UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Exatas

Programa de Pós-Graduação em Estatística

César Macieira

Estimação em modelos de redes de afinidade

Belo Horizonte

2021

César Macieira

Estimação em modelos de redes de afinidade

Dissertação apresentada ao Programa de Pós Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Mestre em Estatística.

Orientador: Prof. Denise Duarte Scarpa Magalhães Alves

Belo Horizonte

2021

Macieira, César

M152e Estimação em modelos de redes de afinidade [manuscrito] / César Macieira. — Belo Horizonte, 2021.
50 f. : il. ; 29cm

Orientador: Denise Duarte Scarpa Magalhães Alves.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 50

1. Estatística – Teses. 2. Grafos aleatórios - Teses. 3. Redes sociais on-line – Teses. 4. Modelos de redes de afinidade-Teses. I. Alves, Denise Duarte Scarpa Magalhães. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz - CRB 6ª Região nº 819.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ATA DA DEFESA DE DISSERTAÇÃO DE MESTRADO DO ALUNO César Macieira, MATRICULADO, SOB O Nº 2019.663.605, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 30 DE ABRIL DE 2021.

Aos 30 dias do mês de abril de 2021, às 14h30, em reunião pública virtual 258 (conforme orientações para a atividade de defesa de dissertação durante a vigência da Portaria PRPG nº 1819) na sala https://us02web.zoom.us/j/84142954905, do Instituto de Ciências Exatas da UFMG, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de dissertação do aluno César Macieira, nº matrícula 2018.663.605, intitulada: "*Estimação em modelos de redes de afinidade* ", requisito final para obtenção do Grau de mestre em Estatística. Abrindo a sessão, a Senhora Presidente da Comissão, Profa. Denise Duarte Scarpa Magalhaes Alves (DEST/UFMG), passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

Profa. Denise Duarte Scarpa Magalhaes Alves
Orientadora – (EST/UFMG)

Prof. Cristiano de Carvalho Santos
(DEST/UFMG)

Profa. Andressa Cerqueira
(DEs- UFSCar)

O resultado final foi comunicado publicamente ao aluno pela Senhora Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 30 de ABRIL de 2021.

Observações:
1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

Resumo

A análise de redes sociais tem como finalidade detectar e mensurar relações entre elementos nos mais diversos setores da sociedade. Por isso, inúmeros métodos são empregados para esse tipo de análise, dentre os quais destacamos o *modelo de redes de afinidade*. Esse modelo de grafos aleatórios considera que a probabilidade de ligação é função de características dos vértices e isso permite uma aproximação maior do modelo a redes reais.

O presente trabalho se propõe a produzir uma metodologia para realizar a estimação dos parâmetros que compõem o *modelo de redes de afinidade*. No caso em que há uma comunidade apenas, utilizamos a função de verossimilhança diretamente para encontrar os estimadores, enquanto que no caso de haver mais de uma comunidade, construímos um algoritmo do tipo Maximização da Expectativa (Expectation Maximization), uma vez que teremos um modelo latente. Certificaremos se o método funciona bem por intermédio de simulações. Propomos uma forma de determinar o número adequado de comunidades da rede de afinidade e de fazer a alocação dos indivíduos a elas. Construiremos grafos para representar as redes de afinidade utilizando duas maneiras diferentes de medir a afinidade entre os indivíduos da rede. Finalmente, realizaremos a aplicação da metodologia apresentada em uma situação real, em um banco de dados coletado com o objetivo de descrever a forma que as pessoas entrevistadas tiveram a vida impactada pela pandemia de Covid-19.

Palavras-chave: Modelos de redes de afinidade, Afinidade, Algoritmo EM, Grafo aleatório

Abstract

The analysis of social networks aims to detect and measure relationships between elements in the most diverse sectors of society. Therefore, numerous methods are used for this type of analysis, among which we highlight the *affinity networks model*. This model of random graphs considers that the probability of connection is a function of characteristics of features and this allows a better approximation of the model to real networks.

The present work proposes to produce a methodology to perform the estimation of the parameters that compose the *model of affinity networks*. In case there is only one community, we use the likelihood function directly to find the estimators, in case there is more than one community, we build an Expectation Maximization algorithm, as we have a latent model. We ensure that the method works well through simulations. We propose a method to determine the appropriate number of communities in the affinity network and allocate individuals to them. Graphs were made to represent affinity networks using two different methods to measure affinity between individuals in the network. Finally, we applied the methodology presented in a real situation, in a data collection in order to describe how the people interviewed had their lives impacted by the Covid-19 pandemic.

Keywords: Affinity Network Models, Affinity, EM algorithm, Random Graph

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A Análise de Redes Sociais (Martino, 2006) é um tema multidisciplinar que contempla as áreas da Sociologia, Psicologia Social e Antropologia (Freeman, 1996), cujo objetivo principal é identificar relações entre pessoas ou organizações de uma rede. Consequentemente, diversas metodologias têm sido utilizadas para mensurar e estabelecer as conexões, dos quais pode-se destacar o modelo em grafos aleatórios.

Um modelo de grafos aleatórios representa a relação entre os elementos acerca de um determinado tema, sendo obtido a partir de um conjunto de vértices e arestas que realizam ligações de forma aleatória. Muitos modelos de grafos utilizam a suposição de que as relações entre elementos são feitas de forma independente, entretanto para a análise de redes sociais este pressuposto é difícil de ser validado. O modelo de redes de afinidade proposto em Pereira (2020) é um método alternativo que considera semelhanças entre dois indivíduos ao estabelecer uma aresta entre eles e também quantifica a força da conexão. Em Pereira (2020) este modelo foi apresentado e estudado, obtendo uma grande família de modelos, onde as conexões são geradas segundo uma função que mensura a afinidade entre os atores da rede, denominada função afinidade. Esse novo modelo de redes permite levar em consideração informações relevantes dos indivíduos na geração das conexões, o que pode levar a uma melhor adaptação a redes reais.

Uma vez definido o modelo de redes de afinidade, assim como as funções afinidades, estudadas suas características e entendidas as suas aplicações, como foi detalhado em Pereira (2020), o próximo passo natural é realizar as estimativas dos parâmetros que compõem esse modelo. No caso em que temos apenas uma comunidade na rede, propomos um estimador para as probabilidades de escolha dos indivíduos em relação a um conjunto de características baseado na função de verossimilhança que construímos para o modelo. Já no caso onde temos mais de uma comunidade na rede e não sabemos a qual delas um indivíduo pertence, construímos um algoritmo EM (Expectation Maximization) seguindo a proposta de Dempster et al.(1977), para determinar estimativas de parâmetros em modelos de mistura ou em modelos com dados faltantes. Desta maneira, aprofundamos e desenvolvemos os trabalhos sobre modelagem de redes sociais, apresentados em Singer (1995), Pereira (2017) e Pereira (2020), onde a ênfase das conexões é colocada nas características dos indivíduos.

Este trabalho visa, portanto, estimar os parâmetros do modelo de grafos aleatórios de redes de afinidade, que serão detalhados no decorrer dessa dissertação. Validaremos o método proposto por meio de simulações. Apresentaremos também uma forma de determinar o número de comunidades adequado em uma rede de afinidades e depois disso propomos uma forma de

alocação dos indivíduos às comunidades encontradas. Construímos os grafos de afinidade entre os indivíduos alocados nas comunidades encontradas considerando duas famílias de funções afinidades apresentadas em Pereira (2020): binária e cardinal.

Por fim, aplicamos a metodologia de estimação proposta a um novo banco de dados que coletamos durante o desenvolvimento desse trabalho. O banco de dados é composto por respostas de indivíduos da comunidade da UFMG, alunos de graduação, de pós-graduação e professores, e também por pessoas da comunidade geral. O questionário tinha uma única pergunta que se referia ao que cada pessoa sentia em relação à pandemia de covid-19, que nos assolou em 2020 e continua nos castigando em 2021. Os dados foram coletados em novembro de 2020. Nosso objetivo era descrever esse sentimento individual e usá-lo para estabelecer conexões entre as pessoas a partir das palavras escolhidas por elas para representar seus sentimentos em relação à pandemia. Também buscamos identificar grupos de pessoas com sentimentos parecidos em relação ao tema estimando os pesos de cada palavra em comunidades diferentes. Nossa inspiração para a construção dessa rede de afinidade veio da Técnica de Associação Livre de Palavras, discutida em Santos (2015) e do trabalho apresentado em Duarte et al. (2020) sobre a rede cognitiva para identificar o pensamento coletivo.

2 Objetivos

Os objetivos propostos para este trabalho são:

1. Criação de um algoritmo para estimação e recuperação das medidas de probabilidades e proporções das comunidades considerando as funções afinidade binária e cardinal;

2. Alocar os indivíduos nas comunidades por meio das funções de verossimilhanças individuais;

3. Realizar simulaçõoes com o intuito de comparar as estimativas com os valores reais.

4. Aplicar a metodologia proposta para dados reais.

3 Modelo de redes de afinidade

O grafo $G = (V, E)$ é definido por um conjunto de vértices, V, e um conjunto de arestas, E, entre os vértices de V. Um grafo aleatório é gerado a partir de uma distribuição de probabilidade sobre um conjunto $\mathcal{G}$ de grafos possíveis com n vértices.

Um dos modelos de grafos aleatórios mais importantes estudados na literatura é o modelo de Erdős-Rényi (1959), que considera que as ligações entre os indivíduos acontecem de maneira independente com uma probabilidade p. No entanto, para análises de redes reais, a suposição de que as arestas entre indivíduos são independentes e identicamente distribuídas é difícil de ser validada. O modelo de redes de afinidade proposto em Pereira (2020) aproxima os modelos probabilísticos de grafos das redes reais, uma vez que as ligações entre os indivíduos são baseadas em características dos vértices (indivíduos).

Este novo modelo de grafos aleatórios atribui valor as conexões entre os indivíduos considerando uma função que quantifica a afinidade entre os elementos da rede. Por meio deste modelo conseguimos estimar a importância de certas características, encontrar subcomunidades e representar a estrutura de relações entre os indivíduos através de grafos. A partir disso, podemos encontrar as principais características de uma comunidade em relação a um determinado tema, o que permite um entendimento melhor dos acontecimentos.

3.1 Definições e notações

Seja $\mathcal{D} = \{w_1, \ldots, w_m\}$ com $m \in \mathbb{N}$ um *Dicionário* conhecido, com m atributos, como por exemplo palavras, frases, caracteres ou números. Suponha que cada indivíduo de uma população de tamanho n escolha um subconjunto de características sem repetição de acordo com uma medida de probabilidade μ. O conjunto de termos escolhidos pelo indivíduo $i, 1 \leq i \leq n$, que chamamos de *Vocabulário*, é representado pelo vetor $D_i = \{w_1^i, w_2^i, \ldots, w_{n_i}^i\}$, onde n_i é o número de informações escolhidas. Podemos representar o vetor *Vocabulário* da seguinte maneira: $U_i = \{U_i^1, \ldots, U_i^m\}$, onde

$$U_i^k = \begin{cases} 1, & \text{se } w_k \in D_i \\ 0, & \text{se } w_k \notin D_i. \end{cases} \tag{1}$$

Desse modo, a matriz $\mathbf{U}_{n\times m}$, que tem como linhas o *Vocabulário* de cada um dos n indivíduos, carrega a informação sobre todas as sentenças escolhidas por todos os indivíduos e é definida como:

$$\mathbf{U}_{N\times m} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,m-1} & u_{1,m} \\ u_{2,1} & u_{2,2} & u_{2,3} & \cdots & u_{2,m-1} & u_{2,m} \\ u_{3,1} & u_{3,2} & u_{3,3} & \cdots & u_{3,m-1} & u_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{n,1} & u_{n,2} & u_{n,3} & \cdots & u_{n,m-1} & u_{n,m} \end{bmatrix}.$$

Além das informações de cada U_i, a matriz $\mathbf{U}$ pode conter informações adicionais, como por exemplo a ordem em que as características foram escolhidas, ou seja, ao invés do conjunto de escolhas, temos $U_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,m_i}]$, um vetor no qual a ordem importa. Para obter $\mathbf{U}$, escreve-se o vetor de postos associado a U_i como $U_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m-1}, u_{i,m}]$ tal que

$$u_{i,j} = \begin{cases} \text{k, se } w_j \in D_i,\ w_j = w_{i,k} \\ 0, \text{ se } w_j \notin D_i \end{cases} \tag{2}$$

de forma que cada elemento diferente de zero em U_i indica o posto da respectiva expressão nas escolhas do i-ésimo indivíduo.

Para exemplificar, suponha uma população composta por $N = 4$ indivíduos que escolhe palavras do dicionário $\mathcal{D} = \{w_1, w_2, w_3, w_4\}$. Considere os conjuntos de escolhas $D_1 = \{w_{i,1}, w_{i,2}\}$, $D_2 = \{w_{i,1}, w_{i,3}, w_{i,2}\}$, $D_3 = \{w_{i,3}, w_{i,4}\}$ e $D_4 = \{w_{i,4}, w_{i,3}, w_{i,1}, w_{i,2}\}$. Os vetores binários associados aos conjuntos de escolhas seriam $U_1 = [1, 1, 0, 0]$, $U_2 = [1, 1, 1, 0]$, $U_3 = [0, 0, 1, 1]$ e $U_4 = [1, 1, 1, 1]$ e a matriz U com $u_{i,j}$ segundo a Equação 5.1.1 seria

$$\mathbf{U}_{4\times 4} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{3}$$

Se houvesse o interesse na ordem de evocação dos termos, para os conjuntos de escolhas U_1, U_2, U_3 e U_4, os vetores de postos vinculados seriam $U_1 = [1, 2, 0, 0]$, $U_2 = [1, 3, 2, 0]$, $U_1 = [0, 0, 1, 2]$ e $U_4 = [3, 4, 1, 2]$ e a matriz U com $u_{i,j}$ segundo a Equação 1 é da forma

$$\mathbf{U}_{5\times 5} = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 1 & 3 & 2 & 0 \\ 0 & 0 & 1 & 2 \\ 3 & 4 & 1 & 2 \end{bmatrix}. \tag{4}$$

3.2 Função afinidade

Definimos a função afinidade como qualquer função $f : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+$ simétrica (isto é, $f(U_i, U_k) = f(U_k, U_i)$, $\forall\ i \neq k$). A ideia é que esta função deve mensurar o quão semelhantes são as escolhas de dois indivíduos. O espaço de todas funções afinidade é muito grande e pode variar de funções que levam a nenhuma conexão entre os indivíduos até aquelas que levam a um grafo completo. Neste trabalho vamos nos restringir a estimar os parâmetros do modelo de afinidade considerando duas funções afinidade fixas: binária e cardinal. As descrições dessas duas funções são dadas na sequência.

3.3 Grafo de afinidade $\mathbf{G}_\lambda$

Seja f uma função afinidade, o grafo de afinidade é definido da seguinte forma:

(I) *Conjunto de vértices*: em uma rede de afinidade, cada indivíduo é representado por um vértice, de forma que o vértice $v_i \in V(G)$ com $i = \{1, 2, \ldots, n\}$ é o vértice associado ao i-ésimo elemento;

(II) *Conjunto de arestas*: o conjunto de arestas $E(G)$ é induzido pelo nível de afinidade entre os indivíduos:

$$E(G) := \{v_i \leftrightarrow v_j \Leftrightarrow f(U_i, U_j) \geq \gamma\},$$

onde $\gamma > 0$ é um ponto de corte e $\leftrightarrow$ significa ter aresta entre elementos. O ponto de corte permite escolher o nível de afinidade que seria necessário para estabelecer uma conexão entre dois indivíduos.

Para gerar um grafo de afinidade, é necessário um conjunto de especificações λ. Denotamos um grafo de afinidade por G_λ, onde $\lambda = (n, m, \mu, f, \gamma)$, n é o número de indivíduos (vértices), m é o tamanho do dicionário, μ é a distribuição de probabilidade sobre D, ou seja, a medida com a qual os indivíduos escolhem as palavras no dicionário, f é a função afinidade e γ é o ponto de corte para a função afinidade.

Além dos parâmetros, cada um dos vértices do grafo $G(\lambda)$ está associado a um vetor aleatório $U_i = [U_{i,1}, U_{i,2}, \ldots, U_{i,m-1}, U_{i,m}]$, onde cada $U_{i,j}$ é uma variável aleatória que segue uma distribuição de probabilidade $\mu_{i,j}$ que está associada à informação sobre w_j (escolha ou não-escolha de w_j, ordem de escolha de w_j dado que w_j foi escolhido, etc.) para o conjunto de características

atribuídas ao i-ésimo indivíduo. Assim, a matriz de adjacências de G_λ é da forma

$$A_{n\times n} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & \cdots & a_{3,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{a,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}. \tag{5}$$

Se $a_{ij} = 1$, os indivíduos i e j estão conectados, carregando toda a incerteza sobre G_λ.

Então para uma função *afinidade* fixa, $f(U_i, U_k) \to \mathbb{R}^+$, aplicada nas linhas da matriz $\mathbf{U}$ induz um grafo aleatório onde os vértices são os n indivíduos e as arestas são colocadas entre aqueles com grande afinidade, $f : (U_i, U_k) > \gamma$, onde γ é o parâmetro de força das conexões. Esse grafo depende, portanto, de um conjunto de parâmetros $\lambda = \big((\mu)_{i\leq n}, f, \gamma\big)$ e o denotamos por G_λ. Como a *afinidade*, f, e o valor de corte γ são fixos, a aleatoriedade de G_λ é decorrente da medida μ com a qual os indivíduos escolhem as palavras em $\mathcal{D}$. Como esta é uma definição desenvolvida para aplicação em redes, convencionamos $f(U_i, U_i) = 0$.

A matriz de adjacências $A_{n\times n}$ de G_λ é definida da seguinte forma

$$a_{ij} = \begin{cases} 1 & \text{se } f(U_i, U_j) > \gamma \\ 0 & \text{caso contrário.} \end{cases}$$

Temos que

$$P(a_{ij} = 1) = P(f(U_i, U_j) > \gamma).$$

A seguir as funções afinidades que trataremos neste trabalho.

3.3.1 Função afinidade binária

Seja $\mathcal{D}_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,m_i-1}, w_{i,m_i}\}$ tal que $\mathcal{D}_i \subset \mathcal{D}$ o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor binário associado a $\mathcal{D}_i$ definido como descrito na Equação (2). Definimos a função afinidade binária entre os elementos i e j do seguinte modo

$$f(U_i, U_j) = \begin{cases} 1, \text{ se } U_{ik} = U_{jk} \text{ para algum k} \\ 0, \text{ se caso contrário.} \end{cases}$$

Desta forma, a função afinidade binária é não-nula sempre que dois indivíduos compartilham pelo menos uma característica em seus conjuntos de escolhas, isto é

$$f(U_i, U_j) = 1 \Leftrightarrow \mathcal{D}_i \cap \mathcal{D}_j \neq \emptyset.$$

Pode-se alocar as respectivas afinidades na matriz de afinidades A

$$A_{n\times n} = \begin{bmatrix} f(U_1,U_1) & f(U_1,U_2) & f(U_1,U_3) & \cdots & f(U_1,U_n) \\ f(U_2,U_1) & f(U_2,U_2) & f(U_2,U_3) & \cdots & f(U_2,U_n) \\ f(U_3,U_1) & f(U_3,U_2) & f(U_3,U_3) & \cdots & f(U_3,U_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(U_n,U_1) & f(U_n,U_2) & f(U_n,U_3) & \cdots & f(U_n,U_n) \end{bmatrix}.$$

Realizando os cálculos para o exemplo genérico construído na subseção 3.1, U_1 e U_2 têm os elementos $w_{i,1}$ e $w_{i,2}$ em comum, logo a posição $a_{1,2}$ da matriz relacionada a função de afinidade binária é 1. Além disso, diagonal principal é composta por $0's$. Realizando os outros cálculos a matriz é dada por

$$A_{4\times 4} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ & 0 & 1 & 1 \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}.$$

3.3.2 Função afinidade cardinal

Seja $\mathcal{D}_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,m_i-1}, w_{i,m_i}\}$ tal que $\mathcal{D}_i \subset \mathcal{D}$ o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor binário associado a $\mathcal{D}_i$ definido como descrito na (2) . Definimos a função afinidade cardinal como

$$f(U_i,U_j) = \begin{cases} k, \text{ se } U_i \text{ e } U_j \text{ têm k termos em comum} \\ 0, \text{ se caso contrário.} \end{cases}$$

Ou seja, a função afinidade cardinal mensura quantas características em comum foram escolhidas pelos indivíduos

$$f(U_i,U_j) = |\mathcal{D}_i \cap \mathcal{D}_j|.$$

Considerando o exemplo construído na subseção 3.1, U_1 e U_2 têm 2 termos em comum, em vista disso o termo $a_{1,2}$ será igual a 2 e a diagonal principal será composta $0's$. Calculando as demais posições, temos a matriz

$$Y_{4\times 4} = \begin{bmatrix} 0 & 2 & 0 & 2 \\ & 0 & 1 & 3 \\ & & 0 & 2 \\ & & & 0 \end{bmatrix}.$$

A próxima seção apresentará distribuições μ com características diversas e técnicas para gerar valores das matrizes de escolhas.

4 Distribuições do vetor de escolhas $\mathbf{U_i}$

A geração do grafo aleatório via modelo de redes de afinidade é feita através das medidas de probabilidade μ que regem o comportamento das escolhas dos indivíduos. A princípio, a distribuição conjunta μ é arbitrária, de forma que o vetor aleatório U_i associado a D_i é tal que

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = P(U_{i,1} = u_{i,1}, U_{i,2} = u_{i,2}, \ldots, U_{i,m} = u_{i,m}).$$

sendo que U_i não possui qualquer restrição, tanto em relação à independência das $U_{i,j} \sim \mu_i$ ou o fato de $U_{i,j}$ serem identicamente distribuídas. Em relação a distribuição de μ_i, não há restrições, havendo até a possibilidade de que μ seja uma mistura finita de distribuições.

Nesta seção serão apresentados exemplos de funções de distribuição de probabilidade conjuntas que podem ser aplicadas sobre $\mathcal{D}$.

4.1 $\mathbf{U_{i,j}}$ binárias e independentes

Considerando $U_{i,j}$ e $U_{i,q}$ independentes $\forall j \neq q$ e a única informação sobre $U_{i,j}$ é se a j-ésima palavra pertence ou não ao vocabulário do i-ésimo indivíduo, isto é, podemos modelar por uma distribuição Bernoulli

$$U_{i,j} \sim Ber(p_j) \Rightarrow P(U_{i,j} = u_{i,j}) = p_j^{u_{i,j}} \cdot (1 - p_j)^{1-u_{i,j}}. \tag{6}$$

Implicando na distribuição conjunta expressa por

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = \prod_{j=1}^{m} P(U_{i,j} = u_{i,j}) = \prod_{j=1}^{m} p_j^{u_{i,j}} \cdot (1 - p_j)^{1-u_{i,j}}. \tag{7}$$

4.2 $\mathbf{U_{i,j}}$ binárias, independentes e identicamente distribuídas

Se $U_{i,j}$ são binárias e independentes, ainda podemos considerar que $U_{i,j}$'s são identicamente distribuídas para todo $j \in \{1, 2, \ldots, m\}$, como consequência $p_j = p$ para todo $j = \{1, 2, \ldots, m\}$, de forma que podemos reescrever a Equação 7 do seguinte modo

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = \prod_{j=1}^{m} P(U_{i,j} = u_{i,j}) = p^{\sum_{j=1}^{m} u_{i,j}} \cdot (1-p)^{m-\sum_{j=1}^{m} u_{i,j}}.$$

5 Inferência para os parâmetros de $\mathbf{G}_\lambda$

Fixadas as funções afinidade binária e cardinal, e estabelecidas as distribuições de probabilidade das matrizes de escolha em cada caso, passamos agora à estimação das medidas de probabilidade que governam as escolhas dos indivíduos. Enfatizamos que a matriz U é a única componente aleatória do modelo, dado que fixamos os demais parâmetros. Portanto, toda informação sobre as probabilidades do indivíduo escolher determinada palavra, ou característica, vem da matriz U. Consideramos, a princípio, que temos apenas uma medida de probabilidade governando as escolhas dos indivíduos, ou seja, existe apenas uma comunidade.

5.1 Cálculo da verossimilhança de $\mathbf{G}_\lambda$

Para encontrarmos um estimador para os parâmetros do modelo vamos escrever a função de verossimilhança do grafo de afinidade G_λ para cada uma das funções de afinidade consideradas. Começamos pela afinidade binária.

5.1.1 Verossimilhança de $\mathbf{G}_\lambda$ com Afinidade binária

Primeiramente, calculamos $P(a_{ij} = 1) = P(f(U_i, U_j) > \delta), 1 \leq i, j \leq m$, que é a probabilidade de cada uma das arestas ser 1, dado pela função afinidade entre dois elementos ser maior que 0. Temos que a probabilidade de ter pelo menos uma aresta entre os indivíduos i e j é dada por

$$\begin{aligned}
P(U_i^k = U_j^k, \text{para algum } k) &= 1 - (P(U_i^k \neq U_j^k, \text{para todo } k)) \\
&= 1 - \prod_{k=1}^{m} (P(U_i^k \neq U_j^k)) \\
&= 1 - \prod_{k=1}^{m} [P(U_i^k = 1, U_j^k = 0) + P(U_i^k = 0, U_j^i = k)] \\
&= 1 - \prod_{k=1}^{m} [p_i^k (1 - p_j^k) + (1 - p_i^k) p_j^k] \\
&= 1 - \prod_{k=1}^{m} [p_i^k + p_j^k - 2 p_i^k p_j^k].
\end{aligned}$$

Se os indivíduos i e j escolhem as palavras com a mesma probabilidade, $p_i^k = p_j^k = p^k$, então

$$P(a_{ij} = 1) = 1 - \prod_{k=1}^{m} [2p^k(1 - p^k)].$$

Então, a probabilidade do grafo de redes de afinidade, G_λ, assumir uma determinada configuração, $\mathcal{G}$, considerando a função afinidade binária, é dada por

$$\begin{aligned} P(G_\lambda = \mathcal{G}) &= \prod_{i=1}^{n} \prod_{j=i+1}^{n} P(a_{ij} = 1) \\ &= \prod_{i=1}^{n} \prod_{j=i+1}^{n} \{1 - \prod_{k=1}^{m} [2p^k(1 - p^k)]\}. \end{aligned} \tag{8}$$

5.1.2 Verossimilhança de G_λ com afinidade cardinal

Vamos calcular a distribuição de probabilidade da afinidade cardinal para dois indivíduos. Definimos $\mu = \{p_1, p_2, p_3, \ldots, p_m\}$, onde $\mu \in [0, 1]^m$, o conjunto das probabilidades de escolha associadas à $U_i = \{U_i^1, U_i^2, \ldots, U_i^{m-1}, U_i^m\}$ de forma que $U_j \sim \mu_j = Ber(p_j)$ e $U_j \perp U_l$. Então temos que, $\forall l, t$,

$$P(f(U_l, U_t) = k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j), \tag{9}$$

onde F_k é o conjunto de todos os subconjuntos com k inteiros que podem ser selecionados de $1, 2, 3, ..., m$.

De forma que $f(U_l, U_t)$ tem distribuição Poisson binomial com parâmetros p^2. Em particular, se $U_j \sim \mu_j = Ber(p),\ \forall j$, então

$$P(f(U_l, U_t) = k) = \binom{m}{k} \cdot (p^2)^k \cdot (1 - p^2)^{m-k},\ k = \{0, 1, \ldots, m\}. \tag{10}$$

Consequentemente, $f(U_l, U_t)$ tem distribuição binomial com parâmetro p^2. Ou seja, a distribuição binomial é um caso particular da distribuição Poisson binomial. Assim, a média e variância de $f(U_l, U_t)$ com f afinidade cardinal são tais que

$$E(f(U_l, U_t)) = \sum_{j=1}^{m} p_j^2 \text{ e } Var(f(U_i, U_k)) = \sum_{j=1}^{m} p_j^2 \cdot (1 - p_j^2). \tag{11}$$

Desse modo,

$$\begin{aligned} P(G_\lambda = \mathcal{G}) &= \prod_{i=1}^{n} \prod_{j=1}^{n} P(f(U_i, U_j) = k_{ij}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{n} \binom{m}{k_{ij}} \cdot (p^2)^{k_{ij}} \cdot (1 - p^2)^{m-k_{ij}},\ k_{ij} = \{0, 1, \ldots, m\}. \end{aligned} \tag{12}$$

5.1.3 Estimador de máxima verossimilhança de $\mu_\mathbf{i}$

Toda a informação sobre a medida de distribuição de U_i está na matriz U. As escolhas dos indivíduos estão armazenadas nela. Se cada indivíduo escolhesse seu vocabulário de acordo com uma μ_i diferente, a matriz U só teria uma realização de cada uma dessas medidas e não seria possível estimar tais medidas, a não ser que dispuséssemos de várias réplicas dela. Então, supondo, a princípio, que todos os indivíduos escolhem seus vocabulários com a mesma lei μ, a verossimilhança para $\mathbf{U}$, considerando um modelo de produto de bernoullis com parâmetros p_i, $1 \leq i \leq m$, ficaria

$$L(\mu, \mathbf{U}) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(U_{i,j} = u_{i,j}) = \prod_{i=1}^{n}\prod_{j=1}^{m} p_j^{u_{i,j}} \cdot (1-p_j)^{1-u_{i,j}}$$
$$= \prod_{j=1}^{m} p_j^{n_j} \cdot (1-p_j)^{n-n_j},$$

onde n_j é o número de vezes que a palavra w_j apareceu na amostra. Daí, segue que o estimador de máxima verossimilhança para a probabilidade de uma certa palavra w_j ser escolhida, $\mu(w_k) = \mu_j$,

$$\hat{p}_j = \frac{n_j}{n} \tag{13}$$

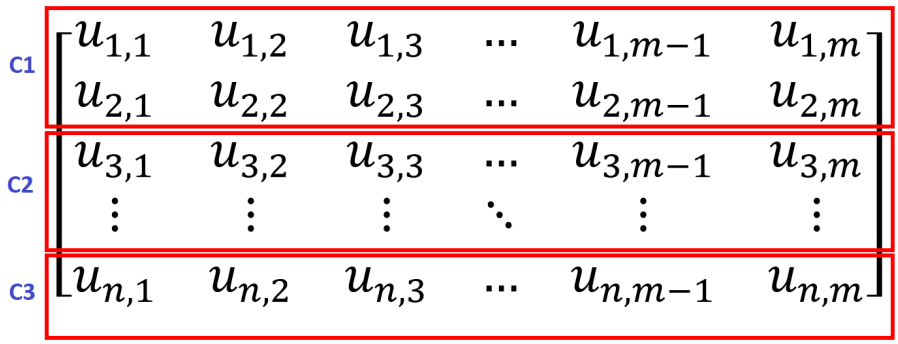
que é o número de vezes que a palavra w_j aparece na matriz U, dividido pelo total de indivíduos.

5.2 Estimação em modelos de redes de afinidade em blocos

Consideremos agora os n indivíduos divididos em comunidades conhecidas $\mathcal{C} = C_1, \ldots, C_Q$, tal que $C_i \cap C_j = \varnothing, \forall\ 1 \leq i, j \leq Q$. Indivíduos pertencentes à mesma comunidade C_q escolhem as palavras em $\mathcal{D}$ de acordo com a mesma medida μ_q, mas de forma diferente que indivíduos de comunidades diferentes. Denotamos o número de indivíduos em cada comunidade por $N_q, 1 \leq q \leq Q$ e o número de arestas por $E_q, 1 \leq q \leq (N_q)^2$. Dessa forma, agora $\lambda = \{(\mu)_{i\in Q}, f, \delta\}$.

A título de ilustração, considere a matriz U com todos os indivíduos conhecidos e divididos em 3 comunidades. Por definição, espera-se uma medida de probabilidade diferente para o

mesmo termo em cada uma das partições.



$$\begin{matrix} C1 \\ C2 \\ C3 \end{matrix} \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,m-1} & u_{1,m} \\ u_{2,1} & u_{2,2} & u_{2,3} & \cdots & u_{2,m-1} & u_{2,m} \\ u_{3,1} & u_{3,2} & u_{3,3} & \cdots & u_{3,m-1} & u_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{n,1} & u_{n,2} & u_{n,3} & \cdots & u_{n,m-1} & u_{n,m} \end{bmatrix} \tag{14}$$

Desta forma, dada a informação de qual comunidade cada indivíduo pertence, a verossimilhança de G_λ para a função afinidade binária fica da forma

$$\begin{aligned} P(G_\lambda = \mathcal{G}|\mathcal{C}) &= \prod_{U_i,U_j \in C_1} P(a_{ij} = 1|C_1) \times \cdots \times \prod_{U_i,U_j \in C_Q} P(a_{ij} = 1|C_Q) \\ &= \prod_{U_i,U_j \in C_1} \left(1 - \prod_{k=1}^{m} [2p_1^k(1-p_1^k)]\right) \times \cdots \times \prod_{U_i,U_j \in C_Q} \left(1 - \prod_{k=1}^{m} [2p_Q^k(1-p_Q^k)]\right), \end{aligned} \tag{15}$$

onde p_q^k, $1 \leq q \leq Q$, é a probabilidade da k na comunidade q, uma vez que os indivíduos são considerados independentes, dada a comunidade.

De forma análoga a (13), temos que o estimador de máxima verossimilhança de μ_q é dado por

$$\hat{p}_q^k = \frac{\sum_{i \in q} U_i^k}{n_q}, \tag{16}$$

onde n_q é o número de indivíduos na comunidade q. O estimador $\hat{P}(G_\lambda = \mathcal{G}|\mathcal{C})$ é obtido levando (16) em (15).

De maneira similar encontramos $\hat{P}(G_\lambda = \mathcal{G}|\mathcal{C})$ para a afinidade cardinal.

5.3 Estimação em modelos de afinidade por blocos quando as comunidades são desconhecidas

Seja $\mathbb{C}_Q$ o conjunto de todas as partições possíveis com tamanho Q do conjunto de vértices de G_λ. Os vértices são alocados a uma partição $\mathcal{C}$ de acordo com probabilidades $\theta_1, \ldots, \theta_Q$. Assim,

$$P(\mathcal{C}) = P(N_1 = n_1, \ldots, N_Q = n_q) = \theta_1^{n_1} \times \ldots \times \theta_Q^{n_Q}.$$

Por outro lado, temos que

$$\begin{aligned}
P(G_\lambda) &= \sum_{\mathcal{C}\in\mathbb{C}_Q} P(G_\lambda, \mathcal{C}) \\
&= \sum_{\mathcal{C}\in\mathbb{C}_Q} P(G_\lambda|\mathcal{C})P(\mathcal{C}) \\
&= \sum_{\mathcal{C}\in\mathbb{C}_Q} P(G_\lambda|\mathcal{C})\theta_1^{n_1} \times \ldots \times \theta_Q^{n_Q} \\
&= \sum_{\mathcal{C}\in\mathbb{C}_Q} \theta_1^{n_1} \times \ldots \times \theta_Q^{n_Q} \prod_{i=1}^{n_1}\prod_{j=1}^{n_1}\{1 - \prod_{k=1}^{m}[2p_1^k(1-p_1^k)]\} \times \ldots \times \prod_{j=1}^{n_Q}\{1 - \prod_{k=1}^{m}[2p_Q^k(1-p_Q^k)]\}.
\end{aligned}$$

Dessa forma, poderíamos utilizar o estimador proposto em (16) para estimar μ, mas precisaríamos propor um estimador também para θ_q,

A estimação dos θ_q (probabilidade de um indivíduo pertencer a comunidade q) a partir de uma distribuição multinomial, onde os indivíduos são alocados às comunidades de maneira independente, traria desvantagem do ponto de vista da interpretação das comunidades. Isso porque as pessoas seriam alocadas às comunidades sem levar em conta a informação sobre as palavras, ou características, que ela escolheu, isto é, estaríamos sorteando os indivíduos nas comunidades com base apenas em proporções, o que não traduz a idéia de afinidade. Assim, θ_q não teria relação com μ e $\mathbf{U}$.

Tendo isso em vista, a proposta de estimação que fazemos para θ_q é baseada em um modelo de mistura de vetores de variáveis aleatórias *Bernoulli(p_{qj})* e leva em conta a informação sobre as palavras escolhidas pelos indivíduos contidas na matriz $\mathbf{U}$. Estimamos as medidas de probabilidade $(\mu_q)_{1\leq q\leq Q}$ e as probabilidades de um indivíduo pertencer a uma classe q, θ_q, conjuntamente, através de um algoritmo EM aplicado na matriz $\mathbf{U}$ diretamente.

5.3.1 Algoritmo EM para estimação dos parâmetros de $\mathbf{G}_\lambda$ com comunidades desconhecidas

Vamos considerar um modelo de misturas para o vetor de escolhas U, isto é, há mais de um conjunto de parâmetros μ que regem o comportamento de U. Seja Z uma variável aleatória, assumindo valores em $\{1, 2, \ldots, Q\}$, que indica o componente da mistura que o indivíduo i pertence e tal que $P(Z = k) = \theta_k$. Desta forma, temos que

$$\begin{aligned}
P(U = u) &= \sum_{k=1}^{Q} P(U = u, Z = k) = \sum_{k=1}^{Q} P(U = u|Z = k)P(Z = k) \\
&= \sum_{k=1}^{Q} \left(\theta_k \prod_{j=1}^{m} p_{j,k}^{u_j}(1 - p_{j,k})^{1-u_j} \right).
\end{aligned} \tag{17}$$

Portanto, a função de verossimilhança para a amostra pode ser escrita por

$$L(\mu, \mathbf{U}, \mathbf{Z}) = \prod_{i=1}^{n} P(U = u) = \prod_{i=1}^{n} \prod_{k=1}^{Q} \left[\theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1 - p_{j,k})^{1-u_{i,j}} \right]. \tag{18}$$

Aplicando o logaritmo na função de verossimilhança amostral

$$\begin{aligned}
l(\mu, \mathbf{U}, \mathbf{Z}) &= \log \left(L(\mu, \mathbf{U}, \mathbf{Z}) \right) \\
&= \log \left(\prod_{i=1}^{n} \prod_{k=1}^{Q} \left[\theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1 - p_{j,k})^{1-u_{i,j}} \right] \right) \\
&= \sum_{i=1}^{n} \log \left(\prod_{k=1}^{Q} \left[\theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1 - p_{j,k})^{1-u_{i,j}} \right] \right)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{n} \sum_{k=1}^{Q} \log (\theta_k) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{Q} \sum_{j=1}^{m} u_{i,j} \log (p_{j,k}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{Q} \sum_{j=1}^{m} (1 - u_{i,j}) \log (1 - p_{j,k}).
\end{aligned} \tag{19}$$

Agora temos que encontrar os valores dos parâmetros que maximizam a função de verossimilhança.

Vamos utilizar o seguinte resultado

$$P(U = u) = \sum_{k=1}^{Q} P(U = u, Z = k) = \sum_{k=1}^{Q} P(Z = k | U = u) P(U = u). \tag{20}$$

Pelo Teorema de Bayes, podemos escrever

$$P(Z = k | U = u) = \frac{P(U = u, Z = k)}{P(U = u)}. \tag{21}$$

e segundo a Equação 17, temos que $\forall\ r \in \{1, 2, \ldots, k\}$

$$P(U = u, Z = r) = \theta_r \prod_{j=1}^{m} p_{j,r}^{u_j} (1 - p_{j,r})^{1-u_j}. \tag{22}$$

Por outro lado,

$$P(U = u) = \sum_{k=1}^{Q} P(U = u, Z = k) = \sum_{k=1}^{Q} \theta_k \prod_{j=1}^{m} p_{j,k}^{u_j} (1 - p_{j,k})^{1-u_j}, \tag{23}$$

de modo que

$$P(Z = k | U = u) = \frac{\theta_k \prod_{j=1}^{m} p_{j,k}^{u_j} (1 - p_{j,k})^{1-u_j}}{\sum_{k=1}^{Q} \theta_k \prod_{j=1}^{m} p_{j,k}^{u_j} (1 - p_{j,k})^{1-u_j}} = T_k. \tag{24}$$

Dado um vetor de escolhas $\mathbf{U}$, o termo T_k se refere a probabilidade de ter sido gerado pela k-ésima parcela da mistura de distribuições que regem U.

Definindo

$$Q(\theta|\theta^{(t)}) = E_{Z|\mathbf{U}}(\log(L(\theta, \mathbf{U}, \mathbf{Z}))) \tag{25}$$

onde (t) representa a informação do valor na iteração anterior. Temos que

$$\begin{aligned} Q(\mu|\mu^{(t)}) &= E_{Z|\mathbf{U}}(\log(L(\mu, \mathbf{u}, \mathbf{Z}))) \\ &= E_{Z|\mathbf{U}}\left(\log\left(\prod_{i=1}^{n} L(\mu, u_i, Z_i)\right)\right) \\ &= E_{Z|\mathbf{U}}\left(\sum_{i=1}^{n} \log\left(L(\mu, u_i, Z_i)\right)\right) \\ &= \sum_{i=1}^{n} E_{Z|\mathbf{U}}\left(\log\left(L(\mu, u_i, Z_i)\right)\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{Q} P(Z_i = k|\mathbf{U_i} = \mathbf{u_i}) l(\mu, u_i, Z_i = k), \end{aligned} \tag{26}$$

então levando (24) em (26) temos

$$\begin{aligned} Q(\mu|\mu^{(t)}) &= \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k} \log\left[\theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1-p_{j,k})^{1-u_{i,j}}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}\log\left(\theta_k\right) + \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}\sum_{j=1}^{m} u_{i,j}\log\left(p_{j,k}\right) + \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}\sum_{j=1}^{m}(1-u_{i,j})\log\left(1-p_{j,k}\right), \end{aligned} \tag{27}$$

onde

$$T_{i,k} = \frac{\theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1-p_{j,k})^{1-u_{i,j}}}{\sum_{k=1}^{Q} \theta_k \prod_{j=1}^{m} p_{j,k}^{u_{i,j}} (1-p_{j,k})^{1-u_{i,j}}}. \tag{28}$$

Para encontrar os estimadores de máxima verossimilhança em $Q(\mu|\mu^{(t)})$, calculamos:

$$\frac{\partial Q(\mu|\mu^{(t)})}{\partial p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}u_{i,j}}{p_{j,k}} - \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j})}{(1-p_{j,k})} = 0 \tag{29}$$

$$\frac{\sum_{i=1}^{n} T_{i,k}u_{i,j}}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j})}{(1-p_{j,k})}$$

$$\frac{1-p_{j,k}}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j})}{\sum_{i=1}^{n} T_{i,k}u_{i,j}}$$

$$\frac{1}{p_{j,k}} - 1 = \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j})}{\sum_{i=1}^{n} T_{i,k}u_{i,j}}$$

$$\frac{1}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j})}{\sum_{i=1}^{n} T_{i,k}u_{i,j}} + 1$$

$$\frac{1}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}(1-u_{i,j}) + \sum_{i=1}^{n} T_{i,k}u_{i,j}}{\sum_{i=1}^{n} T_{i,k}u_{i,j}}$$

$$\frac{1}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k} - T_{i,k}u_{i,j} + T_{i,k}u_{i,j}}{\sum_{i=1}^{n} T_{i,k}u_{i,j}}$$

$$\frac{1}{p_{j,k}} = \frac{\sum_{i=1}^{n} T_{i,k}}{\sum_{i=1}^{n} T_{i,k}u_{i,j}}$$

$$\hat{p}_{j,k} = \frac{\sum_{i=1}^{n} T_{i,k}u_{i,j}}{\sum_{i=1}^{n} T_{i,k}}. \tag{30}$$

Para maximizar θ_k, considere $\sum_{k=1}^{Q} \theta_k = 1$, e utilizaremos a seguir os multiplicadores de Lagrange para encontrar os estimadores. Suponha que a função $f(x, y)$ será maximizada, onde o ponto de otimização (máximo ou mínimo) esteja restrito a $g(x, y) = c$. Pela função de Lagrange, uma nova variável λ será introduzida e será definida por

$$\Lambda(x, y, \lambda) = f(x, y) - \lambda\left(g(x, y) - c\right)$$

de forma que

$$\Lambda(\theta, \mu, u, Z) = Q(\mu|\mu^{(t)}) - \lambda\left(\sum_{k=1}^{Q} \theta_k - 1\right),$$

onde $\Lambda(\theta, \mu, u, Z) = Q(\mu|\mu^{(t)})$ já que $\lambda\left(\sum_{k=1}^{Q} \theta_k - 1\right) = 0$, pois $\sum_{k=1}^{Q} \theta_k = 1$. Destarte, ainda maximizaremos $Q(\mu|\mu^{(t)})$. Derivando $\Lambda(\theta, \mu, u, Z)$ com respeito a θ_k temos

$$\frac{\partial \Lambda(\theta, \mu, u, Z)}{\partial \theta_k} = -\lambda + \frac{\sum_{i=1}^{n} T_{i,k}}{\theta_k} = 0$$

$$\theta_k = \frac{\sum_{i=1}^{n} T_{i,k}}{\lambda}. \tag{31}$$

Todavia, observe que

$$\sum_{k=1}^{Q} \theta_k = \sum_{k=1}^{Q} \frac{\sum_{i=1}^{n} T_{i,k}}{\lambda} = \frac{\sum_{k=1}^{Q}\sum_{i=1}^{n} T_{i,k}}{\lambda} = 1$$

considerando a restrição $\sum_{k=1}^{Q} \theta_k = 1$. Logo,

$$\lambda = \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k} = n,$$

visto que é a soma para todos os indivíduos e para todos os grupos da probabilidade dos indivíduos pertencerem aos grupos. Deste modo, temos que

$$\hat{\theta}_k = \frac{\sum_{i=1}^{n} T_{i,k}}{\lambda} = \frac{\sum_{i=1}^{n} T_{i,k}}{n}. \tag{32}$$

Por conseguinte, para estimar os parâmetros do modelo de mistura de vetores *Bernoulli* é necessário: a matriz u de dimensões $n \times m$ observada, um vetor de parâmetros $\theta^{(0)}$ (os palpites iniciais das proporções das parcelas das misturas) de comprimento Q (o número de grupos que dividirá os dados), uma matriz $\mu^{(0)}$ de dimensões $Q \times m$ (os valores iniciais para os parâmetros μ iniciais que serão usados para iniciar o algoritmo) e uma constante de tolerância ε.
Resumindo o algoritmo de estimação, temos:

Passo E: Esperança

Computar

$$T_{i,k}^{(t)} = \frac{\theta_k^{(t)} \prod_{j=1}^{m} \left(p_{j,k}^{(t)}\right)^{u_{i,j}} \left(1 - p_{j,k}^{(t)}\right)^{1-u_{i,j}}}{\sum_{k=1}^{Q} \theta_k^{(t)} \prod_{j=1}^{m} \left(p_{j,k}^{(t)}\right)^{u_{i,j}} \left(1 - p_{j,k}^{(t)}\right)^{1-u_{i,j}}} \tag{33}$$

e calcular

$$\begin{aligned} Q(\mu|\mu^{(t)}) = &\sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}^{(t)} \log\left(\theta_k^{(t)}\right) + \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}^{(t)} \sum_{j=1}^{m} u_{i,j} \log\left(p_{j,k}^{(t)}\right) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{Q} T_{i,k}^{(t)} \sum_{j=1}^{m} (1 - u_{i,j}) \log\left(1 - p_{j,k}^{(t)}\right). \end{aligned} \tag{34}$$

Passo M: Maximização

Calcular e atualizar os parâmetros

$$p_{j,k}^{(t+1)} = \frac{\sum_{i=1}^{n} T_{i,k}^{(t)} u_{i,j}}{\sum_{i=1}^{n} T_{i,k}^{(t)}} \tag{35}$$

$$\theta_k^{(t+1)} = \frac{\sum_{i=1}^{n} T_{i,k}^{(t)}}{n}. \tag{36}$$

Repetir os passos sucessivamente até que $Q(\mu|\mu^{(t)}) - Q(\mu|\mu^{(t-1)}) < \varepsilon$.

5.4 Alocação dos indivíduos às comunidades da rede de afinidades

Uma vez que o processo de estimação via algoritmo EM tenha sido realizado, passamos agora à etapa de alocar os indivíduos e estimar o número de comunidades com base nas verossimilhanças individuais. Para tal, suponha que n indivíduos pertencentes a Q comunidades possam escolher m palavras, deste modo, a função de verossimilhança individual avaliada usando a medida estimada usando o EM para o elemento i é dada por

$$f(u_i, \tilde{\mu}, \tilde{\theta}) = \sum_{q=1}^{Q} \theta_q P(u_i, \tilde{\mu}_q) = \sum_{q=1}^{Q} \theta_q L(u_i, \tilde{\mu}_q), \tag{37}$$

onde u_i é um vetor composto por $0's$ e $1's$ de tamanho m que representa as escolhas do i-ésimo indivíduo, $\tilde{\mu}_q$ é um vetor que representa a medida de probabilidade estimada na comunidade q e θ_q é a proporção estimada de indivíduos na comunidade q.

Por consequência, como os vetores $u_i's$ são independentes, a função de verossimilhança amostral total pode ser escrita como o produto das funções de probabilidades individuais, isto é,

$$\prod_{i=1}^{n} f(u_i, \tilde{\mu}_q, \tilde{\theta}_q) = \prod_{i=1}^{n} \sum_{k=1}^{Q} \theta_k L(u_i, \tilde{\mu}_q). \tag{38}$$

e a função de log-verossimilhança é dada por

$$l(u_i, \tilde{\mu}_q) = log\Big(\prod_{i=1}^{n} f(u_i, \tilde{\mu}_q, \tilde{\theta}_q) \Big) = \sum_{i=1}^{n} log\Big\{ \sum_{k}^{Q} \theta_q L(u_i, \tilde{\mu}_q) \Big\}.$$

Como ilustração, considerando duas comunidades e dois indivíduos, a função de log-verossimilhança amostral seria dada por:

$$\begin{aligned} \prod_{i=1}^{2} f(u_i, \tilde{\mu}_q, \tilde{\theta}_q) =& \prod_{i=1}^{2} log\Big(\sum_{q=1}^{2} \theta_q L(u_i, \tilde{\mu}_q) \Big) = \\ & = log\big\{ \theta_1 L(u_1, \tilde{\mu}_1) + \theta_2 L(u_1, \tilde{\mu}_2) \big\} \times log\big\{ \theta_1 L(u_2, \tilde{\mu}_1) + \theta_2 L(u_2, \tilde{\mu}_2) \big\} \end{aligned}$$

5.4.1 Seleção do número de comunidades em uma rede de afinidade

O Critério de Informação Bayesiano, BIC na sigla em inglês, é um índice usado para escolher entre dois ou mais modelos alternativos. O BIC foi desenvolvido por Gideon E. Schwarz e publicado em um artigo de 1978, [1] onde ele deu um argumento bayesiano para adotá-lo. O critério baseia-se, em parte, na função de verossimilhança, mas também no número de parâmetros do modelo escolhido. Ao ajustar modelos, é possível aumentar a verossimilhança adicionando parâmetros, mas isso pode resultar em um sobreajuste. O BIC resolve esse problema introduzindo um termo de penalidade para o número de parâmetros no modelo. O modelo com o BIC mais alto é o preferido.

Em modelos de redes de afinidade onde a quantidade de comunidades é desconhecida, é necessário determinar esse número. Chegamos, então, em um problema de seleção de modelos e podemos calcular o BIC considerando o modelo gerado por cada número de comunidades e, assim, escolher aquele que for mais parcimonioso, indicado pelo critério.

Para o caso geral, o BIC é determinado da seguinte maneira

$$BIC_{argmax}(Q) = \big\{ 2l_Q(U, \tilde{\mu}) - \Delta log(n) \big\}, \tag{39}$$

onde Δ é o número de parâmetros do modelo considerado e n é o número de indivíduos. Vale ressaltar que e o BIC é consistente em várias situações (Portela, 2008), ou seja, quando $n \to \infty$ a escolha do modelo sugerida pelo BIC é a correta.

Para um cenário em que deseja-se obter o número ótimo de comunidades sem nenhum conhecimento prévio do mesmo, supondo valores de Q iguais a 1, 2 e 3, o Δ é calculado da seguinte forma para o modelo de produto de *Bernoulli's*:

$$\Delta = Qm + (Q-1) = Q(m+1) - 1$$

onde m é o tamanho do dicionário.

5.4.2 Critério de alocação dos indivíduos às comunidades

As alocações dos indivíduos nas suas respectivas comunidades estimadas são feitas por intermédio das funções de verossimilhanças individuais, uma vez que as medidas de probabilidades e o número ótimo de grupos já foram estimados. Primeiramente calculamos a verossimilhança individual considerando cada um dos grupos e observamos o resultado de maior valor, sendo este o critério para definir a qual comunidade este elemento pertence. Dessa forma, o i-ésimo indivíduo será alocado na comunidade que tiver máximo valor de verossimilhança individual e em caso de empate, sorteamos aleatoriamente a qual agrupamento ele pertence.

Para exemplificar, suponha que o vetor de escolhas do i-ésimo indivíduo seja igual a $u_i = [0, 0, 1]$ e considere que há duas comunidades, com medidas de probabilidades estimadas iguais a $\mu_1 = [0.70, 0.60, 0.20]$ e $\mu_2 = [0.10, 0.30, 0.80]$ para Q_1 e Q_2, respectivamente. Para o grupo 1 o cálculo é dado por

$$L(u_i, \tilde{\mu}_1) = \tilde{\mu}_1^{u_i} = \prod_{j=1}^{m} p_{j,1}^{u_{i,j}} = p_{1,1}^{u_{i,1}} \times p_{2,1}^{u_{i,2}} \times p_{3,1}^{u_{i,3}} = 0.70^0 \times 0.60^0 \times 0.20^1 = 0.20,$$

de forma similar, para o grupo 2 temos

$$L(u_i, \tilde{\mu}_2) = \tilde{\mu}_2^{u_i} = \prod_{j=1}^{m} p_{j,2}^{u_{i,j}} = p_{1,2}^{u_{i,1}} \times p_{2,2}^{u_{i,2}} \times p_{3,2}^{u_{i,3}} = 0.10^0 \times 0.30^0 \times 0.80^1 = 0.80.$$

Por consequência, o i-ésimo indivíduo é alocado à comunidade 2.

5.5 Construção dos grafos de afinidade G_λ

Seguindo o processo, após a alocação dos indivíduos o próximo passo é construir os grafos de afinidade. Para a construção dos grafos é necessário aplicar as funções afinidades (binária e cardinal) na matriz de adjacências, indicando assim se há aresta entre dois elementos.

Vale lembrar que haverá ligação entre os elementos i e j considerando a função afinidade binária se ambos os indivíduos compartilharem ao menos uma palavra. Para a função afinidade cardinal deve-se inicialmente determinar o ponto de corte δ, dessa forma para que haja ligação entre os elementos i e j, eles têm que compartilhar no mínimo δ termos. Sejam os indivíduos 1 e 2 alocados à comunidade Q com vetores de escolhas $U_1 = [0, 1, 0, 1, 0, 1, 1, 0, 1, 0]$ e $U_1 = [1, 1, 1, 0, 1, 0, 1, 0, 0, 1]$. O número de palavras em comum entre eles é 3, se analisarmos a função afinidade binária, concluímos que há aresta entre eles. Entretanto para a função afinidade cardinal temos que verificar o ponto de corte:

- se $\delta = 1$, há aresta entre os indivíduos 1 e 2 (função afinidade binária);
- se $\delta = 2$, há aresta entre os indivíduos 1 e 2;
- se $\delta = 3$, há aresta entre os indivíduos 1 e 2;
- se $\delta = 4$, não há aresta entre os indivíduos 1 e 2.

Para verificar como a escolha do ponto de corte influencia na construção do grafo, considere que 15 indivíduos possam escolher até 10 características para um determinado tema. A matriz de escolhas U pode ser vista a seguir

$$U_{15\times10} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Nas figuras 1 e 2 temos os grafos para diversos pontos de corte, assim pode-se observar que a medida que o ponto de corte aumenta, a quantidade de arestas entre os vértices diminui.

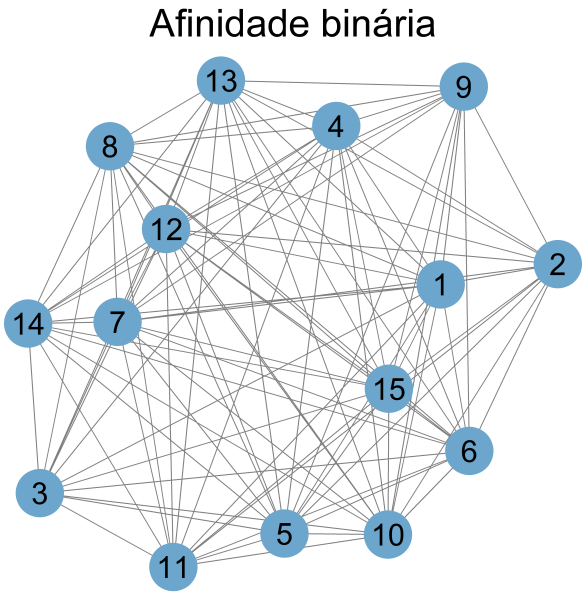


Figura 1: Exemplo de Grafo Função Afinidade Binária

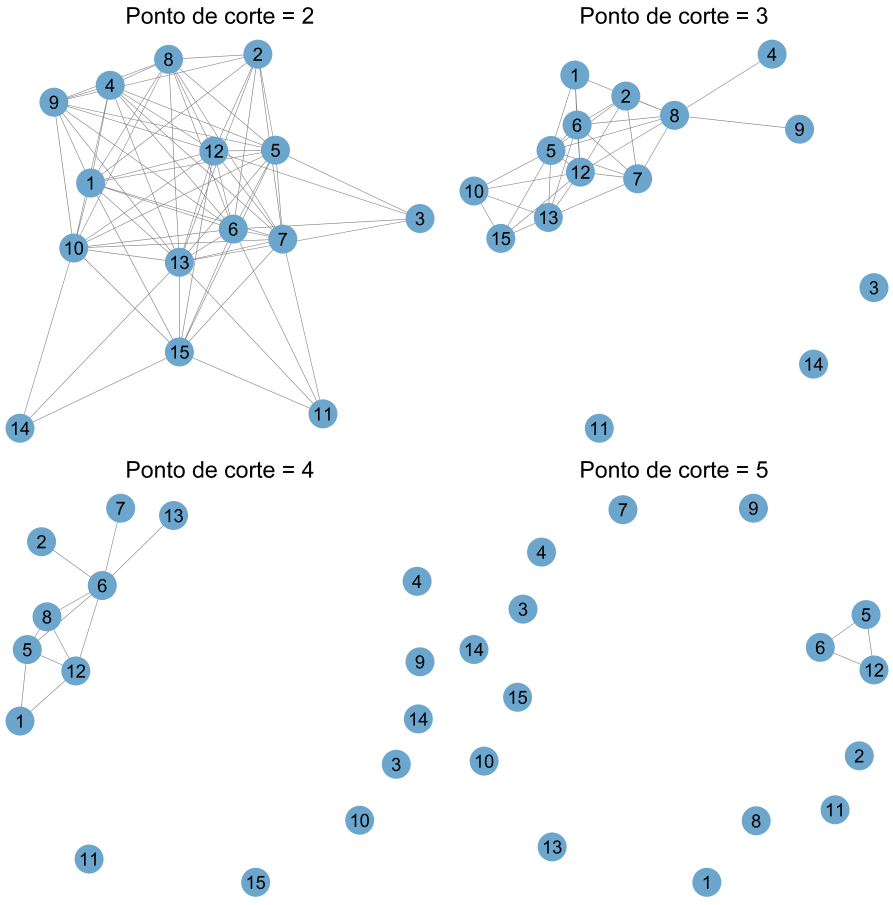


Figura 2: Exemplo de Grafos Funções Afinidades Cardinais

6 Simulações

Este trabalho tem como finalidade principal o desenvolvimento de uma metodologia que permite estimar os parâmetros do modelo de redes de afinidade. Sendo assim, após a realização do processo de estimação devemos validar a metodologia e comprovar a eficiência da mesmo por meio de simulações.

Para esse fim, simulações foram construídas com o foco de recuperar os valores verdadeiros e explorar os mais diversos tipos cenários, isto é, variações das medidas de probabilidades iniciais e reais, do número de vértices, número de simulações e quantidade de palavras. Para inicializar a aplicação do algoritmo, o primeiro passo é gerar as matrizes de escolhas U segundo medidas e proporções reais de indivíduos nas comunidades. Em seguida, deve-se estabelecer os chutes iniciais para os vetores $\boldsymbol{\mu}$ e $\boldsymbol{\theta}$. A fim de ter o máximo de imparcialidade, os valores iniciais do vetor de $\theta's$ considerados seguem uma distribuição uniforme, ou seja, $\theta_{kinicial} = 1/Q$, onde Q é o número total de comunidades.

Todos os códigos foram implementados em R, versão 4.0.3, e executados em um computador com processaor Intel Core i7 sétima geração, 16GB de memória ram e 128GB de SSD. O tempo das simulações em cada cenário teve relação direta com o número de escolhas, indivíduos, comunidades, repetições e valores iniciais, isto é, quanto mais elementos, características, grupos e chutes iniciais distantes dos reais, maior o tempo de convergência. Em casos extremos, alguns cenários convergiram em segundos, enquanto outros demoraram até 13 horas.

Nas Tabelas 1, 2, 3 e 4 é possível encontrar os resultados médios das simulações de diversos cenários, onde θ é a proporção de indivíduos nas C comunidades e μ é a medida de probabilidade para cada um dos grupos. Ademais, entre parênteses constam os desvios-padrões das estimativas médias fornecidas pelo algoritmo e na última coluna têm a proporção de alocação dos indivíduos com base nas medidas estimadas.

6.1 Cenário 1: Alterações dos valores de μ iniciais

Para este cenário foram considerados 60 elementos, 5 palavras e duas comunidades, com medida de probabilidade real para o grupo 1 seguindo distribuição uniforme, ou seja, $\mu_1 = [0.20, 0.20, 0.20, 0.20, 0.20]$, enquanto que para o grupo 2 a medida foi igual a $\mu_2 = [0.60, 0.10, 0.00, 0.00, 0.30]$. Além disso, as proporções reais foram de 30% e 70% para as comunidades 1 e 2, respectivamente e o números de simulações igual a 1000. Deste modo, houve variações nas medidas de probabilidades iniciais.

Isto posto, a Tabela 1 apresenta os resultados das estimativas do primeiro cenário para 5

medidas de probabilidades iniciais diferentes, permitindo inferir que mesmo alterando os valores iniciais o método consegue recuperar as medidas reais. Nota-se que os valores médios de θ mais próximos dos reais foram encontrados na quinta variação das medidas iniciais.

Tabela 1: Resultados das simulações para os parâmetros após aplicação do algoritmo considerando medidas iniciais distintas - Cenário 1

Cenário 1 - 60 vértices			θ	μ					Alocação
Simulações = 1000		C		j = 1	j = 2	j = 3	j = 4	j = 5	
Valores reais		1	0,3000	0,2000	0,2000	0,2000	0,2000	0,2000	-
		2	0,7000	0,6000	0,1000	0,0000	0,0000	0,3000	
1	Valores iniciais	1	0,5000	0,2500	0,2300	0,1000	0,1300	0,2900	-
		2	0,5000	0,6700	0,0100	0,0500	0,0700	0,2000	
	Estimativas	1	0,3299 (0,2098)	0,1511 (0,2037)	0,2774 (0,2493)	0,2446 (0,2551)	0,2438 (0,2545)	0,2363 (0,2109)	0,25
		2	0,6701 (0,2098)	0,6737 (0,2128)	0,0802 (0,0672)	0,0262 (0,0748)	0,0265 (0,0888)	0,2937 (0,1508)	0,75
2	Valores iniciais	1	0,5000	0,3000	0,2300	0,0500	0,0700	0,3500	-
		2	0,5000	0,5000	0,0100	0,1000	0,1000	0,2900	
	Estimativas	1	0,3951 (0,2664)	0,2880 (0,2781)	0,3358 (0,2489)	0,1843 (0,2348)	0,1845 (0,2416)	0,2409 (0,2092)	0,25
		2	0,6049 (0,2664)	0,5122 (0,2690)	0,0518 (0,0578)	0,0933 (0,1974)	0,0894 (0,1835)	0,2859 (0,1818)	0,75
3	Valores iniciais	1	0,5000	0,4500	0,2000	0,1000	0,1500	0,1000	-
		2	0,5000	0,8000	0,0500	0,0500	0,0500	0,0500	
	Estimativas	1	0,2835 (0,1868)	0,1269 (0,1514)	0,2465 (0,2171)	0,2521 (0,2441)	0,3244 (0,3125)	0,2550 (0,2475)	0,27
		2	0,7165 (0,1868)	0,6771 (0,2014)	0,0915 (0,0666)	0,0177 (0,0295)	0,0137 (0,0233)	0,2855 (0,0971)	0,73
4	Valores iniciais	1	0,5000	0,0500	0,6000	0,1000	0,2000	0,0500	-
		2	0,5000	0,4000	0,1000	0,1000	0,2000	0,2000	
	Estimativas	1	0,2771 (0,1910)	0,1059 (0,1714)	0,2733 (0,2510)	0,2726 (0,2743)	0,2770 (0,2774)	0,1640 (0,1770)	0,25
		2	0,7229 (0,1910)	0,6436 (0,1887)	0,0954 (0,0737)	0,0258 (0,0669)	0,0247 (0,0644)	0,3151 (0,1415)	0,75
5	Valores iniciais	1	0,5000	0,1000	0,3000	0,2000	0,3000	0,1000	-
		2	0,5000	0,3000	0,4000	0,1000	0,1000	0,1000	
	Estimativas	1	0,2994 (0,1944)	0,1043 (0,1659)	0,2313 (0,2300)	0,2706 (0,2699)	0,2675 (0,2696)	0,2252 (0,2169)	0,25
		2	0,7006 (0,1944)	0,6810 (0,1946)	0,1125 (0,1064)	0,0202 (0,0491)	0,0190 (0,0557)	0,2940 (0,1230)	0,75

6.2 Cenário 2: Alterações dos valores de μ iniciais e o número de vértices (N)

Neste cenário o objetivo foi avaliar como o processo de estimação funciona para diferentes quantidades de elementos e medidas iniciais. Fixados o número de simulações em 1000, proporções reais de indivíduos em cada comunidade (30% para a comunidade 1 e 70% para a comunidade 2) e medidas reais $\mu_1 = [0.20, 0.20, 0.20, 0.20, 0.20]$ para o grupo 1 e $\mu_2 = [0.60, 0.10, 0.00, 0.00, 0.30]$ para o grupo 2, a tabela 2 mostra os resultados para os casos em que as medidas iniciais são iguais a $\mu_1 = [0.20, 0.20, 0.20, 0.20, 0.20]$ e $\mu_2 = [0.20, 0.20, 0.20, 0.20, 0.20]$.

Portanto, na Tabela 2 têm-se as estimativas médias e desvio-padrão das mesmas. Observa-se que ao aumentar o número de vértices, a exatidão das estimativas aumenta, pois vão se aproximando dos valores reais, indicando que o algoritmo consegue recuperar de forma satisfatória os parâmetros verdadeiros.

Tabela 2: Resultados das simulações para os parâmetros após aplicação do algoritmo considerando número de vértices e medidas iniciais distintos - Cenário 2

Simulações = 1000		θ	μ					Alocação
Parâmetros	C		j = 1	j = 2	j = 3	j = 4	j = 5	
Valores reais	1	0.30	0.20	0.20	0.20	0.20	0.20	-
	2	0.70	0.60	0.10	0.00	0.00	0.30	
Valores iniciais	1	0.50	0.25	0.23	0.10	0.13	0.29	-
	2	0.50	0.67	0.01	0.05	0.07	0.20	
Vértices = 20	1	0,3819 (0,2160)	0,1926 (0,2851)	0,3267 (0,3132)	0,1938 (0,2625)	0,1957 (0,2764)	0,2986 (0,3027)	0,2500
	2	0,6181 (0,2160)	0,6935 (0,2581)	0,0628 (0,0934)	0,0340 (0,1198)	0,0347 (0,1127)	0,2860 (0,2342)	0,7500
Vértices = 60	1	0,3319 (0,2129)	0,1557 (0,2099)	0,2803 (0,2506)	0,2441 (0,2567)	0,2399 (0,2512)	0,2390 (0,2141)	0,2500
	2	0,6681 (0,2129)	0,6693 (0,2158)	0,0786 (0,0664)	0,0294 (0,0891)	0,0292 (0,0937)	0,2904 (0,1488)	0,7500
Vértices = 100	1	0,3073 (0,1991)	0,1438 (0,1769)	0,2531 (0,1935)	0,2505 (0,2318)	0,2558 (0,2313)	0,2052 (0,1813)	0,2400
	2	0,6927 (0,1991)	0,6437 (0,1914)	0,0915 (0,0558)	0,0206 (0,0616)	0,0229 (0,0771)	0,2963 (0,1196)	0,7600
Vértices = 150	1	0,2781 (0,1765)	0,1315 (0,1402)	0,2326 (0,1549)	0,2721 (0,2165)	0,2612 (0,2023)	0,1953 (0,1504)	0,2933
	2	0,7219 (0,1765)	0,6376 (0,1647)	0,0983 (0,0540)	0,0144 (0,0434)	0,0148 (0,0400)	0,2987 (0,0873)	0,7067
Vértices = 200	1	0,2642 (0,1619)	0,1237 (0,1166)	0,2283 (0,1301)	0,2668 (0,1952)	0,2704 (0,1948)	0,1893 (0,1240)	0,2800
	2	0,7358 (0,1619)	0,6300 (0,1495)	0,1006 (0,0377)	0,0121 (0,0178)	0,0119 (0,0162)	0,3001 (0,0648)	0,7200
Vértices = 500	1	0,2585 (0,1113)	0,1369 (0,0811)	0,2152 (0,0637)	0,2353 (0,1012)	0,2371 (0,1031)	0,1856 (0,0682)	0,2360
	2	0,7415 (0,1113)	0,6075 (0,0920)	0,1026 (0,0227)	0,0090 (0,0110)	0,0088 (0,0107)	0,2980 (0,0319)	0,7640
Vértices = 1000	1	0,2617 (0,0732)	0,1583 (0,0610)	0,2107 (0,0406)	0,2242 (0,0597)	0,2238 (0,0607)	0,1893 (0,0438)	0,2320
	2	0,7383 (0,0732)	0,5951 (0,0479)	0,1025 (0,0156)	0,0064 (0,0076)	0,0065 (0,0080)	0,2981 (0,0210)	0,7680
Vértices = 2000	1	0,2732 (0,0529)	0,1714 (0,0395)	0,2061 (0,0267)	0,2136 (0,0396)	0,2136 (0,0388)	0,1935 (0,0273)	0,2410
	2	0,7268 (0,0529)	0,5956 (0,0318)	0,1016 (0,0113)	0,0045 (0,0055)	0,0045 (0,0057)	0,2984 (0,0154)	0,7590

6.3 Cenário 3: Alterações dos valores de μ iniciais, número de comunidades (C) e proporções em cada comunidade (θ)

Vamos considerar fixos o número de simulações e o número de vértices, iguais a 1000 e 100, respectivamente. Para essas simulações o intuito é avaliar a recuperação dos parâmetros reais variando o número e proporções de indivíduos nas comunidades e as medidas de probabilidade reais e iniciais de cada grupo.

A Tabela 3 expõe as estimativas médias e desvio-padrão para o caso em que o número de comunidades é 5 e o número de palavras é 3. Nota-se que mesmo quando as medidas de probabilidades reais e iniciais variam, as estimativas do algoritmo se aproximam dos valores originais, comprovando a eficiência do processo para mais que duas comunidades.

Tabela 3: Resultados das simulações para os parâmetros após aplicação do algoritmo considerando número de comunidades distintos - Cenário 3

100 vértices	Real				Inicial				Estimativas				
Sim. = 1000	θ	μ			θ	μ			θ	μ			Alocação
Comunidades		j = 1	j = 2	j = 3		j = 1	j = 2	j = 3		j = 1	j = 2	j = 3	
1	0,20	0,35	0,50	0,15	0,20	0,25	0,55	0,20	0,1949 (0,0245)	0,2541 (0,1340)	0,6087 (0,0956)	0,1164 (0,1000)	0,11
2	0,20	0,20	0,10	0,70	0,20	0,25	0,15	0,60	0,2113 (0,0259)	0,2143 (0,1107)	0,0853 (0,0812)	0,6885 (0,1246)	0,30
3	0,20	0,55	0,20	0,25	0,20	0,45	0,25	0,30	0,1799 (0,0226)	0,4491 (0,1316)	0,1989 (0,1194)	0,2731 (0,1636)	0,03
4	0,20	0,00	0,75	0,25	0,20	0,10	0,60	0,30	0,2061 (0,0205)	0,0396 (0,0381)	0,6703 (0,0971)	0,2870 (0,1683)	0,29
5	0,20	0,50	0,05	0,45	0,20	0,55	0,10	0,35	0,2079 (0,0178)	0,6363 (0,0857)	0,0434 (0,0456)	0,3868 (0,1806)	0,27
1	0,30	0,85	0,05	0,10	0,20	0,75	0,10	0,15	0,2577 (0,0293)	0,8229 (0,0655)	0,0178 (0,0293)	0,1076 (0,0856)	0,40
2	0,15	0,50	0,10	0,40	0,20	0,45	0,20	0,35	0,1869 (0,0210)	0,4784 (0,0860)	0,1045 (0,1074)	0,3410 (0,1069)	0,13
3	0,20	0,05	0,25	0,70	0,20	0,10	0,10	0,80	0,1737 (0,0287)	0,0923 (0,0918)	0,1597 (0,1095)	0,7861 (0,0906)	0,18
4	0,20	0,00	0,60	0,40	0,20	0,05	0,65	0,30	0,2097 (0,0245)	0,0303 (0,0510)	0,7209 (0,0767)	0,4563 (0,1641)	0,25
5	0,15	0,35	0,40	0,25	0,20	0,40	0,45	0,15	0,1721 (0,0257)	0,3764 (0,1251)	0,3223 (0,1155)	0,1484 (0,1370)	0,04
1	0,15	0,40	0,10	0,50	0,20	0,35	0,15	0,55	0,1663 (0,0212)	0,4889 (0,0907)	0,1597 (0,1513)	0,4676 (0,0946)	0,08
2	0,10	0,10	0,10	0,80	0,20	0,15	0,15	0,70	0,1504 (0,0204)	0,1533 (0,0989)	0,1048 (0,1079)	0,6728 (0,1158)	0,23
3	0,40	0,90	0,05	0,05	0,20	0,80	0,10	0,10	0,3311 (0,0363)	0,9170 (0,0458)	0,0281 (0,0315)	0,0250 (0,0281)	0,47
4	0,25	0,30	0,50	0,20	0,20	0,40	0,45	0,15	0,1887 (0,0277)	0,5076 (0,1077)	0,3927 (0,1149)	0,1111 (0,0951)	0,07
5	0,10	0,00	0,60	0,40	0,20	0,00	0,55	0,45	0,1635 (0,0261)	0,0000 (0,0000)	0,6286 (0,1276)	0,3370 (0,1900)	0,15
1	0,10	0,00	0,70	0,30	0,20	0,10	0,65	0,25	0,1457 (0,0212)	0,1275 (0,0968)	0,6048 (0,1312)	0,1642 (0,1391)	0,12
2	0,20	0,34	0,33	0,33	0,20	0,35	0,30	0,35	0,1886 (0,0204)	0,5135 (0,0767)	0,1556 (0,1356)	0,3915 (0,1322)	0,15
3	0,10	0,10	0,40	0,60	0,20	0,00	0,50	0,50	0,1559 (0,0234)	0,0000 (0,0000)	0,3283 (0,1391)	0,6179 (0,1230)	0,10
4	0,30	0,80	0,10	0,10	0,20	0,60	0,20	0,20	0,2653 (0,0269)	0,8104 (0,0601)	0,1024 (0,0899)	0,0910 (0,0717)	0,35
5	0,30	0,20	0,00	0,80	0,20	0,15	0,15	0,70	0,2444 (0,0251)	0,1812 (0,1019)	0,0343 (0,0413)	0,8334 (0,0659)	0,28

6.4 Cenário 4: Alterações dos valores de μ iniciais, número de palavras (m) e proporções de vértices em cada comunidade (θ)

A Tabela 4 mostra os resultados das 1000 simulações para o cenário em que o número de palavras é igual a 7, o número de indivíduos é 100, considerando duas comunidades com proporções reais e medidas de probabilidade reais e iniciais distintas.

O foco das simulações com essas condições é analisar a exatidão do algoritmo para recuperar os parâmetros reais alterando o número de palavras, proporções de elementos nos grupos e as medidas de probabilidade reais e iniciais de cada grupo. Constata-se que as estimativas obtidas foram próximas dos valores reais, tanto para as proporções de indivíduos nas comunidades, quanto para as medidas de probabilidades.

Tabela 4: Resultados das simulações para os parâmetros após aplicação do algoritmo considerando número de palavras distintos - Cenário 4

100 vértices		Parâmetros								
		θ	μ							Alocação
Simulações = 1000			j = 1	j = 2	j = 3	j = 4	j = 5	j = 6	j = 7	
Valores reais	C1	0.70	0.10	0.10	0.20	0.00	0.10	0.10	0.40	-
	C2	0.30	0.70	0.00	0.00	0.20	0.00	0.10	0.00	
	C1	0.80	0.05	0.05	0.05	0.05	0.80	0.00	0.00	
	C2	0.20	0.00	0.05	0.20	0.60	0.05	0.00	0.10	
	C1	0.40	0.10	0.05	0.10	0.30	0.20	0.05	0.20	
	C2	0.60	0.25	0.60	0.00	0.00	0.00	0.15	0.00	
	C1	0.51	0.10	0.05	0.00	0.05	0.40	0.20	0.20	
	C2	0.49	0.30	0.10	0.20	0.30	0.05	0.00	0.05	
Valores iniciais	C1	0.50	0.15	0.05	0.15	0.05	0.10	0.15	0.35	-
	C2	0.50	0.60	0.05	0.00	0.15	0.05	0.15	0.00	
	C1	0.50	0.10	0.10	0.10	0.05	0.60	0.00	0.05	
	C2	0.50	0.05	0.00	0.15	0.65	0.05	0.05	0.05	
	C1	0.50	0.15	0.05	0.10	0.05	0.05	0.10	0.50	
	C2	0.50	0.20	0.10	0.15	0.25	0.15	0.10	0.05	
	C1	0.50	0.05	0.10	0.05	0.05	0.45	0.15	0.15	
	C2	0.50	0.35	0.15	0.15	0.25	0.00	0.05	0.05	
Estimativas	C1	0,6725 (0,0857)	0,0906 (0,0456)	0,0994 (0,0411)	0,2135 (0,0593)	0,0001 (0,0018)	0,0990 (0,0394)	0,0981 (0,0403)	0,4255 (0,0885)	0,6300
	C2	0,3275 (0,0857)	0,7038 (0,1677)	0,0097 (0,0219)	0,0000 (0,0000)	0,1958 (0,0890)	0,0094 (0,0205)	0,1006 (0,0649)	0,0000 (0,0000)	0,3700
	C1	0,8100 (0,0924)	0,0462 (0,0262)	0,0622 (0,0399)	0,0514 (0,0295)	0,0554 (0,0371)	0,7943 (0,0897)	0,0000 (0,0000)	0,0023 (0,0070)	0,7900
	C2	0,1900 (0,0924)	0,0085 (0,0288)	0,0000 (0,0000)	0,2168 (0,1365)	0,6771 (0,2410)	0,0531 (0,1036)	0,0000 (0,0000)	0,1147 (0,1020)	0,2100
	C1	0,3836 (0,1323)	0,1018 (0,0790)	0,0644 (0,1548)	0,0994 (0,0727)	0,2966 (0,1427)	0,2019 (0,1143)	0,0531 (0,0576)	0,2214 (0,1235)	0,3000
	C2	0,6164 (0,1323)	0,2347 (0,0741)	0,5562 (0,1998)	0,0136 (0,0375)	0,0442 (0,1229)	0,0290 (0,0809)	0,1402 (0,0581)	0,0120 (0,0335)	0,7000
	C1	0,5658 (0,1530)	0,1151 (0,0612)	0,0544 (0,0408)	0,0160 (0,0249)	0,0708 (0,0499)	0,4331 (0,1515)	0,1808 (0,0709)	0,1866 (0,0672)	0,5500
	C2	0.4342 (0.1530)	0.3170 (0.1335)	0.1035 (0.0754)	0.2309 (0.1463)	0.3318 (0.1548)	0.0000 (0.0000)	0.0100 (0.0273)	0.0497 (0.0531)	0.4500

7 Aplicação do modelo de redes de afinidade a dados sobre sentimento a respeito da Covid-19

Por meio das simulações apresentadas na seção anterior verificamos que o método proposto tem funcionamento comprovado, visto que para cenários diversificados as informações reais foram recuperadas satisfatoriamente. Desta forma, podemos aplicar essa metodologia para situações práticas do mundo real, corroborando para o objetivo geral do projeto.

No ano de 2020 teve início a pandemia da covid-19 no Brasil, doença esta causada pelo coronavírus, denominado SARS-CoV-2, que apresenta um espectro clínico variando de infecções assintomáticas a quadros graves, segundo fonte Ministério da Saúde. Fato é que a pandemia vem produzindo repercussões não apenas de ordem biomédica e epidemiológica em escala global, mas também repercussões e impactos sociais, econômicos, políticos, culturais e históricos (Impactos sociais, econômicos, culturais e políticos da pandemia, Fundação Oswaldo Cruz, 2021).

Desta maneira, investigando os efeitos que este surto vem provocando nas pessoas, foi realizada uma pesquisa de opinião simples a respeito da "Percepção de impacto da pandemia de covid-19 na comunidade". Composto por duas perguntas, o estudo tinha o objetivo de descrever como a comunidade se sente afetada em relação à pandemia de covid-19 de uma maneira geral e identificar se existem grupos de pessoas que se sentem afetadas de forma semelhante. Foram construídos dois formulários, um direcionado a pessoas com vínculo com a Universidade Federal de Minas Gerais (UFMG) e outro para pessoas externas (geral).

À vista disso, o respondente assinalava a qual categoria profissional ele pertencia dentre empresário(a), profissional liberal, empregado(a) setor privado, empregado(a) setor público, estudante ou outro para pessoas não vinculadas a UFMG, e professor(a), aluno(a) graduação, aluno(a) pós-graduação ou outro para a comunidade universitária. Em seguida era pedido para responder por meio de palavras (frases não eram possíveis) de que forma a pandemia de covid-19 tem afetado a sua vida.

7.1 Análise descritiva da amostra

Neste projeto 285 pessoas responderam ao formulário, sendo 183 pertencentes a comunidade externa à UFMG e 102 à comunidade interna. A comunidade geral escolheu ao todo 795 palavras e a UFMG 444 palavras distintas.

A Tabela 5 apresenta a análise descritiva das categorias profissionais para os respondentes. Nota-se que dentre aqueles que têm vínculo com a universidade, a maior parte da amostra foi

composta por alunos(as) de graduação (48,04%), ao passo que para não vinculados ao campus, empregado(as) do setor privado foram a maioria (29,51%).

Tabela 5: Caracterização da amostra

	Profissão	Frequência Absoluta (N)	Frequência Relativa (%)
Geral	Empregado(a) setor privado	54	29,51%
	Empregado(a) setor público	40	21,86%
	Empresário(a)	13	7,10%
	Estudante	19	10,38%
	Outro	20	10,93%
	Profissional liberal	37	20,22%
	Total	183	100,00%
UFMG	Aluno(a) graduação	49	48,04%
	Aluno(a) pós-graduação	20	19,61%
	Outro	2	1,96%
	Professor(a)	31	30,39%
	Total	102	100,00%

A Tabela 6 exibe a descrição do número de palavras escolhidas pelas pessoas em cada grupo. Na comunidade geral o número máximo de palavras escolhidas por um indivíduo foi 19, com moda 1 e na UFMG foi 16 o número máximo e moda igual 4.

Tabela 6: Descrição do número de palavras escolhidas

Nº de palavras escolhidas	Geral		UFMG	
	Freq. Absoluta (N)	Freq. Relativa (%)	Freq. Absoluta (N)	Freq. Relativa (%)
1	44	24,04%	16	15,69%
2	15	8,20%	13	12,75%
3	19	10,38%	14	13,73%
4	37	20,22%	17	16,67%
5	18	9,84%	16	15,69%
6	12	6,56%	8	7,84%
7	19	10,38%	5	4,90%
8	4	2,19%	3	2,94%
9	2	1,09%	5	4,90%
10	4	2,19%	2	1,96%
11	2	1,09%	-	-
12	-	-	2	1,96%
14	3	1,64%	-	-
15	1	0,55%	-	-
16	1	0,55%	1	0,98%
17	1	0,55%	-	-
19	1	0,55%	-	-

A Figura 3 e a Tabela 7 expõem as palavras com maior frequência para a comunidade geral. Destaca-se que o termo “Medo” foi escolhido por 63 pessoas (corresponde a 7,92% da quantidade total de palavras mencionadas), seguidos por “Isolamento”, “Ansiedade”, “Tristeza” e “Insegurança”.

Tabela 7: Palavras com maiores frequências - Geral

Palavras	Frequência Absoluta (N)	Frequência Relativa (%)
Medo	63	7,92%
Isolamento	48	6,04%
Ansiedade	41	5,16%
Tristeza	22	2,77%
Insegurança	20	2,52%
Angústia	19	2,39%
Reflexão	19	2,39%
Saudade	19	2,39%
Solidão	19	2,39%
Família	18	2,26%

Figura 3: Nuvem de palavras - Geral

Na Figura 4 e na Tabela 8 estão descritas as principais palavras escolhidas no grupo UFMG, sendo que a palavra “Medo” foi escolhida por 29 pessoas (6,53% da quantidade total de palavras menciondas), seguidas por “Ansiedade” (5,63%), “Isolamento” (5,41%), “Cansaço” (4,95%) e “Preocupação” (3,38%).

Tabela 8: Palavras com maiores frequências - UFMG

Palavras	Frequência Absoluta (N)	Frequência Relativa (%)
Medo	29	6,53%
Ansiedade	25	5,63%
Isolamento	24	5,41%
Cansaço	22	4,95%
Preocupação	15	3,38%
Tristeza	15	3,38%
Solidão	13	2,93%
Insegurança	12	2,70%
Angústia	11	2,48%
Ausência de perspectiva	9	2,03%

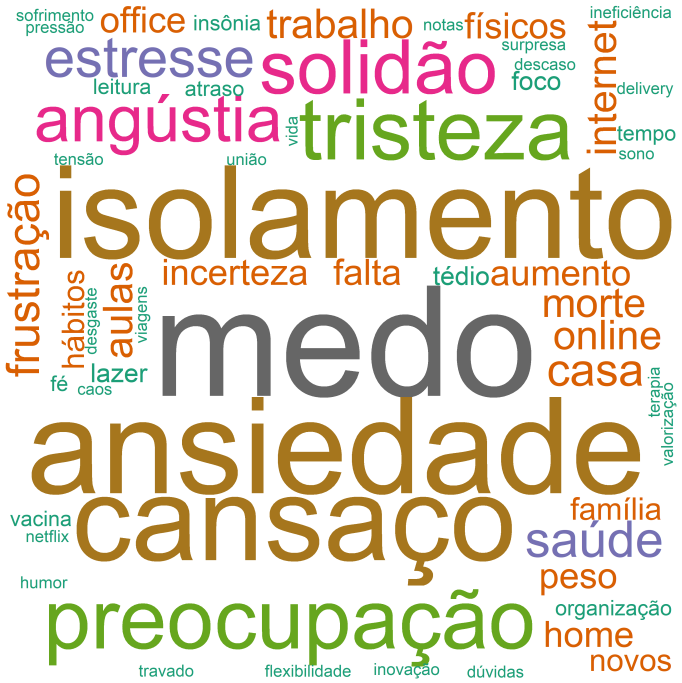


Figura 4: Nuvem de palavras - UFMG

7.2 Seleção do número de comunidades da rede de afinidade

As matrizes de escolhas dos indivíduos das comunidades geral e UFMG têm dimensão 183×145 e 102×123, respectivamente, ou seja, número de indivíduos por total de palavras que compõem o dicionário. A análise conjunta também será feita, denominamos essa comunidade como "total", composta pela junção dos grupos geral e UFMG. A matriz U tem dimensões 285×181, visto que 285 pessoas responderam à pergunta e o dicionário total é composto por 181 termos. Como mencionado na seção 2.1, se a i-ésima pessoa selecionou a m-ésima palavra, o elemento $U_{i,m}$ recebe valor 1, caso contrário recebe 0.

Construídas as matrizes U, o próximo passo é determinar o número ótimo de subcomunidades calculando os valores de BIC por meio do algoritmo EM. Para viabilizar a execução do mesmo fornecendo a menor quantidade de informação possível, as proporções iniciais seguiram distribuição uniforme, ou seja, $1/Q$ para cada grupo. As medidas de probabilidades iniciais das comunidades geral e UFMG foram geradas com base em valores aleatórios de uma distribuição Dirichlet com parâmetros α iguais as frequências relativas das palavras.

Na Tabela 9 é possível observar as principais estimativas das palavras para os casos geral e UFMG considerando uma, duas e três comunidades. A escolha da palavra "medo" foi estimada como 34,43% e 28,43% para as comunidades geral e UFMG, respectivamente, considerando apenas uma subcomunidade.

Tabela 9: Estimativas das medidas de probabilidades - Geral e UFMG

Comunidade	Palavras	Número de comunidades = 1	Número de comunidades = 2		Número de comunidades = 3		
		C1	C1	C2	C1	C2	C3
Geral	Medo	0,3443	0,3060	0,4910	0,2970	0,4930	0,4020
	Isolamento	0,2623	0,2760	0,2090	0,2420	0,2400	0,4120
	Ansiedade	0,2240	0,1750	0,4100	0,1710	0,5040	0,1270
	Tristeza	0,1202	0,0940	0,2180	0,0970	0,1060	0,2730
	Insegurança	0,1093	0,1140	0,0900	0,1230	0,0670	0,0870
	Angústia	0,1038	0,0720	0,2240	0,0790	0,2750	0,0000
	Reflexão	0,1038	0,0920	0,1490	0,0630	0,0850	0,3670
	Saudade	0,1038	0,1310	0,0000	0,1220	0,0000	0,1490
	Solidão	0,1038	0,0840	0,1780	0,0810	0,1460	0,1760
	Família	0,0984	0,0960	0,1070	0,0400	0,0640	0,4880
UFMG	Medo	0,2843	0,4080	0,2140	0,4140	0,0000	0,4130
	Ansiedade	0,2451	0,1610	0,2920	0,3040	0,1830	0,1790
	Isolamento	0,2353	0,3820	0,1530	0,1350	0,2730	0,4730
	Cansaço	0,2157	0,3440	0,1430	0,3210	0,0000	0,2920
	Preocupação	0,1471	0,1870	0,1250	0,1490	0,1290	0,1760
	Tristeza	0,1471	0,0290	0,2140	0,1890	0,1250	0,0580
	Solidão	0,1275	0,1480	0,1160	0,0860	0,1070	0,2920
	Insegurança	0,1176	0,2150	0,0630	0,1500	0,0000	0,2350
	Angústia	0,1078	0,1240	0,0990	0,0870	0,1050	0,1760
	Ausência de perspectiva	0,0882	0,1100	0,0760	0,1510	0,0000	0,0590

Os valores dos BIC's obtidos para as duas situações (geral e UFMG) podem ser vistos na Tabela 10. Pode-se coligir que para os grupos geral e UFMG não há subcomunidades, visto que o maior BIC está no caso 1.

Tabela 10: Resultados dos BIC's

Comunidades	Geral	UFMG
1 comunidade	1055,98	406,16
2 comunidades	294,55	-167,92
3 comunidades	-445,20	-741,41

7.3 Grafos aleatórios para as comunidades geral e UFMG

As Figuras 5 e 6 mostram os grafos aleatórios para as comunidades geral e UFMG resultantes das funções afinidade binária e cardinal.

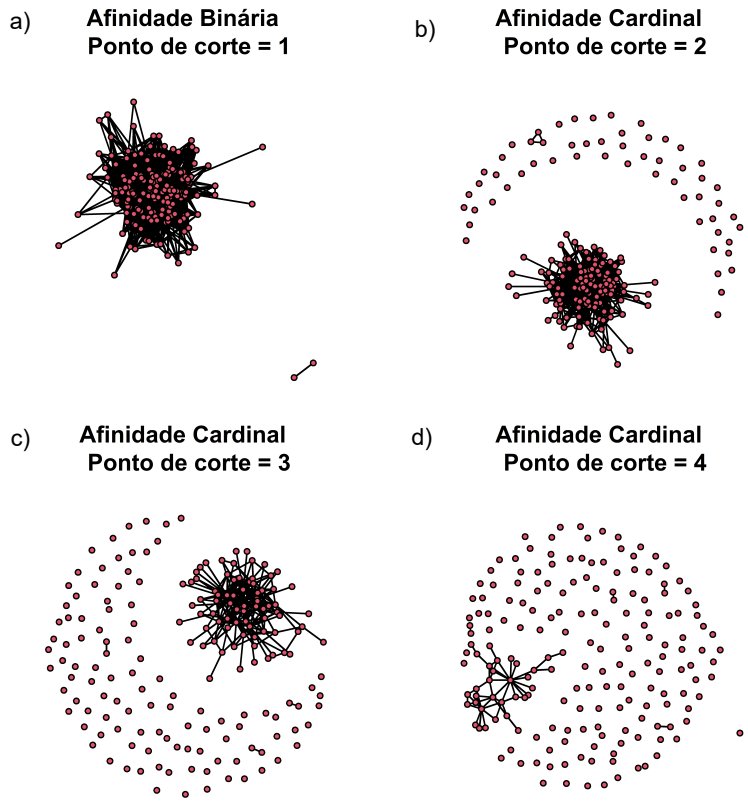


Figura 5: Grafos aleatórios - Geral

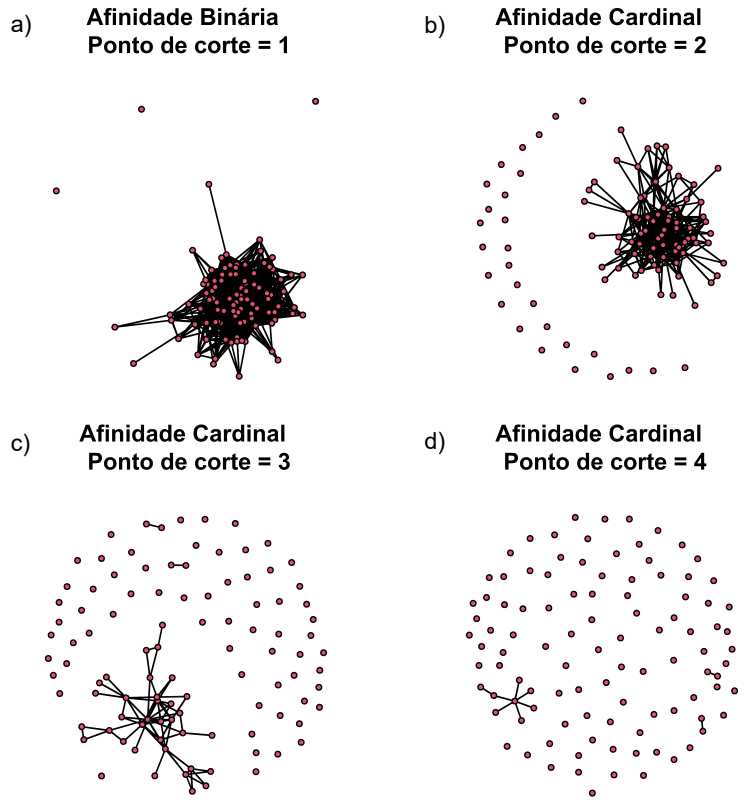


Figura 6: Grafos aleatórios - UFMG

7.4 Aplicação do método para o caso total

Para o caso total que avalia as comunidades geral e UFMG de forma conjunta, especificamos o número de comunidades igual a 2, pois há uma comunidade no geral e uma na UFMG, e as medidas iniciais seguiram distribuição uniforme. Definido o número de comunidades, foi feita a alocação dos indivíduos com base nos cálculos de verossimilhanças individuais, desta maneira 43,51% dos indivíduos pertencem a comunidade 1 e 56,49% a comunidade 2. Posto isso, a Tabela 11 exibe as estimativas das medidas de probabilidades para as duas comunidades no caso total. Verifica-se que para a comunidade 1, o termo "isolamento" teve estimativa de 30,60%, enquanto que para a comunidade 2, 21,50%.

Tabela 11: Estimativas das medidas de probabilidades - Total

Palavras	Comunidade 1	Comunidade 2
Medo	0,3520	0,3020
Isolamento	0,3060	0,2150
Ansiedade	0,2170	0,2420
Tristeza	0,1660	0,1040
Reflexão	0,1460	0,0460
Família	0,1410	0,0320
Empatia	0,1350	0,0180
Solidão	0,1320	0,0980
Angústia	0,1250	0,0910
Novos hábitos	0,1170	0,0490

7.5 Análise descritiva dos grafos aleatórios

A Tabela 12 expõe a análise descritiva dos grafos aleatórios para a função afinidade binária. Nota-se que a maior densidade (o quanto os vértices são porporcionamente conectados) foi para o grafo da comunidade UFMG, a maior transitividade (mede a presença de triângulos na rede – conjunto de três vértices interconectados entre si) para a geral e a maior proximidade global (definida como o inverso da soma das distâncias entre os vértices) foi para o total.

Tabela 12: Análise descritiva dos grafos - Função afinidade binária

Modelos	Vértices	Arestas	Grau máximo	Densidade	Transitividade	Proximidade global
Geral	183	5107	134	0,3067	0,6678	0,6329
UFMG	102	1583	69	0,3081	0,6274	0,6126
Total	285	12217	204	0,3019	0,6587	0,6368

A análise descritiva dos grafos aleatórios para a função afinidade cardinal está na Tabela 13. Verifica-se que conforme o ponto de corte aumenta, todas as demais medidas diminuem, chegando ao ponto de ter somente uma aresta quando o ponto de corte é igual a 5 na comunidade UFMG.

Tabela 13: Análise descritiva dos grafos - Função afinidade cardinal

Ponto de corte	Modelos	Vértices	Arestas	Grau máximo	Densidade	Transitividade	Prox. global
2	Geral	183	1232	82	0,0740	0,4997	0,2540
	UFMG	102	374	38	0,0726	0,4529	0,2607
	Total	285	2882	134	0,0712	0,4817	0,2791
3	Geral	183	258	34	0,0155	0,3120	0,0810
	UFMG	102	69	16	0,0134	0,3790	0,0479
	Total	285	594	58	0,0147	0,3345	0,0961
4	Geral	183	56	14	0,0034	0,2672	0,0135
	UFMG	102	9	6	0,0017	0,0000	0,0036
	Total	285	108	24	0,0027	0,2081	0,0147
5	Geral	183	8	2	0,0005	0,0000	0,0006
	UFMG	102	1	1	0,0002	1,0000	0,0002
	Total	285	11	3	0,0003	0,0000	0,0004

As Figuras 7 e 8 apresentam os grafos aleatórios para o caso total, onde os pontos avermelhados representam os indivíduos alocados na comunidade 1 e e os pontos pretos na comunidade 2. Nota-se que conforme o ponto de corte aumenta, o número de arestas entre os indivíduos diminuem.

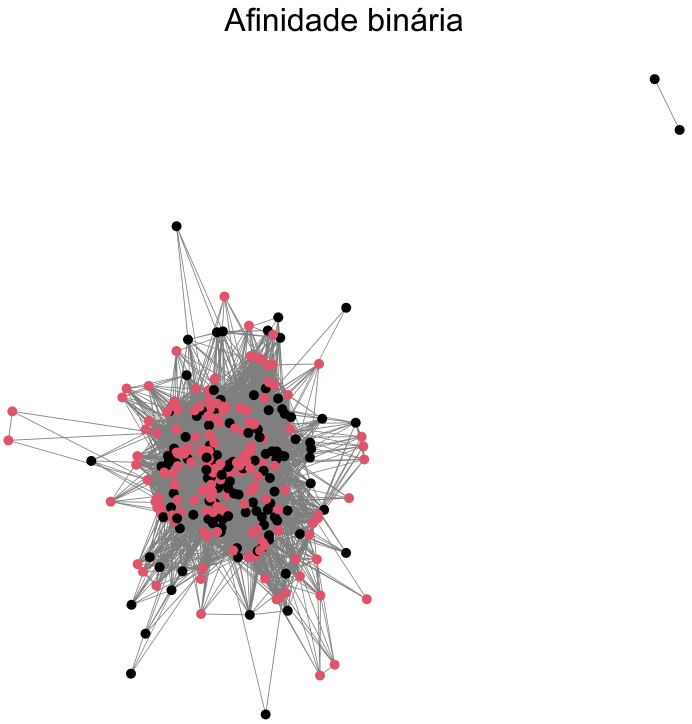


Figura 7: Grafo afinidade binária - Total

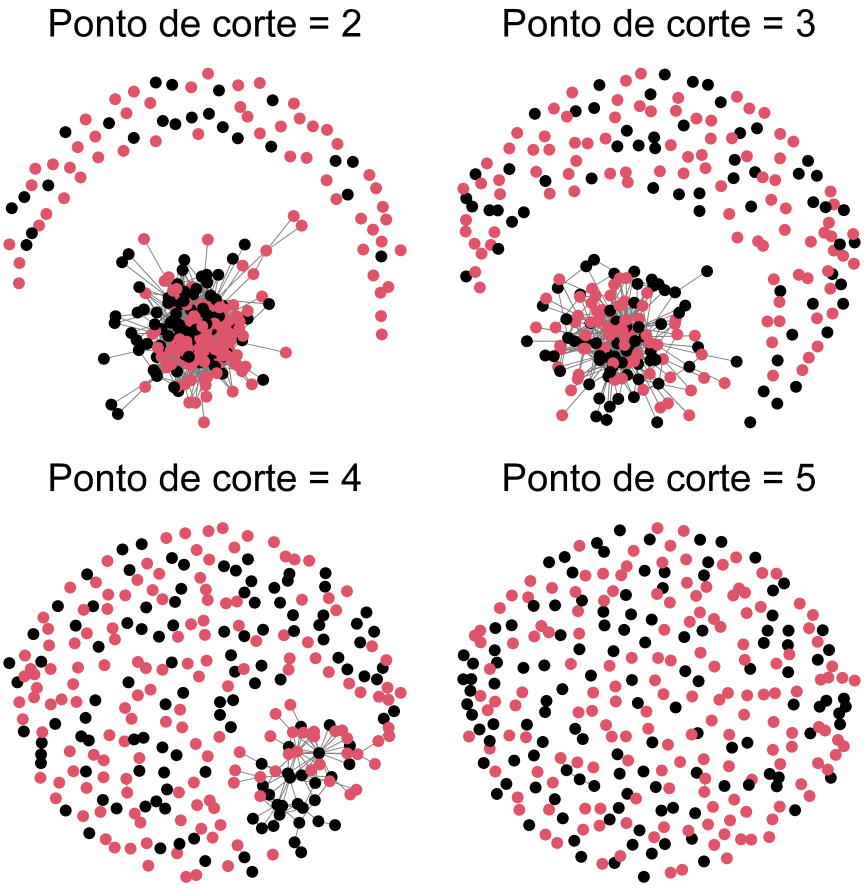


Figura 8: Grafos afinidade cardinal - Total

7.6 Resultados

Na comunidade 1 as palavras com maiores estimativas foram: medo, isolamento, ansiedade, tristeza, reflexão, família, empatia, solidão, angústia e novos hábitos, enquanto para a comunidade 2 foram: medo, ansiedade, isolamento, cansaço, insegurança, tristeza, preocupação, solidão, angústia e saudade.

De modo geral, o sentimento em relação aos impactos provocados pela pandemia foi negativo, porém na comunidade 1 é possível observar que houveram mais aspectos positivos e de reflexão em relação a comunidade 2.

8 Conclusões

Neste trabalho inicialmente foi descrito o modelo de redes de afinidade proposto por Pereira (2020). Descrevemos os componentes do modelo, discutimos suas propriedades, construímos as funções de verossimilhança para alguns casos particulares e apresentamos alguns exemplos.

Em seguida mostramos formas de gerar os vetores U_i que regem o comportamento das escolhas dos indivíduos e o processo de inferência para o grafo G_λ.

Para cumprir o objetivo proposto de construir um processo de estimação com a maior exatidão possível, primeiramente computamos as verossimilhanças para as funções afinidades binária e cardinal, estendemos as redes de afinidade para blocos, obtivemos os estimadores para as medidas de probabilidades e proporções, e por fim, por meio do algoritmo EM modificado, realizamos o processo de estimação dos parâmetros. Para colocar em funcionamento esta metodologia é necessários como entrada a matriz de escolhas U, encontrar o número de comunidades da rede, indicar os valores iniciais das proporções e medidas para cada comunidades e a constante de tolerância (erro).

Após o processo de estimação executado, determinamos o melhor número de comunidades, alocamos os indivíduos às mesmas, aplicamos as funções afinidades e construímos os grafos de afinidade. Para verificar a acurácia da metodologia, realizamos simulações de diversos cenários visando a recuperação das medidas reais e obtivemos valores próximos. Também foi feita a aplicação para dados reais no contexto da pandemia da covid-19, onde encontramos resultados satisfatórios, podendo observar as alterações que as funções afinidades e os pontos de corte provocam ao serem alterados.

Deste modo, comprovamos que a técnica proposta de fato é adequada para estimar e recuperar parâmetros no contexto de modelos de redes de afinidade e pode ser aplicada em diversos tipos se situações reais, que envolvam características individuais ou de grupos, trazendo maior conhecimento sobre o comportamento dos indivíduos em uma rede, suas interações e também sobre as comunidades a que pertencem.

Este trabalho pode ser estendido para outros tipos de funções de afinidade e implementado em outras linguagens, como por exemplo Python, o que traria maiores avanços e visibilidades para esta classe de modelos.

9 Referências

DEMPSTER, A.; LAIRD, N.; RUBIN, D. *Maximum likelihood from incomplete data via the EM algorithm* Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, p. 1-38, 1977.

DUARTE, D; GUEDES, G.R.; RIBEIRO, R.B; SILVA, W.H.P *Representing the collective thinking through cognitive networks*, preprint, 2020.

ERDŐS, P.; RÉNYI, A. *On random graphs*, i. Publicationes Mathematicae (Debrecen), v. 6, p. 290–297, 1959.

FREEMAN, Linton C. *Some antecedents of social network analysis.* Connections, v. 19, n. 1, p. 39-42, 1996.

MARTINO, F; SPOTO, A. *Social Network Analysis: A brief theoretical review and further perspectives in the study of Information Technology.* PsychNology Journal. 4. 53-86, 2006.

PEREIRA, W. H. S. *Representação da estrutura do pensamento coletivo sobre as enchentes do rio doce: conectando indivíduos afins através da teoria dos grafos.* Trabalho de conclusão de curso em Estatística - Bacharelado. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística, 2017.

PEREIRA, W. H. S. *Modelo de rede de afinidades.* Dissertação do mestrado em Estatística. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística, 2020.

PORTELA, J. *Clustering Discrete Data Through the Multinomial Mixture Model*, Communications in Statistics - Theory and Methods, 37:20, 3250-3263, DOI: 10.1080/03610920802162623

SANTOS, G. T.; DIAS, J.M.B. *Teoria das Rrepresentações Sociais: uma abordagem sociopsicológica*, PRACS, vol. 8, no. 1, p. 173-187, 2015

SCHWARZ, G. *Estimating the dimensional of a model.* Annals of Statistics, Hayward, v.6, n.2, p.461-464, Mar. 1978.

SINGER, K. B. *Random Intersection Graphs.* PhD thesis, John Hopkins University (1995) 1997.

VERGÈS, P. *L'évocation de l'argent: une méthode pour la définition du noyau central de la représentation.* Bulletin de Psychologie, 45, 203-209, 1992.

Impactos sociais, econômicos, culturais e políticos da pandemia, Fundação Oswaldo Cruz, 2021. Disponível em: https://portal.fiocruz.br/impactos-sociais-economicos-culturais-e-politicos-da-pandemia. Acesso em: 10.abril.2021