UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS - ICEX
PÓS-GRADUAÇÃO EM ESTATÍSTICA

DISSERTAÇÃO

MODELO DE REDES DE AFINIDADE

ORIENTADORA: PROFA. DRA. DENISE DUARTE
CO-ORIENTADOR: PROF. DR. RODRIGO B. RIBEIRO

AUTOR: WESLEY HENRIQUE SILVA PEREIRA
MATRÍCULA: 2018665078

BELO HORIZONTE
FEVEREIRO DE 2020

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas - ICEx
Pós-Graduação em Estatística

Wesley Henrique Silva Pereira

Dissertação

Modelo de redes de afinidade

Dissertação apresentada ao Programa de Pós-Graduação em Estatística, como requisito parcial para a obtenção do Título de Mestre em Estatística, Instituto de Ciências Exatas da Universidade Federal de Minas Gerais

Orientadora: Profa. Dra. Denise Duarte Scarpa Magalhães Alves
Departamento de Estatística - ICEx - UFMG
Co-orientador: Prof. Dr. Rodrigo Botelho Ribeiro
Pontifícia Universidad Católica de Chile - Departamento de Matemática - Santiago, Chile

Belo Horizonte
Agosto de 2019

Aos meus pais, minha irmã e meus amigos.

“Um novo mandamento vos dou:
que vos ameis uns aos outros;
assim como Eu vos amei; que
dessa mesma maneira tenhais
amor uns para com os outros.”

Bíblia Sagrada, Jo 13:34

“Mas, se ergues da justiça a clava forte,
Verás que um filho teu não foge à luta,
Nem teme, quem te adora, a própria morte!”

Joaquim O. Duque-Estrada
Hino Nacional Brasileiro

RESUMO

Na literatura sobre dados relacionais, uma das abordagens mais populares atualmente é a Análise de Redes Complexas. Consequentemente, análises estatísticas sobre redes sociais buscaram acompanhar este crescimento para atender à esta demanda. Para modelar estatisticamente os fenômenos estudados em redes socais, buscou-se aproveitar modelos probabilísticos introduzidos para a modelagem em grafos. Entretanto, observou-se que redes sociais possuem características que são diferentes dos modelos de grafos aleatórios que possuem arestas independentes. Uma alternativa proposta na literatura aos grafos de arestas independentes é o modelo de grafos aleatórios de interseção (RIG). Neste, os atores da rede estão associados a um conjunto finito de características e são conectados toda vez que compartilham pelo menos uma destas características.

A proposta deste trabalho é apresentar e estudar uma generalização ao modelo de grafos aleatórios de interseção, à qual denominamos de modelo de redes de afinidades. Estendendo as suposições assumidas no RIG, obtemos uma vasta família de modelos, onde as conexões são valoradas segundo uma função que mensura a afinidade entre os atores da rede. Além disso, as conexões são realizadas a partir de um determinado nível desta função afinidade, e não mais através do simples compartilhamento de uma característica. Para exemplificar o estudo do comportamento do modelo de redes de afinidades, elaboramos um estudo simulado baseado em simulações de Monte Carlo em uma das funções de afinidade, realizando *tuning* nos parâmetros geradores do modelo, analisando suas medidas topológicas e comparando as mesmas com as medidas topológicas encontradas em grafos com a mesma distribuição de afinidade, mas com arestas sorteadas independentemente.

Palavras-chave: Afinidade, Redes sociais, Interseção, Grafo aleatório

ABSTRACT

In the literature on relational data, one of the most popular approaches is Complex Network Analysis. Consequently, the statistical analysis of social networks sought to accompany this growth to meet this demand. To statistically model the phenomena studied in social networks, we sought to seize probabilistic models introduced for graph modeling. However, we observed that social networks possess characteristics that are different from random graph models that hold independent edges. An alternative proposed in the literature to independent edge graphs is the random intersection graph model (RIG). In this, network actors are associated with a finite set of features and are connected every time they share at least one of these features.

The purpose of this study is to present and study a generalization to the random intersection graph model, which we call Affinity Network Model. Extending the assumptions made in the RIG, we get a wide family of models, where connections are valued according to a function that measures the affinity between the actors in the network. Beside, connections are made from a certain level of this affinity function, and no longer by simply sharing a characteristic. To illustrate the study of the behavior of the network of affinities model, we have prepared a simulated study based on Monte Carlo simulations in one of the affinity functions: tuning the generating parameters of the model, analyzing its topological measurements and comparing them with the topological measurements found in graphs with the same affinity distribution, but with edges drawn independently.

Keywords: Affinity, Social networks, Intersection, Random graph

SUMÁRIO

1 INTRODUÇÃO

Quando estamos interessados em analisar dados relacionais, uma das abordagens que vem ganhando destaque nos últimos trinta anos é a Análise de Redes Sociais (MATHEUS; SILVA, 2006). Existem diversas formas de realizar análises sobre redes sociais, entre elas abordagens que utilizam modelagens estatísticas. Essas modelagens se aproveitam de vários modelos probabilísticos introduzidos na literatura para modelagem em grafos. No entanto, uma parte significativa dos modelos em grafos como, por exemplo, Erdős e Rényi (1959) assumem que as arestas são independentes. No entanto, de acordo com Newman e Park (2003), há diferenças importantes entre redes sociais e grafos que supõem arestas independentes. No caso das redes sociais, eles argumentam que a probabilidade de dois atores estarem conectados dado que estes atores se conectam a um terceiro ator em comum é maior do que a probabilidade do mesmo evento em grafos com arestas independentes. Esta diferença fundamental tem implicações significativas sobre a configuração das conexões da rede, tornando os modelos convencionais não adequados para a modelagem das redes sociais.

Uma abordagem alternativa para modelos em grafos aleatórios foi proposta por Singer (1995). Neste modelo, as arestas não são aleatórias, mas função de características aleatórias dos atores regidas por variáveis aleatórias Bernoulli independentes e identicamente distribuídas. Sempre que um par de atores compartilha uma ou mais características, eles estarão conectados no grafo. Alguns trabalhos na literatura, como Guedes et al. (2018) utilizam ideias similares para conectar indivíduos com características similares. Neste exemplo em específico, os autores conectam indivíduos através de palavras coletadas pela Técnica de Associação Livre de Palavras (TALP) - uma técnica de captura do pensamento leigo sobre algum tema derivada do Teste de Associação Livre de Jung (1910) - sobre o tema Zika Vírus, onde o conjunto de todas as palavras evocadas pelos entrevistados são equivalentes às características aleatórias do modelo de Singer (1995). Os indivíduos foram conectados sempre que compartilharam ao menos uma palavra, e cada conexão existente foi valorada pelo coeficiente de afinidade cognitiva introduzido por Pereira (2017), considerando ordem de evocação das palavras e diferenças entre as ordens de evocação das palavras de indivíduo para indivíduo. Tanto no trabalho de Pereira (2017) quanto no de Guedes et al. (2018) os autores ainda decompõem a rede gerada em *clusters* denominados comunidades, analisando descritivamente através de métricas da rede quais seriam as palavras mais relevantes para cada comunidade.

Este trabalho objetiva estender tanto o trabalho de Singer (1995) quanto o de Pereira (2017), introduzindo uma família de funções denominada por função afinidade. Em Singer, generalizaremos o modelo violando a suposição de que as características dos indivíduos sejam independentes e identicamente distribuídas e estudaremos o comportamento geral dos grafos aleatórios gerados. Além disso, em Pereira, definiremos o conceito da função de afinidade, uma

vasta família de funções que será responsável tanto por valorar as conexões sob algum critério como permitirá ser mais rigoroso em especificar a partir de qual limiar consideramos que há afinidade suficiente para formar uma aresta no grafo aleatório.

2 OBJETIVOS

Os objetivos propostos para este trabalho são

1. Introduzir o modelo de redes de afinidade, definindo seus parâmetros, apresentando suas principais características e criando um método para gerar grafos aleatórios para este modelo;

2. Mostrar a relação entre o modelo de redes de afinidade e outros modelos de grafos na literatura;

3. Através de simulações via método de Monte Carlo, realizar *tuning* em um dos modelos de redes de afinidade, analisando o comportamento das principais medidas topológicas resultantes e observando a função que cada parâmetro exerce no perfil destas medidas topológicas;

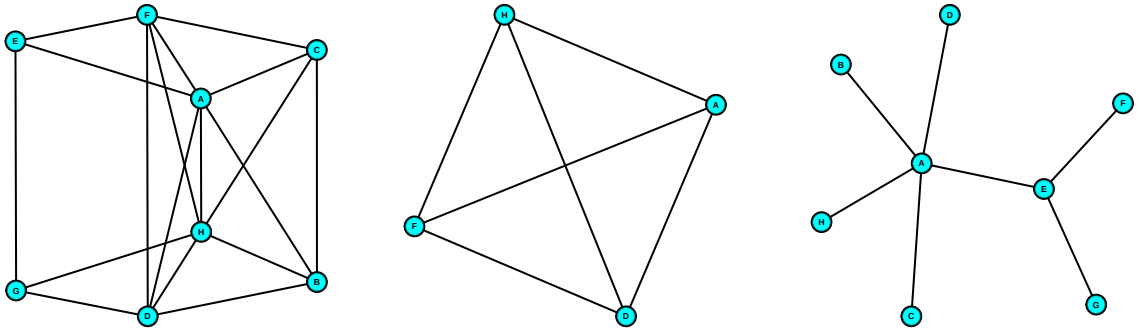
4. Encontrar e explorar diferenças entre as topologias dos modelos de redes de afinidade e modelos cujas arestas são geradas independentemente.

3 ALGUMAS DEFINIÇÕES EM TEORIA DOS GRAFOS

Neste capítulo, serão apresentados alguns conceitos em grafos que serão muito importantes em todo este trabalho. O uso de grafos para descrever uma rede é algo natural na literatura devido à similaridade de suas estruturas. Pode-se dizer que a única diferença entre um grafo e uma rede é a razão para sua existência: enquanto o grafo é apenas um objeto matemático, assim como um número natural, as redes geralmente estão associadas a um fenômeno, isto é, uma rede é uma observação do fenômeno da natureza.

Vamos começar apresentando a definição de um grafo: um grafo $G(V,E)$ é uma coleção de elementos divididos em dois conjuntos: o primeiro conjunto denotado por $V(G) = \{v_1, v_2, \ldots, v_n\}$ contém a unidade fundamental de um grafo denominado *vértice* ou *nó*. O segundo conjunto denotado por $E(G) = \{e_{i,k} : i \neq k\}$ com $i,k = \{1,2,3,\ldots,n\}$ contém as conexões entre um par de vértices. Essas conexões são denominadas *arestas*. Também denotaremos $v_i \to v_k$ se há uma aresta que vai de v_i para v_k e $v_i \nrightarrow v_k$ caso contrário. Ao longo do presente texto, lidaremos com grafos não-direcionados, isto é, $v_i \to v_k \Leftrightarrow v_k \to v_i$. Ou seja, denotaremos $v_i \leftrightarrow v_k$ se existe uma aresta entre v_i e v_k e $v_i \nleftrightarrow v_k$ caso contrário.

Definido o objeto grafo, vamos apresentar algumas das medidas topológicas que utilizaremos ao longo deste trabalho. Para começar, vamos apresentar na Figura 1 alguns exemplos de grafos. A partir daí, vamos avaliar as medidas topológicas tomando tais grafos para mostrar exemplos de mensuração destas medidas topológicas.



(a) Exemplo G_1 **(b)** Exemplo G_2 **(c)** Exemplo G_3

Figura 1 – Alguns exemplos de grafos

(I) A *ordem* de um grafo é a *cardinalidade* de $V(G)$ denotada por $|V(G)|$. Alguns autores definem o *tamanho* de um grafo como a cardinalidade de E(G), denotada por $|E(G)|$. Outros autores o definem como $|V(G)| + |E(G)|$. Neste trabalho será utilizada a primeira

definição. Nos exemplos apresentados na Figura 1, $|V(G_1)| = |V(G_3)| = 8$, $|V(G_2)| = 4$, $|E(G_1)| = 18$, $|E(G_2)| = 6$ e $|E(G_3)| = 7$.

(II) Um *subgrafo* de G é um outro grafo H tal que $V(H) \subset V(G)$, $E(H) \subset E(G)$. Além disso, se $e_{i,k} \in E(G)$, então if $e_{i,k} \in E(H) \Rightarrow \{v_i, v_k\} \in V(H)$. Nos exemplos apresentados na Figura 1, G_1, G_2 e G_3 são subgrafos de G_1.

(a) Se H é um subgrafo de G e $H \neq G$, então H é denominado de subgrafo *próprio*. Nos exemplos apresentados na Figura 1, G_2 e G_3 são subgrafos próprios de G_1.

(b) Se H é um subgrafo de G em que $e_{i,k} \in E(G)$ e $e_{i,k} \in E(H) \Leftrightarrow \{v_i, v_k\} \in V(H)$, então H é denominado subgrafo *induzido*. Nos exemplos apresentados na Figura 1 temos que G_2 é subgrafo induzido de G_1.

(c) Se H é um subgrafo de G e $V(G) = V(H)$, então H é denominado subgrafo *gerador*. Nos exemplos apresentados na Figura 1, G_3 é subgrafo gerador de G_1.

(III) Um grafo G é *completo* se, e somente se, existe uma aresta entre cada par de vértices de G. Nos exemplos apresentados na Figura 1, G_2 é um grafo completo.

(IV) Um *clique* de G é um subgrafo completo de G. Um *k-clique* é um clique de ordem k. O número $\kappa(G)$ é definido como a ordem do maior clique de G. Por conveniência, vamos denotar o número de k-cliques em G como $|\Phi_k(G)|$, onde $\Phi_k(G) \subset G$ é o conjunto de todos os k-cliques em G. Nos exemplos apresentados na Figura 1, G_2 é um 4-clique de G_1. Além disto, $|\Phi_1(G)| = 8$, $|\Phi_2(G)| = 18$, $|\Phi_3(G)| = 14$ e $|\Phi_4(G)| = 4$.

(V) Um grafo é dito *ponderado* se existem pesos associados aos seus vértices ou arestas. Neste trabalho trabalharemos com pesos associados às arestas, denotados por $w = \{w_{i,k} : i \neq j\}$ com $i,k = \{1,2,3,\ldots,n\}$.

(VI) O *grau* ou *valência* de um vértice é o número de arestas que incidem sobre aquele vértice, denotado por $deg(v_i)$. De fato,

$$\sum_i deg(v_i) = 2 \cdot |E(G)|. \tag{3.1}$$

No grafo G_3 apresentado na Figura 1, $deg(A) = 5$, $deg(E) = 3$ e $deg(B) = deg(C) = deg(D) = deg(F) = deg(G) = deg(H) = 1$.

(VII) A *força* de um vértice em um grafo ponderado é definida como a soma dos pesos de todas as arestas que incidem sobre o vértice.

(VIII) Um *passeio* é definido como uma sequência de vértices e arestas tal que

$$(v_{(i_0)}, e_{(i_0,i_1)}, v_{(i_1)}, e_{(i_1,i_2)}, v_{(i_2)}, \ldots, e_{(i_{n-2},i_{n-1})}, v_{(i_{n-1})}, e_{(i_{n-1},i_n)}, v_{(i_n)}). \tag{3.2}$$

O tamanho do passeio é definido como seu número de arestas. Uma *trilha* é definida com um passeio que não repete arestas. Um *caminho* é definido com um passeio que não

repete vértices. Por conveniência, vamos denotar o número de trilhas de G com tamanho k como $|\Psi_k(G)|$, onde $\Psi_k(G) \subset G$ é o conjunto de todos as k-trilhas em G.

(IX) Em um grafo não-ponderado G, a *distância* entre dois vértices é o tamanho do *menor* caminho entre eles. Se tal caminho não existe, definimos sua distância como infinita. No grafo G_3 apresentado na Figura 1, $dist(F,B) = 3$. Já em G_1, $dist(F,B) = 1$.

(X) A *proximidade* de um vértice i é definida como o inverso da soma das distâncias entre i e os outros vértices, isto é,

$$prox(i) = \frac{1}{\sum_{i \neq k} dist(i,k)}. \tag{3.3}$$

No grafo G_3 apresentado na Figura 1, $prox(A) = \frac{1}{5 \cdot 1 + 2 \cdot 2} = \frac{1}{9}$. A imagem da proximidade é trivial em 0 sempre que o grafo é desconexo, já que pelo menos uma de suas distâncias é infinita. Alternativamente, pode ser utilizada a soma do inverso das distâncias ao invés do inverso das somas das distâncias. Se não há um caminho entre i e k, como visto, sua distância é definida como infinita, então convenciona-se $\frac{1}{dist(x,y)} = \frac{1}{\infty} = 0$ de modo que

$$prox(i) = \sum_k \frac{1}{dist(i,k)}. \tag{3.4}$$

(XI) A *excentricidade* do vértice i é definida

$$exc(i) = \max_k dist(i,k). \tag{3.5}$$

isto é, a distância máxima entre ele e qualquer outro vértice de. Nos exemplos apresentados na Figura 1 temos que em G_3, A e E tem excentricidade 2, enquanto os demais vértices tem excentricidade 3.

(XII) Um grafo é dito *conexo* se, e somente se, existe pelo menos um caminho entre cada par de vértices. Nos exemplos apresentados na Figura 1, todos os grafos são conexos.

(XIII) O *centro* de um grafo G é o conjunto de vértices com excentricidade mínima. Um grafo G está *centrado* em v_i se v_i pertence ao seu conjunto central. No grafo G_3 apresentado na Figura 1, A e E pertencem ao seu grupo central.

(XIV) A *densidade* de um grafo G é dada por

$$den(G) = \frac{2 \cdot |E(G)|}{(|V(G)| - 1) \cdot |V(G)|} \tag{3.6}$$

e esta mensura o quanto os vértices são proporcionalmente conectados. Nos exemplos apresentados na Figura 1, $den(G_1) = 0.6428571$, $den(G_2) = 1$ e $den(G_3) = 0.25$.

(XV) A *transitividade* do grafo G é dada por

$$\tau(G) = 3 \cdot \frac{|\Phi_3(G)|}{|\Psi_2(G)|}. \tag{3.7}$$

Em outras palavras, a transitividade corresponde à razão entre o número de triângulos e o número de 2-trilhas presentes no grafo. Nos exemplos apresentados na Figura 1, $\tau(G_1) = 0.6176471$, $\tau(G_2) = 1$ e $\tau(G_3) = 0$.

(XVI) Vamos definir por conveniência $|\psi_k(x)|$ o número de k-trilhas centradas em x, enquanto $|\phi_k(x)|$ o número de k cliques que contém x. O *coeficiente de clustering local* do vértice i é definido por

$$C_i(G) = \frac{|\phi_3(i)|}{|\psi_2(i)|}. \tag{3.8}$$

Se o grau do vértice i é igual a 0 ou 1, considera-se $C_i(G) = 0$. No grafo G_1 apresentado na Figura 1, $C_A(G_1) = C_D(G_1) = C_F(G_1) = C_H(G_1) = 0,6$, $C_B(G_1) = C_C(G_1) = 0,8333$ e $C_E(G_1) = C_G(G_1) = 0,3333$. O *coeficiente de clustering global* do grafo G é definido por

$$C(G) = \frac{1}{n}\sum_i C_i(G), \tag{3.9}$$

de forma que $C(G_1) = 0,5917$.

(XVII) A *distribuição dos graus* de um grafo é definida como o comportamento das proporções dos vértices que tem grau k - denotado nesse trabalho como $p(k)$. Nos exemplos apresentados na Figura 1, em G_3, $p(1) = 0.75$, $p(3) = 0.125$ e $p(5) = 0.125$.

3.1 GRAFOS ALEATÓRIOS

Assim como as variáveis matemáticas têm sua variante aleatória, os grafos como objetos matemáticos também possuem sua variante no campo estocástico. Um *grafo aleatório* é uma distribuição sobre o conjunto dos grafos. Vamos a princípio apresentar um dos modelos seminais em grafos aleatórios. O modelo apresentado nessa seção servirá para introduzir e exemplificar a modelagem de incerteza sobre um objeto relacional como um grafo. Além disto, este modelo será bastante utilizado ao longo deste trabalho para realizar comparações entre seu comportamento e os modelos que serão propostos neste trabalho.

Um bom exemplo de modelo de grafos aleatórios são os modelos Erdős-Rényi. Estes modelos são considerados como os modelos seminais no que se refere à modelagem em grafos aleatórios. Na literatura, o nome Erdős-Rényi faz referência a dois modelos altamente relacionados, mas desenvolvidos independentemente.

3.1.1 MODELO POR ERDŐS-RÉNYI POR ERDŐS E RÉNYI

O primeiro modelo foi introduzido por Erdős e Rényi (1959). Seja $G(V,E)$ um grafo com $|V(G)| = n$. Além disso, vamos definir $E' = E'(G)$ como o conjunto de todas as arestas possíveis em G, enquanto que definimos $E = E(G)$ como uma configuração aleatória de arestas em G.

O número de pares de vértices em G pode ser escrito como $|E'| = \binom{n}{2}$. Consequentemente, o conjunto das partes de E' (isto é, o conjunto de todos os subconjuntos de E') denotado neste trabalho por $\mathscr{P}(E')$ tem cardinalidade

$$|\mathscr{P}(E')| = 2^{|E'|} = \sum_{M=0}^{|E'|} \binom{|E'|}{M}. \tag{3.10}$$

Assim, dando igual probabilidade a todos os elementos deste conjunto, a probabilidade de uma configuração específica $E \in \mathscr{P}(E')$ em G é

$$P[E] = (|\mathscr{P}(E')|)^{-1} = \left(\frac{1}{2}\right)^{|E'|}. \tag{3.11}$$

A probabilidade de escolher uma configuração de arestas em G que tem exatamente M arestas pode ser escrita por

$$P(|E| = M) = \frac{\binom{|E'|}{M}}{2^{|E'|}} \tag{3.12}$$

e a probabilidade condicional de escolher uma configuração específica $E \in \mathscr{P}(E')$ em G dado que $|E| = M$ pode ser escrita como

$$P_{|E|=M}(E) = \frac{1}{\binom{|E'|}{M}} = \frac{M! \cdot (|E'| - M)!}{|E'|!}. \tag{3.13}$$

Geralmente, este modelo aleatório é denotado na literatura por $G(n,M)$.

3.1.2 MODELO POR ERDŐS-RENYI POR GILBERT

O segundo modelo foi introduzido por Gilbert (1959). Sob as mesmas condições mencionadas para o modelo de Erdős e Rényi (1959), $\forall\, e_{i,k} \in E'$, $i,k = \{1,2,3,\ldots,n\}$, $i \neq k$, seja $0 < p < 1$ a probabilidade de $e_{i,k} \in E$. Se as arestas são independentes e identicamente distribuídas, a probabilidade de uma configuração específica $E \in \mathscr{P}(E')$ é

$$P(E) = p^{|E|} \cdot (1-p)^{|E'|-|E|}, \tag{3.14}$$

onde

$$\sum_{E \in \mathscr{P}(E')} P(E) = \sum_{E \in \mathscr{P}(E')} p^{|E|} \cdot (1-p)^{|E'|-|E|} = 1. \tag{3.15}$$

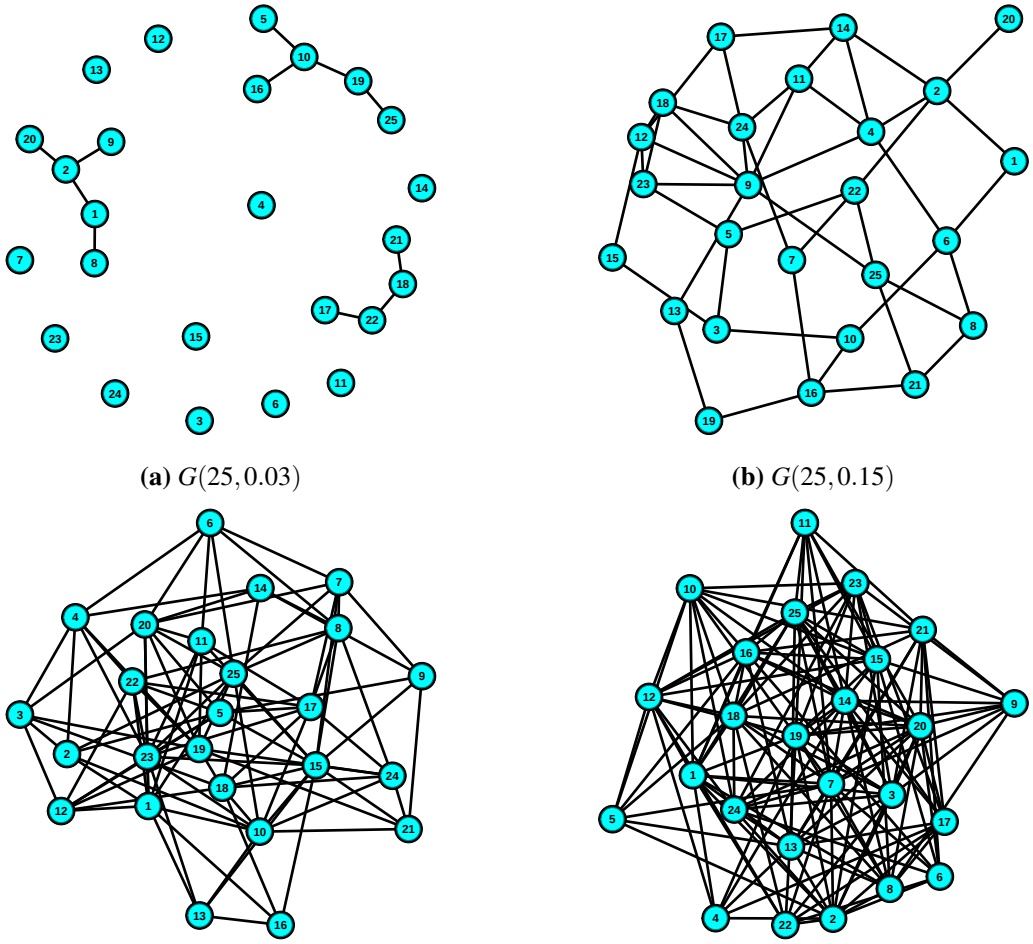
Normalmente, este modelo é denotado por $G(n,p)$. O grafo aleatório $G(n,p)$ tem número esperado de arestas

$$E_X(|E(G)|) = \binom{n}{2} \cdot p \tag{3.16}$$

e a distribuição de graus de qualquer vértice segue uma distribuição binomial da forma

$$P(deg(v_i) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \underset{np<\infty}{\overset{n\to\infty}{\longrightarrow}} \frac{(np)^k e^{-np}}{k!}, \tag{3.17}$$

isto é, segue a distribuição Poisson quando n é suficientemente grande e np é limitado. A Figura 2 apresenta alguns exemplos de grafos gerados através do modelo Erdős-Rényi $G(n,p)$ com diferentes valores para p.



(a) $G(25, 0.03)$ **(b)** $G(25, 0.15)$

(c) $G(25, 0.30)$ **(d)** $G(25, 0.45)$

Figura 2 – Exemplos de grafos aleatórios gerados pelo modelo Erdös-Rényi $G(n, p)$

3.1.3 MODELO BARABÁSI-ALBERT

Um outro modelo de grafos aleatórios muito recorrente na literatura é o modelo de Barabási e Albert (1999). Eles observaram empiricamente que muitas redes de colaboração científica e redes de internet como, por exemplo, hiperlinks em sites apresentavam distribuição de grau que se aproximavam de uma *lei de potência* (DURRETT, 2007). Dizemos que a distribuição de graus $p(k)$ segue distribuição de lei de potência com parâmetro β se

$$p(k) \sim k^{-\beta}. \tag{3.18}$$

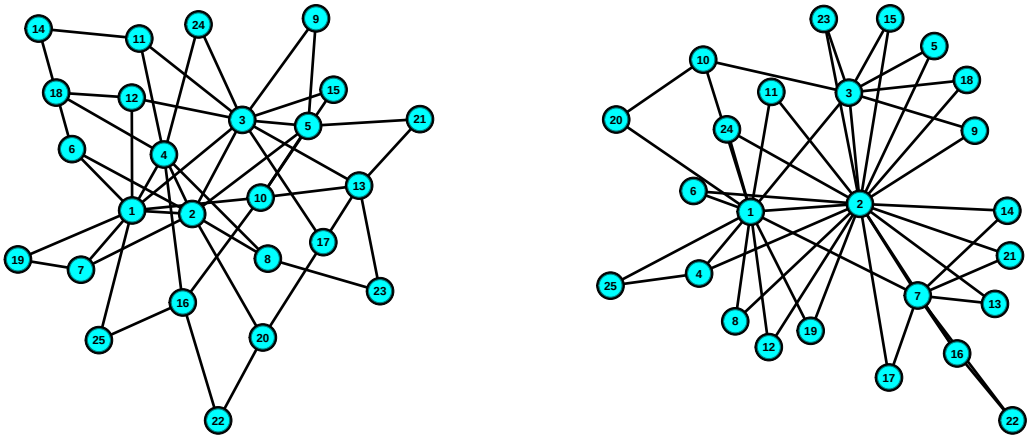
Desta forma, eles propuseram um modelo aleatório baseado em um algoritmo simples para produzir grafos que seguem aproximadamente a distribuição da lei de potência. Apesar da nomenclatura, trata-se mais de um algoritmo para gerar redes *livres de escala* do que propriamente de um "modelo". O grafo é gerado da seguinte maneira:

1. Passo 0: Inicia-se o algoritmo escolhendo um número pequeno m_0 de vértices conectados;

2. Passo t: A cada passo t, adiciona-se um novo vértice conectado a $m \leq m_0 + t - 1$ vértices já presentes na rede. A probabilidade do novo vértice se conectar ao vértice i já presente na rede é proporcional ao grau de v_i. Isto é,

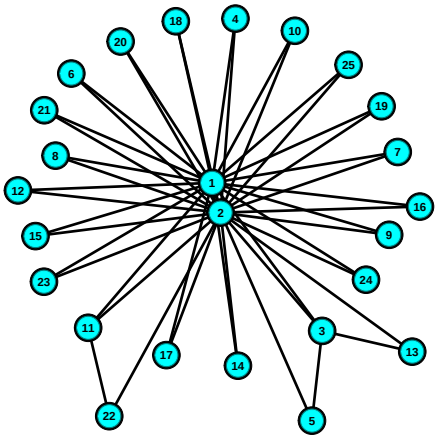
$$p_i = \frac{deg_{t-1}(v_i)}{\sum_j deg_{t-1}(v_k)}. \tag{3.19}$$

Ao fim do algoritmo, teremos $t + m_0$ vértices e $m \cdot t$ arestas na rede. Desta forma, Barabási e Albert (1999) provaram que a distribuição de graus resultante deste algoritmo é livre de escala, uma lei de potência com $\beta = 3$, ou seja, $p(k) \sim k^{-3}$. É interessante ressaltar que este modelo vem da necessidade de solucionar um problema real. Muitos sistemas naturais e artificiais possuem comportamento muito similar ao gerado por este modelo, tornando-o assim um dos modelos mais importantes na literatura sobre redes. A Figura 3 apresenta alguns exemplos de grafos gerados através do modelo Barabási-Albert com diferentes valores para β.

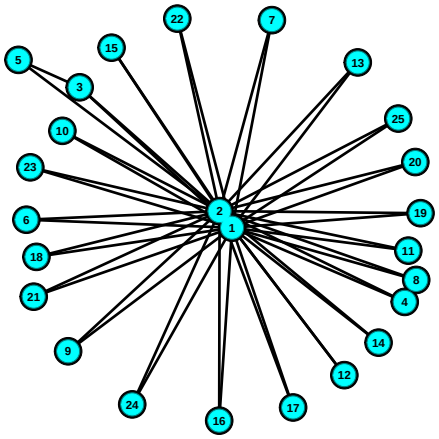


(a) $m_0 = 2, m = 25, \beta = 1$

(b) $m_0 = 2, m = 25, \beta = 2$



(c) $m_0 = 2, m = 25, \beta = 2.5$



(d) $m_0 = 2, m = 25, \beta = 3$

Figura 3 – Exemplos de grafos aleatórios gerados pelo modelo Barabási-Albert

4 MODELO DE REDES DE AFINIDADE

Em muitos modelos propostos na literatura em grafos, há especial interese sobre o comportamento das conexões, tomando como exemplo os modelos propostos por Erdős e Rényi (1959), Gilbert (1959) and Barabási e Albert (1999) mostrados na seção 3.1. Entretanto, muitas vezes estamos interessados em analisar os atributos dos vértices que de fato são responsáveis por realizar - ou não - uma conexão entre cada par de vértices. Neste sentido o trabalho desenvolvido por Singer (1995) introduziu o modelo de grafos aleatórios de interseção, cujo interesse é modelar grafos em que a fonte de incerteza está sobre o vértice. Nos trabalhos que se derivaram sobre os grafos aleatórios de interseção, houve especial interesse em encontrar propriedades topológicas sobre os grafos gerados sob o modelo, como por exemplo o trabalho de Karoński, Scheinerman e Singer-Cohen (1999) que buscavam estudar as probabilidades limiares para as quais um determinado subgrafo induzido estaria presente com alta probabilidade no grafo de interseção, ou mesmo o trabalho de Stark (2004) que se dedica ao estudo da distribuição de graus para o grafo aleatório de interseção.

Apesar de introduzir uma nova abordagem aos modelos de grafos aleatórios, os grafos aleatórios de interseção ainda deixam algumas lacunas para a modelagem de redes reais. Primeiramente, no modelo de grafos de interseção, todas a características que podem ser escolhidas pelo indivíduo são independentes e identicamente distribuídas, o que é um um pressuposto difícil de ser sustentado no mundo real. Além disso, os grafos aleatórios de interseção são baseados apenas no fato de um indivíduo compartilhar ou não pelo menos uma característica, o que é também difícil de sustentar, pois se assim fosse, tendo o conjunto de características {Cruzeiro, Atlético, Raposa, Galo, Futebol}, o mais radical dos atleticanos e o mais radical dos cruzeirenses estariam tão conectados ao escolher a palavra “Futebol” quanto dois cruzeirenses que simultaneamente escolhessem para si as palavras “Cruzeiro” e “Raposa”, para citar um exemplo.

Apresentamos neste trabalho uma proposta para aproximar os modelos probabilísticos de grafos e as redes reais baseadas em características dos vértices, em especial redes sociais. Ao utilizar como base o modelo proposto por Singer (1995), o modelo de redes de afinidade pode ser encarado como uma generalização do mesmo. Trazendo para o caso real específico que inspirou este trabalho - as representações sociais coletadas via TALP - o interesse principal é encontrar o núcleo central, isto é, as evocações que representariam os pensamentos mais importantes e difundidos no pensamento coletivo daquela sociedade sobre um determinado objeto social. Na literatura a respeito do tema, é altamente difundida a utilização dos quadrantes de Vergés (1992), valorando as palavras por frequência e ordem média de evocação. Daí, atribui-se

pontos de corte segundo algum critério, tomando o núcleo central como as palavras mais prioritárias em termo de ordem média e mais frequentes simultaneamente. No trabalho de Pereira (2017), entretanto, além da utilização dos quadrantes de Vergés, o autor conecta os indivíduos por uma rede cognitiva através de evocações em comum e as valora de acordo com sua ordem e frequência de evocação através de uma fórmula própria denominada por *coeficiente de afinidade cognitiva*. Desta forma, a conexão é uma consequência dos atributos dos vértices, que passam a ser a fonte de incerteza ao invés de termos incerteza sobre as conexões. Consequentemente, espera-se que modelos que investigam puramente os comportamentos das arestas sejam inadequados para modelar essa classe de problemas.

Toda estrutura desenvolvida para a construção deste modelo gira em torno de um conceito fundamental: a *função afinidade*. O conceito de função afinidade é um derivado do trabalho desenvolvido por Pereira (2017) e também usado em Guedes et al. (2018), que neste trabalho será generalizado. O coeficiente de afinidade proposto por Pereira (2017) foi aplicado para ponderar as conexões de uma rede cognitiva de indivíduos construída através do pensamento coletivo sobre as Enchentes do Rio Doce. No caso de Guedes et al. (2018), houve a utilização do mesmo coeficiente de afinidade, mas o estudo em questão era uma rede cognitiva de indivíduos construída através do pensamento coletivo sobre o Zika Virus. Em ambos os casos, os dados foram coletados via Técnica de livre associação de palavras (TALP), uma técnica derivada do instrumento desenvolvida por Jung (1910) cujo objetivo é capturar o pensamento leigo difundido na população através de palavras ou expressões (evocações) que cada indivíduo utiliza para dar significado a um termo indutor (tema). Em ambos os instrumentos de coleta utilizados em Pereira (2017) e Guedes et al. (2018), o indivíduo entrevistado foi exposto ao termo indutor e solicitado a, espontaneamente, dizer as 5 palavras que lhe ocorreram à exposição do tema. Após a fase de evocação, foi solicitado que o indivíduo ordenasse tais palavras por importância que o mesmo entendia que as evocações pronunciadas possuíam acerca de seu entendimento sobre o termo indutor.

A primeira generalização em relação aos trabalhos desenvolvidas por Pereira (2017) e Guedes et al. (2018) é que a função afinidade - que contemplará, inclusive, o coeficiente desenvolvido nesses trabalhos - não estará limitada à estrutura da TALP. No entanto, nada impede que façamos tal restrição, mas poderíamos estar interessados em uma função afinidade em que tal suposição não se faça necessária ou sequer fizesse sentido. A segunda generalização também diz respeito à estrutura da TALP utilizada nos trabalhos: o número de evocações pronunciadas pelas pessoas deixa de ter limites pré-estabelecidos, diferentemente dos trabalhos onde, em ambos os casos, o indivíduo estaria restrito a falar entre 1 e 5 palavras. A terceira e mais importante generalização é que o modelo não será limitado ao coeficiente de afinidade desenvolvido por Pereira (2017). Vamos propor uma vasta família de funções que poderão valorar o peso das conexões entre indivíduos mediante algum critério definido pelo pesquisador. Pretende-se também fazer algo diferente dos modelos que trabalham com ponderações, inserindo um limiar de corte, onde o pesquisador pode escolher a partir de que ponto considera-se que a força de

ligação é suficiente para criar uma conexão entre dois indivíduos e não simplesmente assumir que qualquer valor acima de zero representa uma conexão na rede.

4.1 FUNÇÃO AFINIDADE: DEFINIÇÃO E EXEMPLOS

Vamos começar definindo o conceito de função afinidade. Nós chamamos de função afinidade toda função que mensura, segundo algum critério, o quão semelhantes são as escolhas de dois indivíduos. Matematicamente, seja $D = \{d_1, d_2, \ldots, d_{m-1}, d_m\}$ com $m \in \mathbb{N}$ um conjunto de características arbitrárias conhecidas dos indivíduos. Suponha que em uma determinada população, cada indivíduo seja portador ou mesmo possa escolher um subconjunto dessas características. Em outras palavras, às características atribuídas ao i-ésimo indivíduo corresponde ao subconjunto $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,m_i-1}, d_{i,m_i}\}$, com $m_i \in \mathbb{N} \cup \{0\} : m_i \leq m$ tal que $D_i \subset D$ de características.

Se a única informação relevante sobre D_i consiste em conhecer quais características foram escolhidas pelo indivíduo i, podemos associar a cada conjunto D_i um vetor binário $U_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m-1}, u_{i,m}]$ de forma que

$$u_{i,j} = \begin{cases} 1, \text{ se } d_j \in D_i \\ 0, \text{ se } d_j \notin D_i \end{cases} \tag{4.1}$$

e U_i conserva a mesma informação trazida por D_i. Para uma população de tamanho N, nós definimos a *matriz de escolhas* $U_{N\times m}$ como

$$U_{N\times m} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_N \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,m-1} & u_{1,m} \\ u_{2,1} & u_{2,2} & u_{2,3} & \cdots & u_{2,m-1} & u_{2,m} \\ u_{3,1} & u_{3,2} & u_{3,3} & \cdots & u_{3,m-1} & u_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{N,1} & u_{N,2} & u_{N,3} & \cdots & u_{N,m-1} & u_{N,m} \end{bmatrix} \tag{4.2}$$

Vale destacar que a matriz U pode perfeitamente estocar informação adicional, evidentemente preservando a informação de cada D_i. Nos trabalhos de Pereira (2017) e Guedes et al. (2018), por exemplo, os autores estavam interessados na ordem em que essas características foram escolhidas. Naqueles casos, ao invés do conjunto de escolhas, teríamos $D_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,m_i}]$, - agora um vetor ao invés de um conjunto, já que a ordem importa. Para obter U, basta escrever o vetor de postos associado a D_i como $U_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m-1}, u_{i,m}]$ tal que

$$u_{i,j} = \begin{cases} k, \text{ se } d_j \in D_i,\ d_j = d_{i,k} \\ 0, \text{ se } d_j \notin D_i \end{cases} \tag{4.3}$$

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

Para exemplificar, vamos supor uma população hipotética de $N = 5$ indivíduos (uma família de atleticanos fanáticos) que escolhe de D = {Amor,Forte,Galo,Paixão,Vingador} palavras para significar o objeto social "Clube Atlético Mineiro". Então, um possível conjunto de escolhas poderia ser $D_1 = \{Galo, Forte, Vingador\}$ cujo vetor binário associado seria $U_1 = [0,1,1,0,1]$ e uma possível matriz U com $u_{i,j}$ segundo a Equação 4.1 seria

$$U_{5\times5} = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{4.4}$$

Mas, se por exemplo há um interesse em contar com a ordem das características for importante (em particular, na letra do hino do Atlético escrita por Vicente Motta, costa a seguinte frase "Clube Atlético Mineiro, Galo forte vingador"), Então, para o conjunto de escolhas $D_1 = \{Galo, Forte, Vingador\}$ cujo vetor de postos associado seria $U_1 = [0,2,1,0,3]$ e uma possível matriz U com $u_{i,j}$ segundo a Equação 4.3 seria

$$U_{5\times5} = \begin{bmatrix} 0 & 2 & 1 & 0 & 3 \\ 1 & 0 & 2 & 0 & 0 \\ 4 & 3 & 1 & 5 & 2 \\ 1 & 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 & 0 \end{bmatrix} \tag{4.5}$$

Formalmente, definimos a *função afinidade* como qualquer função simétrica f tal que

$$f : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+ \tag{4.6}$$

Apresentaremos nesta seção alguns exemplos simples de função afinidade.

4.1.1 FUNÇÃO AFINIDADE BINÁRIA

Seja $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,m_i-1}, d_{i,m_i}\}$ tal que $D_i \subset D$ o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor binário associado a D_i definido como descrito na Equação 4.1. Definimos a função afinidade binária como

$$f(U_i, U_k) = \begin{cases} 1, \text{ se } U_i \cdot U_k > 0 \\ 0, \text{ se } U_i \cdot U_k = 0 \end{cases} \tag{4.7}$$

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

$$f(U_i, U_k) = 1 \Leftrightarrow D_i \cap D_k \neq \emptyset \tag{4.8}$$

Podemos alocar as respectivas afinidades na matriz

$$A_{n\times n} = \begin{bmatrix} f(U_1,U_1) & f(U_1,U_2) & f(U_1,U_3) & \cdots & f(U_1,U_n) \\ f(U_2,U_1) & f(U_2,U_2) & f(U_2,U_3) & \cdots & f(U_2,U_n) \\ f(U_3,U_1) & f(U_3,U_2) & f(U_3,U_3) & \cdots & f(U_3,U_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(U_n,U_1) & f(U_n,U_2) & f(U_n,U_3) & \cdots & f(U_n,U_n) \end{bmatrix} \tag{4.9}$$

de forma que, considerando o exemplo construído em 4.1, obtemos

$$A_{5\times 5} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{4.10}$$

4.1.2 FUNÇÃO AFINIDADE CARDINAL

Seja $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,m_i-1}, d_{i,m_i}\}$ tal que $D_i \subset D$ o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor binário associado a D_i definido como descrito na Equação 4.1. Definimos a função afinidade cardinal como

$$f(U_i,U_k) = U_i \cdot U_k \tag{4.11}$$

Desta forma, a função afinidade cardinal mensura quantas características em comum foram escolhidas pelos indivíduos, isto é

$$f(U_i,U_k) = |D_i \cap D_k| \tag{4.12}$$

Considerando o exemplo construído em 4.1, nós teríamos

$$A_{5\times 5} = \begin{bmatrix} 3 & 1 & 3 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 \\ 3 & 2 & 5 & 2 & 2 \\ 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 2 & 1 & 2 \end{bmatrix} \tag{4.13}$$

4.1.3 COEFICIENTE DE CONCORDÂNCIA DE JACCARD

Em geral, coeficientes de concordância podem ser encarados com funções afinidades. Um bom exemplo é o coeficiente de Jaccard. Seja $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,m_i-1}, d_{i,m_i}\}$ tal que $D_i \subset D$

o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor binário associado a D_i definido como descrito na Equação 4.1. O coeficiente de concordância de Jaccard (1901) pode ser convertido na seguinte expressão

$$f(U_i, U_k) = \frac{U_i \cdot U_k}{\sum_{j=1}^{m} \max\{U_{i,j}, U_{k,j}\}} \tag{4.14}$$

Desta forma, o coeficiente de concordância de Jaccard mensura qual a proporção das características escolhidas pelos indivíduos que é compartilhada por eles, isto é

$$f(U_i, U_k) = \frac{|D_i \cap D_k|}{|D_i \cup D_k|} \tag{4.15}$$

Considerando o exemplo construído em 4.1, nós teríamos

$$A_{5\times5} = \begin{bmatrix} 1,00 & 0,25 & 0,60 & 0,00 & 0,25 \\ 0,25 & 1,00 & 0,40 & 0,33 & 0,00 \\ 0,60 & 0,40 & 1,00 & 0,40 & 0,40 \\ 0,00 & 0,33 & 0,40 & 1,00 & 0,33 \\ 0,25 & 0,00 & 0,40 & 0,33 & 1,00 \end{bmatrix} \tag{4.16}$$

4.1.4 COEFICIENTE DE AFINIDADE COGNITIVA

Em seu trabalho, Pereira (2017) apresenta um coeficiente de afinidade para calcular o nível de afinidade entre dois indivíduos através das palavras escolhidas pelos mesmos. O coeficiente de afinidade cognitiva foi desenvolvido com a finalidade de valorar a informação captada por questionários construídos sob a Técnica de associação livre de palavras (TALP). Assim sendo, alguns pressupostos foram assumidos:

1. A ordem de importância atribuída às palavras pelo indivíduo é importante;
2. A diferença entre as posições atribuídas às palavras pelo indivíduo é importante;
3. Existe um número M de palavras que devem ser evocadas por cada indivíduo. Se um indivíduo evocar menos que M, então o indivíduo deve ser penalizado por falta de informação.

Nestes termos, eis a definição para o coeficiente de afinidade cognitiva: seja $D_i = \{d_{i,1}, \ldots, d_{i,m_i}\}$ tal que $D_i \subset D$ o conjunto de características atribuídas ao i-ésimo indivíduo e seja U_i o vetor de postos associado a D_i definido como descrito na Equação 4.3. Seja $m_i \leq M$ o número de palavras evocadas pelo i-ésimo indivíduo e seja $m = \max(m_i, m_k)$. Então o coeficiente de afinidade cognitiva pode ser escrito como

$$f(U_i, U_k) = \frac{\sum_{j=1}^{M} (2 \cdot (m+1) - (u_{i,j} + u_{k,j})) \cdot (m - |u_{i,j} - u_{k,j}|) \cdot I(\min\{u_{i,j}, u_{k,j}\} > 0)}{\frac{m^2 \cdot (m+1) \cdot M \cdot (M+1)}{M \cdot (M+1) - (M-m) \cdot (M-m+1)}}, \tag{4.17}$$

No coeficiente de afinidade cognitiva, atribui-se pesos aos postos atribuídos às palavras. Estes é estes pesos são inversamente proporcionais aos postos, isto é, como o vetor de postos pode ser escrito como $[1,2,3,\ldots,m]$, seus respectivos pesos são $[m,m-1,m-2,\ldots,1]$. Desta forma, a primeira parcela do numerador é uma pontuação baseada na soma dos pesos dos postos atribuída às palavras, enquanto a segunda parcela é uma pontuação baseada na diferença entre os pesos dos postos atribuídos às palavras. A terceira parcela garante que a j-ésima parcela do somatório é positiva toda vez que os indivíduos simultaneamente escolhem a j-ésima palavra de D. O denominador do coeficiente tem duas finalidades: manter o coeficiente no intervalo $[0,1]$ e penalizar os indivíduos por falta de informação. Para mais detalhes, consultar Pereira (2017). Daí, considerando o exemplo da Equação 4.13 fixando $M = 5$, temos que

$$A_{5\times5} = \begin{bmatrix} 0,8000 & 0,2222 & 0,7068 & 0,0000 & 0,2667 \\ 0,2222 & 0,6000 & 0,3333 & 0,4000 & 0,0000 \\ 0,7068 & 0,3333 & 1,0000 & 0,1600 & 0,2267 \\ 0,0000 & 0,4000 & 0,1600 & 0,6000 & 0,1500 \\ 0,2667 & 0,0000 & 0,2267 & 0,1500 & 0,6000 \end{bmatrix} \tag{4.18}$$

4.2 GRAFO ALEATÓRIO GERADO PELO MODELO DE REDES DE AFINIDADE

Todo o conceito de função afinidade foi concebido em função de gerar grafos baseados em quão afins estariam as pessoas mediante as características a elas atribuídas. Assim, o grafo aleatório gerado por um modelo de redes de afinidade é tão importante ou mais do que a própria definição da função afinidade. Dado uma função afinidade f, podemos gerar um grafo de afinidade $G(n,m,\mu,f,\gamma) = G(\lambda)$, onde n é o número de vértices, m é o tamanho de D, μ é a distribuição de probabilidade sobre D, f é a função afinidade e γ é o ponto de corte, da seguinte forma:

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

(II) *Matriz de escolhas*: 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 alguma distribuição de probabilidade $\mu_{i,j}$ que está associada à informação sobre d_j (escolha ou não-escolha de d_j, ordem de escolha de d_j dado que d_j foi escolhido, etc.) para o conjunto de características atribuídas ao i-ésimo indivíduo. Assim, a matriz $U_{n\times4m}$ da

forma

$$U_{n\times m} = \begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} & \cdots & U_{1,m-1} & U_{1,m} \\ U_{2,1} & U_{2,2} & U_{2,3} & \cdots & U_{2,m-1} & U_{2,m} \\ U_{3,1} & U_{3,2} & U_{3,3} & \cdots & U_{3,m-1} & U_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ U_{n,1} & U_{n,2} & U_{n,3} & \cdots & U_{n,m-1} & U_{n,m} \end{bmatrix} \tag{4.19}$$

carrega toda a incerteza sobre $G(\lambda)$. É importante destacar que a função afinidade não é uma função aleatória, mas uma função determinística aplicada à variáveis aleatórias presentes na matriz de escolhas $U_{n\times m}$. Em geral, supomos que $U_i \perp U_k$, $\forall\, i \neq k$;

(III) *Conjunto de arestas*: o conjunto de arestas $E(G(\lambda))$ é induzido pelo nível de afinidade entre os indivíduos. Isto é,

$$E(G(\lambda)) := \{v_i \leftrightarrow v_k \Leftrightarrow f(U_i, U_k) \geq \gamma\}, \tag{4.20}$$

onde $\gamma > 0$ é um ponto de corte. A introdução deste ponto de corte contempla, caso desejado, maior rigor no que se refere a qual nível de afinidade seria necessário para estabelecer uma conexão entre dois indivíduos, de forma que

$$\gamma_1 > \gamma_2 \Rightarrow |E(G(n,m,\mu,f,\gamma_1))| \leq |E(G(n,m,\mu,f,\gamma_2))|. \tag{4.21}$$

Convencionamos $f(U_i,U_i) = 0$ para evitar *loops* no na rede gerada, já que em geral estamos interessados em analisar o nível de afinidade entre indivíduos diferentes. Desta forma, se no exemplo da subseção 4.1.4 considerarmos que a matriz da equação 4.18 foi gerado aleatoriamente sob uma função μ aplicada a D, o grafo gerado pela função de afinidade de Pereira com $\gamma = 0,4$ teria como matriz de adjacência

$$A(G(\lambda)) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.22}$$

De fato, considerando uma função distribuição arbitrária μ sobre D, um grafo $G(\lambda)$ satisfaz $0 \leq den(G(\lambda)) \leq 1$ se $0 < \gamma \leq \max\limits_{1\leq i<k\leq n} f(U_i,U_k)$ e $den(G(\lambda)) = 0$ se $\gamma > \max\limits_{1\leq i<k\leq n} f(U_i,U_k)$ para qualquer função afinidade f.

(IV) *Conjunto de pesos*: a imagem de f pode ser vista como pesos relativos à força da conexão entre dois indivíduos, de forma que

$$W(G(\lambda)) := \{f(U_i,U_k) | v_i \leftrightarrow v_k \in E(G(\lambda))\} \tag{4.23}$$

Desta forma, se no exemplo da subseção 4.1.4 considerarmos que a matriz da equação 4.18 foi gerado aleatoriamente sob uma função μ aplicada a D, o grafo gerado pela função de afinidade de Pereira com $\gamma = 0,4$ teria como matriz de adjacência ponderada

$$A_W(G(\lambda)) = \begin{bmatrix} 0,0000 & 0,0000 & 0,7068 & 0,0000 & 0,0000 \\ 0,0000 & 0,0000 & 0,0000 & 0,4000 & 0,0000 \\ 0,7068 & 0,0000 & 0,0000 & 0,0000 & 0,0000 \\ 0,0000 & 0,4000 & 0,0000 & 0,0000 & 0,0000 \\ 0,0000 & 0,0000 & 0,0000 & 0,0000 & 0,0000 \end{bmatrix} \tag{4.24}$$

De maneira prática, conhecer o tipo de informação armazenada em U, a distribuição de probabilidade conjunta μ sobre D, a função afinidade f e o ponto de corte γ são o suficiente para gerar um grafo aleatório de afinidade. A próxima seção será dedicada a apresentar distribuições μ com características diversas, bem como técnicas para gerar valores das mesmas.

4.2.1 ALGUNS EXEMPLOS DE DISTRIBUIÇÕES PARA A MATRIZ DE ESCOLHAS

Uma das condições para a geração do grafo aleatório via modelo de redes de afinidade é o conhecimento da função μ que rege o comportamento das escolhas dos indivíduos. A princípio, a função de distribuição conjunta μ é arbitrária, de forma que o vetor aleatório U_i associado a D_i é tal que

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = P(U_{i,1} = u_{i,1}, U_{i,2} = u_{i,2}, \ldots, U_{i,m} = u_{i,m}) \tag{4.25}$$

de forma que U_i não possui qualquer restrição, quer seja no quesito de independência das $U_{i,j} \sim \mu_j$ ou no quesito das $U_{i,j}$ serem identicamente distribuídas. Assim como tudo relacionado ao modelo de redes de afinidade, há muita liberdade no que concerne a distribuição μ. Não há sequer restrição sobre o fato de a população ter o conjunto D regido por apenas uma distribuição conjunta μ. Pelo contrário, é perfeitamente possível que μ seja uma mistura finita de distribuições μ_r, com $r = \{1,2,\ldots,r\}$. Além disso, μ_r e μ_s poderiam ter características muito distintas: μ_r poderia ser uma distribuição conjunta de variáveis independentes, enquanto μ_s poderia ser uma matriz de processo estocástico, por exemplo. Nesta seção iremos apresentar exemplos de funções de distribuição de probabilidade conjuntas que podem ser aplicadas sobre D.

4.2.1.1 $U_{i,j}$ binárias e independentes

O modelo probabilístico mais simples para $U_{i,j}$ seria aquele em que a única informação relevante sobre $U_{i,j}$ é sua escolha ou não-escolha e $U_{i,j}$ seja independente de $U_{i,h}$, $\forall j \neq h$. Para

apresentá-lo, vamos considerar o vetor aleatório $U_i = [U_{i,1}, U_{i,2}, \ldots, U_{i,m}]$ associado a D_i definido como descrito na Equação 4.1. Se com probabilidade δ_j o i-ésimo indivíduo escolhe $d_j \in D$, podemos escrever

$$U_{i,j}|U_{i,-j} \sim Ber(\delta_j) \Rightarrow P(U_{i,j} = u_{i,j}|U_{i,-j}) = \delta_j^{u_{i,j}} \cdot (1-\delta_j)^{1-u_{i,j}} \tag{4.26}$$

onde $U_{i,-j} = [u_{i,1}, \ldots, u_{i,j-1}, u_{i,j+1}, \ldots, u_{i,m}]$. Em particular, assumindo $U_{i,j} \perp U_{i,q}, \forall j \neq q$, obtemos

$$U_{i,j}|U_{i,-j} = U_{i,j} \sim Ber(\delta_j) \Rightarrow P(U_{i,j} = u_{i,j}) = \delta_j^{u_{i,j}} \cdot (1-\delta_j)^{1-u_{i,j}} \tag{4.27}$$

implicando na distribuição conjunta expressa por

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = \prod_{j=1}^{m} P(U_{i,j} = u_{i,j}) = \prod_{j=1}^{m} \delta_j^{u_{i,j}} \cdot (1-\delta_j)^{1-u_{i,j}} \tag{4.28}$$

Um caso particular importante deste modelo pode ser derivado de imediato assumindo que as variáveis $U_{i,j}$'s são identicamente distribuídas para todo $j \in \{1,2,\ldots,m\}$. Assim, $\delta_j = \delta$, de forma que podemos reescrever a Equação 4.28 da forma

$$U_i \sim \mu \Rightarrow P(U_i = u_i) = \prod_{j=1}^{m} P(U_{i,j} = u_{i,j}) = \delta^{\sum_{j=1}^{m} u_{i,j}} \cdot (1-\delta)^{m-\sum_{j=1}^{m} u_{i,j}} \tag{4.29}$$

4.2.1.2 Um exemplo de μ para $U_{i,j}$ binárias e dependentes

Vamos agora apresentar um exemplo de função de distribuição de probabilidade conjunta onde a escolha das características não são independentes. Este modelo será baseado em um processo utilizando um vetor de probabilidades inicial e uma matriz de transição de probabilidades que será atualizada a cada passo. Vamos definir $q \in \{1,\ldots,m\}$ um conjunto de índices para enumerar qual é a etapa atual do processo. Considere $D = \{$Faculdade, Park, Estatística, Doutorado, Inglês, Luara, Idioma$\}$, bem como o conjunto de características atribuídas ao i-ésimo indivíduo D_i e o vetor U_i associado a D_i. Temos então um conjunto D de tamanho $m = 7$. O primeiro passo é saber quantas características o indivíduo escolheu. A título de exemplo, considere que a distribuição de probabilidade de m_i denotada por $P(m_i = x)$ é descrita pelo vetor

$$P(m_i = x) = \begin{array}{c} \begin{matrix} m_i{=}0 & m_i{=}1 & m_i{=}2 & m_i{=}3 & m_i{=}4 & m_i{=}5 & m_i{=}6 & m_i{=}7 \end{matrix} \\ \left[\begin{matrix} 0,1355 & 0,2710 & 0,2710 & 0,1807 & 0,0903 & 0,0363 & 0,0120 & 0,0034 \end{matrix} \right] \end{array} \tag{4.30}$$

Deve-se então sortear aleatoriamente m_i considerando-se tais probabilidades. Suponha que $m_i = 3$. A partir daí, precisamos sortear a primeira palavra do vetor. Neste exemplo, considere o vetor de probabilidades inicial

$$S(U_{i,j_1} = 1) = \begin{array}{c} \begin{matrix} j_1{=}1 & j_1{=}2 & j_1{=}3 & j_1{=}4 & j_1{=}5 & j_1{=}6 & j_1{=}7 \end{matrix} \\ \left[\begin{matrix} 0,17500 & 0,09998 & 0,10334 & 0,36543 & 0,01604 & 0,23451 & 0,00570 \end{matrix} \right] \end{array} \tag{4.31}$$

Considere a matriz de transição de probabilidades condicionais também a matriz de transições iniciais $T_0(U_{i,j_2} = 1|U_{i,j_1})$

$$T_0(U_{i,j_2}=1|U_{i,j_1}) = \begin{array}{c} \\ j_1{=}1 \\ j_1{=}2 \\ j_1{=}3 \\ j_1{=}4 \\ j_1{=}5 \\ j_1{=}6 \\ j_1{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} j_2{=}1 & j_2{=}2 & j_2{=}3 & j_2{=}4 & j_2{=}5 & j_2{=}6 & j_2{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,000 & 0,091 & 0,045 & 0,230 & 0,003 & 0,631 & 0,000 \\ 0,263 & 0,000 & 0,309 & 0,000 & 0,004 & 0,156 & 0,268 \\ 0,002 & 0,341 & 0,000 & 0,000 & 0,053 & 0,528 & 0,076 \\ 0,139 & 0,781 & 0,000 & 0,000 & 0,000 & 0,080 & 0,000 \\ 0,075 & 0,159 & 0,000 & 0,011 & 0,000 & 0,754 & 0,000 \\ 0,221 & 0,102 & 0,000 & 0,001 & 0,450 & 0,000 & 0,225 \\ 0,002 & 0,768 & 0,000 & 0,002 & 0,228 & 0,000 & 0,000 \end{array}\right] \end{array} \tag{4.32}$$

Para começar o processo, basta então sortear aleatoriamente um valor para j_1 considerando as probabilidades de $S(U_{i,j_1} = 1)$. Probabilisticamente, temos que para $b \in \{1,\ldots,m\}$ e $s(U_{i,b} = 1))$ sendo o elemento de S da posição b,

$$P(U_{i,j_1} = 1) = \prod_{b=1}^{m} s(U_{i,b} = 1)^{I(b=j_1)} \tag{4.33}$$

de forma que U_{i,j_1} segue a distribuição multinomial com parâmetros $\mathscr{M}(1, s(U_{i,b} = 1))$. Vamos supor que sorteamos $j_1 = 4$, de forma que $u_{i,4} = 1$. Logo a matriz de transição deve ser atualizada. Neste modelo, removemos a probabilidade de transição para $u_{i,2}$ e normalizamos a probabilidade por linha, isto é,

$$t_b(U_{i,j_b} = 1|u_{i,j_{b-1}}) = \frac{I(j_b \notin \{j_1,\ldots,j_{b-1}\}) \cdot t_{b-1}(U_{i,j_b} = 1|u_{i,j_{b-1}})}{\sum_{q=1}^{m} I(q \notin \{j_1,\ldots,j_{b-1}\}) \cdot t_{b-1}(U_{i,q} = 1|u_{i,j_{b-1}})} \tag{4.34}$$

Desta forma, temos que $T_1(u_{i,j_1} = 1) = T_0(U_{i,j_2} = 1|u_{i,j_1})$ escrita como

$$T_1(u_{i,j_1}=1) = \begin{array}{c} \\ j_1{=}1 \\ j_1{=}2 \\ j_1{=}3 \\ j_1{=}4 \\ j_1{=}5 \\ j_1{=}6 \\ j_1{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} j_2{=}1 & j_2{=}2 & j_2{=}3 & j_2{=}4 & j_2{=}5 & j_2{=}6 & j_2{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,000 & 0,118 & 0,059 & 0,000 & 0,004 & 0,819 & 0,000 \\ 0,263 & 0,000 & 0,309 & 0,000 & 0,004 & 0,156 & 0,268 \\ 0,002 & 0,341 & 0,000 & 0,000 & 0,053 & 0,528 & 0,076 \\ \mathbf{0,139} & \mathbf{0,781} & \mathbf{0,000} & \mathbf{0,000} & \mathbf{0,000} & \mathbf{0,080} & \mathbf{0,000} \\ 0,076 & 0,161 & 0,000 & 0,000 & 0,000 & 0,763 & 0,000 \\ 0,221 & 0,103 & 0,000 & 0,000 & 0,451 & 0,000 & 0,225 \\ 0,002 & 0,769 & 0,000 & 0,000 & 0,228 & 0,000 & 0,000 \end{array}\right] \end{array} \tag{4.35}$$

Prosseguindo, basta então sortear aleatoriamente um valor para j_2 considerando as probabilidades da linha $j_1 = 4$ em $T_1(u_{i,j_1} = 1)$. Probabilisticamente, temos que para $t_1(U_{i,b} = 1|u_{i,j_1})$ sendo o elemento de T_1 da linha j_1 e coluna b

$$P(U_{i,j_2} = 1|u_{i,j_1}) = \prod_{b \neq j_1} t_1(U_{i,b} = 1|u_{i,j_1})^{I(b=j_2)} \tag{4.36}$$

de forma que U_{i,j_2} condicionado a $u_{i,4}$ segue a distribuição $\mathscr{M}(1, t_1(U_{i,b} = 1|u_{i,4}))$. Vamos supor que sorteamos $j_2 = 2$, de forma que $u_{i,2} = 1$. Logo a matriz de transição deve ser atualizada.

Desta forma, temos que $T_2(u_{i,2}=1) = T_0(U_{i,j_3}=1|u_{i,j_1},u_{i,j_2})$ escrita como

$$T_2(u_{i,j_2}=1) = \begin{array}{c} \\ j_2{=}1 \\ j_2{=}2 \\ j_2{=}3 \\ j_2{=}4 \\ j_2{=}5 \\ j_2{=}6 \\ j_2{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} j_3{=}1 & j_3{=}2 & j_3{=}3 & j_3{=}4 & j_3{=}5 & j_3{=}6 & j_3{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,000 & 0,000 & 0,067 & 0,000 & 0,004 & 0,929 & 0,000 \\ \mathbf{0,263} & \mathbf{0,000} & \mathbf{0,309} & \mathbf{0,000} & \mathbf{0,004} & \mathbf{0,156} & \mathbf{0,268} \\ 0,003 & 0,000 & 0,000 & 0,000 & 0,080 & 0,802 & 0,115 \\ 0,635 & 0,000 & 0,000 & 0,000 & 0,000 & 0,365 & 0,000 \\ 0,091 & 0,000 & 0,000 & 0,000 & 0,000 & 0,909 & 0,000 \\ 0,246 & 0,000 & 0,000 & 0,000 & 0,503 & 0,000 & 0,251 \\ 0,009 & 0,000 & 0,002 & 0,000 & 0,990 & 0,000 & 0,000 \end{array}\right] \end{array} \tag{4.37}$$

Por fim, basta então sortear aleatoriamente um valor para j_3 considerando as probabilidades da linha $j_2 = 2$ em $T_2(u_{i,2}=1)$. Probabilisticamente, temos que

$$P(U_{i,j_3}=1|u_{i,j_1},u_{i,j_2}) = \prod_{b\notin\{j_1,j_2\}} t_2(U_{i,b}=1|u_{i,2})^{I(b=j_3)} \tag{4.38}$$

de forma que U_{i,j_3} condicionado a $u_{i,4},u_{i,2}$ segue a distribuição $\mathscr{M}(1,t_2(U_{i,b}=1|u_{i,2}))$. Vamos supor que sorteamos $j_3=7$, de forma que $u_{i,7}=1$. Assim, obtemos o vetor $u_i=[0,1,0,1,0,0,1]$ equivalente a {Park, Doutorado, Idioma}. De modo geral,

$$P(U_{i,j_1}=1,\ldots,U_{i,j_b}=1) = P(m_i=b)\cdot P(U_{i,j_1}=1)\cdot\cdots\cdot P(U_{i,j_b}=1|u_{i,j_1},\ldots,u_{i,j_{b-1}}) \tag{4.39}$$

Agora, podemos definir a função μ de forma que para $x_b \in \{0,1\}$, temos

$$U_i \sim \mu \Rightarrow P(U_{i,j_1}=x_1,\ldots,U_{i,j_m}=x_m) \tag{4.40}$$

Vale destacar que podemos facilmente utilizar essa função para gerar dados ordenados, bastando usar o número do passo em que a palavra foi escolhida como seu respectivo posto. Este exemplo resultaria em $u_i=[0,2,0,1,0,0,3]$ e consequentemente em $D_i=$ {Doutorado, Park, Idioma}. Na próxima subseção, vamos apresentar uma maneira de criar dados ordenados quando as escolhas são independentes.

4.2.1.3 Um exemplo de μ para U_i com postos

Vamos apresentar agora um exemplo simples de uma função de distribuição composta μ para gerar dados ordenados. Assim como o modelo anterior, este modelo será baseado em uma matriz de transição de probabilidades que será atualizada em cada etapa do processo. Vamos definir $q \in \{1,\ldots,m\}$ um conjunto de índices para enumerar qual é a etapa atual do processo. Considere $D=$ {Faculdade, Park, Estatística, Doutorado, Inglês, Luara, Idioma}, bem como o conjunto de características atribuídas ao i-ésimo indivíduo D_i e U_i^* o vetor de escolhas binárias associadas a D_i. Vamos definir $R(u^*_{i,j_q})$ a variável aleatória para o posto de u^*_{i,j_q}. Logo, queremos encontrar $U_i = R(U_i^*)$. Vamos supor à priori que, em uma determinada população ,

a probabilidade de que j-ésima palavra de u_i^* assuma o posto x em u_i, isto é, $P(R(u_{i,j}^*) = x)$ é descrita pela matriz

$$T(R(u_{i,j}^*) = x) = \begin{array}{c} \\ j{=}1 \\ j{=}2 \\ j{=}3 \\ j{=}4 \\ j{=}5 \\ j{=}6 \\ j{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} x{=}1 & x{=}2 & x{=}3 & x{=}4 & x{=}5 & x{=}6 & x{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,010 & 0,025 & 0,005 & 0,006 & 0,015 & 0,017 & 0,021 \\ 0,023 & 0,009 & 0,000 & 0,015 & 0,000 & 0,001 & 0,157 \\ 0,015 & 0,000 & 0,034 & 0,005 & 0,000 & 0,053 & 0,000 \\ 0,065 & 0,000 & 0,000 & 0,037 & 0,000 & 0,082 & 0,000 \\ 0,054 & 0,004 & 0,015 & 0,011 & 0,027 & 0,018 & 0,000 \\ 0,000 & 0,000 & 0,036 & 0,012 & 0,058 & 0,012 & 0,000 \\ 0,082 & 0,003 & 0,022 & 0,030 & 0,001 & 0,001 & 0,017 \end{array}\right] \end{array} \tag{4.41}$$

de forma que a soma dos elementos de $T(R(u_{i,j}^*) = x)$ é igual a 1. Supondo μ^* binária arbitrária, o resultado de amostras de $U_i^* \sim \mu^*$ será um vetor binário com m_i observações 1 e $m - m_i$ observações 0. Um bom exemplo, para $m = 7$ e $m_i = 3$, é o vetor $u_i^* = [0,1,0,0,1,1,0]$. Daí, tiramos que apenas os postos $h_i = \{0,1,2,3\}$ podem ser escolhidos pois, de acordo com a Equação 4.3 que será seguida neste modelo, os postos das características não escolhidas recebem 0, de forma que tomando $j_1 = 1$, $j_2 = 3$, $j_3 = 4$, $j_4 = 7$ temos que para $x_q \in h_i$ observamos $x_1 = x_2 = x_3 = x_4 = 0$ resultando em $r(u_{i,j_1}^*) = r(u_{i,j_2}^*) = r(u_{i,j_3}^*) = r(u_{i,j_4}^*) = 0$. Dadas as escolhas feitas em u_i^*, a matriz $T(R(u_{i,j}^*) = x)$ deve ser atualizada. Neste modelo de exemplo, vamos apenas remover as probabilidades que não podem ser observadas dado a informação de u_i^* e normalizar os elementos da matriz resultante pela soma de seus elementos, isto é,

$$t_b(R(u_{i,j_b}^*) = x_b) = \frac{I(j_b \notin \{j_1,\ldots,j_{b-1}\}) \cdot I(x_b \in h_i) \cdot [1 - I(x_b > 0) \cdot I(x_b \notin \{x_1,\ldots,x_{b-1}\})] \cdot t(R(u_{i,j_b}^*) = x_b)}{\sum_{q=1}^{m} \sum_{y=1}^{m} I(q \notin \{j_1,\ldots,j_{b-1}\}) \cdot I(y \in h_i) \cdot [1 - I(y > 0) \cdot I(y \notin \{x_1,\ldots,x_{b-1}\})] \cdot t(R(u_{i,q}^*) = y)} \tag{4.42}$$

Definindo $T_5(R(u_{i,j_5}^*) = x_5) = T(R(u_{i,j}^*) = x | r(u_{i,j_1}^*), \ldots, r(u_{i,j_4}^*))$, obtemos a matriz

$$T_5(R(u_{i,j_5}^*) = x_5) = \begin{array}{c} \\ j_5{=}1 \\ j_5{=}2 \\ j_5{=}3 \\ j_5{=}4 \\ j_5{=}5 \\ j_5{=}6 \\ j_5{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} x_5{=}1 & x_5{=}2 & x_5{=}3 & x_5{=}4 & x_5{=}5 & x_5{=}6 & x_5{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ \mathbf{0,161} & \mathbf{0,061} & \mathbf{0,000} & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ \mathbf{0,388} & \mathbf{0,026} & \mathbf{0,107} & 0,000 & 0,000 & 0,000 & 0,000 \\ \mathbf{0,000} & \mathbf{0,000} & \mathbf{0,258} & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \end{array}\right] \end{array} \tag{4.43}$$

de onde podemos escrever uma forma reduzida como

$$T_5^{red}(R(u_{i,j_5}^*) = x_5) = \begin{array}{c} \\ j_5{=}2 \\ j_5{=}5 \\ j_5{=}6 \end{array} \begin{array}{c} \begin{array}{ccc} x_5{=}1 & x_5{=}2 & x_5{=}3 \end{array} \\ \left[\begin{array}{ccc} 0,161 & 0,061 & 0,000 \\ 0,388 & 0,026 & 0,107 \\ 0,000 & 0,000 & 0,258 \end{array}\right] \end{array} \tag{4.44}$$

e a forma vetorial reduzida

$$V_5^{red}(R(u_{i,j_5}^*) = x_5) = \begin{array}{c} \begin{array}{ccccccccc} x_5{=}1,j_5{=}2 & x_5{=}2,j_5{=}2 & x_5{=}3,j_5{=}2 & x_5{=}1,j_5{=}5 & x_5{=}2,j_5{=}5 & x_5{=}3,j_5{=}5 & x_5{=}1,j_5{=}6 & x_5{=}2,j_5{=}6 & x_5{=}3,j_5{=}6 \end{array} \\ \left[\begin{array}{ccccccccc} 0,161 & 0,061 & 0,000 & 0,388 & 0,026 & 0,107 & 0,000 & 0,000 & 0,258 \end{array}\right] \end{array} \tag{4.45}$$

Basta agora sortear o par ordenado (x_5, j_5) considerando as probabilidades contidas na matriz $T_5(R(u^*_{i,j}) = x)$. Em termos probabilísticos, temos para $b \in \{1,\ldots,m\}$, $y \in \{0,\ldots,m_i\}$ e $t_5(R(u^*_{i,b}) = y)$ sendo o elemento de T_5 da linha b e coluna y

$$P(R(u^*_{i,j_5}) = x_5 | r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_4})) = \prod_{b\notin\{j_1,\ldots,j_4\}} \prod_{y\notin\{x_1,\ldots,x_4\}} t_5(R(u^*_{i,b}) = y)^{I(y=x_5, b=j_5)} \quad (4.46)$$

de forma que $R(u^*_{i,j})$ condicionado a $r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_5})$ segue a distribuição multinomial com parâmetros $\mathscr{M}(1, V_5^{red}(R(u^*_{i,j_5}) = x_5))$. Prosseguindo, suponha que sorteamos $(x_5 = 2, j_5 = 5)$ observando consequentemente $r(u^*_{i,5}) = 2$. Uma nova atualização deve ser realizada. Faça $P(R(u^*_{i,j_6}) = x) = P(R(u^*_{i,j}) = x | r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_5}))$. Obtemos a matriz

$$T_6(R(u^*_{i,j_6}) = x_6) = \begin{array}{c} \\ j_6{=}1 \\ j_6{=}2 \\ j_6{=}3 \\ j_6{=}4 \\ j_6{=}5 \\ j_6{=}6 \\ j_6{=}7 \end{array} \begin{array}{c} \begin{array}{ccccccc} x_6{=}1 & x_6{=}2 & x_6{=}3 & x_6{=}4 & x_6{=}5 & x_6{=}6 & x_6{=}7 \end{array} \\ \left[\begin{array}{ccccccc} 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ \mathbf{0,384} & 0,000 & \mathbf{0,000} & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \\ \mathbf{0,000} & 0,000 & \mathbf{0,616} & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 & 0,000 \end{array}\right] \end{array} \quad (4.47)$$

de onde podemos escrever uma forma reduzida como

$$T_6^{red}(R(u^*_{i,j_6}) = x_6) = \begin{array}{c} \\ j_6{=}2 \\ j_6{=}6 \end{array} \begin{array}{c} \begin{array}{cc} x_6{=}1 & x_6{=}3 \end{array} \\ \left[\begin{array}{cc} 0,384 & 0,000 \\ 0,000 & 0,616 \end{array}\right] \end{array} \quad (4.48)$$

e a forma vetorial reduzida

$$V_6^{red}(R(u^*_{i,j_6}) = x_6) = \begin{array}{c} \begin{array}{cccc} x_6{=}1,j_6{=}2 & x_6{=}3,j_6{=}2 & x_6{=}1,j_6{=}6 & x_6{=}3,j_6{=}6 \end{array} \\ \left[\begin{array}{cccc} 0,384 & 0,000 & 0,000 & 0,616 \end{array}\right] \end{array} \quad (4.49)$$

Em termos probabilísticos, temos

$$P(R(u^*_{i,j_6}) = x | r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_5})) = \prod_{b\notin\{j_1,\ldots,j_5\}} \prod_{y\notin\{x_1,\ldots,x_5\}} t_6(R(u^*_{i,b}) = y)^{I(y=x_6, b=j_6)} \quad (4.50)$$

De forma que $R(u^*_{i,j})$ condicionado a $r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_5})$ segue $\mathscr{M}(1, V_6^{red}(R(u^*_{i,j_6}) = x_6))$. Suponha agora que sorteamos $(x_6 = 3, j_6 = 6)$ observando consequentemente $r(u^*_{i,6}) = 3$. Poderíamos re-atualizar a matriz, mas existe apenas uma opção restante: sortear $(x_7 = 2, j_7 = 1)$ observando consequentemente $r(u^*_{i,2}) = 1$. Assim, obtemos o vetor $u_i = [0,1,0,0,2,3,0]$ equivalente a [Park, Inglês, Luara]. De modo geral,

$$P(R(u^*_{i,j_b}) = x_b | r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_{b-1}})) = P(R(u^*_{i,j_b}) = x_b)\cdot I(x_b \notin \{r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_{b-1}})\})\cdot I(x_b > 0) + I(u^*_{i,j_b} = 0)\cdot I(x_b = 0) \quad (4.51)$$

de forma que

$$P(U_{i,j_1} = x_1,\ldots,U_{i,j_m} = x_m) = P(R(u^*_{i,j_1}) = x_1)\cdot\cdots\cdot P(R(u^*_{i,j_m}) = x_m | r(u^*_{i,j_1}),\ldots,r(u^*_{i,j_{m-1}})). \quad (4.52)$$

Agora, podemos definir a função μ de forma que

$$U_i \sim \mu \Rightarrow P(U_{i,j_1} = x_1, \ldots, U_{i,j_m} = x_m). \tag{4.53}$$

Em comparação à variante do modelo apresentado na subseção anterior a qual pode ser utilizada para gerar $U_{i,j}$ com postos, a diferença está em como escolher as características. Neste modelo estamos escolhendo as características segundo um critério arbitrário e então atribuindo postos às mesmas. Este critério arbitrário de escolha inclui os modelos onde $U_{i,j}$ são independentes. O modelo apresentado na subseção anterior escolhe as características considerando necessariamente que existe uma estrutura de dependência entre as características.

4.2.2 GERANDO MATRIZES DE ESCOLHAS U

Nesta seção vamos apresentar alguns resultados que mostrarão como é possível gerar matrizes U fixados alguns parâmetros. Vamos nos concentrar nos modelos de μ apresentados na subseções 4.2.1.1 já que os modelos das subseções 4.2.1.2 e 4.2.1.3 foram exemplificados em suas próprias subseções. Começando pela estrutura de U mais simples, isto é, a estrutura de U onde $U_{i,j}$ são independentes e identicamente distribuídas, nós temos que, considerando um grafo aleatório $G(n,m,\mu,f,\gamma)$ e que com probabilidade δ o i-ésimo indivíduo escolhe a j-ésima palavra, temos que a probabilidade de conexão entre dois indivíduos $p = P(f(U_i,U_k) > 0)$ é definida por

$$p = 1 - (1-\delta^2)^m, \tag{4.54}$$

que vem a ser a probabilidade complementar de não haver nenhuma palavra em comum entre os conjunto de características atribuídas aos indivíduos em questão. Daí, podemos definir a probabilidade de escolha para cada palavra de D em função de m e de p isolando δ na equação 4.54. De fato,

$$\delta = \sqrt{1 - \sqrt[m]{1-p}}. \tag{4.55}$$

O resultado acima pode ser estendido para a estrutura de U como definida na subseção 4.2.1.1 com algumas alterações. Considere que com probabilidade δ_j o i-ésimo indivíduo escolhe a j-ésima palavra, temos que

$$p = 1 - \prod_{j=1}^{m}(1-\delta_j^2). \tag{4.56}$$

Daí, temos

$$\prod_{j=1}^{m}(1-\delta_j^2) = 1 - p \Rightarrow \sum_{j=1}^{m}\log(1-\delta_j^2) = \log(1-p). \tag{4.57}$$

Agora, vamos considerar o vetor de pesos z definido no $(m-1)$-simplex, isto é, $0 < z_j < 1 \; \forall \; j = \{1,2,\ldots,m\}$ e $\sum_{j=1}^{m} z_j = 1$. Daí, podemos escrever

$$\sum_{j=1}^{m}\log(1-\delta_j^2) = \log(1-p)\cdot\sum_{j=1}^{m} z_j = \sum_{j=1}^{m} z_j \cdot \log(1-p), \tag{4.58}$$

de tal maneira que

$$\log(1-\delta_j^2) = z_j \cdot \log(1-p) \Rightarrow 1-\delta_j^2 = \exp\left\{z_j \cdot \log(1-p)\right\} \tag{4.59}$$

e, finalmente,

$$\delta_j = \sqrt{1-\exp\left\{z_j \cdot \log(1-p)\right\}}. \tag{4.60}$$

Em outras palavras, podemos definir a probabilidade δ_j de escolher a j-ésima palavra é proporcional a p e z_j. Assim, podemos comparar também distribuições onde a probabilidade de escolha das características no grafo de afinidade não é a mesma para todas as caracterísitcas, mas que mesmo assim tenham m e p fixados. Vale destacar que δ fixo implica em $z_j = \frac{1}{m}$, $\forall j = \{1,2,\ldots,m\}$.

4.3 MODELOS DE GRAFOS ALEATÓRIOS DE INTERSEÇÃO

Como discutido no preâmbulo deste capítulo, os primeiros modelos aleatórios introduzidos para grafos geralmente possuíam a fonte de incerteza concentrada nas arestas. Em contraste, Singer (1995) introduziu um modelo ao qual nomeou por *Random Intersection Graph* (numa tradução literal, *grafo aleatório de interseção*) (RIG). Nesta seção vamos apresentar este modelo e mostrar que a grafo de afinidade é uma generalização deste modelo.

O modelo RIG, em contraste com os modelos propostos por Erdős e Rényi (1959), Gilbert (1959) and Barabási e Albert (1999) concentra a fonte de incerteza sobre os vértices. Mais precisamente, a fonte de incerteza se concentra sobre os atributos dos vértices. De acordo com a definição em Karoński, Scheinerman e Singer-Cohen (1999), seja n e m inteiros positivos e $\delta \in [0,1]$. Para cada inteiro positivo i com $1 \leq i \leq n$, seja S_i um subconjunto aleatório de $S = 1,2,\ldots,m$ formado selecionando m_i elementos de forma independente e identicamente distribuída com probabilidade δ. Então, há uma conexão entre os vértices i e k se $S_i \cap S_k \neq \emptyset$. Fica claro de imediato que o modelo RIG como introduzido por Singer (1995) é um caso particular da função afinidade, mais precisamente considerando a combinação de $U_{i,j}$ independentes e identicamente distribuídas com a função afinidade f definida na subseção 4.1.1 e ponto de corte que satisfaça $0 \leq \gamma \leq 1$.

Há diversos trabalhos publicados na literatura que se dedicaram a estudar os grafos aleatórios de interseção desde sua introdução em Singer (1995). Estes trabalhos compreendem desde estudos de suas características topológicas, transição de fase, evolução do aparecimento dos mais diversos subgrafos a medida que p cresce, etc. O resultado mais interessante para este trabalho é apresentado por Fill, Scheinerman e Singer-Cohen (2000). Tal estudo demonstra que ao relacionar a ordem de $G(n,m,\delta)$ com o tamanho do número de características m, de forma que $m = g(n) = \lfloor n^\alpha \rfloor$ e ao considerar $\hat{p}$ como a probabilidade assintótica de uma aresta

em particular ocorrer em $G(n,m,\delta)$, então é necessário $\alpha > 6$ para que $G(n,m,\delta)$ e $G(n,\hat{p})$ possuam os mesmos limiares para todas as propriedades, isto é, é necessário que $\alpha > 6$ para que um grafo gerado via RIG se comporte como um grafo gerado via modelo Erdős-Rényi. Tal resultado significa que o número de características à disposição de dois indivíduos deveria ser extremamente maior para que a conexão entre os indivíduos tivessem o mesmo comportamento de uma conexão gerada de forma independente daquelas características. Uma vez que, de acordo com este estudo, uma rede de ordem 10 já implicaria na necessidade de mais de 1 milhão de características em um grafo RIG à disposição para gerar um grafo ER, devido a limitações do tamanho de D em casos reais, seria altamente improvável redes de afinidades ao serem utilizados com finalidade de, por exemplo, representações sociais se comportarem como grafos ER.

5 ESTUDO SIMULADO NO MODELO DE REDES DE AFINIDADE

Um dos objetivos deste trabalho é analisar como os grafos gerados pelos modelos de rede de afinidade se comportam. Dado o conhecimento de f, μ e γ, podemos traçar o perfil do comportamento de qualquer modelo de redes de afinidade. No entanto, existem algumas limitações sobre a forma de obter estes resultados. Como dito anteriormente, os modelos de redes de afinidade compõem uma vasta família de modelos. Há inúmeras combinações possíveis de f, μ e U disponíveis para se construir um modelo de redes de afinidade. Podemos, por exemplo, escolher a distribuição μ com diversas estruturas (independentes, dependentes, identicamente distribuídas ou não), com vários tipos de resultados alocados em U (binários, ordinais, etc.); podemos escolher o vetor de pesos z balanceado ou desbalanceado; podemos particionar U, de forma que exista uma distribuição μ distinta para cada partição; podemos escolher inúmeras funções afinidade, bem como podemos eleger inúmeros pontos de corte. Neste sentido, há enorme flexibilidade nos modelo. Infelizmente, a dificuldade de obter resultados analíticos se torna tão grande quanto ou maior que o nível de flexibilidade do modelo. Desta forma, precisamos recorrer a simulações para reproduzir o comportamento de um modelo de redes de afinidade.

Neste capítulo, vamos construir uma estrutura de simulação para observar medidas topológicas empiricamente através do método de Monte Carlo para um modelo de redes de afinidade. Escolhemos para este estudo μ independente em vários níveis de desbalanceamento, inclusive μ identicamente distribuída. Escolhemos a função afinidade cardinal para este estudo devido à simplicidade da função. Já que imagem da função cardinal é o número de características compartilhadas pelos indivíduos, isto simplifica a escolha dos cenários a serem avaliados, pois sua imagem é discreta. Daí, a configuração de um grafo de interseção apenas se altera em pontos de corte inteiros. Logo, podemos simplesmente observar o comportamento de cada medida topológica para qualquer conjunto discreto dos valores de γ entre 1 até o maior valor observado para a afinidade em toda a amostra de Monte Carlo inclusive, isto é, $\gamma \in \{1, 2, \ldots, \hat{\gamma}_{max}\}$, onde $\hat{\gamma}_{max} = \max\{f(U_i, U_k)\}$, ao invés de avaliar e escolher valores de $\gamma \in (0, \hat{\gamma}_{max}]$ como seria necessário se a imagem de f fosse contínua. Além de comparar as medidas topológicas observadas entre os cenários da função afinidade, ainda vamos comparar o comportamento da função afinidade cardinal com o comportamento de grafos gerados com aresta independentemente distribuídas cuja distribuição de probabilidade conserve a distribuição dos pesos das arestas do grafo $G(\lambda)$.

5.1 METODOLOGIA

Como discutido anteriormente, para gerar grafos aleatórios através do modelo de redes de afinidade de ordem n com m características, é suficiente conhecer a distribuição μ sobre D, o tipo de informação armazenada em U, a função afinidade f e o ponto de corte γ. Embora estejamos nos concentrando em um único modelo de afinidade, nos propusemos a investigar o comportamento do modelo de afinidade cardinal para diversos pontos do espaço paramétrico. Começando por μ, vamos nos restringir a $U_{i,j}$ independentes e $U_{i,j}$ independentes e identicamente distribuídas. É simples gerar $U_{i,j}$ com $U_{i,j}$ i.i.d. uma vez que, fixada uma probabilidade de conexão p e m o tamanho do conjunto D, basta aplicar a Equação 4.55 para obter o valor de δ tal que a probabilidade de conexão seja p. Já no caso em que $U_{i,j}$ é independente, encontrar o vetor de probabilidades de escolha $\Delta = \{\delta_1, \ldots, \delta_m\}$ associado a D requer, além de fixar uma probabilidade de conexão p, que o vetor de pesos z seja gerado. O verdadeiro desafio para a simulação proposta é gerar z garantindo que na amostra de Monte Carlo existam cenários em que todos as probabilidades δ_j sejam bastantes próximas como existam cenários onde um determinado grupo de características tenha probabilidade de escolha muito superiores às demais características. Além disso, é importante que exista um método simples e replicável para gerar vetores Δ. Vamos apresentar agora o método sugerido pelo autor deste trabalho, método este que necessita de duas entradas: o número de características m e o parâmetro de desbalanceamento da distribuição de probabilidades de escolha $\theta \geq 0$. O parâmetro de desbancamento foi idealizado com a intenção de controlar o quanto os elementos de Δ são diferentes, de forma que $\theta = 0$ representaria o desbalanceamento nulo, isto é, $U_{i,j}$ são i.i.d. e, à medida em que θ se afaste de 0, pretendemos colocar mais probabilidade de escolha em um conjunto de características que nas demais características. Para tal, foi desenvolvida uma estratégia para gerar os vetores de z dados m e θ. Seguindo a definição de z na subseção 4.2.2, seja Z um vetor aleatório definido no $(m-1)$-simplex, isto é, $\sum_{j=1}^{m} Z_i = 1$ com $0 \leq Z_i \leq 1 \ \forall \ j$. Podemos então gerar amostras de Z definindo

$$Z \sim Dirichlet(v), \tag{5.1}$$

onde v é um vetor de parâmetros de comprimento m com $v_j > 0 \ \forall \ j$, onde escolher $v_j > v_h$ aumenta a probabilidade de $Z_j > Z_h$. Basta agora gerar o vetor aleatório V de parâmetros a serem utilizados para gerar Z. Vamos definir $V \sim \eta(g(\theta))$, isto é, queremos gerar V de uma função densidade η cujo parâmetro é uma função que depende de θ. Podemos escolher por conveniência uma função que se adeque aos nossos propósitos. Neste sentido, é ideal que a variância de η seja diretamente proporcional a θ, já que elevar θ consequentemente elevaria sua variância. Daí, esperaríamos uma variância maior em V, acarretando em Z desbalanceado. Além disso, é interessante que η não seja superiormente limitada, afim de não limitar o impacto de θ sobre η. Uma das distribuições que atendem este requisito é a distribuição gamma, já que

para $V_j \sim gamma(\alpha, \beta)$ temos

$$f(v, \alpha, \beta) = \frac{v^{\alpha-1} \cdot \exp\{-\frac{v}{\beta}\}}{\beta^{\alpha} \cdot \Gamma(\alpha)},\ v > 0,\ \alpha > 0,\ \beta > 0 \tag{5.2}$$

e

$$E_{V_j}(V_j) = \alpha \cdot \beta \text{ e } Var_{V_j}(V_j) = \alpha \cdot \beta^2. \tag{5.3}$$

Para um maior controle deste processo, propomos $g(\theta)$ de forma que $E_V(v) = f(k)$, isto é, a esperança de V é uma função de k, onde k é uma constante arbitrária, e $f(k)$ não depende de θ. Definindo $(\alpha, \beta) = g(\theta) = (\frac{1}{\theta}, \frac{\theta}{k})$, isto é,

$$V_j \sim gamma(\frac{1}{\theta}, \frac{\theta}{k}), \tag{5.4}$$

onde

$$E_{V_j}(V_j) = \frac{1}{k} \text{ e } Var_{V_j}(V_j) = \frac{\theta}{k^2}. \tag{5.5}$$

Para evitar problemas computacionais, é necessária uma pequena alteração: fazemos $V_j \sim \eta + c$ com $c > 0$ constante. Utilizamos este artifício para evitar que V_j se aproxime demasiadamente de 0, o que pode fazer com os softwares arredondem o valor para 0, que não é um parâmetro válido para a distribuição Dirichlet. O algoritmo a seguir resume os passos necessários para gerar vetores Δ de comprimento m para um nível de desbalanceamento θ utilizando o método sugerido.

- Passo 0: Verificar se $\theta = 0$. Se afirmativo, então faça o vetor de pesos z ser tal que $z_j = \frac{1}{m}\ \forall\ j$ e então ir diretamente para o Passo 4. Caso contrário, ir para o Passo 1.
- Passo 1: Gerar t amostras do vetor $\eta + c$. Vamos denotar a k-ésima amostra de $\eta + c$ por $\eta^{[k]}$. A constante c garante que, por maior que seja $Var_X(\eta_j)$, η_j será inferiormente limitado por c.
- Passo 2: Para cada amostra $\eta^{[k]}$, gerar e ordenar

$$Z^{[k]} \sim Dirichlet(\eta^{[k]}). \tag{5.6}$$

- Passo 3: Calcular o vetor de pesos ordenados z como

$$z_{(j)} = \frac{1}{t} \cdot \sum_{k=1}^{t} Z_{(j)}^{[k]},\ \forall\ j = \{1, \ldots, m\},\ k = \{1, \ldots, t\}. \tag{5.7}$$

- Passo 4: Calcular o vetor de probabilidades de escolha Δ como

$$\delta_{(j)} = \sqrt{1 - \exp\{z_{(j)} \cdot \log(1-p)\}}. \tag{5.8}$$

Para avaliar as diferenças entre as distribuições de grau, vamos utilizar a estatística D de Kolmogorov-Smirnov. Considerando X e Y os vetores contendo os graus dos n vértices dos grafo G e os graus dos s vértices do grafo H. Então,

$$F_X(k) = \frac{1}{n} \cdot \sum_{i=1}^{n} I(X_i < k) \tag{5.9}$$

é a função de probabilidade acumulada empírica de X. A estatística D de Kolmogorov-Smirnov (SMIRNOV; SMIRNOV, 1939) é definida como

$$D_{X,Y} = \max_k |F_X(k) - F_Y(k)| \tag{5.10}$$

Em outras palavras, a estatística D de Kolmogorov-Smirnov é um método não-paramétrico que avalia o quanto duas distribuição são diferentes baseado na distância máxima alcançada pelas suas respectivas distribuições acumuladas nos pontos de observação k.

Finalmente, objetivamos comparar o do modelo de redes de afinidade cardinal com o respectivo modelo de redes de arestas independentes, isto é, queremos observar as diferenças entre o grafo gerado via afinidade cardinal com um grafo cujas arestas, embora geradas independentemente, conservem a distribuição de peso da função afinidade cardinal. Para tal, precisamos calcular a distribuição de probabilidade da afinidade cardinal para dois indivíduos. Vamos definir $\Delta = \{\delta_1, \delta_2, \delta_3, \ldots, \delta_m\}$ onde $\Delta \in [0,1]^m$ 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 $U_j \sim \mu_j = Ber(\delta_j)$ e $U_j \perp U_l$. Então, o conjunto das partes de Δ denotado por $\mathscr{P}(\Delta)$ possui 2^m subconjuntos de Δ com tamanhos que variam entre 0 e m. Definindo $\mathscr{P}^{(r)}(\Delta) \subset \mathscr{P}(\Delta)$ com $r = \{0,1,2,\ldots,m\}$ o subconjunto de todos os elementos de $\mathscr{P}(\Delta)$ de tamanho r. Daí, $\mathscr{P}^{(r)}(\Delta)$ possui $\binom{m}{r}$ subconjuntos de Δ. Defina agora $\mathscr{P}_s^{(r)}(\Delta) \subset \mathscr{P}^{(r)}(\Delta)$ com $s = \{0,1,2,\ldots,\binom{m}{r}\}$ o s-ésimo subconjunto de $\mathscr{P}^{(r)}(\Delta)$. A fim de simplificar a notação, vamos escrever $q(r,s) = \mathscr{P}_s^{(r)}(\Delta)$. Então podemos escrever

$$P(f(U_i,U_k) = x) = \sum_{s=1}^{\binom{m}{x}} \Big(\prod_{\delta_j \in q(x,s)} \delta_j^2 \prod_{\delta_j \notin q(x,s)} (1-\delta_j^2) \Big),\ x = \{0,1,\ldots,m\} \tag{5.11}$$

De forma que $f(U_i,U_k)$ tem distribuição Poisson binomial com parâmetros Δ. Em particular, se $U_j \sim \mu_j = Ber(\delta),\ \forall j$, então

$$P(f(U_i,U_k) = x) = \binom{m}{x} \cdot (\delta^2)^x \cdot (1-\delta^2)^{m-x},\ x = \{0,1,\ldots,m\} \tag{5.12}$$

Consequentemente, $f(U_i,U_k)$ tem distribuição binomial com parâmetro δ. Em sua definição, a distribuição binomial é um caso particular da distribuição Poisson binomial. Assim, a média e variância de $f(U_i,U_k)$ com f afinidade cardinal são tais que

$$E(f(U_i,U_k)) = \sum_{j=1}^{m} \delta_j^2 \text{ e } Var(f(U_i,U_k)) = \sum_{j=1}^{m} \delta_j^2 \cdot (1-\delta_j^2) \tag{5.13}$$

O estudo simulado planejado nesta seção foi realizado através de simulações de Monte Carlo utilizando o software *R*. Os gráficos foram gerados dinamicamente no software *Microsoft Excel*

(os arquivos com os gráficos podem ser encontrados no endereço <https://drive.google.com/drive/folders/1h1LI067XQo4pEs6yW60eAFd_G310q6B_>). Nesta série de comparações, nos focamos em observar determinadas medidas topológicas dos grafos para entender qual o comportamento do modelo de afinidade cardinal para cada um dos cenários escolhidos. Foram analisadas as seguintes medidas topológicas: distribuição quantílica dos graus, grau máximo, força máxima, transitividade, coeficiente de clustering, proximidade, número de componentes, ordem da maior componente e média do tamanho das componentes excluindo-se a maior componente. Para realizar as comparações acima, definiu-se os seguintes parâmetros para estudo:

- Ordem do grafo: $n = 256$;
- Número de características: $m \in \{20, 80, 320\}$;
- Probabilidade de conexão: $p \in \left\{\frac{1}{40}, \frac{2}{40}, \ldots \frac{38}{40}, \frac{39}{40}\right\}$;
- Parâmetro de desbalanceamento: $\theta \in \{0, 4, 16\}$;
- Número de réplicas: $M = 200$;
- Ponto de corte: $\gamma \in \{1, 2, \ldots, 6\}$;
- Constantes: $c = 1 \cdot 10^{-12}$ e $k = \frac{1}{100}$.

5.2 RESULTADOS

Através da metodologia detalhada na seção anterior, obtivemos os resultados apresentados nesta seção. Como dito anteriormente, é difícil encontrar resultados analíticos dado a complexidade do modelo de redes de afinidade. A natureza dependente das arestas do grafo torna difícil inclusive explicitar a função de verossimilhança do modelo. Nos resultados a seguir, vamos apresentar e comparar os perfis comportamentais da função afinidade cardinal e o respectivo modelo de arestas independentes para várias combinações de parâmetros.

5.2.1 DISTRIBUIÇÃO DE PROBABILIDADE

Fixados os parâmetros θ, m e p, foram gerados vetores Δ através do processo mencionado na Seção 5.1. Considerando a distribuição de probabilidade do modelo de afinidade cardinal, foram calculadas as médias e as variâncias de $f(U_i, U_k)$ baseado nos Δ gerados. A Tabela 1 apresenta a esperança do número de conexões por cenário estudado. A Tabela 2, por outro lado, apresenta as variâncias dos mesmos.

Como esperado, a esperança de $f(U_i, U_k)$ no modelo de afinidade cardinal aumenta à medida que p cresce. O mesmo vale para a variância. Um detalhe interessante a ser observado

Tabela 1 – Valores esperados para distribuição de probabilidade do modelo de afinidade cardinal

p	$\theta = 0$			$\theta = 4$			$\theta = 16$		
	m =20	m = 80	m =320	m =20	m = 80	m =320	m =20	m = 80	m =320
0.1	0.1051	0.1053	0.1053	0.1009	0.1038	0.1049	0.0961	0.1009	0.1038
0.2	0.2219	0.2228	0.2231	0.2111	0.2191	0.2220	0.1957	0.2111	0.2189
0.3	0.3535	0.3559	0.3565	0.3322	0.3488	0.3544	0.3043	0.3359	0.3485
0.4	0.5044	0.5092	0.5104	0.4674	0.4980	0.5067	0.4260	0.4741	0.4995
0.5	0.6813	0.6902	0.6924	0.6270	0.6716	0.6869	0.5559	0.6309	0.6743
0.6	0.8956	0.9111	0.9150	0.8079	0.8781	0.9063	0.7148	0.8227	0.8847
0.7	1.1685	1.1950	1.2017	1.0354	1.1480	1.1888	0.8653	1.0508	1.1556
0.8	1.5464	1.5934	1.6054	1.3305	1.5198	1.5842	1.0519	1.3726	1.5346
0.9	2.1750	2.2698	2.2943	1.7990	2.1340	2.2528	1.4453	1.8694	2.1529

Tabela 2 – Variância para distribuição de probabilidade do modelo de afinidade cardinal

p	$\theta = 0$			$\theta = 4$			$\theta = 16$		
	m =20	m = 80	m =320	m =20	m = 80	m =320	m =20	m = 80	m =320
0.1	0.1045	0.1052	0.1053	0.0988	0.1031	0.1047	0.0920	0.0992	0.1033
0.2	0.2194	0.2222	0.2229	0.2028	0.2163	0.2212	0.1781	0.2040	0.2166
0.3	0.3473	0.3543	0.3561	0.3106	0.3420	0.3525	0.2621	0.3185	0.3427
0.4	0.4916	0.5060	0.5096	0.4241	0.4842	0.5029	0.3498	0.4412	0.4882
0.5	0.6581	0.6842	0.6909	0.5565	0.6477	0.6798	0.4284	0.5690	0.6539
0.6	0.8555	0.9007	0.9124	0.6893	0.8350	0.8938	0.5158	0.7213	0.8488
0.7	1.1002	1.1771	1.1972	0.8441	1.0762	1.1683	0.5681	0.8847	1.0943
0.8	1.4268	1.5616	1.5973	1.0261	1.3975	1.5476	0.6052	1.1029	1.4325
0.9	1.9385	2.2054	2.2779	1.2547	1.8925	2.1786	0.7471	1.3877	1.9469

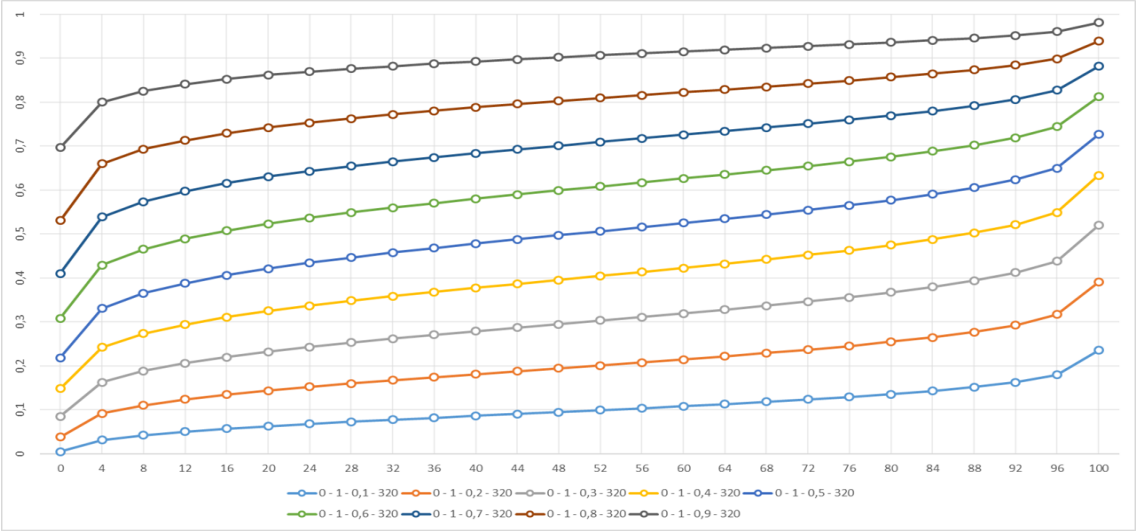
é que, mesmo com 320 características à disposição e $p = 0,9$, $E(f(U_i,U_k)) < 2,3$ nos cenários selecionados, isto é, nesses modelos que se supõem que as escolhas são independentes umas das outras, não são esperadas muitas características compartilhadas entre dois indivíduos. É interessante observar os papéis que θ e m exercem sobre a esperança e a variância. De acordo com a Tabela 1, quanto mais probabilidade concentrarmos em um grupo específico de características, menor é o número esperado de características compartilhadas por dois indivíduos os indivíduos. Por outro lado, aumentar o número de características eleva a esperança do número de características compartilhadas por um indivíduo. Assim, no que se refere à esperança de $f(U_i,U_k)$, θ e m parecem exercer efeitos opostos. Observando a Tabela 2, podemos perceber que, assim como a esperança, elevar o valor de θ implica na redução da variância de $f(U_i,U_k)$, enquanto elevar m implica no aumento da variância de $f(U_i,U_k)$. Baseado nesta informações, vamos ter maior atenção nas próximas análises com os seguintes pares de parâmetros: $(\theta,m) = \{(0,320),(4,80),(16,20)\}$, uma vez que os pares $(0,320)$ e $(16,20)$ representam casos extremos, enquanto o par $(4,80)$ representa o caso médio em termos de esperança e variância de $f(U_i,U_k)$ neste estudo simulado.

5.2.2 DISTRIBUIÇÃO QUANTÍLICA DOS GRAUS

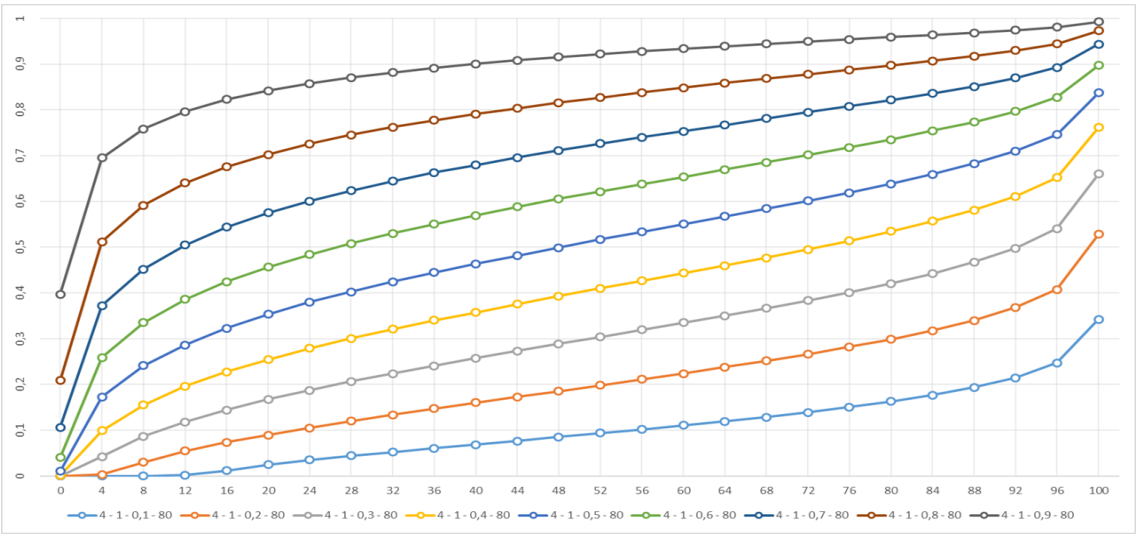
Nesta subseção, vamos apresentar o comportamento quantilico da distribuição de graus para o modelo de afinidade cardinal. A Figura 4 apresenta o perfil de tal distribuição fixados os parâmetros m, θ, p e γ. As curvas apresentadas na Figura 4 indicam que existe um padrão para a distribuição de graus que independe do p escolhido. A curva pode ser descrita como tendo duas fases de crescimento distintas. A medida em que θ cresce e m decresce, a taxa de crescimento da fase de crescimento mais rápido parece aumentar. Podemos perceber a princípio que a taxa de crescimento nos percentis iniciais da distribuição de graus é incrementado quando θ cresce e m decresce. Um outro detalhe muito importante é que o aumento de θ implica em nós de grau 0. Para $p = 0.1$, pelo menos 44% dos nós apresentou grau 0 para o cenário $m = 20$ e $\theta = 16$. Em contraponto, como p está fixado, há a tendência a haver graus mais altos nas curvas deste cenário. Podemos dizer que, ao concentrar muita probabilidade em um grupo específico de características, escolher características daquele grupo implica em maior probabilidade de realizar conexões e obter grau rápido. Por outro lado, não escolher características deste grupo diminui drasticamente a probabilidade do vértice formar conexões.

A Figura 5 apresenta o comportamento da estatística D de Kolmogorov-Smirnov resultante da comparação entre as distribuições de grau da função afinidade cardinal para os conjuntos de parâmetros selecionados. Em todos os valores de p retratados na Figura 5, podemos observar que, como esperado, os modelos $(0,320)$ e $(16,20)$ são os mais afastados uns dos outros. Podemos perceber uma maior aproximação entre as curvas $(0,320)$ e $(4,80)$ do que entre as curvas $(16,20)$ e $(4,80)$. Ao comparar as distâncias destes três cenários com outros cenários que compartilham ou o mesmo m ou o mesmo θ, é possível perceber que para o cenário $(4,80)$, que não está entre casos extremos, elevar ou diminuir θ e m faz com que os efeitos opostos que exercem sobre a distribuição dos graus cancelem um ao outro, fazendo com que a distância de $(0,20)$ e $(16,320)$ sejam menores até mesmo do que modelos que compartilham o mesmo m ou o mesmo θ, o que não é observado para os cenário $(0,320)$ e $(16,20)$, cujos modelos que compartilham um dos parâmetros tem distribuição de grau mais próximas do que os modelos que não compartilham pelo menos um dos parâmetros.

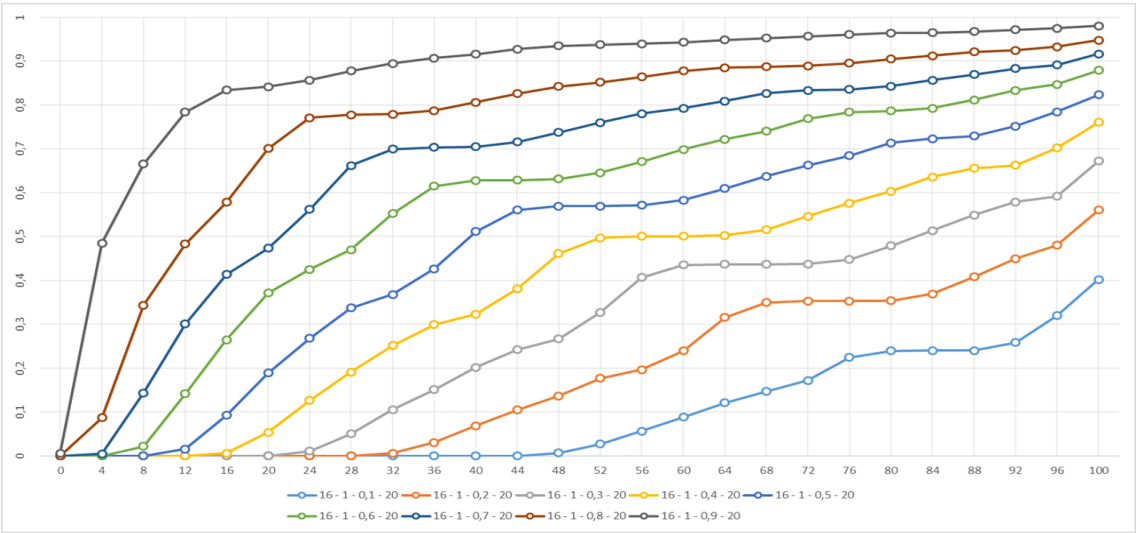
A Figura 6 apresenta o perfil de tal distribuição quantílica dos graus fixados os parâmetros m, θ, conforme variamos os valores de γ. Para uma melhor visualização gráfica, selecionou-se $p = 0,9$. Podemos observar pela Figura 6 alguns padrões de comportamento do grafo para θ, m e p a medida em que γ cresce. Em termos de distribuição de graus, o cenário $(0,320)$ se mostrou o cenário que mais conservou as características da distribuição de graus para $\gamma > 1$. Considerando-se que, para $p = 0,9$ temos $E(f(U_i,U_k)) = 2,2779$ características compartilhadas entre dois indivíduos no cenário $(0,320)$, então a perda de arestas cujo $f(U_i,U_k) = 1$ torna-se menos drástica em sua distribuição de graus quando $\gamma = 2$ do que para $(16,20)$, cuja $E(f(U_i,U_k)) = 1,4453$. É possível perceber que para $\gamma = 4$, ainda não é esperado encontrar vértices de grau 0 no cenário $(0,320)$, enquanto já é esperado encontrar tais vértices para $\gamma = 3$



(a) Cenários: $\theta = 0$, $\gamma = 1$, $p \in \{0.1, \ldots, 0.9\}$ e $m = 320$

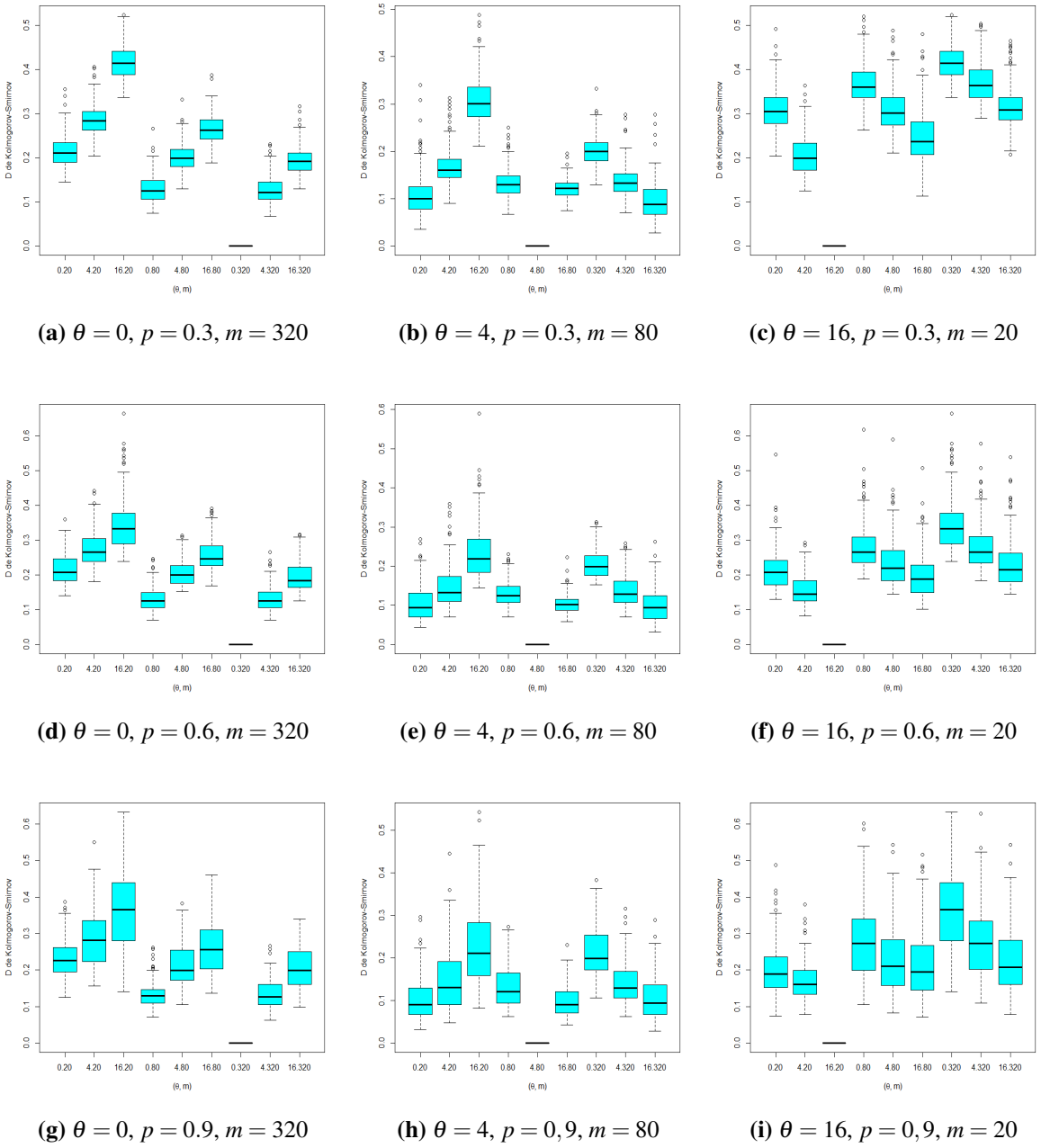


(b) Cenários: $\theta = 4$, $\gamma = 1$, $p \in \{0.1, \ldots, 0.9\}$ e $m = 80$



(c) Cenários: $\theta = 16$, $\gamma = 1$, $p \in \{0.1, \ldots, 0.9\}$ e $m = 20$

Figura 4 – Perfil da distribuição quantílica dos graus para o modelo de afinidade cardinal



(a) $\theta = 0$, $p = 0.3$, $m = 320$ **(b)** $\theta = 4$, $p = 0.3$, $m = 80$ **(c)** $\theta = 16$, $p = 0.3$, $m = 20$

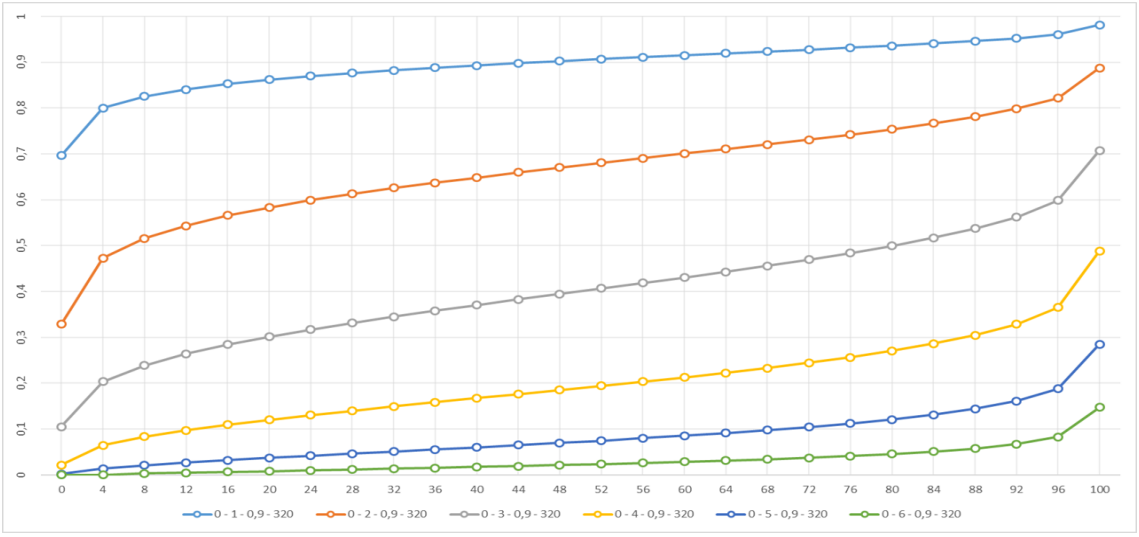
(d) $\theta = 0$, $p = 0.6$, $m = 320$ **(e)** $\theta = 4$, $p = 0.6$, $m = 80$ **(f)** $\theta = 16$, $p = 0.6$, $m = 20$

(g) $\theta = 0$, $p = 0.9$, $m = 320$ **(h)** $\theta = 4$, $p = 0,9$, $m = 80$ **(i)** $\theta = 16$, $p = 0,9$, $m = 20$

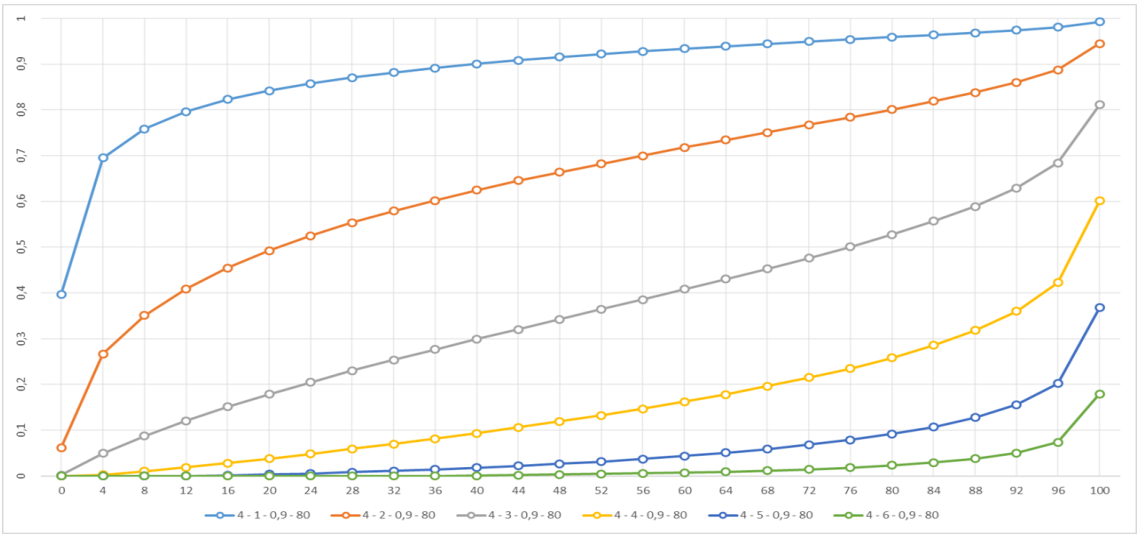
Figura 5 – Estatística D de Kolmogorov-Smirnov para distância entre as distribuições de graus dos modelos de afinidade cardinal

e $\gamma = 2$ em $(4,80)$ e $(16,20)$ respectivamente. Um outro aspecto interessante é que assim como em $\gamma = 1$, a medida em que γ cresce, os graus mais elevados de $(4,80)$ continuam sendo superiores aos graus mais elevados de $(0,320)$, o que não acontece para $(16,20)$. Assim, existem indícios de que quanto mais elevado é o valor de θ, maior é o impacto que o grafo de afinidade sofre à medida que γ cresce.

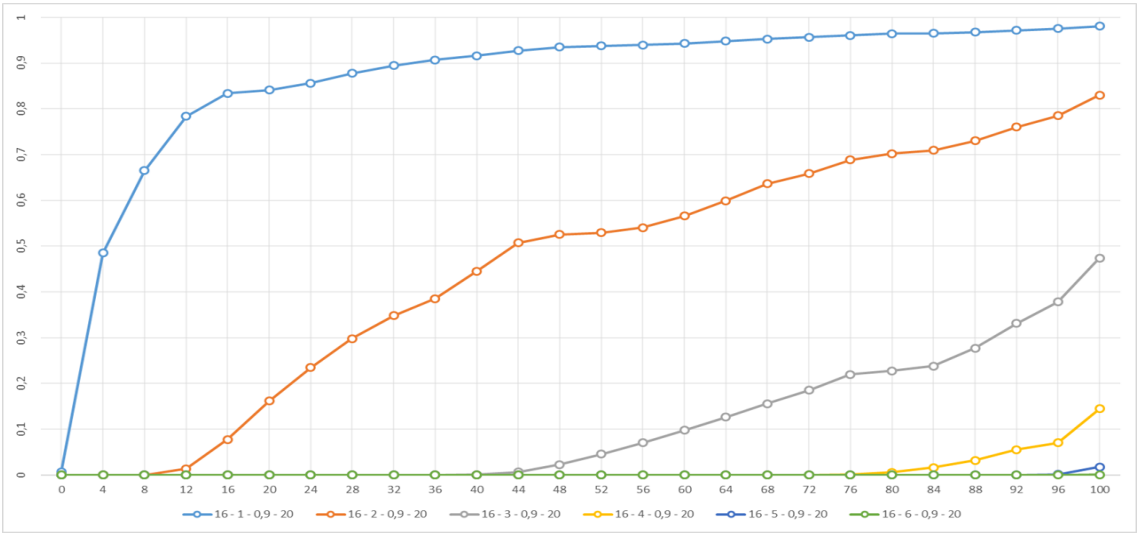
A Figura 7 apresenta o perfil do comportamento da estatística D de Kolmogorov-Smirnov resultante da comparação entre as distribuições de grau da função afinidade à medida em que θ cresce. A distância foi medida sempre entre $G(n,m,\mu,f,\gamma)$ e $G(n,m,\mu,f,\gamma+1)$, isto é,



(a) Cenários: $\theta = 0$, $\gamma \in \{1,\ldots,6\}$, $p = 0{,}9$ e $m = 320$



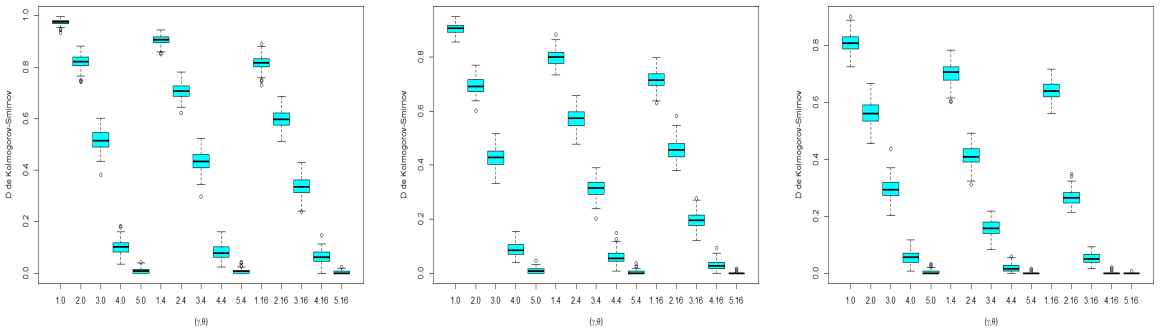
(b) Cenários: $\theta = 4$, $\gamma \in \{1,\ldots,6\}$, $p = 0{,}9$ e $m = 80$



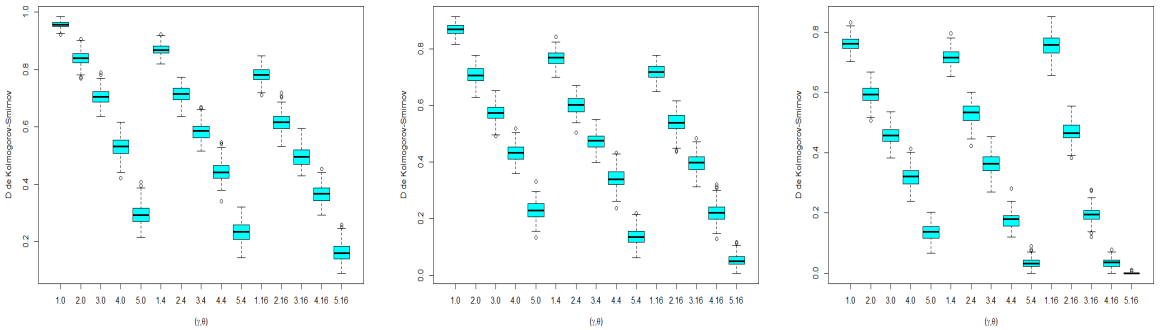
(c) Cenários: $\theta = 16$, $\gamma \in \{1,\ldots,6\}$, $p = 0{,}9$ e $m = 20$

Figura 6 – Perfil da distribuição quantílica dos graus para o modelo de afinidade cardinal dado γ

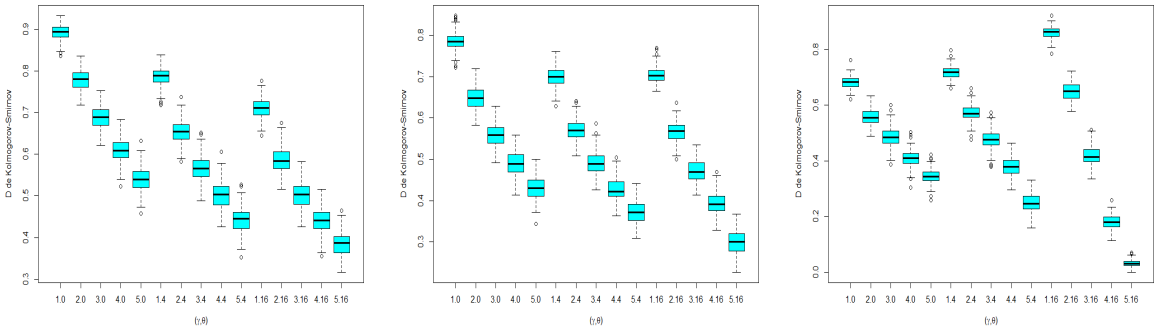
estamos comparando um grafo de afinidade com ponto de corte γ com o grafo que possui $\gamma+1$. Um padrão constante nesta comparação é o fato das diferenças entre as distribuições de grau



(a) $\theta \in \{0,4,16\}$, $p = 0.3$, $m = 320$ **(b)** $\theta = \{0,4,16\}$, $p = 0.3$, $m = 80$ **(c)** $\theta = \{0,4,16\}$, $p = 0.3$, $m = 20$



(d) $\theta = \{0,4,16\}$, $p = 0.6$, $m = 320$ **(e)** $\theta = \{0,4,16\}$, $p = 0.6$, $m = 80$ **(f)** $\theta = \{0,4,16\}$, $p = 0.6$, $m = 20$



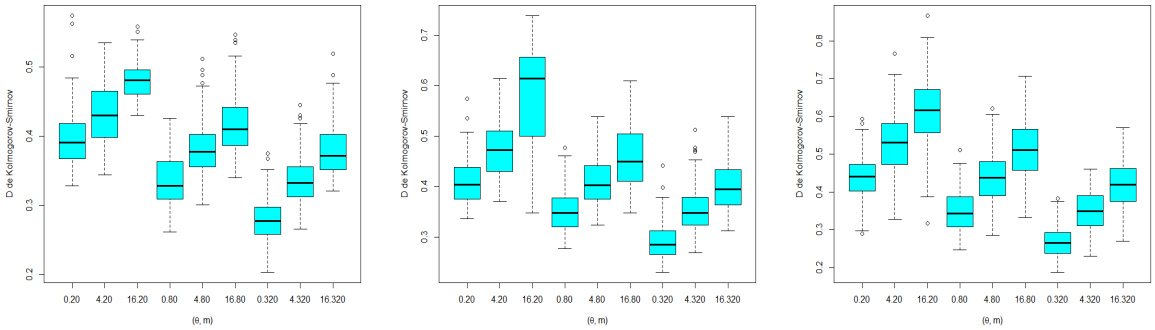
(g) $\theta = \{0,4,16\}$, $p = 0.9$, $m = 320$ **(h)** $\theta = \{0,4,16\}$, $p = 0,9$, $m = 80$ **(i)** $\theta = \{0,4,16\}$, $p = 0,9$, $m = 20$

Figura 7 – Estatística D de Kolmogorov-Smirnov para distância entre as distribuições de graus dos modelos de afinidade cardinal dado γ

de $G(n,m,\mu,f,\gamma)$ e $G(n,m,\mu,f,\gamma+1)$ diminui à medida que γ cresce. Em outras palavras, o impacto que γ exerce sobre a distribuição de graus $G(\lambda)$ vai diminuindo enquanto o mesmo cresce. Além disso, há uma tendência a diferença como um todo ser reduzida a medida que θ cresce. O único cenário que foge a este padrão é o cenário $(16,20)$, onde a diferença é maior que nos demais valores de θ. Isto indica que aliar m pequenos com θ grande garante muitas conexões formado por um grupo pequeno de características, mas estas conexões estão basicamente relacionadas a estas tais características. Caso não ocorra uma conexão neste grupo,

havendo menos características adicionais, há uma menor probabilidade de uma conexão através de características com probabilidade de escolha baixa.

A Figura 8 apresenta o perfil do comportamento da estatística D de Kolmogorov-Smirnov resultante da comparação entre as distribuições de grau da função afinidade e o modelo de arestas independentes cujos pesos foram gerados sob a distribuição Poisson binomial com parâmetros Δ. Podemos perceber que, em concordância ao resultado de Fill, Scheinerman e Singer-



(a) $\theta \in \{0,4,16\}$, $p = 0.3$, $m \in \{20,80,320\}$ **(b)** $\theta \in \{0,4,16\}$, $p = 0.6$, $m \in \{20,80,320\}$ **(c)** $\theta \in \{0,4,16\}$, $p = 0,9$, $m \in \{20,80,320\}$

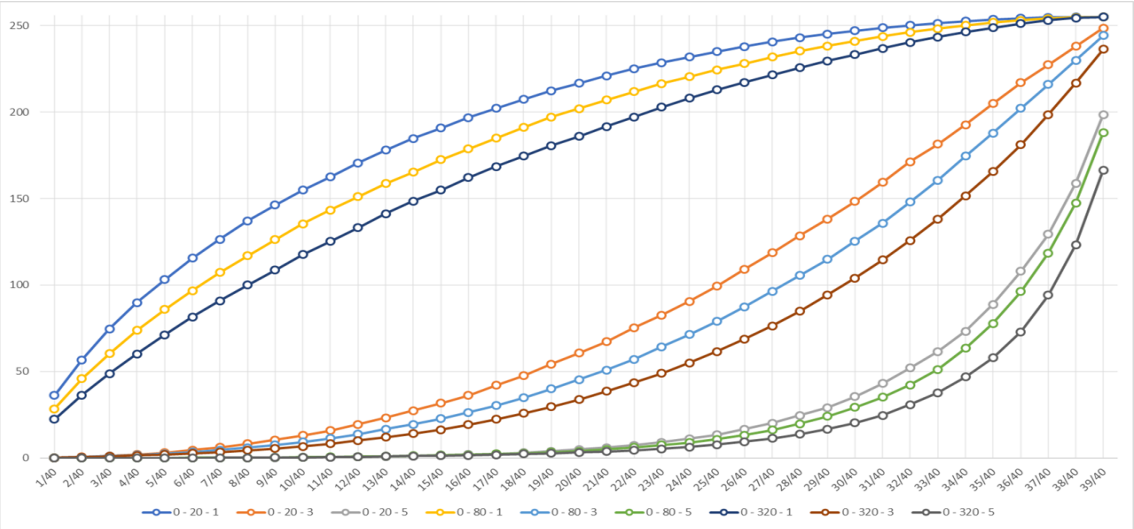
Figura 8 – Estatística D de Kolmogorov-Smirnov para distância entre as distribuições de graus dos modelos de afinidade e o grafo de arestas independentes

Cohen (2000), aumentar o valor de m resulta na diminuição da distância entre a distribuição de graus do modelo de afinidade cardinal e a distribuição de graus do modelo de arestas independentes. Por outro lado, nós podemos perceber que elevar o valor de θ aumenta tal distância. Percebemos que o modelo $(0,320)$ é, para todos os p exibidos, o modelo mais próximo do modelo de arestas independentes. Por outro lado, o modelo $(16,20)$ é o modelo mais distante do modelo de arestas independentes.

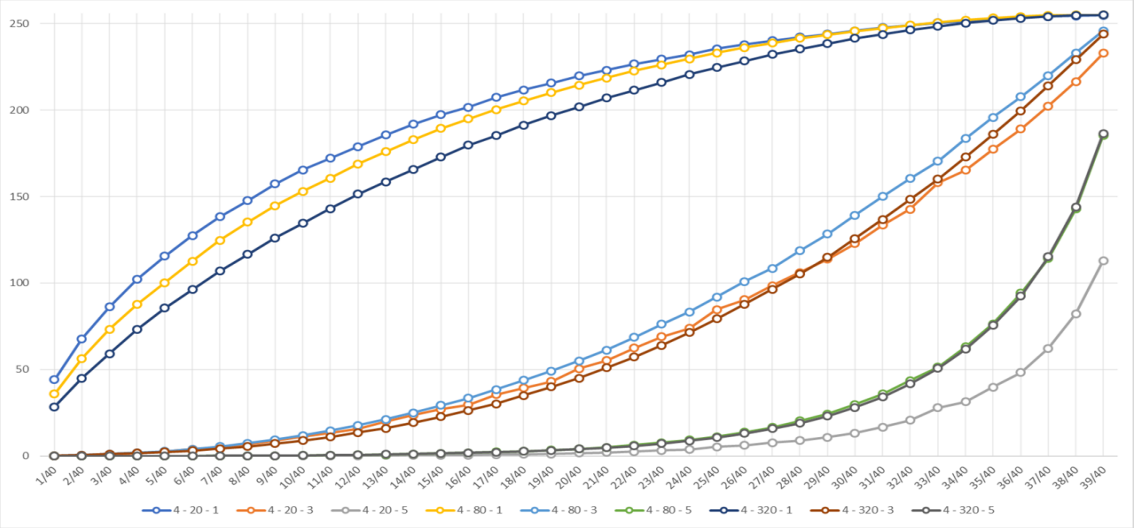
5.2.3 MEDIDAS TOPOLÓGICAS

Vamos agora apresentar os resultados obtidos para os perfis de algumas medidas topológicas do grafo de afinidade cardinal. Os perfis aqui apresentados foram analisados ao longo das probabilidades de conexão p.

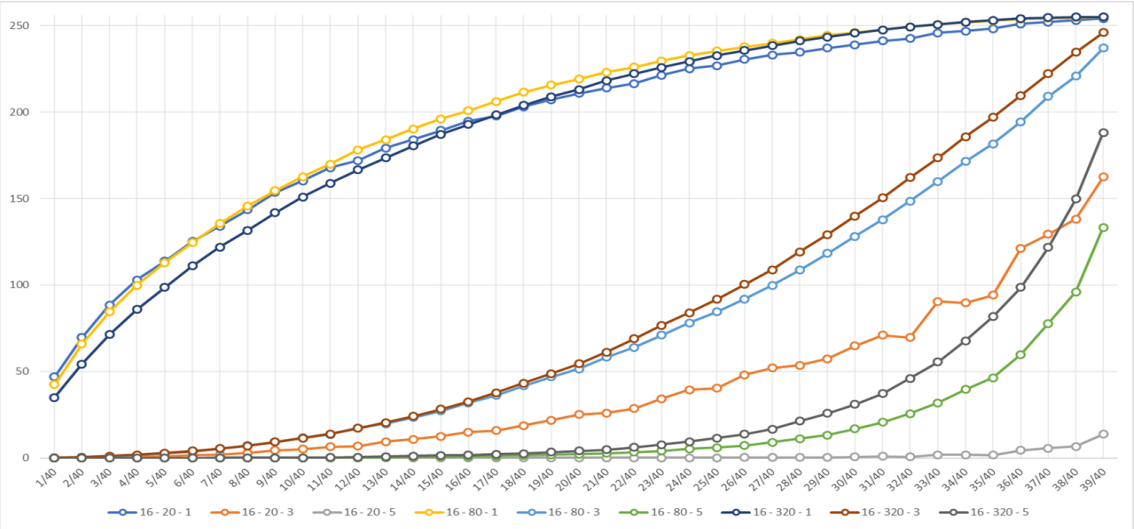
A Figura 9 apresenta o perfil da esperança de Monte Carlo para o grau máximo observado. Para $\theta = 0$, observamos os perfis bastantes comportados. Diminuir m representa, ao longo de p sistematicamente um aumento no grau máximo esperado. As curvas para cada m possuem concavidade voltada para baixo e são crescentes ao longo de p. Mediante ao aumento em γ, podemos observar que o grau máximo esperado se comporta de maneira similar para cada m no sentido de que aumentar diminuir m aumenta o grau máximo esperado. No entanto, a concavidade da curva é invertida para os outros γ exibidos. Há uma tendência de aproximação dos graus máximos esperados para p próximos a 0 e a 1 quando $\gamma = 1$. Nos valores de p



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

Figura 9 – Perfil da esperança de Monte Carlo para o grau máximo para a função afinidade cardinal

intermediários, encontramos um maior afastamento dos graus esperados para os valores de m. Para $\gamma = 3$, observamos este afastamento a partir de $p \geq 0,375$ e $p \geq 0,7$ para $\gamma = 5$.

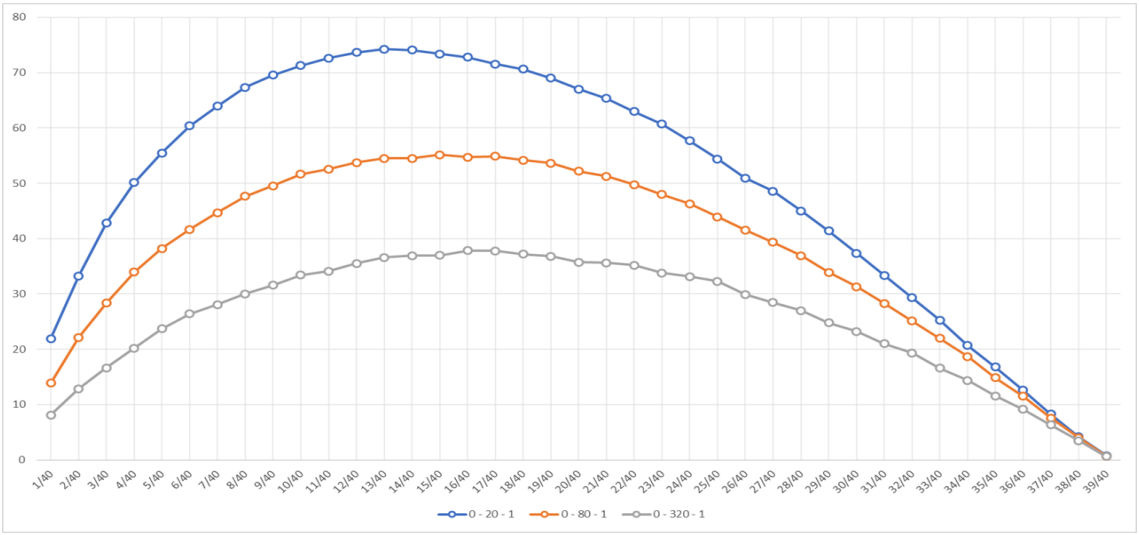
Para $\theta = 4$, observamos padrões distintos dos observados em $\theta = 1$. Existem aproximações entre as curvas dos perfis para $\gamma = 1$. A partir de $p = 0,7$, o grau máximo esperado para $m = 20$ e $m = 80$ parecem ter aproximadamente o mesmo valor esperado. Para $p \geq 0,9$, as três curvas têm basicamente o mesmo comportamento. Para $\gamma = 3$, observamos que as curvas não só se aproximam, como a curva para $m = 20$ cruza as duas outras curvas. Entre $p = 0,25$ e $p = 0,7$, a curva para $m = 20$ fica abaixo da curva $m = 80$ e acima da curva $m = 320$. Para $p > 0,7$, a curva para $m = 20$ fica abaixo das outras duas. Para $p \leq 0,05$, o grau máximo esperado é bastante próximo a zero para os três valores de m. As curvas são convexas mantém padrão similar ao observado para $\theta = 0$.

Para $\theta = 16$, a curva dos perfis para $\gamma = 1$ se aproximam ainda mais. A curva para $m = 20$ cruza as demais curvas. Para $p \leq 0,125$, temos que a esperança do grau máximo aumenta quando diminuímos o m. Por outro lado, entre $p = 0,125$ e $p = 0,425$, a curva para $m = 20$ fica abaixo das demais. Para $p \geq 0,75$, as curvas para $m = 80$ e $m = 320$ têm comportamento bastante similar. Para os demais valores de γ, temos a inversão do cenário, sendo que aumentar m aumenta o grau máximo esperado. Há um decaimento substancial do grau máximo para a curva $m = 20$ em relação aos outros cenários de θ, levando tal curva a se afastar bastante das demais curvas. O mesmo acontece para $\gamma = 3$. As concavidades das curvas mantém padrão similar ao observado para $\theta = 0$.

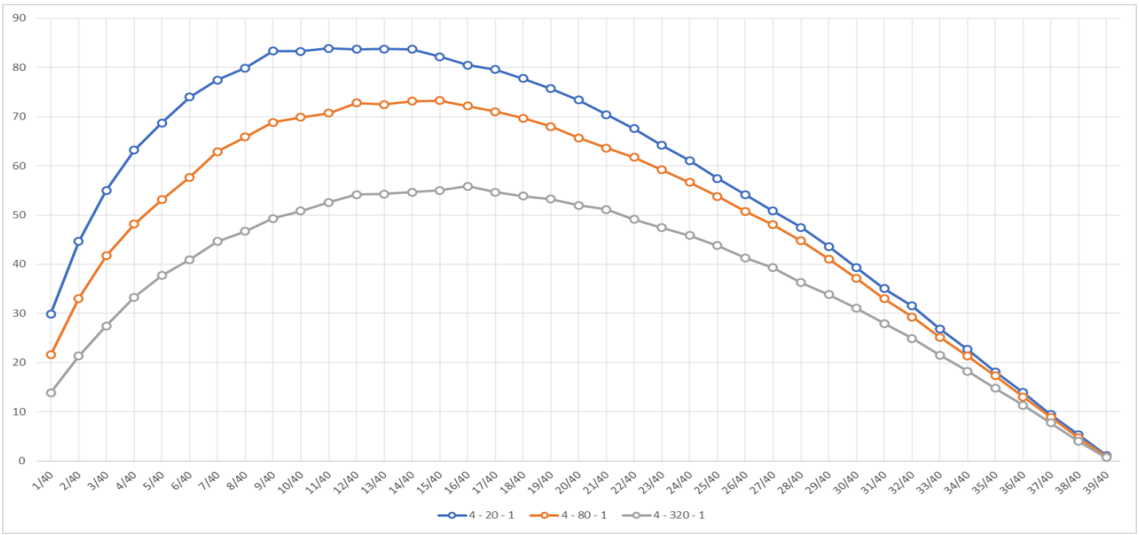
Em linhas gerais, para θ grande, as curvas que mais sofrem com perda em termos de valores esperados para o grau máximo são as curvas com m pequeno. Isto pode ser explicado pelo fato de θ pequeno induzir ligações cuja $E(f(U_i, U_k))$ é menor que as demais. Com menos características disponíveis, há menor probabilidade de haver conexões através de características com probabilidade de escolha baixa. Assim, há maior perda de arestas neste modelo, consequentemente diminuindo o grau máximo esperado.

A Figura 10 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para o grau máximo do modelo de afinidade cardinal e o da esperança de Monte Carlo para o grau máximo para o modelo de arestas independentes. Para $\theta = 0$, nós podemos observar que quanto maior o valor de m, mais próximo sua curva estará do grau máximo da curva do grau máximo do modelo de arestas independentes. Isso vai ao encontro ao resultado de Fill, Scheinerman e Singer-Cohen (2000). Além disso, é interessante observar que a curva se aproxima de uma parábola de concavidade para baixo. A distâncias alcançam o seus máximos próximo de pontos diferentes: cerca de $p = 0,325$ para $m = 20$, $p = 0,4$ para $m = 80$ e $p = 0,425$ para $m = 320$. Este máximo parece mais destacado a medida que m diminui. Nos valores extremos de p mais extremos, as curvas se aproximam, com uma aproximação maior quando $p = 1$ do que quando $p = 0$.

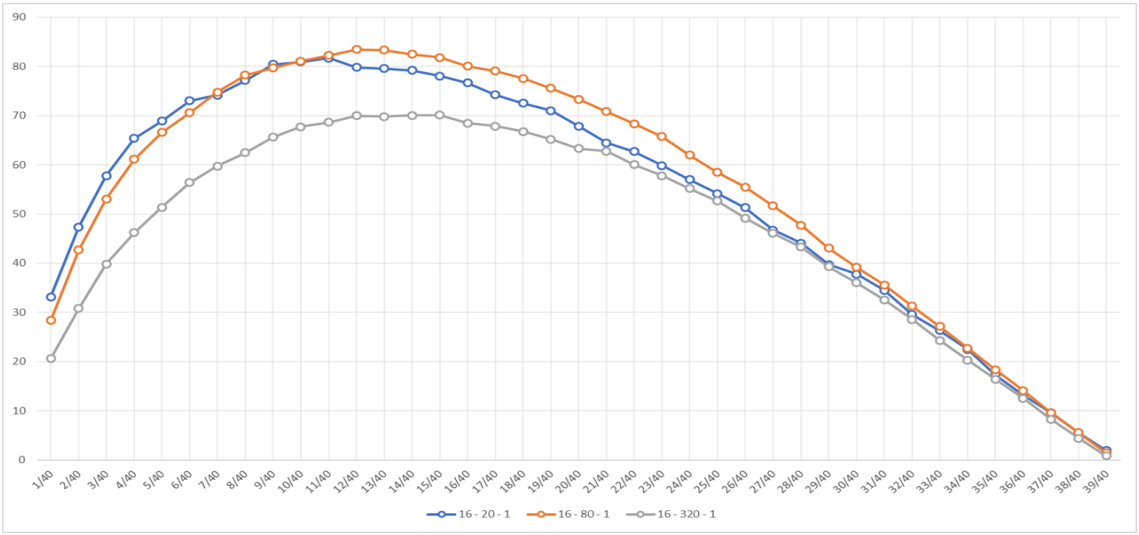
Para $\theta = 4$, as curvas tem comportamento parecido. No entanto, há um aumento na



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

Figura 10 – Perfil da diferença entre as esperanças do grau máximo do modelo de afinidade cardinal e o modelo de aresta independentes

diferença entre o grau máximo do modelo de afinidade cardinal e o grau máximo do modelo de arestas independentes. Além disso, as curvas se aproximam umas das outras, sendo que o o máximo das curvas diminuem para cerca de $p = 0,25$ para $m = 0$, $p = 0,375$ para $m = 80$ e $p = 0,4$ para $m = 320$.

Para $\theta = 16$, as curvas tem comportamento um pouco distinto do que os demais θ. Há uma maior aproximação da curva, bem como aumento das distâncias gerais entre o grau máximo dos modelos de afinidade e dos modelos de arestas aleatórios. No entanto, a curva para $m = 20$ cruza a curva de $m = 80$. Para $p \geq 0,25$, a distância entre a o modelo $m = 20$ para o seu correspondente de arestas independentes é menor do que para $m = 80$. O maximante das curvas é reduzido para cerca de $p = 0,225$ para $m = 0$, $p = 0,3$ para $m = 80$ e $p = 0,35$ para $m = 320$.

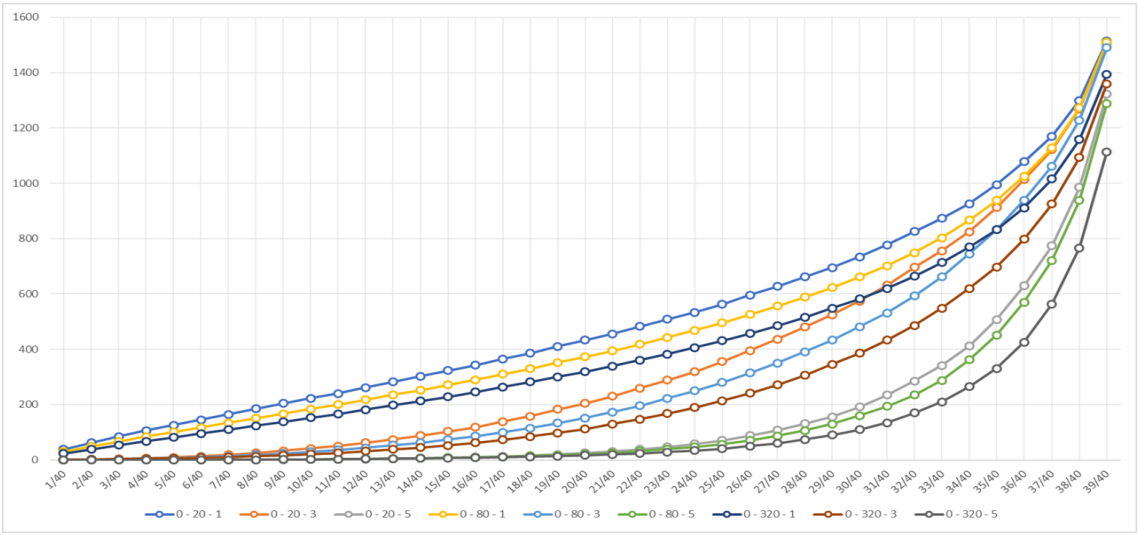
Em termos gerais, podemos perceber que, dados os valores de diferenças positivos, a função de afinidade cardinal tem tendências a ter o grau máximo maior do que o modelo de arestas aleatórias, com esta diferença aumentando quando θ aumenta e diminuindo quando m aumenta.

A Figura 11 apresenta o perfil da esperança de Monte Carlo para a força máxima observada. Assim como observado para o grau máximo, para $\theta = 0$ encontramos um padrão bem comportado das curvas para a força máxima. Para $\gamma = 1$, diminuir m implica em aumentar a força máxima esperada. No entanto, ao contrário do que observamos para o grau máximo, as curvas da força máxima são convexas para todos os γ apresentados. Além disso, as curvas são crescentes ao longo de p. Assim como para o grau máximo, há uma aproximação das curvas em valores próximos a $p = 0$ e $p = 1$ quando para $\gamma = 1$.

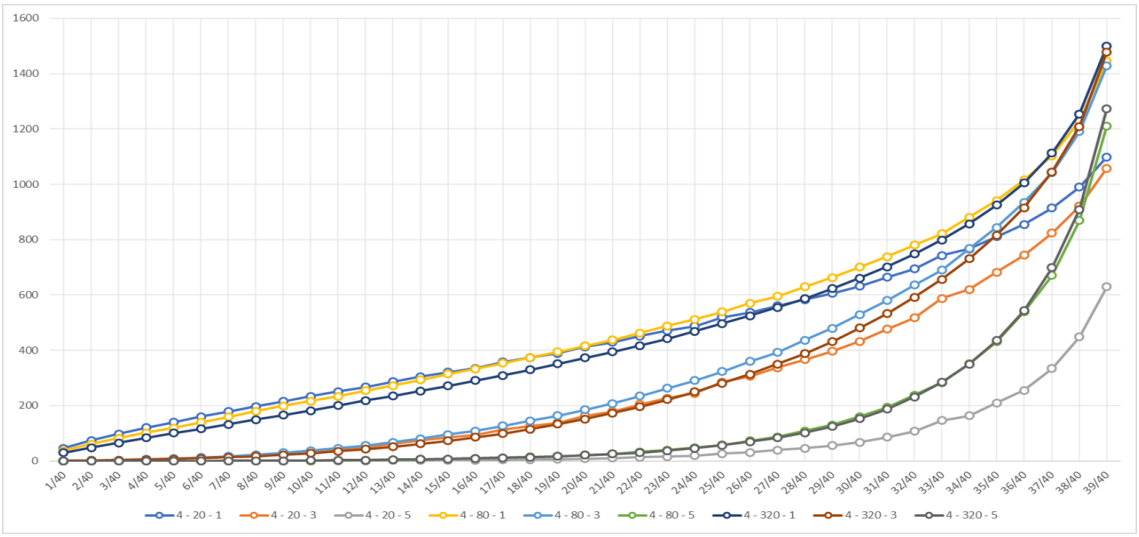
Para $\theta = 4$, já observamos as curvas as quais $m = 20$ cruzando as demais curvas, tendo maior força máxima esperada para pequenos valores de p e menor força máxima esperada para grandes valores de p quando $\gamma = 1$. Para $\gamma = 3$, observamos tal curva com um comportamento bastante similar à curva $m = 320$, ficando abaixo da mesma a partir de $p = 0.65$. Para $\gamma = 5$, observamos a inversão da ordem das curvas, fazendo com que quanto maior m maior a força esperada.

Para $\theta = 16$, observamos para $\gamma = 3$ e $\gamma = 5$ que aumentar m aumenta a força máxima esperada. Para $\gamma = 1$, observamos o cruzamento das curvas, onde a partir de $p = 0,275$ a curva para $m = 20$ tem menor força máxima esperada. A partir de de $p = 0,525$, as curvas $m = 80$ e $m = 320$ se cruzam. Abaixo desta probabilidade de conexão, $m = 80$ tem maior esperança da força máxima, enquanto acima do mesmo isto ocorre para $m = 320$.

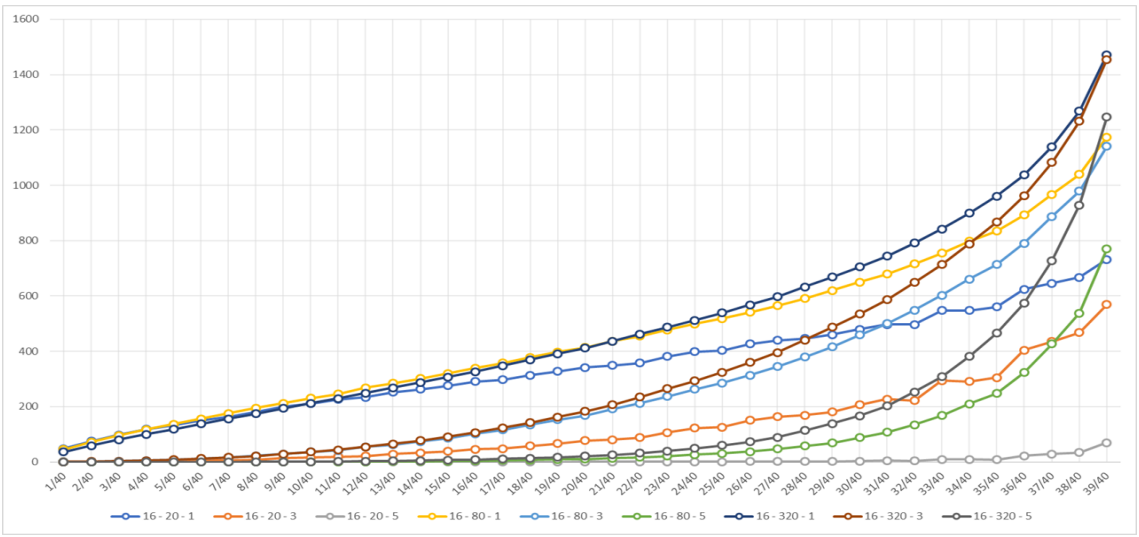
Em linhas gerais, podemos perceber que $m = 20$ se distancia do comportamento comportamento dos outros valores de m à medida que θ aumenta. Além disso, podemos perceber que, com um valor razoável de m, o grau máximo quase não sofre alterações para altos valores de p. Além disso, uma diferença interessante em relação ao grau máximo é o fato de que para p e γ alto, em especial $p = 0,975$, as curvas que mais se aproximam de $m = 320$ com $\gamma = 1$ são $m = 320$ com $\gamma = 3$ e $m = 320$ com $\gamma = 5$. Em outras palavras, a força máxima resistiu melhor



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

Figura 11 – Perfil da esperança de Monte Carlo para a força máxima para a função afinidade cardinal

à perda de aresta do que a perda de características nestes casos.

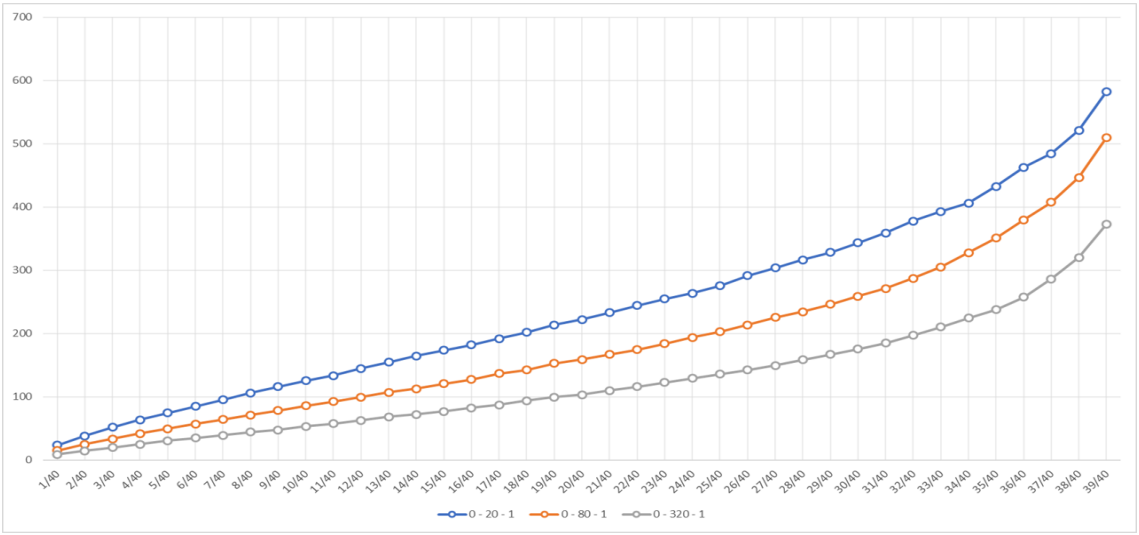
A Figura 12 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para a força máxima do modelo de afinidade cardinal e o da esperança de Monte Carlo para a força máxima para o modelo de arestas independentes. Assim como observado para o grau máximo, a força máxima concorda com o resultado de Fill, Scheinerman e Singer-Cohen (2000): aumentar m quando $\theta = 0$ reduz a distância entre a força máxima esperada para o modelo de afinidade cardinal e a força máxima esperada para a força máxima para o modelo de arestas independentes. Em contraste com o que foi observado para o grau máximo, observamos uma função crescente ao longo de p para a força máxima. Para $\theta = 0$, à medida em que p aumenta, esta distância também aumenta.

Para $\theta = 4$, observamos um aumento na distância entre a força máxima para o modelo de afinidade cardinal e a força máxima para o modelo de arestas aleatórias quando $m = 80$ e $m = 320$. Por outro lado, observamos uma redução desta distância para $m = 20$, especialmente para valores altos de p, fazendo com que esta curva cruze as demais, sendo sua força máxima esperada menor que $m = 80$ para $p = 0,65$ e menor que $m = 320$ para $p = 0,9$.

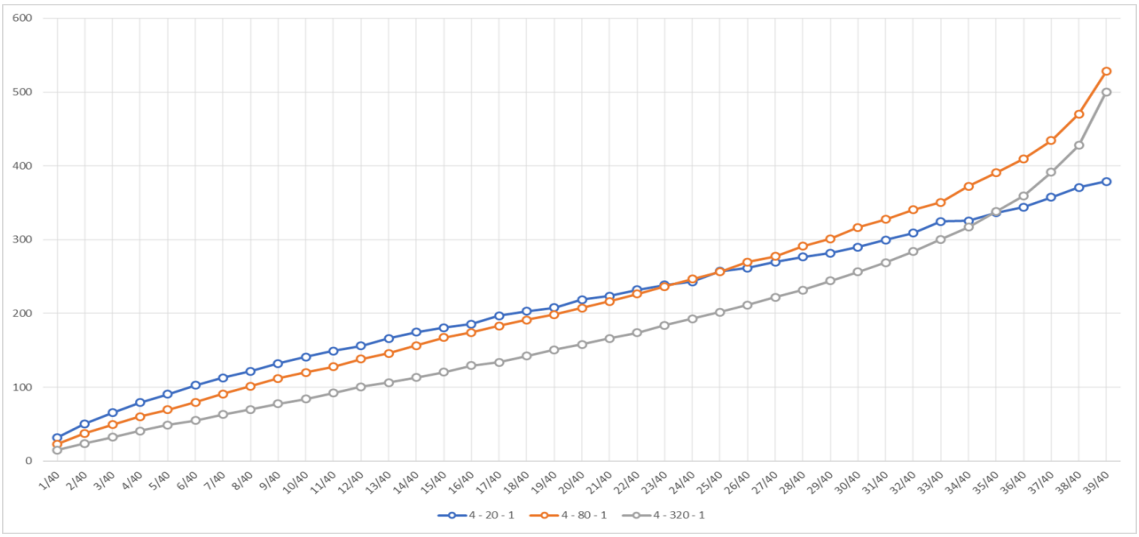
Para $\theta = 16$, observamos mais aumento na distância entre a força máxima para o modelo de afinidade cardinal e a força máxima para o modelo de arestas aleatórias quando $m = 320$, enquanto observamos redução desta distância nas demais curvas. A partir de $p = 0,675$, a curva $m = 320$ apresentou a maior força esperada. O mesmo ocorre a partir de $p = 0,3$ em relação à $m = 20$. Quanto à $m = 80$, a partir de $p = 0,15$, a curva tem força máxima esperada maior que $m = 20$.

Em linhas gerais, podemos observar que θ tem impacto na distância entre os modelos de afinidade e os modelos de arestas independentes. A distância menor de $m = 20$ quando $\theta = 16$ da força máxima esperada para modelos de aresta aleatórias pode estar associada a fatores como a limitação do número de características, bem como ao valor das conexões. Perceba que de acordo com de acordo com a Equação 4.60, estamos dividindo o valor de p em m parcelas. Quanto menor o valor de m, espera-se que maior seja o valor de cada parcela, especialmente se $\theta = 0$, onde estaríamos dividindo em m parcelas iguais. Logo, a probabilidade de escolha de cada palavra neste caso para o mesmo p é maior para $m = 20$ do que para $m = 320$. Isto reflete no fato da força máxima ser maior para $m = 20$. Por outro lado, quando concentramos probabilidade em um grupo específico de características, há um efeito inverso, pois a probabilidade de haver ligações fora deste grupo de características é reduzida quanto menor for m. Em valores absolutos, isto torna o número de características compartilhadas entre indivíduos menor para $m = 20$ do que para valores de m mais baixos, fazendo que com que a distância para o modelo de arestas aleatórios diminua. Uma vez que as distância são positivas, podemos ver que se espera forças máximas mais altas no modelo de afinidade cardinal do que no modelo de arestas aleatórias.

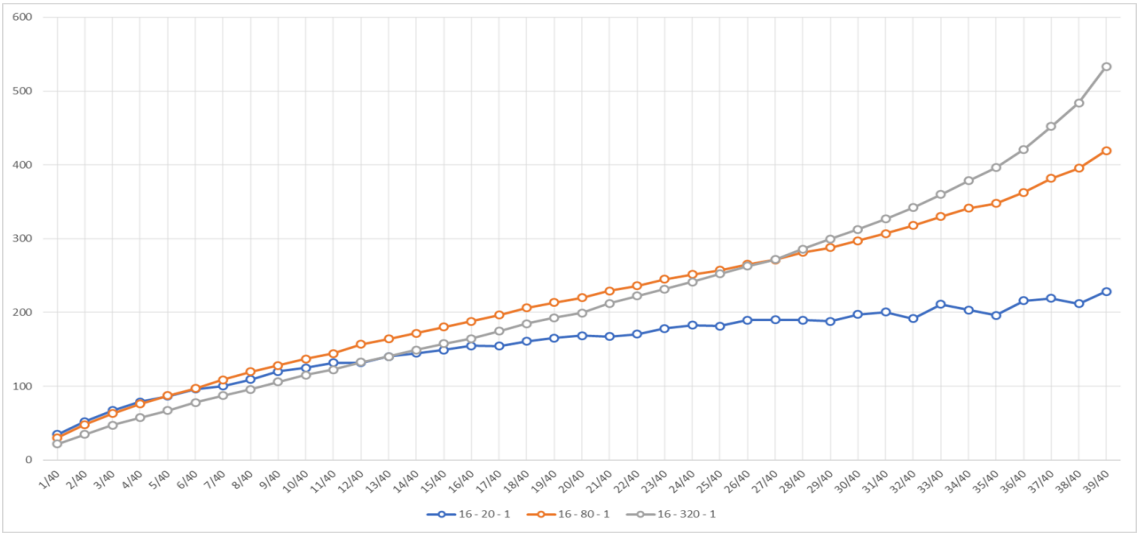
A Figura 13 apresenta o perfil da esperança de Monte Carlo para a transitividade obser-



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

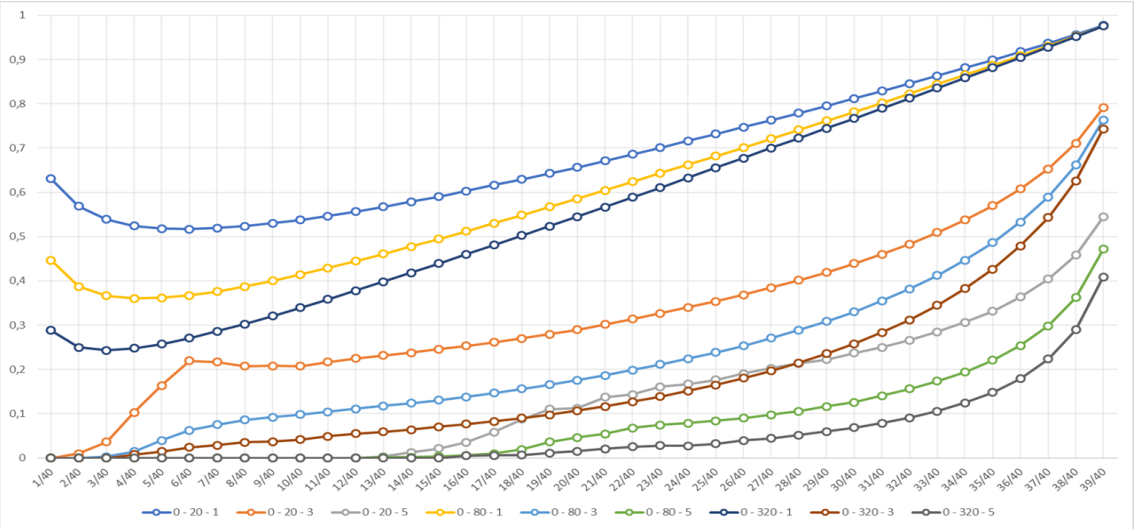
Figura 12 – Perfil da diferença entre as esperanças da força máxima do modelo de afinidade cardinal e o modelo de aresta independentes

vada. Para todos os valores de θ, quando $\gamma = 1$ que quanto maior o valor de m, menor é a transitividade observada. Observamos que existe um comportamento similar a uma parábola com concavidade para cima, com decaimento para os valores pequenos de p, atingindo um ponto mínimo e em seguida crescendo para os demais valores de p. O mínimo é alcançado em cerca de $p = 0.125$ para $m = 20$, $p = 0.1$ para $m = 80$ e $p = 0.075$ para $m = 320$ para $\theta = 1$, $p = 0.175$ para $m = 20$, $p = 0.125$ para $m = 80$ e $p = 0.1$ para $m = 320$ para $\theta = 3$ e $p = 0.225$ para $m = 20$, $p = 0.15$ para $m = 80$ e $p = 0.125$ para $m = 320$. Percebemos que há um aumento da transitividade à medida em que θ cresce.

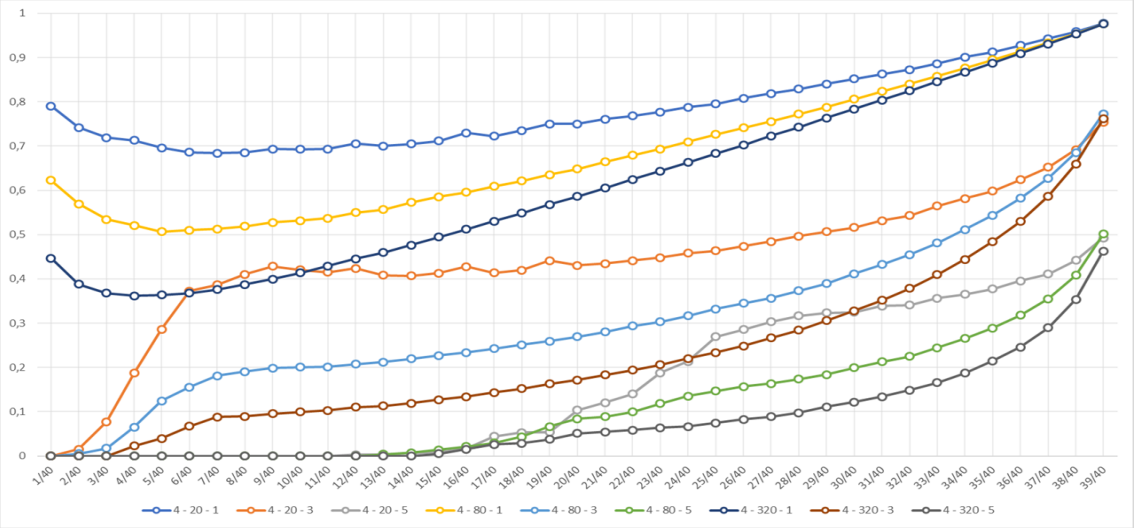
Para os demais valores de γ observamos a existência de transições de fases para a transitividade, cujas quais o ponto de início da fase de crescimento mais rápido é deslocada pra valores mais altos de p conforme γ cresce e m diminui. A exceção ocorre para $m = 20$ quando $\gamma = 5$ e $\theta = 16$, onde há até $p = 0,525$ há transitividade esperada próxima de 0. Para $\theta = 16$, a transitividade para $m = 20$ torna-se descoordenada e possuim um comportamento um pouco distinto das outras curvas, que seguem um padrão similar.

Podemos dizer que concentrar probabilidade em um grupo específico de características faz com que tal grupo seja escolhido mais vezes e então, escolher uma palavra deste grupo aumenta substancialmente a probabilidade de fazer conexões e consequentemente triângulos. Uma explicação plausível para a queda da transitividade nos níveis iniciais de p é a de que os indivíduos para estes p tem maior propensão a escolher menos características. Logo suas conexões seriam formadas através de um grupo bastante reduzido de características. Daí, se estas pessoas escolhem apenas uma palavra, as pessoas que escolhem esta palavra formam triângulos automaticamente (desde que sejam mais de duas pessoas, evidentemente). A medida que p aumenta, os indivíduos tem propensão a escolher mais características. Daí, uma nova possibilidade torna-se mais provável: dois indivíduos que escolhem cada uma palavra que sejam diferentes um do outro e um terceiro indivíduo que escolha as duas. Logo, embora o terceiro indivíduo se conecte com os dois primeiros, este cenário não forma um triângulo. Ao concentrar mais peso em uma palavra, isto é, aumentar θ, espera-se que a probabilidade de, para um p pequeno, o indivíduo escolha uma única palavra aumente. Logo, espera-se que a taxa de queda da transitividade ao longo dos níveis inicias seja menos vertiginosa, o que pode ser observado nas Figura 13.

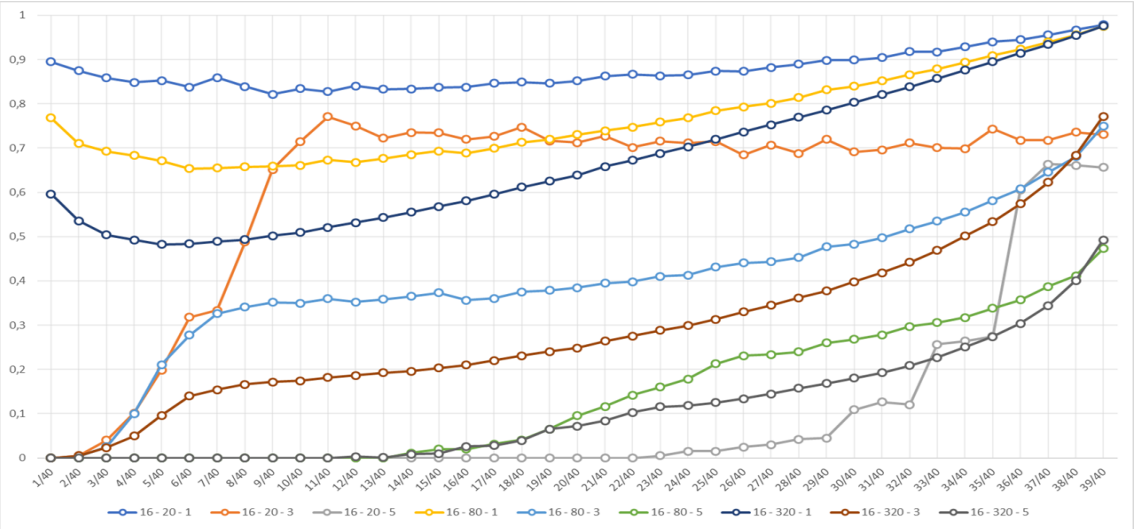
A Figura 14 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para a transitividade do modelo de afinidade cardinal e o da esperança de Monte Carlo para a transitividade para o modelo de arestas independentes. É possível perceber que, em todos os casos, a transitividade para o modelo de afinidade cardinal foi superior ao modelo de arestas independentes. Tal resultado concorda com a argumentação de Newman e Park (2003) para redes sociais. Além disso, em todos os cenários, observamos que a diferença entre as transitividades esperada aumenta ao reduzir m, o que concorda com o resultado de Fill, Scheinerman e Singer-Cohen (2000). A diferença entre as transitividades também aumenta ao aumentar θ. Em linhas gerais, a diferença é decrescente e se aproxima de zero à medida em que p se aproxima



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

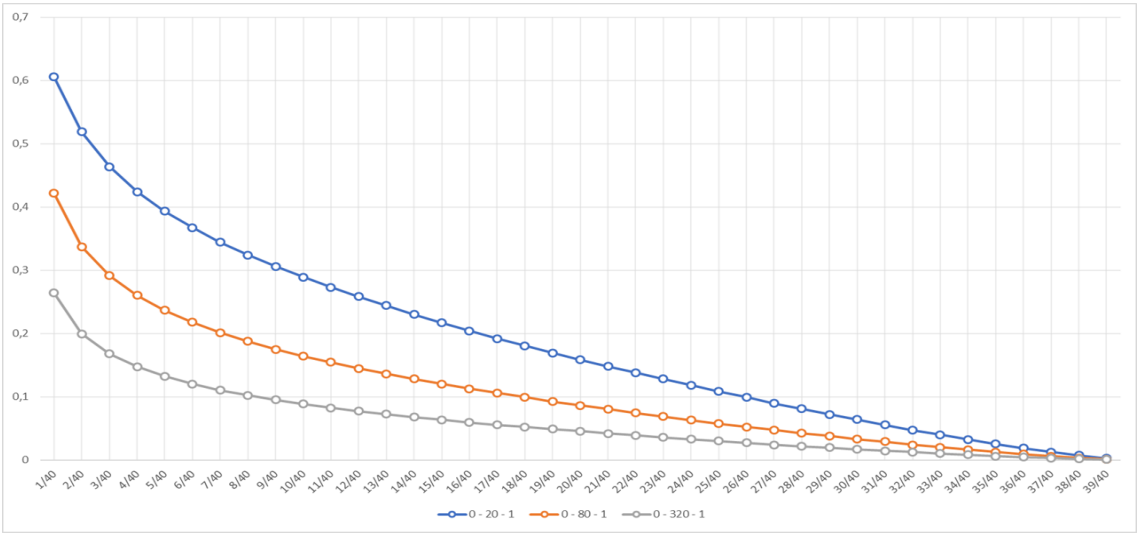


(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

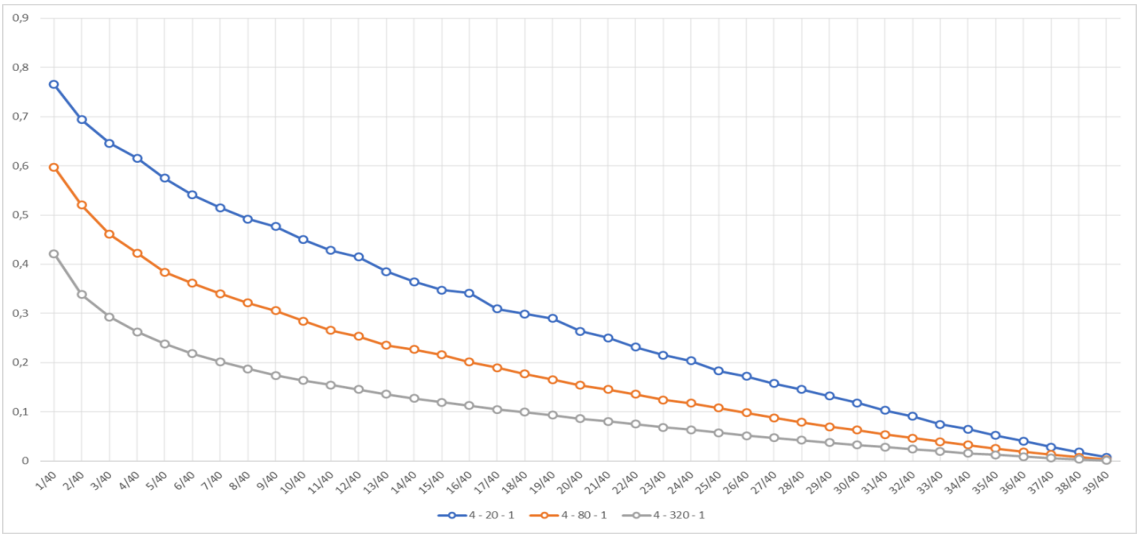


(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

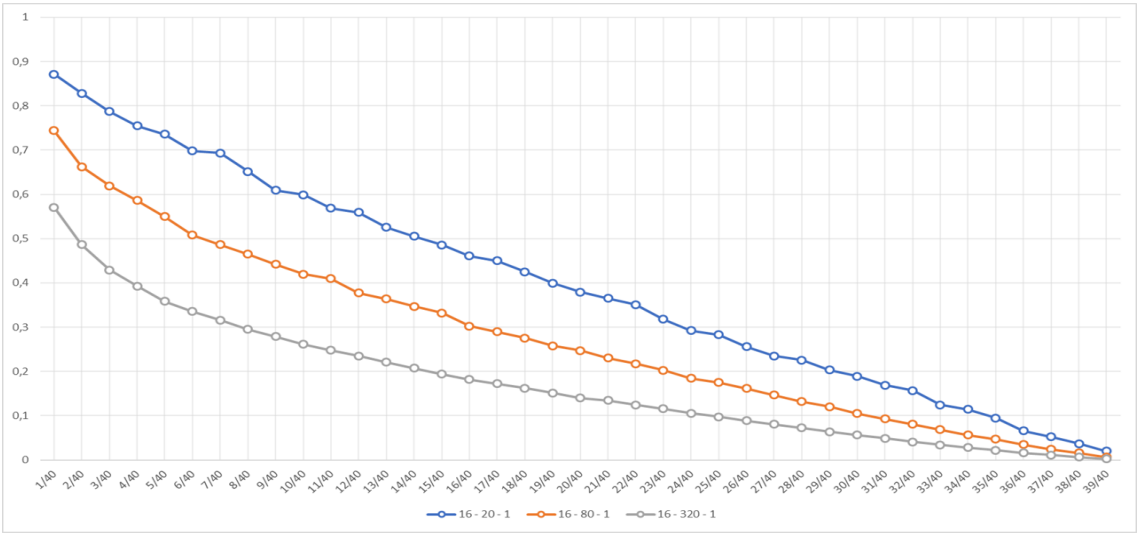
Figura 13 – Perfil da esperança de Monte Carlo para a transitividade da função afinidade cardinal



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$



(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

Figura 14 – Perfil da diferença entre as esperanças da transitividade do modelo de afinidade cardinal e o modelo de aresta independentes

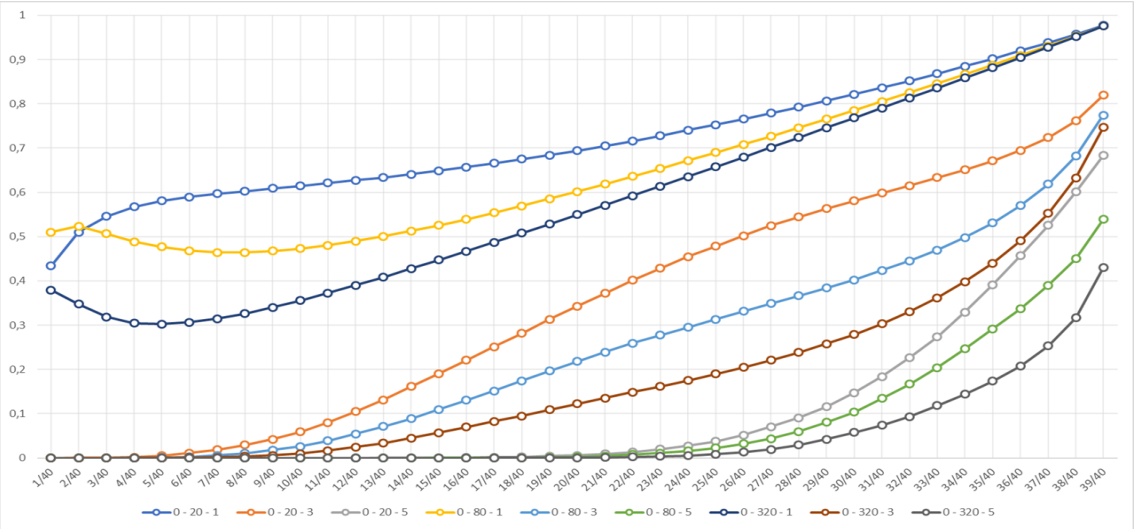
de 1.

A Figura 15 apresenta o perfil da esperança de Monte Carlo para o coeficiente de clustering observado. As curvas para cada valor de m assumem comportamentos distintos nos valores iniciais de p quando $\gamma = 1$. Aparentemente, observamos três comportamentos: crescimento ao longo de p, crescimento seguido de decrescimento e novo crescimento, e decrescimento seguido de crescimento. Não foi encontrada uma explicação intuitiva para este padrão de comportamento, entretanto, estes comportamentos parecem ser controlados pela interação (θ, m). Se θ é alto e m é baixo, o comportamento se aproxima do comportamento de crescimento ao longo de p. Se θ é baixo e m é alto, o comportamento se aproxima do decrescimento nos p iniciais seguido de crescimento, exceto para $m = 20$ onde observamos crescimento em todos os θ. As demais interações apresentam traços de crescimento seguido de decrescimento e novo crescimento. Outro detalhe interessante é que aparentemente, à medida que θ cresce, há a tendência à curva de maior m possuir a maior esperança do coeficiente de clustering, ao menos nos valores iniciais de p.

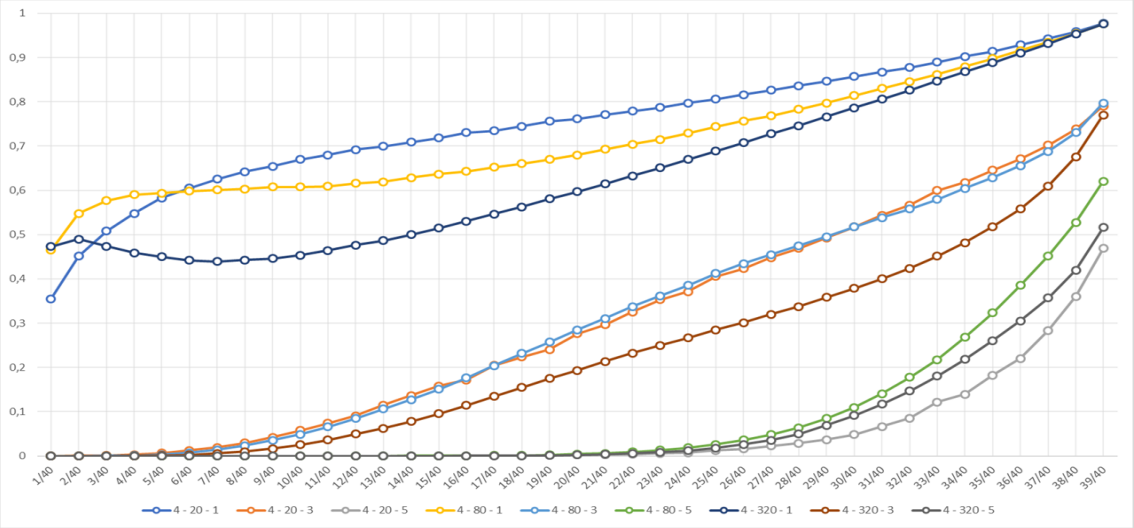
Para os demais valores de γ, observamos padrões mais claros. Observamos que a medida que θ aumenta, a tendência é que a curva de maior m tenha maior coeficiente de clustering esperado ao longo de p, enquanto se há um valor de θ próximo a 0, a tendência é que os menores valores de m possuem maior coeficiente de clustering esperado.

A Figura 16 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para o coeficiente de clustering do modelo de afinidade cardinal e o da esperança de Monte Carlo para coeficiente de clustering para o modelo de arestas independentes. Concordando com os resultados de Fill, Scheinerman e Singer-Cohen (2000) e Newman e Park (2003), o coeficiente de clustering se mostra maior que o observado em grafos de arestas independentes e quanto maior o m quando $\theta = 0$, mais próximo este coeficiente de clustering se aproxima do coeficiente de clustering observado nos grafos de arestas independentes. Podemos observar a mudança de comportamento da diferença à medida do crescimento de θ, sugerindo que o comportamento limite seria aproximadamente de uma parábola quando θ é alto. Neste caso, observamos que para os menores valores de p e θ alto, as funções de maior m tem maior diferença em relação ao modelo de aresta independentes, panorama que se inverte à medida que p cresce.

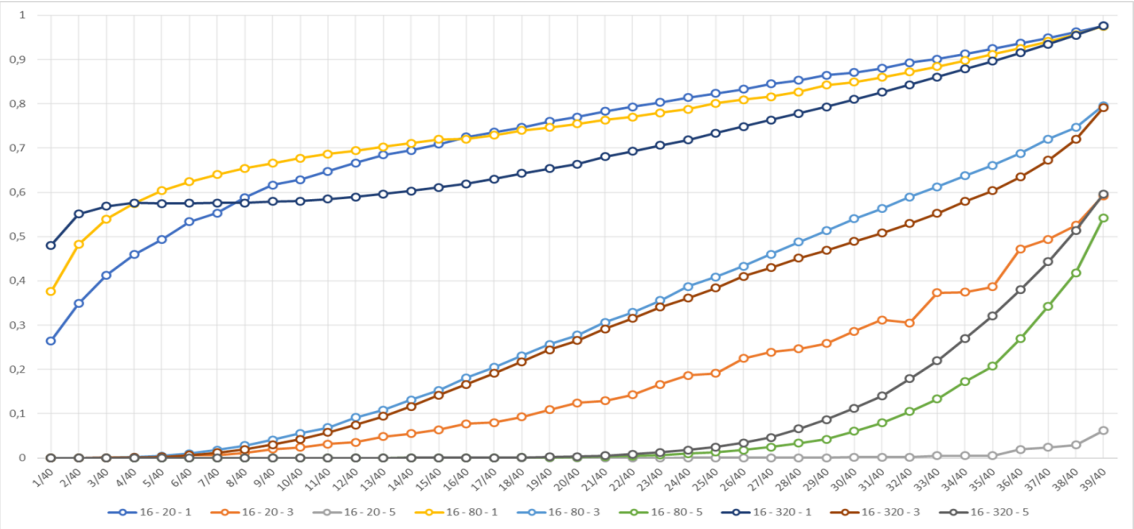
A Figura 17 apresenta o perfil da esperança de Monte Carlo para a proximidade. Podemos observar um padrão na proximidade esperada. Para todos os valores de θ e γ observados, quanto maior o valor de m, maior será a proximidade esperada. Além disso, observamos que a proximidade nos valores iniciais de p é reduzida ao aumentar o valor de θ. Isto significa que concentrar mais probabilidade em um grupo específico de características torna as distâncias maiores. Existe também um afastamento das curvas para os m à medida em que θ e γ crescem. A proximidade em geral cresce à medida em que p cresce. Para $\gamma = 1$, observamos uma curva crescente côncava. Para $\gamma = 3$, observamos pontos de inflexão para m exceto em $m = 20$ para



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

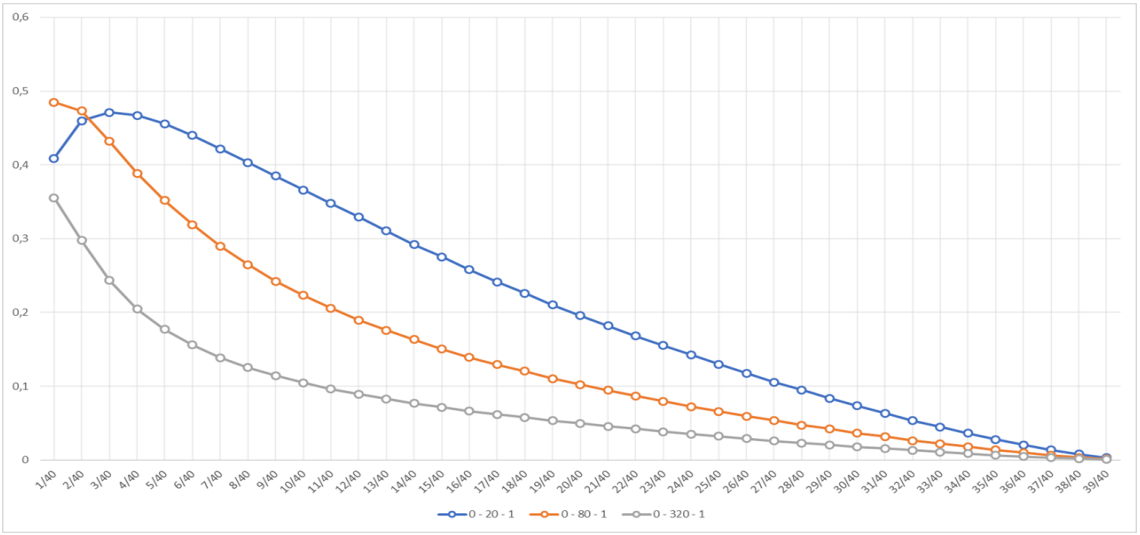


(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

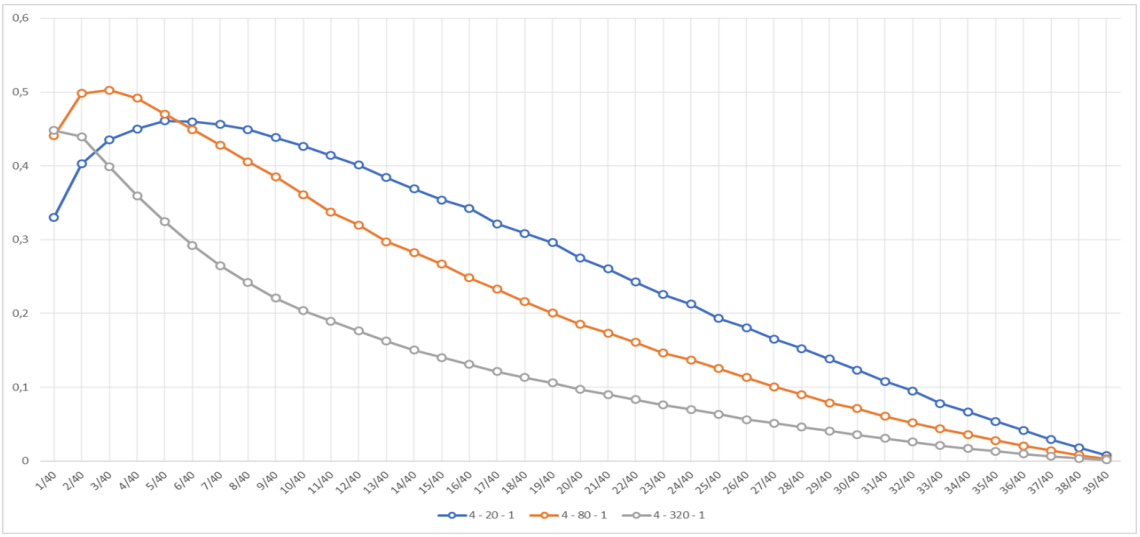


(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

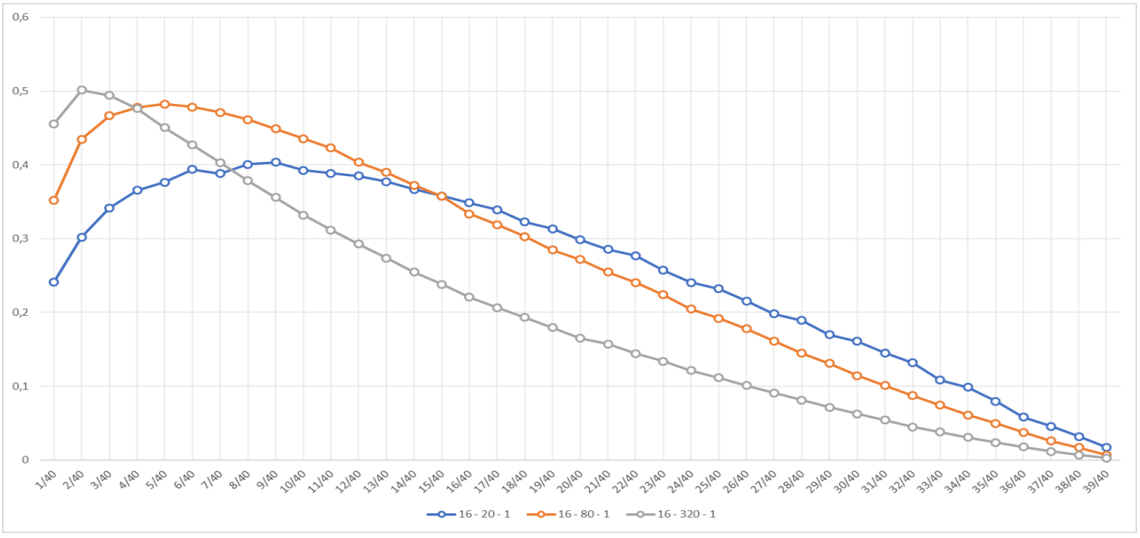
Figura 15 – Perfil da esperança de Monte Carlo para o coeficiente de clustering da função afinidade cardinal



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

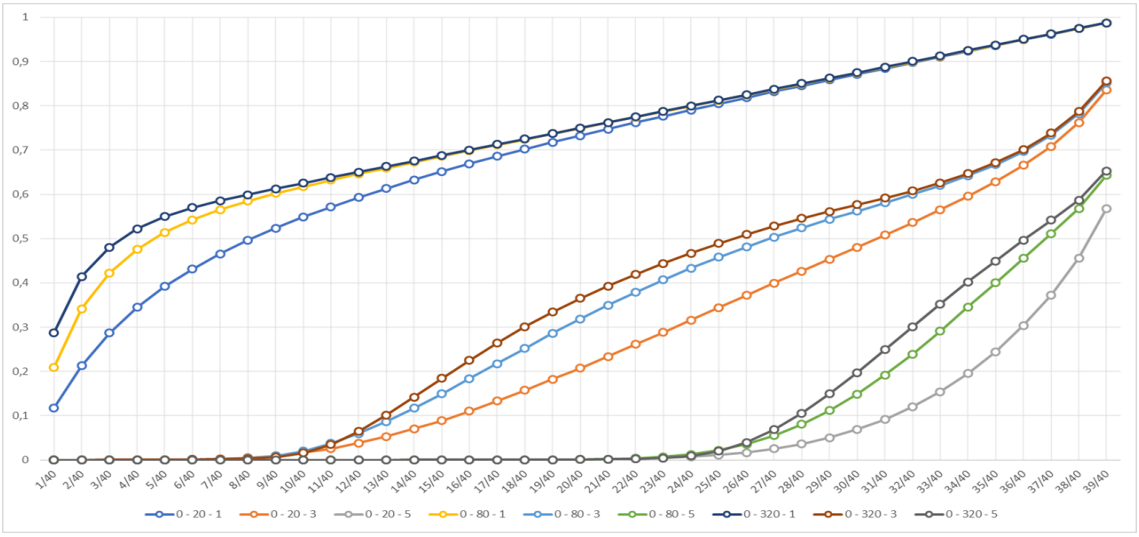


(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

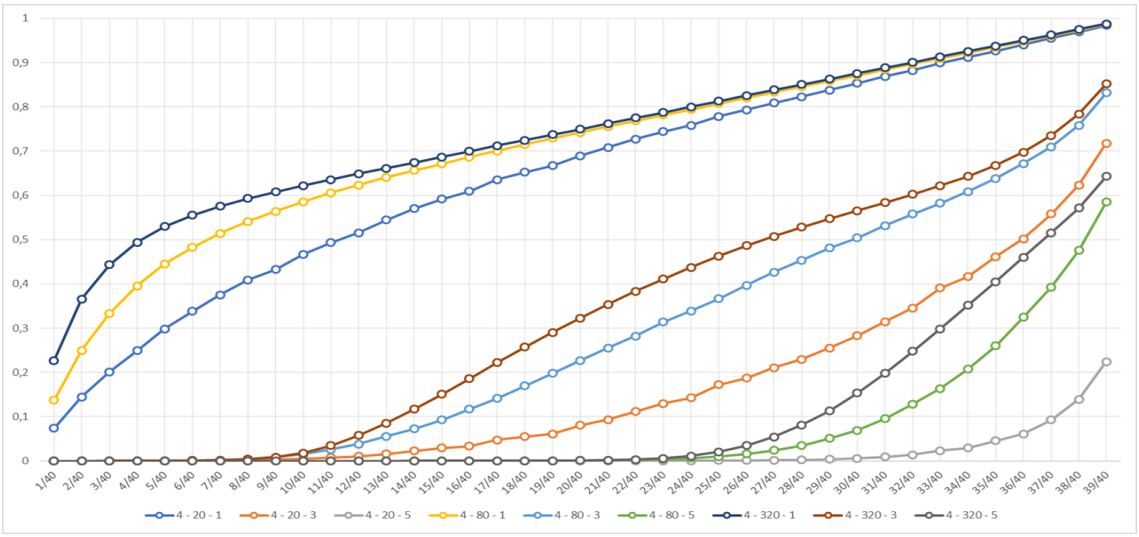


(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

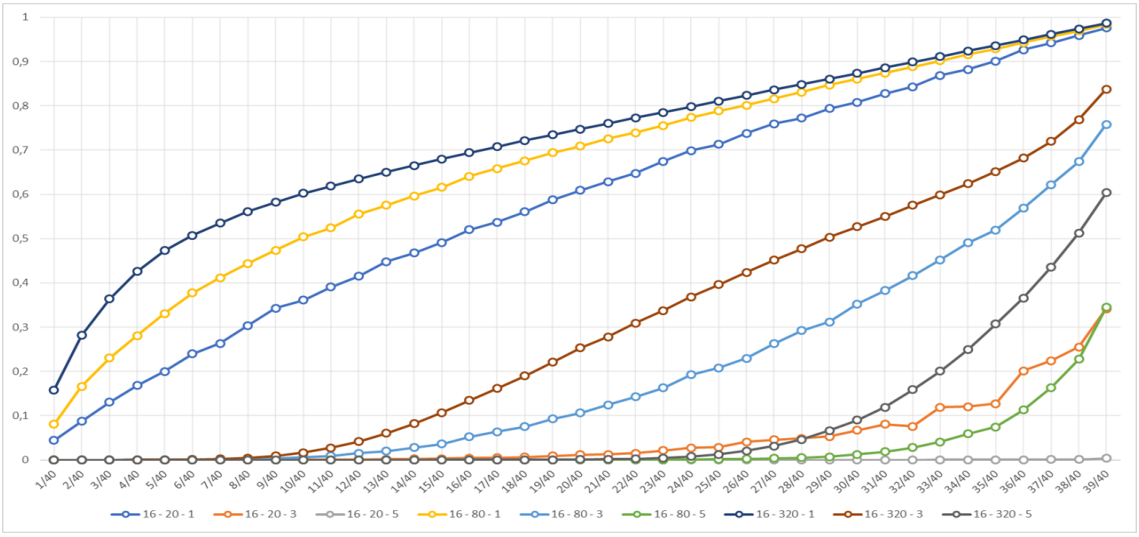
Figura 16 – Perfil da diferença entre as esperanças do coeficiente de clustering do modelo de afinidade cardinal e o modelo de aresta independentes



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

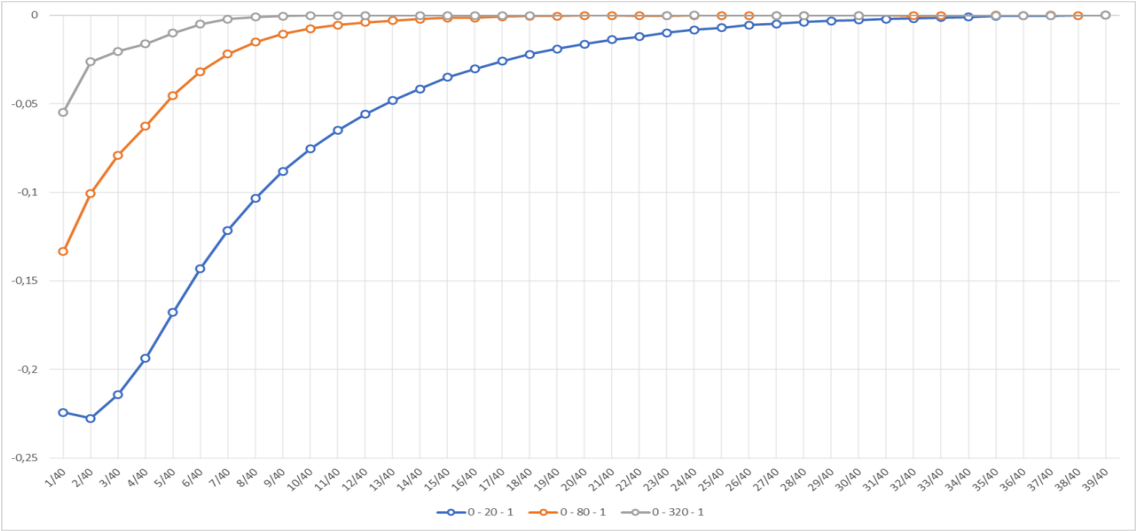
Figura 17 – Perfil da esperança de Monte Carlo para a proximidade da função afinidade cardinal

todos os θ e $m = 80$ para $\theta = 16$, enquanto para $\gamma = 5$ observamos uma curva convexa crescente.

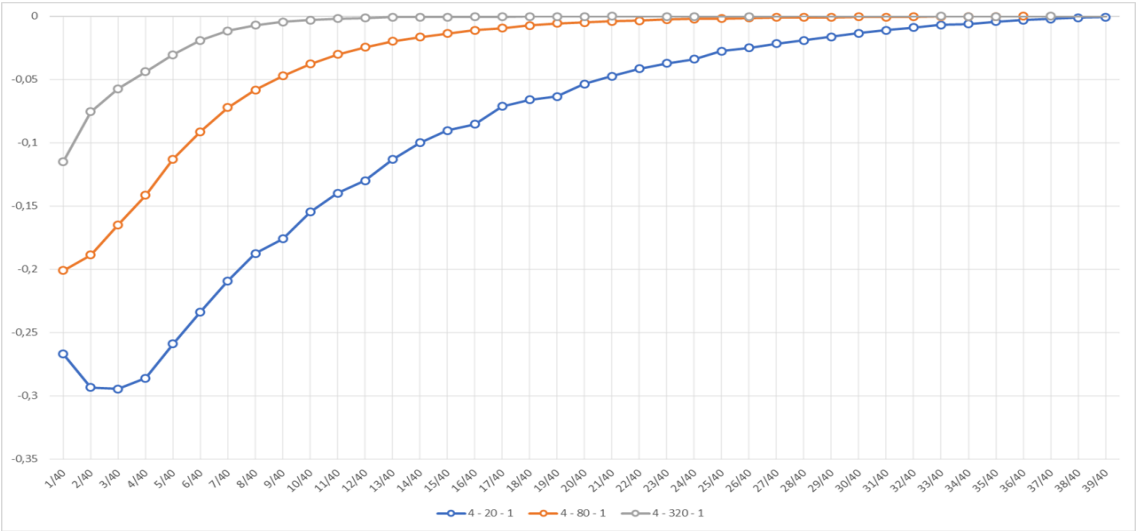
A Figura 18 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para a proximidade do modelo de afinidade cardinal e o da esperança de Monte Carlo para a proximidade para o modelo de arestas independentes. O grafo gerado pelo modelo de afinidade cardinal apresenta proximidade menor do que a observada em grafos gerados via modelo de arestas independentes. Este resultado era esperado, já que ao haver mais componentes, as distâncias observadas serão maiores, incluindo distâncias infinitas e, consequentemente, reduzindo a proximidade baseada no recíproco das distâncias. Percebemos que aumentar θ reduz a proximidade e aumenta sua distância em relação à proximidade de modelos de aresta independentes, enquanto aumentar m surte o efeito oposto, concordando com Fill, Scheinerman e Singer-Cohen (2000). Podemos observar que para m pequeno e mesmo para m grande combinado com θ grande, inicia-se o surgimento de uma mudança de concavidade nos níveis iniciais de p, havendo um aumento da distância quando p é pequeno seguido da redução para os demais p. Além disso, observamos que à medida que θ cresce as curvas para os m se afastam umas das outras.

A Figura 19 apresenta o perfil da esperança de Monte Carlo para o número de componentes. Podemos perceber um padrão para o tamanho relativo esperado da maior componente. Quanto maior p, menor o número esperado de componentes. Além disso, quanto maior o m, menor o número esperado de componentes. Por outro lado, quanto maior θ, maior o número de componentes. Isto significa que, aumentar m aumenta as probabilidades do grafo ser conexo, ao passo que concentrar probabilidade em um grupo específico de características reduz esta probabilidade. Analisando em conjunto com as observações levantadas para a transitividade, podemos perceber que considerando p fixo, o ganho em transitividade é compensado por perda de conectividade, e o nível desta troca é influenciado por m e θ. Para $\gamma = 1$, observamos uma função convexa decrescente, mas para $\gamma > 1$, observamos transições de fase, cujo início da fase de decrescimento rápido começa em p mais alto à medida que γ cresce. Este fase de decrescimento rápido é menor à medida em que θ cresce, se inicia mais tarde para m grande quando θ é pequeno e mais cedo para m grande quando θ é grande. Podemos observar também que as curvas se afastam umas das outras à medida em que θ cresce.

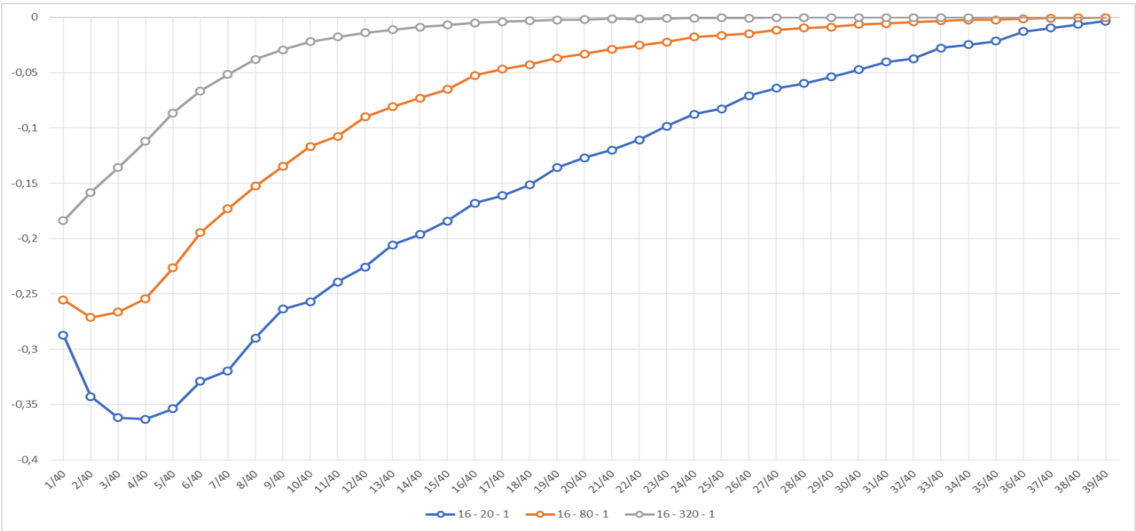
A Figura 20 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para o número de componentes do modelo de afinidade cardinal e o da esperança de Monte Carlo para o tamanho relativo da maior componente para o modelo de arestas independentes. Podemos perceber que o modelo de afinidade cardinal é menos conexo do que o modelo de arestas independentes. Esta diferença é elevada para θ grande, especialmente para m pequeno, concordando com o resultado de Fill, Scheinerman e Singer-Cohen (2000), quanto maior o m, menor a diferença quando $\theta = 0$. Percebemos o decaimento da diferença à medida que p cresce. Como dito anteriormente, combinando esta interpretação com a da transitividade, para um p fixo, a disposição não independente das arestas aumenta a transitividade sacrificando a conectividade do gráfico. Além disso, concentrar probabilidade em um grupo específico de



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

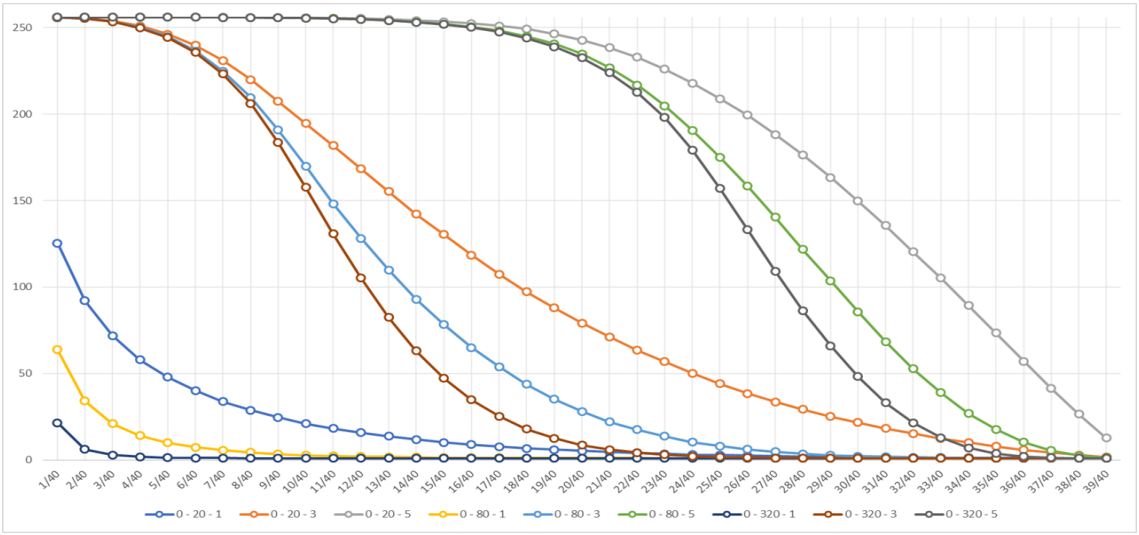


(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

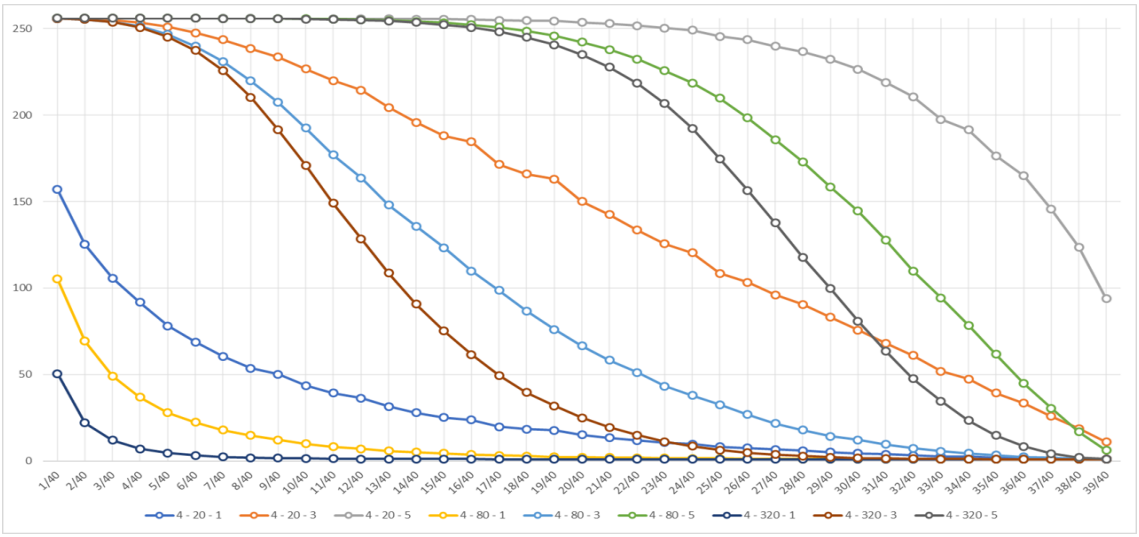


(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

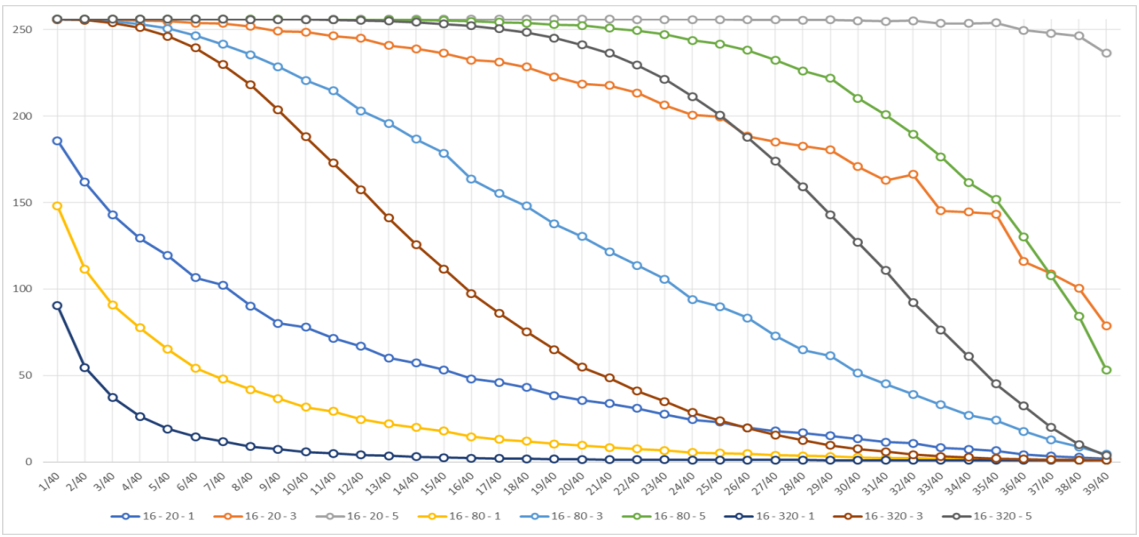
Figura 18 – Perfil da diferença entre as esperanças da proximidade do modelo de afinidade cardinal e o modelo de aresta independentes



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

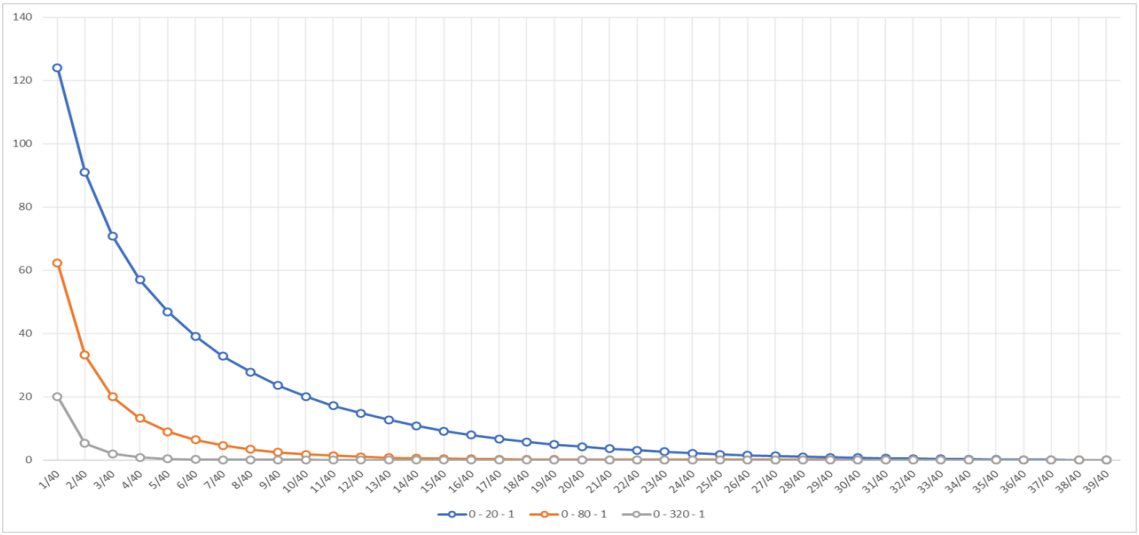


(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

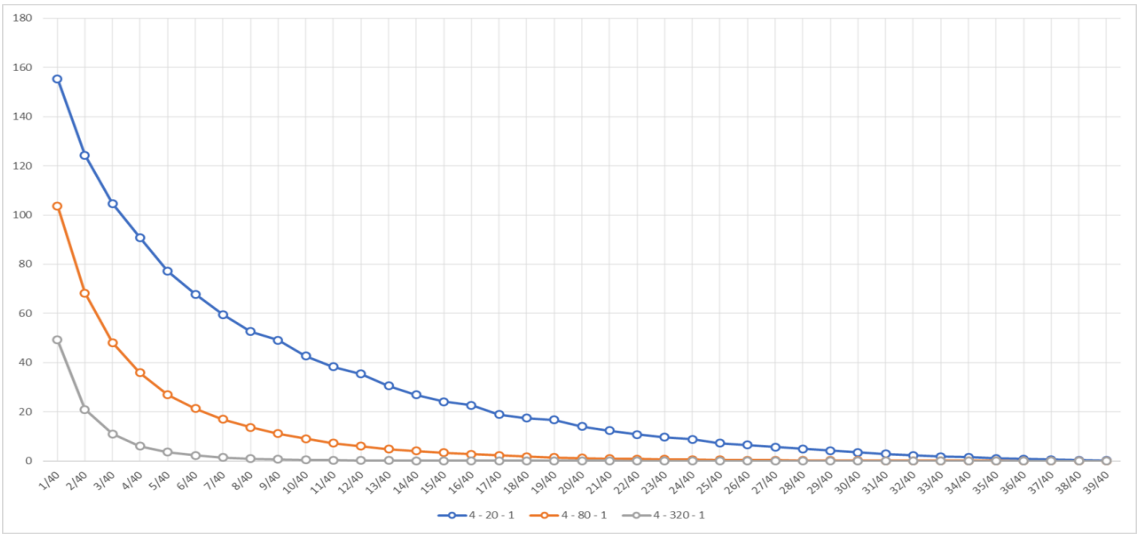


(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

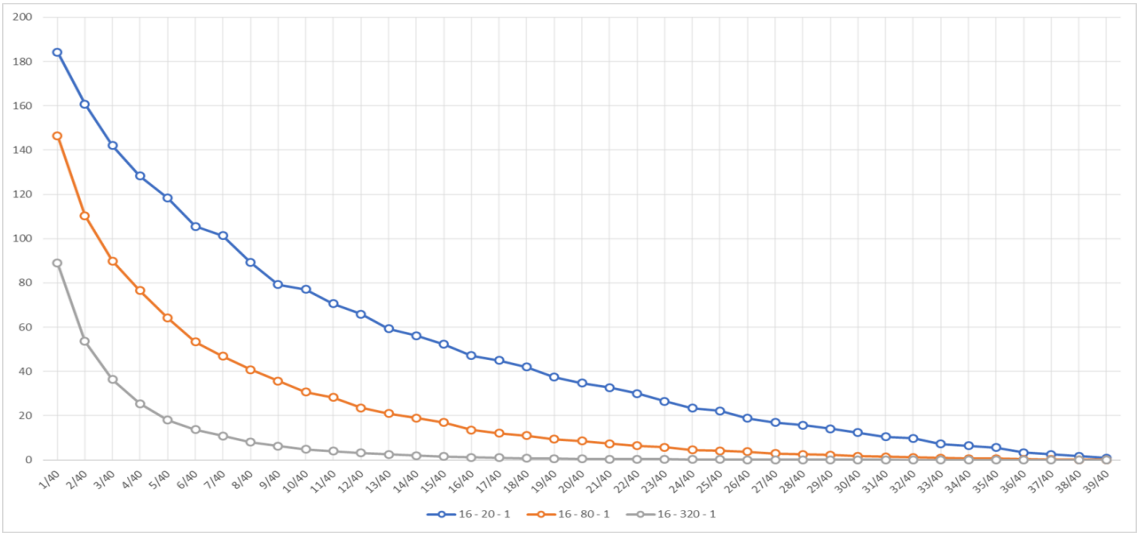
Figura 19 – Perfil da esperança de Monte Carlo para o número de componentes da função afinidade cardinal



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

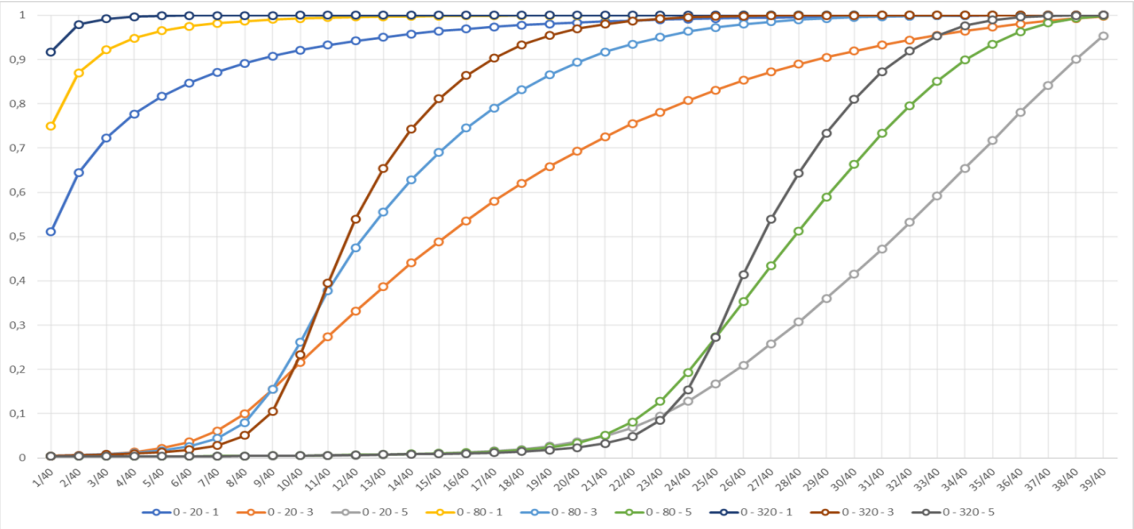
Figura 20 – Perfil da diferença entre as esperanças do número de componentes do modelo de afinidade cardinal e o modelo de aresta independente

características faz com que se formem grupos baseados nestas características, mas muito menos conectados através de outras características, fazendo com que a conectividade decaia ainda mais.

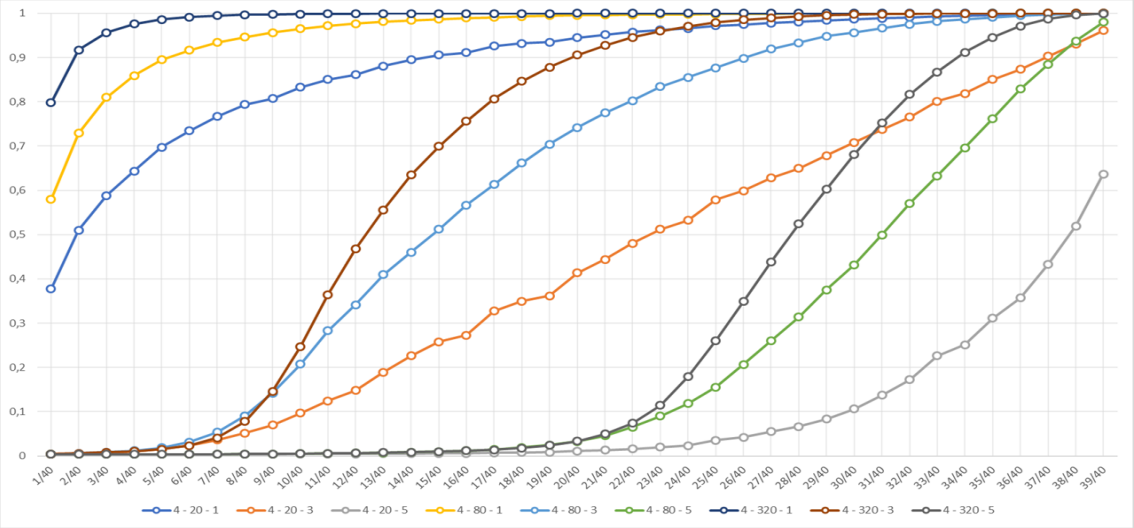
A Figura 21 apresenta o perfil da esperança de Monte Carlo para o tamanho relativo da maior componente. Podemos perceber um relacionamento entre o número de componentes e o tamanho relativo da maior componente. O comportamento dos perfis observados nas duas medidas topológicas parecem espelhados, sendo muito similares na forma, mas em sentido contrário. Quanto maior p, maior o tamanho relativo esperado. Além disso, quanto maior o m, maior o tamanho relativo esperado. Por outro lado, quanto maior θ, menor o tamanho relativo esperado. Isto significa que, aumentar m aumenta as probabilidades do grafo ser conexo, ao passo que concentrar probabilidade em um grupo específico de características reduz esta probabilidade. Analisando em conjunto com as observações levantadas para a transitividade, podemos perceber que considerando p fixo, o ganho em transitividade é compensado por perda de conectividade, e o nível desta troca é influenciado por m e θ. Para $\gamma = 1$, observamos uma função concava crescente, mas para $\gamma > 1$, observamos transições de fase, cujo início da fase de crescimento rápido começa em p mais alto à medida que γ cresce. Este fase de crescimento rápido é menor à medida em que θ cresce, se inicia mais tarde para m grande quando θ é pequeno e mais cedo para m grande quando θ é grande. Podemos observar também que as curvas se afastam umas das outras à medida em que θ cresce.

A Figura 22 apresenta o perfil da diferença entre os perfis da esperança de Monte Carlo para o tamanho relativo da maior componente do modelo de afinidade cardinal e o da esperança de Monte Carlo para o tamanho relativo da maior componente para o modelo de arestas independentes. Assim como observado para o número de componentes, podemos perceber que o modelo de afinidade cardinal é menos conexo do que o modelo de arestas independentes. Esta diferença é elevada para θ grande, especialmente para m pequeno e, concordando com o resultado de Fill, Scheinerman e Singer-Cohen (2000), quanto maior o m, menor a diferença quando $\theta = 0$. Percebemos o decaimento da diferença à medida que p cresce. Como dito anteriormente, combinando esta interpretação com a da transitividade, para um p fixo, a disposição não independente das arestas aumenta a transitividade sacrificando a conectividade do gráfico. Além disso, concentrar probabilidade em um grupo específico de características faz com que se formem grupos baseados nestas características, mas muito menos conectados através de outras características, fazendo com que a conectividade decaia ainda mais.

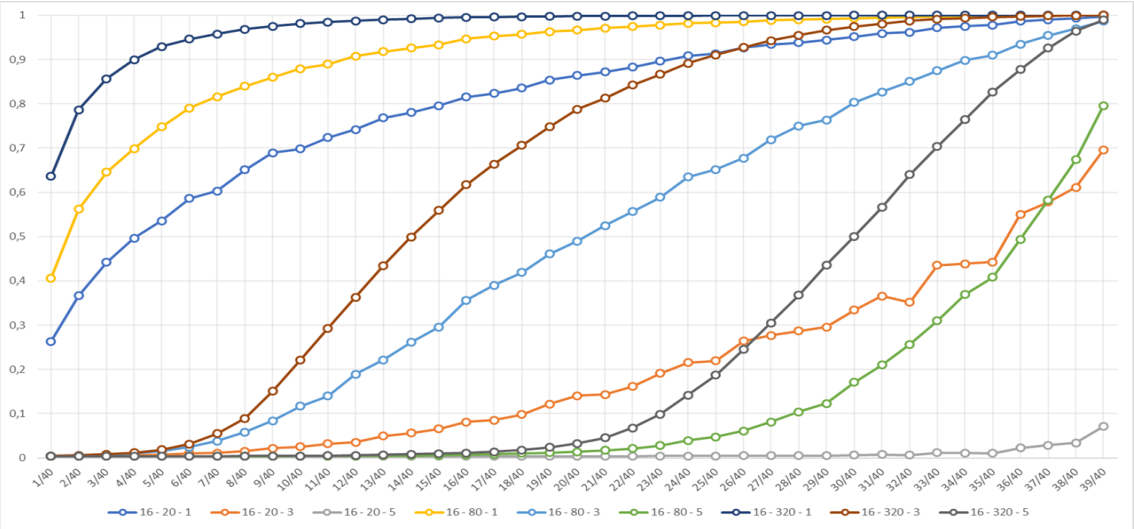
A Figura 23 apresenta o perfil da esperança de Monte Carlo para a média das componentes excluindo-se a maior componente da função afinidade cardinal. Observando a Figura 23, nós podemos observar que ao excluir a maior componente e nos concentrarmos nas componentes restantes, podemos perceber que quando $\gamma = 1$, observamos que as para os valores pequenos de p, observamos que a esperança da média é aproximadamente 1. À medida em que p cresce, observamos o decrescimento da esperança em direção à zero. A velocidade em que o perfil alcança zero parece depender de m e θ: quanto maior m, mais rápido o a média vai para zero.



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

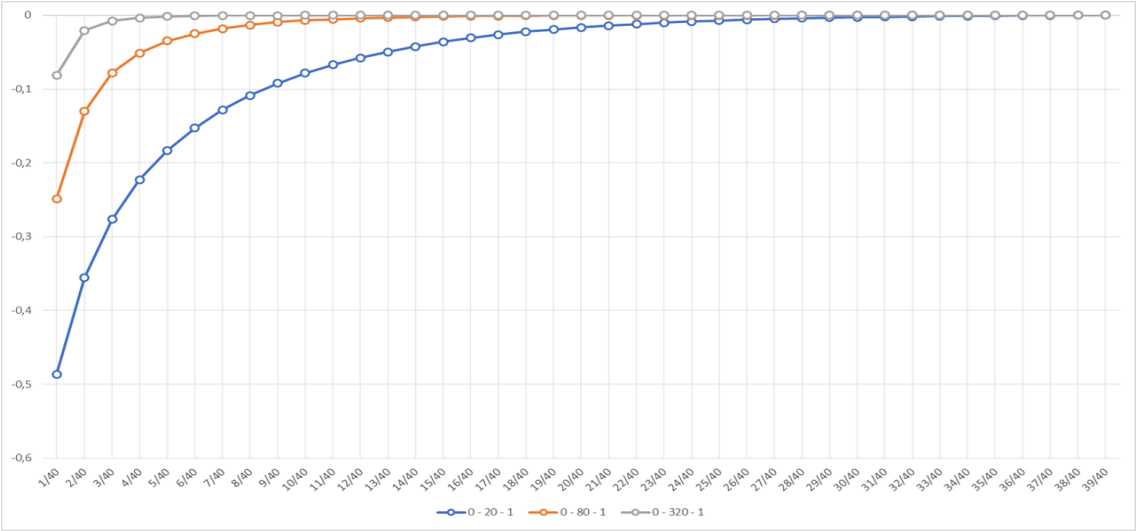


(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

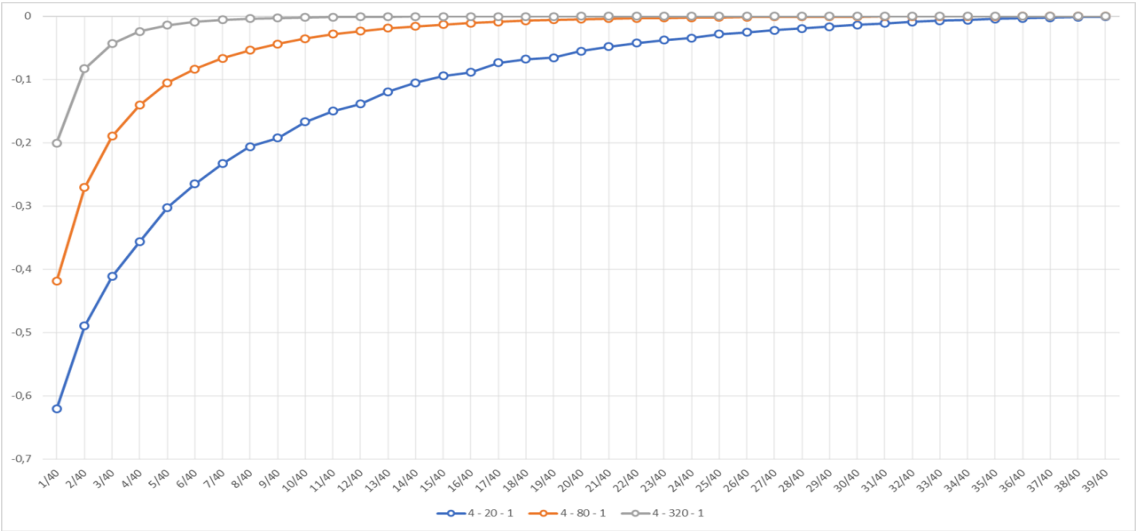


(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

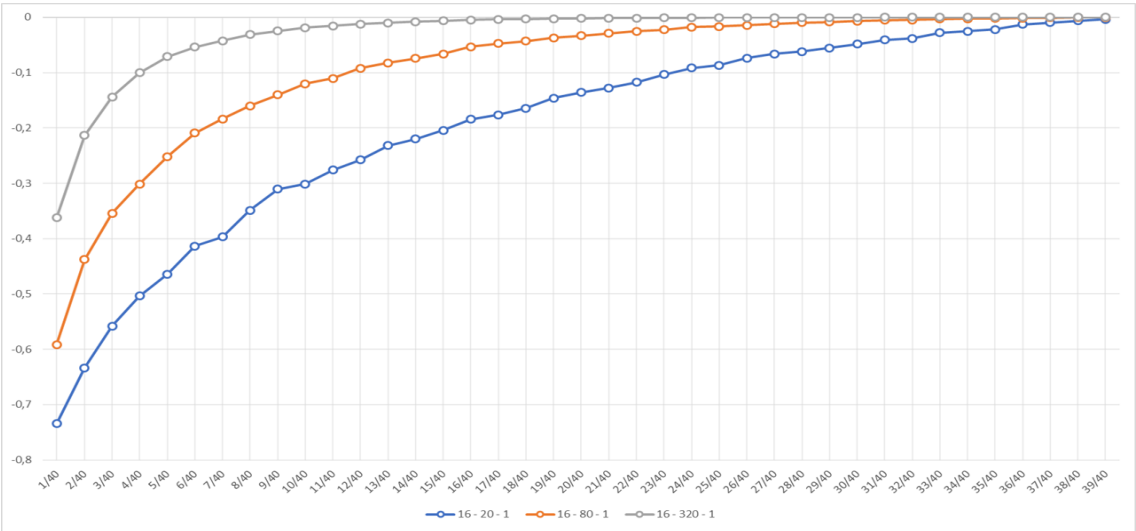
Figura 21 – Perfil da esperança de Monte Carlo para o tamanho relativo da maior componente da função afinidade cardinal



(a) Cenários: $\theta = 0$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

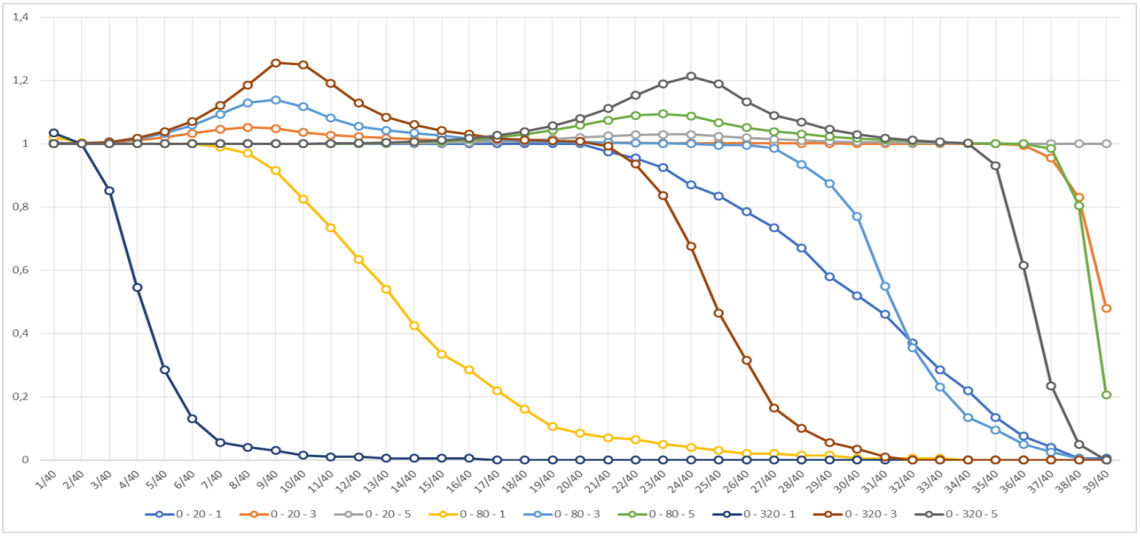


(b) Cenários: $\theta = 4$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

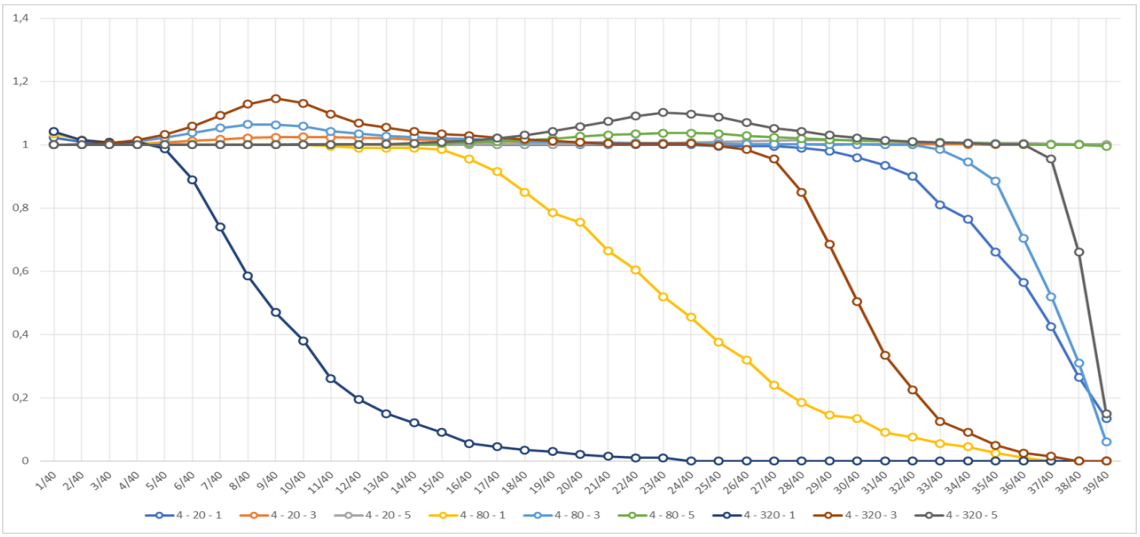


(c) Cenários: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

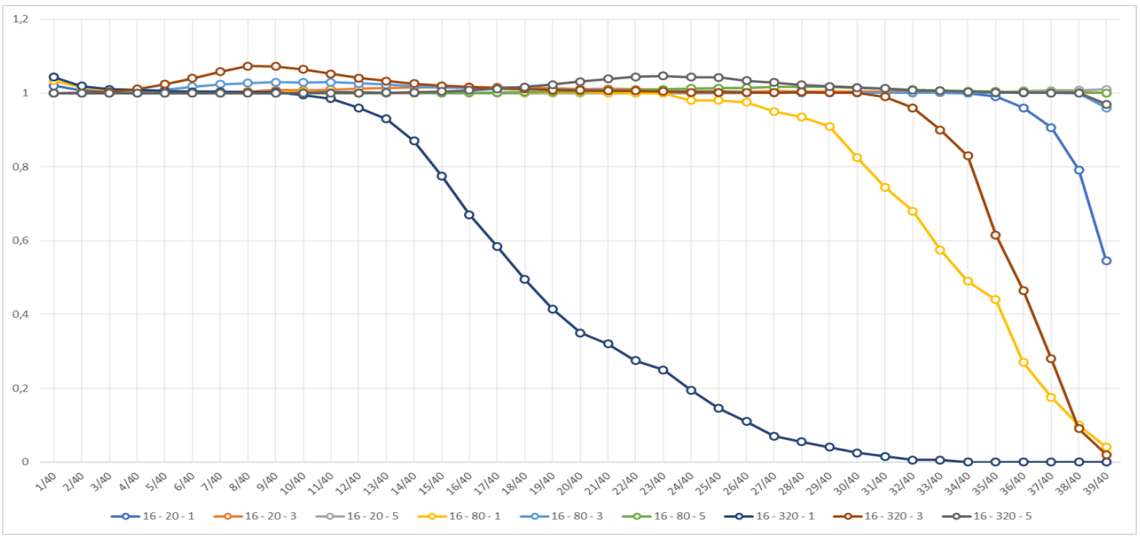
Figura 22 – Perfil da diferença entre as esperanças do tamanho da maior relativo da maior componente do modelo de afinidade cardinal e o modelo de aresta independentes



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



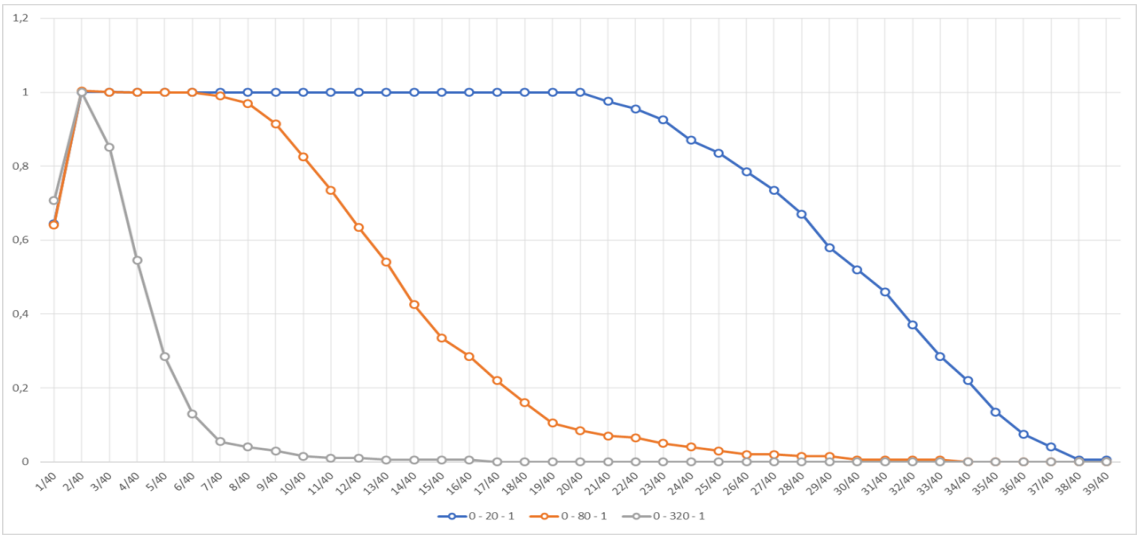
(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

Figura 23 – Perfil da esperança de Monte Carlo para a média das componentes excluindo-se a maior componente da função afinidade cardinal

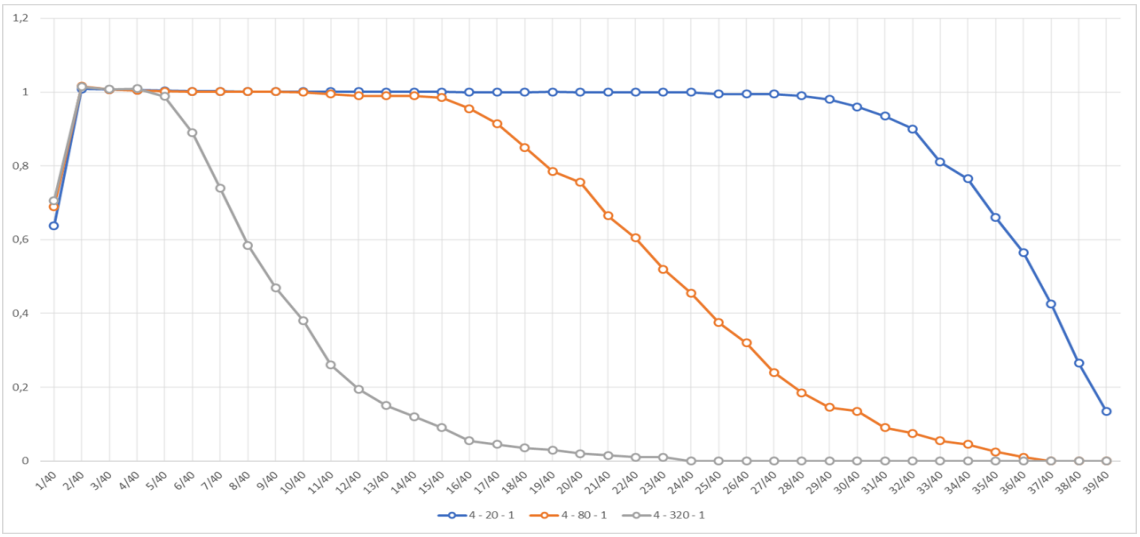
Por outro lado, quanto maior θ, mais devagar a média vai para zero.

Para $\gamma > 1$, percebemos que existe um subintervalo de probabilidade de conexão p em que se espera que, em cada um das componentes restantes, exista mais de um vértice. Mais precisamente, para $\theta = 0$ e $m = 320$ quando nos valores iniciais de p, observamos a média do tamanho das componentes próximo de 1. Quanto maior γ, maior é o valor de p para o início deste subintervalo de crescimento seguido de decrescimento. A seguir, à medida que em que p cresce, observamos um crescimento até um ponto maximante seguido de decrescimento voltando para a esperança próxima de 1. Além disso, aumentar θ e reduzir m parecem aumentar o valor do maximante do subintervalo. Este crescimento seguido de decrescimento apresenta forma de sino, similar ao formato da função de densidade da distribuição t-student. A partir daí, observamos decrescimento da média em direção à zero. Este padrão leva a crer que é esperado que existam, no mínimo, pares de indivíduos que são conectados por um grupo específico de características que é responsável por conectar exclusivamente os mesmos e que a existência destes grupos de características é aumentada quando θ se aproxima de zero e m cresce. Por outro lado, aumentar θ e reduzir m concentrar muita probabilidade a um grupo pequeno de características, implicando na redução da probabilidade de observarmos conexões através de palavras que não às pertencentes a este grupo, de forma que não estar na maior componente implica em maior probabilidade de estar isolado.

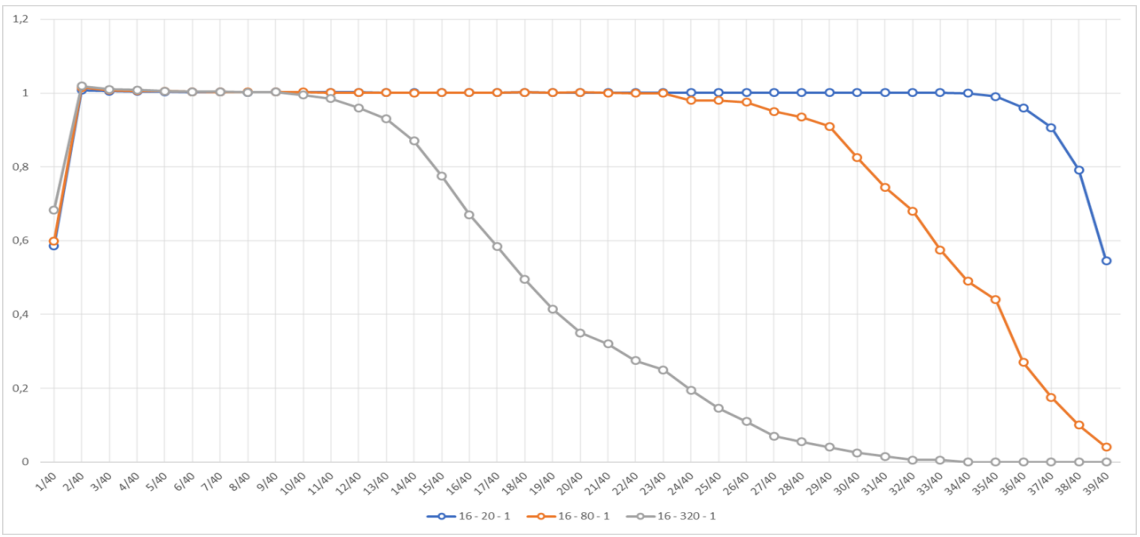
A Figura 24 apresenta o perfil da diferença entre os perfis da diferença entre a esperança de Monte Carlo para a média das componentes excluindo-se a maior componente do modelo de afinidade cardinal e o da esperança de Monte Carlo para a média das componentes excluindo-se a maior componente do modelo de arestas independentes. Assim como observado em outras medidas topológicas ao longo deste trabalho, nós observamos que a esperança da média das componentes excluindo-se a maior componente é maior no modelo de redes de afinidade do que no modelo de arestas independentes, o que leva a conclusão de que o modelo de redes de afinidade é mais desconexo do que o modelo de arestas independentes. Para p muito grandes e muito pequenos, observamos uma tendência aproximação desta distância de 0. Além disso, a diferença é elevada para θ grande, especialmente para m pequeno e, concordando com o resultado de Fill, Scheinerman e Singer-Cohen (2000), quanto maior o m, menor a diferença quando $\theta = 0$.



(a) Cenários: $\theta = 0$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(b) Cenários: $\theta = 4$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$



(c) Cenários: $\theta = 16$, $\gamma \in \{1,3,5\}$ e $m \in \{20,80,320\}$

Figura 24 – Perfil da diferença entre as esperanças da média das componentes excluindo-se a maior componente do modelo de afinidade cardinal e o modelo de aresta independente

6 CONCLUSÕES

Durante este trabalho, apresentamos o modelo de redes de afinidades, construindo uma estrutura sólida baseada na estrutura proposta por Singer (1995), introduzindo novos elementos tais como o como a função afinidade, a violação da suposição de características regidas por distribuições independentes e identicamente distribuídas e os pontos de corte. Apresentamos exemplos de como gerar tais vetores de probabilidades de escolha, bem como gerar os vetores de funções de probabilidades dependentes ou com postos.

O trabalho desenvolvido foi capaz de generalizar o modelo de grafos aleatórios de interseção, tanto no que se refere à função afinidade quanto aos vetores desbalanceados. Foi possível observar que a interação do tamanho de D com o nível de desbalanceamento do vetor de probabilidades de escolha surte efeitos opostos nas medidas topológicas do grafo, bem como na distância destas para um modelo de arestas independentes. Em geral, aumentar o número de características aproximou o modelo de afinidade cardinal de modelos de arestas independentes concordando com Fill, Scheinerman e Singer-Cohen (2000), enquanto o aumento do desbalanceamento do vetor de probabilidades o afastou do modelo de arestas independentes.

Outro resultado empírico observado muito interessante é o de que, fixado um p, os modelos de afinidade têm propensão à maior transitividade e coeficiente de clustering, concordando com Newman e Park (2003), mas o custo por este aumento é a redução no nível de conectividade do grafo. De acordo com o estudo, o grafo de afinidade tende a ser menos conexo do que grafos de arestas independentes. Além disso, concentrar probabilidade em um grupo específico de características aumenta a transitividade, mas torna o valor esperado da afinidade menor e o grafo menos conexo. Além disso, o modelo de afinidades gera graus e forças máximas mais altos, mas sendo p fixo, isso resulta em graus menores observado entre os menores graus da rede.

O estudo aqui desenvolvido pode ser ampliado para qualquer modelo de redes de afinidade. Um avanço interessante seria desenvolver um método para recuperar as probabilidades de escolha dado uma configuração e um modelo de afinidade. A maior dificuldade encontrada para tal foi explicitar a função de verossimilhança da matriz de adjacências. Apesar das tentativas por nós realizadas, não conseguimos expressar tal função de verossimilhança devido ao fato todas as arestas não são independentes e parecem ter uma estrutura de dependência complexa que se torna mais e mais complexa à medida que aumentamos o número de atores e o número de características envolvidos. Este seria um grande avanço que pode ser estudado e desenvolvido em trabalhos futuros. Um outro avanço que pode ser feito seria tentar encontrar resultados analíticos para estes modelos, ou encontrar alguma forma para modelá-los.

REFERÊNCIAS

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, The American Association for the Advancement of Science, v. 286, n. 5439, p. 509, 1999.

DURRETT, R. *Random graph dynamics*. [S.l.]: Cambridge university press Cambridge, 2007. v. 200.

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

FILL, J. A.; SCHEINERMAN, E. R.; SINGER-COHEN, K. B. Random intersection graphs when m= ω (n): an equivalence theorem relating the evolution of the g (n, m, p) and g (n, p) models. *Random Structures & Algorithms*, Wiley Online Library, v. 16, n. 2, p. 156–176, 2000.

GILBERT, E. N. Random graphs. *The Annals of Mathematical Statistics*, JSTOR, v. 30, n. 4, p. 1141–1144, 1959.

GUEDES, G. R. et al. Signifying zika: heterogeneity in the representations of the virus by history of infection. *Cadernos de saude publica*, SciELO Public Health, v. 34, p. e00003217, 2018.

JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, v. 37, p. 547–579, 1901.

JUNG, C. G. The association method. *American Journal of Psychology*, University of Illinois Press, v. 31, p. 219–269, 1910.

KAROŃSKI, M.; SCHEINERMAN, E. R.; SINGER-COHEN, K. B. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, Cambridge University Press, v. 8, n. 1-2, p. 131–159, 1999.

MATHEUS, R. F.; SILVA, A. B. d. O. Análise de redes sociais como método para a ciência da informação. *DataGramaZero-Revista de Ciencia da informacao*, IASI-Instituto de Adaptação e Inserção na Sociedade da Informação, v. 7, n. 2, 2006.

NEWMAN, M. E.; PARK, J. Why social networks are different from other types of networks. *Physical review E*, APS, v. 68, n. 3, p. 036122, 2003.

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.

SINGER, K. *Random intersection graphs*. Tese (Doutorado) — Department of Mathematical Sciences, The Johns Hopkins University, 1995.

SMIRNOV, N.; SMIRNOV, N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. 1939.

STARK, D. The vertex degree distribution of random intersection graphs. *Random Structures & Algorithms*, Wiley Online Library, v. 24, n. 3, p. 249–258, 2004.

VERGÉS, P. L'évocation de l'agent: une méthode pour la dédinition du nouau central d'une representation. *Bulletin de Psychologie*, v. 45, n. 405, p. 203–209, 1992.