Evaluation of Thresholding in the Generation of Textual Networks

Avaliação da Limiarização na Geração de Redes Textuais

João Pedro F. O. Nascimento[1], **Anderson C. S. Oliveira**[1] **Lia H. M. Morita**[1]

[1] Departamento de Estatística – Universidade Federal do Mato Grosso (UFMT)

jpolnasc@gmail.com, anderson.oliveira@ufmt.br, lia.morita@ufmt.br

Abstract. *Textual data can be modeled into graphs using similarity matrices. However, these dense matrices require sparsification, and thresholding is often employed for this purpose. In this study, simulations were conducted to assess the influence of threshold selection on the quality of communities identified by the Leiden algorithm, using the V Measure as the metric. The results showed an increase in the V Measure as the threshold varied until reaching an asymptote. In certain scenarios, an inflection was observed. It was identified that texts with low lexical diversity exhibit asymptotic behavior, suggesting an association with the observed pattern.*
Keywords: *thresholding, textual networks, leiden algorithm, community detection, v measure*

Resumo. *Dados textuais podem ser modelados em grafos utilizando matrizes de similaridade. No entanto, essas matrizes densas exigem esparsificação, e a limiarização é frequentemente usada para este propósito. Neste estudo, simulações foram realizadas para avaliar a influência da escolha do limiar na qualidade das comunidades identificadas pelo algoritmo de Leiden, com a Medida V como métrica. Os resultados mostraram aumento da Medida V conforme variação do limiar até atingir uma assíntota. Em certos cenários, observou-se uma inflexão. Foi identificado que textos de baixa diversidade léxica exibem comportamento assintótico, sugerindo uma associação com o padrão observado.*
Palavras-chave: *limiarização, redes textuais, algoritmo de leiden, detecção de comunidades, medida v*

1. Introdução

Em um cenário cada vez mais saturado de informações, a análise de textos e a detecção de comunidades emergem como áreas de pesquisa cruciais. A habilidade de entender como os dados textuais estão organizados, como se relacionam e como podem ser agrupados em comunidades distintas é fundamental para explorar o vasto repositório de informações à nossa disposição. Essa compreensão não apenas auxilia na extração de conhecimento significativo, mas também desempenha um papel vital em diversas aplicações, desde sistemas de recomendação de conteúdo até a detecção de notícias falsas e análises de sentimentos [Jurafsky and Martin 2009, Liu et al. 2020].

No contexto da representação de grafos, várias abordagens têm sido empregadas, incluindo listas de adjacências, matrizes de incidência e matrizes de adjacências [Bapat 2014]. Para dados textuais, as matrizes de adjacências são frequentemente derivadas da filtração de matrizes de correlação, onde a escolha do limiar é de extrema relevância. No entanto, essa estratégia enfrenta críticas relacionadas à sua capacidade limitada de eliminar correlações espúrias, à natureza arbitrária na escolha do limiar e à possível perda de informações durante o processo de limiarização [Kojaku and Masuda 2019].

Assim, o objetivo central deste estudo é investigar a influência do processo de limiarização na criação de matrizes de adjacências para dados textuais e compreender o impacto dessa estratégia na qualidade das comunidades identificadas. Para essa avaliação, a "Medida V" é empregada como métrica, uma ferramenta consolidada que considera tanto a homogeneidade quanto a completude dos clusters.

2. Materiais e Métodos

2.1. Dados

Utilizou-se o conjunto de dados denominado *The 20 Newsgroups*. Este conjunto é composto por documentos oriundos de grupos de notícias da Usenet, coletados e categorizados especificamente para propósitos de pesquisa e desenvolvimento. Compreende aproximadamente 18.000 mensagens, as quais estão distribuídas em 20 categorias distintas, determinadas pelo conteúdo textual.

Para os propósitos deste estudo, as 20 categorias originais foram reorganizadas em cinco tópicos principais: Ciência, Esportes, Computadores, Religião e Veículos. Mensagens que não se adequavam a nenhuma dessas categorias específicas foram eliminadas da análise.

2.2. Ferramentas de Análise

Para todos os procedimentos de análise dos textos, foram utilizados Python [Python Core Team 2019] e R [R Core Team 2022].

2.2.1. Pré-Processamento

O conjunto de mensagens do *The 20 Newsgroups* foi submetido a uma série de etapas de pré-processamento para garantir a qualidade e a homogeneidade dos dados. As etapas incluíram:

- Remoção de espaços em branco no início e no final dos textos, bem como a exclusão de novas linhas e tabulações.
- Eliminação de caracteres especiais, simplificando a representação textual e mitigando potenciais inconsistências.
- Conversão de todos os caracteres alfabéticos para o formato minúsculo, assegurando uma análise padronizada e garantindo a consistência na identificação de palavras, independentemente da variação entre maiúsculas e minúsculas.

Na etapa de vetorização dos textos, utilizou-se a API de embeddings da OpenAI [OpenAI 2023]. Esta ferramenta converte palavras e frases em vetores numéricos contínuos de alta dimensão (1536 dimensões), refletindo o significado semântico e as relações gramaticais intrínsecas ao texto. O modelo adotado para esta tarefa foi o *text-embedding-ada-002*, reconhecido por sua proficiência em discernir nuances intrincadas da linguagem.

2.2.2. Estudo de Simulação

O estudo de simulação foi estruturado em diversas etapas, repetidas em 100 iterações para cada tópico. As etapas sequenciais foram:

1. **Amostragem:** Em cada tópico, adotou-se uma amostragem estratificada, com cada tópico constituindo um estrato de tamanho fixo de $n = 100$. A título de exemplo, o tópico "Ciência" conta com 4 categorias, o que resulta numa amostra com 400 elementos.
2. **Geração da matriz de correlação:** Com base nos dados amostrados, elaborou-se uma matriz de similaridades, utilizando a correlação de Pearson. Este método fornece um coeficiente que expressa o grau de relação linear entre os vetores. Cada coeficiente nesta matriz indica a correlação entre dois textos específicos da amostra.
3. **Esparsificação da matriz de correlação:** O Método da Limiarização foi empregado para esparsificar a matriz de correlação. Este processo visa acentuar as relações mais significativas e atenuar a complexidade dos dados. Os percentis P dos valores da matriz de correlação foram testados, variando de 1 a 99, para cada tópico.
4. **Detecção de comunidades:** A partir da matriz de correlação esparsificada, procedeu-se à detecção de comunidades. Para cada grafo gerado em cada iteração, utilizou-se o algoritmo Leiden [Traag et al. 2019] com os parâmetros:
    - `objective_function="modularity"`
    - `weights=None`
    - `resolution=1.0`
    - `beta=0.01`
    - `initial_membership=None`
    - `node_weights=None`
5. **Avaliação das comunidades:** A escolha da Medida V como métrica de avaliação deve-se à sua capacidade de combinar dois aspectos fundamentais da clusterização: homogeneidade e completude.
    - *Homogeneidade:* Refere-se a quão puros são os clusters em relação às classes originais. Um cluster é considerado homogêneo se ele contém apenas dados que são membros de uma única classe original.
    - *Completude:* Mede se todos os dados que são membros de uma determinada classe original estão contidos no mesmo cluster.

    Além da Medida V, o número de comunidades detectadas também foi levado em consideração. Isso ajuda a garantir que o algoritmo não estava dividindo excessivamente (sobre-segmentando) ou agrupando de forma muito ampla (sub-segmentando) os dados. O objetivo ao combinar a Medida V com o número de

comunidades é obter uma representação que seja ao mesmo tempo precisa em termos de composição das comunidades e informativa em termos do número de comunidades. [Rosenberg and Hirschberg 2007].

Os resultados das simulações são apresentados por meio de diagramas de dispersão da métrica em função do percentil limiar. Em cada gráfico, é exibida a mediana das métricas em relação ao limiar que serve como referência e a partir dela, determina-se o percentil (limiar) que maximiza a Medida V.

Para avaliar a relação entre o percentil e a Medida V, também foram obtidos índices de diversidade lexical dos textos [Mccarthy and Jarvis 2007]. A diversidade lexical fornece uma perspectiva quantitativa sobre a riqueza e variedade do vocabulário nos textos, indicando a singularidade e distinção de cada comunidade, e ajudando a identificar possíveis sobreposições ou sobre-segmentações.

Concluindo a análise, é essencial não apenas considerar a estrutura matemática das comunidades, mas também seu significado e relevância prática. Por isso, as comunidades detectadas foram avaliadas quanto à sua coesão interna e sua relevância para o contexto geral do estudo, assegurando que cada uma delas possui uma contribuição significativa para a interpretação e compreensão do conjunto de dados.

3. Resultados e Discussão

3.1. Avaliação das Comunidades

Para exemplificar a avaliação das comunidades realizadas, foram selecionados apenas os tópicos de Ciência e Religião.

Na Figura 1, são apresentados gráficos que ilustram a medida V e o número de comunidades em relação ao percentil (limiar) para o tópico em ciência. Observa-se um aumento linear na medida V à medida que o percentil aumenta, indicando uma associação direta entre o aumento do limiar e o desempenho da métrica V. No entanto, essa tendência linear parece atingir um ponto de inflexão para percentis mais elevados, sugerindo uma possível saturação dos benefícios de aumentar o limiar.

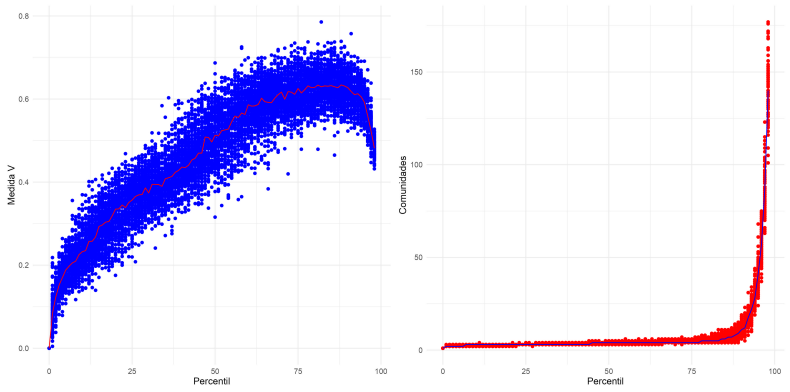

Figura 1. Medida V e número de comunidades em função do percentil (limiar) para o tópico ciência.

O limiar que otimiza a Medida V, representado pelo percentil 88, destaca-se ao apresentar uma mediana de 0.64 e uma mediana de 8 comunidades. Esse ponto de máximo

representa um equilíbrio entre a qualidade das comunidades identificadas e a quantidade delas.

Adicionalmente, no decorrer da análise deste tópico, foi realizado o cálculo do índice de diversidade lexical dos textos vinculados ao tema de Ciência, resultando em um valor total de 103,75. Este índice ressalta que os textos exibem uma notável diversidade lexical, enriquecendo assim o vocabulário abordado. Tal diversidade desempenha um papel fundamental na compreensão da amplitude de tópicos e conceitos contemplados neste contexto.

É importante notar que esse comportamento, com aumento linear seguido de uma possível saturação, também foi observado em outros tópicos, como Política, Esportes e Veículos. Essas descobertas destacam a importância de escolher cuidadosamente o limiar ao realizar análises de detecção de comunidades, pois ele desempenha um papel crucial na qualidade e na quantidade das comunidades identificadas.

O comportamento da Medida V em relação ao limiar para o tópico de Religião (Figura 2) apresenta uma dinâmica que pode ser dividida em duas fases distintas. Inicialmente, à medida que o limiar aumenta, ocorre um crescimento acentuado na Medida V, indicando uma melhoria na precisão da detecção de comunidades e na coesão dos grupos identificados nos dados. No entanto, essa melhoria parece estabilizar após atingir um determinado ponto, criando uma curva de crescimento que se assemelha a uma assíntota.

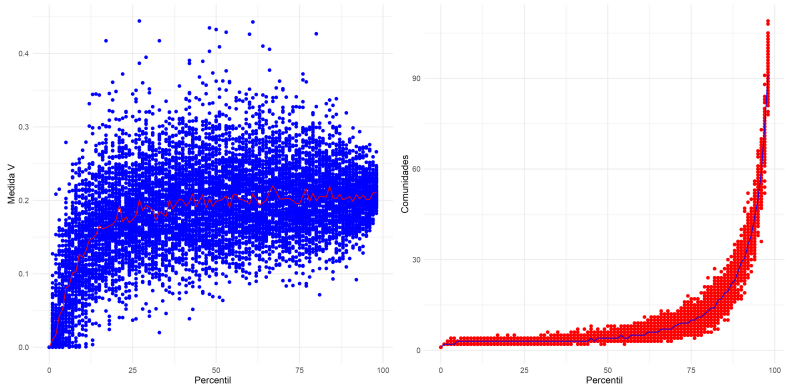

Figura 2. Medida V e número de comunidades em função do percentil (limiar) para o tópico Religião.

O limiar que maximiza a Medida V, representado pelo percentil 84, destaca-se ao apresentar uma mediana de 0.22 e uma mediana de 17 comunidades, o que indica uma divisão significativa dos dados em grupos distintos. Essa combinação de alta coesão e diversidade de comunidades é essencial para uma análise rica e abrangente dos tópicos relacionados à religião.

O índice de diversidade lexical dos textos associados ao tópico de Religião resultou em um valor de 83.85. Esse valor representa uma notável diversidade lexical, embora seja inferior ao observado no tópico de Ciência. Essa diferença sugere que, ao escrever sobre religião, há um uso menor de palavras distintas em comparação com o tópico de Ciência. Essa variação na diversidade lexical pode influenciar o comportamento de assintota observado na medida V.

É interessante notar que esse mesmo padrão de comportamento, com um crescimento inicial rápido seguido de estabilização, também foi observado no tópico relacionado a Computadores. Essas descobertas sugerem que a seleção criteriosa do limiar desempenha um papel crucial na qualidade e quantidade das comunidades identificadas, não apenas para esse tema específico, mas também em outras áreas de estudo, como a informática.

3.2. Obtenção de Comunidades

Na Figura 3, é apresentado o histograma das correlações dentro do tópico de Ciência. Nesse gráfico, observa-se uma queda acentuada nos valores de correlação, indicando um efeito de cauda longa, como representado no gráfico de dispersão log-log, que exibe a correlação em função da frequência. A observação de uma relação linear nesse gráfico de dispersão sugere que a distribuição de correlações segue uma lei de potências. Essa lei de potências na distribuição de correlações pode refletir uma estrutura subjacente relacionada dentro do tópico Ciência.

A importância dessa observação reside na compreensão da organização e interconexão dos textos dentro do tópico da Ciência. A presença de uma lei de potências na distribuição de correlações sugere que alguns conceitos têm correlações mais fortes, desempenhando um papel central, enquanto outros têm correlações mais fracas. Isso pode indicar uma hierarquia ou uma estrutura subjacente que influencia a dinâmica da discussão científica.

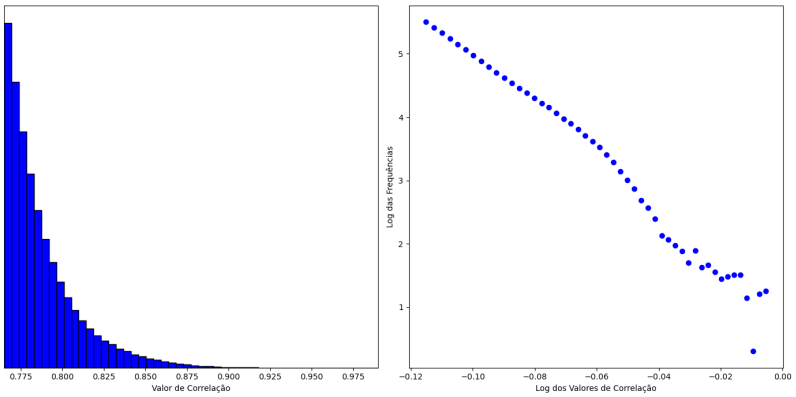

Figura 3. Histograma da Matriz de Correlação e Gráfico de Dispersão Log-Log.

É importante notar que o tópico Ciência é composto por quatro categorias distintas:

- `sci.electronics` - textos relacionados à eletrônica,
- `sci.crypt` - textos relacionados à criptografia,
- `sci.med` - textos relacionados à medicina e saúde,
- `sci.space` - textos relacionados à exploração espacial e astronomia.

Essas categorias, por sua vez, podem conter subcategorias e nuances específicas. O método utilizado na análise é capaz de identificar e mapear essas complexidades dentro do amplo tópico de Ciência, proporcionando uma visão mais detalhada e estruturada da discussão científica.

Nas análises conduzidas, as comunidades identificadas no grafo após a aplicação do algoritmo de Leiden são apresentadas na Figura 4. Foram identificadas cinco comunidades distintas. A Comunidade 1 é predominantemente composta por textos da categoria `sci.electronics` (86%), totalizando 934 nós e apresentando um índice de diversidade lexical de 105.7. A Comunidade 2, ligada à categoria `sci.med` (88%), conta com 996 nós e um índice de diversidade lexical de 106.7. A Comunidade 3 está associada à categoria `sci.crypt` (87%), possuindo 977 nós e um índice de 97.7. A Comunidade 4 corresponde à categoria `sci.space` (90%), compreendendo 918 nós e um índice de diversidade lexical de 105.3.

Por fim, a Comunidade 5, também vinculada à categoria `sci.space`, contém apenas um nó e possui um índice de diversidade lexical de 2.0. É interessante notar a presença dessa comunidade com apenas um nó, associada à categoria `sci.space`. Esse fenômeno pode ser explicado por diversos fatores, como a alta dimensionalidade dos embeddings (1536), que pode ter gerado uma maior distância entre o vetor desse texto e os demais; a falta de contexto, visto que o texto é composto por apenas duas palavras (“sherzer methodology”), o que pode tê-lo tornado um outlier; a influência na matriz de correlação, onde a singularidade do vetor desse texto pode ter levado a correlações baixas; e a sensibilidade do algoritmo de Leiden, que pode ter identificado esse texto como um outlier e, consequentemente, criado uma comunidade separada para ele.

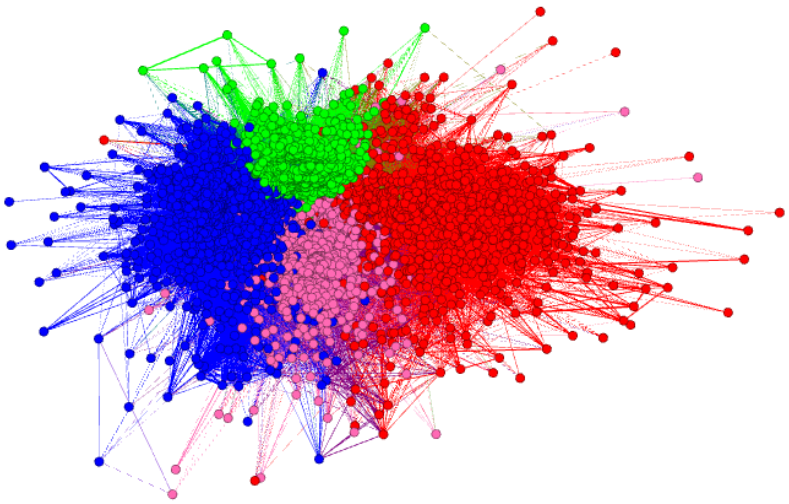

Figura 4. Comunidades identificadas pelo algoritmo de Leiden no tópico Ciência.

Na Figura 5, uma representação dos textos das categorias que compõem o tópico da Ciência é apresentada. Esses textos abrangem as quatro categorias principais: `sci.electronics`, `sci.crypt`, `sci.med` e `sci.space`. O que se destaca ao observar essa representação é a notável semelhança entre as comunidades identificadas aqui e aquelas reveladas pelo algoritmo de Leiden na Figura 4. Isso mostra que as comunidades geradas estão em sintonia com os tópicos originais, sugerindo que o espaço semântico representado pelos textos foi capturado de maneira adequada.

Os índices de homogeneidade e completude entre as comunidades e as categorias originais resultaram em valores de 0.6481 e 0.6474, respectivamente. Esses valores indicam uma correspondência sólida entre as comunidades identificadas e as categorias

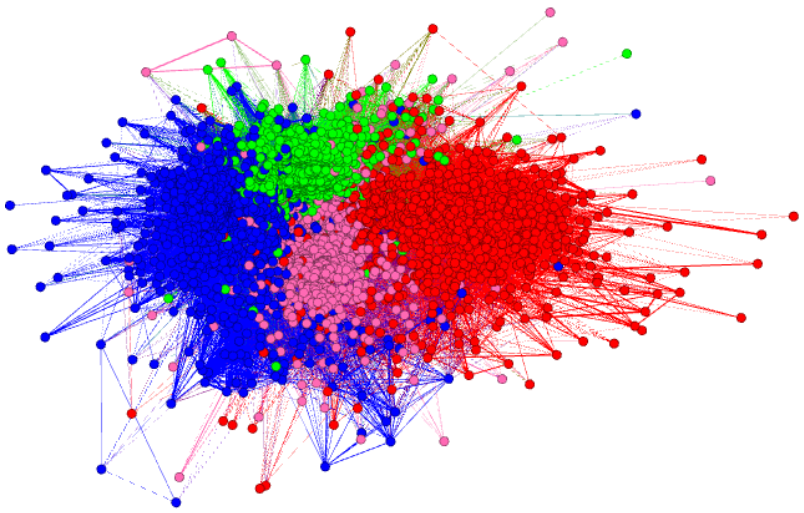

Figura 5. Representação dos textos das categorias que compõem o tópico da Ciência.

originais. Essa correspondência robusta reforça a qualidade da segmentação realizada e a precisão com que as comunidades representam os tópicos originais dentro do amplo tópico da Ciência.

A afinidade entre as comunidades e os tópicos reais é ainda mais evidenciada ao comparar os índices de diversidade lexical, como demonstrado na Tabela 1.

Categoria	Índice da Comunidade	Índice Real
`sci.electronics`	105.7	103.5
`sci.med`	97.7	98.8
`sci.crypt`	97.7	98.8
`sci.space`	105.3	106.2

Tabela 1. Comparação entre a diversidade léxica das comunidades e as categorias reais.

Essa sólida correspondência entre as comunidades e os tópicos originais dentro do tópico Ciência sugere que a técnica utilizada foi capaz de capturar eficazmente a estrutura subjacente aos textos. Resultados semelhantes foram obtidos para os demais tópicos.

4. Considerações Finais

Os resultados derivados da análise do impacto do limiar nas comunidades revelaram dois comportamentos distintos. Primeiramente, notou-se um aumento gradual da Medida V à medida que o percentil aumentava, atingindo um ponto de assíntota. Em segundo lugar, foi observada uma elevação seguida de uma inflexão na mesma medida. Curiosamente, os temas que apresentaram comportamento de assíntota também exibiram os menores valores de diversidade léxica. Isso implica uma possível relação entre a diversidade léxica e a dinâmica da detecção de comunidades, destacando a relevância desse fator na análise de tópicos complexos.

No que diz respeito à detecção de comunidades, os resultados foram particularmente notáveis no contexto do tópico Ciência. Foi possível estabelecer uma cor-

respondência sólida entre as comunidades identificadas e as categorias originais, tanto através das métricas de qualidade quanto na comparação dos índices de diversidade lexical. Isso reforça a capacidade do método utilizado em capturar efetivamente as estruturas subjacentes aos textos e representá-las de maneira coerente.

É relevante destacar a presença de comunidades que continham apenas um nó em determinados casos, como observado no tópico Ciência. Isso pode ser atribuído à esparsidade dos vetores de palavras, à falta de contexto nesses vetores específicos e à sensibilidade do método de Leiden. Essa observação realça os desafios e complexidades inerentes à análise de redes e à seleção adequada de métodos e parâmetros.

A natureza dos dados, especialmente em áreas tão diversificadas e complexas como a ciência, torna a tarefa de detecção de comunidades extremamente desafiadora. A escolha do limiar adequado é essencial para obter uma segmentação significativa, equilibrando a captura de nuances sutis e a evitação de segmentações excessivamente granulares. O estudo também realça a importância de se ter um entendimento profundo dos métodos empregados e do contexto dos dados sob análise. Olhando para o futuro, seria interessante investigar a aplicabilidade da metodologia em outros conjuntos de dados e contextos, bem como explorar abordagens complementares para enriquecer a análise.

Referências

Bapat, R. B. (2014). *Graphs and Matrices*. Springer-Verlag, London, 2nd edition.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Kojaku, S. and Masuda, N. (2019). Constructing networks by filtering correlation matrices: a null model approach. *Proc. R. Soc. A*, 475:20190578.

Liu, X., Jiang, S., Sun, M., and Chi, X. (2020). Examining patterns of information exchange and social support in a web-based health community: Exponential random graph models. *J Med Internet Res*, 22(9):e18062.

Mccarthy, P. and Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. language testing, 24, 459-488. *Language Testing - LANG TEST*, 24:459–488.

OpenAI (2023). *Manual da API de Embeddings da OpenAI*. OpenAI Incorporated.

Python Core Team (2019). *Python: A dynamic, open source programming language*. Python Software Foundation.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.