A probabilistic model for networks generated by actors' characteristics

Abstract

This work presents the affinity network model for random graphs, consisting of a broad family of random graph models depending on some parameters. In this model, we suppose that each individual randomly chooses a set of characteristics that represent him according to a certain probability measure. The connections between two individuals depend on their shared characteristics and are valued according to a function that measures what we call *affinity* in the network. According to the choice of this function, the network's density can vary from sparse to complete graph, causing the model to be very flexible, which makes it suitable to fit with real networks. To illustrate the behavior of the affinity network model, we present a Monte Carlo simulation study. We tune the model's generating parameters, analyze its topological measurements, and compare them with equivalent graphs with edges occurring independently, disregarding actors' characteristics.

Keywords: Random graph models, Social networks, Inference for random graphs

1 Introduction

In this work, we present a random graph model, which we call the *affinity network* model, that considers the attributes of the vertices responsible for making a connection between them. We define a function that measures what we call *affinity* between two vertices, or actors in the network, based on their characteristics and use this function to build the network. The idea is that vertices that share characteristics are alike and therefore are connected in the graph, whose strength will depend on the shape of this *affinity* function that influences the topological structure of the network. We believe that flexibility makes our model suitable to fit with real networks, which is supported by our simulated study of some instances of the model.

As examples of important random graph models based only on the connection probabilities, disregarding vertex features, we can cite [7], [9], and [3]. By ignoring the characteristics of the vertices, these models distance themselves from real network data. Many real networks show properties not covered by these models, such as clustering or transitivity, the propensity for two neighbors of the same vertex to be neighbors, forming a triangle of connections in the network. We refer the reader to [14] for a detailed explanation. Besides, there are several generating

models for Dynamic Weighted Complex Networks. To cite a few, we refer to [2], [16], [12], and in [18] it is presented a comprehensive survey on this topic.

The main inspiration for the affinity network model comes from Social Representations collected via the Free Word Association Technique (TFAW) based on psychoanalytic principles of non-cognitive associations. In this instrument, individuals are asked to freely write five words, or evocations, about a specific subject and then order these words according to their importance to them. The main goal of this technique is to find the central nucleus, that is, the words that would represent the most critical and widespread thoughts in the collective thinking of that society about a given social object. An overview of this subject is presented in [17].

In the literature on the subject, the quadrants of [22], which values words by frequency and the mean order of evocation, are widely used. However, [6] presents an alternative methodology for the construction of a cognitive network through shared evocations and values them according to their order and frequency of evocation through a coefficient we call *cognitive affinity* . In this way, the cognitive network can capture people's collective thinking about a given theme without establishing cutoff points for the frequency and the average order of evocation, as in the analysis of quadrants. The affinity network model extends the family of networks presented in this paper by allowing a broader class of functions linking the vertices.

The affinity network model proposed here can also be seen as a generalization of the Random Intersection Graph model introduced by [19], which also considers that the source of uncertainty is in the vertex. The affinity network model inherits some interesting Random Intersection Graphs' proven properties. The studies that analyzed the Random Intersection Graphs model were particularly interested in finding topological properties. For example, the article [13] presents the threshold probabilities for which a given induced subgraph would be present with a high chance in the intersection graph, and [21] presents a study of its degree distribution.

This paper primarily aims to achieve two key objectives. First, we introduce a novel network model that draws inspiration from the previously mentioned concept of Social Representations. This model broadens the scope of established models and provides remarkable versatility, making it suitable for a broad variety of scenarios. Second, we conduct a simulated study to support the model's capacity to encapsulate network properties typically observed in real-world networks, such as high clustering, small diameter, large maximum degree, etc.

This paper is organized as follows. In the first section, we recall the Erdős-Rényi model that we will use as a basis to make comparisons. Section 2 presents the Affinity Networks model, describing its components in detail. Section 3 contains some examples and usage of the Affinity Network model. In section 4, we present a simulated study of the model for various scenarios showing the behavior of the Affinity Network model concerning the main topological measures for graphs and other essential features. Section 5 presents the conclusions based on the simulations performed, leading us to claim that our model may fit real networks adequately.

1.1 Baseline model

Let $G(V, E)$ be a graph where V is a set of vertices and E is a set of edges connecting them. We consider the collection of all graph configurations, $\mathcal{G}$, with N vertices and a probability

distribution over it. The maximum possible edges for a graph $G \subset \mathcal{G}$ is $n = \binom{N}{2}$. In the model proposed in [7] (ER), a graph with M edges is chosen uniformly at random from $\mathcal{G}$, that is, the probability of choosing a graph in $\mathcal{G}$ that has exactly M edges, $G_{(N,M)}$, is

$$P(G_{(N,M)}) = \frac{\binom{n}{M}}{2^n}$$

A variant of this model was introduced by [9], where a graph is constructed by connecting labeled nodes randomly. Each edge is included in the graph with probability p, independently from every other edge, where the probability for generating each graph that has N nodes and M edges, $G_{(N,p)}$, is

$$P(G_{(N,p)}) = p^M(1-p)^{n-M}.$$

The ER model $G_{(N,p)}$ is the one more commonly used, mainly due to its simplicity and the ease of analysis allowed by the independence of the edges. Therefore, when we refer to an ER model, we refer to the $G_{(N,p)}$ throughout this text.

We have that the expected number of edges in $G(N,p)$ ER model is np and the degree distribution of any particular vertex has a binomial distribution:

$$P(\deg(v) = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k},$$

which converges to a $Poisson(\lambda)$ distribution for large N and $\lambda = Np$ constant.

The introductory article by Erdős and Rényi (1959) has more than 17000 citations in the literature up to now, which gives an idea of its importance to the field of random graph models. Since then, a lot of work has been dedicated to understanding this model and developing the area of random graphs. For an interesting review, we refer the reader to the book by Newman (2010).

However, the ER model does not fit well with data from real networks, in general, due to the complexity of the connections present in the real networks that it cannot capture. It represents an unrealistic situation where the connections between individuals are set independently and uniformly, disregarding all other information. Nevertheless, it is a valuable model used as a baseline for other models. It is interesting to note how much other models deviate from it and manage to capture essential features in real networks, such as small diameter, heavy-tailed degree distribution, high clustering, and transitivity.

2 Affinity network model: presentation

In this section, we formally introduce the affinity network model. The main ideas behind the affinity model are that vertices or actors might have intrinsic characteristics and that the intensity of connections between actors is a function of these characteristics. The intrinsic characteristics are encoded in the *vocabulary matrix*. At the same time, the mechanism controlling the intensity of connections is formalized into the *affinity function*. Thus, we will introduce these concepts in the following subsections to properly formalize our model.

2.1 Vocabulary matrix

We consider a set $\mathcal{D} = \{d_1, d_2, \ldots, d_{m-1}, d_m\}$ with $m \in \mathbb{N}$, which we call *Dictionary*, consisting of all possible characteristics associated to a subject, that can be, for example, words, labels, or features in a population. Suppose that each individual from this population randomly chooses, according to a certain probability measure, a subset of characteristics in $\mathcal{D}$. We denote by $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,m_i-1}, d_{i,m_i}\}$, with $m_i \leq m$ such that $D_i \subset D$, the set of characteristics chosen by the i-th individual and call it *vocabulary*.

For practical purposes, we associate each set D_i with a binary vector $U_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m-1}, u_{i,m}]$ such that

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

In this way, U_i keeps the same information brought by D_i. For a population of size N, we define the *vocabulary matrix* $U_{N\times m}$ as

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

In words, the i-th row of $U_{N\times m}$ stores all the relevant characteristics of the i-th actor/vertex. For example, let's assume a population with $N = 4$ individuals and $\mathcal{D} = \{d_1, d_2, d_3, d_4, d_5\}$, According to Equation 1, a possible U matrix, would look like

$$U_{4\times 5} = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix},$$

where $U_1 = [0, 1, 1, 0, 1]$ means that the first individual has chosen the characteristics d_2, d_3 and d_5.

If there is interest in knowing the order in which the characteristics were chosen, the matrix U can store additional information, preserving each D_i. To get U in this case, we write the order vector associated with D_i as $U_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m-1}, u_{i,m}]$ such that

$$u_{i,j} = \begin{cases} t, \text{ if } d_j \in D_i, \ d_j = d_{i,t} \\ 0, \text{ if } d_j \notin D_i \end{cases}$$

so that each non-zero element in U_i indicates the position of the respective characteristic in the choices of the i-th individual.

For example, the vector $U_1 = [0, 2, 1, 0, 3]$ reveals which characteristics out of the 5 in $\mathcal{D}$ were chosen by individual 1 and in which order. A possible U matrix for four individuals would be

$$U_{4\times 5} = \begin{bmatrix} 0 & 2 & 1 & 0 & 3 \\ 1 & 0 & 2 & 0 & 0 \\ 4 & 3 & 1 & 5 & 2 \\ 1 & 0 & 0 & 2 & 0 \end{bmatrix}.$$

2.2 The affinity function

As mentioned before, in our model, the connection between two vertices is according to a deterministic function, which measures what we call *affinity* between individuals regarding their vocabulary. The idea behind this function is that we would like to say that individuals who have chosen similar characteristics are very likely to be connected and to hold a strong connection. On the other hand, if they share none of a few similarities, they are unlikely to be connected or they are connected but by a weak connection.

The function measuring and determining the intensity of the connections between actors/vertices is what we call the *affinity function*. Mathematically, an *affinity function* is defined as a symmetric function, f, such that

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

We present some interesting examples in the following.

1. **Binary affinity function**

   The binary affinity function is non-null whenever two individuals, i e j, share at least one word in their choice sets, that is

   $$f(U_i, U_j) = 1 \Leftrightarrow D_i \cap D_j \neq \emptyset. \tag{B}$$

2. **Cardinal affinity function**

   The cardinal affinity function measures how many characteristics in common were chosen by two individuals,

   $$f(U_i, U_j) = |D_i \cap D_j|. \tag{C}$$

3. **Jaccard affinity function**

   In general, concordance coefficients can be seen as affinity functions. The agreement coefficient described in [11] measures which proportion of the characteristics chosen by two individuals is shared by them, that is

   $$f(U_i, U_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}. \tag{J}$$

4. **Cognitive affinity coefficient**

   In [6], the authors present an affinity coefficient to calculate the level of affinity between two individuals. Considering $|\mathcal{D}| = M$, $m = max(|U_i|, |U_j|)$, where u_{ij} is the order of the characteristic w_j for the individual i, they propose

$$f(U_i, U_j) = \sum_{l=1}^{M} [2 \cdot (m+1) - (u_{i,l} + u_{j,l})) \cdot (m - |u_{i,l} - u_{j,l}|) \cdot w \cdot I_{(\min\{u_{i,l}, u_{j,l}\} > 0)}]. \tag{CAE}$$

   where I is an indicator function, $I = 1$ if $(min\{u_{i,l}, u_{j,l}\} > 0)$, and 0, otherwise. Furthermore, the coefficient w is a normalization factor that makes the affinity coefficient belong to the [0.1] interval.

   The cognitive affinity coefficient depends on the order and distance between characteristics two individuals choose. We will discuss this choice for the affinity function in detail in Section 3.3.

2.3 The Affinity Network

We call *affinity network* the graph obtained linking the n individuals using the affinity function. Two individuals, i and j, are connected if $f(U_i, U_j) \geq \gamma$, for some value of γ, which is a tuning parameter. The greater the γ, the less dense the affinity network. High values of γ can be used where only edges representing a powerful connection are allowed.

We denote the *affinity network* by $G(\lambda)$, $\lambda = (n, m, \mu, f, \gamma)$, where n is the number of vertices, m is the number of characteristics in $\mathcal{D}$, $\mu = \{\mu_{\mathbf{j}}\}_{\mathbf{1 \leq j \leq m}}$ is the probability distribution over $\mathcal{D}$, f is the affinity function and γ is the cutoff. We notice that $\mathbf{U}$ carries all the uncertainty about $G(\lambda)$ since λ is fixed. We set $f(U_i, U_i) = 0$ to avoid loops in $G(\lambda)$.

More formally, given $\lambda = (n, m, \mu, f, \gamma)$, $G(\lambda)$ is the weighted graph whose vertex set is $V = \{1, 2, \ldots, n\}$ and edge set, E, is

$$E := \{\{i, j\} \,:\, f\,(U_i, U_j) \geq \gamma\} \tag{3}$$

and the weight or intensity of the connection between i and j is given by $f\,(U_i, U_j)$.

Figure 1 below presents an example of an affinity network with 15 actors, $m = 10$, and a cardinal affinity function defined at (C) where the cutoff, γ, changes from 2 to 5. We use $\gamma = 1$, for the binary affinity function.

3 Examples and Usage Cases of The Affinity Networks

To illustrate the model's versatility, we provide several examples and discuss some use cases of the affinity network model in this section. Notice that modeling a certain phenomenon or situation with the affinity network involves choosing the distribution μ's, the proper affinity

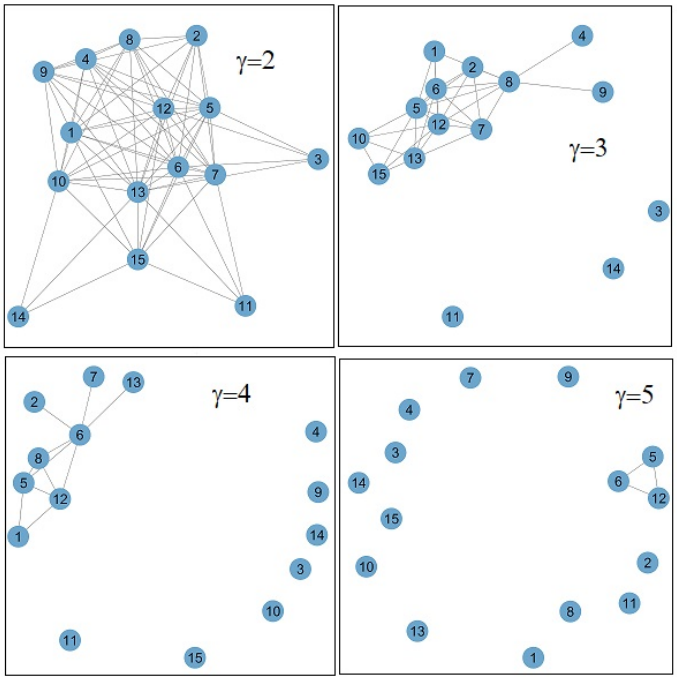


Figure 1: Affinity network with cardinal function, $m = 10$, and cutoff $\gamma = \{2, 3, 4, 5\}$.

function, and the cutoff parameter γ. The choice of μ involves assumptions on how characteristics are assigned to the actors. Whereas the affinity function together with γ drives not only the density of connections in the network but also their intensity/weights.

3.1 A particular important case of the affinity network model

The Random Intersection Graphs (RIG) model introduced by [19] is a particular case of the affinity network model considering the $U_{i,j}$ independent and identically distributed and binary affinity function with cut point $0 \leq \gamma \leq 1$.

There are several interesting works in the literature on RIG models. To cite a few, we refer the reader to [13] , [10], [5] , [4], [1] . These studies analyze topological features, phase transition, and evolution of subgraphs' appearance as the probability of choosing an element grows. An interesting result presented by [8] relates the order of a RIG, n, with the size of the characteristics set, m, such that $m = n^{\alpha}$. Considering $\hat{p}$ as the asymptotic probability of a particular edge, it is shown that is necessary $\alpha > 6$ for a RIG graph to behave like an Erdős-Rényi model. This result is, therefore, valid for affinity network models when $U_{ij} \sim Bernoulli(\delta)$ and the affinity function is binary.

This result means that the dictionary size must be much larger than the number of individuals for the binary affinity network graph to have the same behavior as the ER. For example, for a $n = 10$, a dictionary with more than 1 million characteristics would be needed for a RIG to

behave like an ER. Since the size of $\mathcal{D}$ in real cases is limited, it would be highly unlikely that affinity networks representing social relations would behave as ER graphs.

3.2 Modeling Perceptions on a Topic

A natural application of the affinity model concerns modeling how people perceive a given topic. In this case, the dictionary $\mathcal{D}$ might be a set of adjectives. Then, U_i becomes a list of adjectives used by the i-th individual to describe the topic T. Moreover, we may connect individuals if they have shared at least two adjectives to describe T. In the context of affinity networks, our affinity function is the cardinal function (C) with cut point $\gamma = 2$.

Assumptions on how individuals select their adjectives are encoded on the choices for distributions μ. Additionally, correlations might be added so that the description given by i affects the description given by j. Formally, one might add some correlation between U_i and U_j to accommodate the fact that individuals might influence each other's opinions.

Finally, in this situation, identifying connected components in $G(\lambda)$ leads to interesting insights about the general perception of T. If there are only two connected components in $G(\lambda)$, this indicates the presence of polarization.

3.3 Collective Thinking

A particular case of an affinity network appears in [6], serving as an empirical motivation for us. This work proposes an alternative methodology for constructing a cognitive network through shared evocations by individuals and values them according to their order and frequency of evocation. The authors propose a new function, *cognitive coefficient*, that maps individual cognitive links within a graph structure. This function transforms the data generated by the words chosen for an individual regarding a specific subject into an appropriate relational object for analyzing cognitive networks. The methodology was applied to novel data on evocations about river floods, allowing to find communities inside the network according to their thinking about this subject, identify the most active individuals inside each one, and, therefore, explicit their collective thinking.

3.4 Anti-affinity Network and Political Polarization

The degree of freedom one has in choosing the affinity function leads to other interesting examples. For the proper choice of f, the affinity network can capture not the affinity among the individuals but how different they are. More concretely, consider the affinity function given by

$$f(U_i, U_j) = 1 - \frac{|D_i \cap D_j|}{|D_i \cup D_j|}. \tag{4}$$

The above function gives more weight to those pair of individuals whose set of characteristics have fewer elements in common. The function reaches its maximum whenever i and j do not share any characteristics. And it reaches its minimum when i and j have the exact same set of

characteristics. This way, the network generated by this choice of f connects actors when they are different. This might be seen as an *anti-affinity network*.

A concrete situation in which one might be interested in the anti-affinity network is measuring political polarization. One should expect a graph with two distinct components in a polarized set of individuals. But one interesting question one may raise could be: *how far the individuals on each side of the political spectrum are?* This can be measured by analyzing the weight of the connections on the anti-affinity network. For instance, if, on average, $f(U_i, U_j)$ is close to one, the population is not only polarized, but the two groups are quite distant in the political spectrum.

4 Simulated study of the Affinity Network with cardinal affinity

In this section, we present the second main goal of this paper: a simulated study of the affinity network model. Our goal here is to find a way to create simulations that allow us to understand the behavior of the affinity network model, taking into account several different scenarios. We do this study for a particular affinity function: the cardinal function. Recall that the cardinal affinity is defined as

$$f(U_i, U_j) = |D_i \cap D_j|,$$

that is, f counts the number of characteristics two actors share. Thus, throughout this entire section, the affinity function will be set as the cardinal function.

We point out that the cardinal function is the most natural choice for the affinity function when one desires to connect actors by their affinity since it counts the number of shared characteristics. Also, observe that the cardinal affinity function is closely related to the Jaccard index defined at (J) since the Jaccard index is the cardinal function normalized by the size of the union to make it an index between 0 and 1.

4.1 Generating vocabulary matrix U

Recall that all the uncertainty of the model lies in the vocabulary matrix U. That is, given f and γ; the graph is completely determined by each realization of U. Thus, to simulate a realization of the Affinity Network with the cardinal affinity and some cut point γ, it is enough to know how to generate U.

With the above in mind, in this subsection, we demonstrate results that show how to generate the vocabulary matrices, U, by setting some parameters in simulations. Considering $U_{i,j}$ independent and identically distributed, and $P(U_{ij} = 1) = \delta$ for each characteristic j, we have that the probability of a connection between two individuals $P(f(U_i, U_k) > 0) = p$ is given by

$$p = 1 - (1 - \delta^2)^m, \tag{5}$$

which is the complementary probability that there is no characteristic in common between the set of characteristics attributed to the individuals. Hence, we can define the probability of choice

for each characteristic as a function of m and p by isolating δ in the equation 5.

$$\delta = \sqrt{1 - \sqrt[m]{1-p}}. \tag{6}$$

In the case where the probability of choosing each characteristic is not the same, let $\delta_j = P(U_{ij} = 1)$ be the probability with which the i-th individual chooses the j-th characteristic, then we have

$$p = 1 - \prod_{j=1}^{m}(1 - \delta_j^2). \tag{7}$$

and,

$$\prod_{j=1}^{m}(1 - \delta_j^2) = 1 - p \Rightarrow \sum_{j=1}^{m}\log(1 - \delta_j^2) = \log(1-p). \tag{8}$$

Now, let's consider a weight vector $\mathbf{z}$ such that, $0 < z_j < 1 \; \forall \; j = \{1, 2, \ldots, m\}$ and $\sum_{j=1}^{m} z_j = 1$. The objective of this weight vector is to allow the generation of random variables with different probabilities over the dictionary symbols in an automatic way for each simulation. Then

$$\sum_{j=1}^{m}\log(1 - \delta_j^2) = \log(1-p) \cdot \sum_{j=1}^{m} z_j = \sum_{j=1}^{m} z_j \cdot \log(1-p), \tag{9}$$

such that

$$\log(1 - \delta_j^2) = z_j \cdot \log(1-p) \Rightarrow 1 - \delta_j^2 = \exp\{z_j \cdot \log(1-p)\} \tag{10}$$

and,

$$\delta_j = \sqrt{1 - \exp\{z_j \cdot \log(1-p)\}}. \tag{11}$$

This means we can find the probability δ_j in terms of z_j, which can be generated following some auxiliary probability distribution and p. Thus, we can also compare distributions where the probability of choosing characteristics in the dictionary is not the same for all characteristics but still has fixed m and p, which allows us to make comparisons with other models, including the ER model. We emphasize that δ fixed implies $z_j = \frac{1}{m}, \; \forall j = \{1, 2, \ldots, m\}$.

4.2 Vocabulary distribution with cardinal affinity function

In this subsection, we show that the probability $P(f(U_i, U_j) = k)$, $k \in \{0, 1, \ldots, m\}$ considering a cardinal affinity function has a closed-form, which is one advantage of this affinity function, crucial to understanding the topological features of the affinity network model.

For each individual vocabulary $U_i = \{U_{i,1}, \ldots, U_{i,m}\}$, $\forall i \in \{1, \ldots, N\}$, we have that $U_{il} \sim Ber(\delta_l)$, the probability of any individual to choose the characteristic d_l,$1 \leq l \leq m$, such that $U_i \perp U_j$ for $i \neq j$.

Let $S \subset \mathcal{D}$, such that $S = \{d_1, d_2, \ldots, d_k\}$ be the set of the k common characteristics for each two individuals U_i and U_j, $\forall i \neq j \in \{1, \ldots, N\}$.

Then

$$P(f(U_i,U_j)=k)=\sum_{k}^{m}\sum_{S,|S|=k}(\prod_{d_l\in S}\delta_l^2\prod_{d_h\notin S}(1-\delta_h^2)),\ k=\{0,1,\ldots,m\}. \tag{12}$$

That is $f(U_i,U_j)$ has Poisson-binomial distribution with vector parameter $\Delta=\{\delta_1\ldots,\delta_m\}$. In particular, if $U_{il}\sim Ber(\delta),\ \forall i,l$, then $f(U_i,U_j)$ has a binomial distribution with parameter δ^2,

$$P(f(U_i,U_j)=k)=\binom{m}{k}\cdot(\delta^2)^k\cdot(1-\delta^2)^{m-k},\ k=\{0,1,\ldots,m\}.$$

And the mean and variance of $f(U_i,U_j)$ with cardinal affinity are given by

$$E(f(U_i,U_j))=\sum_{j=1}^{m}\delta_j^2\quad\text{and}\quad Var(f(U_i,U_j))=\sum_{l=1}^{m}\delta_l^2\cdot(1-\delta_l^2) \tag{13}$$

However, the expected number of edges on an affinity network depends on the cutoff parameter. In the case of cardinal affinity, a link between two individuals only happens if the number of common characteristics between these individuals, k, is greater than the established cutoff, γ. Let A_{ij} be the adjacency matrix associated with the affinity network, such that each entry $a_{ij}=1$ if there is an edge between vertices i and j. Then we have

$$a_{ij}=1\iff k\geq\gamma,\quad k,\gamma\in\mathbb{N}.$$

Henceforth,

$$\begin{aligned}P(a_{ij}=1)&=P(f(U_i,U_j)\geq\gamma)\\&=\sum_{l\geq\gamma}^{m}P(f(U_i,U_j)=l)\\&=1-\{\sum_{1\leq l<\gamma}\}P(f(U_i,U_j)=l)\\&=1-\{P(f(U_i,U_j)=1)+\ldots+P(f(U_i,U_j)=\gamma)\\&=p_\gamma\end{aligned}$$

That is, the bigger the γ, the smaller the connection probability, p_γ. Since the expectation $\mathbb{E}(a_{ij})=P(a_{ij}=1)$, the expected number of edges in the affinity network with cardinal function is Np_γ, decreasing when γ increases.

4.3 Simulating the Affinity Network model

This section presents a Monte Carlo simulation study of the topological measurements of the graph generated by an affinity network with a cardinal function. The cardinal function's image is the number of characteristics individuals share. We observe the behavior of each topological measure considered in the study for some values of the parameters.

In addition to comparing the topological measurements observed between the affinity function scenarios, we also compared the behavior of the affinity network with the behavior of graphs generated with the Erdős-Rényi model whose probability distribution preserves the distribution of the weights of the edges of the graph $G(\lambda)$.

4.3.1 Methodology

To generate affinity networks with n vertices and m characteristics is sufficient to know the distribution μ on $\mathcal{D}$, the affinity function f, and the cutoff γ.

Generating $U_{i,j}$ i.i.d is simple, since, having fixed a connection probability p and m the size of the set $\mathcal{D}$, we just apply the Equation 6 to obtain the value of δ such that the connection probability is p.

In the case where $U_{i,j}$ is independent, but not iid, finding the vector of choice probabilities $\Delta = \{\delta_1, \ldots, \delta_m\}$ associated with $\mathcal{D}$ requires generating the weight vector $\mathbf{z}$, in addition to setting a connection probability p.

We now present a method for generating the Δ vector that requires two entries: the number of characteristics m, and an unbalance parameter of the choice probability distribution, $\theta \geq 0$. We introduce this parameter to simulate scenarios with different probability distributions in the dictionary. The unbalance parameter controls how different the probabilities in Δ are so that $\theta = 0$ represents null unbalance, that is, $U_{i,j}$ are i.i.d. and, as θ increases, we put more probability of choice in a specific set of characteristics than in the others.

According to the definition of $\mathbf{z}$ in the subsection 4.1, let $\boldsymbol{Z} = \{z_1, \ldots, z_m\}$ be a random vector with $\sum_{j=1}^{m} z_j = 1$ and $0 \leq z_j \leq 1 \; \forall \; j$. We generate samples of $\boldsymbol{Z}$ using a Dirichlet distribution

$$\boldsymbol{Z} \sim Dirichlet(\boldsymbol{v}),$$

where $\boldsymbol{v} = \{v_1, \ldots, v_m\}$ is a parameter vector of size m with $v_j > 0 \; \forall \; j$.

We want to generate the parameter v from a density function whose parameter is a function that depends on θ, $\boldsymbol{v} \sim \eta(g(\theta))$, such that the η variance is directly proportional to θ, and raising θ would consequently increase its variance. A large variance in $\boldsymbol{v}$ implies an unbalanced $\mathbf{Z}$. Besides, it is interesting that η has not an upper bound. The gamma distribution meets this requirement. Then we consider the random variable V_j as

$$V_j \sim gamma\left(\frac{1}{\theta}, \frac{\theta}{\alpha}\right),$$

where α is an arbitrary constant, chosen to control the process. Therefore

$$E(V_j) = \frac{1}{\alpha} \text{ and } Var(V_j) = \frac{\theta}{\alpha^2}.$$

To avoid computational problems, we make $V_j \sim \eta + c$ with constant $c > 0$. We use this device to prevent v_j from getting too close to 0, causing the software to round the value down to 0, which is not a valid parameter for the Dirichlet distribution. The constant c guarantees that no matter how large $Var_X(\eta_j)$ is, η_j will be always bounded below by c.

The following algorithm summarizes the steps required to generate Δ with length m and an unbalanced level θ.

- Step 0: Verifify if $\theta = 0$. If yes , then take $\boldsymbol{z}$ such that $z_j = \frac{1}{m}\ \forall\ j$ and then go to step 4. Otherwise, go to step 1.
- Step 1: Generate t samples of $\boldsymbol{\eta} + \boldsymbol{c}$. Let us denote the b-th sample by $\boldsymbol{\eta}^{[b]}$.
- Step 2: For each sample $\boldsymbol{\eta}^{[b]}$, generate and order

$$\boldsymbol{Z}^{[b]} \sim Dirichlet(\boldsymbol{\eta}^{[b]}).$$

- Step 3: Calculate the ordered weights vector $\boldsymbol{z}$

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

- Step 4: Calculate the choice probability vector $\boldsymbol{\Delta}$

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

4.4 Results

This section presents the behavioral profiles of the affinity network with cardinal affinity function, which we call cardinal affinity network, and compares them with the respective ER model (ER) for scenarios resulting from combinations of its parameters.

The simulated study presented was carried out using Monte Carlo simulations using the software *R*. The graphics were generated dynamically in the software *Microsoft Excel*.

We also analyze the differences between the cardinal affinity network with a corresponding Erdős-Rényi model, which, although generated independently, retains the same weight distribution.

We start considering the probability distribution of the cardinal affinity network, calculating the expectation and variance of $f(U_i, U_j), \forall i, j \in \{1, \ldots, n\}$, and the degree distribution. After that, we present a series of comparisons, focusing on some topological measures of the graphs

to understand each of the chosen scenarios of the cardinal affinity model's behavior. The following topological measures were analyzed: maximum degree, transitivity, clustering coefficient, proximity, number of components, and the order of the largest component. We set the following parameters for the study:

- Graph order: $n = 256$;
- Number of characteristics: $m \in \{20, 80, 320\}$;
- Probability of connection: $p \in \left\{\frac{1}{40}, \frac{2}{40}, \ldots \frac{38}{40}, \frac{39}{40}\right\}$;
- Unbalance parameter: $\theta \in \{0, 4, 16\}$;
- Number of replicas: $M = 200$;
- Cutoff: $\gamma \in \{1, 2, \ldots, 6\}$;
- Constants: $c = 1 \cdot 10^{-12}$ and $\alpha = \frac{1}{100}$.

4.5 Mean and variance

Table 1 shows the expected values for the number of connections per each scenario and Table 2 presents the variances for $\gamma = 1$.

Table 1: Expected values of the numbers of connections in the cardinal affinity model.

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

As in the case of a RIG, the expectation of $f(U_i, U_j)$ in the cardinal affinity model increases as p grows, as does the variance. It is worth noting that, even in the scenario with 320 characteristics and $p = 0.9$, we have $E(f(U_i, U_j)) < 2.3$, that is, in those models where the choices are independent of each other, we do not expect many characteristics shared between two individuals.

Let's focus on the roles that θ and m play on expectation and variance. According to Table 1, the higher the probability of concentration in a specific group of characteristics, the lower the expected number of characteristics shared by two individuals. On the other hand, increasing the

Table 2: Variance of the number of connections in the cardinal affinity model.

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

number of characteristics raises the expected number of characteristics shared by individuals. Thus, about the expectation of $f(U_i, U_j)$, θ and, m appear to have opposite effects.

Observing Table 2, we can see that raising the value of θ implies reducing the variance of $f(U_i, U_j)$ while raising m implies increasing the variance of $f(U_i, U_j)$. We pay more attention to the following pairs of parameters in the subsequent analysis based on this information. : $(\theta, m) = \{(0, 320), (4, 80), (16, 20)\}$, since $(0, 320)$ and $(16, 20)$ represent extreme cases, while the pair (4.80) represents the average case in terms of expectation and variance of $f(U_i, U_j)$ in this simulated study.

4.6 Degree distribution

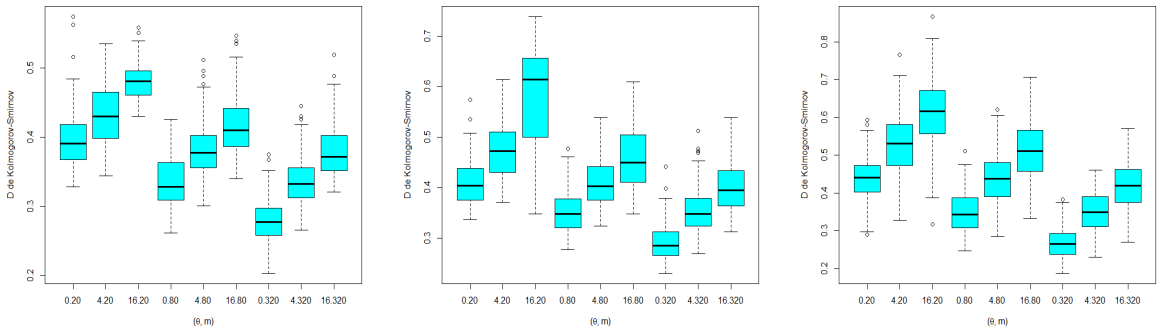
To assess the differences between the degree distributions of the two models, we used the Kolmogorov-Smirnov D statistic. Considering X and Y the vectors containing the degrees of the N vertices of the graph G and the degrees of the s vertices of the graph H. Then,

$$F_X(a) = \frac{1}{N} \cdot \sum_{i=1}^{n} I(X_i < a) \tag{14}$$

is the empirical cumulative probability function of X. The Kolmogorov- statistic, D, [20] is defined as

$$D_{X,Y} = \max_a |F_X(a) - F_Y(a)| \tag{15}$$

Figure 2 presents the behavior profile of the Kolmogorov-Smirnov D statistic resulting from the comparison between the degree distributions of the affinity function and the ER model whose weights were generated under the binomial-Poisson distribution with Δ as parameters. We can see that, as in the case of RIG , [8], increasing the value of m results in a decrease in the distance between the degree distribution of the cardinal affinity model and the degree distribution of the ER model. On the other hand, we can see that raising the value of θ increases that distance. We



(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) $\theta \in \{0, 4, 16\}$, $p = 0,9$, $m \in \{20, 80, 320\}$

Figure 2: Kolmogorov-Smirnov D statistic for the distance between the degree distributions of the cardinal affinity and ER models.

noticed that the scenario with (0.320) is, for all p displayed, is the closest from the ER model and the scenario with (16.20) is the furthest.

4.7 Network features of the Affinity Network model

We present the results obtained for the profiles of some features and topological measures of the affinity network with cardinal function: maximum degree, transitivity, clustering coefficient, proximity, number of components, and the size of the largest component. The profiles were made by increasing the values of the connection probabilities p. For each network feature analyzed, we present two figures: one with the behavior of the affinity network with cardinal function concerning the network feature, and another comparing it with the Erdős-Renyi (ER) model.

For the text's readability, we presented only the scenario results with $\theta = 16$, representing the highest probability concentration around a set of characteristics. The other scenarios with $\theta = 0$ and $\theta = 4$ can be downloaded from
$https://drive.google.com/drive/folders/1pQjD3hjnCLS8gk_pSgw1jSkdZl12d2z8$.

4.7.1 Maximum degree

Figure 3 presents the profiles of the maximum degrees of the cardinal affinity function for the scenarios where γ and m vary and $\theta = 16$.The connection probabilities $p, 1/40 \leq p \leq 40/40$ are on the X-axis. The maximum degrees are displayed on the Y-axis.

For θ large, the curves representing the larger loss in terms of expected values for the maximum degree are those with small m. This loss can be explained because small θ induces scenarios with $E(f(U_i, U_j))$ less than the others. There are fewer connections through characteristics with a low choice probability and fewer characteristics available.

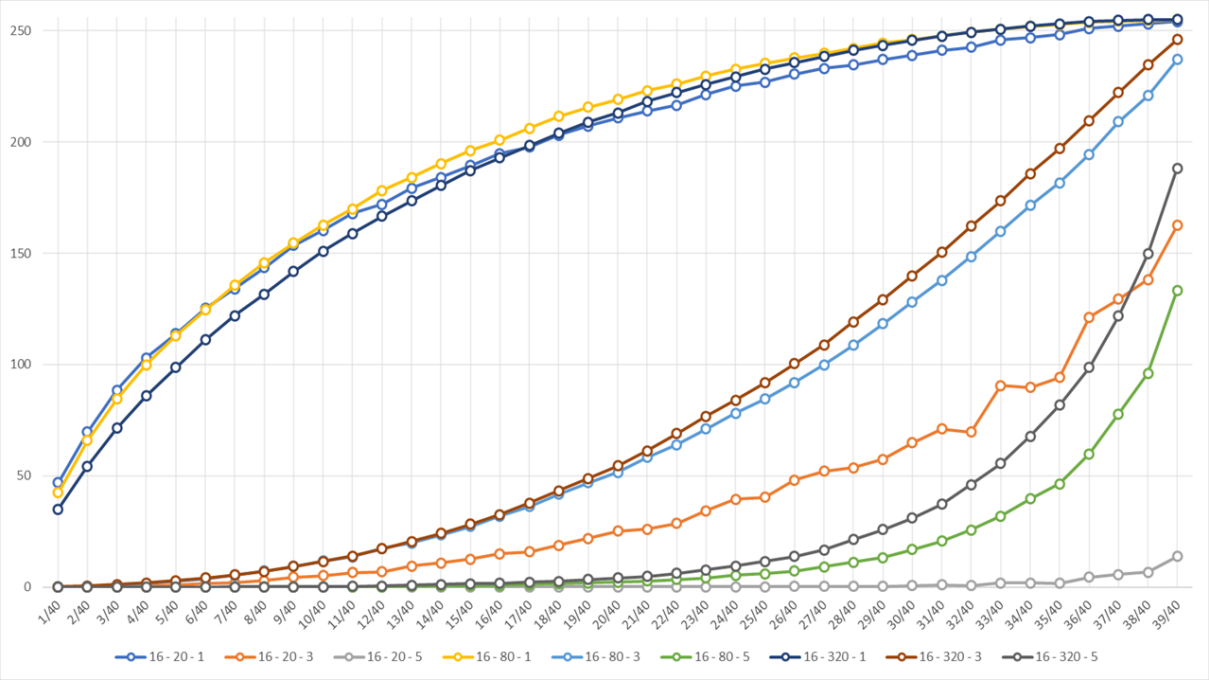


Figure 3: Monte Carlo expectation profile for the cardinal affinity function maximum degree. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ e $m \in \{20, 80, 320\}$

Figure 4 presents the difference profile between the Monte Carlo expected profiles for the maximum degree of the cardinal affinity and the ER model for $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$.

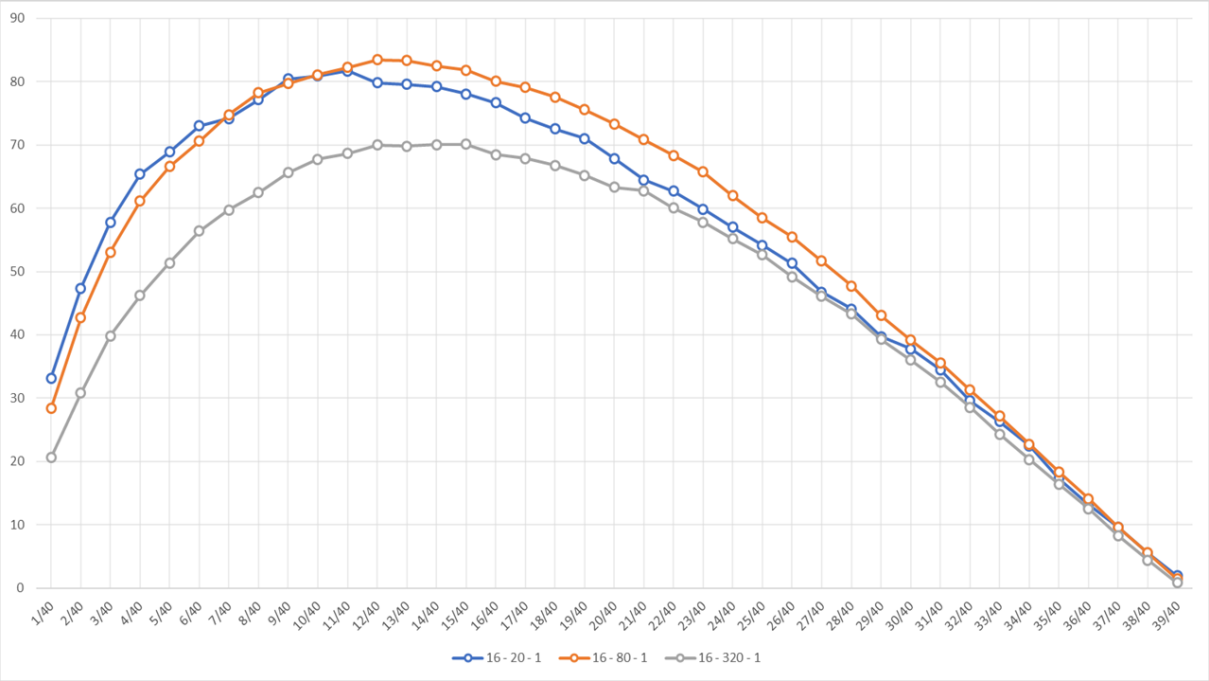


Figure 4: Profile of the difference between the Monte Carlo expected maximum degree of the cardinal affinity and the ER model. Scenarios: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

We noticed that the cardinal affinity function tends to have a maximum degree greater than the random edge model, where the difference increases when θ increases and decreases when m increases.

4.7.2 Transitivity

Figure 5 shows the transitivity profiles of the cardinal affinity function for the scenarios $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma \in \{1, 3, 5\}$. The connection probabilities $p, 1/40 \leq p \leq 40/40$ are on the X-axis. The transitivity coefficient, based on the relative number of triangles in the graph, compared to the total number of connected triples of nodes, is displayed on the Y-axis.

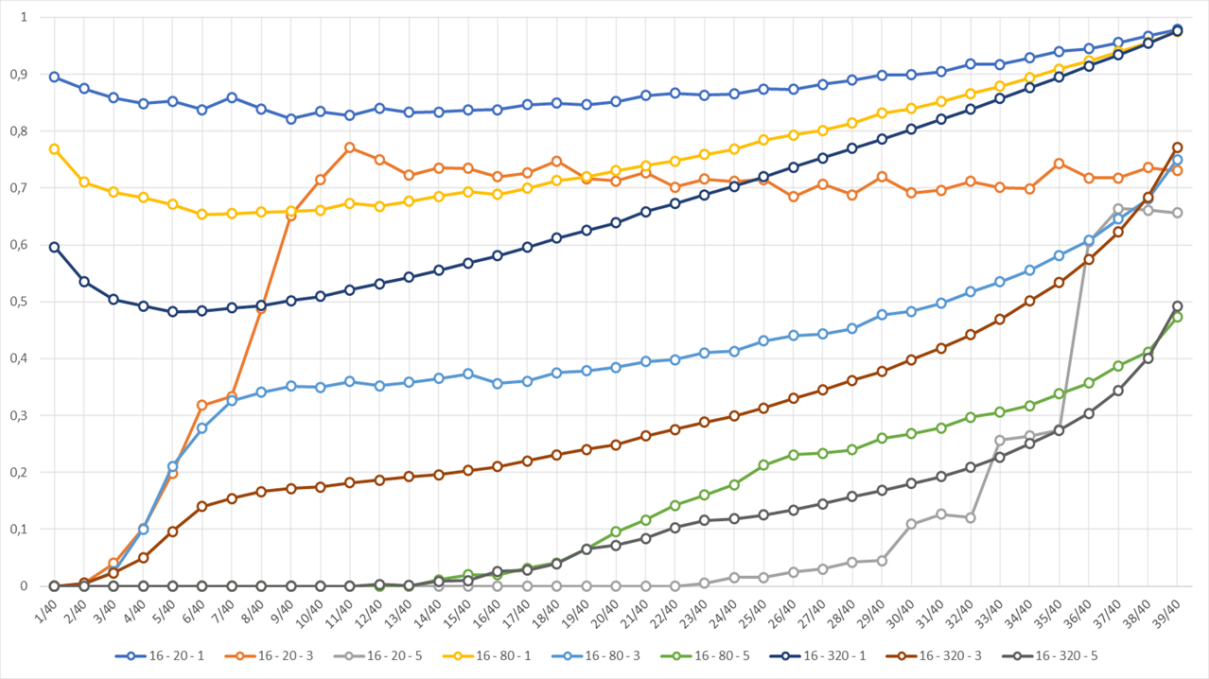


Figure 5: Transitivity profiles of the cardinal affinity function. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ e $m \in \{20, 80, 320\}$

For all values of θ, when $\gamma = 1$, the higher the value of m, the lower the observed transitivity. Besides, there is an increase in the transitivity as θ grows. We observed that concentrating probability on a specific group of characteristics causes that group to be chosen more often, and then choosing a characteristic from this group increases the probability of making connections and, consequently, triangles. A plausible explanation for the drop in transitivity at the initial p levels is that individuals for these p are more likely to choose fewer characteristics. Soon their connections would be formed through a tiny group of characteristics. Hence, if these people choose only one word, the people who choose this word automatically form triangles (as long as there are more than two people, of course). As p increases, individuals are more likely to choose more characteristics. Hence, a new possibility becomes more likely: two individuals who choose characteristics different from each other and a third individual who chooses both. Therefore, although the third individual connects with the first two, this scenario does not form a triangle. By concentrating more weight on a characteristic, that is, increasing θ, it is expected

that the probability that, for a small p, the individual will choose a single characteristic increases. Therefore, it is expected that the rate of drop in transitivity over the initial levels will be lower.

Figure 6 presents the profile of the difference between the Monte Carlo expectation profiles for the transitivity of the cardinal affinity and ER models for $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$.

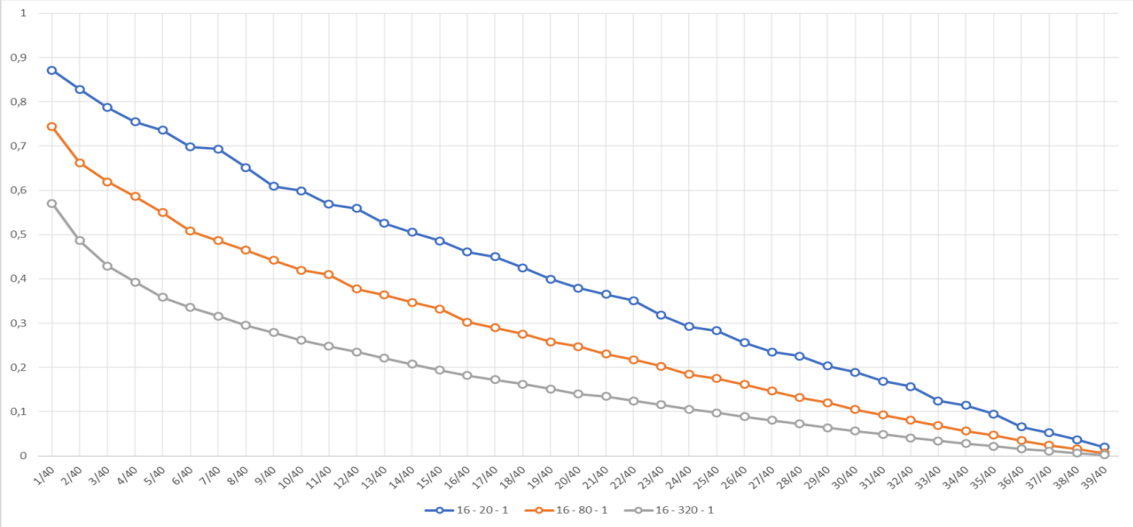


Figure 6: Difference between the Monte Carlo expectation profiles for the transitivity of the cardinal affinity and ER models. Scenarios: $\theta = 16$, $m \in \{20, 80, 320\}$ e $\gamma = 1$

In all cases, we noticed that the transitivity for the cardinal affinity model was superior to the ER model. This result agrees with the argument of [15] for social networks. Also, in all scenarios, we observed that the difference between the expected transitivity increases by reducing m, which agrees with the result of [8]. The difference between transitivity increases by increasing θ while decreasing to zero as p approaches 1.

4.7.3 Clustering

Figure 7 presents the profile of Monte Carlo's expectation for the observed clustering coefficient of the cardinal affinity function for $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$. The connection probabilities $p, 1/40 \leq p \leq 40/40$, are on the X-axis. The clustering coefficient is displayed on the Y-axis.

The curves for each value of m assume different behaviors in the initial values of p when $\gamma = 1$. We observed three behaviors: growth over p, growth followed by a decrease and new growth, and decrease followed by growth.

We observe more evident patterns for the other values of γ. We note that as θ increases, the trend is for the higher m curve to have a higher expected clustering coefficient over p. At the same time, if a value of θ is close to 0, the trend is that the lower values of m have a higher expected clustering coefficient.

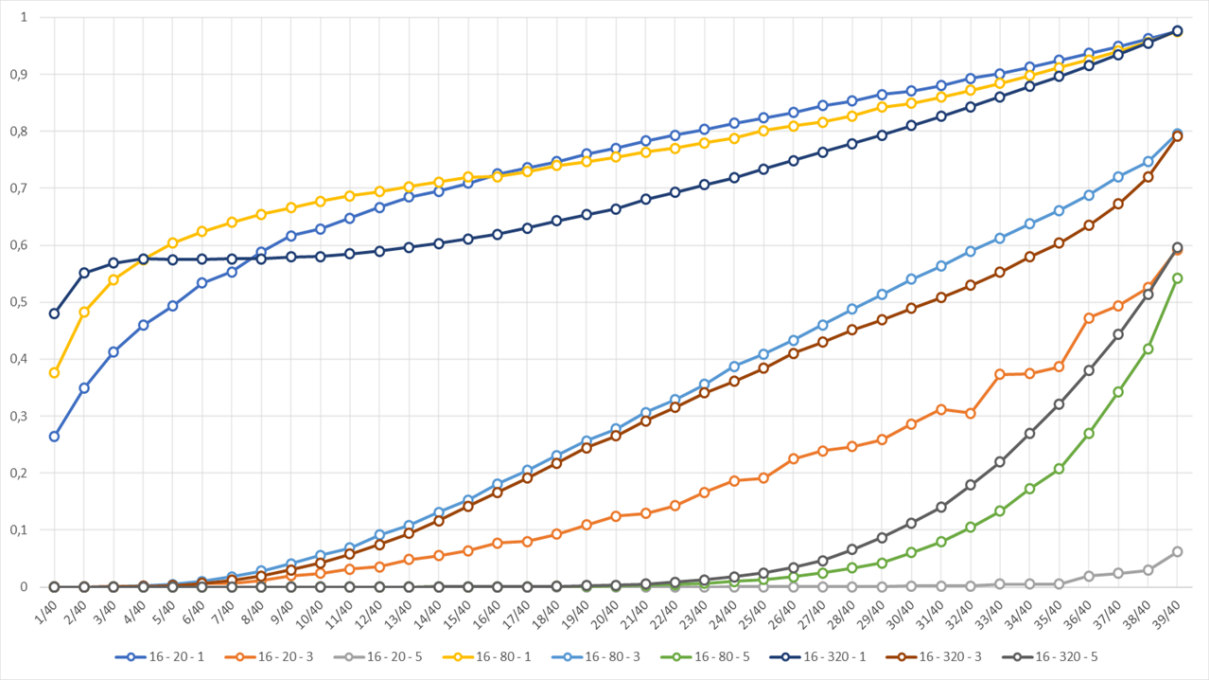


Figure 7: Profile of Monte Carlo's expectation for the observed clustering coefficient of the cardinal affinity function. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$

Figure 8 presents the Difference between the Monte Carlo expectation clustering coefficient for the cardinal affinity model and the ER model for $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$.

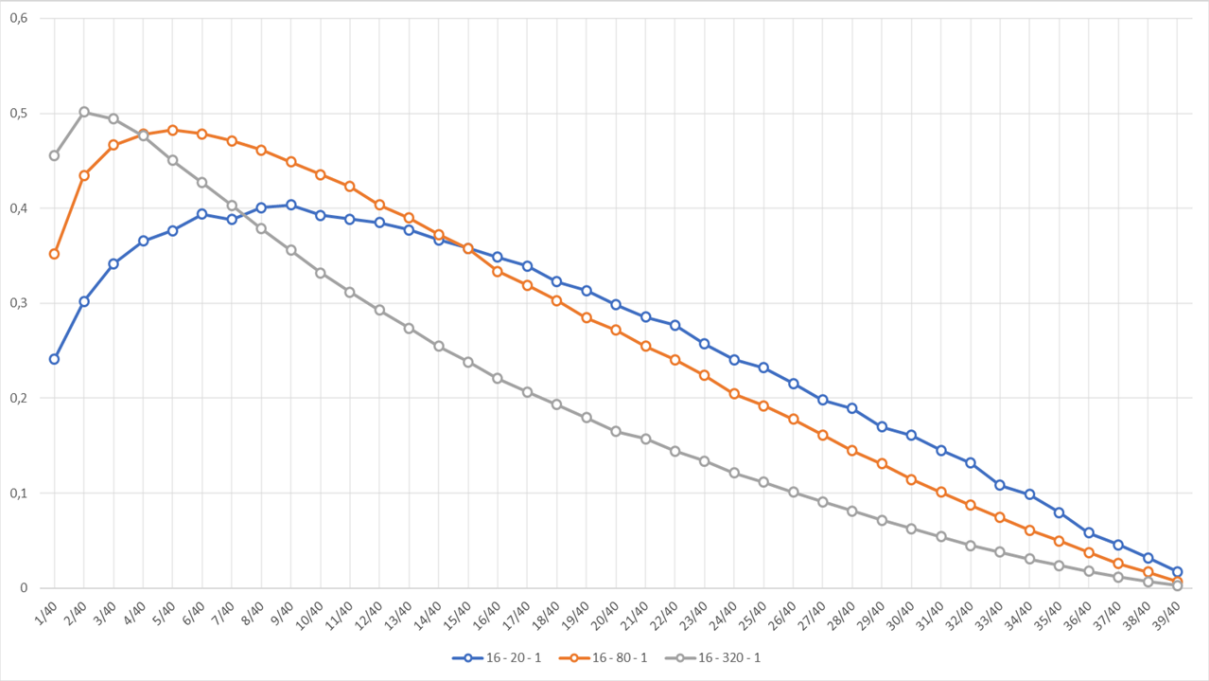


Figure 8: Difference between the Monte Carlo expectation clustering coefficient for the cardinal affinity model and the ER model. Scenarios: $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$

The clustering coefficient is higher than that observed in the ER model, and the higher the m, when $\theta = 0$, the closer they become. We can observe the change in behavior of the difference as the θ grows, suggesting that the limit behavior would be approximately one parable when θ is high. In this case, we observe that for the lowest values of p and θ high, the highest m functions have a more significant difference with the ER model, a panorama that is inverted as p grows.

4.7.4 Closeness

Figure 9 presents the profile of Monte Carlo's expectation for the proximity of the cardinal affinity function for $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$. The connection probabilities $p, 1/40 \leq p \leq 40/40$, are on the X-axis. The closeness is displayed on the Y-axis.

We observed a pattern in the expected proximity. For all observed values of θ and γ, the higher the value of m, the greater the expected proximity. Also, we observe that the proximity in the initial values of p reduces when θ increases, which means that concentrating more probability on a specific group of characteristics makes the distances longer. There is also a shift away from the curves for the m as θ and γ grow. Proximity, in general, grows as p grows.

Figure 10 presents the difference between the Monte Carlo expectation profiles for the prox-

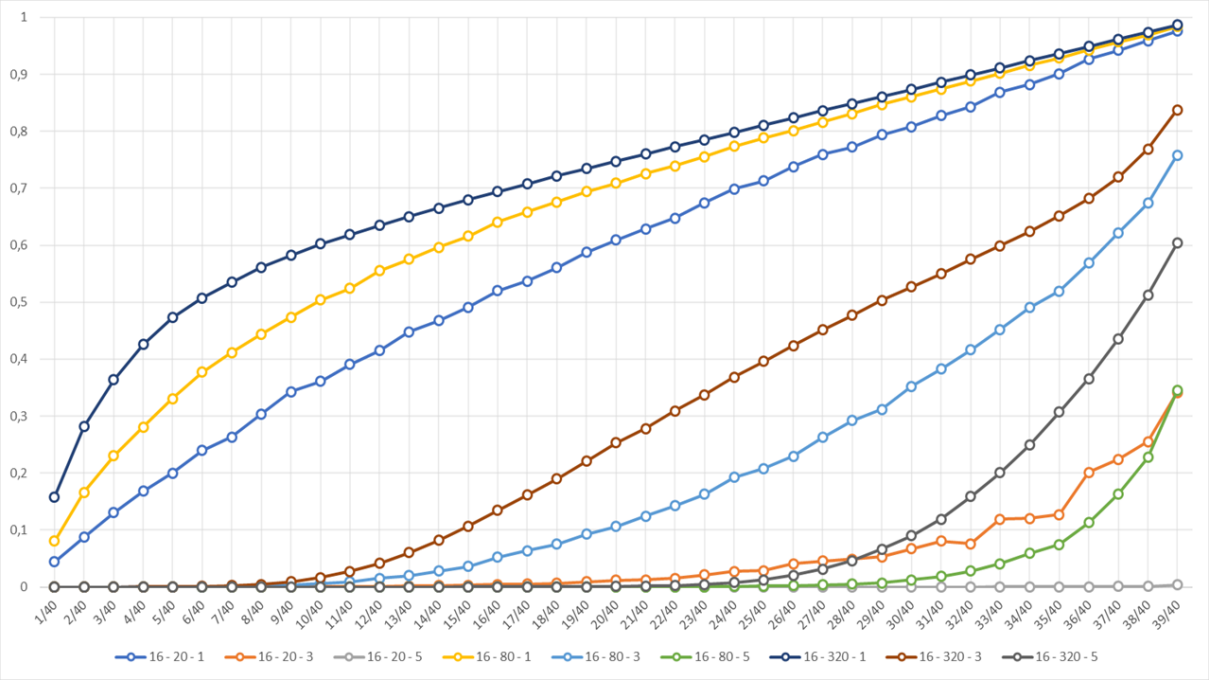


Figure 9: Profile of Monte Carlo's expectation for the proximity of the cardinal affinity function. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$

imity of the cardinal affinity model and ER models' proximity. $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$

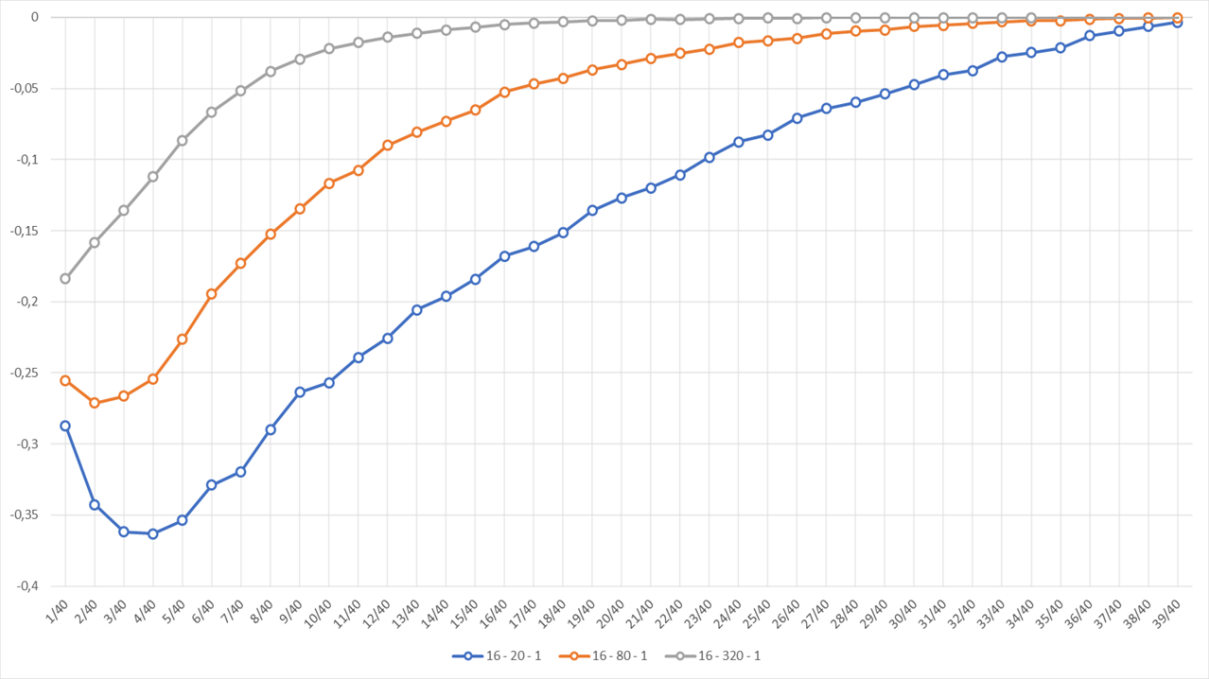


Figure 10: Difference between the Monte Carlo expectation profiles for the proximity of the cardinal affinity model and ER models' proximity. Scenarios: $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$

The graph generated by the cardinal affinity model is less close than in the ER model. We found that increasing θ reduces the proximity and increases its distance from the closeness of ER models while increasing m has the opposite effect. We can observe that for small m and even for large m combined with large θ, a change of concavity begins in the initial levels of p, with an increase in the distance when p is small followed by the reduction for the other p. Besides, as θ grows, the curves for the m move away from each other.

4.7.5 Number of components

Figure 11 presents the profile of Monte Carlo's expected number of components of the affinity network with cardinal function for $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$. We can see a pattern for the expected relative size of the largest component. The higher p, the lower the expected number of components. Besides, the higher the m, the lower the expected number of components. On the other hand, the higher θ, the greater the number of components, which means that increasing m increases the probabilities of a connected graph while concentrating probability on a specific group of characteristics reduces this probability.

Analyzing the observations raised for transitivity, we can see that considering p fixed, a loss

of connectivity offsets the gain concerning transitivity, and the exchange level is influenced by m and θ.

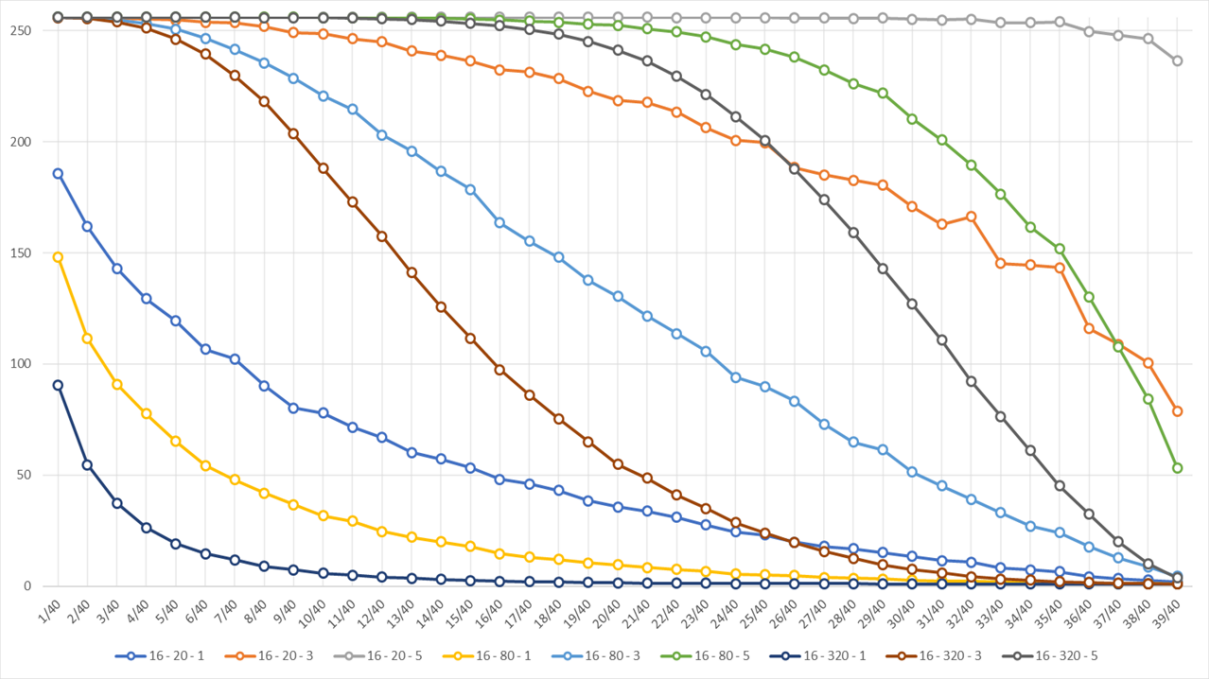


Figure 11: Profile of Monte Carlo's expected number of components of the affinity network with cardinal function. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$

Figure 12 presents the difference between the Monte Carlo expectation of the number of components in the cardinal affinity model and the ER model for $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$. The connection probabilities $p, 1/40 \leq p \leq 40/40$, are on the X-axis. The expected number of components is displayed on the Y-axis.

We observed that the cardinal affinity model is less connected than the ER model. This difference is increased to θ large, especially for small m. The higher the m, the smaller the difference when $\theta = 0$. We notice a decline in the difference as p grows. As stated earlier, combining this interpretation with that of transitivity, for a fixed p, the non-independent arrangement of edges increases transitivity by sacrificing the graph's connectivity. Besides, concentrating probability on a specific group of characteristics causes groups' formation but is much less connected with other characteristics, causing connectivity to decline further.

4.7.6 Largest component size

Figure 13 presents the Monte Carlo expectation profile for the largest component's relative size of the affinity network with affinity function for $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$. The connection probabilities $p, 1/40 \leq p \leq 40/40$, are on the X-axis. The expected large com-

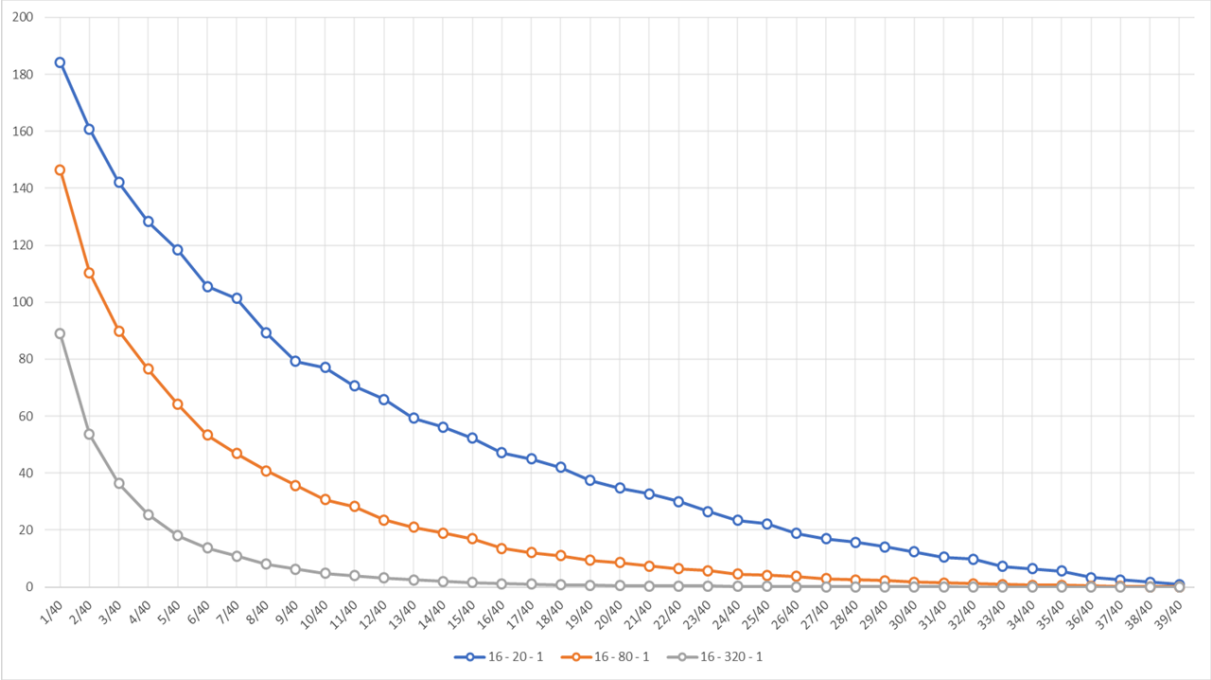


Figure 12: Difference between the Monte Carlo expectation of the number of components in the cardinal affinity model and the ER model Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$

ponent size is displayed on the Y-axis.

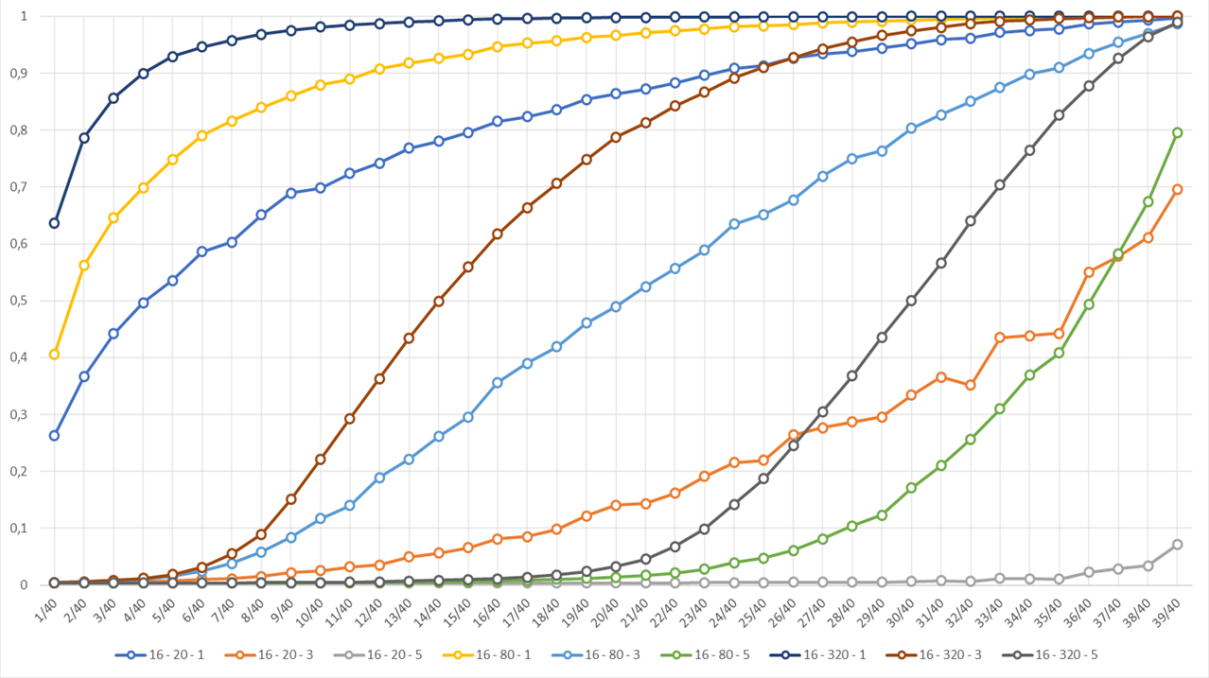


Figure 13: Monte Carlo expectation profile for the largest component's relative size of the affinity network with affinity function. Scenarios: $\theta = 16$, $\gamma \in \{1, 3, 5\}$ and $m \in \{20, 80, 320\}$

We observed a relationship between the number of components and the relative size of the largest component. The behavior of the two topological measurements' profiles seems mirrored, being very similar in shape but in the opposite direction. The higher p, the larger the expected relative size. Besides, the higher the m, the larger the expected relative size. On the other hand, the larger θ, the smaller the expected relative size, which means that increasing m increases the probability of a connected graph while concentrating probability on a specific group of characteristics reduces this probability.

Figure 14 presents the profiles of the difference between the Monte Carlo expectation for the relative size of the largest component of the cardinal affinity model and the ER model for $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$.

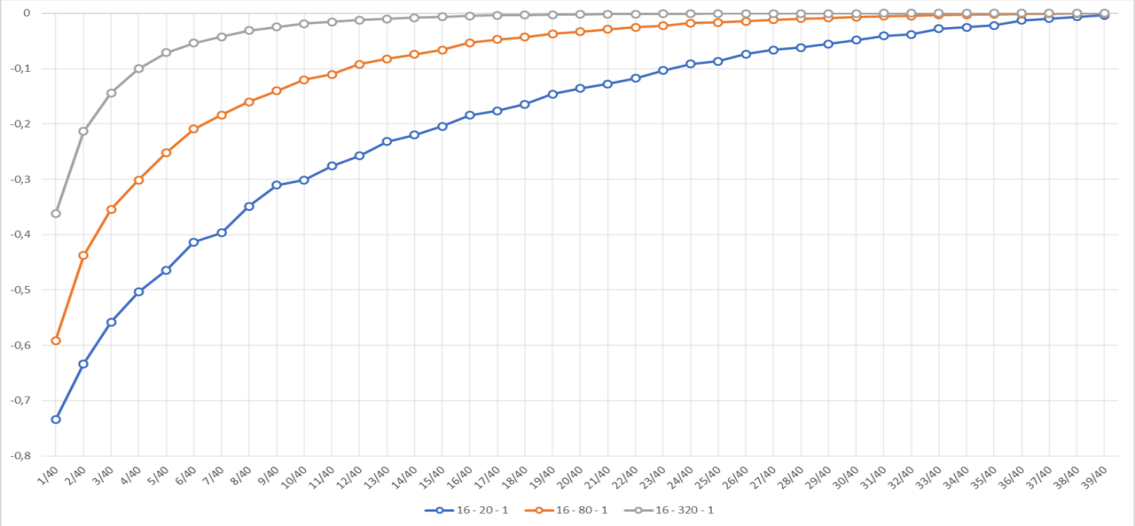


Figure 14: Difference between the Monte Carlo expectation for the relative size of the largest component of the cardinal affinity model and the ER model. Scenarios: $\theta = 16$, $m \in \{20, 80, 320\}$ and $\gamma = 1$

5 Conclusions

We finish this paper with a discussion about the results obtained. The discussion concerns two aspects of our work. Firstly, how the parameters introduced in our methodology affect the topology of the affinity network. Secondly, we address whether affinity networks fit with real-life networks.

5.1 The effect of the parameters on the topology of $G(\lambda)$

We found that the size of the set of characteristics, $\mathcal{D}$, and the level of unbalance of the vector of choice probabilities, Δ, have opposite effects on the affinity network's topological measures and also on the distance between the affinity network model with cardinal function and ER model. Increasing the number of characteristics in the dictionary makes the two models closer, as in the RIG model. In contrast, the increased unbalance of the probability vector moved them away.

We also noted that graphs generated via cardinal affinity have a maximum degree, maximum strength, transitivity, and clustering coefficient higher than the ER model. The cost for this increase is the reduction in the network's connectivity level. On the other hand, they have a smaller size of the larger component and also smaller proximity.

5.2 The Affinity Network and real-life eetworks

As for real-life networks, we begin by highlighting the type of properties observed in real-life networks. By analyzing empirical data Watts and Strogatz (1998) and Barabási and Albert

(1999) observed that real-life networks have small diameters and heavy-tailed degree distributions. M. E. J Newman (2001) investigated networks of scientific collaboration, finding that these networks have high clustering coefficients. In 2006, Ahn et al. investigated real-life networks with more than 10 million nodes, such as MySpace and Orkut finding high clustering coefficients.

As our results suggest, for proper choices of the parameters, Affinity Networks with the cardinal affinity are capable of capturing features observed in real-life networks. All these characteristics make this model a good candidate to fit with a real network. Therefore, we claim that the affinity network model with a high unbalance parameter is a good candidate to fit networks based on instruments that collect characteristics of the individuals, like those who collect words, which inspired this work.

The study developed here can be extended to models with other affinity functions and dependence structures, but we leave it for future work.

Acknowledgements: The authors want to thank CAPES for its financial support.

References

[1] Frank G Ball, David J Sirl, and Pieter Trapman. Epidemics on random intersection graphs. *The Annals of Applied Probability*, 24(3):1081–1128, 2014.

[2] Jayanth R Banavar, Francesca Colaiori, Alessandro Flammini, Amos Maritan, and Andrea Rinaldo. Topology of the fittest transportation network. *Physical Review Letters*, 84(20):4745, 2000.

[3] Albert-Laszlo Barabási and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[4] Mindaugas Bloznelis and Julius Damarackas. Degree distribution of an inhomogeneous random intersection graph. *the electronic journal of combinatorics*, 20(3):P3, 2013.

[5] Maria Deijfen and Willemien Kets. Random intersection graphs with tunable degree distribution and clustering. *Probability in the Engineering and Informational Sciences*, 23(4):661–674, 2009.

[6] Denise Duarte, Gilvan Ramalho Guedes, Wesley H. Silva Pereira, and Rodrigo B. Ribeiro. Representing collective thinking through cognitive networks. *Journal of Complex Network*, 10, issue 6, 2022.

[7] Paul Erdős and Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[8] James Allen Fill, Edward R Scheinerman, and Karen B Singer-Cohen. 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*, 16(2):156–176, 2000.

[9] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[10] Erhard Godehardt and Jerzy Jaworski. Two models of random intersection graphs for classification. In *Exploratory data analysis in empirical research*, pages 67–81. Springer, 2003.

[11] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[12] W Jeżewski. Scale-free properties of weighted networks with connectivity-driven topology. *Physica A: Statistical Mechanics and its Applications*, 354:672–680, 2005.

[13] Michał Karoński, Edward R Scheinerman, and Karen B Singer-Cohen. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8(1-2):131–159, 1999.

[14] Mark EJ Newman. Random graphs with clustering. *Physical review letters*, 103(5):058701, 2009.

[15] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

[16] Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, Gábor Szabó, M Argollo De Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New journal of physics*, 9(6):179, 2007.

[17] Geovane Tavares dos Santos and José Manuel de Barros Dias. Teoria das Rrepresentações Sociais: uma abordagem sociopsicológica. *PRACS*, 8(1):173–187, 2015.

[18] Akrati Saxena. *Evolving Models for Dynamic Weighted Complex Networks*, pages 177–208. Springer Singapore, Singapore, 2022.

[19] Karen Singer. *Random intersection graphs*. PhD thesis, Department of Mathematical Sciences, The Johns Hopkins University, 1995.

[20] Nikolai Smirnov and NV Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. 1939.

[21] Dudley Stark. The vertex degree distribution of random intersection graphs. *Random Structures & Algorithms*, 24(3):249–258, 2004.

[22] P Vergés. L'évocation de l'agent: une méthode pour la dédinition du nouau central d'une representation. *Bulletin de Psychologie*, 45(405):203–209, 1992.