Análise de Traços de Personalidade em Mídias Sociais por meio de Redes Complexas

Matheus Henrique dos Santos

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Matheus Henrique dos Santos

Análise de Traços de Personalidade em Mídias Sociais por meio de Redes Complexas

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Profª Drª Fabíola Souza Fernandes Pereira
Coorientador: Profª Drª Elaine Ribeiro de Faria Paiva

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S237 Santos, Matheus Henrique dos, 1994-
2025 Análise de Traços de Personalidade em Mídias Sociais por meio de Redes Complexas [recurso eletrônico] / Matheus Henrique dos Santos. - 2025.

Orientadora: Fabíola Souza Fernandes Pereira.
Coorientadora: Elaine Ribeiro de Faria Paiva.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2025.548
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Pereira, Fabíola Souza Fernandes,1987-, (Orient.). II. Paiva, Elaine Ribeiro de Faria,1980-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada "**Análise de Traços de Personalidade em Mídias Sociais por meio de Redes Complexas**" por **Matheus Henrique dos Santos** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação.**

Uberlândia, 28 de agosto de 2025

Orientador: _____
Profa. Dra. Fabíola Souza Fernandes Pereira
Universidade Federal de Uberlândia

Coorientador: _____
Profa. Dra. Elaine Ribeiro de Faria Paiva
Universidade Federal de Uberlândia

Banca Examinadora:

_____
Prof. Dr. Claudio Douglas Gouveia Linhares
Universidade de Linnaeus

_____
Profa. Dra. Maria Camila Nardini Barioni
Universidade Federal de Uberlândia

Este trabalho é dedicado à memória de Renato Alves.

Agradecimentos

Agradeço primeiramente a Deus, por ter me permitido vivenciar essa grande experiência do mestrado, que sempre foi um dos meus sonhos de vida. Por ter me amparado nos momentos de dificuldade e dúvida, e por me dar forças para chegar até o fim.

Agradeço também à minha orientadora, Fabíola, e à coorientadora, Elaine. Muito obrigado por todo o apoio ao longo deste período e, principalmente, por ouvirem minhas angústias e não permitirem que eu desistisse dessa jornada. Vocês foram — e são — incríveis.

Agradeço à minha família e aos amigos, por estarem ao meu lado desde o início dessa trajetória, compreendendo meus momentos de ausência e sempre me incentivando a seguir em frente. Sem vocês, com toda certeza, eu não teria chegado até aqui.

Em especial, gostaria de agradecer ao Tiago Kerr, um grande amigo que embarcou comigo nessa jornada e, em muitos momentos, me apoiou e me ajudou a encontrar os melhores caminhos a seguir.

Por fim, agradeço a todos que, direta ou indiretamente, me deram forças e foram suporte para a conclusão do mestrado e a realização desse grande sonho.

"Compreender o outro é o início de toda transformação."
(Carl Rogers)

Resumo

O crescimento das mídias sociais e o volume de dados gerados por seus usuários têm despertado o interesse dos pesquisadores da área de ciência de dados para compreender padrões comportamentais e traços psicológicos manifestados nesses ambientes. Grande parte dos estudos existentes concentra-se no uso de modelos tradicionais de classificação com o objetivo de prever o temperamento dos usuários. No entanto, tais abordagens apresentam uma limitação relevante: não capturam de forma satisfatória as relações e interações entre os usuários, nem as dependências entre os atributos extraídos. O emprego de redes complexas busca superar essa restrição. Este trabalho propõe analisar e correlacionar os temperamentos/traços de personalidade de usuários de mídias sociais a partir de dados obtidos nessas plataformas, por meio da utilização de redes complexas. Diferentemente das abordagens tradicionais de classificação, a proposta visa a modelagem de dados textuais e comportamentais extraídos da plataforma *X* (antigo *Twitter*) em estruturas de rede, permitindo a detecção de comunidades e a análise de clusters de usuários com características semelhantes. Foram utilizadas duas bases de dados contendo publicações de usuários rotulados com tipos de personalidade, passando por etapas de pré-processamento textual, extração de atributos psicológicos via LIWC, discretização, normalização e construção de redes com base em medidas de similaridade. A análise dos clusters mais puros revelou padrões distintos entre as dimensões, especialmente Extroversão/Introversão (E/I) e Intuição/Sensação (N/S), sugerindo que determinados traços compartilham características linguísticas e sociais observáveis nas redes. Os resultados sustentam a hipótese de que as redes complexas são capazes de identificar padrões estruturais associados à personalidade que sejam complementares às análises estatísticas tradicionais, contribuindo para novos caminhos de pesquisa.

Palavras-chave: Redes Complexas. MBTI. Mídias Sociais. Personalidade. LIWC.

Abstract

The growth of social media and the volume of data generated by its users have sparked the interest of data science researchers in understanding behavioral patterns and psychological traits manifested in these environments. Much of the existing research focuses on using traditional classification models to predict user temperaments. However, these approaches have a significant limitation: they do not adequately capture the relationships and interactions between users, nor the dependencies between the extracted attributes. The use of complex networks seeks to overcome this limitation. This work proposes to analyze and correlate the temperaments/personality traits of social media users based on data obtained from these platforms, using complex networks. Unlike traditional classification approaches, the proposal aims to model textual and behavioral data extracted from the X platform (formerly Twitter) into network structures, enabling the detection of communities and the analysis of clusters of users with similar characteristics. Two databases containing posts from users labeled with personality types were used, undergoing stages of textual preprocessing, extraction of psychological attributes via LIWC, discretization, normalization, and network construction based on similarity measures. Analysis of the purest clusters revealed distinct patterns among the dimensions, especially Extraversion/Introversion (E/I) and Intuition/Sensation (N/S), suggesting that certain traits share linguistic and social characteristics observable in networks. The results support the hypothesis that complex networks are capable of identifying structural patterns associated with personality that complement traditional statistical analyses, contributing to new avenues of research.

Keywords: Complex Networks. MBTI. Social Media. Personality. LIWC.

Lista de ilustrações

Lista de tabelas

Lista de siglas

LIWC *Linguistic Inquiry Word Count*

MBTI *Myers-Briggs Type Indicator*

Sumário

Capítulo 1

Introdução

Desde o início do século XXI, o reconhecimento do valor e da relevância das mídias sociais cresceu exponencialmente, impulsionado pelos avanços da internet e pela capacidade de conectar pessoas para além das barreiras geográficas convencionais (CROMITY, 2012). Essas transformações ampliaram significativamente o volume de interações e a disseminação de informações, despertando o interesse de pesquisadores em compreender padrões de comportamento, preferências e tendências que emergem nessas plataformas (JUE; MARR; KASSOTAKIS, 2010; LIMA; CASTRO, 2014).

As mídias sociais tornaram-se ambientes privilegiados para análises psicológicas, uma vez que o conteúdo publicado pelos usuários frequentemente reflete seus interesses, opiniões e estados emocionais. Estudos anteriores apontam que essas interações podem revelar características importantes da personalidade e do temperamento (PLANK; HOVY, 2015; LIMA; CASTRO, 2019), contribuindo para a compreensão da tomada de decisões e do processamento de informações em diferentes contextos, incluindo saúde, negócios e psicologia (PLANK; HOVY, 2015). Nesse sentido, a análise de traços de personalidade e temperamentos apresenta relevância não apenas científica, mas também prática, possibilitando aplicações em múltiplas áreas.

Grande parte dos trabalhos existentes na literatura busca classificar usuários em categorias de temperamento ou traços de personalidade utilizando métodos tradicionais de aprendizado de máquina, com base em atributos textuais e comportamentais (PLANK; HOVY, 2015; CLARO, 2018; LIMA; CASTRO, 2019; ALMEIDA; GOYA, 2023). Embora esses métodos apresentem resultados importantes, em geral eles analisam atributos de forma isolada, o que pode limitar a identificação de padrões complexos e a compreensão das interações entre variáveis (MARTINS, 2022; ALCANTARA et al., 2024). Além disso, os resultados numéricos oriundos dessas abordagens nem sempre fornecem representações intuitivas e de fácil interpretação.

Nesse contexto, as redes complexas surgem como uma alternativa promissora, pois permitem modelar não apenas atributos individuais, mas também as interações entre eles, possibilitando a identificação de padrões emergentes em estruturas de dados comple-

xos (CARNEIRO, 2017; BARABáSI; PóSFAI, 2016). O uso de métricas topológicas e de técnicas de detecção de comunidades oferece uma perspectiva complementar às análises tradicionais, permitindo explorar relações locais e globais nos dados de mídias sociais. Trabalhos recentes reforçam esse potencial ao mostrar a aplicabilidade das redes complexas na descoberta de estruturas e correlações em diferentes domínios (GUL et al., 2022; UCHIYAMA, 2022; AQUINO; STROELE; SOUZA, 2020).

Dessa forma, a presente pesquisa busca avançar além das abordagens tradicionais, empregando métodos de análise de redes complexas para investigar traços de personalidade e temperamentos em mídias sociais. O diferencial deste estudo está em explorar não apenas a previsão ou classificação dos usuários, mas também a identificação de padrões estruturais e perfis emergentes em comunidades, ampliando as possibilidades de análise em dados sociais. Do ponto de vista da Computação, a contribuição reside na aplicação de técnicas de modelagem em grafos a dados textuais e comportamentais, reforçando o papel das redes complexas como ferramenta computacional robusta para problemas de mineração de dados, aprendizado de máquina e descoberta de padrões.

1.1 Objetivos

Diante do desafio citado anteriormente, o principal objetivo do trabalho é analisar e correlacionar os temperamentos/traços de personalidade de usuários de mídias sociais a partir de dados obtidos nessas plataformas, através da utilização de redes complexas. Para atingir o objetivo geral desta pesquisa, foram traçados os seguintes objetivos específicos:

- ❏ Definir a forma mais eficiente de representação de dados tabulares em redes complexas dentro do contexto específico desse trabalho considerando a modularidade e a pureza dos clusters;

- ❏ Construir e analisar redes complexas que permitam analisar dados comportamentais e textuais, estabelecendo relações entre os usuários a partir das características extraídas das mídias sociais.

1.2 Hipótese

Este trabalho possui a seguinte hipótese:

- ❏ As redes complexas permitem identificar padrões e correlações entre temperamento/traços de personalidade e informações textuais ou comportamentais extraídas de mídias sociais, que vão além daquelas captadas apenas por análises estatísticas convencionais.

O avanço será mensurado a partir de métricas de redes (como modularidade e pureza de clusters), bem como pela comparação direta entre os achados estatísticos e os padrões estruturais obtidos nas redes.

1.3 Contribuições

As principais contribuições esperadas para este trabalho são:

- ❑ Aplicação experimental de redes complexas como ferramenta para investigar padrões associados aos traços de personalidade e temperamento de usuários de mídias sociais;
- ❑ Apresentação de abordagens para transformar dados tabulares em redes complexas, ampliando o repertório metodológico da área de Computação;
- ❑ Contribuição para o entendimento da relação entre as características das mídias sociais e o temperamento/traços de personalidade dos usuários;
- ❑ Geração de descobertas sobre a relação entre atributos comportamentais/linguísticos e as dimensões do *Myers-Briggs Type Indicator* (MBTI) (MYERS, 1985), com potencial aplicação em áreas como psicologia, marketing, educação e recursos humanos;
- ❑ Análise de pontos positivos e limitações da interseção entre dados de mídias sociais e características individuais dos usuários, oferecendo insumos para pesquisas futuras.

1.4 Organização da Dissertação

Este trabalho está organizado da maneira a seguir. O Capítulo 2 apresenta os principais conceitos que serão utilizados no decorrer do trabalho. O Capítulo 3 discorre os trabalhos do estado da arte relacionados ao trabalho descrito aqui. O Capítulo 4 aborda o desenvolvimento do trabalho. O Capítulo 5 mostra os experimentos e a análise dos resultados obtidos. Por fim, o Capítulo 6 tece as principais conclusões deste presente trabalho e os principais trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os conceitos fundamentais que sustentam o desenvolvimento deste trabalho, organizados em três seções. A Seção 2.1 aborda a teoria dos temperamentos, enquanto a Seção 2.2 explora os conceitos relacionados às redes complexas. Finalmente, a Seção 2.3 trata da utilização de dados de redes sociais, incluindo sua estrutura, processo de pré-processamento e outros aspectos relevantes.

2.1 Teoria dos Temperamentos

Há um interesse duradouro entre os estudiosos pelas questões que influenciam o comportamento humano, ao longo dos séculos. Igualmente, as noções de temperamento, personalidade e caráter têm sido frequentemente discutidas desde tempos antigos (VOLPI, 2004). A ideia de que os seres humanos sejam agrupados em classes ou tipos tem estado presente desde a antiguidade entre os pesquisadores da personalidade humana (PASQUALI, 2000).

A investigação das disparidades entre as pessoas tem sempre atraído a atenção de teóricos, pesquisadores e leigos. Variadas dimensões, traços ou atributos são reconhecidos em indivíduos conforme a abordagem teórica e o interesse envolvido (ITO; GUZZO, 2002).

O médico e filósofo Galeno, responsável pelo desenvolvimento da primeira tipologia de temperamento baseado na teoria de Hipócrates, estabeleceu e explicou quatro temperamentos primários, que são amplamente reconhecidos tanto por especialistas quanto por pessoas comuns, e são denominados de acordo com os humores predominantes no corpo (AIKEN, 2009): 1) tipo sanguíneo, caracterizado por indivíduos atléticos e enérgicos, nos quais o humor corporal predominante é o sangue; 2) tipo colérico, pessoas facilmente irritáveis, cuja bile amarela predomina; 3) tipo melancólico, indivíduos tristes e deprimidos que mostram um excesso de bile negra; e 4) tipo fleumático, pessoas cronicamente fatigadas e com movimentos lentos, que têm um excesso de fleuma.

O psiquiatra suíço Carl Gustav Jung propôs um modelo que ficou conhecido como modelo de Jung (JUNG, 1923). Esse modelo é uma teoria psicológica que descreve os tipos

psicológicos com base em diferentes formas de funcionamento cognitivo e preferências de personalidade. Jung propôs que as pessoas têm diferentes maneiras de perceber o mundo e tomar decisões, e categorizou essas diferenças em quatro dimensões principais: extroversão/introversão, sensação/intuição, pensamento/sentimento e julgamento/percepção.

O MBTI é um instrumento elaborado por Isabel Myers e sua mãe, Katherine Briggs, que reorganizaram as teorias de temperamento propostas por Jung para torná-las mais acessíveis ao público em geral (VALLANDER, 2023). Segundo (PASQUALI, 2000), o MBTI se apresenta numa série de formas diferentes, e em uma delas ele mede até 16 tipos que resultam da combinação de 4 polaridades.

As quatro polaridades propostas pelo modelo MBTI são: extroversão (E) vs. introversão (I), que indicam a orientação predominante da energia, seja para o mundo exterior — pessoas, atividades e objetos — ou para o mundo interior, relacionado a ideias, emoções e impressões; sensação (S) vs. intuição (N), que refletem a preferência por adquirir informações de forma concreta, por meio dos sentidos, ou de maneira abstrata, por meio de percepções inconscientes, conforme proposto por (JUNG, 1923); pensamento (T) vs. sentimento (F), que dizem respeito à tendência de organizar e estruturar informações com base em critérios lógicos ou em valores pessoais, influenciando a tomada de decisão; e, por fim, julgamento (J) vs. percepção (P), que se referem à preferência por um estilo de vida planejado e estruturado, em contraste com uma abordagem mais flexível e espontânea.

Combinando estas quatro polaridades, geram-se os 16 tipos de temperamento.

Com base na teoria de Hipócrates,o autor (KEIRSEY, 1998) propôs um modelo de temperamentos associado ao MBTI, classificando os indivíduos em quatro categorias: Guardiães (SJ), correspondentes ao tipo melancólico descrito por Hipócrates, que se concentram no dever, comércio e economia, preservando tradições e hierarquia, representando aproximadamente 45% da população; Artesões (SP), relacionados ao tipo sanguíneo, voltados para a arte, a estética e a liberdade, caracterizados pela busca da espontaneidade e da diversão, correspondendo a cerca de 35% da população; Racionais (NT), equivalentes ao colérico, que se orientam pela ciência e pela teoria, prezando pela competência, liderança e objetividade, abrangendo aproximadamente 5% da população; e, por fim, Idealistas (NF), vinculados ao tipo fleumático, com foco no espiritual e no ético, que buscam a paz, a harmonia e a autorrealização, representando cerca de 5% da população (PASQUALI, 2000).

(AKISKAL; AKISKAL, 2005) propuseram uma abordagem mais contemporânea dos temperamentos afetivos, que se baseia nos estudos de (KRAEPELIN, 1921) e (KRETSCHMER, 1936). Esse conceito foi desenvolvido a partir de considerações teóricas e observações clínicas, e foi formulado como um instrumento chamado Avaliação do Temperamento de Memphis, Pisa, Paris e San Diego (TEMPS). Inicialmente, foi apresentado como um instrumento semiestruturado (TEMPS-I, versão de entrevista) para avaliar os temperamentos depressivo, ciclotímico, hipertímico e irritável. Mais tarde, foi criada a versão do

autoquestionário (TEMPS-A), que requer respostas simples de "sim"ou "não"e consiste em 110 itens (109 para homens), incluindo a adição de um temperamento ansioso (RIHMER et al., 2010).

Segundo (AKISKAL; AKISKAL, 2005), os temperamentos podem ser classificados em cinco tipos. O hipertímico é caracterizado por energia, otimismo e autoconfiança, atributos que favorecem liderança e exploração (WOODRUFF et al., 2011). O depressivo envolve sensibilidade ao sofrimento, perseverança e confiabilidade, embora com dificuldades em liderança e relações interpessoais (WOODRUFF et al., 2011; DEMBIŃSKA-KRAJEWSKA; RYBAKOWSKI, 2014). O ansioso reflete predisposição excessiva à preocupação, associada à sobrevivência em contextos de parentesco (WOODRUFF et al., 2011). O ciclotímico apresenta instabilidade de humor, mudanças rápidas de energia e autoestima, além de impacto nas relações sociais (WOODRUFF et al., 2011; DEMBIŃSKA-KRAJEWSKA; RYBAKOWSKI, 2014). Por fim, o irritável compartilha aspectos do ciclotímico, mas com maior energia e menor empatia, sendo associado a ceticismo, crítica e comportamento mais sombrio (DEMBIŃSKA-KRAJEWSKA; RYBAKOWSKI, 2014).

(WOODRUFF et al., 2011) criaram uma versão resumida da tradução do TEMPS-A para o português. A tradução recebeu o nome de TEMPS-RIO. O TEMPS-RIO é composto por 45 questões de verdadeiro ou falso, sendo relacionadas aos cinco temperamentos: ciclotímico, irritável, hipertímico, depressivo e ansioso, mais o temperamento preocupado (MARTINS, 2022).

Algumas outras classificações foram criadas ao longo dos anos por outros pesquisadores, mas optou-se por manter na fundamentação teórica apenas os que são pertinentes aos conjuntos de dados que serão utilizados para desenvolvimento do trabalho.

2.2 Redes Complexas

A disciplina da ciência de redes surgiu com o propósito de compreender e representar sistemas do mundo real que podem ser conceptualizados como redes, ou seja, sistemas compostos por componentes individuais interconectados por relações muitas vezes complexas. As redes permeiam diversas áreas, desde a internet e o sistema de transporte até as mídias sociais e a propagação de doenças. Ao longo das últimas décadas, diversos modelos surgiram para representar sistemas reais e explorar sua natureza por meio de métricas de redes (MERENDA, 2023).

Uma rede é um grafo composto por um conjunto de vértices (ou nós) interligados por um conjunto de arestas (ou arcos), estabelecendo relações entre os vértices de acordo com o problema em questão (BARABáSI; PóSFAI, 2016). Além disso, o grafo pode ser direcionado ou não. Em um grafo direcionado (dígrafo), cada aresta possui uma direção, conectando um vértice de origem a um vértice de destino. Exemplos de dígrafos incluem representações de chamadas telefônicas e fluxo de mensagens de e-mails, nos quais as

mensagens são enviadas de um indivíduo para outro (METZ et al., 2007).

É relevante ressaltar que nem todo grafo pode ser classificado como uma rede complexa, já que essa categorização só é aplicável se o grafo apresentar certas propriedades topológicas específicas não encontradas em grafos simples. Algumas dessas propriedades são brevemente descritas a seguir (METZ et al., 2007).

É possível identificar na literatura diversas medidas e modelos (topologias) utilizadas para caracterizar a estrutura das redes complexas. Normalmente essas medidas são usadas para analisar características e propriedades estatísticas que descrevem o comportamento do sistema em rede e também sua estrutura, enquanto a criação de modelos de rede geralmente se relaciona com o entendimento do significado dessas propriedades (CARNEIRO, 2017).

2.2.1 Modelos de Redes Complexas

Segundo (CARNEIRO, 2017), o principal propósito dos modelos de redes é prever a evolução de redes do mundo real, como a internet, redes biológicas e mídias sociais. Esses modelos de redes possibilitam a criação de redes com propriedades controláveis, como grau e coeficiente de aglomeração, entre outras. Portanto, o emprego desses modelos simplifica a investigação sobre como a estrutura da rede influencia sua dinâmica, sendo particularmente útil para compreender processos como a propagação de doenças e notícias falsas. Nesta seção, serão apresentados alguns dos modelos de redes mais importantes.

Os principais modelos de redes complexas são: Redes Regulares, Redes Aleatórias, Redes de Pequeno-Mundo, Redes Livres de Escala e Redes Geométricas. A seguir, apresenta-se uma breve definição de cada modelo (Figuras 1, 2 e 3).

1. **Redes Regulares:** representam o paradigma mais elementar das redes, nas quais todos os nós possuem o mesmo número de conexões, Figura 1. Em geral, apresentam alta densidade de ligações e grande diâmetro (MERENDA, 2023).

2. **Redes Aleatórias:** estabelecem conexões entre vértices de forma aleatória, a partir de parâmetros definidos (CARNEIRO, 2017), Figura 1. São analisadas em conjuntos, a fim de identificar propriedades típicas, e podem ser geradas, por exemplo, pelo modelo de Erdõs-Rényi, no qual cada par de vértices é conectado com probabilidade $p \in [0, 1]$.

3. **Redes de Pequeno-Mundo:** combinam alto coeficiente de aglomeração e baixo diâmetro, sendo uma alternativa ao padrão aleatório (LOPES, 2011; MERENDA, 2023), Figura 1. Baseiam-se em uma rede regular na qual arestas são reconectadas com probabilidade p, criando "atalhos" que reduzem a distância média entre vértices, fenômeno relacionado aos "seis graus de separação".

Figura 1 – Representação de uma Rede Regular, Rede de Pequeno-Mundo e Rede Aleatória.

(Fonte: Elaborada pelo autor)

4. **Redes Livres de Escala:** caracterizam-se pela *conexão preferencial*, em que novos vértices tendem a se ligar a vértices de maior grau (CARNEIRO, 2017), Figura 2. Diferente dos modelos aleatórios ou de pequeno-mundo, seguem uma distribuição de graus em lei de potência, como no modelo Barabási-Albert (BA), reproduzindo melhor propriedades observadas em sistemas reais (MERENDA, 2023).

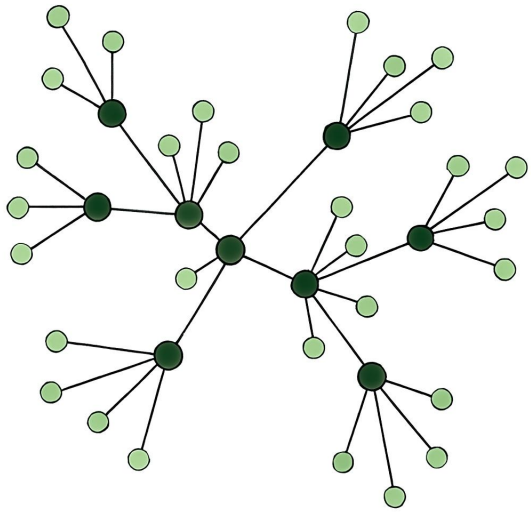


Figura 2 – Rede Livre de Escala obtida a partir do modelo Barabási-Albert.

(Fonte: Adaptado de (CARNEIRO, 2017))

5. **Redes Geométricas:** também chamadas de espaciais ou geográficas, atribuem uma posição espacial definida a cada nó (COMIN; F, 2018). São comuns em sistemas reais como redes de cidades, estradas ou aeroportos, pois, além da topologia, consideram informações geométricas como posição e distância Euclidiana. Diversos

algoritmos podem ser utilizados para sua construção, sendo três dos mais representativos ilustrados na Figura 3.

O primeiro é o *Grafo Geométrico Aleatório*, em que vértices são distribuídos em um espaço métrico e conectados de acordo com a proximidade espacial. O segundo é o modelo de *Voronoi*, no qual o espaço é particionado em células, e as conexões surgem a partir da adjacência dessas regiões, refletindo, por exemplo, a organização de cidades ou regiões de influência. Já o modelo de *Waxman* estabelece conexões com base em uma função decrescente da distância, o que permite simular redes onde ligações longas são possíveis, mas menos prováveis — como ocorre em cabos de telecomunicações ou rotas aéreas (MERENDA, 2023).

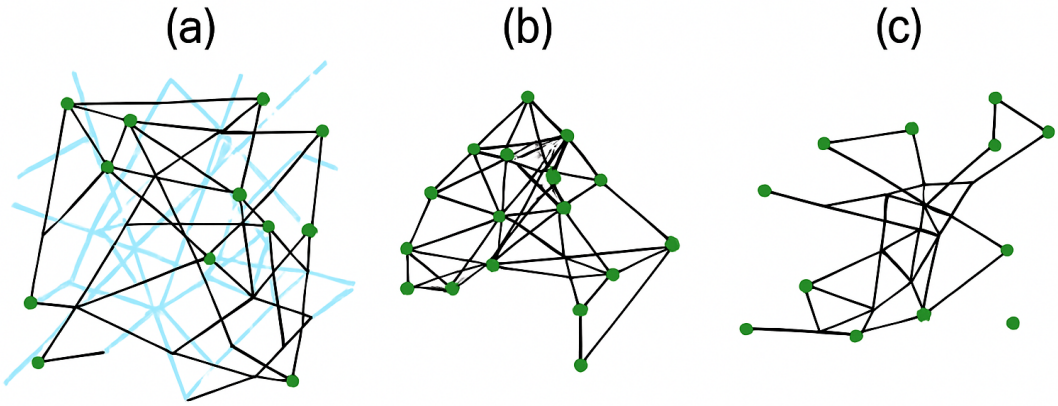


Figura 3 – Três modelos de Redes Geométricas: (a) Rede de Voronoi, incluindo as células de Voronoi. (b) Rede Geométrica Aleatória. (c) Rede de Waxman.

(Fonte: Adaptado de (MERENDA, 2023))

2.3 Medidas de Redes Complexas

No contexto do mundo real, é frequente se deparar com redes e desejamos investigar algumas de suas características para obter uma compreensão mais profunda do que estamos lidando. Isso se torna especialmente relevante quando se lida com grandes redes que exibem estruturas aparentemente complexas e imprevisíveis.

A análise de redes representa um campo de pesquisa que se apoia em diversas ferramentas e métodos matemáticos. Ela é considerada uma subárea da mineração

de grafos e abrange uma variedade de métricas usadas em diversas disciplinas para investigar redes. Algumas métricas se baseiam nos graus dos vértices, enquanto outras adotam uma abordagem mais detalhada com estatísticas de distância. Um conceito crucial empregado para descrever muitas redes do mundo real é o agrupamento, além das métricas relacionadas à centralidade e importância, especialmente relevantes para as mídias sociais (STEEN, 2010).

A seguir são listadas algumas medidas obtidas a partir de redes complexas:

a) **Grau dos vértices/da rede:** Conforme apresentado por (CARNEIRO, 2017), se a rede for uma rede não direcionada, o grau de casa vértice será dado pelo número total de vértices adjacentes a ele, ou seja, pelo número total de arestas. Caso a rede for direcionada, o grau do vértice será dado pela soma do total de vértices de entrada e de saída. A partir desses conceitos, é possível calcular o grau médio da rede, que será dado pela média aritmética do grau de todos os vértices.

b) **Coeficiente de Agrupamento:** Segundo (METZ et al., 2007), os agrupamentos intrínsecos às redes são quantificados por meio do coeficiente de aglomeração, também conhecido como fenômeno de transitividade. Esse fenômeno ocorre quando um vértice A está conectado a um vértice B, e o vértice B está conectado a um vértice C, aumentando as chances de o vértice A também estar conectado ao vértice C. Em outras palavras, a transitividade indica a presença de um número elevado de triângulos na rede, isto é, conjuntos de três vértices conectados uns aos outros. Para compreender melhor, pode-se considerar a analogia com uma rede social, onde se A é amigo de B e B é amigo de C, existem grandes chances de A e C também serem amigos.

c) **Assortatividade:** A assortatividade é uma medida que avalia a propensão dos vértices a se conectarem com aqueles que são, de alguma forma, semelhantes. Por exemplo, a assortatividade do grau indica a tendência dos vértices se ligarem a outros vértices com grau semelhante (MERENDA, 2023). Inicialmente, o coeficiente de assortatividade r representa a correlação de Pearson dos graus entre pares de nós conectados. Esse coeficiente varia de -1 a 1, onde valores positivos indicam que pares de vértices diretamente conectados têm maior probabilidade de se comportar de maneira semelhante, enquanto valores negativos indicam maior probabilidade de comportamentos distintos entre os vértices conectados (CARNEIRO, 2017).

d) **Medidas de Distância:** Segundo (CARNEIRO, 2017), algumas medidas de distância da rede são frequentemente utilizadas na literatura, como, por exemplo:

- ❏ Distância geodésica: A distância entre dois vértices i e j é o número de arestas do menor caminho entre eles;
- ❏ Excentricidade: A maior distância de um vértice i para qualquer outro vértice no grafo é denominada excentricidade;
- ❏ Diâmetro: O diâmetro de um grafo é a excentricidade máxima de qualquer vértice do grafo;
- ❏ Raio: Ao contrário do diâmetro, o raio de um grafo representa a excentricidade mínima de qualquer vértice do grafo.

e) **Medidas de Centralidade:** Segundo (MERENDA, 2023), a centralidade avalia a importância de um vértice na propagação de informações na rede. Para ilustrar esse conceito, este autor apresentou como exemplo o aeroporto de Congonhas em São Paulo. Comparado a um aeroporto em uma cidade de porte médio, Congonhas tem um volume de passageiros e uma variedade de rotas aéreas muito maiores. Uma falha em Congonhas poderia causar transtornos significativos, afetando dezenas de outros aeroportos e centenas de rotas em um efeito cascata. Em uma rede de comunicação, por exemplo, a relevância de um vértice pode ser estabelecida pelo número de caminhos mais curtos dos quais ele faz parte, já que isso pode indicar sua carga de trabalho no processamento e encaminhamento de mensagens (STEEN, 2010). Portanto, a mensuração da centralidade dos vértices é essencial para compreender a estrutura e o funcionamento da rede. Existem várias maneiras de medir a centralidade, algumas das mais conhecidas são: *degree centrality*, *closeness centrality*, *betweenness centrality* e *eigenvector centrality*.

f) **Page Rank:** O PageRank, medida introduzida por Brin e Page, é uma medida de centralidade baseada em passeios aleatórios em um grafo. Ele representa um modelo de caminhada aleatória, onde um agente move-se pelo grafo selecionando aleatoriamente conexões de saída em cada nó. Além disso, o agente pode saltar aleatoriamente para qualquer outro nó do grafo, independente das conexões existentes, com uma certa probabilidade. Isso evita problemas de convergência. O valor do PageRank de um nó é a probabilidade de o agente estar nesse nó, refletindo sua importância com base no número e na relevância das conexões de entrada (CARNEIRO, 2017).

2.4 Comunidades em Redes Complexas

A ciência contemporânea das redes, também conhecida como *network science*, tem desempenhado um papel fundamental na análise das redes complexas, permitindo compreender sua estrutura, propriedades e aplicações em diferentes áreas do conhecimento. Uma das características notáveis e distintivas das redes complexas é a

existência de comunidades. A ideia por trás das comunidades é simples: cada comunidade pode ser entendida como um conjunto de vértices altamente conectados entre si dentro de um subgrafo, enquanto mantém poucas conexões com o restante da rede (SILVA, 2012). A Figura 4 ilustra esse conceito, representando uma rede dividida em diferentes grupos de vértices (comunidades) que apresentam forte coesão interna e fraca conexão externa.

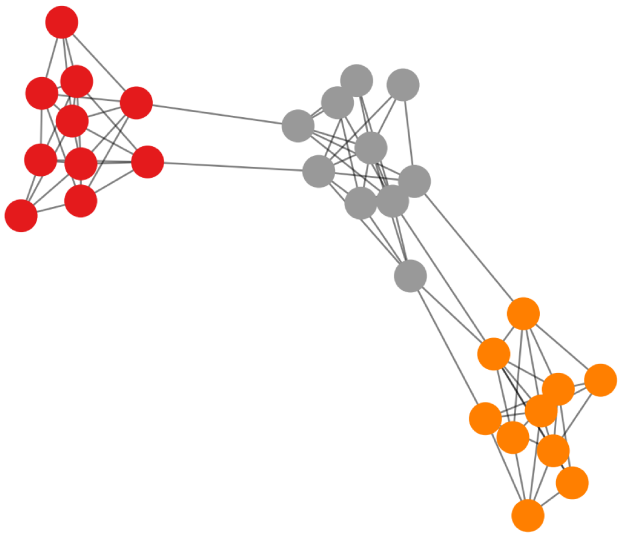


Figura 4 – Exemplo didático de uma rede com comunidades, destacando vértices mais densamente conectados internamente e com poucas conexões externas.

(Fonte: Elaborada pelo autor)

Foram propostas várias abordagens distintas para detectar comunidades em redes complexas (FORTUNATO, 2010). Uma categoria de algoritmos amplamente empregados se baseia na otimização de uma métrica chamada modularidade (SILVA, 2012)).

Outro conceito essencial nesse campo de estudo é a detecção de vértices sobrepostos (FORTUNATO, 2010). Esses vértices são definidos como pertencentes a mais de uma comunidade ou categoria simultaneamente. Por exemplo, em uma rede social, uma pessoa naturalmente faz parte da empresa em que trabalha e também do grupo que representa sua família. Nesse contexto, a identificação de vértices e comunidades sobrepostas é crucial para a análise de dados em geral.

2.4.1 Métricas para Avaliação de Comunidades em Redes Complexas

Diversos autores apresentam métricas estatísticas frequentemente utilizadas para avaliar a qualidade das comunidades identificadas em redes complexas (LINHA-

RES et al., 2020; YIN et al., 2015; YANG; ALGESHEIMER; TESSONE, 2016; NEWMAN, 2016). Embora a modularidade seja amplamente reconhecida, outros indicadores, como precisão, *recall* e *F-Measure*, também se mostram eficientes nesse contexto (YIN et al., 2015). A seguir, apresentamos os conceitos dessas métricas comumente empregadas.

A *precisão* avalia a proporção de comunidades corretamente detectadas em relação ao total de comunidades identificadas, variando de 0 (pior) a 1 (melhor). Sua fórmula é expressa por:

$$Precisão(R,T) = \frac{1}{p}\sum_{i=1}^{p}\max_{j}\left(\frac{|Ri \cap Tj|}{|Ri|}\right) \tag{1}$$

onde $T = \{T1, T2, T3, ..., Tq\}$ representa o conjunto de comunidades reais na rede (como, por exemplo, uma turma de estudantes ou pessoas que vivem na mesma cidade), e $R = \{R1, R2, R3, ..., Rp\}$ é o conjunto de comunidades identificadas por um algoritmo de *clustering*. Apesar de sua utilidade, a precisão não é uma métrica definitiva para avaliar a detecção de comunidades de maneira isolada. Por exemplo, ao considerar cada nó como uma comunidade separada, a precisão alcançaria seu valor máximo (YIN et al., 2015).

O *recall* é uma métrica que também varia de 0 (pior) a 1 (melhor) e indica a proporção de comunidades corretamente identificadas em relação ao número total de comunidades reais $|Tj|$. Sua fórmula é dada por:

$$Recall(R,T) = \frac{1}{q}\sum_{j=1}^{q}\max_{i}\left(\frac{|Ri \cap Tj|}{|Tj|}\right) \tag{2}$$

Contudo, assim como a precisão, o *recall* não é uma métrica perfeita para avaliar a detecção de comunidades de forma independente. Por exemplo, ao agrupar todos os nós em uma única comunidade, o *recall* atingirá seu valor máximo (YIN et al., 2015).

A *F-Measure* é uma métrica que combina as métricas de precisão e *recall*, proporcionando um equilíbrio entre ambas. Seu valor também varia de 0 (pior) a 1 (melhor) e é calculado como a média harmônica dessas duas medidas. Quando as comunidades detectadas se assemelham às reais, a F-Measure tende a se aproximar de 1 (YIN et al., 2015).

$$F - Measure(R,T) = \frac{2 \cdot Precision(R,T) \cdot Recall(R,T)}{Precision(R,T) + Recall(R,T)} \tag{3}$$

Embora a F-Measure exija a presença de um conjunto de referência (comunidades reais), a *modularidade* pode ser calculada para qualquer rede, mesmo na ausência

de um conjunto de referência (LINHARES et al., 2020). O valor da modularidade varia de -1 (pior) a 1 (melhor) e é utilizado para medir a qualidade ou a força de uma determinada divisão da rede.

$$Modularidade = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{4}$$

onde m representa o número total de arestas na rede, A_{ij} é o valor da matriz de adjacência que indica a presença de uma aresta entre os nós i e j ($A_{ij} = 1$ se existir aresta, $A_{ij} = 0$ caso contrário), e k_i é o grau do nó i (isto é, o número de arestas conectadas a i). O símbolo δ representa o delta de Kronecker, que assume o valor 1 quando os nós i e j pertencem à mesma comunidade, e 0 caso contrário. Valores de modularidade entre 0,3 e 0,7 geralmente indicam uma boa segmentação ou uma estrutura comunitária forte (NEWMAN, 2016).

2.5 Representação de Dados Tabulares em Redes Complexas

A conversão de dados tabulares em redes complexas é uma etapa fundamental na análise de dados estruturados, permitindo a aplicação de técnicas de teoria de grafos para entender interações e relações entre entidades. Dados tabulares, como aqueles provenientes de planilhas ou bancos de dados relacionais, geralmente contêm registros e atributos que precisam ser transformados em uma estrutura de rede para análise em termos de nós e arestas. As redes complexas oferecem uma maneira poderosa de modelar esses dados e identificar padrões de interação entre as entidades (BARABáSI; PóSFAI, 2016).

Em um conjunto de dados tabulares, as linhas representam entidades e as colunas contêm atributos dessas entidades. Para transformar esses dados em uma rede, é necessário definir uma maneira de estabelecer conexões entre as entidades com base em seus atributos. As representações de redes podem ser obtidas por diferentes abordagens, como a construção de grafos ponderados ou direcionados, dependendo das interações e dos dados envolvidos (UCHIYAMA, 2022).

Uma abordagem comum para essa transformação é a construção de uma matriz de adjacência. Cada linha e coluna dessa matriz representa um nó na rede, e os valores nas células indicam a presença ou o peso de uma conexão entre os nós. Por exemplo, em um conjunto de dados de redes sociais, onde as linhas representam usuários e as colunas representam interações (como curtidas ou comentários), pode-se estabelecer uma conexão entre dois nós (usuários) quando houver uma interação entre eles, como um comentário ou uma menção. Esse tipo de abordagem permite

representar de maneira explícita as interações entre os dados (TABASSUM et al., 2023).

Outra técnica de conversão envolve o uso de medidas de similaridade ou distância entre as entidades. Por exemplo, se temos um conjunto de dados tabular com características numéricas de usuários (como idade, localização e número de postagens), podemos calcular a similaridade entre esses usuários (usando, por exemplo, a similaridade do cosseno ou a distância Euclidiana) e, com isso, estabelecer arestas entre eles com base em seus valores de similaridade. Essa abordagem é comumente utilizada em problemas de *clustering* e agrupamento de usuários em redes sociais, onde a similaridade entre os registros define as conexões entre os nós da rede (AQUINO; STROELE; SOUZA, 2020; UCHIYAMA, 2022).

Além disso, em alguns casos, as relações entre as entidades podem ser definidas com base em uma regra de negócios ou um critério específico. Por exemplo, em um banco de dados de transações financeiras, pode-se representar uma rede onde os nós são as contas bancárias e as arestas representam transações entre elas. Esse tipo de representação é fundamental para detectar padrões de comportamento financeiro ou de fraude, utilizando redes complexas. A análise de redes de transações financeiras pode revelar padrões de comportamento, como a identificação de crimes de lavagem de dinheiro (MOLLER, 2024).

Em resumo, a representação de dados tabulares em redes complexas é um passo crucial para a análise e visualização de interações e relações em conjuntos de dados estruturados. Dependendo do contexto e dos objetivos da análise, diferentes abordagens podem ser utilizadas para estabelecer as conexões entre as entidades, seja por meio de similaridade, distância ou regras de negócio específicas. Ao adotar a transformação de dados tabulares em redes, é possível aplicar uma variedade de algoritmos de análise de redes, como a detecção de comunidades e análise de centralidade, para entender melhor as interações e os padrões subjacentes dos dados.

2.6 Dados de Redes Sociais

Os dados extraídos de redes sociais têm sido amplamente utilizados em pesquisas para entender comportamentos, sentimentos e interações entre indivíduos. Essas plataformas geram um volume imenso de dados não estruturados, que podem incluir textos, imagens, vídeos, interações sociais (curtidas, comentários, compartilhamentos), além de informações demográficas e de rede (seguidores, amigos, grupos) (LIMA; CASTRO, 2014). No entanto, trabalhar com esses dados apresenta diversos desafios devido à sua complexidade e à natureza não estruturada da informação (PIRES, 2024).

Os dados textuais, por exemplo, frequentemente precisam ser pré-processados para tornarem-se utilizáveis em modelos analíticos. O pré-processamento de dados textuais em redes sociais envolve uma série de etapas, incluindo a remoção de *stop words*, a normalização de texto (como a conversão de palavras para minúsculas e a remoção de pontuação), e a tokenização, que separa o texto em unidades menores, como palavras ou frases (SAIF; HE; ALANI, 2012). Essa etapa é crucial, pois os dados brutos coletados de plataformas como *X* e Facebook frequentemente contêm muito "ruído", ou seja, informações irrelevantes ou redundantes, que podem prejudicar as análises subsequentes (PIRES, 2024).

Além disso, uma das principais dificuldades no pré-processamento de dados de redes sociais é a diversidade linguística. Em plataformas globais como o *X*, usuários de diferentes origens e com diferentes níveis de proficiência linguística podem produzir textos que variam consideravelmente em estilo, vocabulário e gramática (SAIF; HE; ALANI, 2012). A variação de linguagem, incluindo gírias, abreviações e erros gramaticais, pode dificultar a tarefa de análise automática e exigir abordagens especializadas para lidar com esses casos.

Outro desafio significativo é a presença de dados desbalanceados, especialmente em tarefas de classificação como a análise de sentimentos ou a previsão de tendências. Por exemplo, em um conjunto de dados de sentimentos extraídos do *X*, pode haver uma quantidade muito maior de postagens neutras em comparação com postagens claramente positivas ou negativas. Isso pode fazer com que os modelos de aprendizado de máquina apresentem desempenho inferior ao tentar prever categorias minoritárias (KAUR; PANNU; MALHI, 2019). Técnicas de balanceamento de dados, como o *undersampling* e o *oversampling*, são frequentemente utilizadas para mitigar esse problema, mas ainda assim exigem cuidado para evitar a perda de informações ou o *overfitting* (BANERJEE et al., 2020).

Além das dificuldades com dados textuais, redes sociais também geram dados de rede (*network data*), que descrevem as interações entre os usuários. Essas interações podem ser representadas por grafos, onde os nós são os usuários e as arestas representam relações como amizade, seguimento ou interação. O pré-processamento de dados de rede envolve a construção de grafos de interações, com a remoção de nós e arestas irrelevantes, além da normalização de métricas de centralidade e conectividade (NEWMAN, 2016). A análise desses grafos é essencial para entender a estrutura das redes sociais e as comunidades formadas dentro delas.

Em resumo, os dados de redes sociais oferecem grandes oportunidades para análises, mas também impõem desafios significativos no que diz respeito ao pré-processamento. A remoção de ruídos, o tratamento da diversidade linguística e a correção de desbalanceamento de dados são alguns dos obstáculos enfrentados pelos pesquisadores

(SAIF; HE; ALANI, 2012). No entanto, com as técnicas de pré-processamento adequadas, esses dados podem ser transformados em informações valiosas para a compreensão de comportamentos humanos e padrões de interação social.

2.7 Considerações Finais

Este capítulo apresentou a fundamentação teórica necessária para o desenvolvimento do trabalho, contemplando conceitos relacionados à análise de traços de personalidade e temperamentos, bem como os modelos psicológicos que os descrevem. Foram discutidas também as bases conceituais de redes complexas e sua aplicação em diferentes contextos, além das técnicas de análise textual empregadas em mídias sociais. Essa revisão forneceu os elementos essenciais para sustentar as escolhas metodológicas adotadas, servindo de suporte para as análises realizadas ao longo da pesquisa. O próximo capítulo apresenta os principais trabalhos relacionados utilizados como referência para desenvolvimento da pesquisa.

Capítulo 3

Trabalhos Relacionados

Nesta seção, será apresentada uma revisão dos principais trabalhos existentes na literatura que estão relacionados a este projeto, que visa utilizar redes complexas para analisar a relação entre personalidade e temperamento a partir de dados obtidos de mídias sociais. Além disso, será destacada a contribuição do mesmo em relação a esses trabalhos existentes.

É possível identificar diversos estudos que buscam conhecimento por meio de mídias sociais com diferentes focos, como pesquisas de opinião ou mesmo questões políticas ((TABASSUM et al., 2018; ZAFARANI; ABBASI; LIU, 2014)). No entanto, poucos estudos buscaram compreender a relação entre mídias sociais e temperamento e/ou traços de personalidade. Os trabalhos existentes sobre o tema concentram-se mais na classificação de um usuário em um determinado temperamento, enquanto a análise de padrões e perfis dentro de cada temperamento ou traço de personalidade é escassa ((MARTINS, 2022; ALCANTARA et al., 2024)). Além disso, até onde sabemos, não foram encontrados na literatura estudos que apliquem redes complexas a este tópico.

3.1 Trabalhos sobre Temperamento do Usuário

Alguns estudos exploram o traço de personalidade e o temperamento dos usuários em mídias sociais, buscando compreender como esses aspectos influenciam o comportamento online. A seguir, serão apresentados os principais trabalhos que investigam essa relação, destacando suas abordagens, descobertas e contribuições.

Os autores (PLANK; HOVY, 2015) conduziram um estudo com o objetivo de comparar a distribuição dos traços de personalidade do sistema Myers-Briggs entre os usuários do X e a população em geral, além de desenvolver modelos preditivos para classificar cada dimensão desse sistema. Eles monitoraram mensagens que mencionavam algum dos 16 tipos de traços associados à palavra Briggs, resultando em um

corpus de 1,2 milhão de tweets de 1.500 usuários diferentes. As mensagens foram estruturadas usando n-gramas e informações como gênero, contagem de tweets e número de seguidores. Utilizando regressão logística, os resultados mostraram que a distinção entre I - E (INTROVERTIDO - EXTROVERTIDO) e F - T (SENTIR - PENSAR) pode ser bem modelada, enquanto as outras duas dimensões apresentaram desafios. As características linguísticas predominam no poder preditivo dos modelos, mas as informações adicionais contribuem significativamente. Embora a distribuição dos tipos de personalidade de Myers-Briggs no *X* seja diferente da população em geral, reflete o uso real da linguagem e possui amostras estatisticamente significativas.

A autora (CLARO, 2018) utilizou o framework TECLA para identificar o temperamento dos usuários no *X* de acordo com o modelo de temperamento de Keirsey. O estudo empregou o dataset TwiSty-PT.json, que contém uma base de dados da literatura chamada *Twisty*, disponibilizada pelo *A Computational Linguists & Psycholinguistics Research Center* (CLiPS) (VERHOEVEN; DAELEMANS; PLANK, 2016). Este conjunto de dados inclui atributos textuais, atributos relacionados aos usuários e resultados do MBTI. Foram aplicados algoritmos de classificação baseados em distância (*K-Nearest Neighbors*, KNN), classificação baseada em árvore (Florestas Aleatórias, *RandomF*) e classificadores baseados em função (*Support Vector Machine*, SVM). Os resultados mostraram que os temperamentos artesão e guardião treinados com SVM obtiveram as melhores acurácias, seguidos pelo algoritmo *Random Forest*, com desempenho próximo ao do SVM. Para a classificação binária, as maiores acurácias médias foram para os temperamentos artesão e guardião, também com destaque para o algoritmo SVM, enquanto as menores acurácias médias foram observadas para o temperamento idealista.

Os autores (LIMA; CASTRO, 2019) propuseram um framework para prever temperamento e tipos psicológicos a partir de uma análise linguística e comportamental de dados do *X*, utilizando o modelo de David Keirsey e o MBTI. Os dados foram os mesmos propostos por (PLANK; HOVY, 2015), com 1,2 milhão de tweets de 1500 usuários classificados de acordo com o tipo psicológico de Myers-Briggs. Para avaliar o modelo Keirsey, cada tipo de MBTI foi mapeado, recebendo uma das quatro classificações: Artesão, Guardião, Idealista ou Racional. Diversos algoritmos foram utilizados, incluindo *AdaBoost*, *Bagging*, J48, *Naive Bayes*, *Random Forest* e SVM. Os resultados indicaram que Florestas Aleatórias com a técnica LIWC podem prever com alta precisão os temperamentos Artesão, Guardião, Idealista e Racional, além de resultados mais satisfatórios para o MBTI utilizando também *Random Forests*.

Em seu trabalho, (ALMEIDA; GOYA, 2023) buscaram construir um conjunto de dados com tweets em português rotulados com traços de personalidade dominantes e avaliar seu potencial em modelos clássicos de aprendizado de máquina. A coleta

de dados começou com a divulgação do questionário Big Five nas mídias sociais, seguida pela seleção e limpeza dos perfis e a obtenção das últimas 60 interações de cada usuário no *X*. Utilizando algoritmos como *Naive Bayes*, SVM, Árvore de decisão, KNN, Regressão Logística e *Random Forest*, eles avaliaram os dados com e sem balanceamento, aplicando a técnica SMOTE. Os dados textuais foram convertidos em numéricos usando TF-IDF e *stemming*, com considerações para unigram e bigram. Observou-se que os algoritmos tiveram dificuldade em diferenciar personalidades sem o balanceamento, enquanto o balanceamento teve um impacto positivo no desempenho dos modelos. A Regressão Logística, usando TF-IDF com unigram, obteve os melhores resultados, com 0,97 de AUC, 0,78 de MCC (Coeficiente de Correlação de Matthews) e 0,76 de F1-score.

Diferentemente dos demais trabalhos citados anteriormente, que objetivam criar modelos para classificar o temperamento/traço de personalidade dos usuários de mídias sociais, (MARTINS, 2022) propôs um método computacional distinto, que visa analisar as postagens e comportamentos dos usuários do Instagram em contraste com seus temperamentos. Utilizando uma base de dados própria composta por voluntários discentes da Universidade Federal de Uberlândia, o estudo investigou a relação entre o temperamento do usuário (determinado pelo questionário TEMPS-RIO) e atributos da mídia social, como quantidade de curtidas, número de publicações, entre outros. Três abordagens foram adotadas para verificar essas relações: testes estatísticos, análise de gráficos *boxplot* e utilização de uma árvore de decisão como modelo descritivo. Os resultados indicam que usuários com temperamento depressivo tendem a compartilhar mais legendas com sentimentos positivos em comparação com usuários hipertímicos, irritados e preocupados. Além disso, os padrões de publicação durante a semana e nos fins de semana são semelhantes entre os temperamentos irritáveis e preocupados, bem como entre os usuários ciclotímicos e depressivos. Por fim, os usuários ansiosos tendem a usar mais emojis nas legendas do Instagram do que os usuários deprimidos e irritados.

O trabalho de (ALCANTARA et al., 2024) propôs um método para contrastar dados das rede sociais online X e Instagram com a percepcão de suporte social, medida pelo questionario ESSS, e temperamento, medido pelo questionário TEMPS-RIO. O trabalho utilizou bases próprias criadas com dados de voluntários da UFU, coletando informações do X, Instagram e questionários sobre temperamento e suporte social. A base continha dados como seguidores, seguidos, curtidas, postagens e textos das postagens. O autor propôs um método não supervisionado baseado em léxico (LeXPAPC) para classificar a polaridade de postagens e realizou uma análise contrastando os comportamentos nas redes sociais X e Instagram com os resultados dos questionários. Comparado ao Vader e SentiStrength, o LeXPAPC teve resultados superiores, embora sua abrangência tenha sido inferior, pois os métodos compara-

dos classificam todas as entradas. A análise revelou indícios de uma relação entre o comportamento dos usuários nas plataformas, seus temperamentos e percepção de suporte social.

Com exceção do trabalho de (MARTINS, 2022) e (ALCANTARA et al., 2024), que objetivaram correlacionar os dados de mídias sociais com o temperamento dos usuários, todos os demais trabalhos utilizaram bases de dados de mídias sociais para buscar classificar o temperamento/traço de personalidade dos usuários utilizando modelos tradicionais de classificação, limitando-se à utilização de atributos comportamentais dos usuários nas mídias sociais e atributos relacionados aos textos publicados pelos mesmos em suas redes. Porém, nenhum dos trabalhos listados utilizaram-se de redes complexas para realização das suas análises e criação dos modelos, sendo este o principal diferencial deste projeto de pesquisa.

Apesar da vasta literatura que emprega técnicas de aprendizado de máquina, especialmente modelos de classificação, para prever ou identificar traços de personalidade em mídias sociais, tais abordagens costumam se restringir a processos de categorização individual dos usuários. Nesse contexto, surge um desafio relevante: esses métodos não exploram de forma adequada as relações e interações entre os usuários ou entre os atributos extraídos. O uso de redes complexas busca superar essa limitação, pois permite não apenas analisar variáveis isoladas, mas também identificar padrões emergentes, correlações e estruturas comunitárias que não seriam facilmente detectados por classificadores tradicionais. Dessa forma, a aplicação de redes complexas neste trabalho representa tanto um avanço metodológico quanto uma contribuição significativa para ampliar a compreensão dos vínculos entre temperamento, traços de personalidade e comportamento em mídias sociais.

A Tabela 1 apresenta, de forma resumida, os trabalhos relacionados a temperamento do usuário que foram utilizados como referência para este trabalho.

Tabela 1 – Resumo dos trabalhos relacionados a temperamento do usuário em redes sociais.

Referência	**Objetivo**	**Base de Dados**	**Idioma**	**Tipo de Temperamento**	**Tarefa de Aprendizado**	**Algoritmos Utilizados**
(PLANK; HOVY, 2015)	Comparar a distribuição do MBTI entre usuários do X e a população geral, além de criar modelos preditivos para classificar suas dimensões.	Dados coletados do X	Inglês	MBTI	Classificação	Reg. Logist.
(CLARO, 2018)	Identificar o temperamento dos usuários no X aplicando o framework TECLA.	Dataset TwiSty-PT	Português	Keirsey	Classificação	KNN Random Forest SVM
(LIMA; CASTRO, 2019)	Propor um framework para prever temperamento e tipos psicológicos a partir de uma análise linguística e comportamental de dados do X.	Dados coletados por (PLANK; HOVY, 2015)	Inglês	MBTI e Keirsey	Classificação	AdaBoost Bagging J48 Naive Bayes Random Forest SVM
(MARTINS, 2022)	Propor um método que visa analisar as postagens e comportamentos dos usuários do Instagram em contraste com seus temperamentos.	Dados coletados do X e Instagram.	Português	TEMPS-RIO	Classificação	Testes Estatísticos Boxplot Árvore de Decisão
(ALCANTARA et al., 2024)	Propor um método para contrastar dados do X e Instagram com a percepcão de suporte social e temperamento.	Dados coletados do X e Instagram.	Português	TEMPS-RIO	Classificação	Testes Estatísticos LeXPAPC
(ALMEIDA; GOYA, 2023)	Construir um conjunto de dados com tweets em português rotulados com traços de personalidade e avaliar seu potencial em modelos de classificação.	Dados coletados do X	Português	Big Five	Classificação	Naive Bayes SVM Árvore de decisão KNN Reg. Logist. Random Forest

3.2 Trabalhos sobre Redes Complexas

O uso de redes complexas tem se consolidado como uma abordagem eficaz para modelar e compreender sistemas complexos do mundo real (CARNEIRO, 2017). Diversos trabalhos exploram essa ferramenta em diferentes domínios, investigando desde interações sociais e políticas até sistemas de transporte e contextos culturais. Em comum, essas pesquisas buscam identificar padrões estruturais, detectar comunidades e extrair informações relevantes a partir da topologia da rede.

No campo político, por exemplo, (BRITO; SILVA; AMANCIO, 2020) analisam a evolução do sistema brasileiro a partir da rede de votação da Câmara dos Deputados, revelando padrões de coalizão e isolamento partidário. Em outra vertente, voltada para redes sociais e educacionais, (LINHARES et al., 2020) propuseram uma metodologia de análise estatística e visual para avaliação de comunidades em diferentes interações sociais, enquanto (AQUINO; STROELE; SOUZA, 2020) investigaram o engajamento estudantil por meio da construção de redes a partir de dados educacionais. Esses trabalhos evidenciam o potencial das redes complexas em capturar dinâmicas sociais e comportamentais.

Outras contribuições concentram-se no aprimoramento de algoritmos e técnicas de análise. (GUL et al., 2022), por exemplo, compararam diferentes métodos de detecção de comunidades em redes de múltiplos domínios, destacando a eficácia do algoritmo multiescala. Já (LINHARES et al., 2020) ressaltaram a importância de associar métricas numéricas à visualização interativa, reforçando que a interpretação dos resultados depende de múltiplas perspectivas.

Em contextos mais específicos, redes complexas têm sido aplicadas para compreender fenômenos culturais e esportivos. (UCHIYAMA, 2022) analisaram a estrutura de livros por meio de redes de similaridade entre parágrafos, na tentativa de prever o sucesso editorial, enquanto (FÉLIX et al., 2019) exploraram a rede global de transferências de jogadores de futebol, evidenciando o papel diferenciado de países desenvolvidos e em desenvolvimento.

Apesar da diversidade de aplicações, nota-se uma limitação importante: nenhum dos trabalhos revisados dedica-se à análise de temperamentos ou traços de personalidade em mídias sociais. Essa lacuna destaca a relevância do presente estudo, que busca aplicar conceitos e métodos da ciência das redes a um novo campo de investigação, contribuindo para o avanço da literatura.

A tabela 2 apresenta, de forma resumida, os trabalhos relacionados a redes complexas que foram utilizados como referência para este trabalho.

Tabela 2 – Resumo dos trabalhos relacionados a redes complexas.

Referência	**Objetivo**	**Tipo de Base de Dados**	**Construção da Rede**	**Medidas Analisadas**
(BRITO; SILVA; AMANCIO, 2020)	Propor um framework baseado em redes complexas para analisar a evolução do sistema político brasileiro.	Conexões entre indivíduos	Nós: Deputados Arestas: Similaridade de voto	Distância topológica e medidas de dissimilaridade
(LINHARES et al., 2020)	Apresentar uma metodologia destinada a realizar uma análise estatística e visual de estruturas comunitárias em redes.	4 bases de conexões entre indivíduos	Nós: Indivíduos Arestas: Interação entre eles	Modularidade, Precisão, Recall e F-Measure
(GUL et al., 2022)	Identificar o algoritmo mais eficaz para detectar comunidades em redes complexas do mundo real de diversos tamanhos e domínios.	9 bases de conexões entre indivíduos	Nós: Indivíduos Arestas: Interação entre eles	Modularidade
(AQUINO; STROELE; SOUZA, 2020)	Aplicar mineração de dados em redes complexas para analisar o engajamento comportamental dos alunos, gerando indicadores para auxiliar gestores na tomada de decisões.	Dados tabulares	Nós: Alunos Arestas: Similaridade coseno entre eles	Grau dos nós
(UCHIYAMA, 2022)	Identificar padrões capazes de classificar um livro de sucesso pela abordagem de redes complexas.	Conjunto de livros (dados textuais)	Nó: Parágrafo do livro Aresta: Similaridade Coseno entre eles	Diâmetro, excentricidade, betweenness, clustering, closeness, índice de correlação e assinatura de recorrência
(FÉLIX et al., 2019)	Utilizar técnicas de redes complexas e análises sociais para avaliação de uma rede de transferências de jogadores entre países.	Dados tabulares	Nós: País Arestas: Vendas de jogadores entre eles	Densidade, Reciprocidade, Assortatividade, Grau máximo de entrada e saída, Diâmetro

3.3 Considerações Finais

Este capítulo apresentou uma revisão dos principais trabalhos relacionados ao estudo de temperamentos e traços de personalidade em mídias sociais, bem como pesquisas que empregam redes complexas em diferentes domínios. Verificou-se que, embora existam avanços significativos no uso de algoritmos de aprendizado de máquina para classificação de traços de personalidade e temperamentos, essas abordagens permanecem limitadas à categorização individual de usuários. Por outro lado, as aplicações de redes complexas têm demonstrado grande potencial em diferentes áreas, mas ainda não foram exploradas para analisar esse tipo de fenômeno psicológico. Nesse contexto, o presente trabalho busca preencher essa lacuna, propondo a aplicação de redes complexas na investigação da relação entre personalidade, temperamento e comportamento em mídias sociais. O próximo capítulo apresenta o método adotado para atingir esse objetivo.

Capítulo 4

Método para extrair padrões em diferentes tipos de traços de personalidade das mídias sociais

Nesta seção, são apresentados os procedimentos metodológicos adotados para a análise dos dados. A pesquisa foi estruturada em etapas que incluem a seleção e pré-processamento dos dados, transformação dos atributos, construção das redes complexas e análise estatística. O objetivo foi investigar a relação entre os traços de personalidade (MBTI) e o comportamento dos usuários nas redes sociais, utilizando técnicas de análise de redes complexas e testes estatísticos para validar os achados.

4.1 Visão Geral do Método

O objetivo deste trabalho é investigar as relações entre os temperamentos/traços de personalidade e o comportamento dos usuários nas redes sociais, utilizando redes complexas para representar dados tabulares extraídos dessas plataformas. Especificamente, a pesquisa busca analisar e comparar diferentes abordagens de representação de dados tabulares em redes, construir redes complexas para modelar dados comportamentais e textuais, e realizar análises visuais e estatísticas dos clusters gerados. A Figura 5 ilustra o processo metodológico adotado, destacando as principais etapas que envolvem a seleção de dados, pré-processamento, construção das redes e análise dos resultados. As etapas detalhadas de cada fase serão apresentadas nas seções seguintes.

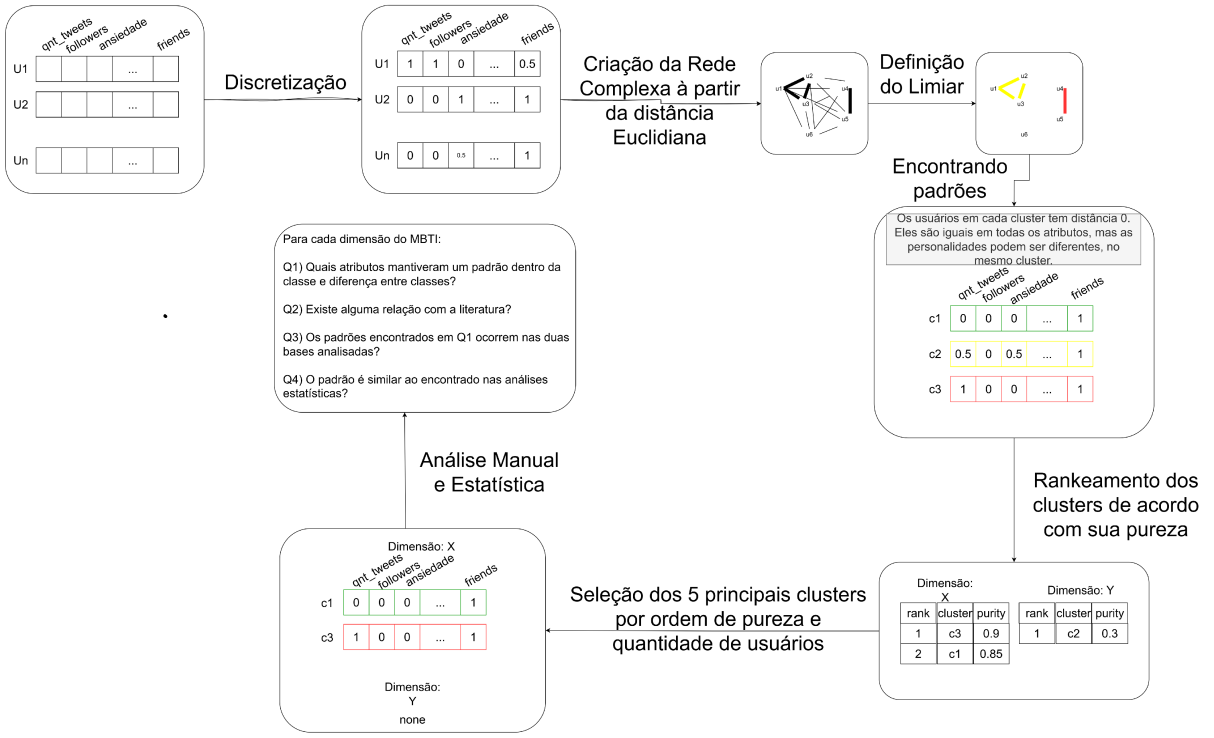


Figura 5 – Visão geral das etapas do trabalho.

(Fonte: Elaborada pelo autor)

4.2 Seleção da Base de Dados

Para aplicar o método proposto, é necessário utilizar dados de mídias sociais, contendo conteúdo textual gerado pelos usuários, bem como atributos sociais, como número de seguidores, engajamento e outras métricas de interação. Idealmente, os dados também devem incluir atributos demográficos (como gênero) e informações relacionadas ao temperamento ou traços de personalidade dos usuários. Nesta primeira etapa, foram selecionadas bases de dados que atendessem a esses requisitos para a aplicação do método. Em uma das bases selecionadas para este trabalho, todos os textos eram do idioma inglês, enquanto a segunda possuía textos em idiomas variados.

É importante destacar que há uma escassez de conjuntos de dados que combinem todas essas informações de forma completa e integrada, especialmente quando se trata do traço de personalidade MBTI, que foi utilizado neste trabalho. No entanto, mesmo que alguns desses atributos estejam ausentes, é possível adaptar o método às variáveis identificadas.

4.3 Pré-processamento

As etapas de pré-processamento realizadas para tratar a base de dados foram:

- ❏ Limpeza e tratamento de textos;
- ❏ Extração e Normalização das características do LIWC.

A primeira etapa foi a limpeza e o tratamento dos textos. Palavras irrelevantes, pontuação, números e caracteres especiais foram removidos para garantir que apenas o texto relevante fosse mantido. Além disso, os textos foram padronizados para letras minúsculas e espaços extras foram eliminados. Esse processo é essencial para garantir que a extração de características pelo LIWC ocorra de forma eficiente.

Em seguida, foi utilizado o método *Linguistic Inquiry Word Count* (LIWC), proposto por (TAUSCZIK; PENNEBAKER, 2010), que é um método de análise de texto que conta palavras em categorias psicologicamente significativas. O modelo original do LIWC cria 73 novos atributos textuais em relação ao texto analisado. Porém, para este trabalho optou-se por manter apenas os 9 atributos psicológicos relacionados aos processos sociais e afetivos, descritos a seguir:

- ❏ ***social***: Número de termos relacionados a processos sociais.
- ❏ ***family***: Número de termos relacionados a família.
- ❏ ***friend***: Número de termos relacionados a amizade.
- ❏ ***affect***: Número de termos relacionados a processos afetivos.
- ❏ ***posemo***: Número de termos relacionados a sentimentos positivos.
- ❏ ***negemo***: Número de termos relacionados a sentimentos negativos.
- ❏ ***anx***: Número de termos relacionados a ansiedade.
- ❏ ***anger***: Número de termos relacionados a raiva.
- ❏ ***sad***: Número de termos relacionados a tristeza.

Como cada usuário tinha um número diferente de textos e também tamanhos diferentes, optou-se por analisar a proporção de termos psicológicos analisados, e não a quantidade absoluta. Dessa forma, foi possível comparar de forma mais adequada quais usuários eram mais semelhantes em relação a essas características. Para isso, as características psicológicas foram ajustadas calculando o número de termos de cada processo dividido pelo número de palavras nos textos daquele usuário.

4.4 Transformação dos Atributos

Após o pré-processamento, os atributos foram transformados para o intervalo [0,1], evitando viés devido às diferentes escalas das variáveis na distância Euclidiana que serão utilizadas durante a criação de redes complexas (ver Seção 4.5). Isso possibilitou o uso de algoritmos baseados na distância Euclidiana. O redimensionamento para o intervalo [0,1] foi feito pela normalização dos atributos através da re-escala dos mesmos. No entanto, utilizando essa transformação, ao calcular a distância, as características binárias assumem maior peso que as demais. O atributo gênero, presente apenas no primeiro conjunto de dados, teve maior peso que as demais.

Isso ocorre porque, para a maioria dos atributos, os valores estão concentrados em um intervalo, e não dispersos entre 0 e 1. Para mitigar esse efeito, três abordagens foram testadas para discretizar as características em categorias menores:

- **Transformação 1:** Cada atributo foi transformado em uma variável classificada em apenas três categorias, recebendo os valores 0, 0.5 ou 1 (com exceção da variável sexo que já era uma variável binária). A partir do menor e maior valor dentro daquele atributo, foram calculados os limites a serem considerados para cada categoria. Para o cálculo dos limites foi utilizada a fórmula apresentada a seguir:

  $\text{limite}_1 = MIN + (MAX - MIN)/3 \;\; \text{limite}_2 = MIN + 2 * (MAX - MIN)/3$

  Sendo MAX o maior valor no atributo e MIN o menor valor no atributo. Caso o valor observado esteja entre MIN e o $limite_1$, ele recebe o valor 0. Caso esteja entre $limite_1$ e $limite_2$ ele recebe o valor 0.5. Caso seja maior que $limite_2$ ele recebe o valor 1.

- **Transformação 2:** Cada atributo foi transformado em uma variável classificada em apenas cinco categorias, recebendo os valores 0, 0.25, 0.5, 0.75 ou 1 (com exceção da variável sexo que já era uma variável binária). A partir do menor e maior valor dentro daquele atributo, foram calculados os limites a serem considerados para cada categoria. Para o cálculo dos limites foi utilizada a fórmula apresentada a seguir:

  $\text{limite}_1 = MIN + (MAX - MIN)/5 \;\; \text{limite}_2 = MIN + 2 * (MAX - MIN)/5$
  $\text{limite}_3 = MIN + 3 * (MAX - MIN)/5 \;\; \text{limite}_4 = MIN + 4 * (MAX - MIN)/5$

  Sendo MAX o maior valor no atributo e MIN o menor valor no atributo. Caso o valor observado esteja entre MIN e o $limite_1$, ele recebe o valor 0. Caso esteja entre $limite_1$ e $limite_2$ ele recebe o valor 0.25. Caso esteja entre $limite_2$ e $limite_3$ ele recebe o valor 0.5. Caso esteja entre $limite_3$ e $limite_4$ ele recebe o valor 0.75. Caso seja maior que $limite_4$ ele recebe o valor 1.

- ❑ **Transformação 3:** Cada atributo foi transformado em uma variável classificada em apenas três categorias, recebendo os valores 0, 0.5 ou 1 (com exceção da variável sexo que já era uma variável binária). Para essa classificação foi utilizado o KMens, com K = 3, em cada atributo. A determinação do tamanho de K foi feito a partir do método de cotovolo, para cada uma das variáveis.

Após a transformação dos dados, os conjuntos de dados finais utilizados para a criação das redes foram compostos pelos usuários e pelos atributos de cada um deles, além do traço de personalidade MBTI, que foi utilizado para a análise de cada uma de suas dimensões. Em seguida, foi calculada a matriz de Distância Euclidiana entre os usuários para cada uma das bases. O cálculo da matriz foi importante para a etapa de Criação da Rede, pois foi usada para criar as arestas das redes complexas.

4.5 Geração das redes complexas

Três redes complexas foram geradas com base nas diferentes estratégias de discretização aplicadas. Cada nó representa um usuário e as arestas não direcionadas são criadas quando a distância Euclidiana entre dois usuários é menor ou igual ao limite estabelecido. Durante a construção das redes, foram testados diversos valores de distância para definir esse limite, com o objetivo de identificar o parâmetro que gerasse os melhores resultados na formação dos clusters.

Várias abordagens foram avaliadas para o cálculo do limiar, incluindo a distância média entre os usuários, o valor médio subtraído de duas vezes o desvio padrão, e a definição de limites com base nos cinco vizinhos mais próximos. Após esses testes, o valor limite zero demonstrou ser o mais eficaz, resultando em redes com clusters mais bem definidos e usuários com características homogêneas dentro de cada grupo. A escolha da distância zero, embora tenha sido precedida de outras tentativas, provou ser o ajuste mais adequado para os dados analisados. Esse processo de experimentação foi crucial para otimizar a construção das redes complexas, mesmo que alguns testes iniciais não tenham apresentado os resultados esperados.

As redes complexas foram criadas como grafos não direcionados e visualizadas na ferramenta *Gephi*[1] (versão 0.10) (BASTIAN; HEYMANN; JACOMY, 2009). Para a construção e extração das medidas das redes, foi utilizada a biblioteca *IGraph*[2] do *Python*. No Gephi, o processo de criação da rede envolve a importação de duas bases de dados: uma contendo as relações de arestas entre os nós e outra com os atributos que serão vinculados a cada nó. Após a importação dos dados, a rede é gerada e pode ser segmentada por cores com base em atributos específicos. Além

[1] <https://gephi.org/>
[2] <https://igraph.org/>

disso, é possível ajustar o *layout* da rede para otimizar sua visualização, permitindo uma análise mais clara dos clusters formados. A configuração do Gephi envolve, principalmente, a escolha do algoritmo de *layout* para posicionamento dos nós, bem como a definição de parâmetros para a segmentação da rede e a análise de suas métricas.

4.6 Seleção da Melhor Rede

O objetivo é identificar qual rede segmenta melhor os traços de personalidade. Primeiro, calculou-se a modularidade (NEWMAN, 2018), uma métrica comum da estrutura de redes que mede a força da divisão de uma rede em módulos (BARABáSI; PóSFAI, 2016). Importante ressaltar que, como a modularidade é calculada após a criação das redes, os usuários que são desconsiderados para a construção das mesmas não são levados em conta no cálculo dessa medida.

Em seguida, a pureza p de cada cluster c foi calculada em relação a cada temperamento/traço de personalidade x. A pureza p de um cluster c é definida como:

$$p_x = \max Q_x/Q_c \tag{5}$$

Onde Q_x é o número de usuários do temperamento x dentro do cluster c, e Q_c é o número total de usuários no cluster c. Quanto mais usuários do mesmo temperamento o cluster tiver, mais puro ele será.

Adotou-se empiricamente que um cluster é puro se sua pureza for maior ou igual a 70% (para cada dicotomia). Após avaliar cada uma das três redes complexas, com base nas medidas de modularidade e pureza p (para cada dicotomia), foi selecionada a rede que apresentou um bom valor de modularidade e a maior proporção de clusters puros para cada uma das dicotomias das dimensões do MBTI.

4.7 Seleção e Análise dos Clusters

Da melhor rede, alguns clusters foram selecionados para uma análise de suas características e padrões. O objetivo dessa análise é interpretar os clusters e buscar identificar padrões entre os usuários desses grupos e as dimensões do MBTI atribuídas a eles.

Para cada rede gerada, os clusters foram classificados com base na pureza p e no número de usuários pertencentes a cada cluster, considerando cada par de dicotomias nas dimensões do MBTI. É relevante destacar que, dentro de cada cluster, os usuários apresentam uma distância Euclidiana de 0 entre si, ou seja, eles são

idênticos em todos os atributos discretizados da base de dados pré-processada. No entanto, a classe atribuída a cada usuário nas dicotomias de suas dimensões pode variar dentro do mesmo cluster. Por exemplo, em um único cluster, usuários podem apresentar diferentes valores para a dicotomia E/I, mesmo que seus valores em todos os outros atributos discretizados sejam idênticos. Em função disso, a classificação dos clusters deve considerar não apenas a homogeneidade em relação aos atributos, mas também a pureza das dicotomias, ou seja, onde a maioria dos usuários pertence a mesma classe na dicotomia analisada.

A última etapa consistiu em uma análise manual dos clusters gerados para cada dimensão do MBTI. O processo de análise seguiu os seguintes passos:

a) Ranqueamento dos clusters em cada classe de cada dimensão, considerando a quantidade de usuários e a pureza de cada cluster.

b) Seleção dos cinco clusters mais puros e populosos para cada categoria, como, por exemplo, os cinco clusters da classe E e os cinco da classe I na dimensão E/I.

c) Extração das categorias discretizadas de cada variável para os clusters selecionados.

d) Identificação de padrões significativos dentro dos clusters da mesma classe, considerando as variáveis, e diferenças visíveis entre as classes em cada dimensão.

O objetivo principal dessa análise foi identificar padrões e diferenças nas variáveis de cada dimensão do MBTI.

Para busca de padrões entre os usuários, os dados extraídos das variáveis para os clusters puros foram inseridos em uma tabela para facilitar a analise manual. A Figura 6 mostra um recorte de como ficou a estrutura da tabela.

Para cada variável em cada dimensão foi avaliado se todos os clusters possuiam o mesmo valor entre uma categoria e se esse valor divergia da outra categoria. Por exemplo, nessa Figura, todos os cluster da categoria E possuiam valor intermediário (0,5), enquanto todos da classe I possuíam valor baixo (0).

4.8 Análises Estatísticas

Este trabalho utilizou testes estatísticos em dois momentos distintos de avaliação, cada um com objetivos específicos:

				VARIÁVEIS "ORIGINAIS" DO TWITTER					VARIÁVEIS LIWC								
Dimensão	Classe	nº do Cluster	Usuários puros da classe	qtde_tweets	followers_count	statuses_count	favorites_count	listed_count	affect	posemo	negemo	anx	anger	sad	social	family	friend
E/I	I	60	6	0,5	0	0	0	0	0,5	0,5	0,5	0	0	0	0,5	0	0,5
E/I	I	80	13	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0,5	0	0	0
E/I	I	116	9	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0
E/I	I	205	6	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0	0,5	0	0
E/I	I	531	8	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0	0	0	0
E/I	E	73	9	0,5	0	0	0	0,5	0,5	0,5	0	0	0	0	0,5	0	0
E/I	E	140	10	0,5	0	0	0	0	1	1	0	0	0	0	0,5	0	0
E/I	E	190	8	0,5	0	0	0	0	0,5	1	0	0	0	0	0,5	0	0,5
E/I	E	275	14	0,5	0	0	0	0	0,5	1	0	0	0	0	0,5	0	0
E/I	E	338	12	0,5	0	0	0	0	0	0	0	0	0	0	1	0	0
N/S	N	121	15	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0,5	0,5	0	0
N/S	N	201	18	0,5	0	0	0	0	0,5	0,5	0	0	0	0	0,5	0	0,5
N/S	N	208	27	0,5	0	0	0	0	0,5	0,5	0	0	0	0	0,5	0	0
N/S	N	275	15	0,5	0	0	0	0	0,5	1	0	0	0	0	0,5	0	0
N/S	N	322	18	0,5	0	0	0	0	0,5	0,5	0,5	0	0,5	0,5	0,5	0	0
N/S	S	135	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N/S	S	423	5	0,5	0,5	0	0	0	0	0	0	0	0	0	0	0	0
N/S	S	431	20	0	0	0	0	0	0	0	0	0	0	0	1	0	0
N/S	S	456	4	0,5	1	0	0	0	0	0	0	0	0	0	0	0	0
N/S	S	463	6	0,5	0	0	0	0	1	1	0	0	0	0	1	0	0
T/F	T	80	11	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0,5	0	0	0
T/F	T	203	13	0,5	0	0	0	0	0	0,5	0	0	0	0	0,5	0	0
T/F	T	389	13	0,5	0	0	0	0	0	0	0	0	0	0	0,5	0	0
T/F	T	449	10	0,5	0	0	0	0	0	0,5	0	0	0	0	0,5	0	0,5
T/F	T	531	10	0,5	0	0	0	0	0,5	0,5	0,5	0,5	0	0	0	0	0
T/F	F	61	7	0	0	0	0	0	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0	0,5

Figura 6 – Exemplo da tabela utilizada para análise manual dos clusters puros.

(Fonte: Elaborada pelo autor)

a) **Objetivo 1:** Avaliar a possibilidade de identificar padrões entre os atributos da base de dados e os temperamentos/traços de personalidade sem o uso de redes complexas, utilizando para isso um teste estatístico.

b) **Objetivo 2:** Analisar os resultados dos clusters puros gerados pela rede, identificados em cada dimensão, e verificar a existência de diferenças significativas entre as classes.

Para alcançar o **objetivo 1**, foram realizadas análises estatísticas utilizando a base de dados TECLA em seu formato pré-processado (sem discretização), considerando as variáveis extraídas da plataforma *X* e os atributos psicológicos obtidos por meio do método LIWC. O objetivo principal dessa análise foi identificar quais variáveis apresentavam comportamentos estatisticamente distintos ao se comparar, isoladamente, os pares de categorias das dimensões do MBTI.

O teste de hipóteses utilizado foi o teste U de *Mann–Whitney*, que foi empregado para relacionar cada variável individualmente com as dimensões do MBTI. O teste U de *Mann–Whitney* é uma alternativa não paramétrica ao teste t de *Student*, sendo apropriado para comparar dois grupos em relação a uma variável ordinal, intervalar ou de razão, especialmente quando os dados não atendem aos pressupostos paramétricos exigidos pelo teste t, como a normalidade, mesmo para variáveis contínuas (GAUTHIER et al., 2007).

Para atingir o **objetivo 2**, com base na análise manual dos clusters apresentada na Seção 4.7, foi realizada uma nova rodada de testes de hipóteses, com o intuito de proporcionar uma análise mais robusta a partir de todos os clusters gerados. Essa etapa teve como objetivo verificar a presença de indícios estatísticos de diferenças entre os clusters em cada dimensão para as variáveis analisadas. Como as variáveis

já haviam sido discretizadas, utilizou-se o teste de independência qui-quadrado, que permite quantificar o grau de associação (ou dependência) entre duas variáveis categóricas (BUSSAB; MORETTIN, 2010).

4.9 Considerações Finais

Este capítulo apresentou o método desenvolvido para analisar os traços de personalidade dos usuários de mídias sociais utilizando redes complexas e testes estatísticos. Foram descritas as etapas de pré-processamento, transformação dos atributos, construção das redes e análise dos clusters gerados. O objetivo foi identificar padrões de comportamento associados aos temperamentos/traços de personalidade, utilizando técnicas de discretização e análise estatística para validar os resultados. O próximo capítulo apresenta os experimentos realizados e a análise dos resultados obtidos.

Capítulo 5

Experimentos e Análise dos Resultados

Este capítulo descreve os experimentos realizados neste trabalho, bem como os resultados obtidos. A Seção 5.1 detalha as bases de dados utilizadas, fornecendo informações sobre as fontes e as variáveis extraídas. A Seção 5.2 aborda a avaliação das estratégias de discretização das variáveis aplicadas na base TECLA e em seguida a análise da rede da base Kaggle. As Seções 5.3 e 5.4 apresentam os resultados das análises dos clusters das redes complexas da base TECLA e Kaggle, respectivamente, enquanto a Seção 5.5 faz uma comparação entre as duas bases. Por fim, a Seção 5.6 explora as análises estatísticas aplicadas para validar as hipóteses do trabalho.

5.1 Seleção da Base de Dados

Duas bases de dados relevantes foram selecionadas para a pesquisa, priorizando informações de usuários de mídias sociais com traços de personalidade ou temperamentos marcantes, além de atributos sociais como curtidas e número de seguidores. Vale ressaltar que houve dificuldade em encontrar bases de dados que contivessem, em simultâneo, informações sociais e informações relacionadas ao temperamento ou traços de personalidade dos usuários nas redes sociais.

As bases de dados utilizadas para o desenvolvimento deste trabalho têm como rótulo o traço de personalidade MBTI. Conforme apresentado por (PLANK; HOVY, 2015), o MBTI se apresenta em uma série de formas diferentes, sendo que em uma delas ele mede até 16 tipos que resultam da combinação de 4 dicotomias. As quatro dicotomias são: E/I (extroversão/introversão), N/S (sensação/intuição), V/F (pensamento/sentimento) e J/P (julgamento/percepção). A combinação dessas quatro polaridades gera os 16 tipos de traços de personalidade.

A primeira base de dados escolhida, chamada TECLA, proposta por (PLANK; HOVY, 2015), reuniu dados de 1500 usuários do *X* (anteriormente *Twitter*) que mencionaram um dos 16 tipos de personalidade do MBTI e incluíram as palavras Briggs ou Myers. Entre 100 e 2000 *tweets* foram coletados por usuário, totalizando 1,2 milhão de *tweets*.

Além dos últimos *tweets*, foram extraídas também algumas outras variáveis do usuário, sendo elas:

- ***followers_count***: Número de seguidores que a conta possuía no momento da extração.
- ***statuses_count***: Número de *Tweets* (incluindo *retweets*) emitidos pelo usuário até o momento da extração.
- ***favourites_count***: O número de *Tweets* que este usuário havia curtido durante a vida da conta até o momento da extração.
- ***listed_count***: O número de listas públicas das quais o usuário era membro até o momento da extração.
- **quantidade_tweets**: Quantidade de *tweets* que foram extraídos pelos autores para cada usuário.
- ***sex***: Sexo do usuário.
- ***mbti_personality***: Traço de personalidade MBTI daquele usuário.

A Tabela 3 apresenta a quantidade de usuários que possuem cada uma das categorias, em cada dimensão, do MBTI.

Tabela 3 – Proporção de usuários de cada categoria nas dimensões do MBTI na base de dados TECLA.

Dimensão	Quantidade	Proporção
E/I	539 / 961	35,9% / 64,1%
N/S	1.162 / 338	77,5% / 22,5%
T/F	624 / 876	41,6% / 58,4%
J/P	882 / 618	58,8% / 41,2%

O segundo conjunto de dados, chamado Kaggle, foi proposto por (RAI, 2022). Ele contém 8.328 usuários do *X* que relataram seus tipos de personalidade MBTI em suas descrições de perfil. Utilizando uma API, foram extraídos os últimos 200 *tweets* desses usuários, bem como diversas outras variáveis, incluindo as mesmas apresentadas na base de dados TECLA, com exceção da variável sexo. A Tabela 4 apresenta a quantidade e proporção de cada categoria, pra cada dimensão do MBTI, na base de dados Kaggle.

Tabela 4 – Proporção de usuários de cada categoria nas dimensões do MBTI na base de dados Kaggle.

Dimensão	Quantidade	Proporção
E/I	3.692 / 4.636	44,3% / 55,7%
N/S	6.319 / 2.009	75,9% / 24,1%
T/F	3.799 / 4.529	45,6% / 54,4%
J/P	4.441 / 3.887	53,3% / 46,7%

A base de dados Kaggle contém textos em diversos idiomas. Como o LIWC possui um dicionário em inglês, optou-se por manter apenas os textos nesse idioma. Para tanto, foi adotado o algoritmo de detecção de idioma da biblioteca *langdetec* [1] em Python.

Estatísticas dos atributos originais e daqueles extraídos pelo LIWC foram analisadas. Como os usuários possuíam diferentes quantidades e tamanhos de textos, os atributos psicológicos foram normalizados pela proporção de termos, permitindo comparações mais justas entre os usuários de cada base de dados. As Tabelas 5 e 6 mostram, respectivamente, as principais estatísticas dos atributos originais de ambas as bases de dados e dos atributos extraídos do LIWC.

Tabela 5 – Estatísticas dos atributos originais das bases de dados TECLA e Kaggle.

Base	Estatísticas	qtde_tweets	followers_count	statuses_count	favorites_count	listed_count	sex
	Tipo da Variável	Numérica	Numérica	Numérica	Numérica	Numérica	Binária
Base TECLA	Média	789	1642	16322	4663	36	-
	Desvio Padrão	550	14631	26363	9650	259	-
	Valor Mínimo	43	3	56	0	0	0
	25% Quartil	247	133	2268	211	2	0
	50% Quartil	774	309	7372	1241	6	1
	75% Quartil	1069	717	18514	4305	17	1
	Valor Máximo	1994	510168	258427	107889	8776	1
Base Kaggle	Média	195	2990	24747	17901	53	-
	Desvio Padrão	18	17843	42580	33725	177	-
	Valor Mínimo	0	0	20	0	0	-
	25% Quartil	198	529	3151	1276	7	-
	50% Quartil	200	982	9979	5624	19	-
	75% Quartil	200	1904	27741	19078	45	-
	Valor Máximo	200	951791	656839	585221	5727	-

Tabela 6 – Estatísticas dos atributos extraídos pelo LIWC nas bases de dados TECLA e Kaggle.

Base	Estatísticas	affect	posemo	negemo	anx	anger	sad	social	family	friend
	Tipo da Variável	Numérica	Numérica	Numérica	Numérica	Numérica	Numérica	Numérica	Numérica	Numérica
Base TECLA	Média	678	462	211	26	76	43	849	41	41
	Desvio Padrão	520	375	175	23	77	36	665	44	38
	Valor Mínimo	9	5	1	0	0	0	8	0	0
	25% Quartil	191	127	51	7	18	11	237	9	11
	50% Quartil	625	413	181	21	53	37	740	31	33
	75% Quartil	967	649	308	37	109	62	1192	59	58
	Valor Máximo	2853	2461	1039	161	526	222	4264	557	362
Base Kaggle	Média	130	90	38	6	13	8	240	8	8
	Desvio Padrão	68	52	25	5	11	6	134	7	6
	Valor Mínimo	0	0	0	0	0	0	0	0	0
	25% Quartil	92	58	20	2	4	4	154	2	3
	50% Quartil	143	94	38	5	11	8	251	6	7
	75% Quartil	176	123	55	8	19	12	328	11	11
	Valor Máximo	454	423	187	62	100	77	813	96	115

[1] <https://pypi.org/project/langdetect/>

5.2 Análises das Redes

Como o objetivo da criação da rede complexa é encontrar uma rede que alcance a melhor segmentação dentro de cada uma das quatro dimensões do MBTI, a estratégia para seleção da melhor rede, discutida na Seção 4.6 foi adotada para a base de dados TECLA e, após identificado o melhor método, foi replicado para a base Kaggle.

Três redes distintas foram construídas com base em diferentes estratégias de transformação/discretização dos atributos. A etapa de discretização mostrou-se necessária, uma vez que a simples normalização dos dados fez com que variáveis binárias, como o sexo, tivessem influência desproporcional na medida de similaridade entre os usuários. Esse viés ocorreu porque, para a maioria dos atributos, os valores estavam concentrados em faixas estreitas do intervalo [0,1], comprometendo a equidade na comparação entre perfis. Na sequência, são apresentados os resultados obtidos para cada uma das três redes geradas.

5.2.1 Resultados da Transformação 1

Na primeira proposta de transformação, todas as variáveis numéricas foram transformadas agrupando seus elementos em três categorias. Cada uma dessas classes foram calculadas a partir do menor e do maior valor de cada atributo. Nesse modelo foram identificados 164 clusters. O valor da modularidade foi de 0.939, que é um alto valor, próximo ao valor máximo dessa medida (1.0). Conceitualmente, a modularidade mede a qualidade de uma determinada divisão da rede. Seu valor varia de 0, que representa uma estrutura totalmente aleatória, até 1, indicando a presença de comunidades bem definidas na rede (CARNEIRO, 2017).

Após o processamento e criação da rede, 201 usuários não foram considerados. Esses usuários não tiveram distância Euclidiana de valor 0 com nenhum dos outros. Logo, a rede final possuía 1.299 nós. Optou-se por utilizar, inicialmente, o valor 0 de distância para identificar os grupos onde o conjunto de variáveis fossem iguais para todos os usuários.

Considerando a dimensão E/I, representada pelas cores verde/vermelho, nota-se que os usuários da classe I são a maioria (63,9%). Neste caso, 18,29% dos clusters da dimensão E são puros, em contraste com 51,83% da classe I. Considerando a dimensão N/S, cores rosa/azul, a maioria dos usuários são da classe N (74,84%). Para essa categoria 70,73% dos clusters são puros, enquanto apenas 2,44% da classe I o são. Já para a dimensão T/F, representada pelas cores laranjado/verde, 58,66% dos usuários são da classe F. 9,15% dos clusters são puros para a dimensão T, quanto 27,44% são para a classe F. Por fim, na dimensão J/P, representada por

azul/amarelo, tem-se a maioria dos usuários na classe J (59,51%). De todos os clusters, 31,10% são puros para a classe J, enquanto para a classe P são apenas 8,54%. A Tabela 7 apresenta de uma forma sumarizada os resultados apresentados anteriormente.

Tabela 7 – Resultados da rede obtida a partir da primeira transformação das variáveis.

Dimensão	Legenda das Cores	% de usuários na rede	% clusters "puros"	Média Geral de "pureza"
E/I	verde/vermelho	36,10% / 63,90%	18,29% / 51,83%	35,06%
N/S	rosa/azul	74,84% / 21,56%	70,73% / 2,44%	36,59%
T/F	laranjado/verde	41,34% / 58,66%	9,15% / 27,44%	18,29%
J/P	azul/amarelo	59,51% / 40,49%	31,1% / 8,54%	19,82%

Um dos critérios para seleção de clusters na análise visual é a quantidade de usuários nos grupos puros, para entender a sua representatividade. Essa proposta tem como objetivo estudar com mais detalhes os grupos mais representativos. Entre os clusters com predominância da classe E, 7 apresentam no mínimo 10 elementos. Para a classe I, esse número aumenta para 15. No caso da classe N, 26 clusters possuem pelo menos 10 elementos, enquanto para a classe S, nenhum grupo tem mais de 3 elementos predominantes. Nos clusters com predominância da classe F, 8 têm no mínimo 10 elementos. Para a classe T, apenas 1 grupo supera os 10 elementos, enquanto os restantes têm, no máximo, 7. Em relação à classe J, 5 clusters possuem pelo menos 10 elementos, ao passo que, na classe P, nenhum grupo ultrapassa os 5 elementos.

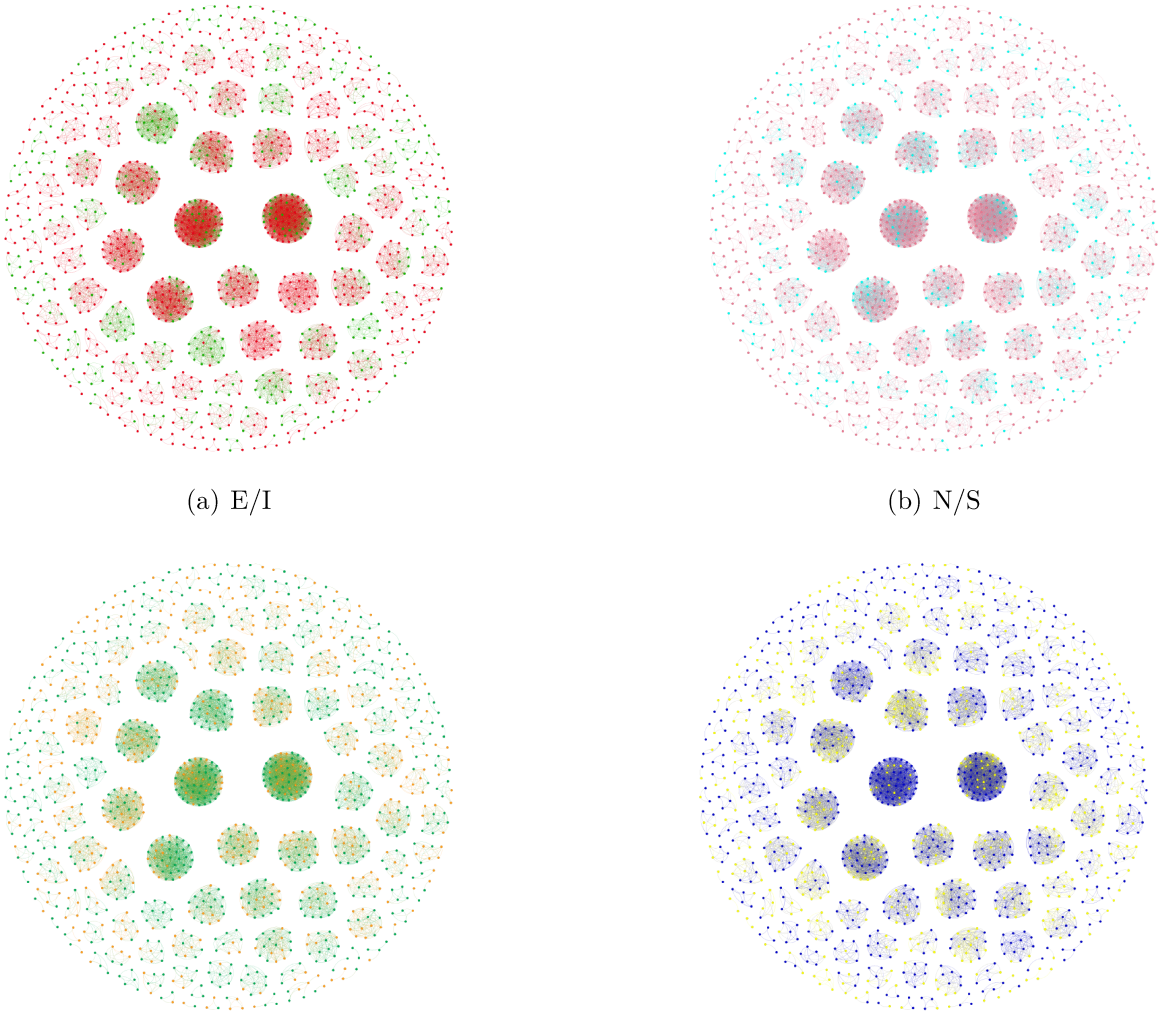
A Figura 7 apresenta a plotagem dessa rede segmentada para cada um dos conjuntos de dimensões.

5.2.2 Resultados da Transformação 2

Na segunda proposta de transformação, todas as variáveis numéricas foram transformadas agrupando seus elementos em cinco classes. Cada uma dessas classes foram calculadas a partir do menor e do maior valor de cada atributo. Nesse modelo foram identificados 217 clusters. Além disso, o valor da modularidade foi de 0.971.

Após o processamento e criação da rede, 890 usuários não foram considerados. Esses usuários não tiveram distância Euclidiana de valor 0 com nenhum dos outros. Logo, a rede final possuía 610 nós.

Na dimensão E/I, representada pelas cores verde e vermelho, observa-se que a maioria dos usuários pertence à classe I (63,9%). Neste contexto, 21,20% dos clusters da classe E são puros, em comparação com 47,00% na classe I. Na dimensão N/S, com as cores rosa e azul, a classe N é predominante, representando 79,18% dos usuários. Para essa classe, 60,83% dos clusters são puros, enquanto apenas 4,15% dos clusters



(a) E/I (b) N/S

(c) T/F (d) J/P

Figura 7 – Rede Complexa com a separação das dimensões utilizando a primeira proposta de transformação das variáveis.

(Fonte: Elaborada pelo autor)

da classe S apresentam essa característica. Na dimensão T/F, representada por laranja e verde, 58,03% dos usuários estão na classe F, com 31,34% dos clusters sendo puros, frente a 16,60% na classe T. Por fim, na dimensão J/P, destacada pelas cores azul e amarelo, a classe J conta com a maioria dos usuários (62,79%), com 36,41% dos clusters sendo puros, contra apenas 9,22% na classe P. A Tabela 8 apresenta de uma forma sumarizada os resultados apresentados anteriormente.

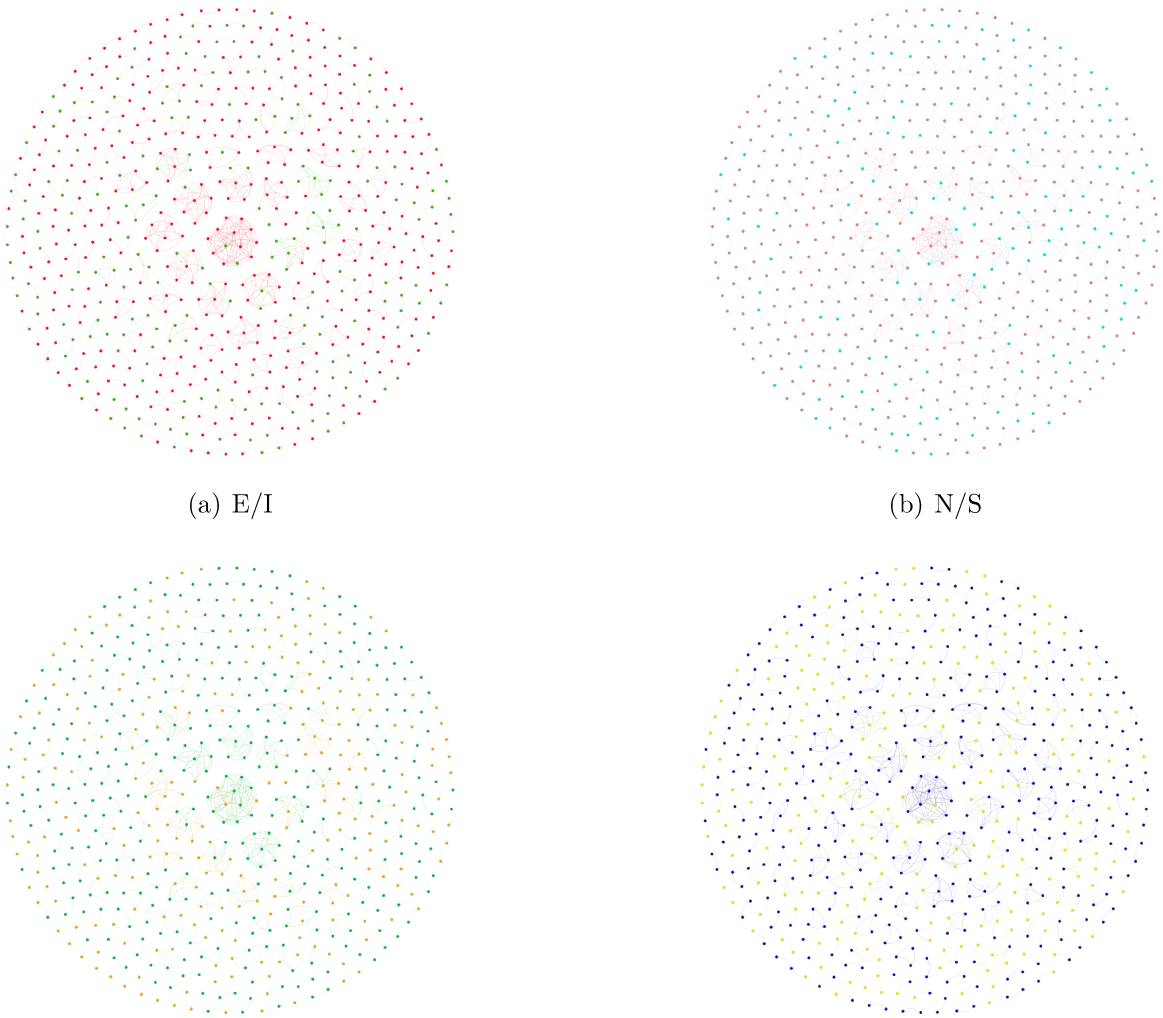
Nos clusters onde a classe E predomina, nenhum grupo tem mais de 7 elementos. Para a classe I, apenas um grupo conta com 11 elementos, enquanto os demais não ultrapassam 8. Na classe N, um grupo possui 13 elementos, enquanto os demais têm, no máximo, 8. Todos os clusters da classe S possuem apenas 2 elementos, ambos pertencentes à classe. Nos clusters dominados pela classe F, um grupo contém 13

Tabela 8 – Resultados da rede obtida a partir da segunda transformação das variáveis.

Dimensão	**Legenda das Cores**	**% de usuários na rede**	**% clusters "puros"**	**Média Geral de "pureza"**
E/I	verde/vermelho	36,10% / 63,90%	21,20% / 47,00%	34,10%
N/S	rosa/azul	79,18% / 20,82%	60,83% / 4,15%	32,49%
T/F	laranjado/verde	41,97% / 58,03%	16,60% / 31,34%	23,96%
J/P	azul/amarelo	62,79% / 37,21%	36,41% / 9,22%	22,81%

elementos, e os outros não passam de 8. Para a classe T, nenhum grupo ultrapassa 7 elementos. Da mesma forma, nenhum cluster das classes J e P tem mais de 7 elementos.

A Figura 8 apresenta a plotagem dessa rede segmentada para cada um dos conjuntos de dimensões.



(a) E/I (b) N/S

(c) T/F (d) J/P

Figura 8 – Rede Complexa com a separação das dimensões utilizando a segunda proposta de transformação das variáveis.

(Fonte: Elaborada pelo autor)

5.2.3 Resultados da Transformação 3

Na terceira proposta de transformação, todas as variáveis numéricas foram transformadas agrupando seus elementos em três classes. Adotou-se o K-Means para clusterização de cada um desses atributos. Nesse modelo foram identificados 118 clusters. Além disso, o valor geral da modularidade obtido foi de 0.939.

Após o processamento e criação da rede, 1193 usuários não foram considerados. Esses usuários não tiveram distância Euclidiana de valor 0 com nenhum dos outros. Logo, a rede final possuía 307 nós.

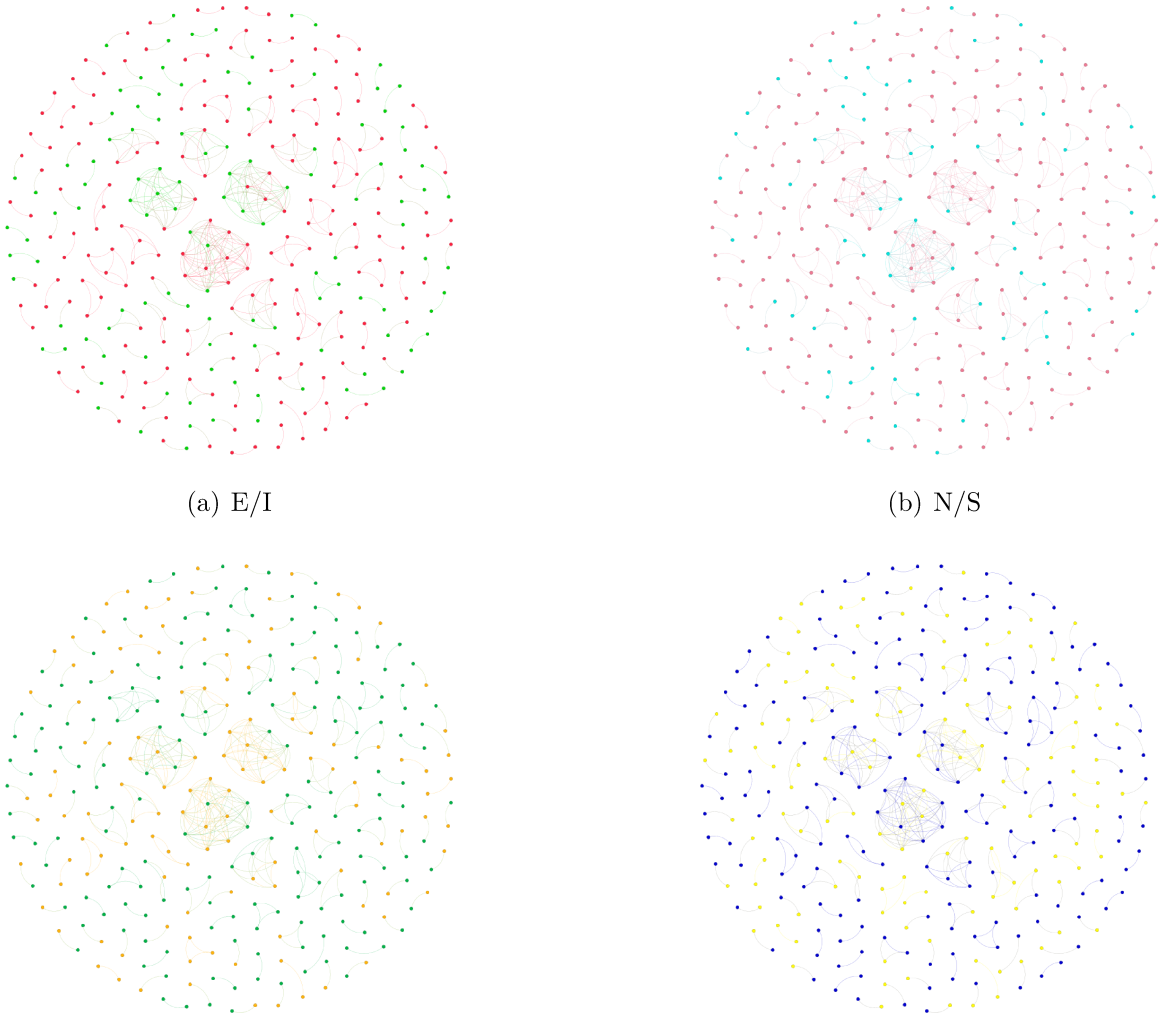
Na dimensão E/I, representada pelas cores verde e vermelho, a classe I prevalece, com 62,21% dos usuários. Nessa categoria, 39,83% dos clusters são puros, enquanto na classe E, apenas 18,64% apresentam essa característica. Na dimensão N/S, destacada pelas cores rosa e azul, a maioria dos usuários pertencem à classe N, correspondendo a 78,83%. Para essa classe, 63,56% dos clusters são puros, em contraste com apenas 4,24% na classe S. Na dimensão T/F, identificada pelas cores laranja e verde, 57,65% dos usuários estão na classe F, com 31,36% dos clusters sendo puros, em comparação com 12,71% na classe T. Finalmente, na dimensão J/P, representada por azul e amarelo, a classe J tem a maioria, com 58,63% dos usuários, e 32,21% dos clusters são puros, contra 15,25% na classe P. A Tabela 9 apresenta de uma forma sumarizada os resultados apresentados anteriormente.

Tabela 9 – Resultados da rede obtida a partir da terceira transformação das variáveis.

Dimensão	**Legenda das Cores**	**% de usuários na rede**	**% clusters "puros"**	**Média Geral de "pureza"**
E/I	verde/vermelho	37,79% / 62,21%	18,64% / 39,83%	29,24%
N/S	rosa/azul	78,83% / 21,17%	63,56% / 4,24%	33,90%
T/F	laranjado/verde	42,35% / 57,65%	12,71% / 31,36%	22,03%
J/P	azul/amarelo	58,63% / 41,37%	32,20% / 15,25%	23,73%

Nos clusters dominados pela classe E, nenhum grupo possui mais de 9 elementos. Para a classe I, um grupo contém 12 elementos, enquanto os demais não ultrapassam 4. Na classe N, um dos clusters tem 10 elementos, e os outros não excedem 9. Na classe S, todos os grupos possuem apenas 2 elementos, ambos pertencentes à classe. Nos clusters da classe F, nenhum grupo tem mais de 5 elementos. Para a classe T, um grupo contém 10 elementos, enquanto os restantes não passam de 4. Nos clusters com predominância da classe J, nenhum tem mais de 4 elementos, o mesmo ocorrendo com a classe P.

A Figura 9 apresenta a plotagem dessa rede segmentada para cada um dos conjuntos de dimensões.



(a) E/I (b) N/S

(c) T/F (d) J/P

Figura 9 – Rede Complexa com a separação das dimensões utilizando a terceira proposta de transformação das variáveis.

(Fonte: Elaborada pelo autor)

5.2.4 Comparativo entre as transformações

Entre os três métodos utilizados para transformar as variáveis, apenas a transformação 1 apresentou um equilíbrio mais adequado, resultando em uma maior proporção de clusters com pelo menos 70% de elementos puros em cada dimensão, além de gerar grupos com maior número de usuários, o que possibilitou análises mais consistentes. Dessa forma, identificou-se que esta foi a abordagem mais apropriada para a base TECLA. Importante ressaltar que, por mais que tenha sido necessário a remoção de alguns usuários em cada uma das transformações, a proporção de usuários em cada categoria, de cada dimensão do MBTI, foi representativa em relação à proporção da base original e completa.

Considerando que ambas as bases possuem natureza semelhante, compostas por da-

dos textuais e atributos sociais extraídos de usuários do *X* (antigo *Twitter*), adotou-se o mesmo método para o conjunto Kaggle, a fim de manter a comparabilidade entre as análises e assegurar a consistência metodológica nos resultados.

Todas as variáveis numéricas foram transformadas agrupando seus elementos em três categorias. Cada uma dessas classes foi calculada com base nos menores e maiores valores de cada atributo. Neste modelo, foram identificados 164 clusters com valor de modularidade de 0,939 para a base de dados TECLA, que representa um bom valor. Conceitualmente, a modularidade mede a qualidade de uma determinada divisão da rede. Seu valor varia de 0, que representa uma estrutura completamente aleatória, a 1, indicando a presença de comunidades bem definidas na rede.

5.2.5 Análise da Rede da Base Kaggle

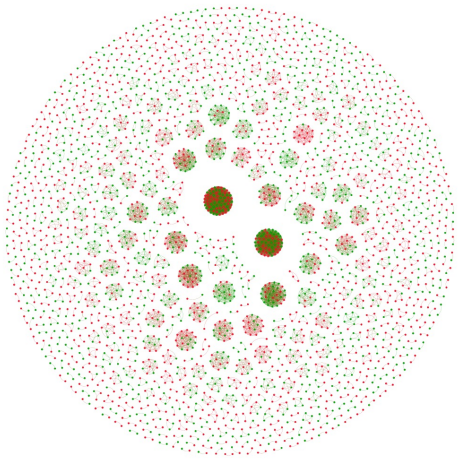
Conforme apresentado na Seção 5.2.4, uma vez identificado que a melhor rede foi obtida a partir da Transformação 1, optou-se por adotar o mesmo procedimento para a construção da rede referente à base Kaggle. Especificamente para essa base de dados, os *outliers* foram desconsiderados no momento do cálculo dos limites de discretização. Tal decisão foi motivada pelo elevado número de instâncias presentes na base, o que resultava em variáveis com *outliers* que tendiam a enviesar o cálculo dos limites e, consequentemente, comprometer o resultado final.

Sem a remoção dos *outliers*, observou-se que a maioria dos atributos apresentava concentrações elevadas na menor classe discretizada, o que acarretava na formação de um número excessivo de pares de usuários com distância Euclidiana igual a zero. Esse comportamento implicava na geração de uma rede demasiadamente densa, dificultando a aplicação dos algoritmos de análise e comprometendo a visualização adequada da rede gerada.

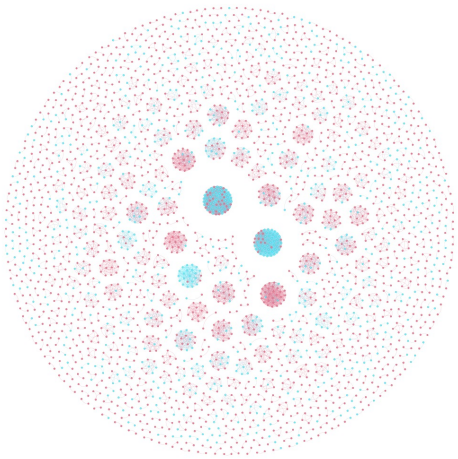
Após o processamento e a criação das redes, 5.428 usuários para a base Kaggle foram desconsiderados, uma vez que esses usuários não apresentaram distância Euclidiana igual a 0 para nenhuma outra base. Portanto, a rede Kaggle final possui 2.900 nós. A Figura 10 mostra a rede segmentada para cada um dos conjuntos de dimensões. As redes foram criadas como grafos não direcionados e visualizadas na ferramenta Gephi (versão 0.10) (BASTIAN; HEYMANN; JACOMY, 2009), enquanto a biblioteca *IGraph* [2] foi utilizada para construir e extrair medidas.

Para os conjuntos de dados selecionados, a pureza do cluster foi calculada para cada dicotomia em cada dimensão do MBTI. Por exemplo, a pureza do cluster c em relação à dicotomia E da dimensão E/I (extroversão/introversão) é calculada dividindo o número de usuários com dicotomia E dentro desse cluster pelo número total de elementos nesse cluster.

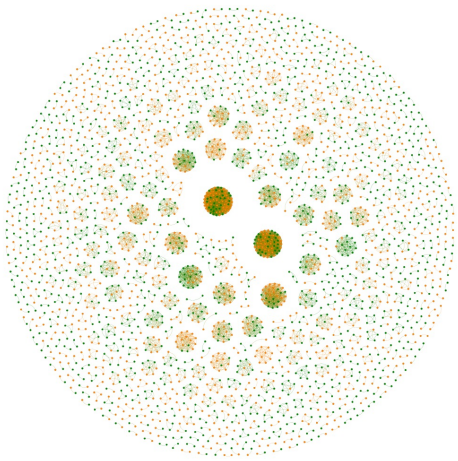
[2] <https://igraph.org/>



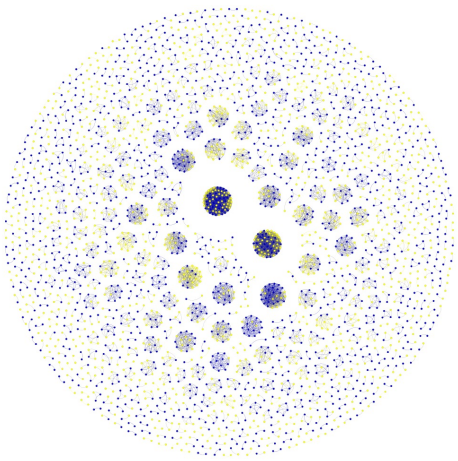
(a) E/I



(b) N/S



(c) T/F



(d) J/P

Figura 10 – Rede Complexa com a separação das dimensões utilizando a primeira proposta de transformação das variáveis aplicada na base Kaggle.

(Fonte: Elaborada pelo autor)

Nessa base de dados, os usuários da classe I (vermelho) também predominam na dimensão E/I, com 53,4%. Nesse contexto, 27,49% dos clusters são puros para a classe I, enquanto 17,18% são para a classe E (verde). Na dimensão N/S (rosa/azul), a classe N mantém a maioria, com 72,86%. Entre os clusters, 57,94% são puros para a classe N e apenas 6,04% para a classe S. Na dimensão T/F (laranja/verde), os usuários da classe F representam 50,03%. A pureza dos clusters é de 25% para a classe F e 21,68% para a classe T. Por fim, na dimensão J/P (azul/amarelo), a classe J lidera com 53,45%. Entre os clusters, 24,76% são puros para a classe J e 18,13% para a classe P. A Tabela 10 apresenta um resumo dos resultados apresentados anteriormente. Importante ressaltar que, mesmo após a alta remoção de usuários para criação da rede a proporção de usuários em cada categoria, de cada dimensão do MBTI, é similar ao da base original do Kaggle.

Entre os clusters da classe E, 5 atingem 10 ou mais elementos, enquanto na classe I, esse número é de apenas 3. A classe N possui 23 clusters que excedem esse número, enquanto na classe S, apenas 2 atendem ao critério. Para a classe F, 4 clusters possuem pelo menos 10 elementos, e na classe T, esse número é 6. Entre os clusters da classe J, 14 possuem pelo menos 10 elementos, e na classe P, apenas 1 grupo atinge esse nível.

Tabela 10 – Resultados da rede para a base Kaggle.

Dimensão	**Legenda das Cores**	**% de usuários na rede**	**% clusters "puros"**	**Média Geral de "pureza"**
E/I	verde/vermelho	46,60% / 53,40%	17,18% / 27,49%	22,33%
N/S	rosa/azul	72,86% / 27,14%	57,93% / 6,04%	31,99%
T/F	laranjado/verde	49,97% / 50,03%	21,68% / 25,00%	23,34%
J/P	azul/amarelo	53,45% / 46,55%	24,76% / 18,13%	21,45%

5.3 Análise dos clusters na base de dados TECLA

Dentre os clusters puros em cada dimensão, cinco de cada classe foram selecionados para identificar um padrão entre seus usuários. O critério para a escolha dos clusters foi priorizar aqueles com o maior número de elementos, para que as análises pudessem ser feitas com base em um número maior de usuários.

Para a dimensão E/I (Extroversão/Introversão), identificou-se que usuários da classe E publicam um número elevado de *tweets*, enquanto os da classe I 83% publicam um número menor (intermediário), o que corrobora com (FILHO et al., 2010), que afirma que pessoas da classe E estão disponíveis para buscar interação por meio de relacionamentos sociais, enquanto as da classe I buscam menos interação. Além disso, usuários da classe I publicam um número baixo de termos positivos.

Para os clusters selecionados na análise da dimensão N/S (Intuição/Sensação), não foram identificadas, de forma visual, diferenças ou padrões significativos em relação a nenhuma das variáveis consideradas.

Analisando a dimensão T/F (Pensamento/Sentimento), 66% os clusters da classe T foram compostos por usuários do sexo masculino, enquanto os da classe F, totalmente por usuários do sexo feminino. Culturalmente, existe uma forte associação do sexo masculino com a razão, assim como do sexo feminino com a emoção (MARTINS-SUAREZ; SOUSA, 2016). No entanto, alguns autores, como (BOURDIEU, 2010), defendem a desconstrução dessa associação, afirmando que isso dá às mulheres a oportunidade de agir e participar do cenário social e político.

Por fim, para a dimensão J/P (Julgamento/Percepção), os usuários da classe J publicaram um número baixo de termos negativos em seus textos, enquanto os da

classe P publicaram um número intermediário. Em relação a termos raivosos, todos os clusters analisados da classe J publicam uma baixa quantidade, enquanto da classe P dividem-se entre baixo e intermediário. O Julgamento envolve a tomada de decisão direta, com preferência por conclusões rápidas e baixa tolerância à incerteza, enquanto a Percepção se caracteriza por uma abordagem mais analítica, priorizando a compreensão do problema e a busca por informações antes da tomada de decisão (FILHO et al., 2010).

5.4 Análise dos clusters na base de dados Kaggle

Na dimensão E/I (Extroversão/Introversão), observou-se que tanto os usuários da classe E quanto os da classe I publicam um número intermediário de *tweets*. Os usuários da classe E apresentaram baixa ocorrência de termos negativos, raivosos e tristes em seus textos, enquanto os da classe I publicaram um número intermediário desses termos. Além disso, os usuários da classe I também apresentaram um número intermediário de termos positivos.

Na dimensão N/S (Intuição/Sentimento), em relação aos termos positivos e afetivos, os usuários da classe N apresentaram um número intermediário, enquanto os da classe S demonstraram baixa frequência desses termos.

Para os clusters selecionados na análise da dimensão T/F (Pensamento/Sentimento), não foram identificadas, de forma visual, diferenças ou padrões significativos em relação a nenhuma das variáveis consideradas.

Por fim, na dimensão J/P (Julgamento/Percepção), observou-se que os usuários da classe J publicaram, em sua maioria, um número baixo de termos positivos, enquanto na classe P a quantidade é intermediária ou alta. Os usuários da classe J apresentaram um número baixo de publicações, enquanto na classe P o número foi intermediário em 60% dos clusters. Em relação aos termos positivos e sociais, os usuários da classe J apresentaram uma quantidade baixa na maioria dos clusters, enquanto na classe P a quantidade foi intermediária ou alta.

5.5 Comparação entre as redes Kaggle e TECLA

A comparação entre os resultados apresentados nas Seções 5.3 e 5.4, em relação às redes Kaggle e TECLA, revelou semelhanças e diferenças:

E/I (Extroversão/Introversão): A base TECLA evidenciou um padrão mais marcante, com a classe E apresentando um número elevado de publicações, em consonância com o perfil socialmente mais ativo desse grupo, conforme discutido por (FILHO et al., 2010). Além disso, na base TECLA, os usuários da classe I apresen-

taram um número baixo de termos positivos, enquanto na Kaggle essa frequência foi intermediária. Na base Kaggle, a classe E também se destacou pela baixa incidência de termos negativos, raivosos e tristes, característica não observada de forma destacada na base TECLA.

N/S (Intuição/Sensação): Na base TECLA, não foram identificados padrões visuais significativos para esta dimensão. Já na base Kaggle, observou-se que usuários da classe N apresentaram uma frequência intermediária de termos positivos e afetivos, enquanto os da classe S demonstraram baixa frequência desses termos. Esse contraste evidencia que a base Kaggle ofereceu melhores indícios de distinção entre as classes N e S em termos de linguagem emocional.

T/F (Pensamento/Sentimento): Na base TECLA, observou-se uma relação entre gênero e a dimensão T/F, com a predominância de usuários do sexo masculino nos clusters da classe T, e do sexo feminino nos da classe F, em consonância com os achados de (MARTINS-SUAREZ; SOUSA, 2016). No entanto, na base Kaggle não foram identificados padrões significativos entre os grupos para esta dimensão, sugerindo que a relação entre os atributos analisados e os perfis T/F pode variar entre os conjuntos de dados. Importante ressaltar que na base Kaggle não havia a variável sexo do usuário.

J/P (Julgamento/Percepção): A base TECLA revelou que usuários da classe J publicam menos termos negativos e raivosos, enquanto os da classe P apresentaram uma frequência intermediária. Na base Kaggle, observou-se que usuários da classe J publicaram, em sua maioria, um número baixo de termos positivos e sociais, além de um volume reduzido de publicações. Em contrapartida, os usuários da classe P apresentaram uma frequência intermediária ou alta desses termos. Ambas as bases, portanto, indicam diferenças de padrão textual entre as classes J e P, com destaque para uma comunicação mais positiva e ativa por parte dos usuários do grupo P.

As discrepâncias observadas entre os resultados das bases TECLA e Kaggle podem ser atribuídas a diversos fatores metodológicos. Em primeiro lugar, destaca-se a diferença no volume de dados: enquanto a base Kaggle considera apenas os 200 *tweets* mais recentes por usuário, a base TECLA contempla até 2.000 publicações, proporcionando uma representação textual mais ampla e robusta. Além disso, há uma distinção importante quanto à homogeneidade linguística: os textos da base TECLA são exclusivamente em inglês, ao passo que a base Kaggle apresenta publicações em diversos idiomas, o que demandou um processo adicional de filtragem para manter apenas os textos em língua inglesa.

Outro ponto relevante diz respeito à amostragem dos clusters analisados. Foram selecionados apenas alguns clusters puros por dicotomia, o que implica que muitos outros clusters, possivelmente ricos em informações, não foram explorados.

Considerando-se o grande número de clusters gerados em ambas as bases e o fato de que nem todas as variáveis apresentam 100% de predominância de uma única classe, a análise visual de todos os agrupamentos torna-se limitada e subjetiva. Além disso, não se sabe em termos de gênero e idade se as bases são muito diferentes. Isso pode ser também um fator importante que impacta nas diferenças identificadas.

5.6 Análises Estatísticas

Este trabalho propõe a utilização de testes estatísticos em dois momentos distintos de avaliação, cada um com objetivos específicos:

a) **Objetivo 1:** Avaliar a possibilidade de identificar padrões entre os atributos da base de dados e os temperamentos/traços de personalidade sem o uso de redes complexas, utilizando para isso um teste estatístico.

b) **Objetivo 2:** Analisar os resultados dos clusters puros gerados pela rede, identificados em cada dimensão, e verificar a existência de diferenças significativas entre as classes.

A fim de avaliar a hipótese do trabalho (Objetivo 1), de que as redes complexas conseguem identificar padrões e correlações entre temperamento/traço de personalidade e informações textuais/comportamentais extraídas de mídias sociais além das identificadas por meio de análise estatística, foram realizadas análises estatísticas na base de dados pré-processada (sem discretização). As análises foram feitas apenas para a base de dados TECLA. A primeira etapa foi a realização de um teste de hipóteses, por meio do teste U de *Mann–Whitney*, para avaliar se cada variável possuía divergência significativa em relação a cada dimensão do MBTI.

A Tabela 11 apresenta um resumo do teste realizado. Para cada dimensão, foi marcado com um X quando o resultado do teste forneceu um p-valor inferior a 5%, indicando haverem evidências significativas para se afirmar que o par de categorias daquela dimensão possuíam um comportamento diferente.

A realização do teste de hipóteses permitiu identificar que diversas variáveis apresentaram diferenças estatisticamente significativas ao se comparar os pares de categorias em cada dimensão do MBTI. Na dimensão E/I, 70% das variáveis testadas mostraram significância estatística. Para a dimensão N/S, esse percentual foi de apenas 20%. Já na dimensão T/F, 73,3% das variáveis apresentaram diferenças significativas, enquanto na dimensão J/P esse número foi de 26,7%.

Ao se comparar esses resultados obtidos na base original com aqueles provenientes da análise realizada sobre os dados da rede complexa — após o processo de discretização — observa-se uma redução expressiva na quantidade de variáveis com diferenças

Tabela 11 – Resultado das variáveis com resultado significativo para o teste de hipóteses.

Dimensão	E/I	N/S	T/F	J/P
qtde_tweets	X	-	-	-
followers_count	X	-	X	-
statuses_count	X	X	-	-
favorites_count	-	-	X	X
listed_count	-	-	-	-
affect	-	-	X	-
posemo	X	-	X	X
negemo	X	-	-	X
anx	-	-	X	-
anger	X	-	X	X
sad	-	-	X	-
social	X	X	X	-
family	X	-	X	-
friend	X	-	X	-
sex	-	X	X	-

entre as categorias de cada dimensão. Essa redução pode estar associada ao próprio processo de discretização, que tende a agrupar múltiplas instâncias sob categorias comuns, diminuindo a sensibilidade para detectar variações mais sutis.

Para fim de um melhor entendimento das diferenças encontradas pelo teste, a Tabela 12 apresenta, para os resultados da Tabela 11 qual a "direção"da diferença, ou seja, qual categoria apresenta maior valor em detrimento da outra. Para essa análise foi utilizada a mediana, e não a média, visto que essa medida está mais associada aos cálculos nos testes não paramétricos (GAUTHIER et al., 2007).

Após a aplicação do teste U de Mann–Whitney, observou-se que várias variáveis apresentaram diferenças estatisticamente significativas entre as categorias de cada dimensão do MBTI. Destacam-se principalmente as dimensões E/I e T/F, com mais de 70% das variáveis mostrando distinções claras entre os grupos, enquanto N/S e J/P apresentaram resultados mais discretos. Esses achados confirmam que, mesmo sem o uso de redes complexas, é possível identificar padrões relevantes entre atributos comportamentais/linguísticos e as dimensões da personalidade, ainda que de forma mais limitada.

Por fim, foi realizado o último de teste de hipóteses em relação aos clusters puros obtidos a partir da criação da rede (Objetivo 2). Na análise visual foi amostrado uma pequena quantidade a partir do montante gerado. O objetivo com a análise estatística, neste caso, é identificar se, ao analisar o todo ainda temos resultados semelhantes em relação à análise visual da rede, ou se divergem. A Tabela 13 apresenta para quais variáveis o resultado do teste de independência foi significativo,

Tabela 12 – Direção da diferença identidade no teste de hipóteses, em relação a cada dimensão do MBTI.

Dimensão	E/I	N/S	T/F	J/P
qtde_tweets	E >I	-	-	-
followers_count	E >I	-	F >T	-
statuses_count	I >E	S >N	-	-
favorites_count	-	-	F >T	P >J
listed_count	-	-	-	-
affect	-	-	F >T	-
posemo	E >I	-	F >T	J >P
negemo	I >E	-	-	P >J
anx	-	-	F >T	-
anger	I >E	-	T >F	P >J
sad	-	-	F >T	-
social	E >I	N >S	F >T	-
family	E >I	-	F >T	-
friend	E >I	-	F >T	-
sex	-	-	-	-

indicando que há evidências para afirmar diferença significativa entre os grupos de cada dimensão.

Tabela 13 – Resultado das variáveis com resultado significativo para o teste de hipóteses a partir dos clusters.

Dimensão	E/I	N/S	T/F	J/P
qtde_tweets	X	-	-	-
followers_count	-	-	-	-
statuses_count	-	-	-	-
favorites_count	-	-	-	-
listed_count	-	-	-	-
affect	-	X	-	-
posemo	-	-	-	-
negemo	-	X	-	X
anx	-	-	-	-
anger	-	-	-	X
sad	-	-	-	-
social	-	-	-	-
family	-	-	-	-
friend	-	-	-	-
sex	-	-	X	-

Confrontando os resultados obtidos por meio do teste estatístico com a análise visual realizada a partir das amostras de clusters da base TECLA, foram identificadas as seguintes semelhanças e divergências em relação a cada dimensão do MBTI:

- **Dimensão E/I**: A variável *qtde_tweets* apresentou diferenças consistentes entre as duas categorias, tanto na análise estatística quanto na avaliação visual;
- **Dimensão N/S**: Nenhum padrão claro foi observado visualmente nos clusters analisados, entretanto, a análise estatística indicou diferenças significativas entre as categorias N e S para as variáveis *affect* e *negemo*;
- **Dimensão T/F**: A variável *sexo* foi a única a apresentar diferenças consistentes em ambas as abordagens, sugerindo uma associação robusta com essa dimensão;
- **Dimensão J/P**: As variáveis *negemo* e *anger* demonstraram diferenças significativas tanto na análise visual quanto na estatística.

Os testes de independência aplicados sobre os clusters puros revelaram que algumas das diferenças detectadas visualmente foram corroboradas estatisticamente, como nos casos de *qtde_tweets* (E/I), *negemo* e *anger* (J/P), além da associação de *sexo* com T/F. Em contrapartida, padrões não identificados visualmente emergiram na análise estatística, como em *affect* e *negemo* (N/S). Dessa forma, os resultados reforçam a relevância da análise com redes complexas: ao estruturar os dados em clusters, torna-se possível confirmar ou ampliar evidências que seriam mais difíceis de capturar apenas por análises visuais ou estatísticas tradicionais.

A aplicação do teste estatístico se mostra fundamental para reduzir possíveis vieses decorrentes da análise de uma amostra limitada de clusters, que pode não refletir adequadamente o comportamento geral da base. Além disso, o uso de métodos estatísticos permite uma avaliação mais abrangente e objetiva, minimizando interpretações subjetivas e garantindo maior rigor na identificação de padrões relevantes.

5.7 Limitações

Apesar das contribuições apresentadas, este trabalho possui algumas limitações que devem ser reconhecidas e discutidas de forma transparente, a fim de contextualizar os resultados obtidos e orientar investigações futuras.

- **Bases de dados limitadas e rotulagem indireta:** as análises foram conduzidas a partir de bases específicas (TECLA e Kaggle), compostas por usuários rotulados de acordo com o MBTI. A rotulagem indireta por meio de autoidentificação ou coleta via palavras-chave pode introduzir vieses de representatividade e comprometer a generalização dos achados.
- **Escopo reduzido de modelos psicológicos:** o trabalho concentrou-se na tipologia do MBTI, com suporte ao modelo de temperamentos de Keirsey. Em-

bora relevante, outras abordagens (como Big Five ou TEMPS-RIO) poderiam complementar e validar a robustez dos padrões observados.

- **Dependência de atributos textuais e linguísticos:** os dados analisados foram majoritariamente textuais, extraídos das publicações dos usuários. Características não textuais (como padrões de rede social, horários de interação, imagens ou multimídia) não foram incorporadas, o que pode limitar a diversidade de indicadores psicológicos captados.
- **Transformação dos dados em redes complexas:** a proposta de construção de redes baseou-se em medidas de similaridade e discretização, que embora fundamentadas, envolvem escolhas metodológicas que impactam diretamente a formação de comunidades. Diferentes técnicas de transformação poderiam gerar redes com propriedades distintas.
- **Validação restrita:** os resultados foram avaliados a partir de métricas de modularidade, pureza de clusters e testes estatísticos. Embora esses métodos sejam adequados, não houve comparação direta com outros métodos de modelagem baseados em redes, nem validação com especialistas da área de psicologia.
- **Generalização dos achados:** os padrões encontrados referem-se especificamente às bases utilizadas e ao contexto da plataforma X (Twitter). É necessário cautela ao extrapolar os resultados para outros contextos ou diferentes mídias sociais.

5.8 Considerações Finais

Os resultados obtidos com a aplicação da Transformação 1 evidenciaram o potencial das redes complexas em representar e segmentar usuários com base em traços psicológicos e comportamentais. A rede construída apresentou alta modularidade (0.939) e formou 164 clusters distintos, muitos deles com elevada pureza em ao menos uma dimensão do MBTI. As análises dos clusters mais representativos demonstraram padrões recorrentes de comportamento e linguagem entre os usuários de um mesmo tipo, especialmente nas dimensões E/I e N/S, como diferenças na frequência de postagens, uso de termos afetivos e indicadores de socialização. Essas evidências, somadas à coerência com achados prévios da literatura (como no caso da dimensão T/F associada ao gênero), reforçam que a abordagem baseada em redes complexas é capaz de capturar estruturas latentes associadas à personalidade, complementando e enriquecendo os métodos estatísticos tradicionais.

Capítulo 6

Conclusão

O objetivo deste trabalho foi analisar e correlacionar os temperamentos/traços de personalidade de usuários de mídias sociais a partir de dados obtidos nessas plataformas, através da utilização de redes complexas. A abordagem adotada neste estudo não visou a classificação dos usuários, mas sim a identificação de padrões a partir da modelagem de redes construídas com base em dados textuais e comportamentais extraídos da plataforma *X* (*Twitter*).

Foram utilizadas duas bases de dados contendo, além dos textos publicados pelos usuários, informações relacionadas ao MBTI. A metodologia envolveu as etapas de pré-processamento textual, extração de atributos psicológicos com o método LIWC, normalização dos dados, discretização das variáveis e construção de redes complexas com base na similaridade entre os usuários. Em seguida, foram aplicados algoritmos de detecção de comunidades e analisadas as métricas de modularidade, pureza e distribuição dos clusters para investigar a relação entre os grupos e as dimensões que compõe os traços de personalidade do MBTI.

A análise visual e estatística dos clusters mais puros revelou padrões distintos entre os grupos, especialmente nas dimensões E/I e N/S, indicando que determinados traços de personalidade compartilham similaridades comportamentais e linguísticas que se refletem nas estruturas das redes. As redes construídas com a transformação 1 apresentaram os melhores resultados em termos de modularidade e pureza, sugerindo que essa transformação favorece a segmentação de usuários com características psicológicas semelhantes.

Testes estatísticos complementares mostraram que, embora algumas variáveis continuem apresentando diferenças significativas entre os tipos de personalidade após a discretização e criação das redes, há uma redução na quantidade de variáveis significativas em relação à base original. Isso indica que a transformação dos dados e a modelagem em rede priorizam padrões mais globais e menos sensíveis à variabilidade individual de cada atributo.

Por fim, os resultados sustentam a hipótese central do trabalho: redes complexas permitem identificar padrões de personalidade e agrupamentos relevantes a partir de dados comportamentais e linguísticos extraídos de mídias sociais, indo além das abordagens estatísticas tradicionais. O método proposto se mostrou eficiente para a segmentação e análise visual de grupos de usuários com características psicológicas semelhantes, contribuindo para o avanço das pesquisas na interseção entre redes complexas, psicologia e ciência de dados.

6.1 Principais Contribuições

Os experimentos realizados ao longo deste trabalho permitiram validar as contribuições propostas e reforçar a relevância da abordagem desenvolvida. Por meio da aplicação de estratégias de discretização, modelagem de redes complexas e análise dos clusters gerados, foi possível identificar padrões associados aos traços de personalidade e temperamento de usuários de mídias sociais.

A utilização das redes complexas demonstrou-se eficaz na segmentação de usuários com base em similaridade de atributos psicológicos e comportamentais, resultando em uma modularidade elevada (0.939) e clusters com alta pureza em dimensões como E/I e N/S. Esses resultados evidenciam o potencial da metodologia para revelar clusters coerentes e informativos em conjuntos de dados sociais.

Além disso, os resultados obtidos permitiram identificar padrões linguísticos e comportamentais distintos entre alguns dos grupos analisados, em consonância com achados da literatura, validando a capacidade da metodologia em capturar nuances relevantes de cada traço. A análise dos clusters mais puros forneceu evidências adicionais da relação entre traços de personalidade e comportamentos observados nas redes sociais.

Dessa forma, as contribuições desta pesquisa se consolidam não apenas pela proposta metodológica, mas também pela validação experimental dos resultados, pela disseminação científica.

6.2 Contribuições em Produção Bibliográfica

Como contribuição adicional, destaca-se a publicação de um artigo derivado deste trabalho, aceito e apresentado no evento ASONAM[1], intitulado "*Analysis of User Temperament and Personality Traits in Social Media through Complex Networks*", reforçando a originalidade e relevância da pesquisa. Ademais, o código-fonte utili-

[1] <https://asonam.cpsc.ucalgary.ca/2025/>

zado nos experimentos foi disponibilizado publicamente, promovendo a reprodutibilidade e o avanço de estudos futuros [2].

6.3 Trabalhos Futuros

A partir das análises e limitações observadas neste estudo, diversas direções podem ser exploradas em pesquisas futuras:

- **Exploração de diferentes modelos de personalidade ou temperamento:** Este trabalho utilizou a tipologia MBTI como referência. Investigações futuras podem aplicar a abordagem proposta utilizando outros modelos consagrados, como o Big Five ou o TEMPS, possibilitando comparações entre metodologias e perfis traçados.
- **Ampliação do conjunto de dados:** Uma limitação enfrentada foi a escassez de bases que integrem traços de personalidade e dados sociais. Trabalhos futuros podem focar na construção ou curadoria de novas bases de dados que combinem textos, atributos sociais e informações psicológicas mais completas e diversas.
- **Adoção de novas estratégias de discretização de variáveis:** A aplicação de outras técnicas de discretização pode revelar diferentes padrões de comportamento e oferecer novos insights sobre a relação entre características dos usuários e o uso das mídias sociais.
- **Refinamento da modelagem de redes:** A geração das redes neste estudo considerou apenas usuários com distância Euclidiana igual a zero (características idênticas após discretização). Novas pesquisas podem explorar diferentes critérios de similaridade e limiares, utilizando pesos nas arestas para refletir diferentes graus de proximidade entre os usuários.
- **Incorporação de técnicas de aprendizado de máquina:** A partir das comunidades detectadas nas redes, é possível treinar modelos supervisionados ou não supervisionados para classificar ou predizer traços de personalidade com base em padrões topológicos, contribuindo com abordagens híbridas entre redes complexas e inteligência artificial.

[2] Código fonte disponível em: <https://github.com/mtshnq/asonam2025>

Referências

AIKEN, L. R. **Psychological testing and assessment**. [S.l.]: Pearson Education India, 2009.

AKISKAL, H. S.; AKISKAL, K. K. Temps: Temperament evaluation of memphis, pisa, paris and san diego. **Journal of affective disorders**, v. 85, n. 1-2, 2005.

ALCANTARA, C. de et al. Análise do temperamento e percepção do suporte social em redes sociais online. **iSys-Brazilian Journal of Information Systems**, v. 17, n. 1, p. 11–1, 2024.

ALMEIDA, L. D.; GOYA, D. Detection of big five personality traits in twitter user's profiles based on textual posts. In: SBC. **Anais do XX Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2023. p. 227–241.

AQUINO, B.; STROELE, V.; SOUZA, J. Análise do engajamento dos alunos em ambientes virtuais de aprendizagem para detecção de comunidade. In: SBC. **Simpósio Brasileiro de Informática na Educação (SBIE)**. [S.l.], 2020. p. 952–961.

BANERJEE, A. et al. Synthetic minority oversampling in addressing imbalanced sarcasm detection in social media. **Multimedia Tools and Applications**, Springer, v. 79, n. 47, p. 35995–36031, 2020.

BARABáSI, A.-L.; PóSFAI, M. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266 1107076269. Disponível em: <http://barabasi.com/networksciencebook/>.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. **Gephi: An Open Source Software for Exploring and Manipulating Networks**. 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

BOURDIEU, P. A dominação masculina. In: **A dominação masculina**. [S.l.: s.n.], 2010. p. 158–158.

BRITO, A. C. M.; SILVA, F. N.; AMANCIO, D. R. A complex network approach to political analysis: Application to the brazilian chamber of deputies. **Plos one**, Public Library of Science San Francisco, CA USA, v. 15, n. 3, p. e0229928, 2020.

BUSSAB, W. d. O.; MORETTIN, P. A. Estatística básica. In: **Estatística básica**. [S.l.: s.n.], 2010. p. xvi–540.

CARNEIRO, M. G. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. Dissertação (Tese de Doutorado) — Universidade de São Paulo, 2017.

CLARO, C. F. **Um estudo de caso sobre o modelo de temperamento de Keirsey**. Dissertação (Dissertação de Mestrado) — Universidade Presbiteriana Mackenzie, 2018.

COMIN, C.; F, C. L. da. Spatial networks: When topology meets geometry (cdt-3). **ResearchGate**, 2018.

CROMITY, J. The impact of social media in review. **New Review of Information Networking**, Taylor & Francis, v. 17, n. 1, p. 22–33, 2012.

DEMBIŃSKA-KRAJEWSKA, D.; RYBAKOWSKI, J. The temperament evaluation of memphis, pisa and san diego autoquestionnaire (temps-a)–an important tool to study affective temperaments. **Psychiatr Pol**, v. 48, n. 2, p. 261–76, 2014.

FÉLIX, L. G. et al. A social network analysis of football with complex networks. In: SBC. **Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)**. [S.l.], 2019. p. 47–50.

FILHO, J. F. R. et al. Características da personalidade de estudantes de ciências contábeis: Análise do conhecimento baseado no modelo myers-briggs type indicator (mbti). **Contabilidade Gestão e Governança**, v. 13, n. 2, 2010.

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

GAUTHIER, T. D. et al. **Introduction to environmental forensics**. [S.l.]: Elsevier Academic Press, 2007.

GUL, H. et al. A systematic analysis of community detection in complex networks. **Procedia Computer Science**, Elsevier, v. 201, p. 343–350, 2022.

ITO, P. d. C. P.; GUZZO, R. S. L. Diferenças individuais: temperamento e personalidade; importância da teoria. **Estudos de Psicologia (Campinas)**, SciELO Brasil, v. 19, p. 91–100, 2002.

JUE, A. L.; MARR, J. A.; KASSOTAKIS, M. E. **Mídias sociais nas empresas**. [S.l.]: Editora Évora, 2010.

JUNG, C. G. **Psychological Types**: [tipos psicológicos]. [S.l.: s.n.], 1923. v. 6.

KAUR, H.; PANNU, H. S.; MALHI, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 52, n. 4, p. 1–36, 2019.

KEIRSEY, D. Please understand me ii: Temperament, character, intelligence. **(No Title)**, 1998.

KRAEPELIN, E. **Manic-depressive insanity and paranoia**. [S.l.]: E. & S. Livingstone, 1921.

KRETSCHMER, E. **Psychique und character**. London: Kegan, Paul, Trench, Trubner and Co. Ltd, 1936.

LIMA, A. C. E.; CASTRO, L. N. D. A multi-label, semi-supervised classification approach applied to personality prediction in social media. **Neural Networks**, Elsevier, v. 58, p. 122–130, 2014.

_____. Tecla: A temperament and psychological type prediction framework from twitter data. **Plos one**, Public Library of Science San Francisco, CA USA, v. 14, n. 3, p. e0212844, 2019.

LINHARES, C. D. et al. Visual analysis for evaluation of community detection algorithms. **Multimedia Tools and Applications**, Springer, v. 79, p. 17645–17667, 2020.

LOPES, F. M. **Redes complexas de expressão gênica: síntese, identificação, análise e aplicações**. Dissertação (Tese de Doutorado) — Universidade de São Paulo, 2011.

MARTINS, L. M. **Análises de publicações da rede social Instagram em contraste com o temperamento human**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Uberlândia, 2022.

MARTINS-SUAREZ, F. C.; SOUSA, J. M. M. Homem razão e mulher emoção: Uma análise da relação dicotômica entre homens e mulheres na visão dos assentados (as). **Ambivalências**, v. 4, n. 7, p. 288–308, 2016.

MERENDA, J. V. B. d. S. **Reconhecimento de padrões em redes complexas usando caminhadas determinísticas do turista**. Dissertação (Tese de Doutorado) — Universidade de São Paulo, 2023.

METZ, J. et al. Redes complexas: conceitos e aplicações. 2007.

MOLLER, V. O. **Análise das redes de crimes de lavagem de dinheiro**. Tese (Doutorado) — Universidade de São Paulo, 2024.

MYERS, I. B. **A guide to the development and use of the Myers-Briggs type indicator: Manual**. [S.l.]: Consulting Psychologists Press, 1985.

NEWMAN, M. **Networks**. [S.l.]: Oxford university press, 2018.

NEWMAN, M. E. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. **arXiv preprint arXiv:1606.02319**, 2016.

PASQUALI, L. Os tipos humanos: a teoria da personalidade. **differences**, v. 7, p. 359–378, 2000.

PIRES, R. F. A. d. S. **Gestão de dados não estruturados na era da transformação digital: Práticas e impactos na eficiência organizacional**. Dissertação (Mestrado), 2024.

PLANK, B.; HOVY, D. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In: **Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis**. [S.l.: s.n.], 2015. p. 92–98.

RAI, S. **Twitter MBTI Personality Types Dataset**. Kaggle, 2022. Disponível em: <https://www.kaggle.com/ds/701530>.

RIHMER, Z. et al. Current research on affective temperaments. **Current opinion in psychiatry**, LWW, v. 23, n. 1, p. 12–18, 2010.

SAIF, H.; HE, Y.; ALANI, H. Semantic sentiment analysis of twitter. In: SPRINGER. **The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11**. [S.l.], 2012. p. 508–524.

SILVA, T. C. **Machine learning in complex networks: modeling, analysis, and applications**. Dissertação (Tese de Doutorado) — Universidade de São Paulo, 2012.

STEEN, M. V. Graph theory and complex networks. **An introduction**, v. 144, p. 1–287, 2010.

TABASSUM, S. et al. Social network analytics and visualization: Dynamic topic-based influence analysis in evolving micro-blogs. **Expert Systems**, Wiley Online Library, v. 40, n. 5, p. e13195, 2023.

_____. Social network analysis: An overview. **WIREs Data Mining and Knowledge Discovery**, v. 8, n. 5, p. e1256, 2018. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1256>.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. **Journal of language and social psychology**, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010.

UCHIYAMA, A. I. **Predição de sucesso de livros por meio de uma abordagem de redes complexas**. 2022. Trabalho de Conclusão de Curso (MBA) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. Disponível em: <https://bdta.abcd.usp.br/directbitstream/d4a805ef-1d98-4e69-a013-e0011e8502cd/Alana%20Uchiyama_TCC_MBA_Alana_Uchiyama_206878.pdf>.

VALLANDER, S. The enneagram and myers-briggs within a neuroscientific framework. OSF, 2023.

VERHOEVEN, B.; DAELEMANS, W.; PLANK, B. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In: **Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)**. [S.l.: s.n.], 2016. p. 1632–1637.

VOLPI, J. H. Particularidades sobre o temperamento, a personalidade e o caráter, segundo a psicologia corporal. **Centro Reichiano**, 2004.

WOODRUFF, E. et al. Validation of the brazilian brief version of the temperament auto-questionnaire temps-a: the brief temps-rio de janeiro. **Journal of affective disorders**, Elsevier, v. 134, n. 1-3, p. 65–76, 2011.

YANG, Z.; ALGESHEIMER, R.; TESSONE, C. J. A comparative analysis of community detection algorithms on artificial networks. **Scientific reports**, Nature Publishing Group UK London, v. 6, n. 1, p. 30750, 2016.

YIN, C. et al. A method for community detection of complex networks based on hierarchical clustering. **International Journal of Distributed Sensor Networks**, SAGE Publications Sage UK: London, England, v. 11, n. 6, p. 849140, 2015.

ZAFARANI, R.; ABBASI, M. A.; LIU, H. **Social Media Mining: An Introduction**. New York, NY, USA: Cambridge University Press, 2014. ISBN 1107018854, 9781107018853.