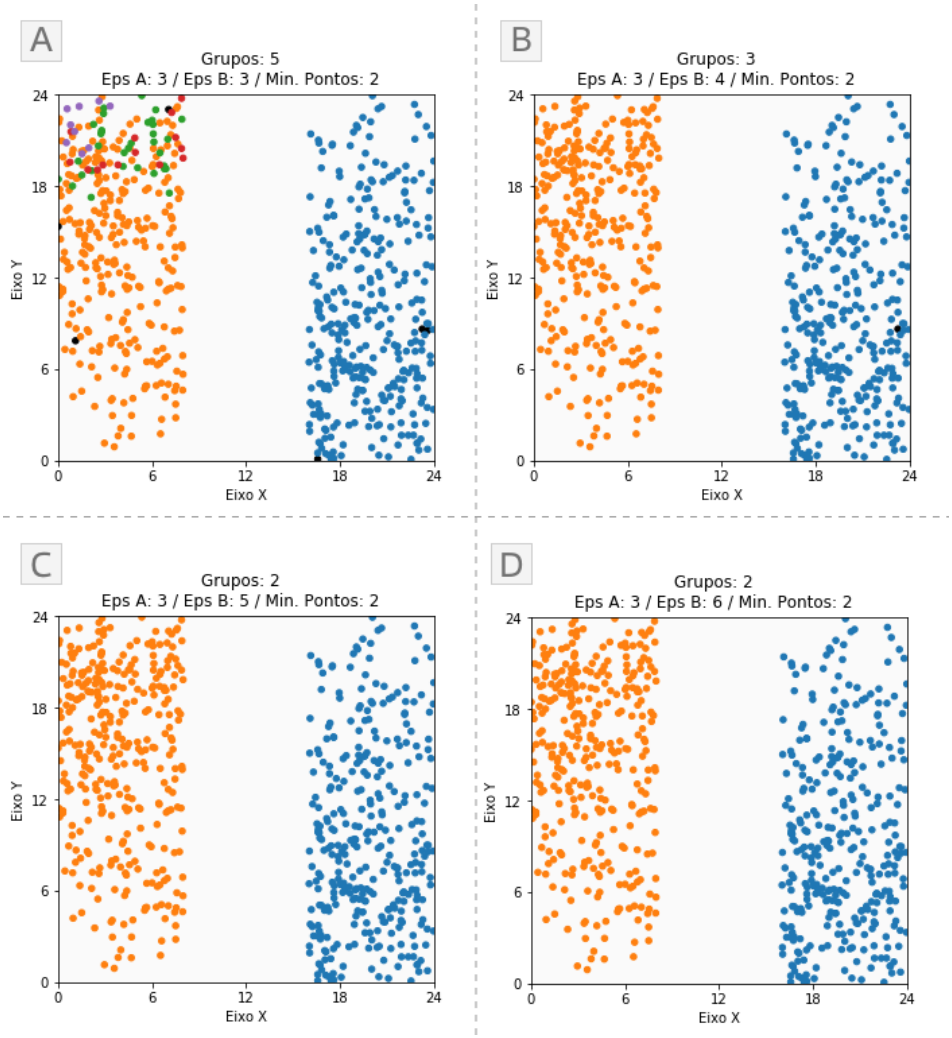
Figura A.4 - Modelo de eventos paralelos



A) 661 eventos separados em dois grupos se deslocando em direções opostas e mudando na quantidade de eventos ao longo do tempo. B) os mesmos eventos adicionando a dimensão temporal no eixo vertical.

Fonte: Produção do autor

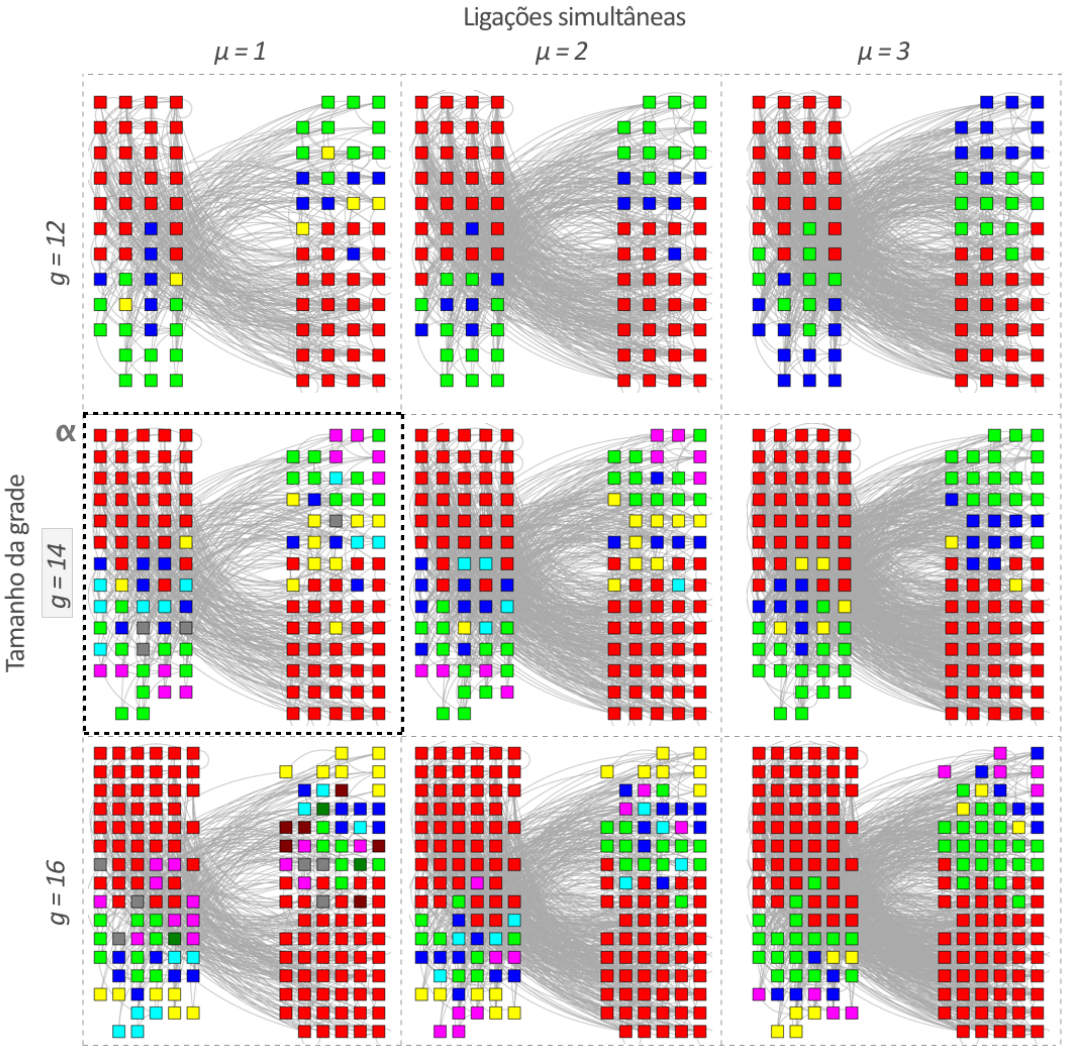
Figura A.5 - Eventos paralelos - Método ST-DBSCAN



Quatro imagens mostrando os dois conjuntos de eventos após usar o método ST-DBSCAN. De forma clara, o método encontra predominantemente dois grupos.

Fonte: Produção do autor

Figura A.6 - Eventos paralelos - Método cronológico

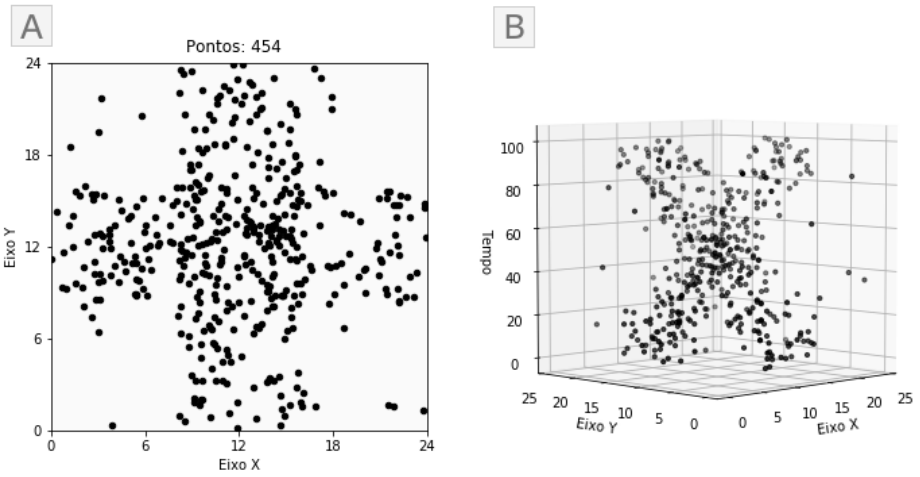


Nove imagens mostrando os resultados após aplicar o método cronológico no modelo de eventos paralelos, considerando entre 1 e 3 ligações simultâneas (μ), e tamanhos de grade (g) de 12, 14 e 16. α) Imagem mostrando o resultado com a configuração recomendado pelo método.

Fonte: Produção do autor

A.3 Cruzamento de eventos com ruido

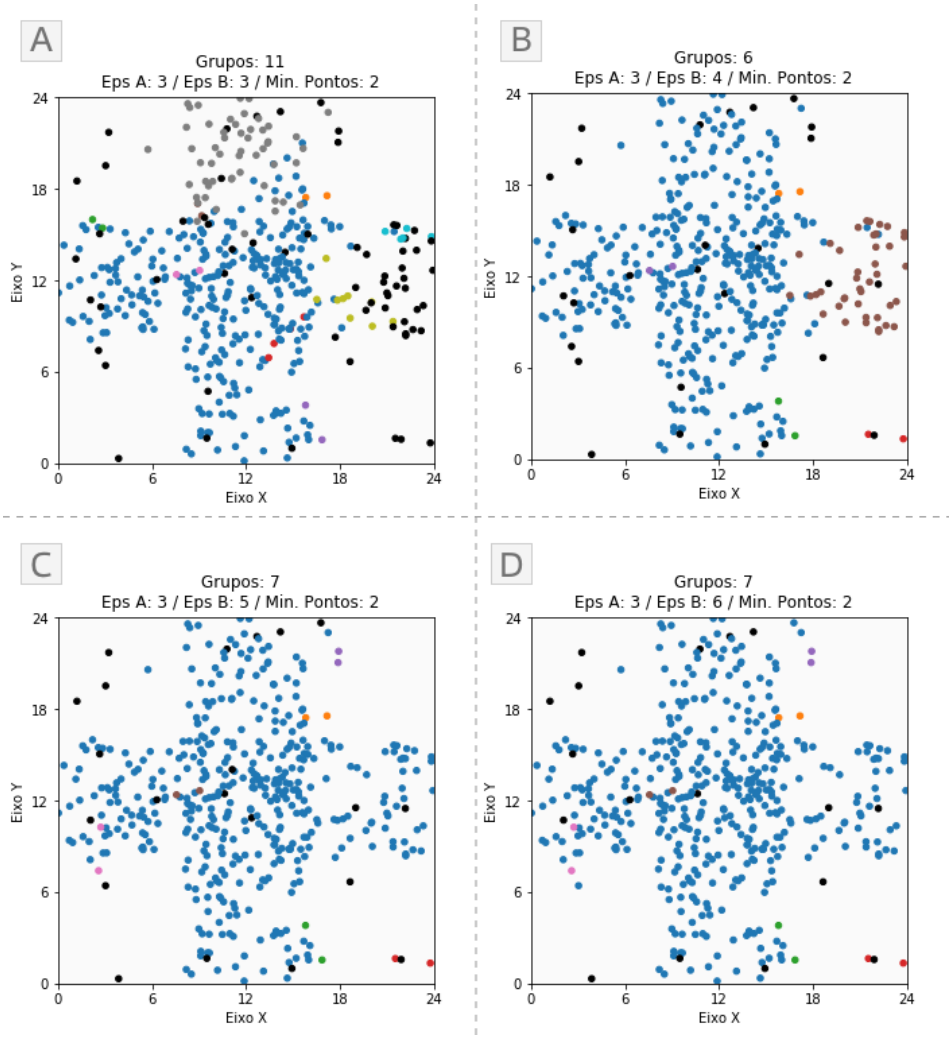
Figura A.7 - Modelo de cruzamento de eventos com ruido



A) 454 eventos gerados, composto por dois grupos se deslocando em formato cruzado, sendo 10% dos eventos removidos e adicionados de forma aleatória. B) os mesmos eventos adicionando a dimensão temporal no eixo vertical.

Fonte: Produção do autor

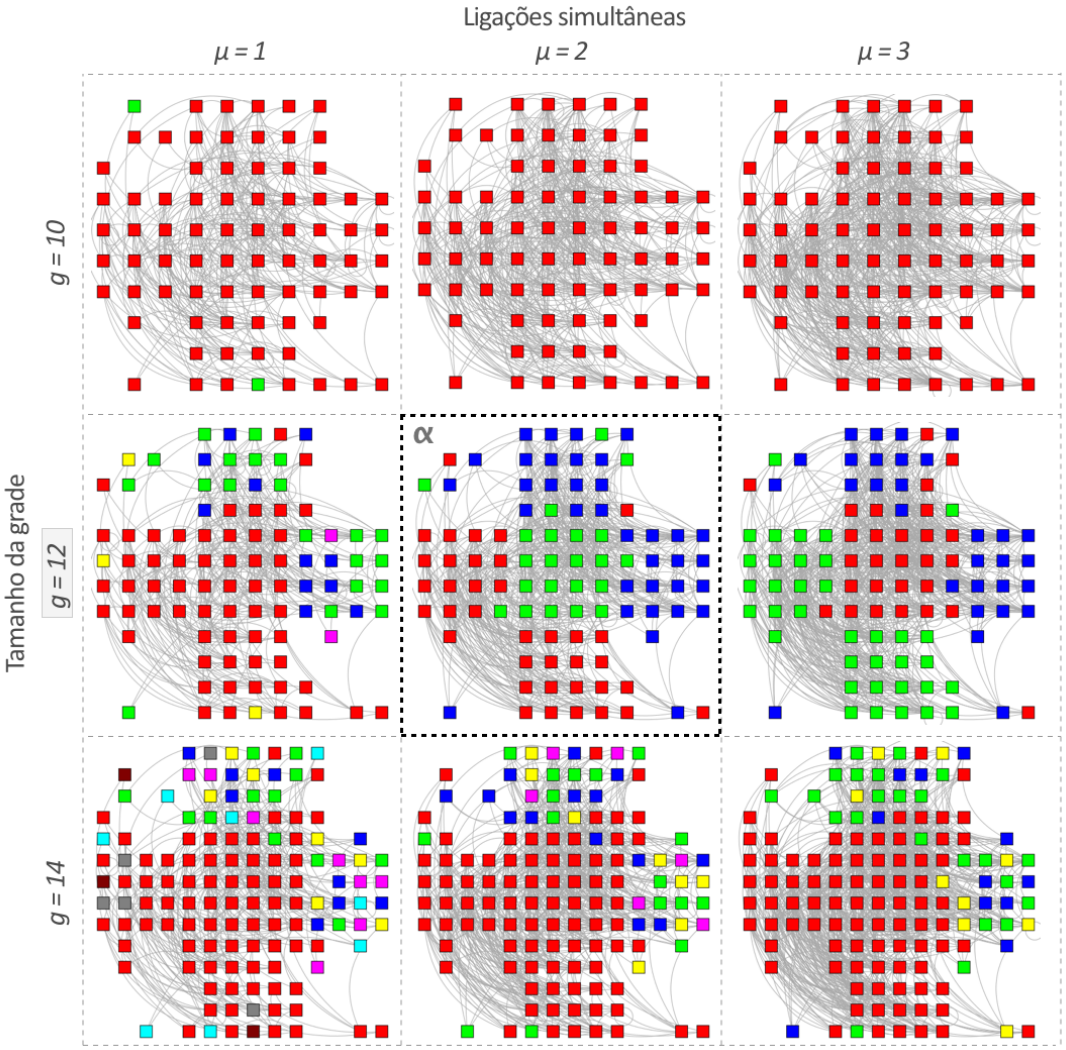
Figura A.8 - Cruzamento de eventos com ruido - Método ST-DBSCAN



Quatro imagens com os 454 eventos agrupados pelo método ST-DBSCAN usando diversos parâmetros. Apesar do ruido, o método consegue identificar o grupo principal, formado pelos eventos de cor azul.

Fonte: Produção do autor

Figura A.9 - Cruzamento de eventos com ruido - Método cronológico

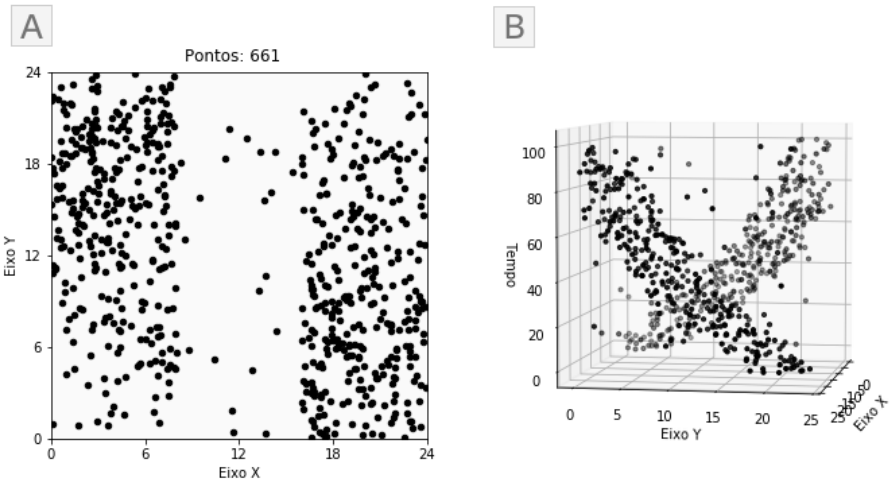


Nove imagens mostrando diversas comunidades a partir da variação das ligações simultâneas e o tamanho da grade. α) Resultado da configuração segundo a recomendação do método.

Fonte: Produção do autor

A.4 Eventos paralelos com ruido

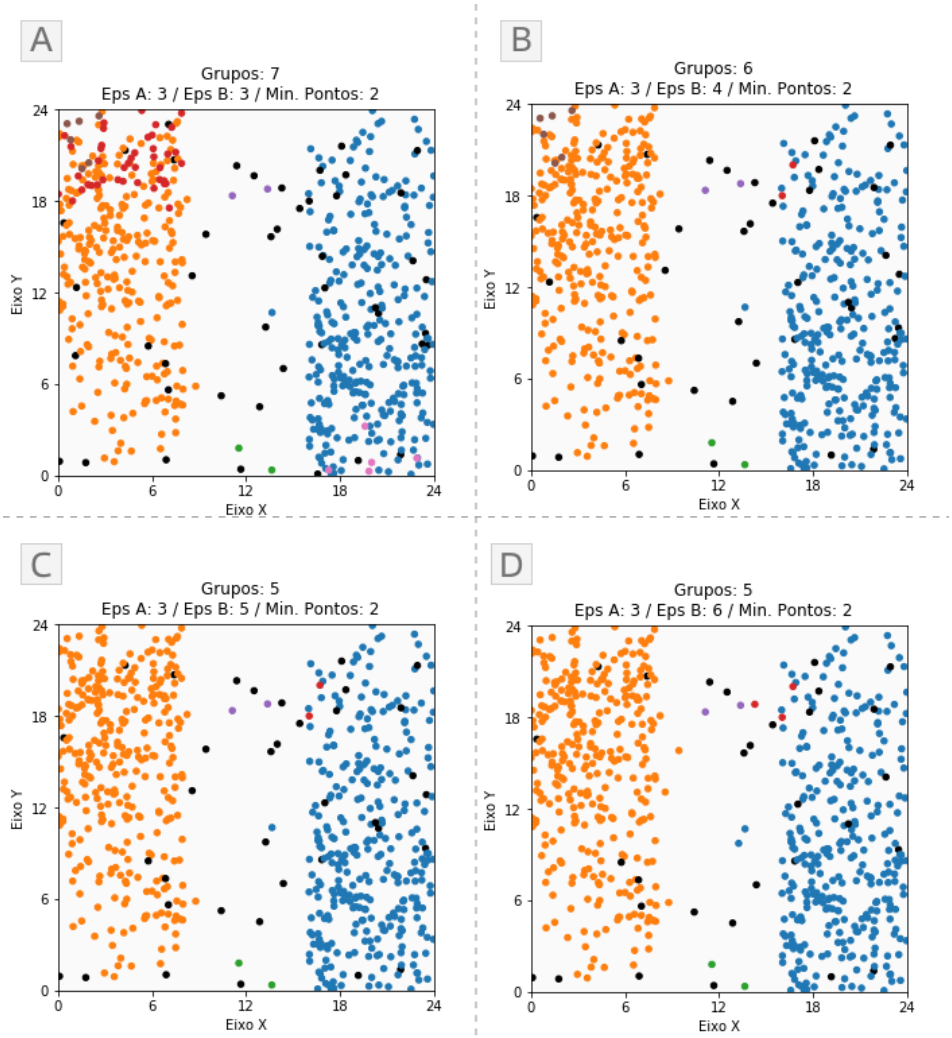
Figura A.10 - Modelo de eventos paralelos com ruido



A) 661 eventos separados em dois grupos se deslocando em direções opostas e mudando na quantidade de eventos ao longo do tempo, considerando que 10% destes foram removidos e adicionados de forma aleatória, para adicionar ruido ao modelo. B) os mesmos eventos adicionando a dimensão temporal no eixo vertical.

Fonte: Produção do autor

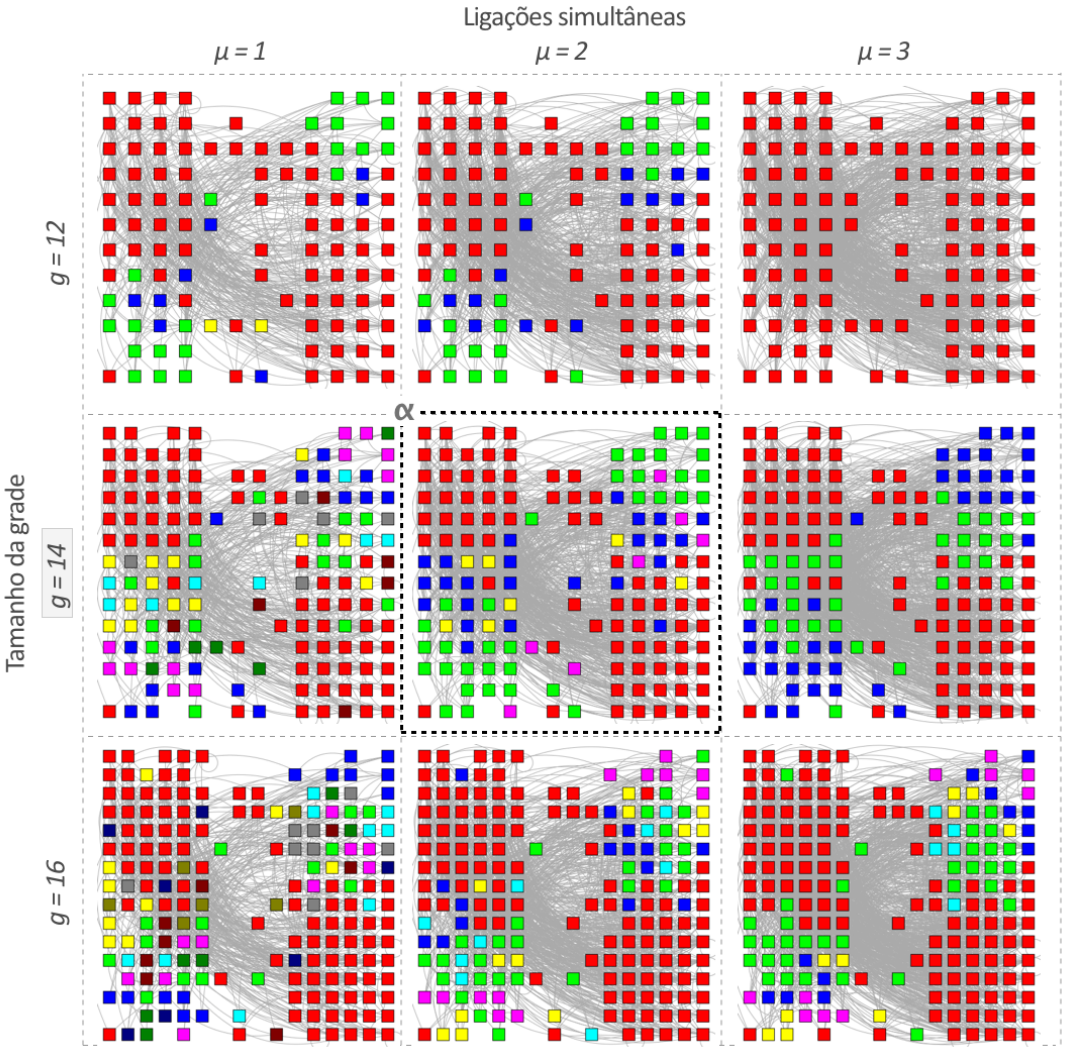
Figura A.11 - Eventos paralelos com ruido - Método ST-DBSCAN



Quatro imagens mostrando o resultado dos agrupamentos após aplicar o método ST-DBSCAN usando diversos parâmetros. Apesar do ruido, o método identifica claramente os dois grupos principais, na cor laranja e azul.

Fonte: Produção do autor

Figura A.12 - Eventos paralelos com ruido - Método cronológico



Nove imagens mostrando os resultados após aplicar o método cronológico no modelo de eventos paralelos e com ruido, considerando entre 1 e 3 ligações simultâneas (μ), e tamanhos de grade (g) de 12, 14 e 16. α) Imagem mostrando o resultado com a configuração recomendado pelo método.

Fonte: Produção do autor

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.