UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

ANÁLISE E MODELAGEM DE REDES GEOLOCALIZADAS: UMA PROPOSTA BASEADA EM REDES MULTICAMADAS VOLTADA PARA O ESTUDO DE COMUNIDADES

Eric Oliveira Leal

Orientador
Jefferson Elbert Simões

RIO DE JANEIRO, RJ - BRASIL
JANEIRO DE 2025

ANÁLISE E MODELAGEM DE REDES GEOLOCALIZADAS: UMA PROPOSTA BASEADA EM REDES MULTICAMADAS VOLTADA PARA O ESTUDO DE COMUNIDADES

Eric Oliveira Leal

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado por:

_____

Jefferson Elbert Simões - UNIRIO

_____

Laura de Oliveira Fernandes Moraes - UNIRIO

_____

Pedro Nuno de Souza Moura - UNIRIO

RIO DE JANEIRO, RJ - BRASIL
JANEIRO DE 2025

Catalogação informatizada pelo(a) autor(a)

L433 Leal, Eric Oliveira
Análise e modelagem de redes geolocalizadas: uma proposta baseada em redes multicamadas voltada para o estudo de comunidades / Eric Oliveira Leal. -- Rio de Janeiro : UNIRIO, 2025.
109p

Orientador: Jefferson Elbert Simões.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2025.

1. Ciência de redes. 2. Redes geolocalizadas. 3. Comunidades em redes. I. Simões, Jefferson Elbert, orient. II. Título.

Aos meus pais Soraia Mesquita de Oliveira Leal e Francisco Roberto Milhomes Leal que sempre me apoiaram e ao meu companheiro Edjalma Gomes Jesus Damasceno.

Agradecimentos

Assim como o a ciência que baseia esse estudo mostra, as conexões entre nós, coisas, pessoas ou o que quer que seja, são elementos poderosos e podem mudar toda uma estrutura completa e complexa. Por esse motivo eu gostaria de agradecer a todas a conexões fortes, fracas, de longa data ou que surgiram há pouco tempo, pois graças a elas eu cheguei até aqui.

Em especial quero agradecer ao meu orientador Jefferson Elbert Simões que foi fundamental para que eu conseguisse concluir esse trabalho e que se tornou um amigo no processo.

Agradeço aos membros da minha banca Laura de Oliveira Fernandes Moraes e Pedro Nuno de Souza Moura que me inspiraram durante a graduação.

Aos melhores amigos que pude conhecer nessa instituição e melhor grupo que me ajudou durante todo esse período e que pretendo ficar próximo para sempre, Eduardo dos Santos Gonçalves, Vívian Rique Gil Ferraro e Priscilla Aparecida Matias de Souza.

Por fim, agradeço a todos que me acompanharam mais uma vez nessa jornada de graduação.

Every choice we make and every road we take
Every interaction starts a chain reaction
We're both affected when we least expect it
And then when we touched then it all connected

Natasha Bedingfield - *Touch*

RESUMO

O estudo da ciência de redes é uma área relativamente nova quando comparada a outras disciplinas científicas, mas seu avanço tem demonstrado grande relevância, tanto pela simplicidade na construção de suas estruturas quanto pela capacidade de analisar o comportamento de diversos sistemas. Entre as variáveis mais importantes para determinados tipos de redes está a localização geográfica, que adiciona um novo espectro de possibilidades analíticas. Apesar de sua relevância em redes como malhas ferroviárias, grades de distribuição de energia e até mesmo redes sociais, a análise de dados geográficos em ciência de redes ainda carece de uma base formalmente consolidada, especialmente no que diz respeito à noção de comunidades. Este trabalho busca preencher essa lacuna ao propor um método estruturado para avaliar as comunidades identificadas em redes que integram componentes geográficas e funcionais. Partindo dessa ideia foi feito o levantamento de *datasets* que representassem redes com nós associados a informações geográficas. Esse processo enfrentou limitações devido à escassez de dados com essa configuração específica, o que levou à escolha de um conjunto de dados disponível do *Twitter*. Com esses dados, foi conduzida uma análise abrangente da base, incluindo a geração de uma rede de menções entre usuários. A partir dessa rede, foram extraídas métricas básicas, como o coeficiente de clusterização, além de uma análise estrutural dos graus, que apresentaram um comportamento característico de lei de potência. Por fim, foram investigadas as comunidades sociais e os agrupamentos geográficos para identificar a correlação entre essas dimensões. Em seguida é proposto um modelo que se baseia em redes multicamadas para gerar uma rede de referência de duas camadas com parâmetros ajustados, a qual a primeira camada é puramente geográfica e a segunda uma camada funcional de outra natureza. Este modelo pode ser usado para analisar as comunidades ajustando os parâmetros para considerar a influência maior de uma camada ou outra. O modelo foi testado utilizando os modelos de Watts-Strogatz e Waxman e foram usadas as métricas do Índice de Jaccard e *Adjusted Mutual Information* de semelhança entre conjuntos para avaliar o impacto nas comunidades identificadas em uma camada em relaçao a outra.

Palavras-chave: Redes Geolocalizadas, Comunidades em Redes, Redes Multicamadas.

ABSTRACT

The study of network science is a relatively new field compared to other scientific disciplines, but its advancement has shown great relevance due to both the simplicity of constructing its structures and the ability to analyze the behavior of various systems. Among the most important variables for certain types of networks is geographic location, which adds a new spectrum of analytical possibilities. Despite its relevance in networks such as railway grids, power distribution grids, and even social networks, the analysis of geographic data in network science still lacks a formally consolidated foundation, especially regarding the notion of communities. This work aims to fill this gap by proposing a structured method to evaluate the communities identified in networks that integrate geographic and functional components. Based on this idea, a survey of datasets representing networks with nodes associated with geographic information was conducted. This process faced limitations due to the scarcity of data with this specific configuration, leading to the choice of an available Twitter dataset. With this data, a comprehensive analysis of the base was conducted, including the generation of a mention network among users. From this network, basic metrics such as the clustering coefficient were extracted, as well as a structural analysis of degrees, which presented a characteristic power-law behavior. Finally, social communities and geographic clusters were investigated to identify the correlation between these dimensions. Subsequently, a model based on multilayer networks is proposed to generate a two-layer reference network with adjustable parameters, where the first layer is purely geographic and the second is a functional layer of a different nature. This model can be used to analyze communities by adjusting the parameters to consider the greater influence of one layer or another. The model was tested using the Watts-Strogatz and Waxman models, and the Jaccard Index and Adjusted Mutual Information metrics of set similarity were used to evaluate the impact on the communities identified in one layer in relation to another.

Keywords: Geolocated Networks, Network Communities, Multilayer Networks.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Redes complexas estão presentes em diversas áreas de estudo, como a matemática, física, química, etc. Esses sistemas compartilham propriedades emergentes que vão além da soma de seus componentes, como alta conectividade, formação de comunidades e eficiência na transmissão de informações. A análise dessas redes permite compreender fenômenos complexos que envolvem interações entre entidades e, frequentemente, os fatores subjacentes que governam sua organização (VIEGAS et al., 2023).

Entre os diversos tipos de redes, destaca-se o papel da localização geográfica como uma dimensão fundamental em muitos sistemas. Em redes de transporte, como malhas rodoviárias e aeroviárias, as conexões são diretamente influenciadas pela proximidade espacial. Em redes sociais, a geografia também desempenha um papel significativo, já que indivíduos que compartilham espaços físicos tendem a interagir com maior frequência. No entanto, Tillema, Dijst e Schwanen (2010) apontam que as redes sociais não se limitam à proximidade geográfica, sendo também moldadas por outros fatores, como interesses comuns, afiliação social e interações virtuais.

Apesar dessa relevância, a maioria dos estudos de redes complexas tende a analisar separadamente os fatores geográficos e sociais, tratando-os como dimensões independentes. Essa abordagem fragmentada pode negligenciar o impacto combinado dessas variáveis na formação e na estrutura das redes. Assim, compreender como a localização espacial interage com outras formas de conexão é essencial para modelar sistemas que representam melhor a realidade e para explorar propriedades que surgem dessa combinação (FACCHINETTI-MANNONE, 2019).

Diante desse contexto, este trabalho se propõe a explorar a integração entre redes baseadas na geografia e redes influenciadas por outras interações. Essa perspectiva oferece uma abordagem mais holística, capaz de capturar dinâmicas mais ricas e complexas que refletem tanto os agrupamentos locais quanto as conexões que transcendem barreiras

espaciais.

1.1 Motivação

Este trabalho foi motivado pela hipótese de que redes complexas, que representam diferentes tipos de interações, apresentam uma correlação significativa quando analisadas em conjunto com a posição geográfica de seus nós. A geografia frequentemente desempenha um papel importante na formação de conexões, especialmente em redes sociais, onde é intuitivo supor que indivíduos próximos geograficamente têm maior probabilidade de interagir. Essa proximidade física tende a criar comunidades que combinam tanto conexões locais quanto padrões espaciais maiores.

A partir dessa premissa, surge a necessidade de explorar modelos que combinem de maneira eficaz a localização geográfica e outros fatores de interação para servirem como *benchmark* na análise de comunidades de redes geolocalizadas, sendo capazes de configurar de forma controlada a influência da característica geográfica e da outra natureza. Esse interesse é especialmente relevante porque, em muitos sistemas complexos, as conexões não são exclusivamente determinadas pela proximidade espacial. Outros fatores, como afinidades sociais, interesses compartilhados ou dinâmicas de rede, também desempenham papéis cruciais. Por exemplo, enquanto a proximidade geográfica pode facilitar interações iniciais, fatores sociais podem criar conexões que transcendem barreiras espaciais.

Dessa forma, este trabalho pretende propor e avaliar um modelo que consiga avaliar as comunidades de uma rede com duas camadas — espacial e de outra natureza. Esse modelo é projetado para ser um metamodelo, ou seja, um *framework* que usará de outros modelos para estudar as características das comunidades em redes geolocalizadas, sem estar diretamente ligado a um dataset específico.

Além disso este modelo permitirá estudar como cada aspecto influencia a estrutura da rede e como a combinação deles pode revelar padrões mais complexos e realistas de conectividade. Além disso, ao incluir essas duas perspectivas, busca-se compreender melhor a formação de comunidades, a propagação de informações e os fatores que governam a organização de redes em contextos variados.

1.2 Objetivos

1.2.1 Objetivos principais

O desenvolvimento deste trabalho tem como objetivo principal estudar a estrutura de comunidades identificadas em redes considerando duas concepções diferentes, por um lado a componente geográfica e por outro uma informação de outra natureza.

1.2.2 Objetivos específicos

Para o desenvolvimento deste trabalho, será necessário analisar as estruturas social e geográfica encontradas em redes reais, utilizando bases de dados consolidadas, conforme sua disponibilidade e adequação às exigências do estudo.

Com os dados coletados, a rede social digital será estruturada e analisada utilizando métricas da ciência de redes. Essa análise será realizada por meio de uma ferramenta computacional adequada, como `graph-tool`, `NetworkX` ou outra similar, para identificar o comportamento das comunidades identificadas.

Por fim, será proposto um modelo de rede multicamadas, que possua uma correlação controlada entre os atributos geográfico e funcional da rede. Através desse modelo é que será possível gerar as redes que servirão de base para os testes de comunidades.

1.3 Estrutura do texto

Este trabalho está estruturado em seis capítulos. O primeiro, de introdução, apresenta a motivação e os objetivos principais e específicos. O segundo capítulo, de fundamentação, discute os conceitos fundamentais de Teoria de Grafos, Ciência de Redes, Comunidades em Redes, Modelos de Redes e Fundamentos de Geolocalização.

No terceiro capítulo, é realizada uma análise exploratória dos dados coletados em um artigo, com foco nas comunidades da rede social e nos agrupamentos por localização. O quarto capítulo descreve um modelo que serve como ponto de referência para o estudo de comunidades em redes geolocalizadas.

No capítulo seguinte, é realizada uma análise do modelo utilizando métricas para avaliar a similaridade entre as comunidades identificadas nas duas camadas da rede. O trabalho é finalizado com o capítulo de conclusão, onde são apresentadas as considerações finais e sugestões para futuros trabalhos.

2. Fundamentação Teórica

2.1 Fundamentos de Teoria de Grafos

Conforme descrito por Barabási e Pósfai (2016), um grafo é uma estrutura matemática que representa um conjunto de elementos conectados por relações binárias. Esses elementos são denominados vértices, enquanto as conexões entre eles são chamadas de arestas. Um grafo pode ser formalmente definido como $G(N, L)$, onde N é um conjunto finito e não vazio de vértices, e L é o conjunto de arestas que conectam pares de vértices distintos em N. As arestas podem ser representadas como pares ordenados ou não ordenados, dependendo da natureza da relação entre os vértices. Por exemplo, uma aresta $e = (u, v)$ conecta os vértices u e v.

Considerando um grafo simples, em que um vértice não se conecta a si mesmo e nem faz múltiplas conexões com um mesmo vértice, o grau de um vértice u, denotado como $k(u)$, é definido como o número de arestas que conectam u a outros vértices no grafo. Por exemplo, na Figura 1 o vértice 1 está conectado a cinco outros vértices, então $k(u) = 5$. Outro conceito relevante é o de caminho, que consiste em uma sequência de vértices conectados por arestas, permitindo alcançar um vértice de destino a partir de um vértice inicial. Um caminho com n vértices possui $n - 1$ arestas.

Também podem surgir ciclos, que ocorrem quando um caminho retorna ao vértice inicial após passar por outros vértices que não tenham sido visitados anteriormente. Formalmente, um ciclo é definido como um caminho $u_1, u_2, \ldots, u_k, u_{n+1}$, onde $u_1 = u_{n+1}$ e $n \geq 3$. Caso nenhum ciclo seja encontrado no grafo, este é classificado como acíclico.

Um grafo é dito conexo se existe pelo menos um caminho entre qualquer par de vértices. Caso contrário, ele é classificado como desconexo, indicando que há vértices isolados ou grupos de vértices sem ligação entre si.

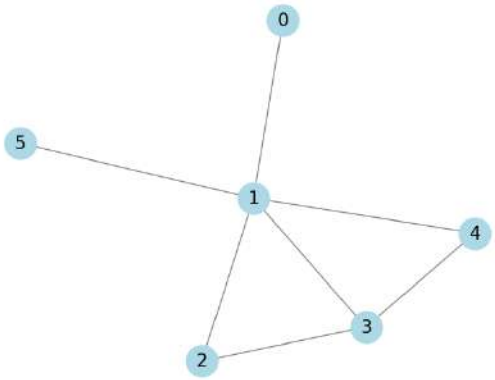


Figura 1: Exemplo de grafo com 6 nós e 7 arestas

Com relação à direção das arestas, os grafos podem ser classificados como:

Não-direcionados : as arestas não possuem orientação, ou seja, uma conexão entre os vértices u e v é bidirecional: $e = u, v$.

Direcionados : as arestas possuem orientação, representando uma relação unidirecional: $e = (u, v)$ indica uma ligação de u para v, mas não necessariamente de v para u.

A teoria de grafos destaca-se como uma área de grande relevância para o desenvolvimento de estudos em diversas disciplinas. Uma de suas propriedades mais notáveis é sua alta versatilidade de modelagem, permitindo sua aplicação em áreas que variam desde o estudo de reações moleculares na Química e a simulação de estruturas atômicas complexas na Física até a solução de problemas relacionados à Inteligência Artificial (IA) na Ciência da Computação (KAUR; TRIPATHI; VERMA, 2008).

Além disso, os grafos oferecem uma maneira intuitiva de visualizar dados complexos, simplificando sua compreensão. Por exemplo, King, Aboudina e Shalaby (2019) apresentam exemplos claros dessa característica, como a facilidade de se localizar em um mapa de metrô ou trem apenas observando as conexões entre as estações. Essa capacidade de representar informações de forma visual e acessível ressalta o impacto dos grafos em áreas multidisciplinares.

Na ciência de redes, a teoria de grafos desempenha um papel essencial, fundamentando suas premissas e análises. No entanto, é importante destacar diferenças marcantes entre as duas abordagens. Enquanto a teoria de grafos segue um rigor matemático mais formal, a ciência de redes evoluiu incorporando conhecimentos de diversas áreas, como Física e Ciência de Dados, combinando fundamentos matemáticos com aspectos empíricos e práticos (BARABÁSI; PÓSFAI, 2016).

A teoria dos grafos tradicionalmente envolve construções mais rígidas, analisando um número menor de dados e focando em propriedades estáticas. Por outro lado, as redes tratam frequentemente de volumes de dados extremamente grandes, alcançando bilhões de nós e enlaces, além de apresentarem dinâmicas mais complexas, frequentemente associadas à análise temporal. Essa distinção entre a teoria de grafos e a ciência de redes ilustra como ambas evoluíram para atender a diferentes necessidades de modelagem e análise.

2.2 Fundamentos de Ciência de Redes

Com boa parte da teoria dos grafos definidas, este trabalho dispõe agora de material suficiente para explorar a ciência de redes, continuando ainda pelas definições presentes em Barabási e Pósfai (2016). Na ciência de redes, a nomenclatura difere ligeiramente do apresentado para os grafos: os vértices são comumente chamados de nós, e as arestas são referidas como elos. Este trabalho vai adotar no seu desenvolvimento os nomes **nós** e **arestas**.

As redes representam sistemas complexos e como são estruturas baseadas em grafos, elas são projetadas de forma semelhante, definindo como os nós se ligam uns aos outros e que comportamentos podem ser analisados a partir dessa interação. Existem atributos associados às redes que são essenciais para a realização de diversos tipos de estudos. Um dos mais triviais e extremamente importante é o **grau médio** $\langle k \rangle$, obtido pela média dos graus de todos nós que é equivalente ao dobro do número de arestas L dividido pela quantidade de nós na rede N, como observado na Equação 2.1 (BARABÁSI; PÓSFAI, 2016).

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{2.1}$$

A **densidade** é um conceito fundamental para a análise de redes, pois representa o nível de interconectividade entre os nós, refletindo o quão unidos eles estão por meio de suas conexões. Essa métrica é particularmente útil para compreender a estrutura e a dinâmica de diferentes tipos de redes, permitindo identificar padrões e características gerais de agrupamento.

De acordo com Barabási e Pósfai (2016), a densidade é definida como a proporção de arestas existentes em relação ao número máximo de arestas possíveis entre todos os nós

da rede. Seu cálculo varia dependendo de a rede ser direcionada ou não direcionada:

Para redes não direcionadas, a densidade (d) é dada pela Equação 2.2:

$$d = \frac{2L}{N(N-1)} \tag{2.2}$$

Para redes direcionadas, onde as arestas possuem direção, o cálculo é ajustado para a Equação 2.3:

$$d = \frac{L}{N(N-1)} \tag{2.3}$$

A partir desses conceitos, é possível perceber que o grafo da Figura 1 tem grau médio $\langle k \rangle \simeq 2.334$ e densidade $d \simeq 0.467$.

A densidade assume valores no intervalo de 0 a 1. Quando $d = 0$, a rede não possui nenhuma conexão entre seus nós, sendo completamente desconexa. Por outro lado, quando $d = 1$, a rede apresenta a estrutura de um grafo completo, em que cada nó está conectado a todos os outros.

Essa métrica é amplamente utilizada para avaliar a conectividade global de redes, comparando diferentes sistemas e identificando padrões de agrupamento. Redes densas, por exemplo, podem indicar uma forte interação entre os nós, enquanto redes esparsas refletem interações limitadas ou especializadas.

Dentro das redes, existem subgrupos de nós que estão interconectados, ou seja, em uma partição da rede, existe pelo menos um caminho que conecta todos os nós dessa partição. Esses conjuntos de nós são conhecidos como componentes conexas, análogas aos grafos conexos descritos na Seção 2.1.

Entre essas componentes, uma em particular se destaca pela sua relevância: a **Componente Conexa Gigante** (*Giant Connected Component* - GCC). Ela é definida como a única componente que contém uma fração significativamente grande dos nós da rede em relação ao total de nós. A formação dessa componente está intimamente relacionada a um parâmetro crítico c, definido como a razão entre o número médio de arestas L e o número de nós N, isto é, $p_c = L/N$. À medida que p_c aumenta, o grafo atravessa diferentes fases de conectividade.

Na fase inicial, quando $p_c < 1/2$, a rede é composta predominantemente por pequenas componentes, consistindo de árvores isoladas e ciclos simples. No ponto crítico, quando

$p_c = 1/2$, ocorre uma transição importante, na qual começa a surgir uma estrutura mais conectada e uma fração significativa dos nós se agrupa em uma única componente, esse ponto crítico é válido para o modelo proposto por Erdos e Renyi (1960) e não necessariamente atende a todas as redes possíveis. Após essa transição, quando $p_c > 1/2$, a GCC emerge, contendo a maior parte dos nós da rede, enquanto as outras componentes permanecem pequenas e isoladas.

A conectividade em redes é uma propriedade fundamental que sustenta muitos estudos na área, sendo frequentemente analisada por meio de métricas que avaliam os agrupamentos ou *clusters* presentes. Conforme descrito por Barabási e Pósfai (2016), o **Coeficiente de Clusterização Local** (C_i) é uma métrica que mede o grau de interconexão entre os vizinhos de um nó específico. Para um nó i com grau k_i, C_i é definido pela Equação 2.4:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.4}$$

onde L_i representa o número de conexões efetivas entre os k_i vizinhos do nó i. O valor de C_i varia entre 0 e 1:

$C_i = 0$ se nenhum dos vizinhos do nó i estiver vinculado entre si.

$C_i = 1$ se os vizinhos do nó i formam um subgrafo completo, ou seja, todos eles estão ligados entre si.

O coeficiente C_i reflete a densidade de conexões locais na vizinhança de um nó: quanto mais interconectados forem os vizinhos de i, maior será o valor de C_i.

Para avaliar o grau de agrupamento da rede como um todo, utiliza-se o Coeficiente de Clusterização Médio ($\langle C \rangle$), que é a média dos valores de C_i calculados para todos os nós $i = 1, ..., N$ da rede. A fórmula é dada pela Equação 2.5:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.5}$$

.

O coeficiente de clusterização médio pode ser interpretado como a probabilidade de que dois nós da rede, ambos conectados a um terceiro nó comum, estejam também conectados entre si, formando assim um triângulo. Essa métrica fornece uma visão global

da densidade de agrupamento presente na rede, sendo amplamente utilizada para analisar a estrutura de redes complexas.

No desenvolvimento de uma rede aleatória, é comum observar que alguns nós recebem um grande número de arestas, enquanto outros possuem poucos ou até mesmo nenhum. Essa aleatoriedade na distribuição de conexões é descrita pela distribuição de grau p_k, que representa a probabilidade de um nó escolhido aleatoriamente ter exatamente grau k (BARABÁSI; PÓSFAI, 2016).

Com base nessa premissa, a distribuição binomial é frequentemente utilizada para descrever redes aleatórias. A probabilidade de um nó i possuir exatamente k arestas é determinada pelo produto de três componentes fundamentais:

1. A probabilidade dos k arestas estarem presentes (p_k)
2. A probabilidade dos $(N-1-k)$ arestas restantes não existirem ($(1-p)^{N-1-k}$)
3. A quantidade de maneiras que se pode escolher k arestas em $N-1$ arestas possíveis que um nó pode ter ($\binom{N-1}{k}$).

Com base nesses componentes, Barabási e Pósfai (2016) demonstra que a distribuição de grau de uma rede aleatória segue a distribuição binomial proposta por Erdös e Rényi (1959) que diz que cada par de nós (i,j) em um grafo é conectado com probabilidade $p \in (0,1)$, independentemente dos outros nós. Essa probabilidade pode ser expressa como:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.6}$$

.

Essa formulação descreve o comportamento probabilístico do grau dos nós em redes aleatórias, refletindo como a estrutura dessas redes é moldada pela probabilidade de conexão p_k e pelo número total de nós. Em redes muito grandes ($N \to \infty$) e com baixa probabilidade de conexão ($p \ll 1$), a distribuição binomial se aproxima de uma distribuição de Poisson.

Como apresentado por Barabási e Pósfai (2016), as redes reais geralmente são muito esparsas o que leva a terem um grau médio ($\langle k \rangle$) significativamente menor que o número total de nós (N). Por esse motivo, a distribuição de Poisson oferece uma representação mais adequada para essas estruturas. Essa distribuição é definida pela equação:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.7}$$

A Equação 2.7, em conjunto com a Equação 2.6, descreve a distribuição de grau de uma rede aleatória. Barabási e Pósfai (2016) expõem também que como ambas representam a mesma quantidade, possuem características semelhantes.

Tanto na distribuição binomial quanto na de Poisson, o pico está centralizado em torno de $\langle k \rangle$, que corresponde ao grau médio da rede. À medida que o valor de p_k aumenta, tornando a rede mais densa, o grau médio também cresce, deslocando o pico da distribuição para a direita. Além disso, a dispersão da distribuição — observada na largura da curva — é diretamente controlada pela densidade da rede. Redes mais densas apresentam distribuições mais amplas, refletindo uma maior variabilidade nos graus dos nós.

Essa relação entre densidade, grau médio e dispersão destaca como a conectividade afeta diretamente a estrutura das redes, permitindo uma análise detalhada de suas propriedades estatísticas. Redes esparsas, por exemplo, tendem a exibir menos variação nos graus dos nós, enquanto redes mais densas revelam uma maior heterogeneidade nas conexões (BARABÁSI; PÓSFAI, 2016).

Além do exposto, quando se trata de redes reais, uma outra propriedade fica bem evidente. A propriedade *Scale-Free* é uma característica de que, em muitas redes reais, o número de conexões (ou o grau) dos nós segue uma distribuição de potência, também conhecida como distribuição *power-law*.

Como o próprio nome sugere a distribuição de grau destas redes não possui uma escala bem definida diferente de outras como a distribuição normal ou a exponencial que apresentam um valor médio explícito.

Em redes com essa propriedade, a maioria dos nós possui poucas conexões, enquanto um pequeno número de nós chamados de *hubs*, é altamente conectado. Esses agrupamentos são essenciais para a conectividade e o funcionamento da rede, além de tornaram as redes mais robustas a falhas aleatórias já que a remoção de nós pouco conectados não afeta tanto a conectividade geral da rede (BARABÁSI; PÓSFAI, 2016).

O conceito de Redes Multicamadas é o responsável por estudar sistemas complexos com múltiplos tipos de interações ou relações entre seus elementos. De Domenico et al. (2013) explicam que diferentemente das redes simples (monocamada), em que há apenas um tipo de conexão entre os nós, as redes multicamadas são compostas por várias camadas,

cada uma representando um tipo diferente de relação. Cada camada possui sua própria topologia, com nós e arestas específicos, e os mesmos nós podem estar presentes em múltiplas camadas, mas com conexões potencialmente diferentes.

Além das conexões dentro de cada camada, podem existir conexões entre nós de camadas diferentes, chamadas de conexões intercamadas. Um mesmo par de nós pode estar conectado de formas diferentes em camadas distintas, capturando a natureza multifacetada das interações. Processos dinâmicos em uma camada podem influenciar e ser influenciados por dinâmicas em outras camadas. As redes multicamadas permitem modelar de forma mais realista diversos sistemas complexos, como redes sociais com diferentes tipos de relações (amizade, trabalho, família), redes de transporte com múltiplos modos (aéreo, rodoviário, ferroviário), redes biológicas com interações em diferentes escalas (genética, metabólica, proteica) e redes de infraestrutura interdependentes (energia elétrica, telecomunicações, água) (BERLINGERIO et al., 2011).

O estudo de redes multicamadas tem ganhado grande interesse nos últimos anos, permitindo avanços na compreensão de sistemas complexos em diversas áreas do conhecimento. Sua estrutura mais rica possibilita capturar aspectos antes negligenciados em abordagens de redes simples, levando a *insights* importantes sobre o funcionamento e vulnerabilidades de sistemas reais (KIVELA et al., 2014).

2.3 Comunidades em redes

Segundo a Hipótese de Conectividade e Densidade, comunidades são subgrafos localmente densos e conectados, o que significa que os membros de uma comunidade estão ligados diretamente ou através de outros membros do mesmo grupo (conectividade), e a probabilidade de conexão entre os nós dentro de uma comunidade é maior do que entre nós de comunidades diferentes (densidade) (BARABÁSI; PÓSFAI, 2016).

As comunidades podem ser classificadas de diferentes maneiras, variando de estruturas rigorosamente definidas, como cliques, a formas mais gerais:

Cliques : Um clique é um subgrafo completo em que todos os nós estão conectados uns aos outros. Embora sejam exemplos ideais de comunidades densas, cliques grandes são raros em redes reais, tornando essa definição excessivamente restritiva.

Comunidades Fortes : Um subgrafo onde cada nó tem mais conexões dentro da comunidade do que fora dela.

Comunidades Fracas : Um subgrafo no qual o número total de conexões internas do subgrafo seja maior que o número de conexões externas. Essa definição é menos rígida e reflete melhor a natureza das redes reais.

Uma métrica amplamente utilizada para avaliar a qualidade das partições de comunidades em uma rede é a **modularidade**. Essa métrica mede a diferença entre o número de conexões dentro de uma comunidade em comparação com o número esperado em uma rede aleatória com o mesmo grau médio. Uma partição com alta modularidade indica comunidades bem definidas, enquanto valores baixos sugerem ausência de agrupamentos significativos.

A partir de uma rede composta por N nós e L arestas, cada comunidade c é formada por um total de N_c nós conectados por L_c arestas, onde $c = 1, ..., n_c$. Nesse contexto, Barabási e Pósfai (2016) apresentam a fórmula para calcular a modularidade de cada comunidade. Esse cálculo leva em conta o diagrama real da rede, representado pela matriz de adjacência (A_{ij}), que indica a presença de uma conexão direta entre os nós i e j, e o número esperado de conexões (p_{ij}) entre os mesmos nós caso os arestas da rede fossem distribuídos de forma completamente aleatória. A Equação 2.8 mostra então como é estruturado esse cálculo.

$$M_c = \frac{1}{2L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij}) \tag{2.8}$$

Barabási e Pósfai (2016) mostram que esta equação pode ser simplificada na forma da Equação 2.9:

$$M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 \tag{2.9}$$

onde:

L_c é a quantidade de arestas dentro da comunidade;

k_c é o grau total na comunidade.

Para generalizar a ideia da formação das comunidades, considera-se uma partição completa da rede em n_c comunidades. Essa abordagem busca avaliar como a densidade de conexões dentro de cada comunidade se compara à densidade esperada em uma rede aleatória com a mesma estrutura de grau. Para isso, define-se que a modularidade (M) da partição é obtida somando as contribuições de todas as n_c comunidades, considerando

a diferença entre a fração de arestas que efetivamente conectam os nós dentro de uma comunidade e a fração esperada para uma rede aleatória. Matematicamente, pode ser expressa pela Equação 2.10:

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \tag{2.10}$$

A modularidade geral apresenta algumas propriedades principais agrupadas em um exemplo na Figura 2 e definidas como:

1. **Valor máximo e qualidade da partição**: Quanto maior for o valor de M, melhor a partição reflete a estrutura das comunidades. A partição com a modularidade máxima, por exemplo, $M = 0.41$ (Figura 2a), captura com precisão comunidades bem definidas. Já partições com valores menores de modularidade indicam estruturas menos coerentes e podem se desviar das comunidades observáveis (Figura 2b). É importante notar que a modularidade não pode exceder 1, sendo este o limite teórico superior.

2. **Casos limite**:

   Toda rede como uma única partição : Quando toda a rede é considerada uma única partição observa-se que $M = 0$ (Figura 2c), pois a densidade de conexões dentro da comunidade é igual à densidade esperada para uma rede aleatória. Nesse caso, os dois termos na definição de M se cancelam.

   Cada nó como uma partição isolada : Se cada nó é tratado como uma comunidade separada, M assume valores negativos (Figura 2d). Isso ocorre porque $L_c = 0$ (não há arestas internas a uma partição de nó único) e a soma dos n_c termos é dominada pelos componentes negativos, resultando em $M < 0$.

A identificação de comunidades em redes é um problema desafiador devido ao crescimento exponencial do número de partições possíveis à medida que o número de nós aumenta. Métodos de força bruta para avaliar todas as partições são inviáveis em redes grandes, exigindo o desenvolvimento de algoritmos eficientes que possam identificar comunidades sem inspecionar todas as possibilidades.

Ao longo do avanço da Ciência de Redes vários algoritmos foram planejados para a detecção de comunidades, aqui serão apresentados o Algoritmo Guloso de Maximização da Modularidade e o Algoritmo de Louvain.

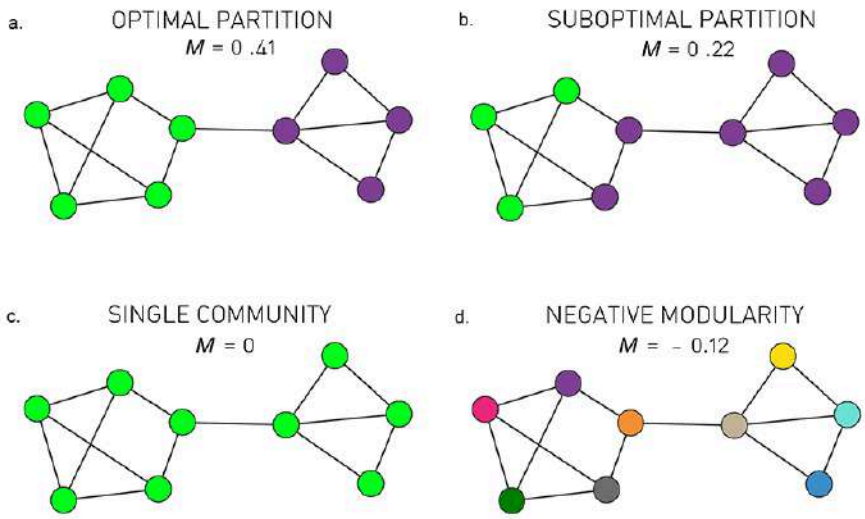


Figura 2: Exemplos de partições com diferentes modularidades. a. Partição ótima, b. Partição subótima, c. Comunidade única, d. Modularidade negativa (BARABÁSI; PÓSFAI, 2016)

2.3.1 Algoritmo de Maximização da Modularidade

O algoritmo de maximização da modularidade é uma busca local que identifica as comunidades possíveis procurando quais agrupamentos de nós com seus vizinhos levam à melhor melhoria da modularidade da rede. Ele segue um padrão guloso fazendo escolhas locais em cada passo para agrupar comunidades que proporcionam o maior aumento na modularidade, sem considerar o impacto a longo prazo dessas decisões.

Este algoritmo possui um tempo de processamento $O(Ld \log N)$, em que d representa a profundidade do dendrograma que descreve a estrutura de comunidade. Considerando que várias redes reais são esparsas e hierárquicas, ou seja $L \sim N$ e $d \sim \log N$, então no pior caso o algoritmo executa em tempo praticamente linear, definido por $O(N \log^2 N)$ (CLAUSET; NEWMAN; MOORE, 2004).

Para explicar o algoritmo de detecção de comunidades apresentado por Clauset, Newman e Moore (2004), é necessário estabelecer algumas definições fundamentais. A matriz de adjacência (A_{uv}) representa as conexões entre os nós u e v na rede. Seu valor é 1 se os nós estão conectados e 0 caso contrário.

Considerando agora uma divisão da rede em comunidades, onde cada nó pertence a uma comunidade específica identificada por c_u ou c_v, a fração de arestas que conecta nós da comunidade i à comunidade j pode ser representada por e_{ij}, que é calculada utilizando a Equação 2.11:

$$e_{ij} = \frac{1}{2L} \sum_{uv} A_{uv} \delta(c_u, i) \delta(c_v, j) \tag{2.11}$$

onde L é o número total de arestas da rede e $\delta(x, y)$ é a função delta de Kronecker, que assume o valor 1 se $x = y$ e 0 caso contrário. A fração de extremidades das arestas conectadas a uma comunidade i, por sua vez, é denotada por a_i e definida pela Equação 2.12:

$$a_i = \frac{1}{2L} \sum_{u} k_u \delta(c_u, i) \tag{2.12}$$

onde k_u é o grau do nó u, ou seja, o número de arestas ligadas a ele.

O algoritmo inicia com cada nó da rede em uma comunidade individual com $e_{ij} = 1/2L$ se i e j estão conectados e 0 caso contrário, e $a_i = k_i/2L$ para cada nó. Com isso os autores mostram através da Equação 2.13 que o aumento potencial na modularidade ΔQ da conexão entre os dois nós.

$$\Delta Q_{ij} = \begin{cases} 1/2L - k_i k_j/(2L)^2 & \text{se } i \text{ e } j \text{ são conectados} \\ 0 & \text{caso contrário.} \end{cases} \tag{2.13}$$

Com base nessas definições, o algoritmo segue um processo iterativo que busca maximizar a modularidade (Q) da rede. Em cada etapa, considera-se a fusão de pares de comunidades. Para isso, calcula-se o ΔQ, caso dois pares de comunidades i e j sejam fundidos.

Entre todos os pares possíveis, escolhe-se aquele que resulta no maior aumento em ΔQ, e as comunidades correspondentes são unidas em uma única. O processo é repetido iterativamente até que nenhuma fusão adicional seja capaz de aumentar a modularidade. Esse método garante que as comunidades sejam agrupadas de forma a capturar a estrutura da rede da maneira mais eficiente possível, dentro dos limites do modelo.

Este algoritmo apresenta uma abordagem eficiente para a detecção de comunidades em redes, especialmente por sua simplicidade e velocidade, o que o torna adequado para redes de tamanho moderado. Sua principal vantagem reside na capacidade de maximizar a modularidade de forma iterativa, oferecendo resultados que frequentemente refletem bem a estrutura de comunidades em redes reais.

No entanto, a busca local pode levar a máximos locais, impedindo a identificação da melhor partição global. Além disso, seu desempenho tende a se degradar em redes

muito grandes ou densas, e a dependência exclusiva da modularidade como métrica pode ser limitada, especialmente em redes com estruturas complexas ou comunidades muito pequenas. Ainda assim, o algoritmo guloso permanece uma ferramenta valiosa no estudo de redes complexas.

2.3.2 Algoritmo de Louvain

O Método Louvain é uma abordagem hierárquica e eficiente para a detecção de comunidades, amplamente utilizada devido à sua escalabilidade e à sua capacidade de lidar com redes muito grandes. Ele também busca maximizar a modularidade através de uma busca local na vizinhança de cada nó, porém utilizando múltiplos níveis (BLONDEL et al., 2008).

Traag (2015) apresenta que o algoritmo original de Louvain tinha um tempo de execução de $O(L)$, enquanto sua proposta reduziria essa eficiência seria reduzida para $O(N \log\langle k\rangle)$.

O algoritmo descrito por Blondel et al. (2008) é dividido em duas etapas que se repetem de forma iterativa, começando com uma rede ponderada composta por N nós. Na etapa inicial, cada nó é atribuído a sua própria comunidade, resultando em uma partição em que o número de comunidades é igual ao número de nós.

Posteriormente, para cada nó u, são examinados seus v vizinhos. Calcula-se o impacto na modularidade ao transferir u de sua comunidade atual para a comunidade de v. O nó u é então movido para a comunidade que maximiza o ganho de modularidade, aplicando uma regra de desempate caso haja igualdade, mas somente se esse ganho for positivo. Se nenhum ganho positivo for identificado, u permanece na comunidade original. Esse procedimento é repetido de forma sequencial para todos os nós, podendo ser revisitado várias vezes, até que não seja mais possível melhorar a modularidade. A primeira etapa termina quando se atinge um máximo local, isto é, nenhuma realocação adicional de nós pode aumentar a modularidade.

A segunda etapa do algoritmo envolve a construção de uma nova rede, cujos nós correspondem às comunidades identificadas na primeira etapa. Para criar essa rede, os pesos das arestas entre os novos nós são definidos como a soma dos pesos das conexões entre os nós das comunidades correspondentes na rede original. Conexões entre nós pertencentes à mesma comunidade resultam em laços para essa comunidade na nova rede.

Após completar essa segunda etapa, o algoritmo pode ser reiniciado aplicando novamente a primeira fase na rede ponderada resultante. Esse processo iterativo continua

até que não seja possível obter mais melhorias na modularidade.

O método Louvain combina simplicidade e eficiência computacional com uma abordagem iterativa que permite lidar com redes de grande escala. Sua capacidade de refinar comunidades ao longo de várias iterações, reconstruindo a rede em cada etapa, garante uma maior flexibilidade e precisão na detecção de comunidades.

Entre suas vantagens, destaca-se a capacidade de encontrar boas aproximações de máximos globais de modularidade, além de sua escalabilidade para redes extensas. Apesar disso, o Louvain também possui limitações, como a possibilidade de convergir para diferentes soluções dependendo da ordem dos nós avaliados, além de ser suscetível ao problema de resolução, que dificulta a identificação de comunidades muito pequenas.

2.4 Modelos de redes

A ciência de redes busca construir modelos que sejam capazes de obter as propriedades fundamentais de redes reais, fornecendo ferramentas para compreender e simular a complexidade inerente a esses sistemas. O problema na análise dessas redes é que diferentemente de estruturas regulares, como teias de aranha, a maioria das redes reais inicialmente aparenta ser caótica e desordenada. Essa aparente aleatoriedade é abordada pela teoria das redes aleatórias, que constrói e caracteriza redes em que as conexões são estabelecidas de forma puramente estocástica.

Um dos mais influentes é o modelo Barabási-Albert, que descreve a formação de redes livres de escala, caracterizadas por uma distribuição de grau em lei de potência. Nesse tipo de rede, poucos nós concentram um número elevado de conexões (*hubs*), enquanto a maioria dos nós possui poucas conexões, refletindo fenômenos como o crescimento dinâmico e a ligação preferencial, em que novos nós têm maior probabilidade de se conectar a nós mais conectados (BARABÁSI; PÓSFAI, 2016).

O modelo opera de forma iterativa, começando com um pequeno grafo conectado e adicionando novos nós que se conectam aos existentes com probabilidade proporcional ao grau dos nós. Xu e Zhang (2023) explicam que o modelo possui propriedades como dominância de *hubs*, alta conectividade e o efeito ”pequeno mundo”, o que o torna aplicável a redes como Internet, redes sociais e citações científicas. Apesar de suas contribuições, o modelo tem limitações, como a suposição de homogeneidade inicial e a exclusão de fatores externos, o que motivou o desenvolvimento de modelos mais sofisticados para abordar a diversidade e a complexidade das redes reais.

Embora exista uma simplicidade estrutural, composta apenas por nós e arestas, o verdadeiro desafio na modelagem de redes está na escolha de como estabelecer as conexões entre os nós. A complexidade das redes reais surge a partir dessa conectividade, e a filosofia subjacente às redes aleatórias é assumir que a melhor maneira de reproduzir essa complexidade é conectando os nós de forma aleatória.

Dessa perspectiva, uma rede aleatória pode ser definida como um modelo $G(N, p)$ composto por N nós, em que cada par de nós está conectado com uma probabilidade p. Esse conceito simples gera redes que capturam algumas propriedades estatísticas de sistemas reais e serve como ponto de partida para análises mais sofisticadas, incluindo o estudo da transição de conectividade, distribuição de grau e robustez estrutural.

As redes regulares e aleatórias calculadas pelos padrões matemáticos não explicam totalmente os fenômenos observados nas redes reais observadas, por este motivo foram desenvolvidos diversos modelos que fossem adaptados para cada necessidade estudada.

2.4.1 Modelo Watts-Strogatz

O artigo de Watts e Strogatz (1998) marcou um avanço significativo na compreensão de redes complexas, introduzindo um modelo inovador que combina características de regularidade e aleatoriedade de forma controlada. Este modelo fornece uma abordagem intuitiva e poderosa para gerar redes que exibem propriedades únicas, situando-se entre estruturas completamente regulares e redes aleatórias.

A construção do modelo é engenhosamente simples, porém profundamente impactante. Partindo de um reticulado regular em forma de anel, onde cada um dos n nós da rede está conectado a seus k vizinhos mais próximos. Em seguida, aplica-se um procedimento de reconexão aleatória: cada aresta é reconfigurada com uma probabilidade p, criando atalhos na rede. Este parâmetro p permite uma transição suave entre uma rede totalmente regular ($p = 0$) e uma completamente aleatória ($p = 1$), conforme ilustrado na Figura 3.

No artigo original, as redes resultantes foram analisadas por duas métricas fundamentais:

Comprimento de caminho característico $L_{(p)}$: Mede a separação média entre dois nós quaisquer na rede.

Coeficiente de clusterização $C_{(p)}$: Quantifica o grau de clusterização ou formação de grupos locais na rede.

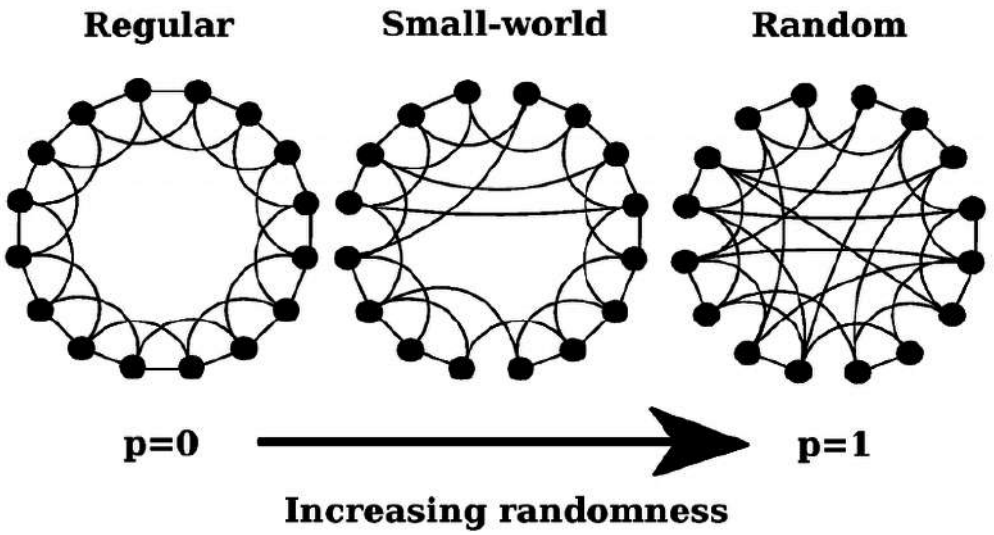


Figura 3: Processo de randomização das arestas (WATTS; STROGATZ, 1998).

A descoberta mais marcante do modelo é que, para valores intermediários de p, as redes apresentam simultaneamente um comprimento de caminho $L_{(p)}$ tão pequeno quanto em redes aleatórias e um coeficiente de clusterização $C_{(p)}$ significativamente maior. Essa combinação peculiar caracteriza as chamadas ”redes de mundo pequeno”, que conciliam alta eficiência de conexão global com fortes agrupamentos locais.

A relevância do modelo é evidenciada pela sua aplicação em diversos sistemas reais, como a rede de distribuição de energia e a rede neural do nematoide C. *elegans*. Além disso, suas propriedades têm sido exploradas em áreas como a propagação de doenças infecciosas, redes de transporte e sistemas sociais (SCHUSTER, 2011; GHOMSHEH; KAMANDI, 2022; BAKKEN; INIEWSKI, 2014).

Apesar de sua contribuição revolucionária, o modelo Watts-Strogatz apresenta algumas limitações. Ele não reproduz a distribuição de grau em forma de lei de potência observada em muitas redes reais, tratando todos os nós de maneira uniforme e ignorando a heterogeneidade intrínseca de sistemas complexos. Além disso, o processo de reconexão aleatória pode não refletir com precisão os mecanismos que formam conexões no mundo real, sendo incapaz de ponderar a importância relativa das arestas.

Ainda assim, o modelo de Watts-Strogatz permanece uma ferramenta essencial para o estudo de redes complexas, servindo como base para o desenvolvimento de abordagens mais sofisticadas. Sua importância é evidenciada por sua ampla adoção e impacto acadêmico, tendo sido citado em mais de 56.000 trabalhos científicos[1]. Mesmo com as limitações, ele continua sendo uma referência fundamental no campo da ciência de redes

[1] https://scholar.google.com/scholar?cites=4992186057191694547&as_sdt=2005&sciodt=0,5&hl=pt-BR

e um marco no entendimento das propriedades das redes de mundo pequeno.

2.4.2 Modelo de Waxman

O modelo de Waxman é utilizado para gerar redes que simulam propriedades de conectividade em sistemas reais entre cada par de nós, particularmente em redes de comunicação. Diferentemente de outros modelos que se concentram apenas em topologias regulares ou aleatórias, este modelo utiliza probabilidades dependentes da distância para criar redes com características realistas, como densidade variável e conexões de longo alcance (WAXMAN, B., 1988).

A rede é definida como um grafo $G(N, L)$, onde N é o conjunto de nós, que possuem rótulos indicando suas posições geográficas, distribuídos uniformemente em um plano bidimensional e L é o conjunto de arestas que conectam pares de nós com uma probabilidade que depende da distância espacial entre eles. Ao construir o grafo para esta rede, é utilizada a probabilidade ($p_{(u,v)}$) que define se uma aresta deve ou não ser colocada entre dois nós u e v. O valor de $p_{(u,v)}$ é obtido utilizando a Equação 2.14 a seguir:

$$P_{(u,v)} = \beta \cdot e^{\left(-\frac{d_{(u,v)}}{\alpha L}\right)} \tag{2.14}$$

onde:

$d(u, v)$ é a distância euclidiana entre os nós u e v;

L é a distância máxima entre quaisquer dois nós na rede;

β é o parâmetro que determina a probabilidade base de conexão na rede, é a probabilidade obtida quando $d_{u,v} = 0$;

α é o parâmetro que ajusta a relação entre a probabilidade de conexão e a distância, favorecendo conexões locais (α pequeno) ou distribuindo conexões mais uniformemente (α grande).

O modelo de Waxman apresenta características notáveis que o tornam altamente aplicável a sistemas reais. Primeiramente, ele combina conexões locais e globais, em que nós geograficamente próximos têm maior probabilidade de se conectarem, mas também há a presença de atalhos que ligam nós distantes, promovendo uma conectividade eficiente em escala global.

Além disso, o modelo oferece grande flexibilidade, pois, ajustando os parâmetros α e β, é possível gerar redes mais densas ou esparsas, com diferentes níveis de dependência

da distância geográfica. Essa versatilidade é especialmente relevante para aplicações em redes de comunicação, como redes de computadores ou redes de sensores, onde a distância física desempenha um papel importante na determinação dos custos e da viabilidade das conexões.

Este modelo foi originalmente desenvolvido para simular o roteamento de conexões multiponto em redes de pacotes, com foco na eficiência e na minimização de custos. Um exemplo de aplicação é o problema da árvore de Steiner, que busca minimizar o custo de interconectar um conjunto de nós em um grafo (WAXMAN, B. M., 1989).

O modelo é útil em situações de roteamento dinâmico, em que os nós podem entrar ou sair de uma conexão ao longo do tempo. Sua abordagem probabilística fornece uma base robusta para estudar a eficiência de algoritmos de roteamento e analisar redes que combinam características de conectividade local e global. Essa capacidade o torna especialmente relevante em contextos como redes de comunicação e infraestrutura de rede (LIN et al., 2018).

Apesar de sua eficácia e ampla aplicabilidade, o modelo de Waxman apresenta algumas limitações. Ele assume que a probabilidade de conexão depende exclusivamente da distância entre os nós, ignorando a heterogeneidade intrínseca dos nós ou a influência de outros atributos que podem ser determinantes na formação de conexões.

Essa simplicidade pode limitar sua capacidade de capturar dinâmicas de redes mais complexas, como redes sociais, onde fatores como interesses compartilhados, preferências ou hierarquias desempenham papéis importantes. Por essas razões, embora o modelo seja uma ferramenta poderosa para redes geográficas, ele pode exigir ajustes ou extensões para representar adequadamente redes com maior diversidade estrutural.

2.5 Fundamentos de geolocalização

A história de evolução dos mapas se funde com a própria evolução da sociedade, visto que os registros dos primeiros mapas mostram como as comunidades se preocupavam em saber sobre a sua localização. Carvalho e Araújo (2008) ilustram o primeiro mapa encontrado que representa a região que hoje é conhecida como o Iraque. Este mapa foi talhado em barro cozido com elementos que aparentam ser uma cadeia de montanhas e um rio. A Figura 4 ilustra o mapa citado.

Além disso era extremamente importante saber sobre sua localização, então com o tempo formas de se obter essa localização foram propostas e posteriormente melhoradas.

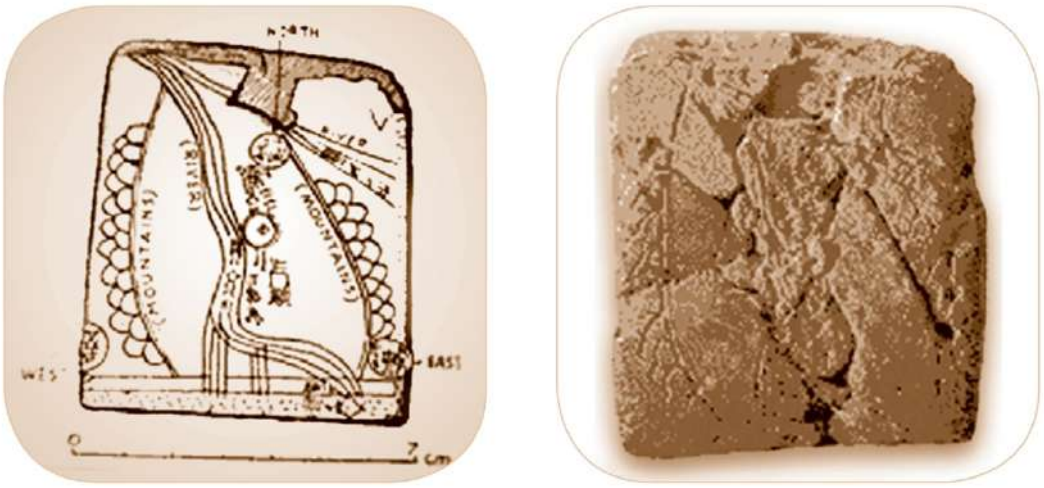


Figura 4: À esquerda, representação do primeiro mapa encontrado. À direita o mapa esculpido em barro (CARVALHO; ARAÚJO, 2008).

Começando pela observação das estrelas, foi desenvolvida a astronomia de posição que consistia em saber em que ponto da Terra a pessoa se encontrava a partir de astros previamente conhecidos no espaço visível.

Birney, Gonzalez e Oesper (2006) descrevem a Terra como uma esfera em rotação, com um eixo que conecta os polos Norte e Sul. O Equador é definido como o plano perpendicular a este eixo, passando pelo centro da esfera.

Para localizar um ponto A na superfície terrestre, dois planos adicionais são considerados: o meridiano, que passa pelo ponto A e pelos polos, e o paralelo, que é paralelo ao Equador. Esses elementos geométricos formam a base para as coordenadas geográficas.

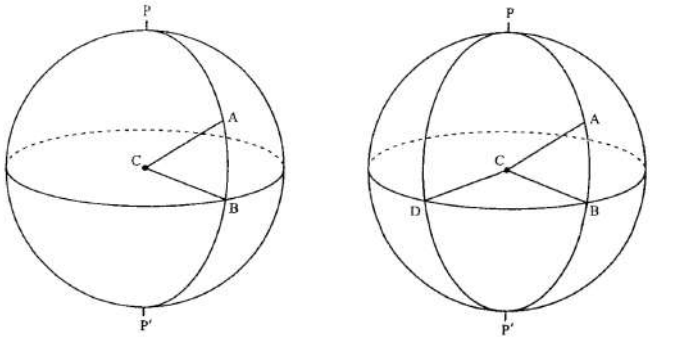
Os meridianos são linhas imaginárias que unem os polos. O meridiano que passa por *Greenwich*, Londres, foi adotado em 1884 como a linha de longitude 0°, Ridpath (2012). A longitude é medida como a distância angular de um ponto até esse meridiano, paralelamente ao Equador, variando de $180°$ a leste a $-180°$ a oeste.

A latitude, por sua vez, é definida como o ângulo formado entre um ponto no globo e a linha do Equador, ISO Central Secretary (2022). Varia de $0°$ a $90°$, do Equador aos polos, usando as letras N e S ou valores positivos e negativos para indicar os hemisférios Norte e Sul, respectivamente.

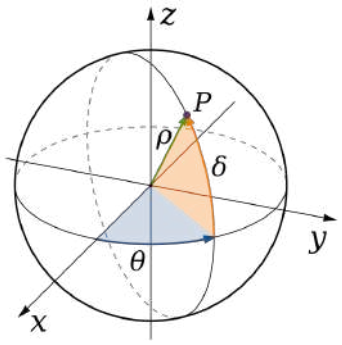
Para calcular as coordenadas de um ponto A, traça-se uma curva na superfície da esfera que liga os polos e passa pelo ponto desejado. O ângulo $B\widehat{C}A$, em que C é o centro

da esfera e B é a interseção da curva com o plano equatorial, representa a latitude. O ângulo $D\widehat{C}B$, formado entre o plano do meridiano de *Greenwich* e o plano que passa por B, representa a longitude.

Este sistema permite a localização de um ponto na superfície terrestre através da observação dos astros. As Figuras 5a e 5b ilustram, respectivamente, o esquema para o cálculo de latitude e longitude em uma esfera, e a representação de um ponto P no globo com latitude δ e longitude Θ.



(a) À esquerda, a latitude ($B\widehat{C}A$). À direita, a longitude ($D\widehat{C}B$) (BIRNEY; GONZALEZ; OESPER, 2006)



(b) Ponto P na superfície terrestre (WIKIPEDIA CONTRIBUTORS, 2008).

Embora o Sistema de Coordenadas Geográficas (SCG) seja bem definido e coeso com a realidade do planeta, sua representação em produtos cartográficos, como mapas, não pode ser feita diretamente nesse formato. Isso ocorre porque é necessário transformar a superfície esférica da Terra em um plano, o que inevitavelmente gera deformações na estrutura original.

Para lidar com essas deformações, diversos métodos de projeção foram desenvolvidos, cada um com o objetivo de preservar características específicas, como ângulos, áreas ou distâncias. Esses métodos deram origem aos Sistemas de Coordenadas Projetadas (SCP), que utilizam um sistema métrico cartesiano para representar as coordenadas (x, y), facilitando medições, cálculos e análises em um espaço bidimensional (D'ALGE, 2001).

Lapaine e Frančula (2022) explicam que essas projeções podem ser classificadas em três grupos principais, de acordo com a propriedade preservada:

Projeções Conformes : Mantêm os ângulos dos objetos representados, preservando sua forma, mas não suas áreas ou distâncias. (Figura 6a)

Projeções Equivalentes : Conservam as proporções de área dos objetos, permitindo análises precisas de tamanho relativo. (Figura 6b)

Projeções Equidistantes : Garantem a proporção das distâncias entre pontos específicos, embora nem todos os pontos mantenham suas relações de distância. (Figura 6c)

(a) Projeção conforme **(b)** Projeção equivalente **(c)** Projeção equidistante

Figura 6: Projeções por propriedades mantidas (IBGE, 2024).

Além dessas, as projeções também podem ser classificadas com base em sua superfície e aspecto. No que diz respeito à superfície, destacam-se:

Projeções Cilíndricas : Produzem mapas retangulares, onde linhas horizontais representam os paralelos (círculos paralelos ao Equador) e linhas verticais representam os meridianos. (Figura 7a)

Projeções Cônicas : Apresentam os meridianos como linhas retas que convergem em um ponto comum, com os paralelos desenhados como curvas concêntricas. Essas projeções formam um setor circular semelhante a um cone aberto. (Figura 7b)

Projeções Azimutais (ou Polares) : Similar às cônicas, mas representam o mapa em um círculo completo, com os meridianos irradiando de um ponto central. (Figura 7c)

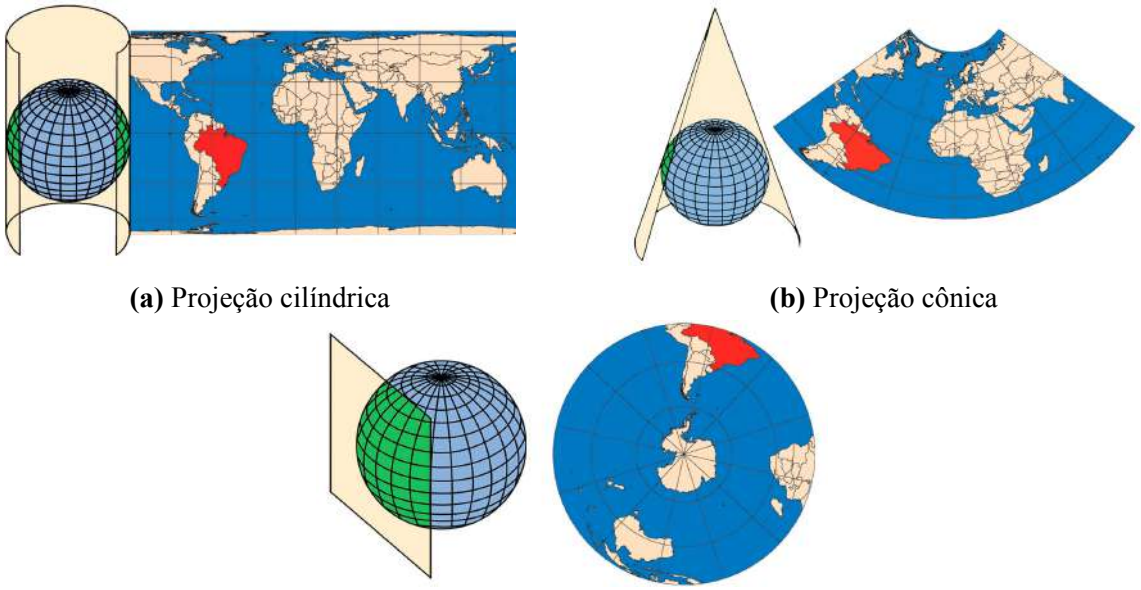
Quanto ao aspecto, as projeções são classificadas em:

Normais : O ponto de projeção coincide com o eixo da esfera. (Figura 8a)

Transversais : O ponto de projeção é perpendicular ao eixo da esfera. (Figura 8b)

Oblíquas : O ponto de projeção forma um ângulo arbitrário com o eixo da esfera. (Figura 8c)

Cada projeção combina uma dessas classificações de forma e aspecto. Por exemplo, uma projeção Equidistante Azimutal Normal (Figura 9) mostra o globo a partir de um



(a) Projeção cilíndrica

(b) Projeção cônica

(c) Projeção azimutal ou polar

Figura 7: Projeções por superfície (IBGE, 2024).

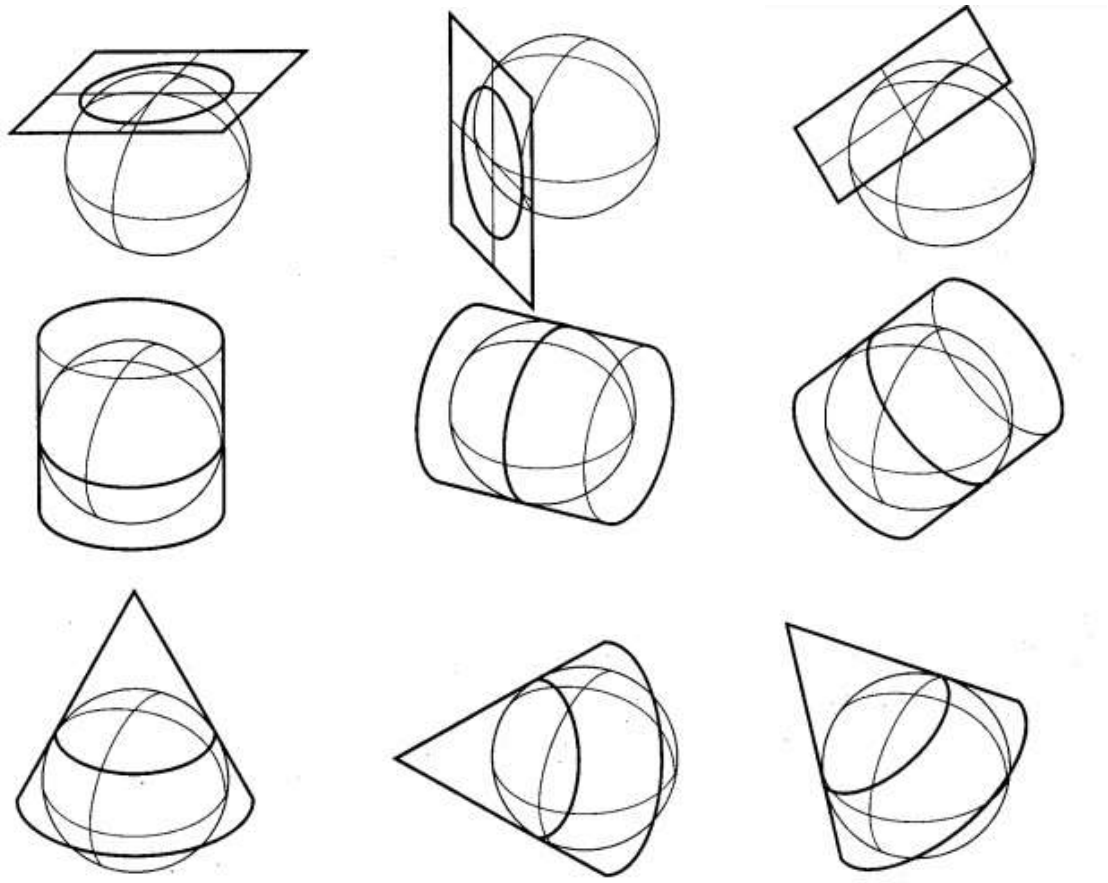
ponto central. Um exemplo clássico dessa projeção apresenta a Groenlândia no centro, com os continentes ao seu redor, e a Antártica na borda externa do mapa.

Ao longo dos anos, muitas projeções foram desenvolvidas, mas a Universal Transversa de Mercator (UTM) destaca-se como a mais utilizada, especialmente para mapas de pequena e média escala. Recomendada pela *International Union of Geodesy and Geophysics* (IUGG), a projeção UTM é cilíndrica, transversal e conforme. Essa configuração projeta o globo em um cilindro com eixo perpendicular ao eixo terrestre, preservando a forma dos objetos representados, embora suas dimensões sejam escaladas (OLIVEIRA; SILVA, 2012).

O sistema UTM divide a superfície terrestre em 60 fusos, cada um com aproximadamente 670 km de largura, equivalente a 6° de longitude, numerados de 1 a 60, a partir do anti-meridiano de *Greenwich* em direção ao leste.

Com os avanços tecnológicos e a padronização dos sistemas de coordenadas, surgiram novas formas de determinar a localização na Terra no sistema geográfico. Atualmente, o método mais comum é o uso do *Global Navigation Satellite System* (GNSS), um termo genérico que descreve sistemas de posicionamento, navegação e cronometragem baseados em constelações de satélites e receptores terrestres (NATIONAL COORDINATION OFFICE FOR SPACE-BASED POSITIONING, NAVIGATION, AND TIMING, 2022).

Entre os sistemas GNSS disponíveis, destacam-se o BeiDou, desenvolvido pela China, que opera com uma constelação de 35 satélites; o Galileo, da União Europeia, composto



(a) Projeção normal **(b)** Projeção transversal **(c)** Projeção oblíqua

Figura 8: Projeções por aspecto (VIEIRA et al., 2004).

por 24 satélites; e o GLONASS, da Rússia, que também conta com 24 satélites em sua constelação. Cada um desses sistemas é projetado para oferecer precisão e cobertura global, desempenhando um papel fundamental em diversas aplicações, como navegação, monitoramento ambiental e agricultura de precisão.

No entanto, o mais conhecido é o Global Positioning System (GPS), desenvolvido pelos Estados Unidos. O GPS conta com 31 satélites operacionais, garantindo que pelo menos 24 estejam ativos em 95% do tempo, fornecendo sinais para localização e cronometragem. Esses satélites orbitam a uma altitude de aproximadamente 20,200 km, completando duas passagens sobre os mesmos pontos a cada dia, garantindo cobertura global para os usuários.

Por fim, a geolocalização desempenha um papel central na representação e análise de fenômenos espaciais, conectando conceitos fundamentais de sistemas de coordenadas, projeções cartográficas e tecnologias modernas de posicionamento. A transição de sistemas de coordenadas geográficas para sistemas projetados permite a manipulação precisa de dados espaciais, enquanto métodos de projeção garantem que características

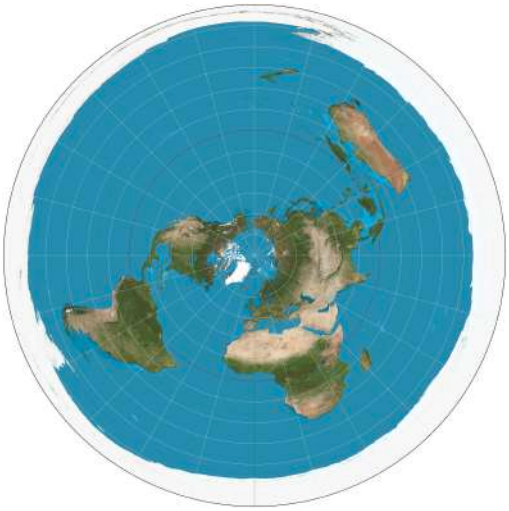


Figura 9: Projeção Equidistante Azimutal Normal

críticas sejam preservadas de acordo com o objetivo do mapa ou estudo. Além disso, avanços em sistemas globais de navegação por satélite reforçam a capacidade de coletar e utilizar dados geoespaciais em escala global. Com essas ferramentas e fundamentos, a geolocalização se estabelece como uma base essencial para aplicações que vão desde o planejamento urbano e ambiental até a análise de redes e sistemas complexos.

3. Análise Exploratória

Para iniciar este trabalho, foi realizado um levantamento na internet em busca de *datasets* que representassem redes sociais e contivessem dados geolocalizados vinculados às entidades definidas como nós da rede. A ideia central era analisar as estruturas das comunidades identificadas para compreender o impacto do componente geográfico dentro das redes sociais.

Embora existam muitos estudos na área de redes sociais e algumas bases de dados com volumes consideráveis de informações, a maioria delas não apresentava dados geográficos ou não possuía um número suficiente de entidades para análise. Por conta dessas limitações, as opções disponíveis foram reduzidas. Assim,este trabalho utilizou parte dos dados e código disponibilizados por Tang e Painho (2023) que apresentam em seu artigo um *framework* para explorar as relações entre conteúdo e localização em dados gerados por usuários de fontes baseadas em espaço (*Twitter*) e baseadas em lugar (*Google Places* e *OpenStreetMap* - OSM) em ambientes urbanos. O objetivo principal é analisar até que ponto as informações encontradas em atividades de mídia social georreferenciadas correspondem ao contexto espacial.

Os autores propuseram uma metodologia inovadora que combina diversas técnicas analíticas para aprimorar os estudos sobre as relações entre conteúdo textual e localização geográfica. Essa abordagem foi aplicada para investigar a correlação entre informações compartilhadas em redes sociais e características do ambiente urbano, com um estudo de caso realizado na cidade de Lisboa, Portugal.

A metodologia utiliza o modelo de tópicos BERTopic, que combina *embeddings* com algoritmos de agrupamento, para extrair temas a partir do conteúdo textual agregado. A análise foi conduzida separadamente em dois conjuntos de dados: *tweets* georreferenciados (dados baseados no espaço) e informações de pontos de interesse extraídas do *Google Places* e *OpenStreetMap* (dados baseados no conceito de lugar).

Os tópicos identificados nas duas fontes foram comparados por meio da métrica de similaridade de cosseno, permitindo avaliar a proximidade semântica entre os temas extraídos.

Além disso, foi realizada uma análise de *hotspots* utilizando o método Getis-Ord G*, que identifica áreas estatisticamente significativas associadas a cada tópico, destacando regiões com alta concentração de temas relevantes. Para medir a sobreposição espacial entre tópicos similares extraídos das duas fontes, os autores aplicaram o índice de Jaccard, fornecendo uma métrica quantitativa da interseção espacial.

A proposta desse *framework* oferece uma abordagem reprodutível e eficiente para explorar as complexas relações entre conteúdo digital e ambiente urbano. Ao integrar modelagem de tópicos, análise espacial e métricas de similaridade, o método permite estudar como as informações compartilhadas em plataformas digitais refletem e interagem com o contexto geográfico e social de diferentes cidades. Essa metodologia não apenas enriquece o entendimento das dinâmicas urbano-digitais, mas também oferece uma base para estudos comparativos entre diversos cenários urbanos.

Os resultados obtidos por Tang e Painho (2023) mostraram que a correlação entre o conteúdo online e o contexto urbano varia significativamente dependendo dos perfis temáticos. Tópicos como futebol e aeroporto apresentaram alta correlação entre conteúdo e localização, enquanto outros temas mostraram relações mais complexas. O estudo concluiu que as relações conteúdo-localização são intrincadas e dependem fortemente das assinaturas temáticas, destacando a importância de uma análise cuidadosa ao usar conteúdo gerado por usuários em estudos urbanos.

O processo desenvolvido nas seções a seguir pode ser estruturado em etapas sequenciais que integram a limpeza dos dados, processamento dos dados e as análises realizadas. Inicialmente, foi conduzida uma análise exploratória do dataset, na qual foram identificadas as principais características da base de dados, incluindo a presença de *tweets* georreferenciados e informações disponíveis sobre as menções entre usuários.

Em seguida, vai ser proposta a construção da rede de menções, na qual cada conta do *Twitter* é representada como um nó e as interações via menção são modeladas como arestas direcionadas. Antes da criação da rede, será necessário um pré-processamento dos dados, que envolve a filtragem dos *tweets* para manter apenas aqueles que estejam georreferenciados.

Com a rede construída, serão calculadas métricas estruturais básicas, incluindo a distribuição de graus e o coeficiente de clusterização. Além disso, foi realizada a

identificação da Componente Conexa Gigante (CCG), que vai ser o principal foco da análise, uma vez que concentra a maior parte das interações e reflete melhor a estrutura social do conjunto de dados. No aspecto geográfico, será determinada a localização representativa dos usuários, utilizando a mediana das coordenadas de suas postagens.

Para investigar os agrupamentos da rede e compreender se existe uma correlação visível entre eles, será aplicado o algoritmo de Louvain para a identificação das comunidades baseadas nas conexões da rede social. Paralelamente, os agrupamentos geográficos utilizarão o método *K-means* com otimização do número de grupos pelo *elbow method*, visando identificar padrões espaciais entre os usuários.

Esse fluxo de processamento está representado pelo diagrama da Figura 10. Ele integra tanto análises estruturais da rede social quanto análises espaciais, permitindo uma compreensão mais abrangente das interações entre usuários do Twitter no contexto geográfico estudado.

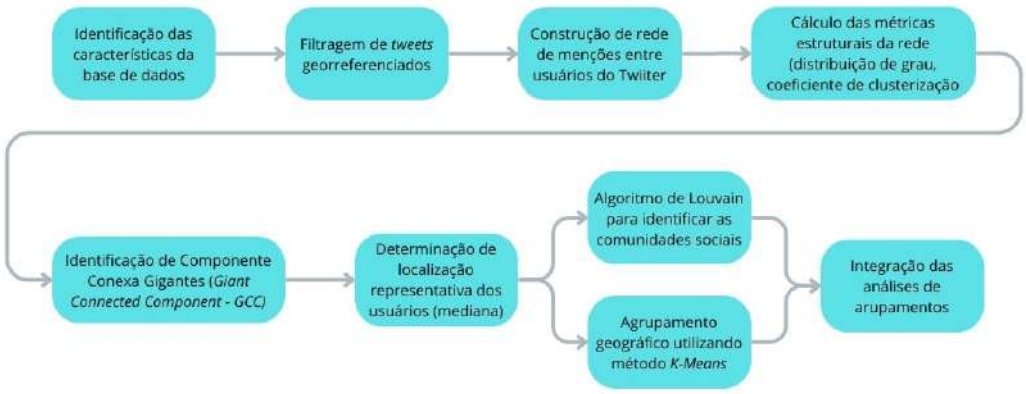


Figura 10: Fluxo de atividades desenvolvidas na Análise Exploratória

3.1 *Dataset* utilizado

O estudo de Tang e Painho (2023) utilizou dados coletados e pré-processados para garantir qualidade e relevância. No caso do *Twitter*, foram utilizados *tweets* georreferenciados dentro de um raio de 40 km do centro de Lisboa, filtrados para incluir apenas mensagens em português e limitar a contribuição por usuário, evitando a superrepresentação de determinados indivíduos. Já os dados do *Google Places* e OSM incluíram pontos de interesse, nomes de lugares e avaliações de usuários localizados dentro dos limites da cidade.

Os dados textuais foram agregados em uma grade hexagonal que cobria toda a cidade,

onde cada célula foi tratada como um ”documento”para a modelagem de tópicos. Esse processo permitiu uma análise espacial detalhada das informações. Adicionalmente, foi realizado um pré-processamento padrão, incluindo a remoção de caracteres especiais, emojis, URLs e outros elementos irrelevantes, garantindo maior clareza e consistência nos textos analisados.

Devido às diferenças nos objetivos entre este estudo e a pesquisa original, algumas adaptações foram realizadas. Diversas bibliotecas anteriormente empregadas não foram necessárias, uma vez que o presente trabalho não utiliza recursos de inteligência artificial nem processamento de linguagem natural (PLN). Foram mantidos principalmente os *tweets* com seus respectivos atributos, além dos métodos de leitura de arquivos no formato “.JSONL” e da conversão dessas informações para tabelas, onde cada linha representa um *tweet* individual.

Além disso, as coordenadas geográficas fornecidas no *dataset* estavam expressas em graus decimais, correspondendo a um sistema de referência geográfico comum, mas pouco adequado para cálculos mais avançados de distância ou análise espacial. Para superar essa limitação e facilitar o processamento posterior, foi realizada uma conversão para um sistema de referência projetado em unidades métricas. A conversão foi feita utilizando a biblioteca *utm* do *Python*, que transformou os valores de latitude e longitude para coordenadas em metros dentro do sistema UTM (*Universal Transverse Mercator*). Esse procedimento não só simplificou a interpretação das distâncias como também tornou os cálculos mais precisos e diretos.

Para o processamento dos dados, foram empregadas as seguintes bibliotecas com finalidades específicas:

Pandas 2.1.2 : Criação e manipulação de objetos em formato de tabelas;

Tweepy : Acesso à API do *Twitter* para recuperação dos nomes de usuários a partir do ID do autor do *tweet*;

Além dessas, a NetworkX versão 3.3 foi amplamente utilizada e desempenhou um papel central no processamento das redes. Essa biblioteca foi responsável por toda a manipulação, desde a criação das redes e cálculo de métricas básicas até a identificação das comunidades. NetworkX é uma biblioteca robusta e versátil para análise e exploração de grafos e redes complexas, desenvolvida inicialmente em 2002.(HAGBERG; SCHULT; SWART, 2008)

Entre suas principais características, destaca-se a flexibilidade na manipulação de

estruturas de dados, permitindo trabalhar com diversos tipos de grafos, incluindo grafos simples, direcionados, multigrafos e grafos com auto-aresta (em que um nó se conecta a ele mesmo). Além disso, os nós das redes podem ser representados por qualquer objeto *hashable* (objetos como *string* e inteiros que não mudam) do Python, oferecendo alta adaptabilidade na modelagem dos dados.

Outro diferencial importante do NetworkX é sua integração com outras bibliotecas do ecossistema Python. Ela aproveita bibliotecas como NumPy e SciPy para cálculos matriciais e operações de álgebra linear, além de oferecer interfaces para ferramentas de visualização como Matplotlib e Graphviz. Essa integração facilita a combinação de suas funcionalidades com outras ferramentas analíticas, tornando-a uma escolha poderosa para análise de redes e grafos em ambientes computacionais modernos.

Após o carregamento dos dados no ambiente do *Jupyter Notebook*, foram conduzidas análises iniciais para levantar estatísticas básicas sobre os *tweets*. Inicialmente foram encontradas na base de dados 10,088,848 postagens, das quais 2/3 contêm o metadado de localização. A Figura 12a ilustra a proporção entre os dados com e sem a *geotag*.

Baseando-se no código de Tang e Painho (2023), foram excluídas linhas da tabela que não continham informações de localização geográfica ou apresentavam coordenadas repetidas. Como resultado, obteve-se um total de 3,454,314 *tweets*, associados a 60,812 usuários únicos, o que corresponde a uma média de 56.8 *tweets* por usuário. A Figura 11 apresenta a distribuição dos tweets por usuário dessa base de dados, a partir dela é possível perceber que poucos usuários fizeram muitas postagens e vice-versa.

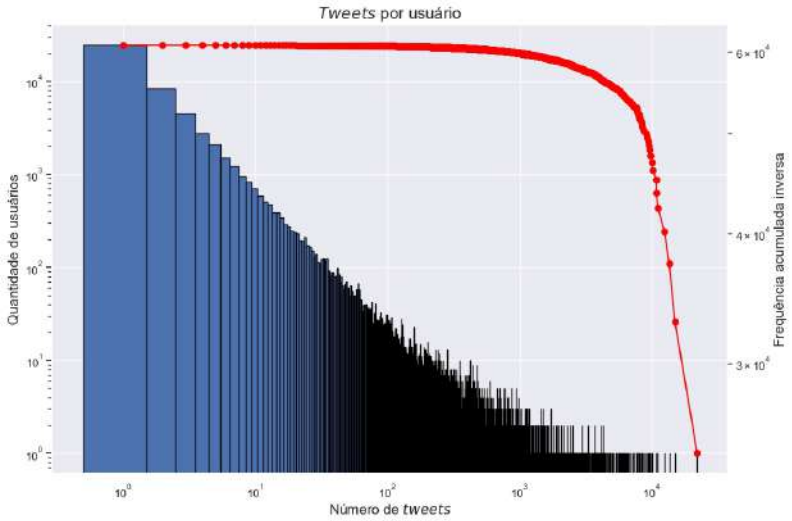


Figura 11: Histograma e CCDF dos *tweets* por usuários.

Em seguida, foram analisadas as menções da coluna *entities_mention*, que registra os usuários mencionados em cada postagem. Foi desenvolvido um método para associar a

cada usuário uma lista de contas mencionadas. Essa análise revelou 514,752 menções distribuídas entre 103,270 contas mencionadas, número aproximadamente 70% superior ao total de usuários únicos na base de dados. Esse dado sugere a existência de contas mencionadas que não possuem informações de geolocalização ou postagens associadas. Ainda considerando as menções e sua relação com os usuários únicos identificados, foi calculado que cada usuário mencionou, em média, 30.5 menções por usuários, sendo mencionadas 6.7 contas distintas.

Num levantamento posterior foi realizada a contagem total de *tweets*, incluindo os que tivessem quaisquer tipos de repetição, com e sem valor no campo de *geotag*, campo este que é usado para obter os valores de latitude e longitude de forma direta.

A base de dados continha ainda informações sobre anotações, que fornecem uma ideia do contexto do *tweet*, anotações de entidade são baseadas em algo que está escrito explicitamente no texto da postagem e anotações de contexto que surgem a partir da análise do texto em conjunto com um par domínio e entidade para descobrir o tópico do assunto. As anotações de entidade são divididas em:

Pessoa : o texto é sobre uma pessoa (Fernanda Torres, Walter Salles, Ed);

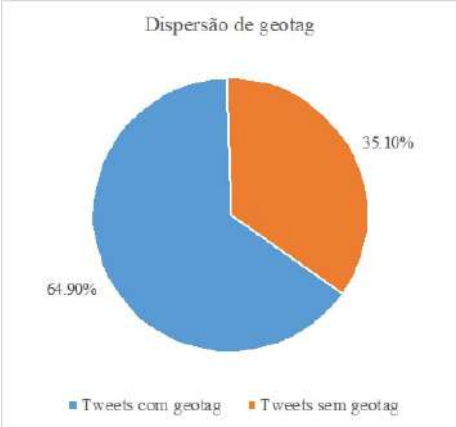
Lugar : o texto é sobre algum lugar (Rio de Janeiro, Curitiba, Bristol);

Produto : o texto fala de uma marca de produto (Google Chrome, Redragon);

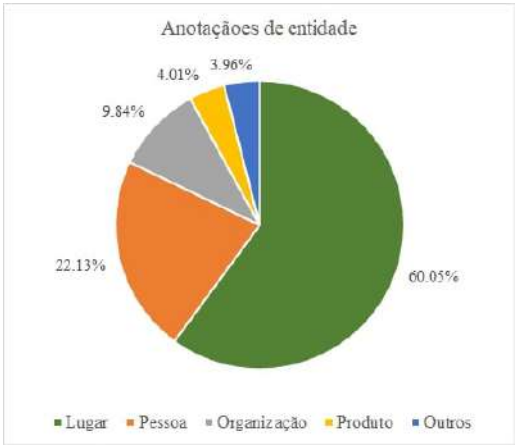
Organização : o texto é sobre uma empresa (esri, Cacau Show);

Outros : o texto não se encaixa nas outras definições (Oscar 2025, Carnaval).

Neste trabalho, foram coletadas as anotações de entidade dos *tweets* que resultaram em uma divisão mostrando que a grande maioria se concentra no assunto “Lugar”. A Figura 12b ilustra como os dados estão divididos entre as anotações.



(a) Gráfico em pizza da dispersão de *tweets* com e sem *geotag*



(b) Gráfico em pizza da dispersão de anotações de entidade dos *tweets*

3.2 Primeira Análise: Rede de Menções

Devido às dificuldades no uso e aquisição de informações da API oficial do *Twitter* que impõe várias barreiras financeiras nos seus dados, inclusive para o desenvolvimento de pesquisas científicas, o trabalho realizado aqui mudou parte do seu desenvolvimento. Neste ponto, a ideia é criar uma rede direcionada em que cada nó defina um usuário identificado pelo seu *username* e as arestas são colocados no sentido da menção, ou seja, um usuário u menciona outro usuário v desse modo é incluído o elo (u, v) apenas, já que a direção da menção importa nesse caso.

A tabela contendo os *tweets* incluía um atributo que armazenava o identificador único dos autores das postagens. No entanto, esse identificador não permitia estabelecer conexões diretas entre os autores e as pessoas mencionadas, uma vez que o identificador das contas mencionadas não estava disponível. Para contornar essa limitação, foi necessário recorrer à API do Twitter, que possibilita a recuperação do nome de usuário a partir do identificador único (ID), mesmo em contas gratuitas. Essa funcionalidade era essencial para o desenvolvimento de uma rede de menções, já que trabalhar apenas com IDs tornaria a análise pouco prática e desconexa com os objetivos do estudo.

Para simplificar o processo de interação com a API, foi utilizada a biblioteca Tweepy, que fornece uma interface eficiente para acessar diversos *endpoints* e manipular informações do Twitter. A configuração da biblioteca exigiu o fornecimento de credenciais de autenticação, incluindo: *Consumer Key*, *Consumer Secret*, *Access Token*, *Access Token Secret* e *Bearer Token*. Todas essas chaves foram obtidas por meio da página de desenvolvedores do Twitter, utilizando um projeto gratuito, o que possibilitou o acesso

necessário para a extração dos dados complementares.

O uso do Tweepy facilitou a integração da API ao fluxo de trabalho, permitindo recuperar os nomes de usuários mencionados e, assim, criar uma rede de menções consistente e alinhada aos objetivos da pesquisa.

Apesar dessas técnicas de enriquecimento do *dataset*, ainda houve problemas relacionados ao fato de que a API limita a quantidade de buscas que podem ser realizadas quando se usa a conta gratuita. Com isso, foram processados aproximadamente mil linhas a cada quinze minutos.

A rede foi construída considerando que cada nome de usuário presente na base de dados representava um nó, enquanto cada menção de um usuário a outro era modelada como uma aresta direcionada entre esses nós. Após a montagem do grafo, foram analisadas algumas de suas métricas básicas, e a partir de sua visualização (*plot*), identificou-se a presença de diversas conexões isoladas, ou seja, arestas formadas apenas entre pares de nós sem outras ligações.

Apesar disso, também foi notada a existência de uma Componente Conexa Gigante (CCG), que englobava a maior parte dos nós conectados (aproximadamente 72%), indicando que, embora a rede como um todo seja pouco conectada, há uma subestrutura principal que concentra a maior parte das interações. A Tabela 1 apresenta as métricas obtidas tanto para o grafo completo quanto para sua CCG. Os resultados mostram que, embora as conexões degradadas (aquelas isoladas ou pouco conectadas) estejam presentes, sua influência na avaliação geral da rede é limitada, sendo a CCG o elemento predominante na análise estrutural.

Tabela 1: Estatísticas dos grafos iniciais

Grafo	Completo	Componente gigante
Número de nós	123,532	89,100 (72.13%)
Número de arestas	174,656	150,212 (86.00%)
Grau médio	2.8277	3.3717
Densidade	1.1445×10^{-5}	1.8921×10^{-5}
CC Global	0.0027	0.0027
CC Médio	0.0142	0.0191

Essas observações ressaltam a importância de focar na CCG para análises mais detalhadas, uma vez que ela concentra a maior parte da dinâmica e da conectividade da rede.

Ao analisar a rede correspondente à CCG, foi realizada a plotagem da Distribuição Complementar Cumulativa (CCDF) para os graus de entrada e saída dos nós, conforme

mostrado na Figura 13. A análise revelou a presença clara do efeito de cauda pesada, característico de redes sociais. Esse efeito descreve a concentração desigual de conexões, onde poucos nós apresentam graus extremamente altos, enquanto a maioria dos nós possui graus baixos.

Esse comportamento é particularmente evidente nos nós que representam menções, indicando a existência de *hubs*, ou seja, usuários que são mencionados com alta frequência. Esses *hubs* desempenham um papel central na estrutura da rede, atuando como pontos de alta interação e conectividade. No contexto da rede analisada, a grande escala do grafo amplifica essa característica, evidenciando uma forte centralização das atividades em torno de um pequeno número de nós-chave. Essa estrutura de centralização reflete padrões típicos de redes sociais digitais, onde poucos atores possuem grande influência ou visibilidade, enquanto a maioria dos participantes tem interações mais limitadas.

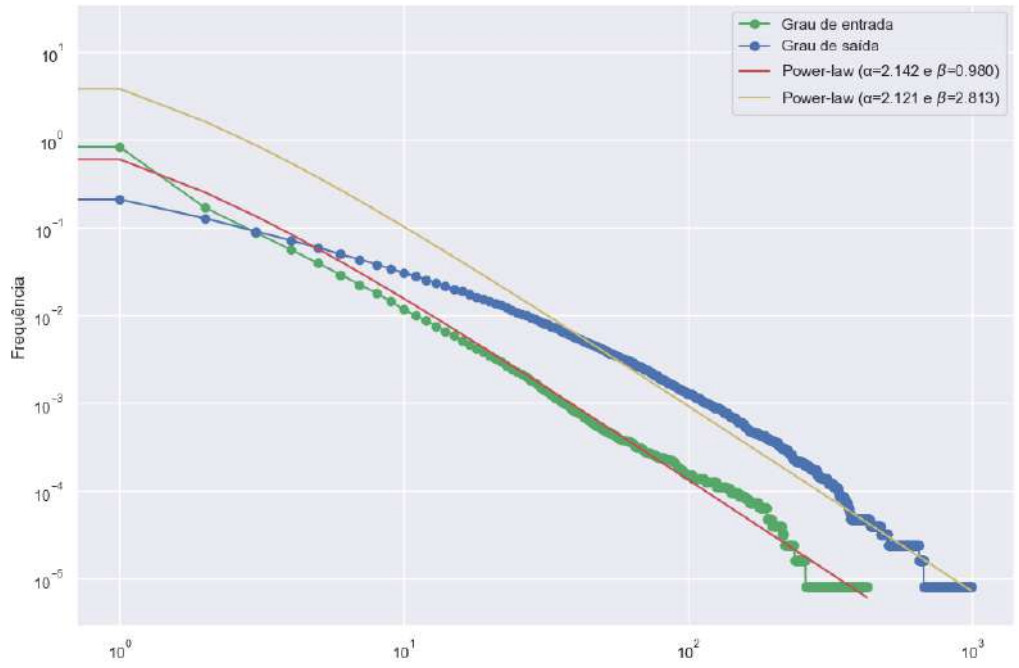


Figura 13: Histograma da Distribuição Complementar Cumulativa (CCDF) dos graus de entrada e saída.

Com base nos dados analisados, as curvas obtidas foram ajustadas a uma distribuição do tipo lei de potência, utilizando uma interpolação polinomial para minimizar o erro quadrático. Dessa forma, foram calculados os valores da potência (α) para os graus de entrada e saída da rede. O valor estimado para o grau de entrada foi de 2.332, enquanto para o grau de saída foi de 2.121. Esses resultados corroboram a hipótese de que a rede segue um padrão de organização típico de redes complexas, com forte hierarquia e centralização em poucos nós que exercem maior influência na dinâmica geral.

Com o objetivo de estudar a variável geográfica da rede, foi necessário determinar uma localização representativa para os nós, visto que, como eles representam os nomes dos usuários e alguns destes postaram de diferentes locais, havia variação nas posições dos nós. Para isso, foi adotado o cálculo da mediana das coordenadas de todas as postagens feitas por um mesmo autor, uma vez que a mediana é menos sensível a *outliers* e oferece uma estimativa mais robusta da localização central de suas atividades. Esse processo resultou na redução do número de pontos para um total de 116.087, conforme mostrado na Tabela 2.

Tabela 2: Estatísticas do grafo filtrado

Grafo	
Número de nós	116,087
Número de arestas	20,035
Grau médio	0.3451
Densidade	1.487×10^{-6}
CC global	0.0321
CC médio	0.0036

A figura 14 mostra que a distribuição dos graus no grafo após a filtragem mantém a tendência de uma curva de uma distruibção *power law*, com $\alpha = 2.769$ para os graus de entrada e $\alpha = 2.381$ para os graus de saída.

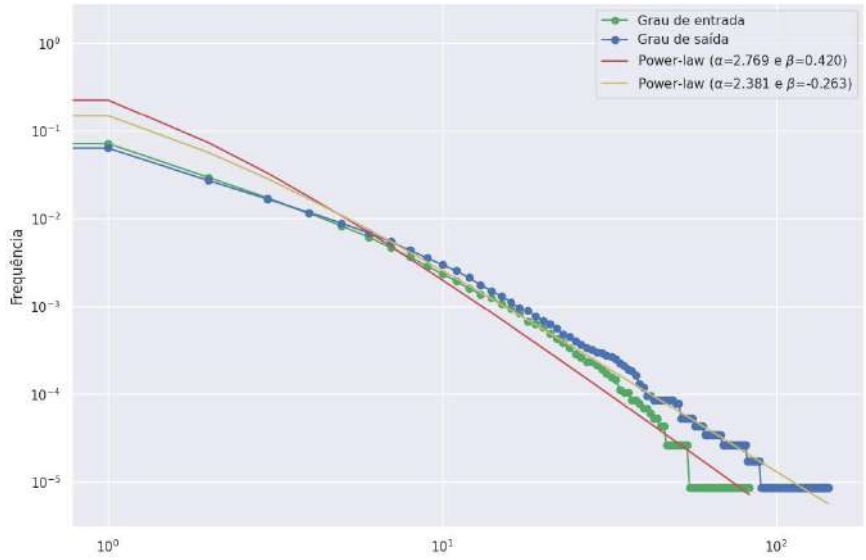


Figura 14: Distribuição Complementar Cumulativa (CCDF) dos graus de entrada e saída do grafo filtrado

Com as coordenadas já transformadas, os dados foram exportados para o software QGis, uma ferramenta de Sistema de Informação Geográfica (SIG). Essa exportação

permitiu a realização de análises espaciais mais detalhadas, além de visualizações que evidenciaram a distribuição geográfica dos usuários. No QGis, foi possível gerar mapas que destacam as posições médias (calculadas pela mediana) dos usuários, revelando padrões interessantes na ocupação espacial dos dados analisados.

Esse processamento também possibilitou a criação de camadas temáticas, que ajudaram a identificar concentrações regionais de usuários e padrões de movimentação ou aglomeração. A Figura 15 ilustra a distribuição espacial desses usuários, destacando as localizações representativas derivadas do conjunto de dados analisados. Este mapa não só reforça a importância de uma análise geográfica detalhada, mas também evidencia como a metodologia adotada foi eficaz para lidar com os desafios da manipulação de grandes volumes de dados espaciais.

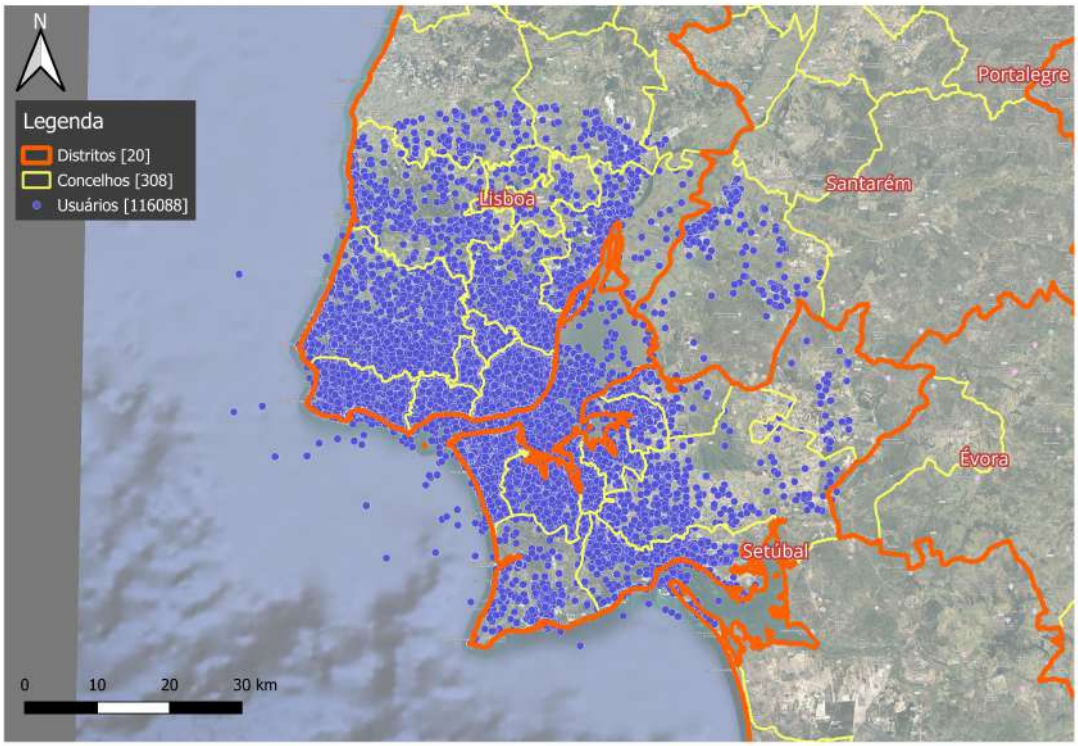


Figura 15: Distribuição geográficas dos usuários

Como é possível ver na imagem apresentada alguns pontos que representam os usuários do *Twitter* estão dispostos dentro do mar, este fenômeno ocorre devido à falta de acurácia dos receptores GNSS presentes nos dispositivos móveis utilizados, juntamente com o fato de utilizar a mediana dos dados que ocasionalmente pode gerar essa inconsistência.

A abordagem adotada, portanto, demonstra o valor de combinar ferramentas de programação, como *Python*, com aplicativos especializados, como o QGIS, para explorar

com profundidade tanto os aspectos geográficos quanto os comportamentais do conjunto de dados. Essa integração de métodos foi essencial para garantir uma análise precisa e visualmente clara da dinâmica espacial envolvida.

3.3 Segunda Análise: Estrutura de comunidades

Uma área de grande interesse deste trabalho é a análise do comportamento das comunidades, tanto aquelas intrínsecas à estrutura da rede quanto os agrupamentos formados pela proximidade espacial dos pontos geográficos. Entretanto, não foi possível encontrar uma forma de fazer uma análise conjunta destes dois conceitos, portanto foi necessário utilizar dois métodos de agrupamento dos dados.

Na Tabela 3 são descritas as principais diferenças entre as duas grandezas, bem como as formas de identificar esses grupos e quais os métodos que este trabalho vai utilizar para cada um.

Tabela 3: Distinções entre comunidades sociais e agrupamentos geográficos

Comunidades sociais	Agrupamentos geográficos
Baseado nas interações	Baseado na localização geográfica
Identificado por algoritmos de detecção	Definido por algoritmos de agrupamento
Algoritmo de Louvain	Método *K-Means* otimizado pelo *elbow method*

Para iniciar essa investigação, foi realizada uma análise detalhada do coeficiente de clusterização para cada nó da rede que possui apenas nós com dado geográfico descrita na Seção 3.2. Esse coeficiente é uma medida importante para compreender o grau de conectividade local e a formação de subgrupos na rede.

Os valores obtidos foram visualizados por meio do gráfico apresentado na Figura 16 que exibe um gráfico que segue a ideologia do CCDF onde o eixo x tem os valores de coeficiente de clusterização presentes no gráfico e o eixo y tem a probabilidade maior ou igual daquele valor ocorrer. Neste gráfico foram desconsiderados os nós com coeficiente igual 0, que representam nós isolados ou com apenas uma aresta. A partir da análise do gráfico, observa-se que a probabilidade associada ao coeficiente de clusterização diminui à medida que o valor do coeficiente aumenta.

Esse comportamento é característico de redes onde apenas uma pequena fração dos nós apresenta altos níveis de interconectividade entre seus vizinhos, enquanto a maioria dos nós possui coeficientes de clusterização mais baixos. Essa distribuição reflete a estrutura heterogênea da rede, onde a formação de grupos fortemente conectados ocorre com maior

frequência em regiões específicas da rede, enquanto outras áreas permanecem menos densamente conectadas.

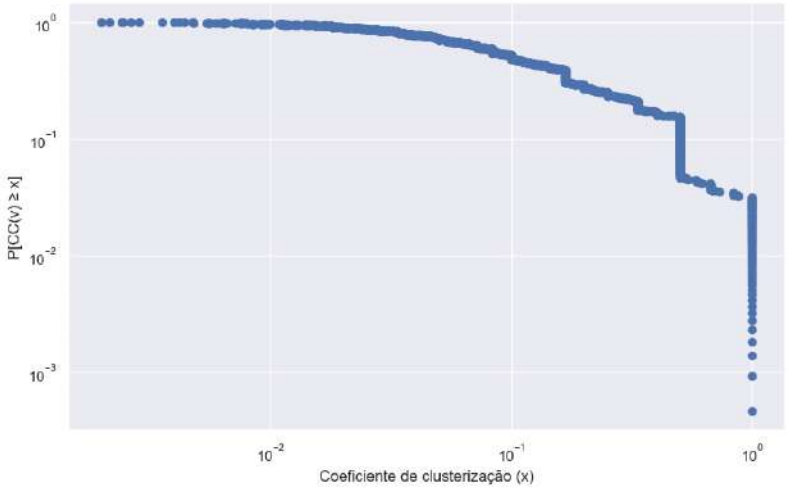


Figura 16: CCDF do coeficiente de clusterização dos nós

Ainda visando analisar o comportamento da rede, foi plotado na Figura 17 a distribuição do coeficiente de clusterização para os diferentes valores de grau dos nós, realçando as propriedades esperadas em uma rede real, onde o coeficiente decresce à medida que o valor do grau aumenta.

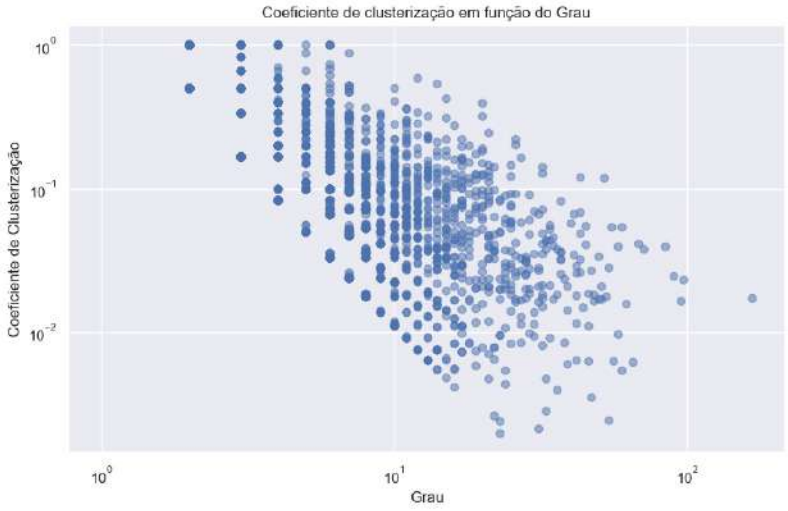


Figura 17: Coeficiente de clusterização por grau

Após essa análise, torna-se necessário pensar em métodos para avaliar os agrupamentos, considerando a componente geográfica. Quando não se está avaliando uma estrutura de rede, algumas análises podem exigir o uso de métodos de clusterização, que são técnicas não supervisionadas de aprendizagem amplamente utilizadas em *data mining*, *machine learning* e reconhecimento de padrões. Esses métodos têm como objetivo agrupar dados similares enquanto separam dados divergentes em diferentes *clusters*,

funcionando de forma análoga à ideia de comunidades em redes (SAEED; AL AGHBARI; ALSHARIDAH, 2020).

Um dos métodos de clusterização mais populares é o *K-Means*, amplamente utilizado por sua simplicidade e eficiência. Conforme apresentado por MacQueen (1967), o *K-Means* é um método baseado em partições que divide os dados em K *clusters* pré-definidos, onde cada dado é atribuído ao grupo cuja média está mais próxima. Essa abordagem permite uma segmentação clara dos dados, mas exige que o número K de grupos seja definido previamente, o que pode representar um desafio dependendo da natureza do conjunto de dados.

Para determinar o número ideal de *clusters*, foi aplicado o método do cotovelo (*elbow method*). Essa técnica avalia a soma das distâncias quadradas dentro dos *clusters* para diferentes valores de K e busca identificar o ponto em que o ganho marginal na redução da variabilidade interna diminui significativamente. Esse ponto, que forma uma curva com um ”cotovelo”, representa o número ideal de agrupamentos para os dados analisados. A Figura 18 ilustra essa técnica aplicada aos dados, evidenciando o ponto ideal de K para a análise em questão.

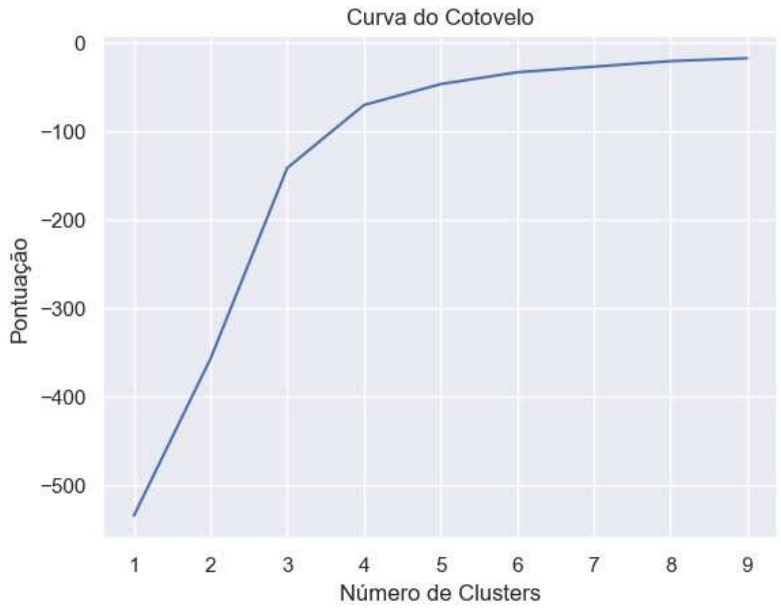


Figura 18: Curva do cotovelo para as coordenadas dos usuários

Com base nos resultados do método do cotovelo, identificou-se que o conjunto de dados comporta quatro *clusters* principais. Em seguida, aplicou-se o algoritmo *K-means* para realizar a classificação e atribuição de cada ponto a um dos *clusters* definidos.

Com todos os nós devidamente espacializados, foi possível realizar uma análise comparativa entre os *clusters* formados com base nos dados geográficos e as comunidades

identificadas a partir das conexões estabelecidas na rede. As Figuras 19 e 20 ilustram as duas abordagens para aglutinar os dados.

Na Figura 19, é exibida a segmentação em múltiplas comunidades geradas pelo método de detecção de comunidades de Louvain. Para evitar que o mapa gerado ficasse sobrecarregado e dificultasse a visualização, optou-se por não exibir os nós isolados, resultando em uma representação mais limpa e focada nos nós conectados.

Já na Figura 20, observa-se uma divisão clara em quatro conjuntos distintos, conforme calculado utilizando o método do cotovelo. Nesse caso, todos os nós estão visíveis, já que não há análise de interações que dependam de relações de rede, mas apenas de proximidade geográfica. Como o agrupamento foi feito exclusivamente com base na localização espacial, nenhum nó foi isolado, permitindo uma visualização completa e representativa de todos os elementos do conjunto de dados.

Essas visualizações destacam as diferenças metodológicas e os resultados obtidos em cada abordagem, evidenciando como a rede e a espacialização influenciam na organização e interpretação dos dados.

Os resultados obtidos indicam que os dados analisados não atendem às expectativas iniciais de que os agrupamentos geográficos teriam alta correlação com as comunidades sociais. Pelo contrário, foi identificada uma significativa disparidade entre esses dois tipos de agrupamento, evidenciando que as conexões sociais não seguem necessariamente padrões espaciais diretos.

Essa discrepância torna evidente a necessidade de um método capaz de gerar comunidades que considerem simultaneamente as dimensões geográficas e sociais de maneira significativa. Essa abordagem integrada é essencial para capturar a complexidade subjacente das redes que combinam diferentes naturezas de interação.

Nos capítulos seguintes, será apresentado um *framework* desenvolvido especificamente para a análise de redes multicamadas, que integra dados geográficos e não geográficos. Esse *framework* possibilita o controle da relevância atribuída a cada aspecto no contexto do estudo, permitindo maior flexibilidade na modelagem e análise. Além disso, serão explorados os resultados gerados a partir de estatísticas detalhadas das comunidades de cada rede, oferecendo *insights* mais profundos sobre a interação entre os diferentes fatores estruturais.

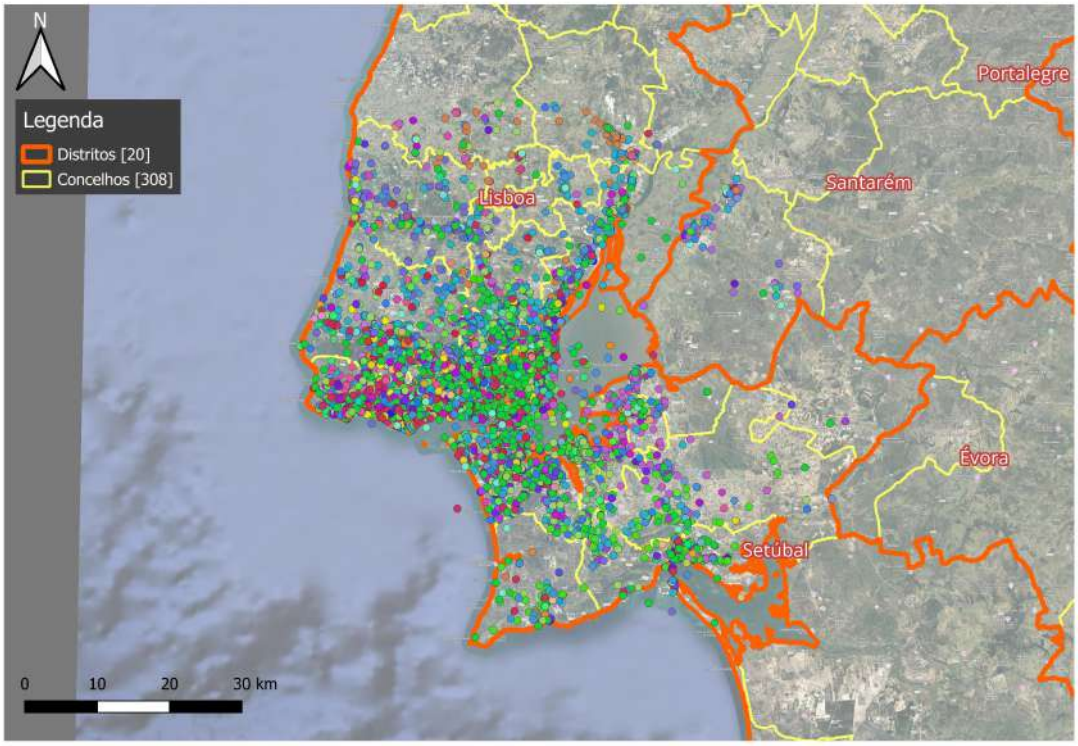


Figura 19: Distribuição dos nós em comunidades de Louvain

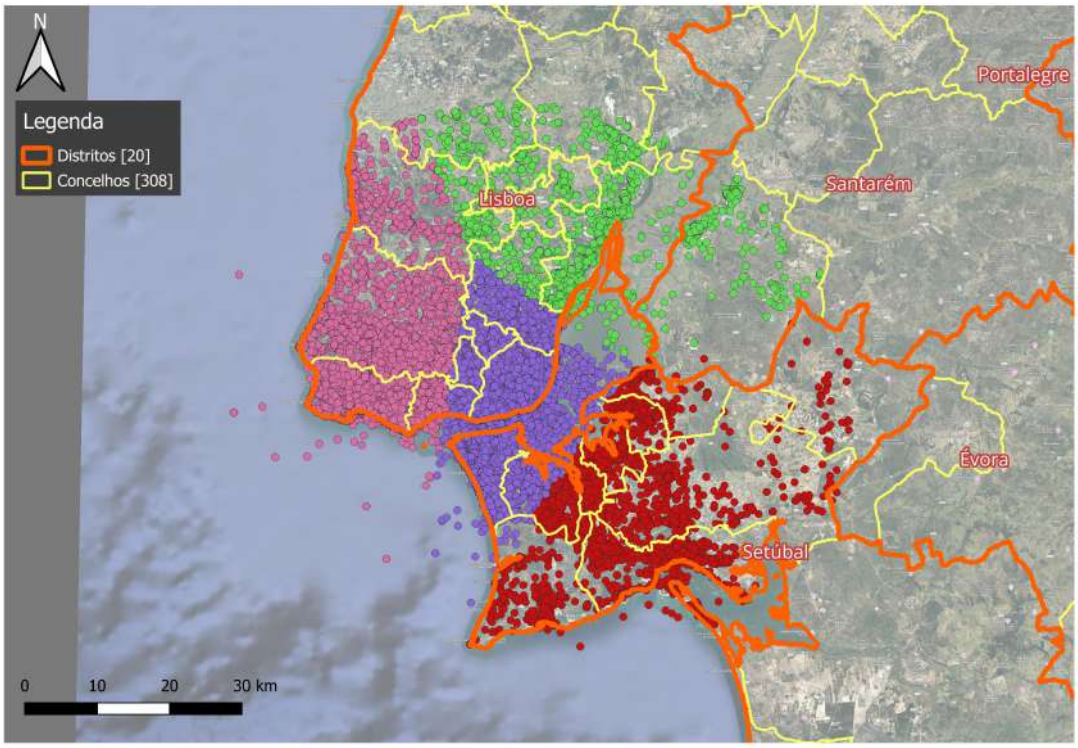


Figura 20: Distribuição dos nós em conjuntos utilizando *K-means*

4. Definição de Modelos para Redes Geolocalizadas

O estudo de redes complexas frequentemente lida com uma ampla variedade de atributos associados aos seus nós. Por exemplo, ao representar uma malha ferroviária como uma rede, os nós podem corresponder às estações, e cada estação pode ter atributos associados, como a quantidade média de passageiros que circulam por ela, as linhas atendidas, entre outros.

Uma característica particularmente relevante desses atributos é a presença de informações geográficas, que atribuem uma localização espacial aos nós. Redes que incorporam esse tipo de dado podem ser definidas como Redes Geolocalizadas. Essas redes possuem uma estrutura em duas camadas: uma camada que contém informações geográficas, representando a localização espacial dos nós, e outra camada funcional, que pode refletir interações de natureza diversa, como conexões sociais, fluxos ferroviários, trocas de informação, entre outras.

O grande interesse em redes geolocalizadas está em sua capacidade de integrar a análise de características espaciais e funcionais em um único modelo, eliminando a necessidade de empregar métodos distintos para identificar agrupamentos em cada dimensão. Essa integração torna a representação dos dados mais flexível e eficaz, sendo particularmente relevante para aplicações em sistemas como redes de computadores sem fio, onde a localização física e as conexões funcionais desempenham papéis complementares. Como resultado, redes geolocalizadas oferecem uma visão mais rica e detalhada tanto da estrutura quanto das dinâmicas subjacentes aos sistemas que representam, permitindo uma análise mais abrangente e precisa.

4.1 Estruturação do modelo

O modelo proposto por este trabalho tem o objetivo de modificar gradualmente a estrutura de uma rede geográfica de modo que ela se assemelhe progressivamente com uma rede de uma outra natureza não geográfica. Este modelo é projetado a partir da ideia de que ele seja utilizado como um metamodelo, em que a ideia dele esteja bem estruturada e a construção das redes seja baseada em algum outro modelo.

Com isso é importante salientar que não é um modelo idealizado para representar uma rede real, mas sim como uma forma de estudar as características de comunidades que podem ser observadas em redes geolocalizadas. Sendo assim ele é proposto para ser um *benchmark* na análise das comunidades desse grupo de redes. A proposta, a partir disso, está dividida em dois principais elementos de análise.

O primeiro elemento diz respeito à representação de redes geolocalizadas por meio de uma estrutura de duas camadas. Uma das camadas é baseada na localização espacial dos nós, enquanto a outra reflete uma natureza funcional, como interações sociais ou fluxos de transporte. O segundo envolve o uso de modelos estocásticos para gerar e analisar essas redes, com controle explícito da similaridade entre as camadas. Para isso, foram utilizados os modelos de Watts-Strogatz e Waxman, que fornecem as bases teóricas para os modelos desenvolvidos neste estudo. (WATTS; STROGATZ, 1998; WAXMAN, B., 1988)

O controle da similaridade entre as camadas é realizado por meio de um parâmetro p, que representa a proporção de arestas que serão reconectadas em relação à configuração original. Quando $p = 0$, as duas camadas são idênticas, refletindo uma estrutura completamente geográfica. À medida que esse valor aumenta, a segunda camada se torna progressivamente mais diferente, refletindo conexões de natureza não geográfica. Esses modelos funcionam como um ponto de referência, permitindo avaliar como as diferenças entre as camadas impactam tanto a estrutura global da rede quanto as propriedades específicas de cada camada.

O modelo proposto segue uma abordagem estruturada composta por três passos:

1. **Posicionamento dos Nós**: Todos os nós são dispostos em um círculo unitário, de forma que a distância entre pares de nós seja uniforme. Essa configuração forma um círculo regular, garantindo simetria na distribuição espacial inicial.

2. **Conexões na Primeira Camada**: Na camada inicial, os nós são conectados com base em um parâmetro d, que define a distância máxima para estabelecer uma

conexão. Dessa forma, apenas os nós cuja separação seja menor ou igual a d estarão conectados, criando uma rede essencialmente geográfica.

3. **Construção da Segunda Camada**: A partir da rede estabelecida na primeira camada, uma segunda camada é gerada realocando as arestas existentes com uma probabilidade p. Essa etapa introduz aleatoriedade na rede, permitindo que conexões não sejam mais restritas à proximidade espacial.

4.2 Modelo Watts-Strogatz de duas camadas

Após a definição geral do modelo, implementa-se uma versão específica baseada no modelo de Watts-Strogatz, em que cada camada da rede é configurada como $WS(n, k, p)$. Nesse modelo, os parâmetros representam, respectivamente, o número de nós (n), a quantidade de vizinhos que cada nó deve conectar (k) e a probabilidade de reconexão das arestas (p). Essa abordagem permite estudar a transição entre uma rede puramente geográfica e uma rede com características de ”pequeno mundo”ou aleatórias, ajustando adequadamente os parâmetros para cada camada.

Uma característica essencial do modelo WS é sua estrutura regular inicial, definida pelo parâmetro k, que determina quais nós devem ser conectados em função de sua proximidade. Essa característica cria redes simétricas e organizadas, com conexões bem definidas.

No contexto deste trabalho, essa ideia é adaptada para integrar características geográficas. Primeiramente, os n nós são dispostos em um círculo unitário regular, e cada nó é conectado a k vizinhos localizados a uma distância d. O cálculo de k é feito utilizando a Lei dos Cossenos, conforme a Equação 4.1, desconsiderando o valor do raio, uma vez que o círculo é unitário:

$$k = \sqrt{a^2 + b^2 - 2ab\cos\theta} \tag{4.1}$$

onde a e b são iguais a 1 e $\theta = \frac{\pi k}{n}$. Essa abordagem assegura que a rede inicial gerada seja equivalente ao modelo WS com probabilidade $p = 0$, ou seja, uma rede puramente geográfica. Essa equivalência permite comparações diretas entre configurações espaciais e redes com níveis crescentes de aleatoriedade.

Com a primeira camada estabelecida, é construída a segunda camada para introduzir desvios da configuração geográfica original. Esse processo é realizado aplicando o

parâmetro p, que define a proporção das conexões que devem ser realocadas em relação ao número total de arestas da rede. A introdução dessa aleatoriedade permite simular o afastamento entre uma rede geográfica e uma rede com características mais abstratas ou não-geográficas.

Após a construção das duas camadas e considerando as divergências observadas de uma camada para a outra, é possível avaliar o impacto das alterações estruturais nas comunidades de cada camada. Isso possibilita uma compreensão mais detalhada de como a transição entre redes geográficas e aleatórias afeta a organização interna da rede e suas dinâmicas subjacentes.

4.3 Modelo Waxman de duas camadas

De forma análoga foi feita uma versão do modelo utilizando o descrito no modelo de Waxman configurando a camada inicial da rede com os parâmetros N, α, β e L. Esses parâmetros determinam, respectivamente, o número de nós, a intensidade da conectividade em função da distância, a probabilidade máxima de conexão e o diâmetro máximo da área em que os nós estão distribuídos.

Esse modelo oferece uma abordagem flexível para analisar a transição de uma rede puramente geográfica para uma rede mais heterogênea. Diferentemente do modelo de Watts-Strogatz, o modelo de Waxman introduz uma variabilidade no número de conexões entre os nós, uma vez que a probabilidade de conexão depende diretamente da distância entre os pares de nós. Essa característica gera uma estrutura levemente assimétrica, refletindo maior realismo em redes geográficas.

A construção das redes baseadas no modelo de Waxman, configuradas em duas camadas, começa com a disposição regular de todos os N nós ao longo de um círculo unitário. Em seguida, as arestas entre os pares de nós (ij) são alocadas com base na probabilidade $p_{i,j}$, descrita pela equação:

$$p_{i,j} = \beta e^{-\frac{d_{i,j}}{\alpha L}} \tag{4.2}$$

Nesta primeira camada, o valor de L é fixado em 2, correspondente ao diâmetro do círculo unitário. O parâmetro β é definido como 1, garantindo que nós sobrepostos ($d_{i,j} = 0$) estejam sempre conectados. O valor de α é ajustado para que o grau médio da rede seja equivalente ao valor de k utilizado no modelo WS. Esse ajuste assegura que as camadas

geográficas iniciais de ambos os modelos sejam comparáveis e consistentes para fins de análise.

A segunda camada é gerada aplicando a proporção p de arestas a serem aleatorizadas. Isso permite que a rede transite gradualmente para uma estrutura menos dependente da geografia, aproximando-se de uma configuração mais abstrata ou aleatória.

Com ambas as camadas finalizadas, é possível analisar as comunidades identificadas em cada uma delas. Essa análise fornece *insights* valiosos sobre as diferenças e semelhanças entre redes geográficas e redes cujas estruturas são influenciadas por outros fatores além da localização, destacando como a aleatoriedade e a geografia impactam a organização das comunidades.

5. Avaliação dos Modelos

O *framework* proposto oferece uma abordagem flexível e robusta para a geração e análise de redes, permitindo ajustes nos parâmetros que influenciam diretamente suas características estruturais. Essa flexibilidade possibilita a criação de redes com configurações que vão desde estruturas baseadas exclusivamente na proximidade geográfica até aquelas onde as conexões refletem fatores mais complexos, como relações sociais, físicas ou de outra natureza.

O objetivo das análises performadas neste capítulo é estudar os dois modelos (Watts-Strogatz e Waxman) em função dos parâmetros de grau médio e proporção de arestas aleatorizadas. Além disso também será avaliada a similaridades entre as comunidades que serão geradas em cada camada da rede geolocalizadas.

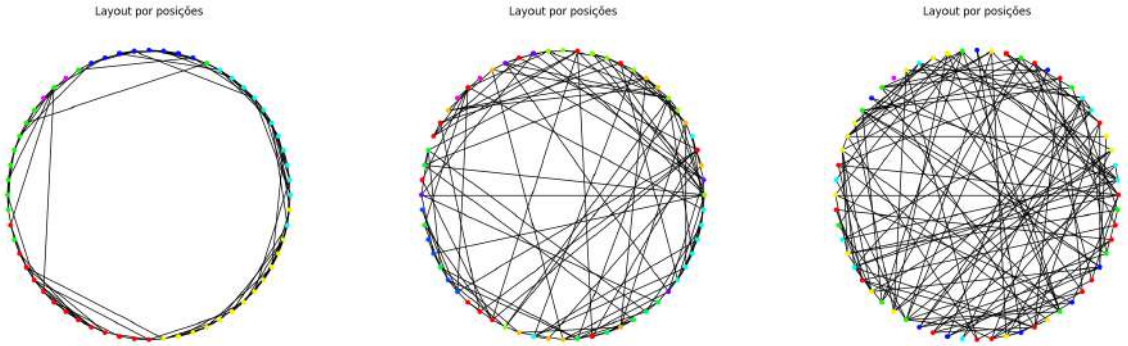
Todos os testes feitos neste trabalho foram realizados utilizando um desktop com 32GB de memória RAM, um HD de 1TB e dois SSD que totalizam juntos 384GB, um processador AMD Ryzen 7 3700X 8-Core Processor de 3.60 GHz com 8 núcleos e o sistema operacional foi o Windows 11 versão 23H2 de 64bits. Além disso, devido a dificuldades de processamento, o teste de cada um dos dois modelos só pôde ser realizado uma única vez, o que impediu a geração de médias e desvio-padrão dos resultados.

Para ilustrar o funcionamento do modelo baseado em Waxman, a Figura 21 apresenta uma rede de 60 nós com grau médio $\langle k \rangle = 4$, gerada em diferentes estágios. A rede é representada por sua localização espacial e, ao lado, por um método de posicionamento do `NetworkX`. Na Figura 21a, com $p = 0$, observa-se que as arestas estão bem próximas umas das outras, já que são estabelecidas com base na proximidade geográfica.

Na Figura 21b, metade das arestas foi aleatorizada ($p = 0.5$), o que resulta em arestas mais dispersas, afastando-se do parâmetro geográfico inicial. Finalmente, quando $p = 1$, ou seja, quando todas as arestas foram completamente aleatorizadas, a rede apresenta uma configuração desorganizada, com muitas arestas sobrepostas, conforme mostrado na

Figura 21c.

Além disso, a modularidade das camadas diminui conforme aumenta o nível de aleatorização, indicando que redes mais aleatórias apresentam divisões de partições menos claras. Os valores de modularidade observados foram 0.59826 no primeiro caso, 0.45033 no segundo e 0.41153 no último. Os nós também são coloridos de acordo com a comunidade que pertencem, o que mostra que na primeira camada as cores estão bem agrupadas, enquanto na última existe uma mistura acentuada das comunidades no espaço.



(a) Rede com $p = 0$ de aleatorização

(b) Rede com $p = 0.5$ de aleatorização

(c) Rede com $p = 1$ de aleatorização

Figura 21: Funcionamento do modelo de Waxman em duas camadas. A rede 21a representa a camada geográfica inicial e as outras duas as camadas derivadas dela por aleatorização

O código dos dois modelos utilizados juntamente com os testes realizados nesse trabalho e os métodos acessórios para visualização e cálculo das métricas estão disponibilizados em um repositório aberto no *GitHub*[1].

5.1 Métricas de avaliação de similaridade

Para quantificar o impacto da randomização sobre a estrutura da rede e avaliar as mudanças na organização das comunidades, foram aplicadas duas métricas o Índice de Jaccard e o *Adjusted Mutual Information* (AMI) que se mostraram fundamentais para analisar a evolução da estrutura comunitária, capturando como a randomização das arestas afetou a organização das comunidades, desde uma configuração fortemente influenciada pela proximidade espacial até uma rede com conexões mais abstratas e menos dependentes da localização.

[1] https://github.com/lealeric/geolocated-networks-metamodel

5.1.1 Avaliação do Índice de Jaccard

O índice de similaridade, proposto em 1901, foi originalmente desenvolvido para quantificar a colocalização de flora alpina, com particular interesse no estudo da diversidade de espécies. Esse índice, que agora leva o nome de seu criador, pode ser representado em notação de teoria dos conjuntos como:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5.1}$$

onde A e B são quaisquer dois conjuntos, e $|A|$ e $|B|$ representam, respectivamente, o número de elementos de cada conjunto. (F. COSTA, 2021)

Essa métrica é amplamente utilizada para avaliar a similaridade entre conjuntos, onde valores próximos a 1 indicam uma alta sobreposição, representando grande interseção entre os elementos, enquanto valores próximos a 0 sugerem uma sobreposição mínima ou inexistente.

5.1.2 Avaliação do AMI

O AMI é uma métrica baseada na Teoria da Informação, mais especificamente na Informação Mútua (MI) que se trata de uma medida que retorna quanta informação uma variável aleatória contém sobre outra variável aleatória. Essa medida é definida utilizando duas variáveis aleatórias X e Y que possuem uma função de massa de probabilidade conjunta $p(x, y)$ e funções de massa de probabilidade marginal $p(x)$ e $p(y)$. (COVER, 2006)

O AMI avalia a concordância entre as estruturas das comunidades identificadas, levando em consideração o acaso. Diferentemente do Índice de Jaccard, a AMI ajusta os resultados para evitar que estruturas aleatórias gerem pontuações infladas, proporcionando uma medida mais robusta de similaridade.

Esta métrica foi definida por Vinh, Epps e Bailey (2009) na Equação 5.2, que utiliza um sentido estocástico para ajustar as medidas do MI. Especificamente quando o valor do AMI é igual a 1 observa-se que os dois conjuntos são idênticos e quanto mais próximo de 0 esse valor, mais próximo do valor esperado de MI ele chega.

$$AMI_{max}(U, V) = \frac{I(U, V) - E\{I(U, V)\}}{\max\{H(U), H(V)\} - E\{I(U, V)\}}, \tag{5.2}$$

onde:

I(U,V) é a Informação Mútua entre U e V;

E{I(U,V)} é o valor esperado do MI;

H(U) e H(V) representam a incerteza associada a cada variável.

Com base nessas premissas e na natureza das informações analisadas, espera-se observar uma tendência de queda nos valores do AMI à medida que aumenta a quantidade de arestas aleatorizadas. Esse comportamento é esperado, pois as redes geográficas tendem a preservar comunidades distintas das formadas em redes de outras naturezas.

5.2 Avaliação do modelo de Watts-Strogatz de duas camadas

No primeiro cenário, utilizou-se o modelo Watts-Strogatz, reconhecido por sua capacidade de simular redes do tipo pequeno mundo e que, nesse caso, oferece a flexibilidade necessária para transitar gradualmente entre conexões estritamente geográficas e aleatórias. O número total de nós foi fixado em $n = 50,000$, garantindo uma análise consistente e escalável. Além disso, foram testados cinco valores distintos para o parâmetro k, para avaliar diferentes configurações estruturais: 10, 20, 40, 80 e 160, resultando em cinco cenários de rede independentes.

Inicialmente, as redes foram configuradas com probabilidade de reconfiguração das arestas ($p = 0$), assegurando que todas as conexões fossem puramente geográficas. Nesse cenário, os nós foram conectados com base em sua proximidade espacial, criando uma rede altamente regular e estruturada. A densidade da rede foi calculada como 2×10^{-4}, e para $k = 10$, o grafo continha 250.000 arestas, resultando, como esperado, em um grau médio igual a 10. O mesmo procedimento foi repetido para os demais valores de k, com as principais métricas estruturais das redes resumidas na Tabela 4.

Tabela 4: Métricas básicas das redes geográficas baseadas no algoritmo de Watts-Strogatz

k	N° de Arestas	Densidade	CC Global
10	250,000	0.0002	0.6667
20	500,000	0.0004	0.7105
40	1,000,000	0.0008	0.7308
80	2,000,000	0.0016	0.7405
160	4,000,000	0.0032	0.7453

Para explorar a transição de redes geográficas para redes menos dependentes da

proximidade espacial, as arestas foram progressivamente reconfiguradas com base em diferentes valores da probabilidade p. Foram testados os seguintes níveis de randomização: 0.01%, 0.02%, 0.04%, 0.08%, 0.1%, 0.2%, 0.4%, 0.8%, 1%, 2%, 4%, 8% e 16%. Essas porcentagens representam a fração de arestas que foram aleatoriamente reassociadas, permitindo que a rede evoluísse de uma configuração puramente geográfica para uma estrutura progressivamente mais aleatória.

Por exemplo, em uma rede com $k = 10$ e 250.000 arestas, reconfigurar 0.01% das conexões equivale a modificar 25 arestas. Esse processo foi replicado para todas as configurações de k testadas, com cada rede resultante comparada à rede geográfica original, que foi mantida como referência para a análise. Além disso a identificação de comunidades para cada uma das redes foi feita utilizando o algoritmo de maximização da modularidade.

Considerando o tamanho da rede, este algoritmo se mostrou muito custoso, tendo um tempo maior de identificação de comunidades a cada aumento da aleatorização. Para serem geradas todas as comunidades necessárias para a análise, o processamento levou em torno de 36 horas.

Com base nos dados obtidos durante a geração das redes geográficas e suas versões derivadas pela aleatorização das arestas, espera-se que a matriz de índices, que compara as comunidades originais com as randomizadas, apresente diferenças significativas em relação a uma matriz diagonal. Isso ocorre porque as comunidades de natureza não espacial tendem a divergir substancialmente das comunidades formadas na rede geolocalizada.

Na Figura 22, são apresentados três gráficos que representam mapas de calor do Índice de Jaccard, calculado entre as comunidades identificadas ao considerar $k = 10$ e aplicar randomizações de 0.1%, 0.8% e 8% das arestas, respectivamente.

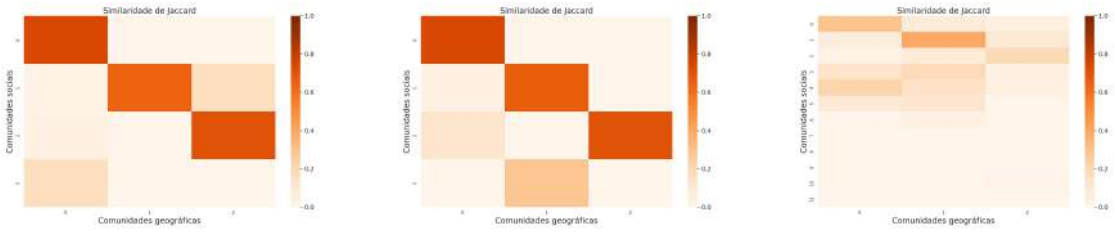
Em cada gráfico, as comunidades obtidas após a randomização das arestas estão representadas no eixo y, enquanto as comunidades geográficas da rede original estão no eixo x. Os gráficos à esquerda apresentam a saída ordenada padrão, enquanto os gráficos à direita foram permutados para posicionar os maiores valores do índice na diagonal principal da matriz, facilitando a interpretação visual.

Esses mapas de calor revelam como a reorganização das arestas impacta diretamente a formação das comunidades. Na primeira camada, quando a rede é baseada na localização geográfica dos nós e as comunidades geradas mostram o agrupamento geográfico dos nós, mas, à medida que as arestas são randomizadas e a segunda camada é gerada, as

comunidades geradas pela rede divergem das comunidades da primeira camada.

Na primeira randomização, com 0.1% das arestas alteradas (Figura 22a), os valores do Índice de Jaccard variam entre 0 e 0.6. Quando a proporção das arestas randomizadas chega a 0.8% (Figura 22b), o índice atinge até 0.8 o que vai contra a ideia de que camadas com uma aleatorização maior das arestas influencia em comunidades diferentes, isso ocorre porque o experimento só foi realizado uma única vez e não tem um intervalo de confiança para analisar, além de que o modelo WS é muito regular e produz comunidades muito aleatórias. No entanto, com uma maior randomização, como na alteração de 16% das arestas (Figura 22c), observa-se uma diminuição acentuada nos valores do índice, que não ultrapassam 0.5.

Também é possível notar que todas as células apresentam um índice bem baixo ainda mais quando comparadas às outras porcentagens que apresentam alguns pontos de alta correlação. Isso demonstra que, conforme a rede se distancia da estrutura original, as comunidades se tornam menos correlacionadas com a organização inicial baseada na localização geográfica.



(a) $p = 0.1\%$ **(b)** $p = 0.8\%$ **(c)** $p = 16\%$

Figura 22: Matriz de índices de Jaccard na forma de *heatmaps* para a rede Watts-Strogatz de duas camadas com $k = 10$

Na Figura 23, são apresentados gráficos semelhantes aos da figura anterior, mas agora considerando a rede com $k = 160$ e os mesmos níveis de randomização de arestas: 0.1%, 0.8% e 16%.

O comportamento observado para $k = 160$ é semelhante ao analisado anteriormente, mas com algumas diferenças sutis. Nos gráficos, os índices máximos de Jaccard atingem valores de aproximadamente 0.5, 0.8 e 0.4 para as proporções de randomização de 0.1% (Figura 23a), 0.8% (Figura 23b)e 16% (Figura 23c), respectivamente. Esses valores refletem que, para $k = 160$, as comunidades se tornam menos correlacionadas às comunidades geográficas originais à medida que aumenta o nível de randomização das arestas.

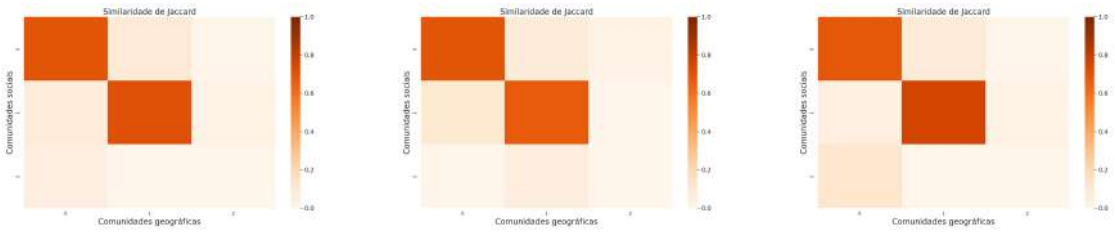
Nesses experimentos os gráficos mostram uma matriz identidade mais clara, com 2

comunidades apresentando valores do índice de Jaccard comparando com as comunidades da primeira camada superiores a 0.4. Essa característica é observada devido ao modelo que gera as comunidades de forma aleatória devido à simetria da sua distribuição de arestas.

Com isso, no caso de 0.1% de randomização, a maior parte das comunidades ainda guarda uma relação significativa com a estrutura geográfica inicial. Quando a proporção de arestas randomizadas aumenta para 0.8%, observam-se picos de similaridade próximos de 0.8, indicando que algumas comunidades ainda são bastante semelhantes, enquanto outras começam a divergir devido à maior influência das conexões aleatórias. Por fim, com 16% de randomização, os índices de similaridade não ultrapassam valores de 0.4, o que evidencia que a estrutura da rede já está se distanciando da organização geográfica inicial.

Essa análise reforça a tendência observada anteriormente: quanto maior o valor de k e a proporção de randomização, mais a influência da localização geográfica dos nós é enfraquecida, e a rede adquire características mais próximas de uma estrutura social ou aleatória, em que as conexões não são mais determinadas majoritariamente pela proximidade física. Esses resultados destacam a sensibilidade da rede às mudanças estruturais e a complexidade em medir a persistência das comunidades em diferentes níveis de randomização.

Além disso, com valores de k maiores, é possível observar uma redução na quantidade de comunidades detectadas, como era esperado. Esse comportamento se deve ao aumento expressivo no número de conexões entre os nós, o que fortalece a coesão da rede e dificulta a segmentação em grupos menores e distintos. Nessas configurações, seria necessário um número significativamente maior de aleatorizações para impactar substancialmente a estrutura da rede e, assim, promover um aumento considerável no número de comunidades formadas. Essa característica reflete o papel do grau médio elevado na manutenção de redes mais densas e menos fragmentadas.



(a) $p = 0.1\%$ **(b)** $p = 0.8\%$ **(c)** $p = 16\%$

Figura 23: Matriz de índices de Jaccard na forma de *heatmaps* para a rede Watts-Strogatz de duas camadas com $k = 160$

A partir da análise apresentada na Figura 24, observa-se que os valores de AMI obtidos

para diferentes níveis de randomização das arestas mostram uma considerável flutuação conforme a rede se distancia da configuração original. No entanto, para todos os valores de k analisados, o AMI converge para valores próximos de 0 à medida que a aleatorização se intensifica.

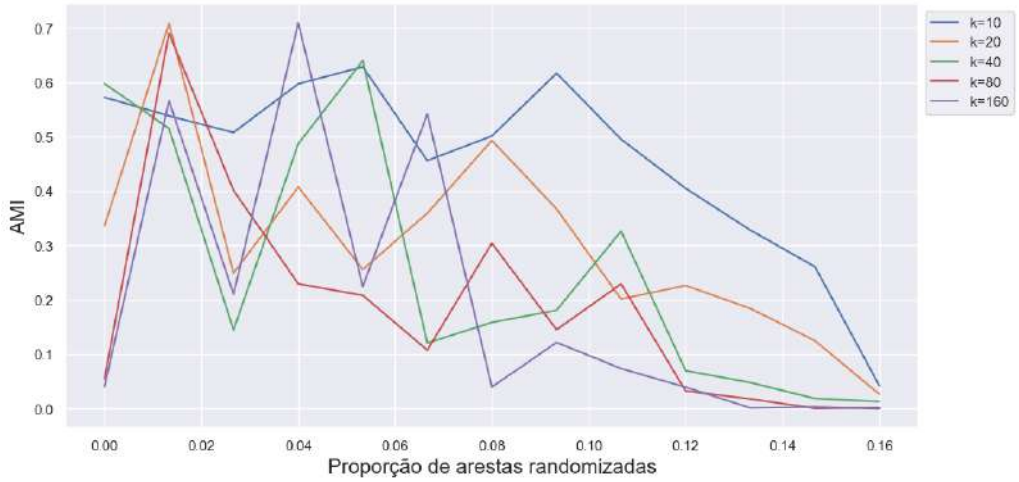


Figura 24: Valores de AMI para o modelo de Watts-Strogatz

Esse comportamento indica que, embora as similaridades entre as comunidades possam variar durante as etapas intermediárias de randomização, quando a segunda camada atinge um grau de aleatoriedade elevado ($p = 16\%$), as comunidades formadas nas duas camadas tornam-se substancialmente diferentes. Esse resultado evidencia que, com o aumento da aleatorização, as características originais da camada geográfica perdem influência sobre a estrutura da rede.

5.3 Avaliação do modelo de Waxman de duas camadas

Para o segundo cenário, foi utilizada uma abordagem baseada no modelo de Waxman, que emprega uma probabilidade dependente da distância entre nós para determinar as conexões. Como descrito na Seção 4.3, os nós neste modelo foram posicionados de forma regular, que é essencial para criar redes que mantenham características geográficas, enquanto o modelo ajusta as conexões com base nos parâmetros α e β.

No experimento, o número de nós foi fixado também em $n = 50,000$ a fim de manter uma forma de comparação direta com o modelo anterior. O parâmetro β foi definido como 1, enquanto os valores de α foram otimizados usando o método de Newton como descrito na Equação 5.3, para que o grau médio das redes geradas fosse aproximadamente o mesmo das redes baseadas no modelo Watts-Strogatz. Os valores calculados para α são compostos por duas raízes e o *framework* utiliza a média entre ele, todos esses valores

estão resumidos na Tabela 5.

$$E[d_v] = 2 \cdot \sum_{i=1}^{N/2} (e^{\frac{-d_{(i)}}{2\alpha}}) - e^{\frac{-d_{(\frac{N}{2})}}{2\alpha}}, \tag{5.3}$$

Tabela 5: Valores de α para a geração da rede geográfica baseada no modelo de Waxman

Grau médio esperado	α min.	α max.	α
10	3.446205×10^{-4}	3.446213×10^{-4}	3.446209×10^{-4}
20	6.592346×10^{-4}	6.592354×10^{-4}	6.592350×10^{-4}
40	1.287795×10^{-3}	1.287796×10^{-3}	1.287796×10^{-3}
80	2.544544×10^{-3}	2.544545×10^{-3}	2.544545×10^{-3}
160	5.057769×10^{-3}	5.057770×10^{-3}	5.057769×10^{-3}

Com base nessas configurações, foram geradas cinco redes distintas para cada valor de α. Essas redes foram analisadas quanto às suas métricas básicas, apresentadas na Tabela 6. Esses resultados servem como base para comparar o comportamento das redes geográficas geradas pelos dois modelos, permitindo avaliar como diferentes abordagens afetam a estrutura e a formação de comunidades nas redes.

Tabela 6: Métricas básicas das redes geográficas baseadas no algoritmo de Waxman

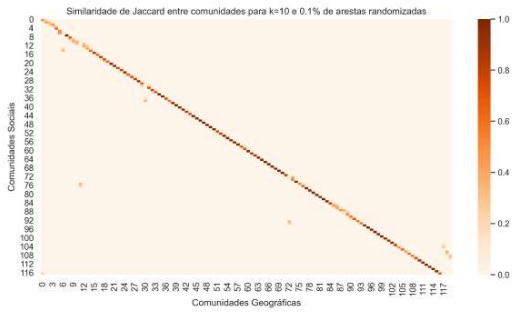
Grau médio esperado	N° de Arestas	Grau médio	Densidade	CC Global	CC Médio
10	249,622	9.9849	0.0002	0.3252	0.3358
20	500,155	20.0062	0.0004	0.3481	0.3543
40	1,000,183	40.0073	0.0008	0.3614	0.3647
80	2,000,180	80.0072	0.0016	0.3677	0.3694
160	4,001,606	160.0642	0.0032	0.3714	0.3723

Para todas as redes geradas no modelo de Waxman, foi realizada a aleatorização das arestas considerando uma proporção da quantidade total de arestas, utilizando as mesmas porcentagens descritas na Seção 5.2. Devido às limitações de processamento, o algoritmo de Louvain foi adotado para a identificação das comunidades nas redes deste modelo, visto que enquanto o outro algoritmo levava horas para executar uma camada da rede, o Louvain conseguia obter os resultados levando em torno de 1 a 2 minutos por camada. Essa escolha impactou diretamente a quantidade de comunidades detectadas, apresentando diferenças significativas em relação ao modelo de Watts-Strogatz.

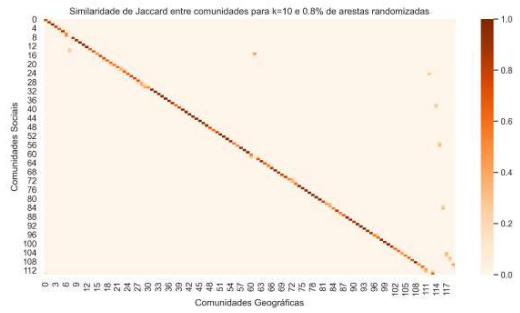
A análise das matrizes de Jaccard teve início com a avaliação do comportamento das comunidades geradas. A Figura 25 revela como os índices de Jaccard estão dispostos para redes com grau médio esperado $k = 10$ e probabilidades de aleatorização $p = 0.1$, $p = 0.8$ e $p = 16$.

De imediato, nota-se que os gráficos obtidos diferem consideravelmente do modelo de Watts-Strogatz. A quantidade de comunidades é significativamente maior, tanto na primeira camada quanto na segunda. Além disso, a matriz apresenta uma configuração mais próxima de uma Matriz Identidade, com uma diagonal principal bem definida. Isso indica que, mesmo após a aleatorização das arestas, as comunidades geradas em cada camada permanecem compostas pelos mesmos nós, sugerindo uma forte correlação entre a localização geográfica dos nós e sua variável funcional.

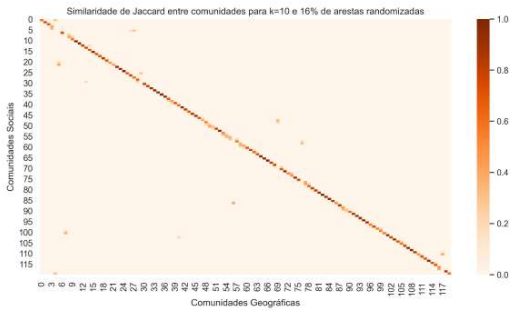
Ainda assim, observam-se nuances interessantes. Nas Figuras 25a e 25b, muitas comunidades apresentam valores de Jaccard iguais a 1, o que implica que as partições são idênticas em ambas as camadas. Entretanto, na Figura 25c, essa correspondência diminui, evidenciando que a aleatorização começa a alterar a estrutura das comunidades entre as camadas. Contudo, essa mudança ocorre de maneira mais sutil em comparação ao comportamento observado no modelo de Watts-Strogatz, possivelmente devido às diferenças estruturais e ao método de identificação de comunidades adotado.



(a) $p = 0.1\%$

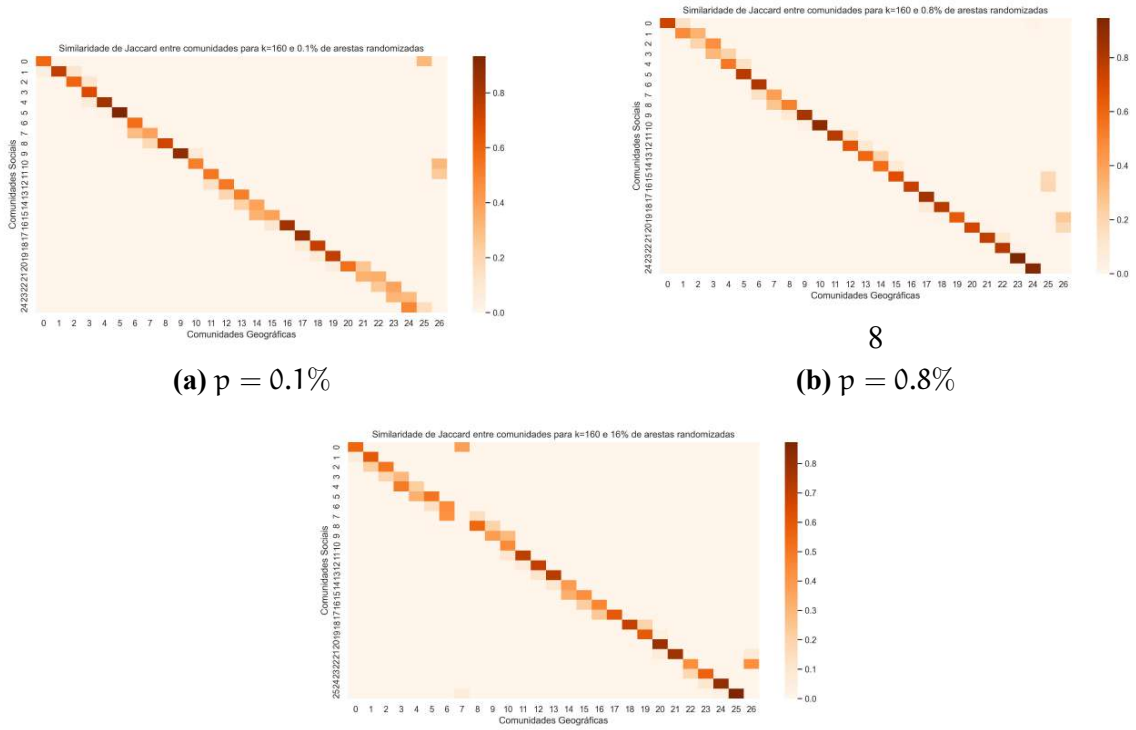


(b) $p = 0.8\%$



(c) $p = 16\%$

Figura 25: Matriz de índices de Jaccard na forma de *heatmaps* para a rede Waxman de duas camadas com $k = 10$



8

(a) p = 0.1%

(b) p = 0.8%

(c) p = 16%

Figura 26: Matriz de índices de Jaccard na forma de *heatmaps* para a rede Waxman de duas camadas com k = 160

Ao analisar a rede com grau médio k = 160, observa-se um comportamento semelhante ao observado para graus menores: os valores de Jaccard permanecem muito altos quando as aleatorizações são pequenas, indicando uma forte correspondência entre as comunidades das duas camadas. À medida que a aleatorização aumenta, os valores de Jaccard diminuem ligeiramente, mas ainda se mantêm bem concentrados em torno da diagonal principal, evidenciando que a estrutura das comunidades permanece altamente correlacionada, mesmo com o aumento da aleatoriedade.

Além disso, percebe-se uma redução significativa no número de comunidades identificadas. Essa diminuição é explicada pela maior densidade da rede, uma vez que, em redes mais densas, os nós tendem a formar grupos maiores e mais interconectados, resultando em menos comunidades distintas. Esse comportamento é esperado, já que redes com um grau médio mais alto possuem uma maior sobreposição entre as conexões, o que naturalmente reduz a fragmentação da rede em subgrupos.

A avaliação do AMI torna ainda mais evidente a diferença entre os dois algoritmos de identificação de comunidades. A Figura 27 ilustra a variação dessa métrica para todas as probabilidades de aleatorização consideradas em todos os graus médios analisados. Os resultados mostram que os valores de AMI variam dentro de uma faixa bastante restrita,

entre 0.8 e 0.96, indicando que as comunidades geradas para as diferentes camadas são quase idênticas, mesmo com a introdução de aleatoriedade nas arestas.

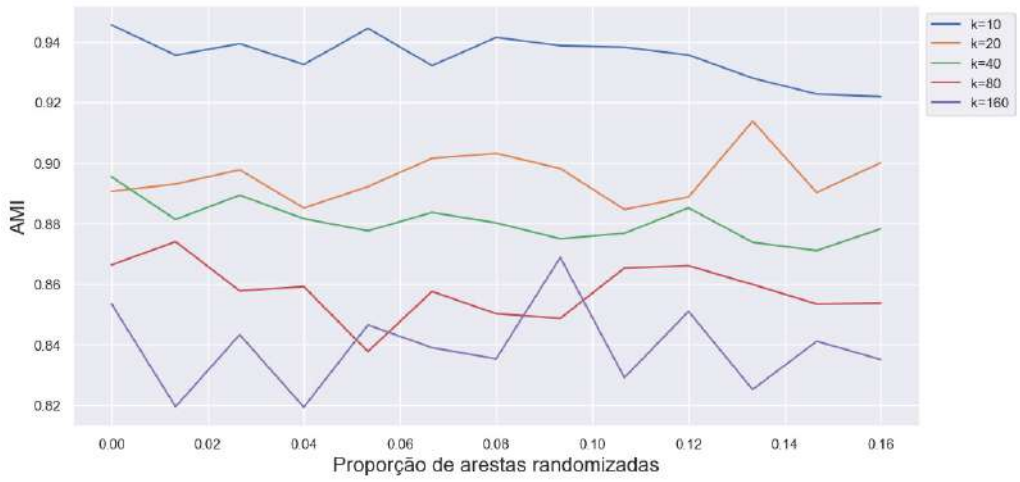


Figura 27: Valores de AMI para o modelo de Waxman

Apesar disso, o gráfico mostra que a faixa de valores calculados diminui e a variabilidade aumenta, conforme o grau médio aumenta. Isso indica que redes mais densas podem ser mais influenciadas pela aleatorização das arestas gerando comunidades mais divergentes entre as camadas.

6. Conclusão

Este trabalho demonstrou que, ao contrário do que se espera das relações sociais básicas, a localização das pessoas não é um fator determinante para o aumento das suas conexões. Isso foi evidenciado pela análise dos mapas da rede social do *Twitter*, que mostrou uma distribuição não homogênea no espaço. Além disso, observou-se que a rede segue o comportamento típico de uma rede real, apresentando uma distribuição de graus que segue uma lei de potência bem definida.

No que diz respeito à clusterização dos dados, foram utilizados dois métodos distintos para analisar as componentes social e geográfica. Ficou claro como a natureza dos dados influencia esses agrupamentos: enquanto a componente social gerou comunidades de tamanhos variados, sem uma conexão espacial aparente, os agrupamentos espaciais se caracterizaram por um número reduzido de grupos, com divisões bem definidas.

Com base nesse entendimento, foi possível propor uma metodologia capaz de equilibrar as duas dimensões de interesse deste estudo. O *framerowk* foi desenvolvido de forma que possa auxiliar no teste de hipóteses sobre redes geolocalizadas, fornecendo um ambiente controlado para avaliar a influência da geografia na estrutura comunitária das redes. Sua aplicação prática pode tornar análises mais tangíveis em áreas como simulações de redes de transporte urbano, análises de redes sociais com influência geográfica e estudos de propagação de informações baseados em localização.

Os resultados dos testes controlados, realizados com dados gerados internamente, demonstraram que a correlação entre as comunidades identificadas nas camadas de redes geolocalizadas tende a diminuir à medida que a camada funcional se aproxima de uma configuração de rede aleatória. O modelo apresentado conseguiu evidenciar que, partindo de uma estrutura inicial onde a localização espacial era o fator predominante na formação das conexões, a rede pode evoluir para uma configuração em que esse aspecto se torna secundário na conectividade dos nós.

Essa transição foi claramente observada por meio das métricas de similaridade utilizadas para comparar as comunidades entre as camadas que permitiram quantificar o grau de divergência entre as estruturas comunitárias, destacando como a introdução de aleatoriedade nas arestas impacta a correlação entre as camadas. Os resultados reforçam a capacidade do modelo proposto de capturar as mudanças na organização das comunidades, à medida que a influência geográfica é reduzida em favor de outros fatores funcionais.

Por fim, este trabalho apresentou um modelo eficiente para a geração de redes, adequado para a análise de dados que possuem uma componente geográfica associada a outros tipos de dados. Essa abordagem oferece uma forma de visualizar dados não espaciais, atribuindo-lhes um significado espacial relevante.

6.1 Trabalhos futuros

Uma das principais direções para trabalhos futuros é a ampliação da análise com a inclusão de *datasets* mais diversificados e representativos. A coleta de dados adicionais, abrangendo diferentes contextos geográficos e sociais, permitirá avaliar a robustez do modelo proposto em uma variedade maior de cenários, além de enriquecer a generalização dos resultados obtidos.

Outra possibilidade é a experimentação com algoritmos mais avançados para a identificação de comunidades que tenham maior eficácia na detecção de comunidades em redes grandes e complexas, podendo fornecer *insights* adicionais sobre a estrutura das camadas e a correlação entre elas. A utilização de algoritmos mais robustos poderá ainda melhorar a precisão e a confiabilidade das análises realizadas.

Por fim, é recomendável realizar múltiplas iterações do cálculo do AMI para cada configuração de rede, de forma a capturar o comportamento médio da métrica. Essa abordagem reduzirá o impacto de variações aleatórias nos resultados e fornecerá uma visão mais detalhada e precisa sobre a similaridade entre as comunidades das camadas geográfica e funcional.

Referências

BAKKEN, David (David Edward); INIEWSKI, Krzysztof. Smart grids : clouds, communications, open source, and automation. In: SMART grids : clouds, communications, open source, and automation. 1st edition. Boca Raton: Taylor & Francis, 2014. cap. 2. (Devices, Circuits, and Systems). ISBN 9781351831413.

BARABÁSI, Albert-László; PÓSFAI, Márton. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266 1107076269. Disponível em: <http://barabasi.com/networksciencebook/>.

BERLINGERIO, Michele et al. Foundations of Multidimensional Network Analysis. In: PROCEEDINGS of the 2011 International Conference on Advances in Social Networks Analysis and Mining. USA: IEEE Computer Society, 2011. (ASONAM '11), p. 485–489. ISBN 9780769543758. DOI: 10.1109/ASONAM.2011.103. Disponível em: <https://doi.org/10.1109/ASONAM.2011.103>.

BIRNEY, D Scott; GONZALEZ, Guillermo; OESPER, David. **Observational astronomy**. Cambridge, UK: Cambridge University Press, 2006.

BLONDEL, Vincent D et al. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p10008, out. 2008. DOI: 10.1088/1742-5468/2008/10/P10008. Disponível em: <https://dx.doi.org/10.1088/1742-5468/2008/10/P10008>.

CARVALHO, Edilson Alves de; ARAÚJO, Paulo César de. **Leituras cartográficas e interpretações estatísticas I: geografia**. Segunda edição. Natal, RN: Edufrn, 2008.

CLAUSET, Aaron; NEWMAN, M. E. J.; MOORE, Cristopher. Finding community structure in very large networks. **Physical Review E**, American Physical Society (APS), v. 70, n. 6, dez. 2004. ISSN 1550-2376. DOI: 10.1103/physreve.70.066111. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.70.066111>.

COVER, Thomas M. **Elements of information theory**. Hoboken, NJ: John Wiley & Sons, 2006.

D'ALGE, Julio Cesar Lima. Cartografia para geoprocessamento. In: INTRODUÇÃO à ciência da geoinformação. São José dos Campos, SP: INPE, 2001.

DE DOMENICO, Manlio et al. Mathematical Formulation of Multilayer Networks. **Phys. Rev. X**, American Physical Society, v. 3, p. 041022, 4 dez. 2013. DOI: 10.1103/PhysRevX.3.041022. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevX.3.041022>.

ERDOS, Paul; RENYI, Alfred. On the evolution of random graphs. **Publ. Math. Inst. Hungary. Acad. Sci.**, v. 5, p. 17–61, 1960.

ERDÖS, P; RÉNYI, A. On Random Graphs I. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

F. COSTA, Luciano da. **Further Generalizations of the Jaccard Index**. [S.l.: s.n.], 2021. arXiv: 2110.09619 [cs.LG]. Disponível em: <https://arxiv.org/abs/2110.09619>.

FACCHINETTI-MANNONE, Valérie. A methodological approach to analyze the territorial appropriation of high-speed rail from interactions between actions and representations of local actors. **European Planning Studies**, Taylor & Francis (Routledge), v. 27, n. 3, p. 461–482, 2019. DOI: 10.1080/09654313.2018.1562653. Disponível em: <https://hal.science/hal-02066155>.

GHOMSHEH, Maliheh; KAMANDI, Ali. C. elegans Neural Network Analysis. **Journal of Algorithms and Computation**, University of Tehran, v. 54, n. 2, p. 71–91, 2022. ISSN 2476-2776. DOI: 10.22059/jac.2022.90482. eprint: https://jac.ut.ac.ir/article_90482_2ff8cbb760522e52219d98530ff4080b.pdf. Disponível em: <https://jac.ut.ac.ir/article_90482.html>.

HAGBERG, Aric A.; SCHULT, Daniel A.; SWART, Pieter J. Exploring Network Structure, Dynamics, and Function using NetworkX. In_____. **Proceedings of the 7th Python in Science Conference**. Pasadena, CA USA: [s.n.], 2008. P. 11–15.

IBGE. **As projeções cartográficas**. [S.l.: s.n.], 2024. https://atlasescolar.ibge.gov.br/cartografia/21733-as-projecoes-cartograficas.html. Acessedo em: 10/01/2025.

ISO CENTRAL SECRETARY. **SyStandard representation of geographic point location by coordinates**. en. Geneva, CH, 2022. Disponível em: <https://www.iso.org/standard/75147.html>.

KAUR, Gurusharan; TRIPATHI, Namrata; VERMA, Mona. Applications of Graph Theory in Science and Computer Science. **International Journal of Advances in Engineering and Management (IJAEM)**, v. 2, p. 736, 2008.

KING, David; ABOUDINA, Aya; SHALABY, Amer. Evaluating transit network resilience through graph theory and demand-elastic measures: Case study of the Toronto transit system. **Journal of Transportation Safety & Security**, v. 12, p. 1–21, fev. 2019. DOI: 10.1080/19439962.2018.1556229.

KIVELA, M. et al. Multilayer networks. **Journal of Complex Networks**, Oxford University Press (OUP), v. 2, n. 3, p. 203–271, jul. 2014. ISSN 2051-1329. DOI: 10.1093/comnet/cnu016. Disponível em: <http://dx.doi.org/10.1093/comnet/cnu016>.

LAPAINE, Miljenko; FRANČULA, Nedjeljko. Map Projections Classification. **Geographies**, v. 2, n. 2, p. 274–285, 2022. ISSN 2673-7086. DOI: 10.3390/geographies2020019. Disponível em: <https://www.mdpi.com/2673-7086/2/2/19>.

LIN, Frank Yeong-Sung et al. Resource Allocation and Multisession Routing Algorithms in Coordinated Multipoint Wireless Communication Networks. **IEEE Systems Journal**, v. 12, n. 3, p. 2226–2237, 2018. DOI: 10.1109/JSYST.2017.2687102.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press. [S.l.: s.n.], 1967. Disponível em: <https://api.semanticscholar.org/CorpusID:6278891>.

NATIONAL COORDINATION OFFICE FOR SPACE-BASED POSITIONING, NAVIGATION, AND TIMING. **GPS - The Global Positioning System: Space Segment**. [S.l.: s.n.], 2022. https://www.https://www.gps.gov/systems/gps/space/ [Acessado em: 28/06/2024].

OLIVEIRA, Roberto; SILVA, Daniel. SISTEMAS DE PROJEÇÃO TRANSVERSA DE MERCATOR NO GEORREFERENCIAMENTO DE IMÓVEIS RURAIS. In_____. **Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**. Recife, PE: [s.n.], jan. 2012.

RIDPATH, Ian. **A dictionary of astronomy**. 2nd revised edition. New York, USA: Oxford University Press, USA, 2012.

SAEED, Mozamel M; AL AGHBARI, Zaher; ALSHARIDAH, Mohammed. Big data clustering techniques based on Spark: a literature review. en. **PeerJ Comput Sci**, United States, v. 6, e321, nov. 2020.

SCHUSTER, Peter. Networks in biology: Handling biological complexity requires novel inputs into network theory. English. **Complexity**, Wiley-Blackwell, v. 16, n. 4, p. 6–9, 2011. ISSN 1076-2787. DOI: 10.1002/cplx.20375.

TANG, Vicente; PAINHO, Marco. Content-location relationships: a framework to explore correlations between space-based and place-based user-generated content. **International Journal of Geographical Information Science**, Taylor & Francis, v. 37, n. 8, p. 1840–1871, 2023.

TILLEMA, Taede; DIJST, Martin; SCHWANEN, Tim. Face-to-face and electronic communications in maintaining social networks: the influence of geographical and relational distance and of information content. **New Media & Society**, v. 12, p. 965–983, 2010. Disponível em: <https://api.semanticscholar.org/CorpusID:9630679>.

TRAAG, V. A. Faster unfolding of communities: Speeding up the Louvain algorithm. **Physical Review E**, American Physical Society (APS), v. 92, n. 3, set. 2015. ISSN 1550-2376. DOI: 10.1103/physreve.92.032801. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.92.032801>.

VIEGAS, Eduardo et al. **A complexity perspective on the geographical location of companies: How distance reduce trade between firms**. [S.l.: s.n.], 2023. arXiv: 2311.15760 [physics.soc-ph]. Disponível em: <https://arxiv.org/abs/2311.15760>.

VIEIRA, Antônio José Berutti et al. **Cartografia**. Curitiba, PR: UFPR, 2004.

VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: PROCEEDINGS of the 26th annual international conference on machine learning. [S.l.: s.n.], 2009. P. 1073–1080.

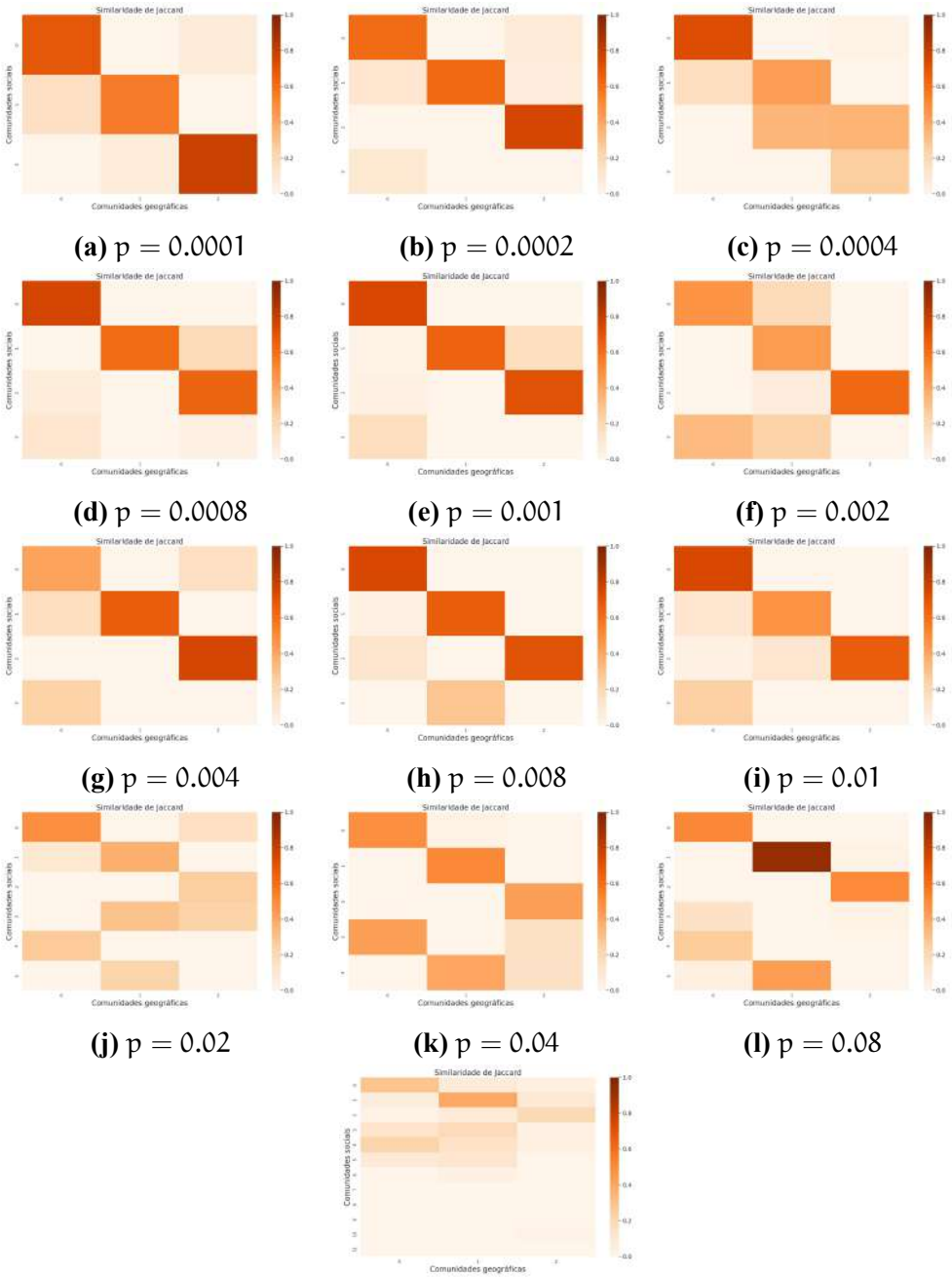
WATTS, Duncan J.; STROGATZ, Steven H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, jun. 1998. ISSN 1476-4687. DOI: 10.1038/30918. Disponível em: <https://doi.org/10.1038/30918>.

WAXMAN, B.M. Routing of multipoint connections. **IEEE Journal on Selected Areas in Communications**, v. 6, n. 9, p. 1617–1622, 1988. DOI: 10.1109/49.12889.

WAXMAN, Bernard M. **New Approximation Algorithms for the Steiner Tree Problem**. [S.l.]: Washington University, Department of Computer Science, 1989.

WIKIPEDIA CONTRIBUTORS. **A diagram of spherical coordinates**. [S.l.: s.n.], 2008. https://commons.wikimedia.org/wiki/File:Spherical_Coordinates_(Latitude,_Longitude).svg. [Acessado em: 15/12/2024].

XU, Wanyue; ZHANG, Zhongzhi. Optimal Scale-Free Small-World Graphs with Minimum Scaling of Cover Time. **ACM Transactions on Knowledge Discovery from Data**, Association for Computing Machinery (ACM), v. 17, n. 7, p. 1–19, abr. 2023. ISSN 1556-472X. DOI: 10.1145/3583691. Disponível em: <http://dx.doi.org/10.1145/3583691>.



(a) $p = 0.0001$ **(b)** $p = 0.0002$ **(c)** $p = 0.0004$

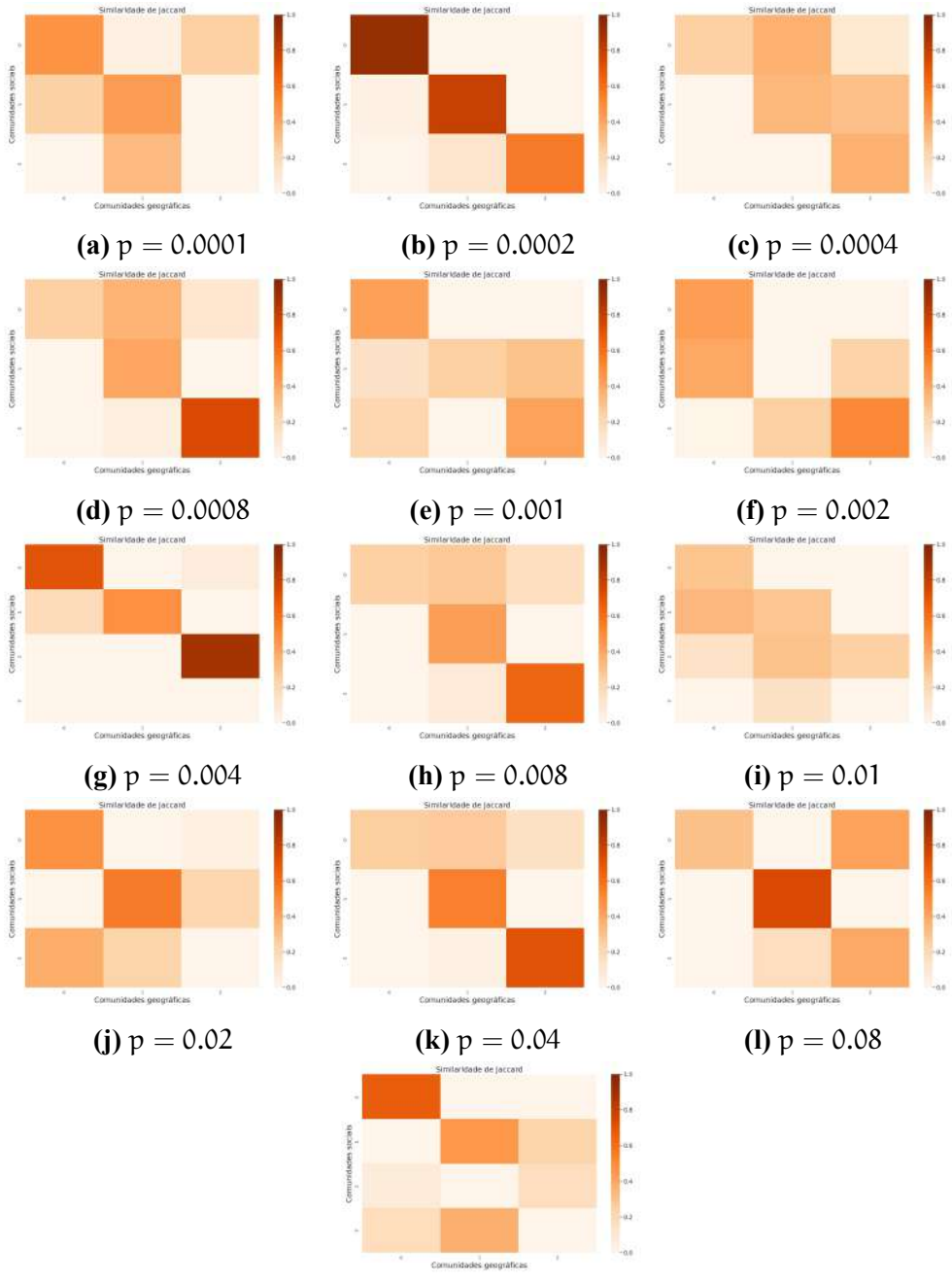
(d) $p = 0.0008$ **(e)** $p = 0.001$ **(f)** $p = 0.002$

(g) $p = 0.004$ **(h)** $p = 0.008$ **(i)** $p = 0.01$

(j) $p = 0.02$ **(k)** $p = 0.04$ **(l)** $p = 0.08$

(m) $p = 0.16$

Figura 28: Matrizes de Jaccard para $k = 10$.



(a) $p = 0.0001$ **(b)** $p = 0.0002$ **(c)** $p = 0.0004$

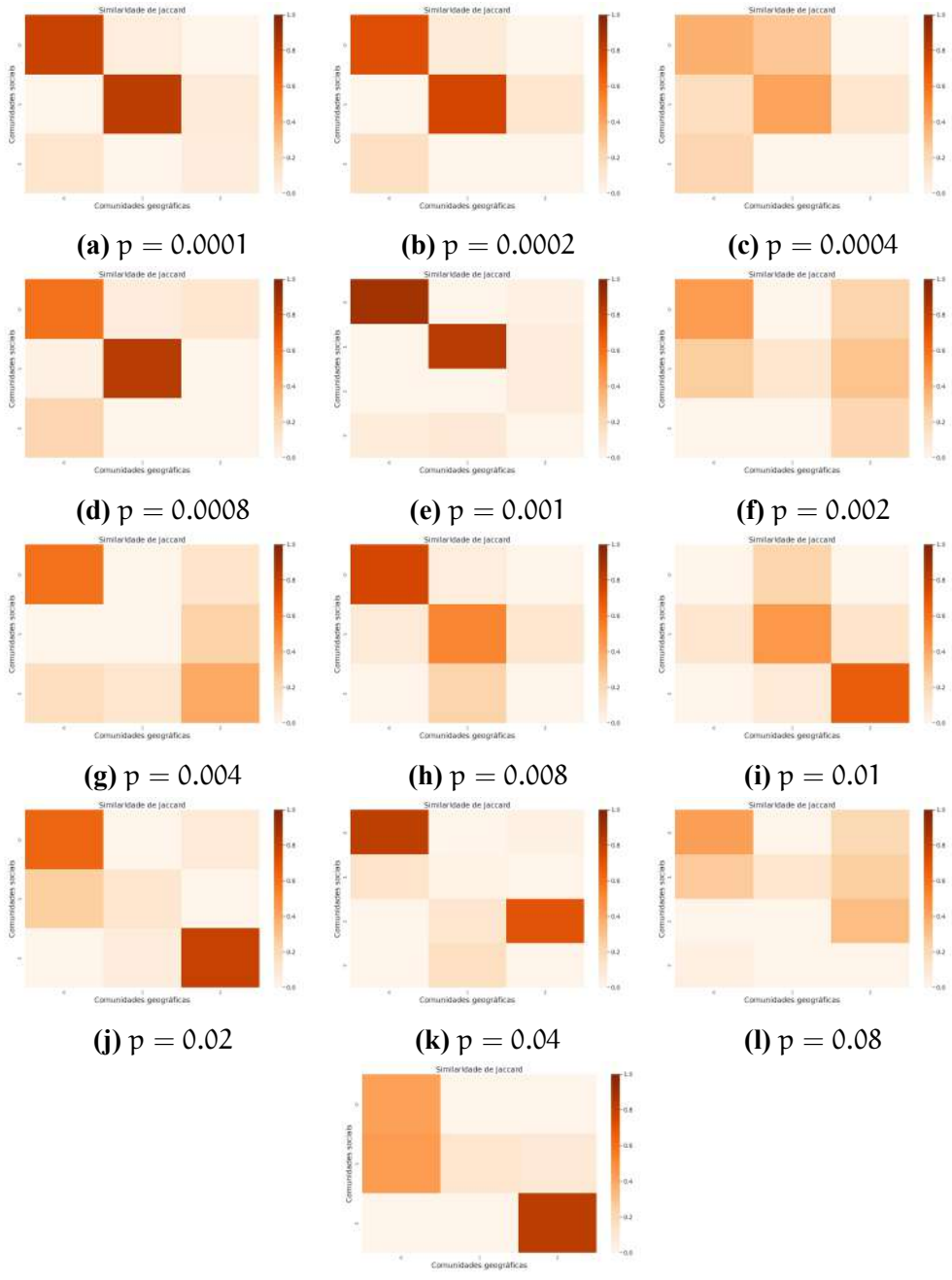
(d) $p = 0.0008$ **(e)** $p = 0.001$ **(f)** $p = 0.002$

(g) $p = 0.004$ **(h)** $p = 0.008$ **(i)** $p = 0.01$

(j) $p = 0.02$ **(k)** $p = 0.04$ **(l)** $p = 0.08$

(m) $p = 0.16$

Figura 29: Matrizes de Jaccard para $k = 20$.



(a) $p = 0.0001$ **(b)** $p = 0.0002$ **(c)** $p = 0.0004$

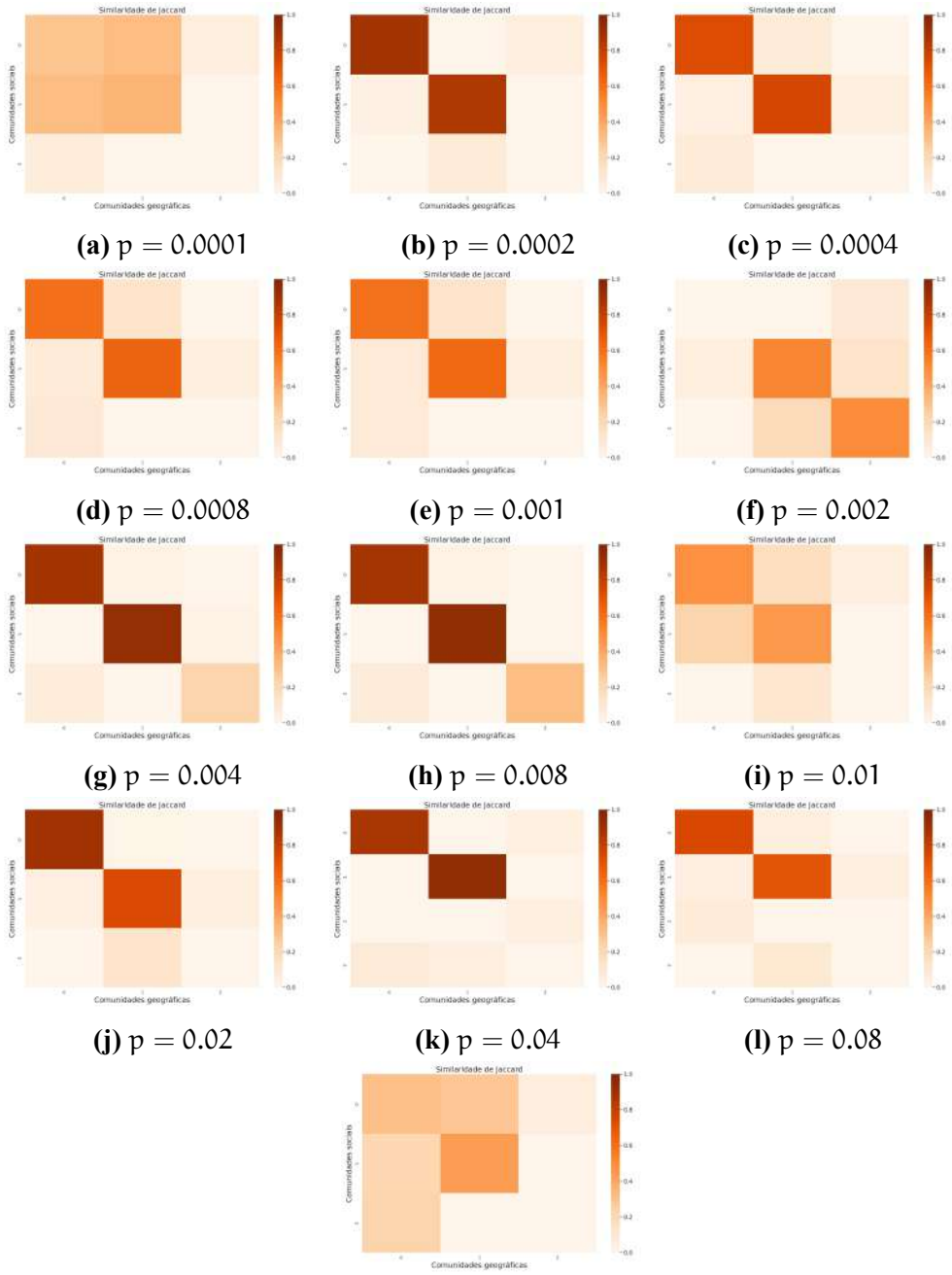
(d) $p = 0.0008$ **(e)** $p = 0.001$ **(f)** $p = 0.002$

(g) $p = 0.004$ **(h)** $p = 0.008$ **(i)** $p = 0.01$

(j) $p = 0.02$ **(k)** $p = 0.04$ **(l)** $p = 0.08$

(m) $p = 0.16$

Figura 30: Matrizes de Jaccard para $k = 40$.



(a) $p = 0.0001$ **(b)** $p = 0.0002$ **(c)** $p = 0.0004$

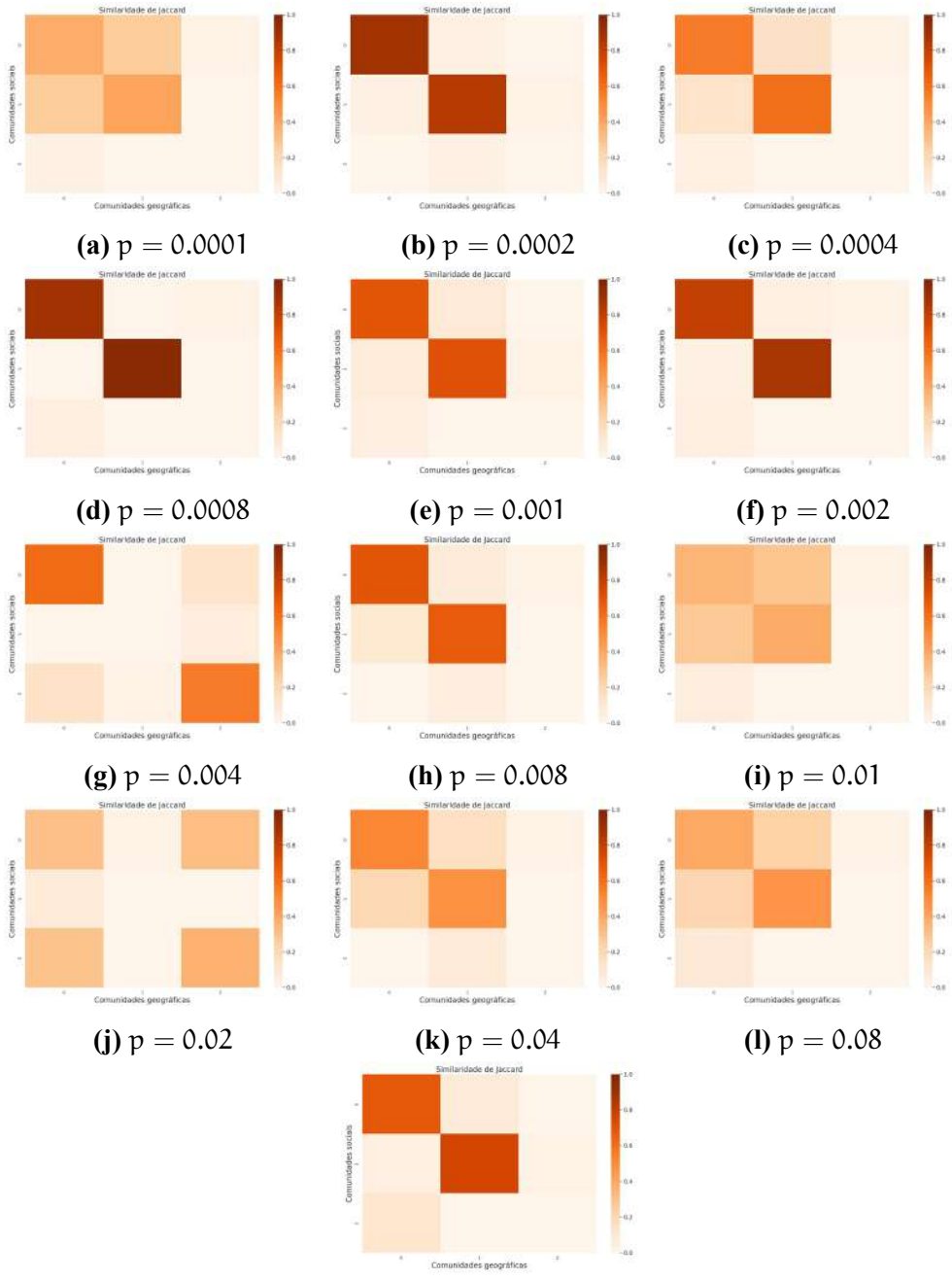
(d) $p = 0.0008$ **(e)** $p = 0.001$ **(f)** $p = 0.002$

(g) $p = 0.004$ **(h)** $p = 0.008$ **(i)** $p = 0.01$

(j) $p = 0.02$ **(k)** $p = 0.04$ **(l)** $p = 0.08$

(m) $p = 0.16$

Figura 31: Matrizes de Jaccard para $k = 80$.



(a) $p = 0.0001$ **(b)** $p = 0.0002$ **(c)** $p = 0.0004$

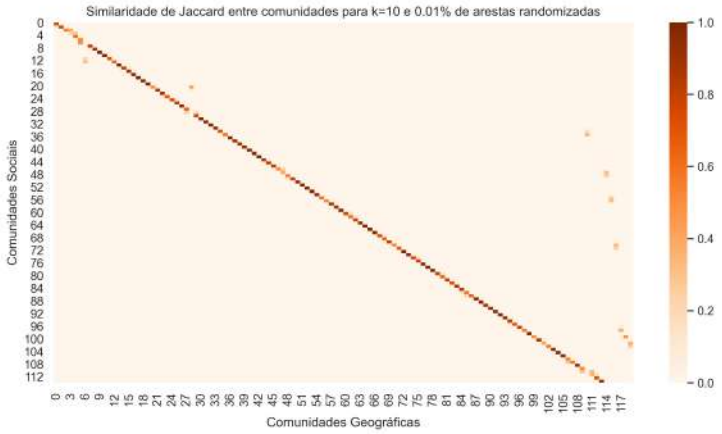
(d) $p = 0.0008$ **(e)** $p = 0.001$ **(f)** $p = 0.002$

(g) $p = 0.004$ **(h)** $p = 0.008$ **(i)** $p = 0.01$

(j) $p = 0.02$ **(k)** $p = 0.04$ **(l)** $p = 0.08$

(m) $p = 0.16$

Figura 32: Matrizes de Jaccard para $k = 160$.



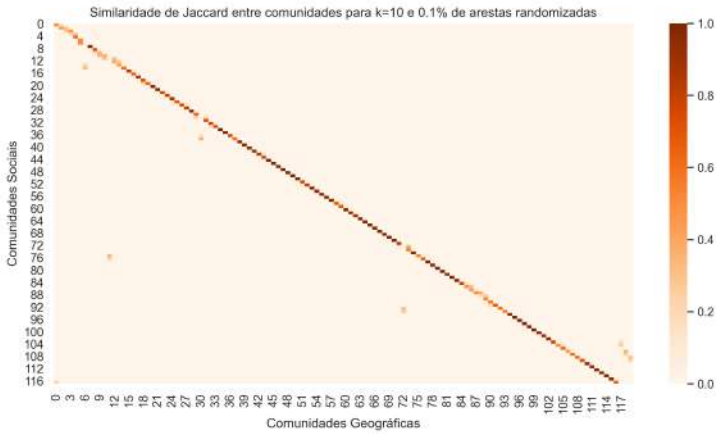
(a) $p = 0.0001$

(b) $p = 0.0002$

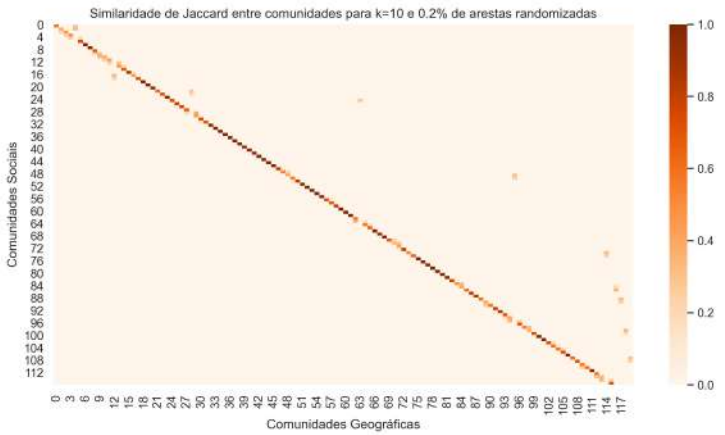
(c) $p = 0.0004$

Figura 33: Matrizes de Jaccard para $k = 10$ (Parte 1).

(a) $p = 0.0008$

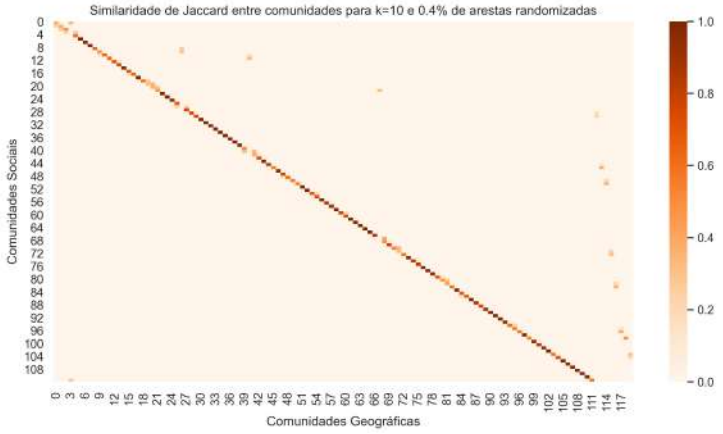


(b) $p = 0.001$

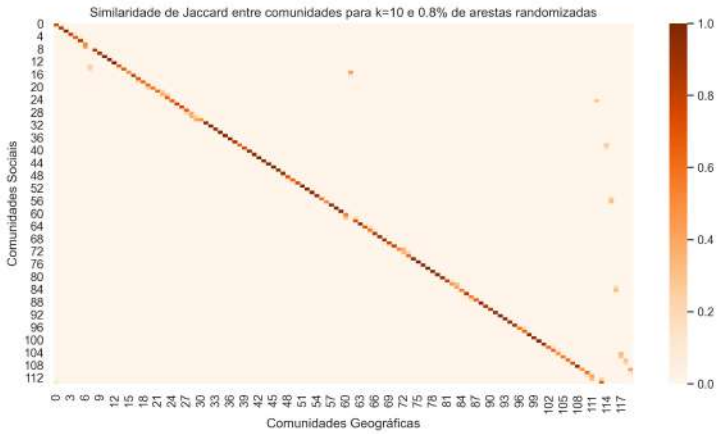


(c) $p = 0.002$

Figura 34: Matrizes de Jaccard para $k = 10$ (Parte 2).



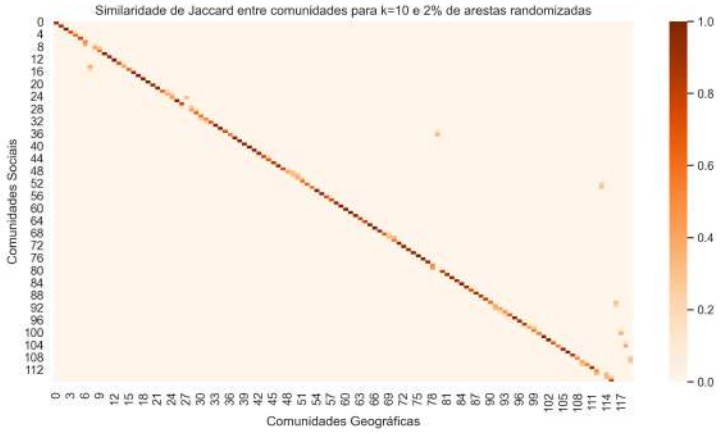
(a) $p = 0.004$



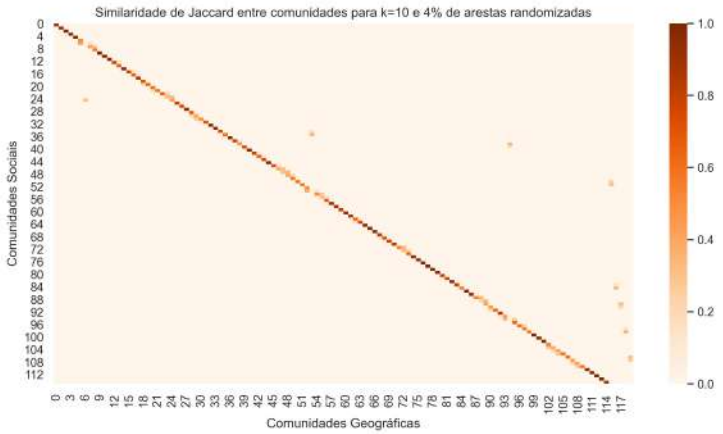
(b) $p = 0.008$

(c) $p = 0.01$

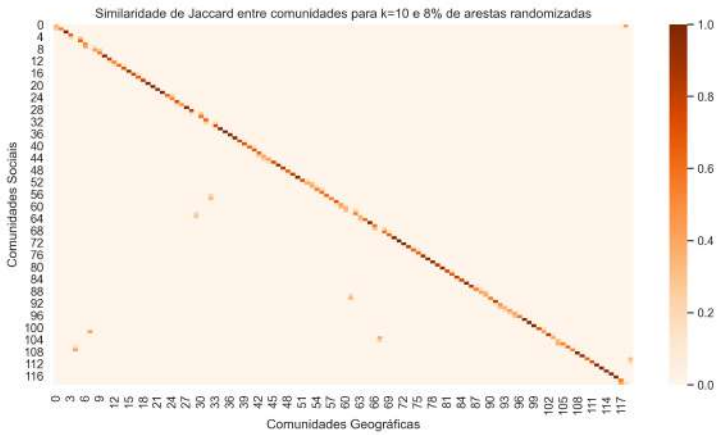
Figura 35: Matrizes de Jaccard para $k = 10$ (Parte 3).



(a) $p = 0.02$



(b) $p = 0.04$

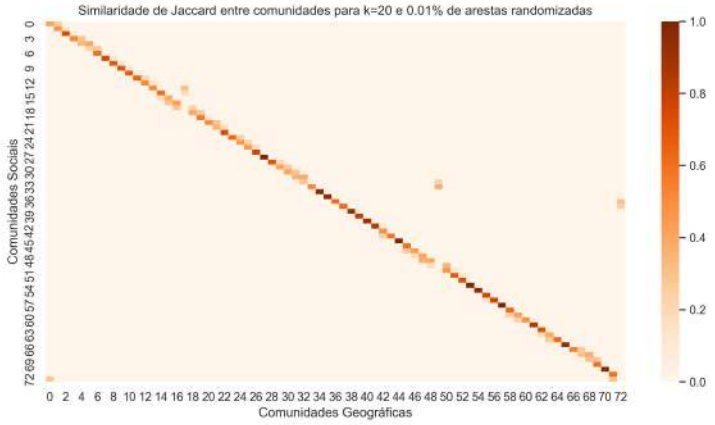


(c) $p = 0.08$

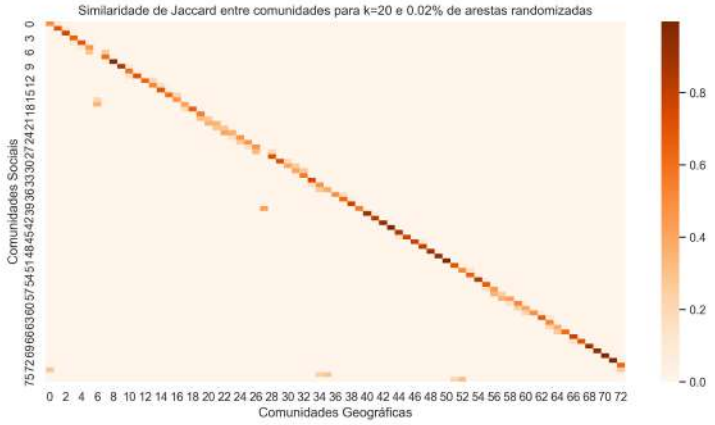
Figura 36: Matrizes de Jaccard para $k = 10$ (Parte 4).

(a) $p = 0.16$

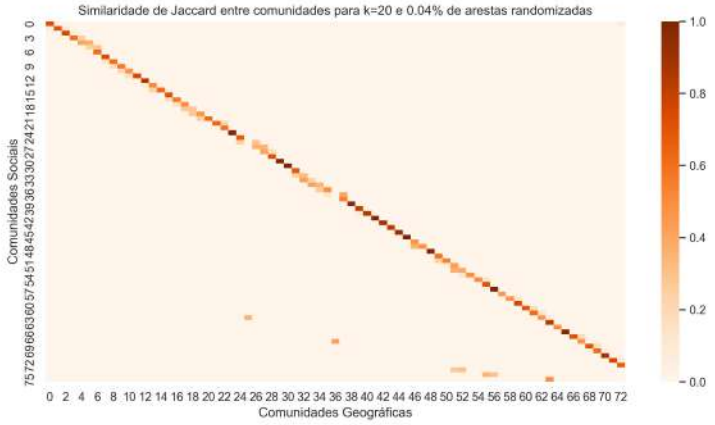
Figura 37: Matrizes de Jaccard para $k = 10$ (Parte 5).



(a) $p = 0.0001$

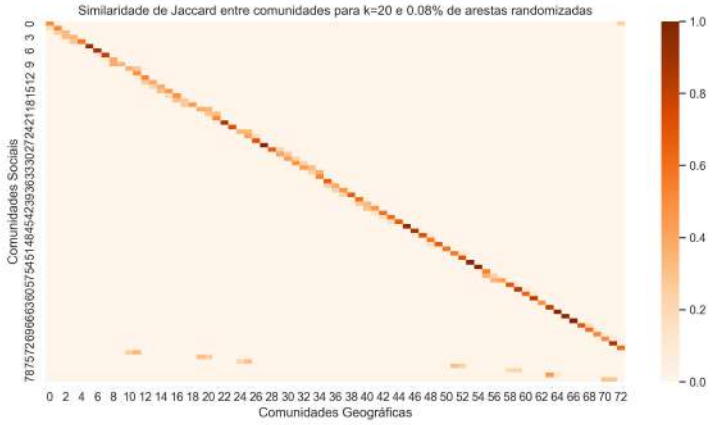


(b) $p = 0.0002$

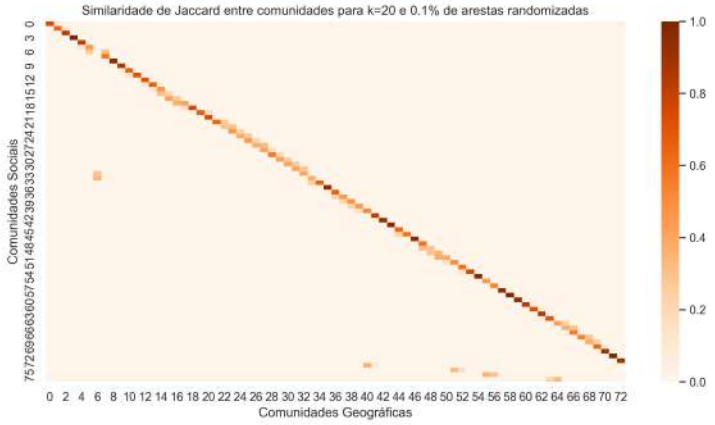


(c) $p = 0.0004$

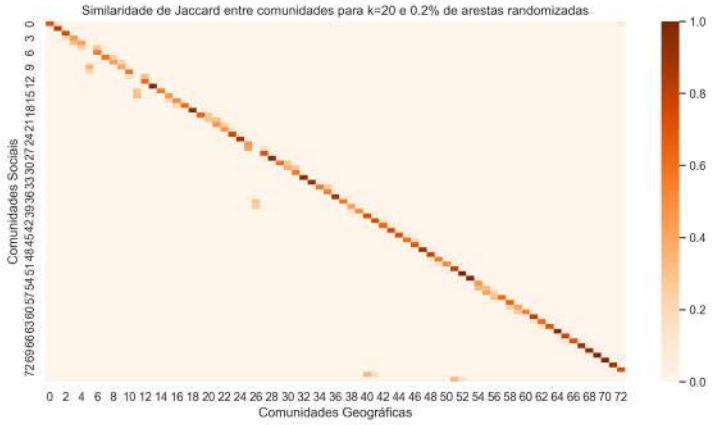
Figura 38: Matrizes de Jaccard para $k = 20$ (Parte 1).



(a) $p = 0.0008$

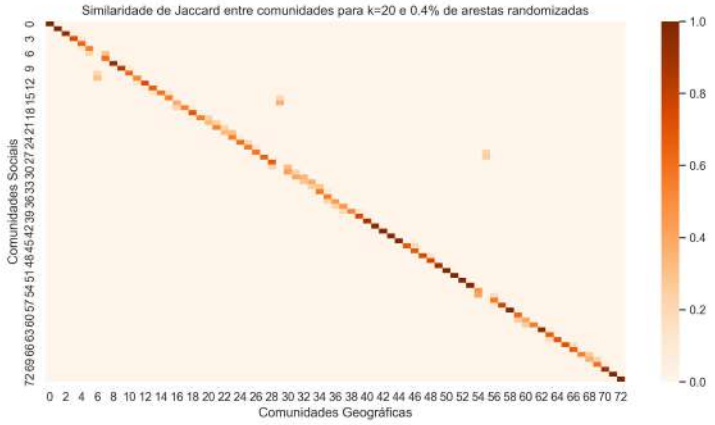


(b) $p = 0.001$

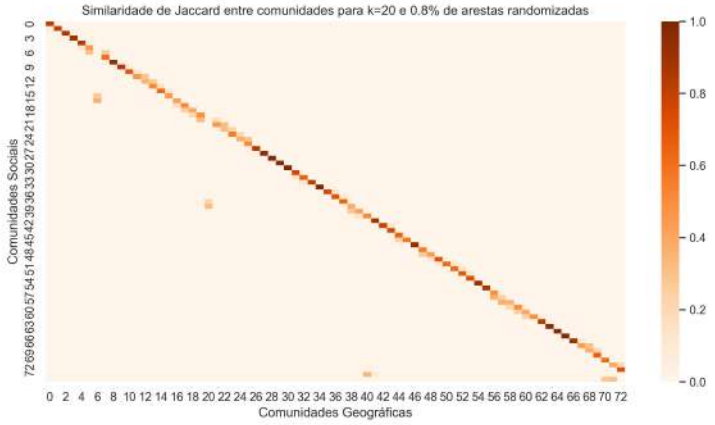


(c) $p = 0.002$

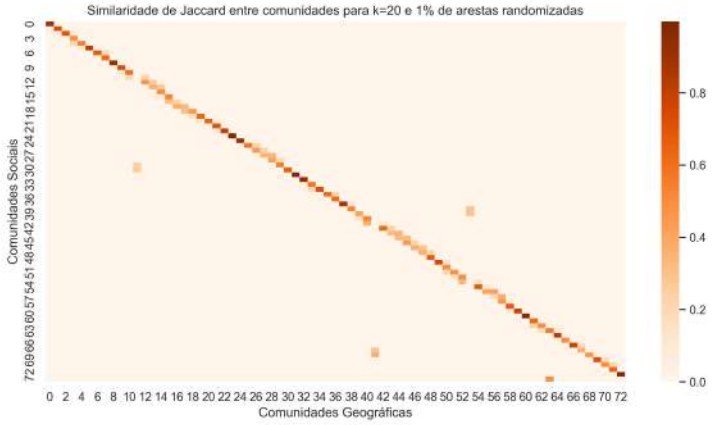
Figura 39: Matrizes de Jaccard para $k = 20$ (Parte 2).



(a) $p = 0.004$

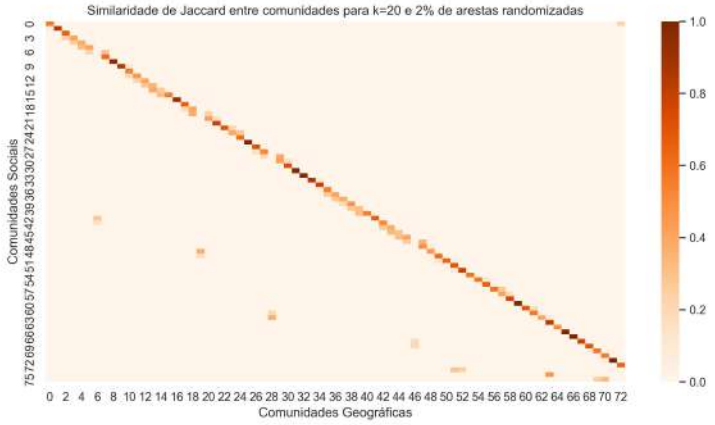


(b) $p = 0.008$

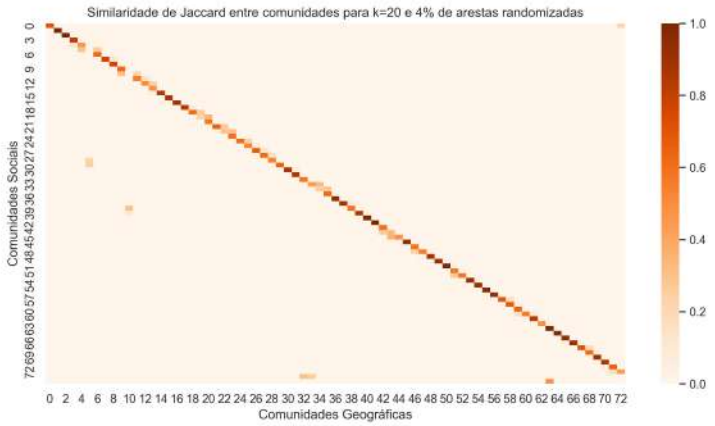


(c) $p = 0.01$

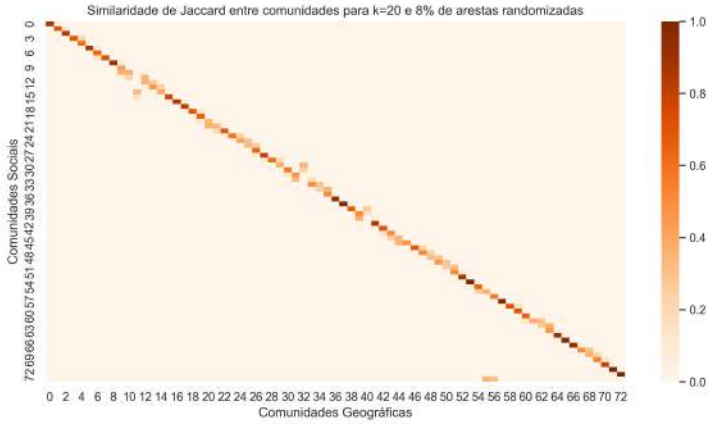
Figura 40: Matrizes de Jaccard para $k = 20$ (Parte 3).



(a) $p = 0.02$

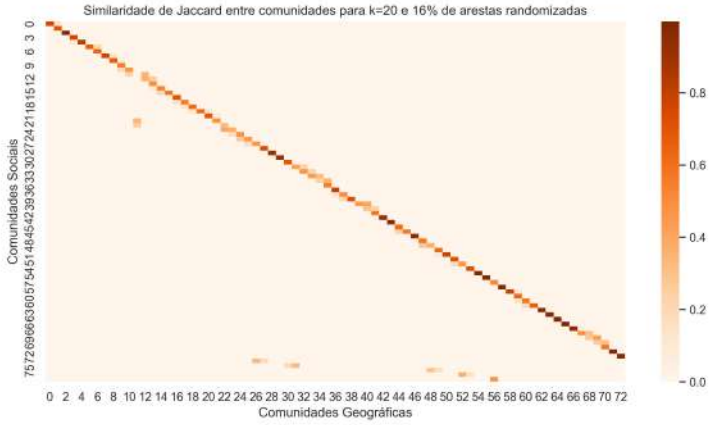


(b) $p = 0.04$



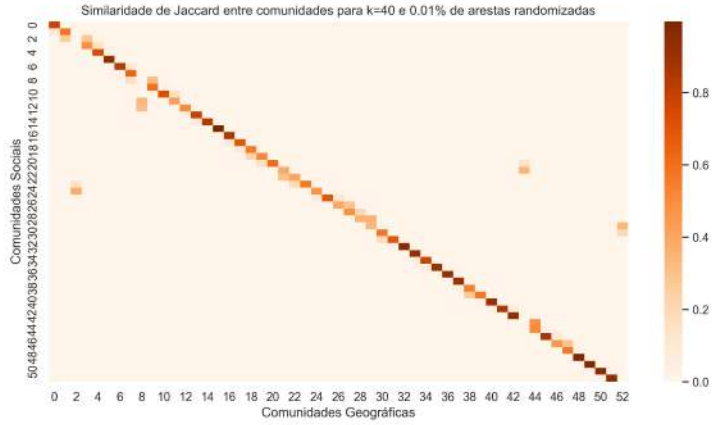
(c) $p = 0.08$

Figura 41: Matrizes de Jaccard para $k = 20$ (Parte 4).

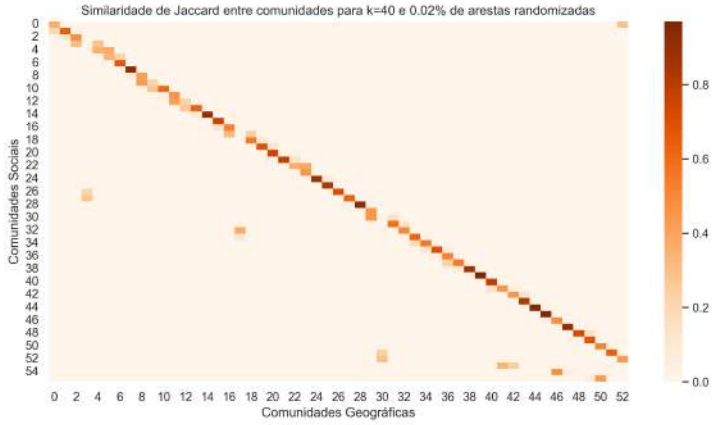


(a) $p = 0.16$

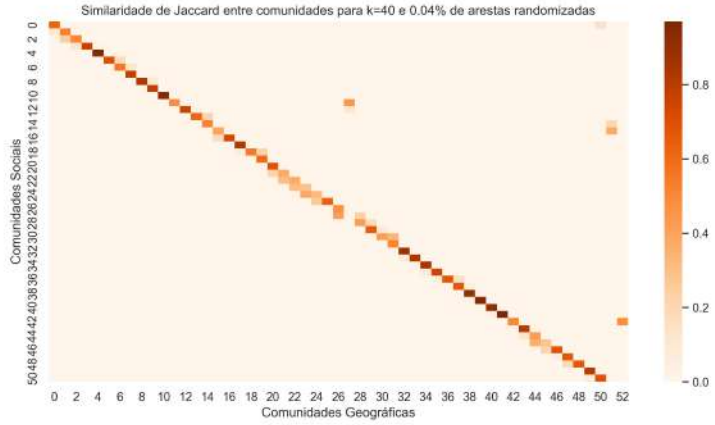
Figura 42: Matrizes de Jaccard para $k = 20$ (Parte 5).



(a) $p = 0.0001$

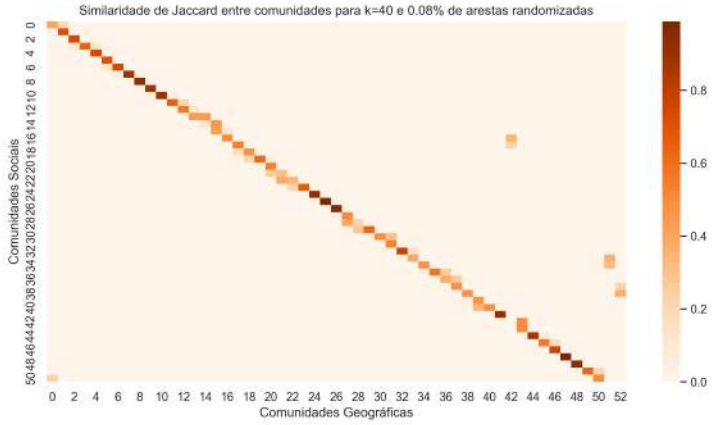


(b) $p = 0.0002$

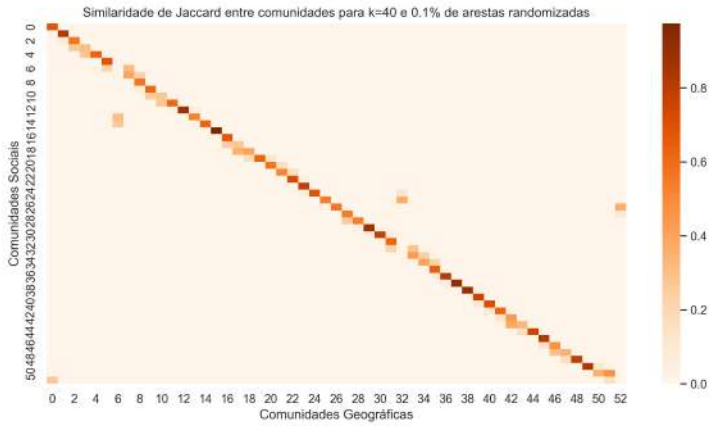


(c) $p = 0.0004$

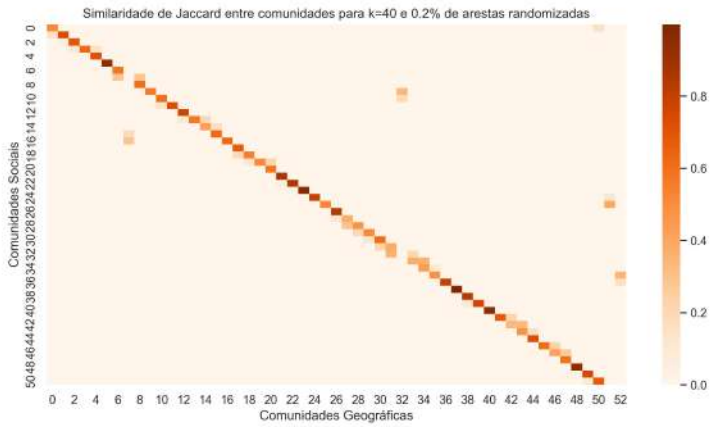
Figura 43: Matrizes de Jaccard para $k = 40$ (Parte 1).



(a) $p = 0.0008$

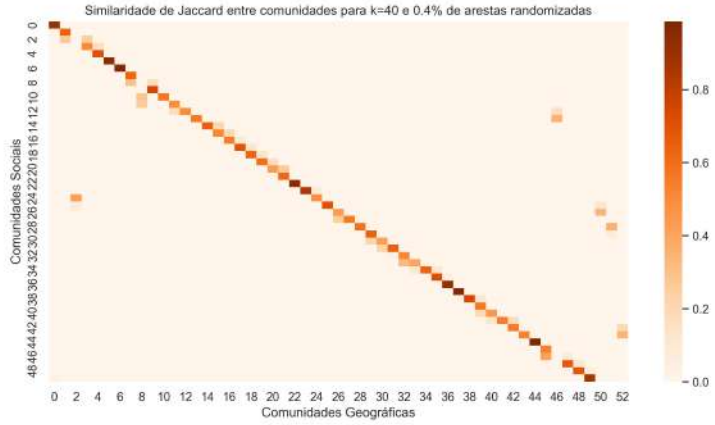


(b) $p = 0.001$

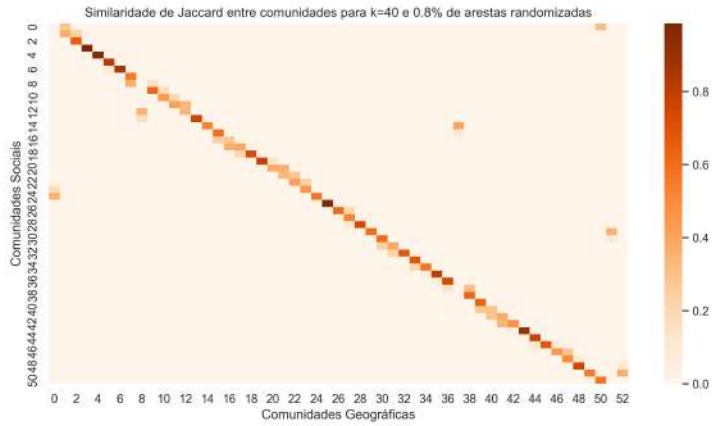


(c) $p = 0.002$

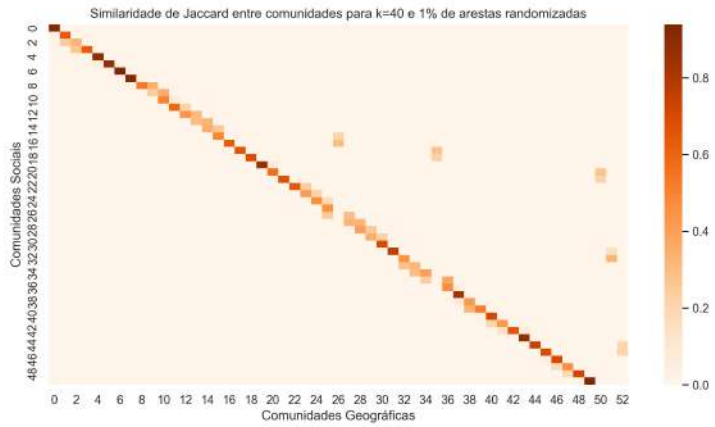
Figura 44: Matrizes de Jaccard para $k = 40$ (Parte 2).



(a) $p = 0.004$

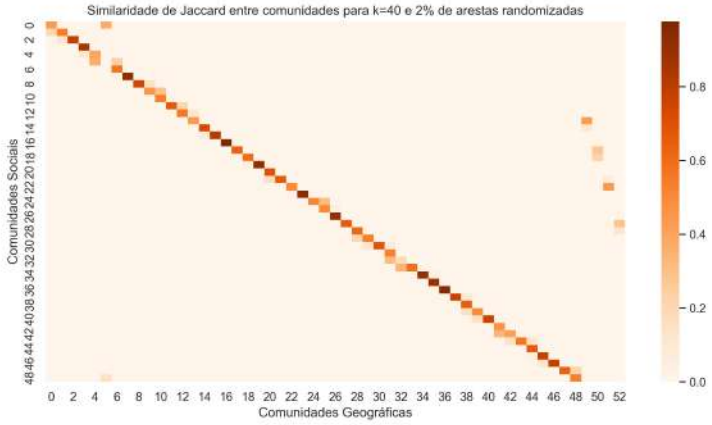


(b) $p = 0.008$

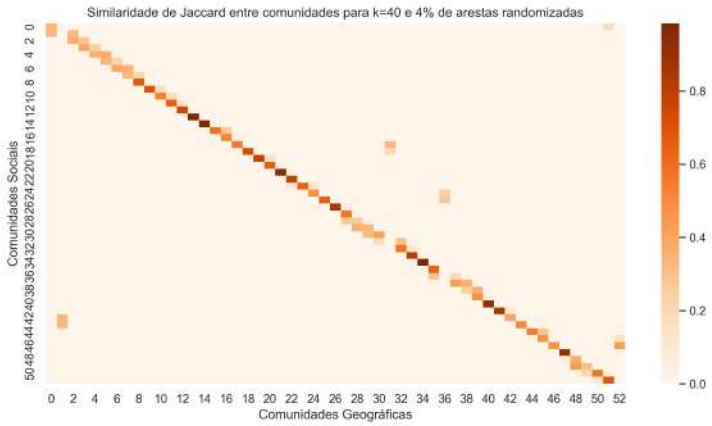


(c) $p = 0.01$

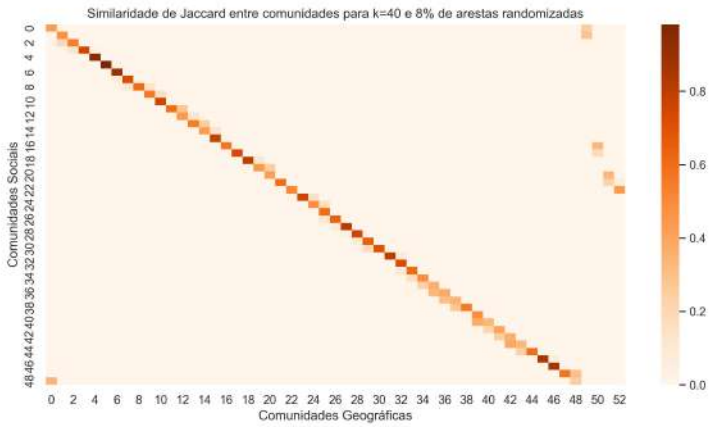
Figura 45: Matrizes de Jaccard para $k = 40$ (Parte 3).



(a) $p = 0.02$

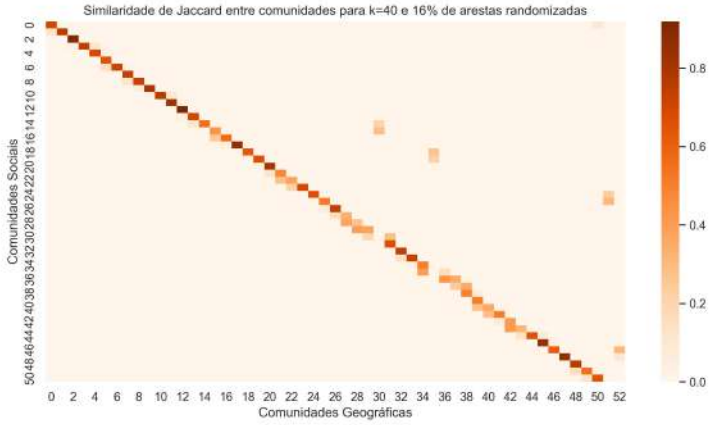


(b) $p = 0.04$



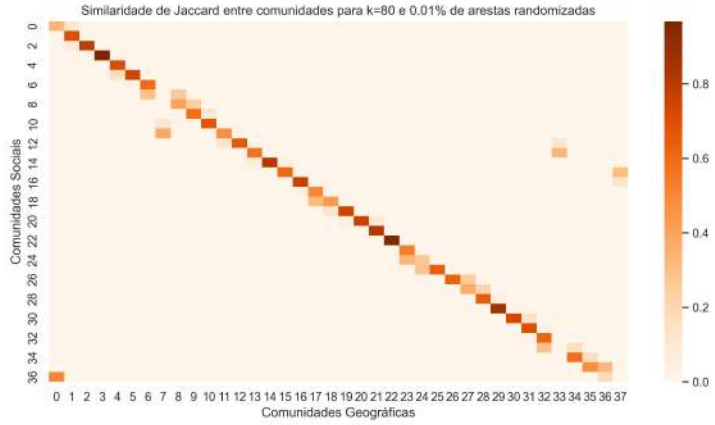
(c) $p = 0.08$

Figura 46: Matrizes de Jaccard para $k = 40$ (Parte 4).

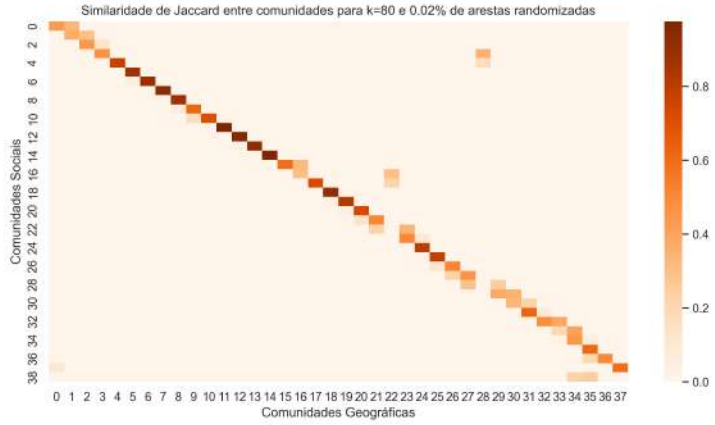


(a) $p = 0.16$

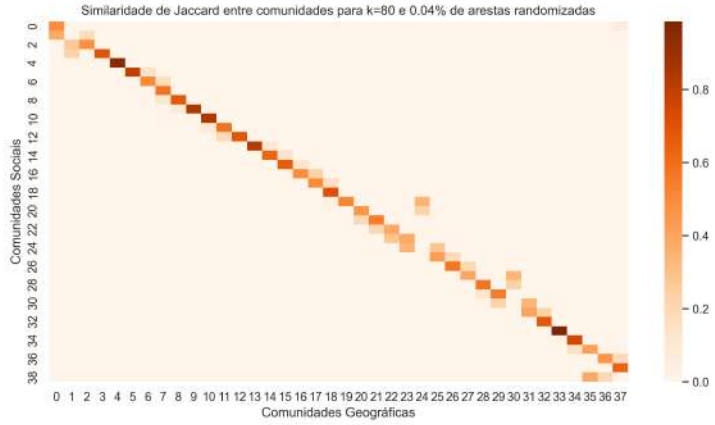
Figura 47: Matrizes de Jaccard para $k = 40$ (Parte 5).



(a) $p = 0.0001$

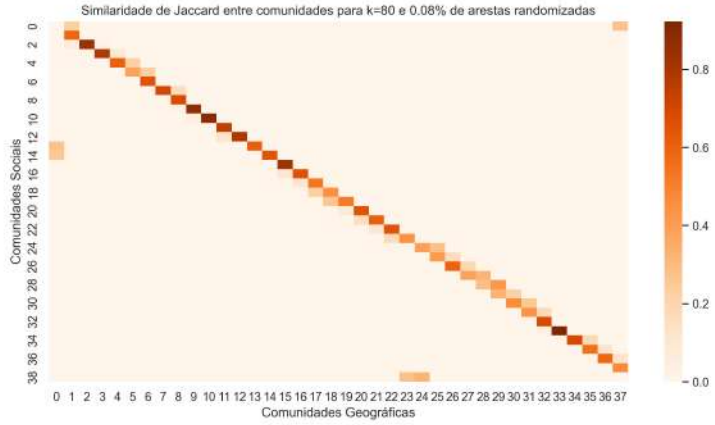


(b) $p = 0.0002$

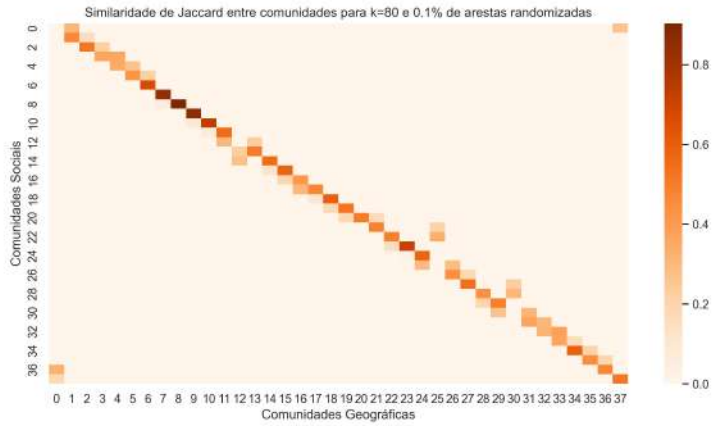


(c) $p = 0.0004$

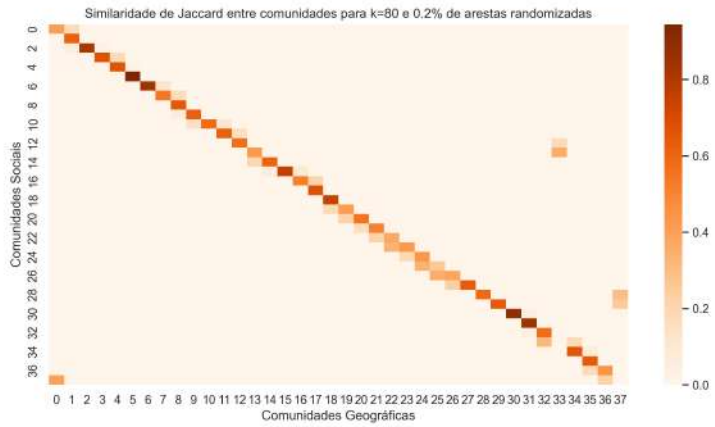
Figura 48: Matrizes de Jaccard para $k = 80$ (Parte 1).



(a) $p = 0.0008$

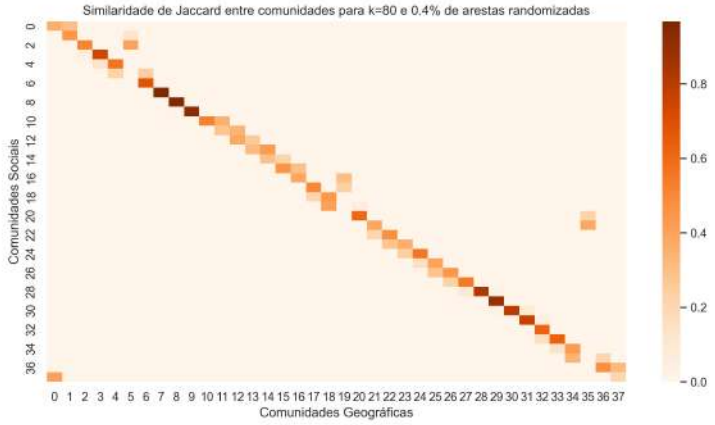


(b) $p = 0.001$

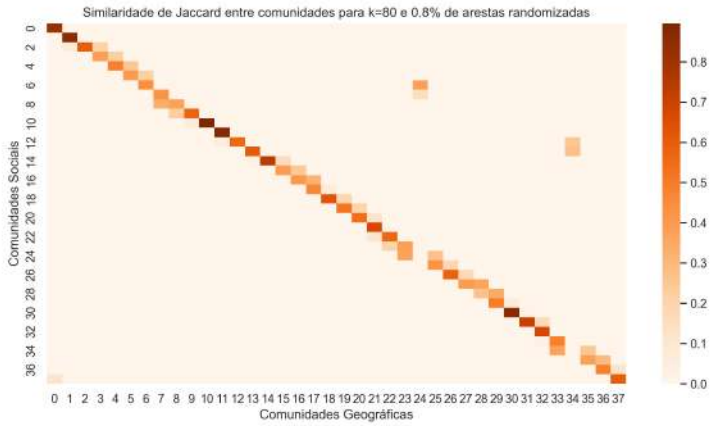


(c) $p = 0.002$

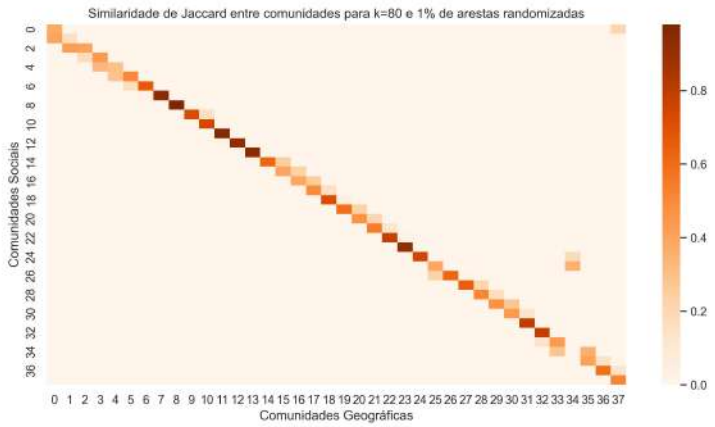
Figura 49: Matrizes de Jaccard para $k = 80$ (Parte 2).



(a) $p = 0.004$

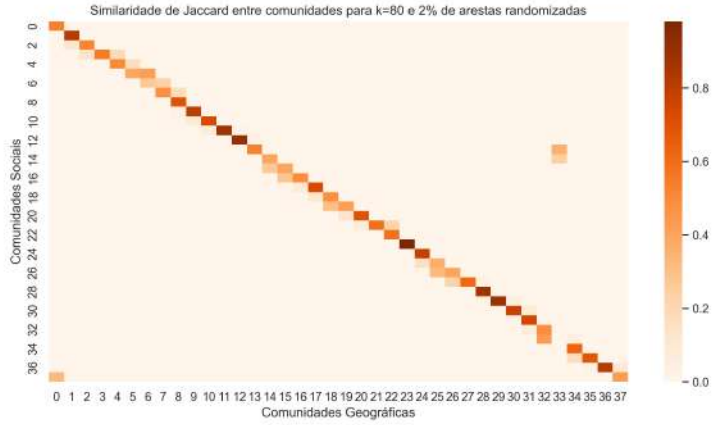


(b) $p = 0.008$

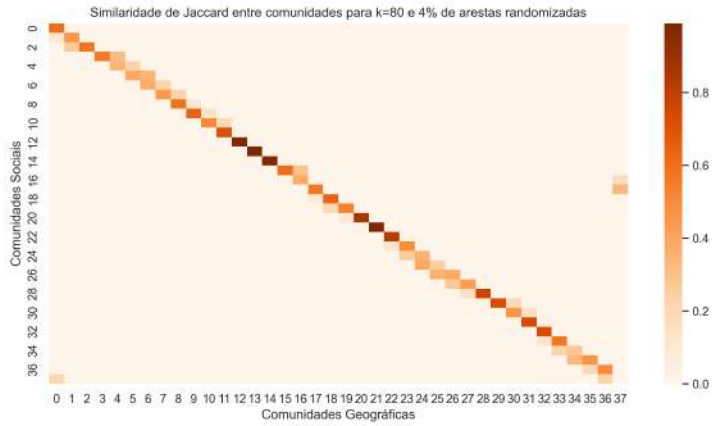


(c) $p = 0.01$

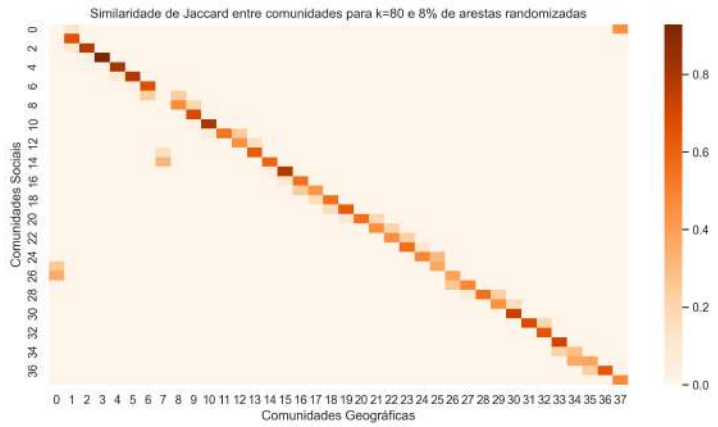
Figura 50: Matrizes de Jaccard para $k = 80$ (Parte 3).



(a) $p = 0.02$

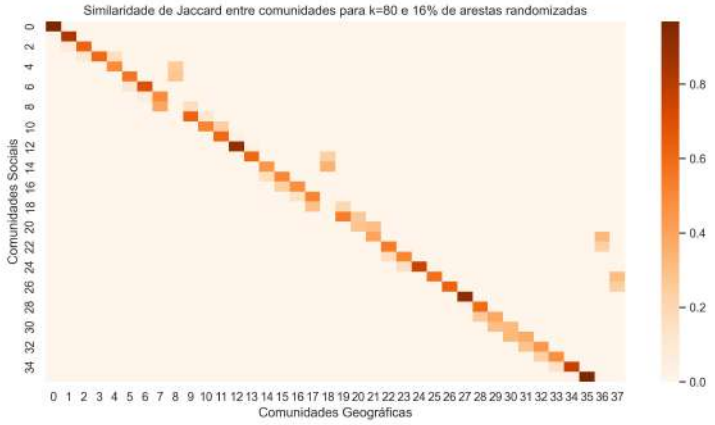


(b) $p = 0.04$



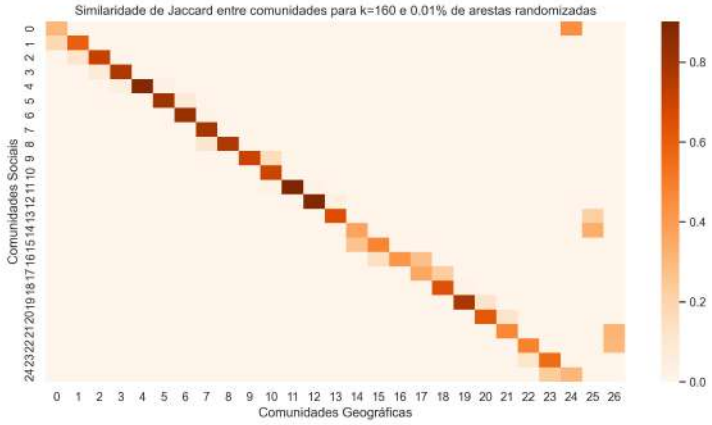
(c) $p = 0.08$

Figura 51: Matrizes de Jaccard para $k = 80$ (Parte 4).

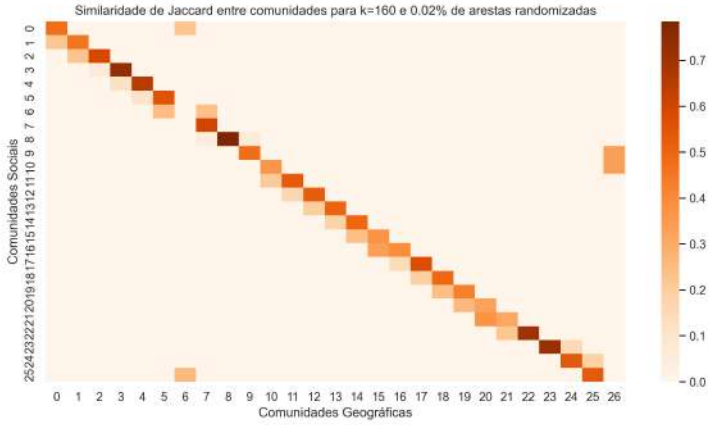


(a) $p = 0.16$

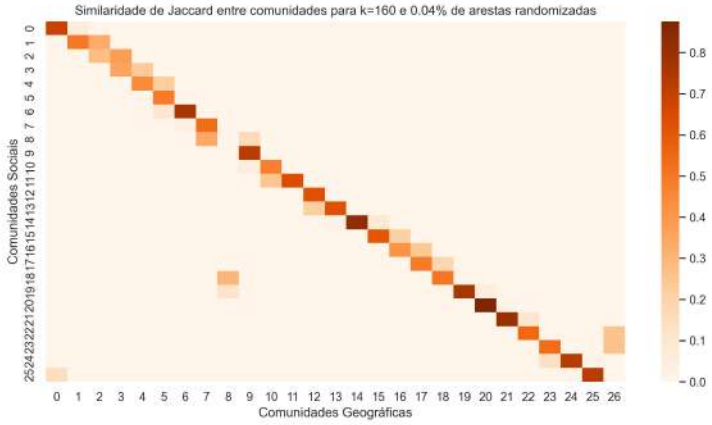
Figura 52: Matrizes de Jaccard para $k = 80$ (Parte 5).



(a) $p = 0.0001$

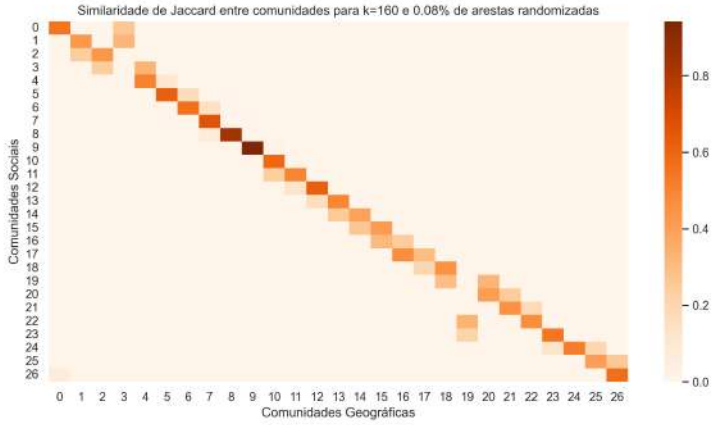


(b) $p = 0.0002$

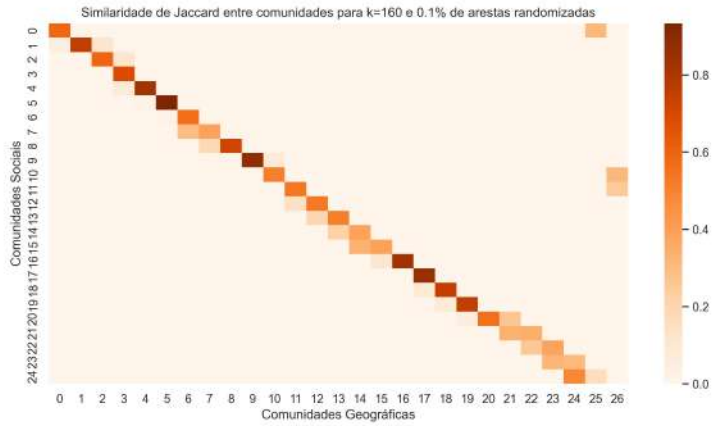


(c) $p = 0.0004$

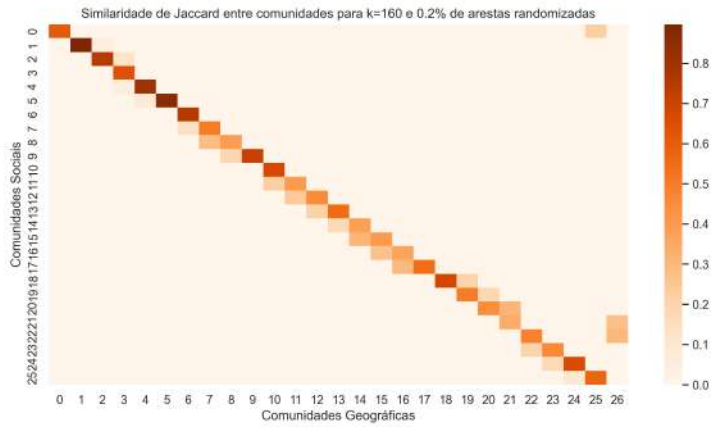
Figura 53: Matrizes de Jaccard para $k = 160$ (Parte 1).



(a) $p = 0.0008$

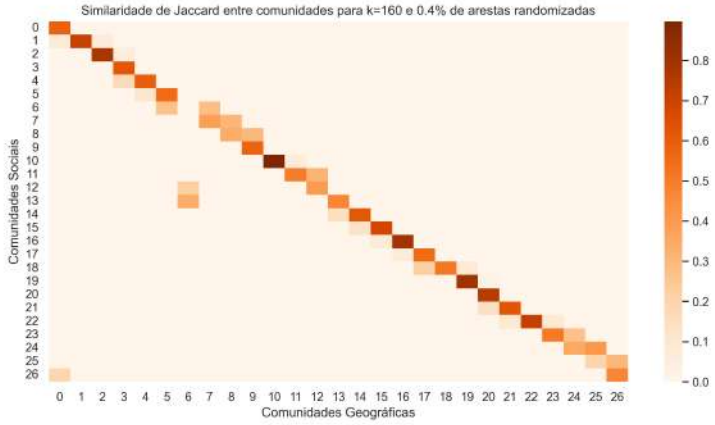


(b) $p = 0.001$

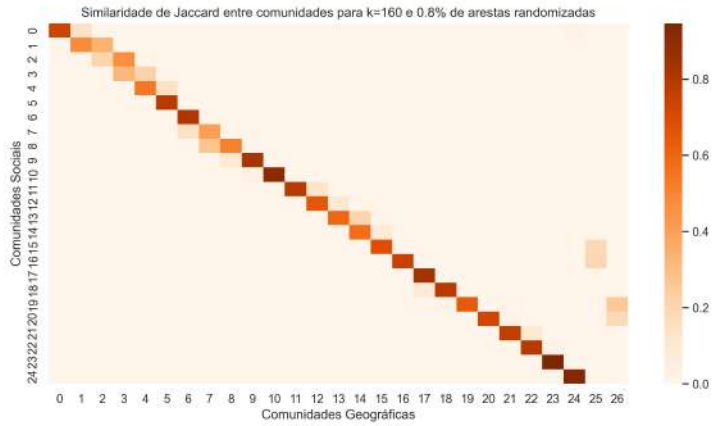


(c) $p = 0.002$

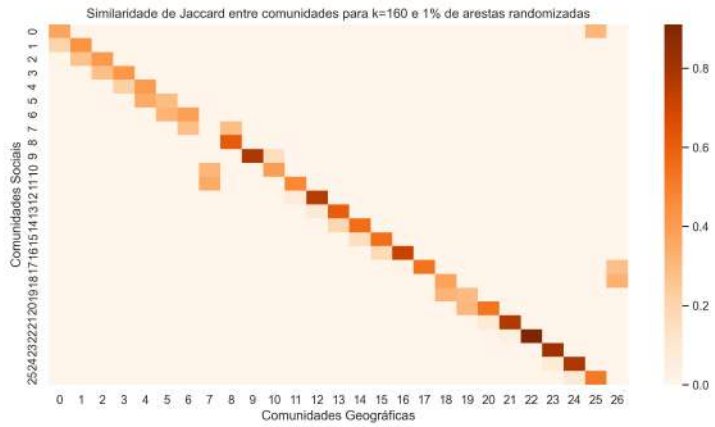
Figura 54: Matrizes de Jaccard para $k = 160$ (Parte 2).



(a) $p = 0.004$

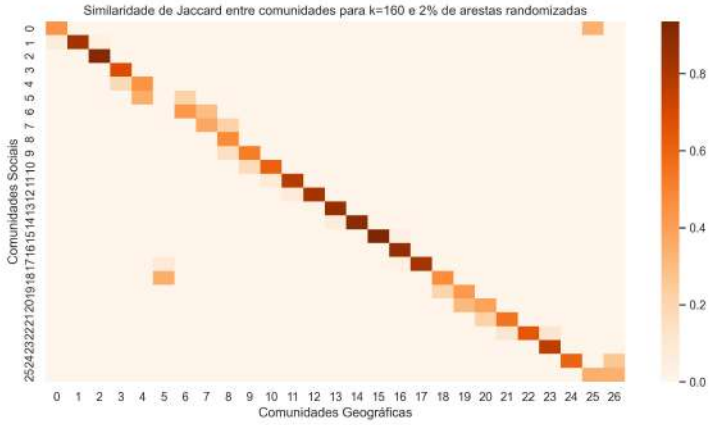


(b) $p = 0.008$

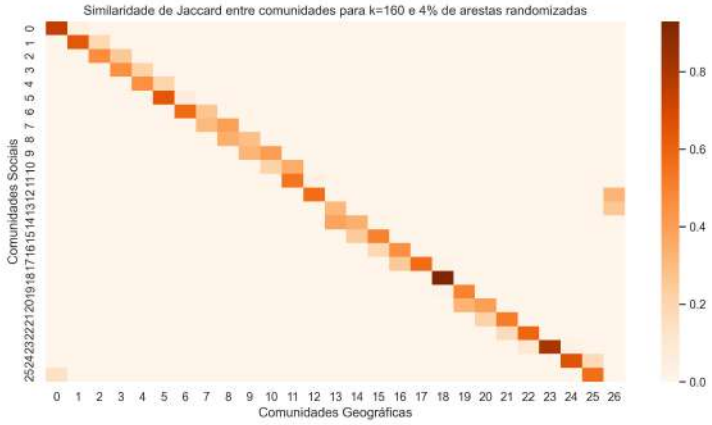


(c) $p = 0.01$

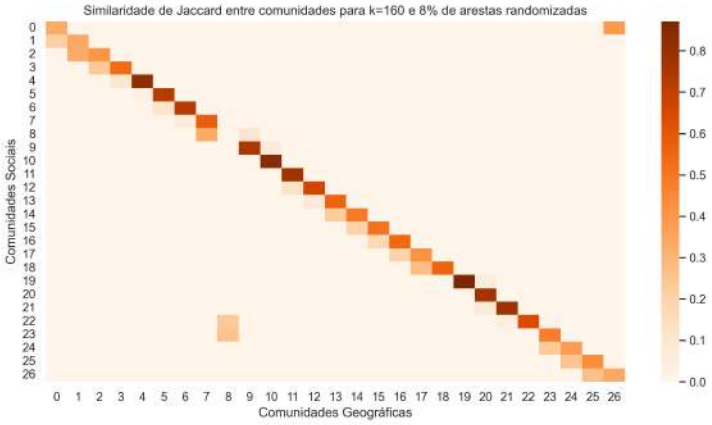
Figura 55: Matrizes de Jaccard para $k = 160$ (Parte 3).



(a) $p = 0.02$

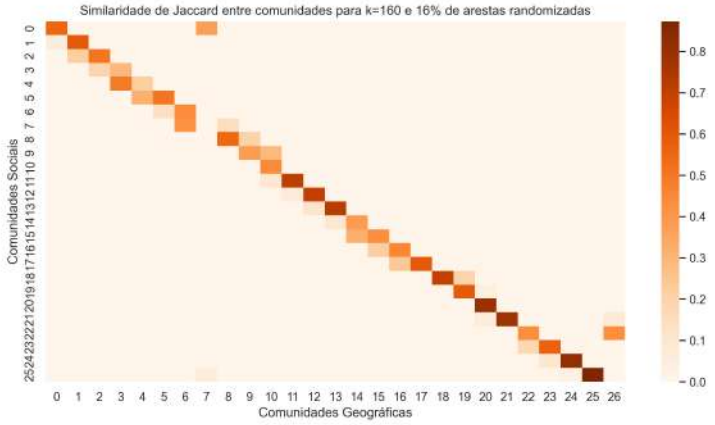


(b) $p = 0.04$



(c) $p = 0.08$

Figura 56: Matrizes de Jaccard para $k = 160$ (Parte 4).



(a) $p = 0.16$

Figura 57: Matrizes de Jaccard para $k = 160$ (Parte 5).