Center for
Mathematical Economics
Working Papers

515

August 2016

Continuous Homophily and Clustering in Random Networks

Florian Gauer and Jakob Landwehr

Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
http://www.imw.uni-bielefeld.de/wp/
ISSN: 0931-6558

Continuous Homophily and Clustering in Random Networks

Florian Gauer[*] and Jakob Landwehr[†]

First Version: July 2014
This Version: August 2016

Abstract

We propose a random network model incorporating heterogeneity of agents and a continuous notion of homophily. Unlike the vast majority of the corresponding economic literature, we capture homophily in terms of similarity rather than equality by assuming that the probability of linkage between two agents continuously decreases in the distance of their characteristics. A homophily parameter directly determines the strength of this effect. As a main result, we show that for any positive level of homophily our model exhibits clustering, that is an increased probability of linkage given a common neighbor. As opposed to this, the seminal Bernoulli Random Graph model à la Erdős and Rényi (1959) is comprised as the limit case of no homophily. Moreover, simulations indicate that, although the average distance between agents increases in homophily, the well-known small-world phenomenon is preserved even at high homophily levels. We finally provide a possible application in form of a stylized labor market model, where a firm can hire a new employee via the social network.

Keywords: Random Graphs, Homophily, Clustering, Small-World Phenomenon, Network Formation, Labor Market Search

JEL-Classification: D85, J64, Z13.

[*] Center for Mathematical Economics, Bielefeld University, postbox 100131, D-33501 Bielefeld, Germany. Email: fgauer86@gmail.com, phone: +49 521 106 4918.

[†] Center for Mathematical Economics, Bielefeld University, postbox 100131, D-33501 Bielefeld, Germany. Email: jakob.landwehr@uni-bielefeld.de, phone: +49 521 106 4918.

1 Introduction

Suppose you own a firm and want to fill an open vacancy through the social contacts of one of your current employees. Whom would you ask to recommend someone? Most probably you would address the worker who would himself perform best in the position in question. While this seems to be intuitively reasonable, why do we expect it to be optimal? One important reason is that people tend to connect to similar others. This phenomenon is known as homophily (Lazarsfeld and Merton, 1954).

In this paper, we introduce a continuous notion of homophily based on incorporating heterogeneity of agents into the Bernoulli Random Graph (BRG) model as examined by Erdős and Rényi (1959). To this end, we propose a two-stage random process which we call *Homophilous Random Network* model. First, agents are assigned characteristics independently drawn from a continuous interval and second a network realizes, linking probabilities being contingent on a homophily parameter and the pairwise distance between agents' characteristics. This enables us to account for homophily in terms of similarity rather than equality of agents, capturing the original sociological definition instead of the stylized version up to now commonly used in the economic literature.

As a first result, we determine the expected linking probabilities between agents (Proposition 1) as well as the expected number of links (Corollary 2). We then calculate the expected probability that an agent has a certain number of links (Proposition 2), showing that the according binomial distribution of the original BRG model is preserved to some degree. Further, we establish a threshold theorem for any given agent to be connected (Proposition 3). For all these (and further) results we demonstrate that the BRG model is comprised as the limit case of no homophily and we thus provide a generalization thereof. As a main result, we show that in our model homophily induces clustering (Theorem 1), two stylized facts frequently observed in real-world networks which are not captured by the BRG model.[1] Furthermore, clustering proves to be strictly increasing in homophily. As a second important feature of our model, two simulations indicate that, although the average distance between agents increases in homophily, the well-known small-world phenomenon is preserved even at high homophily levels.[2] We finally provide an application of the Homophilous Random Network model within a stylized labor market setting to answer the introductory questions.

[1] A network exhibits clustering if two individuals with a common neighbor have an increased probability of being connected.

[2] The small-world phenomenon describes the observation that even in large networks on average there exist relatively short paths between two individuals.

In the literature the presence of homophily has been established in a wide range of sociological and economic settings. Empirical studies on social networks discovered strong evidence for the similarity of connected individuals with respect to age (see e.g. Verbrugge, 1977; Marsden, 1988; Burt, 1991), education (see e.g. Marsden, 1987; Kalmijn, 2006), income (see e.g. Laumann, 1966, 1973), ethnicity (see e.g. Baerveldt et al., 2004; Ibarra, 1995) or geographical distance (see e.g. Campbell, 1990; Wellman, 1996). For an extensive survey see McPherson et al. (2001). In recent years, economists have developed an understanding of the relevance of network effects in a range of economic contexts. Thus, bearing in mind the presence of homophily in real-world networks can be of great importance for creating meaningful economic models.

There already exists a strand of economic literature examining homophily effects in different settings (see e.g. Currarini et al., 2009). Most of the models assume a finite type space and binary homophily in the sense that an agent prefers to connect to others that are of the same type while not distinguishing between other types.[3] Thus, these models rather capture the idea of equality than of similarity. However, in reality people are in many respects neither "equal" nor "different". We therefore believe that a notion that provides an ordering of the "degree of similarity" with respect to which an agent orders his preference for connections can capture real-world effects more accurately. This gives rise to a continuous notion of homophily in networks.

This approach is followed by Gilles and Johnson (2000) and Iijima and Kamada (2014), who examine strategic, deterministic models of network formation. In both models individual utility is shaped directly by homophily such that individuals connect if (and only if) they are sufficiently similar. Iijima and Kamada (2014) consider the extreme case of purely homophilous utility functions, entailing that a high level of homophily is directly identified with efficiency. As opposed to this, in our random graph model, a novel continuous homophily measure is incorporated as a parameter that may be freely chosen to reflect a broad range of possible situations. In their multi-dimensional framework, Iijima and Kamada (2014) examine clustering and the average path length as functions of the number of characteristics agents take into account when evaluating their social distance to others. In contrast, we investigate the direct relation between homophily and these network statistics. The differences in methodology especially lead to opposing results concerning the small-world phenomenon. While in Iijima and Kamada (2014) small worlds only arise if agents

[3] For several homophily measures of this kind see Currarini et al. (2009).

disregard a subset of characteristics, we show that this phenomenon is well present in our one-dimensional setting.

Besides the presence of homophily, stylized facts such as the small-world phenomenon and high levels of clustering have indeed been empirically identified in real-world networks (see e.g. Milgram, 1967; Watts and Strogatz, 1998). As in many cases these networks are very large and remain unknown for an analysis, typically random networks are used as an approximation.[4] This constitutes a challenge to design the random network formation process in a way to ensure it complies with the observed stylized facts.

Since the seminal work of Erdős and Rényi (1959), who developed and analyzed a random graph model where a fixed number out of all possible bilateral connections is randomly chosen, a lot of different models have been proposed (see e.g. Wasserman and Pattison, 1996; Watts and Strogatz, 1998; Barabási and Albert, 1999). The most commonly used until today is the BRG model where connections between any two agents are established with the same constant probability. It has been shown that for large networks this model is almost equal to the original model of Erdős and Rényi (1959) (for details see Jackson, 2006; Bollobás, 2001).[5] It is well understood that this model reproduces the small-world phenomenon but does not exhibit clustering. Also, a notion of homophily is not present as the described random process does not rely on individual characteristics. The latter is also true for the small-world model proposed by Watts and Strogatz (1998). Starting from a network built on a low-dimensional regular lattice, they reallocate randomly chosen links and obtain a random network showing a small-world phenomenon. According to their notion this encompasses an increased level of clustering. However, the socio-economic causality of this occurrence remains uncertain. In this regard our model can to some extend serve as a socio-economic foundation of the work of Watts and Strogatz (1998). An approach to generate random graphs more similar to ours is proposed by the recently emerging graph-theoretic literature on random intersection graphs (see e.g. Karonski et al., 1999). Here, each node is randomly assigned a set of features. Connections are then established between any two nodes sharing a given number of features. It has been shown that the resulting graphs also exhibit clustering (Bloznelis, 2013).

In general, not much work has yet been dedicated to the incorporation of homophily into random networks. However, some papers exist that include similar ideas. Closest to our work is perhaps Jackson (2008a), who analyzes the impact of

[4] For instance, this might be of interest, when investigating the formation of opinions, buying decisions, social mobility, the spreading of information or diseases, etc. in societies.

[5] In fact, the BRG model rather than their original one is nowadays also known as the Erdős-Rényi model.

increasing homophily on network statistics such as clustering and the average distance of nodes. A finite number of types, linking probabilities between these, as well as agents' expected degrees are exogenously given. While Jackson (2008a) uses a predefined partition of agents into groups and then considers a random network model based on Chung and Lu (2002), we consider a two-stage random process where (instead of only the network) also agents' characteristics are determined randomly and which yields a generalization of the BRG model. Additionally considering the concrete functional form of homophilous linking probabilities, our model is immediately available as an approximation tool for large societies. However, the major difference between the two papers is revealed by a contradictory result on the average distance between agents. While Jackson (2008a) finds that the average distance is invariant with respect to changes in homophily, our Simulation 1 indicates that it increases in homophily (see Section 5).[6]

Another, however less closely related paper in this strand of literature is the one by Golub and Jackson (2012) who also assume a finite number of types as well as the linking probabilities between them to be exogenously given. Based on this they analyze the implications of homophily in the framework of dynamic belief formation on networks. Bramoullé et al. (2012) combine random link formation and local search in a sequentially growing society of heterogeneous agents and establish a version of binary homophily along with a degree distribution.

In all cases, besides the concrete continuous notion of homophily, a major distinction of our approach is the sequential combination of two random processes where agents' characteristics are considered as random variables that influence the random network formation. We thus account for the fact that in many applications, in which the network remains unobserved, it seems unnatural to assume that individual characteristics, which in fact may depict attitudes, beliefs or abilities, are perfectly known.

We conclude this paper by providing an application of our model for the labor market, proposing an analysis of the introductory question: When is it optimal for a firm to search for a new employee via the contacts of a current employee? We assume the characteristic of each worker to be her individual ability to fill the open vacancy and use our Homophilous Random Network model as an approximation of

[6] In fact, our results indicate that his finding crucially depends on the assumption that – as opposed to our setting (see Section 3) – agents' degrees do not depend on homophily. Note that, being based on Chung and Lu (2002), this assumption is inevitable in his model. However, we think that this is up for discussion as in many applications (political attitude, income, social status, etc.) agents with extreme characteristics will tend to have fewer links than those with intermediate characteristics and the extent of this difference will heavily depend on the actual level of homophily.

the workers' network. Given an agent and her characteristic, we determine the expected characteristic of a random contact (Proposition 4). This gives rise to a simple decision rule stating in which constellations firms should hire via the social network. In particular, given sufficiently high levels of homophily and the current employee's ability, it proves to be optimal to always hire via the social network.

Within the job search literature, Horváth (2014) and Zaharieva (2013) incorporate homophily among contacts into job search models. However, these models are again based on a binary concept of homophily and do not include an explicit notion of networks. This research strand traces back to the work of Montgomery (1991), who was the first to address this issue. Finally, to some extent, our application captures an idea proposed by Ioannides and Loury (2004) to combine this class of models with a random network setting à la Erdős-Rényi.[7]

The rest of the paper is organized as follows. In Section 2 we set up the model. Section 3 reveals basic properties of homophilous random networks while results on clustering can be found in Section 4. In Section 5 we simulate the model focusing on the small-world phenomenon. Section 6 contains the labor market application and Section 7 concludes. Proofs of most results are provided in the appendix.

2 The Model

We set up a model of random network formation where first each agent is randomly assigned a continuous characteristic which then influences the respective linking probabilities. We refer to this as the *Homophilous Random Network* model. Consider a set of agents $N = \{1, 2, ..., n\}$. A *connection* or *(undirected) link* between two agents $i, j \in N$ is denoted by $ij = ji := \{i, j\}$. By $g^N := \{ij \mid i, j \in N\}$ we denote the complete network, that is the network where any two agents are connected. Then, we let $\mathbb{G} := \{g \mid g \subseteq g^N\}$ be the set of all possible *non-directed graphs* or *networks.* Further, we define $N_i(g) := \{j \in N \mid ij \in g\}$ to be the set of *neighbors* of agent i in network g, and let $\eta_i(g) := |N_i(g)|$ denote the number of her neighbors. This is sometimes also referred to as the *degree* of agent i. Each agent is assigned a *characteristic* p_i where the vector $p = (p_1, p_2, ..., p_n)$ denotes a certain realization of the random

[7] Ioannides and Loury (2004, p. 1068) state "It would be interesting to generalize the model of social structure employed by Montgomery, by assuming groups of different sizes. For example, one may invoke a random graphs setting (Paul Erdős and Alfred Rényi 1960; Ioannides 1997), where a fraction of the entire economy may be in groups whose sizes are denumerable but possibly large."

variable $P = (P_1, P_2, ..., P_n)$. The underlying distribution of each P_i is assumed to be standard uniform. Hence, all P_i are identically and independently distributed.

Subsequent to the assignment of characteristics a random network forms. Here, based on the Bernoulli Random Graph (BRG) model as introduced by Erdős and Rényi (1959), we assume the following variation. The *linking probability* of two agents $i, j \in N$ is given by

$$q(p_i, p_j) := \lambda a^{|p_i - p_j|}, \tag{1}$$

where the *scaling parameter* $\lambda \in [0, 1]$ and the *homophily parameter* $a \in [0, 1]$ are exogenously given and independent of agents i and j. Note that, in situations where the vector of characteristics is unknown, $q(P_i, P_j)$ is a random variable such that the linking probability $q(p_i, p_j)$ is in fact a conditional probability. Figure 1 depicts the linking probabilities $q(p_i, p_j)$ for different homophily parameters a, first as a function of the distance of characteristics and second as a function of p_j for given $p_i = 0.25$. As in our model λ simply serves as a scaling parameter corresponding to the linking probability in the BRG model, in Figure 1 it is fixed to one for simplicity.

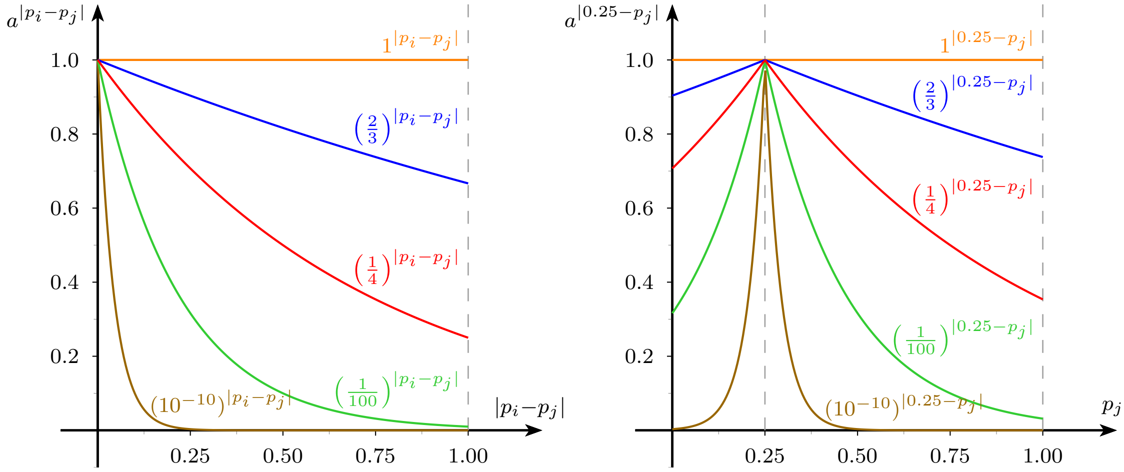

Figure 1: Left: Linking probability for all distances of characteristics for several homophily parameters a; Right: Linking probabilities for an agent with characteristic $p_i = 0.25$ for several homophily parameters a

Let us shortly elaborate on the role of the homophily parameter a. Observe that the linking probability q is decreasing in $|p_i - p_j|$ as a takes values only in $[0, 1]$. In particular, for $a = 1$ the model is equal to the BRG model as all linking probabilities are equal to λ and hence independent of the agents' characteristics. On the contrary, if we have $a = 0$, then solely agents with identical characteristics

$p_i = p_j$ get connected with probability λ while all other linking probabilities are zero. Insofar, the parameter a serves as a measure of homophily in the model. Here, lower parameter values correspond to a higher homophily level in the network. The notion at hand measures homophily in a continuous instead of a binary manner since the distance function $|\cdot|$ is continuous. Note, however, that an increase in homophily which leads to a decreased linking probability then also implies a decreased ex-ante expected degree of agents. Whenever suitable, one may therefore choose the scaling parameter λ dependent on a such that the expected degree is kept constant for any level of homophily (see Remark 1).

3 Basic Properties of Homophilous Random Networks

This section constitutes a foundation for the upcoming main results. To this end, we first need to collect several important properties of the Homophilous Random Network model, such as the expected linking probabilities and the number of links of agents. Moreover, we discuss a threshold theorem for an agent to be isolated. This is of particular importance for the labor market application provided in Section 6. Throughout this section we explore, on the one hand, situations in which the realization of one considered agent $i \in N$ is known while all others are not and, on the other hand, situations in which the whole vector of characteristics is unknown. In any case we demonstrate that the BRG model is recuperated as the limit case of no homophily and we thus provide a generalization thereof.

We start by determining the expected linking probabilities for two given agents $i, j \in N$ in the following proposition.

Proposition 1. *Given agent i's realized characteristic $P_i = p_i$ while all other characteristics p_{-i} are unknown, the expected probability that a certain link ij forms is*

$$\mathbb{E}^P\left[\mathbb{P}^G\left(ij \in G \mid P\right) \;\middle|\; P_i = p_i\right] = \frac{\lambda}{\ln(a)}\left(a^{p_i} + a^{1-p_i} - 2\right) =: \varphi(\lambda, a, p_i). \qquad (2)$$

If the vector p is unknown, the expected probability that the link ij forms is

$$\mathbb{E}^P\left[\mathbb{P}^G\left(ij \in G \mid P\right)\right] = \frac{2\lambda}{\ln(a)^2}\Big(a - 1 - \ln(a)\Big) =: \Phi(\lambda, a). \qquad (3)$$

The proof of Proposition 1 as well as all subsequent proofs can be found in the appendix. It is straightforward to understand that the function φ indeed has to depend on characteristic p_i as it makes a difference whether p_i tends to the center or to the boundaries of the interval $[0,1]$. In what follows, a characteristic of the former (latter) kind is said to be "intermediate" ("extreme"). The closer p_i is to 0.5 the smaller is the expected difference with respect to other agents' characteristics, hence, the higher is the expected linking probability φ. In particular, it is $\arg\max_{p_i} \varphi = 0.5$ and $\arg\min_{p_i} \varphi = \{0,1\}$ for all $a \in (0,1)$. To this respect, it is obvious that $\varphi(\lambda, a, 0) \leq \Phi(\lambda, a) \leq \varphi(\lambda, a, 0.5)$ for all $\lambda, a \in [0,1]$. Also, it is important to note that the expected linking probability is decreasing in homophily, that is for all $a \in (0,1]$ we have

$$\frac{\partial}{\partial a}\Phi(\lambda, a) = \frac{\partial}{\partial a}\left[2\lambda\frac{a-1-\ln(a)}{\ln(a)^2}\right] = 2\lambda\frac{2(1-a)+\ln(a)(1+a)}{a\ln(a)^3} > 0.^8$$

To verify intuition that our model reproduces the BRG model as a limit case and to gain insights on the behavior in boundary cases, the following corollary is concerned with the limits of the expected linking probabilities with respect to the homophily parameter a.

Corollary 1. *For maximal homophily, i.e. for $a \to 0$, the expected linking probability is*

$$\lim_{a\to 0} \varphi(\lambda, a, p_i) = \lim_{a\to 0} \Phi(\lambda, a) = 0. \tag{4}$$

In case of no homophily, i.e. for $a \to 1$, the expected linking probability is

$$\lim_{a\to 1} \varphi(\lambda, a, p_i) = \lim_{a\to 1} \Phi(\lambda, a) = \lambda. \tag{5}$$

As usual, a proof is provided in the appendix. Maximal homophily in this model means that only agents with identical characteristics would have a strictly positive linking probability. However, since the standard uniform distribution has no mass point, such two agents do not exist with positive probability. Therefore, both according expected linking probabilities φ and Φ tend to zero. In case of no homophily, as mentioned before, the model indeed reproduces the BRG model such that all linking probabilities are alike, independent of individual characteristics p.

Based on Proposition 1, we also immediately get the expected number of links of an agent.

[8] We indeed can include the value $a = 1$ here as it happens to be a removable discontinuity of the derivative. On the contrary, at $a = 0$ the right-handed derivative is infinity as the expected number of links is zero with probability one.

Corollary 2. *The expected number of links of an agent i with given characteristic $P_i = p_i$ is*

$$\mathbb{E}^P \Big[\mathbb{E}^G \left[\eta_i(G) \mid P \right] \ \Big| \ P_i = p_i \Big] = (n-1)\varphi(\lambda, a, p_i). \tag{6}$$

Similarly, if p is unknown, we have

$$\mathbb{E}^P \Big[\mathbb{E}^G \left[\eta_i(G) \mid P \right] \Big] = (n-1)\Phi(\lambda, a). \tag{7}$$

A proof of this corollary is omitted as it is clear that all expected linking probabilities are independent and, hence, the result follows directly from the proof of Proposition 1. Observe that from this result, we can also calculate the ex-ante expected number of links in a network to be

$$\frac{n(n-1)}{2}\Phi(\lambda, a).$$

Together with Corollary 1 this gives that the ex-ante expected number of links is zero for maximal homophily while in case of no homophily, again as in the BRG model, one gets $\lambda n(n-1)/2$ links in total in expectation. More generally, agents' ex-ante expected degree in a network depends on the level of homophily and the number of agents in the network. For reasons of comparability, however, it is sometimes required to keep agents' ex-ante expected number of links fixed (see e.g. Section 5). In this context, consider the following remark.

Remark 1. *For $a \in (0,1)$ and $n \in \mathbb{N}$ consider some number $\eta^{exp} \in \mathbb{R}_+$ and the function*

$$\bar{\lambda}(a, n, \eta^{exp}) := \frac{\eta^{exp} \ln(a)^2}{2(n-1)(a-1-\ln(a))}.$$

Note that, for $\frac{\eta^{exp}}{n}$ reasonably small, we have $\bar{\lambda}(a, n, \eta^{exp}) \in [0,1]$.[9] *Any agent's ex-ante expected number of links is then given by*

$$\mathbb{E}^P \Big[\mathbb{E}^G \left[\eta_i(G) \mid P \right] \Big] \Big|_{\lambda = \bar{\lambda}(a,n,\eta^{exp})} = (n-1)\Phi(\bar{\lambda}(a, n, \eta^{exp}), a) \equiv \eta^{exp},$$

meaning that it is invariant with respect to changes in homophily as well as in the number of agents.

However, for the expected number of links of an agent i with given characteristic

[9] Also, note that for $a = 1$, i.e. for the limit case of no homophily, it is consistent to define $\bar{\lambda}(1, n, \eta^{exp}) := \lim_{a \to 1} \bar{\lambda}(a, n, \eta^{exp}) = \frac{\eta^{exp}}{n-1}$.

$P_i = p_i$ we calculate

$$\mathbb{E}^P\left[\mathbb{E}^G\left[\eta_i(G) \mid P\right] \;\middle|\; P_i = p_i\right]\Big|_{\lambda=\bar{\lambda}(a,n,\eta^{exp})} = (n-1)\varphi(\bar{\lambda}(a,n,\eta^{exp}),a,p_i)$$
$$= \eta^{exp}\frac{\ln(a)(a^{p_i}+a^{1-p_i}-2)}{2(a-1-\ln(a))}, \qquad (8)$$

meaning that it still depends on a and p_i.

Thus, it is important to note that, in our model, fixing agents' ex-ante expected degree does not imply that an agent's expected degree, given her realized characteristic, is invariant with respect to changes in the level of homophily.[10] Indeed, it seems plausible to expect agents with an extreme characteristic (i.e. with p_i close to zero or one in our model) to have fewer links than agents with an intermediate characteristic.[11] And moreover, this effect's magnitude should heavily depend on the actual level of homophily. This can indeed be observed in Figure 2 where an agent's expected degree as in (8) is plotted for $\eta^{exp} = 1$ as a function of her realized characteristic $p_i \in [0,1]$ and the level of homophily $a \in (0,1)$.

At $a = 1$, which depicts the limit case of no homophily, independently of her characteristic, any agent has an expected degree of η^{exp}. If a decreases, that is if homophily increases, then the expected number of links of agents with an intermediate characteristic (with an extreme characteristic) increases (decreases). This is because for an agent with an intermediate characteristic there will be more agents with a similar characteristic than for an agent with an extreme characteristic.[12] However, the higher the level of homophily, the stricter is the interpretation of similarity in our model. This implies that, from some (relatively high) level of homophily on, there is an expanding range of characteristics for which agents' expected degrees reapproach. On the contrary, for sufficiently extreme characteristics, the effect described first keeps dominating and becomes even stronger as homophily increases. Finally, for the limit case of maximal homophily, we have

$$\lim_{a\to 0}\mathbb{E}^P\left[\mathbb{E}^G\left[\eta_i(G) \mid P\right] \;\middle|\; P_i = p_i\right]\Big|_{\lambda=\bar{\lambda}(a,n,\eta^{exp})} = \begin{cases} \eta^{exp} & \text{for} \quad p_i \in (0,1) \\ \frac{1}{2}\eta^{exp} & \text{for} \quad p_i \in \{0,1\} \end{cases}.$$

[10] This is in contrast to the model considered by Jackson (2008a), where any agent's degree with given characteristic is fixed.

[11] In other words, one should expect the degree of an agent i with characteristic p_i to be decreasing in $|p_i - 0.5|$.

[12] The reason for this is simply that $\mathcal{U}[0,1]$ has a finite support, meaning that, for extreme characteristics, nearby other ones are rather onesided whereas for intermediate characteristics it will typically be the case that there are nearby other ones being smaller and greater.

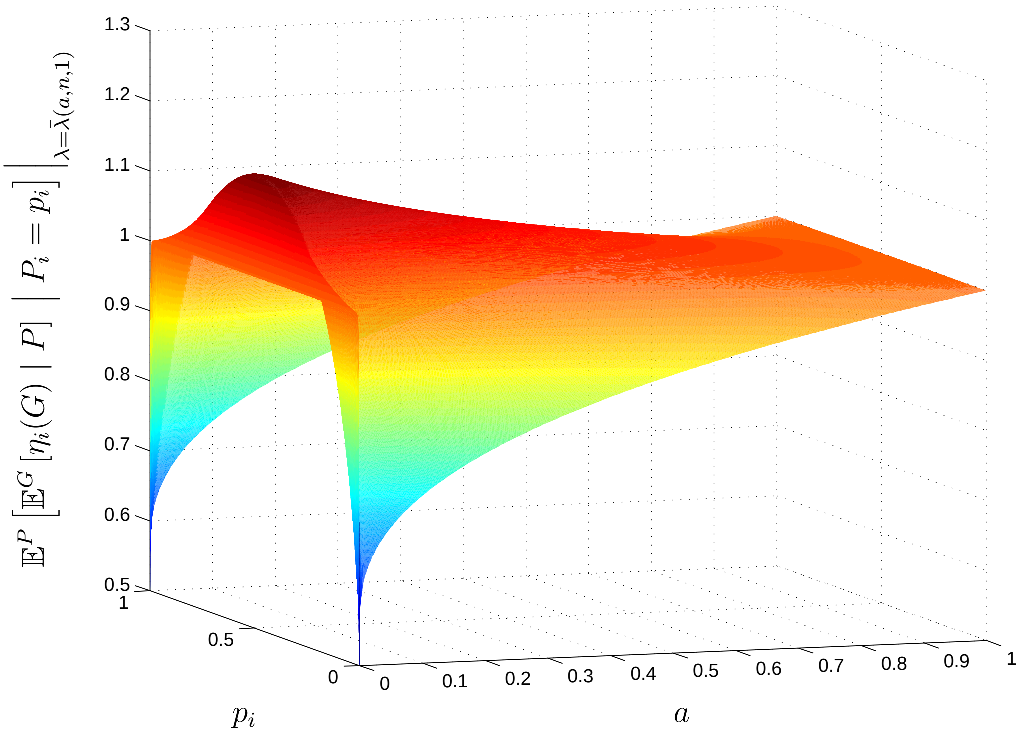

Figure 2: Expected degree of an agent with realized characteristic p_i, given homophily level $a \in (0,1)$ and the ex-ante expected degree being fixed at one (created with MATLAB, 2014)

To sum up, we have that, for all levels of homophily $a \in (0,1)$, there exist two ranges of extreme characteristics (one close to zero and the other one close to one) where agents have relatively few links in expectation. While these ranges shrink in homophily, the expected degrees of the respective agents decrease even further. Apparently, the latter is of great importance when considering the average distance between agents at different levels of homophily (see Simulation 1 in Section 5).

In what follows, we calculate the expected probability for an agent with given characteristic to have a certain number of links. This entails that the model inherits a version of the binomial distribution known from the BRG model.

Proposition 2. *The expected probability that an agent i with given characteristic $P_i = p_i$ has $k \in \{0, 1, ..., n-1\}$ links is given by*

$$\mathbb{E}^P\left[\mathbb{P}^G\left(\eta_i(G) = k \mid P\right)\Big|P_i = p_i\right] = \binom{n-1}{k} \cdot \varphi(\lambda, a, p_i)^k \cdot (1 - \varphi(\lambda, a, p_i))^{n-k-1}. \tag{9}$$

Observe that this form can be interpreted as a binomial distribution with parameters $\varphi(\lambda, a, p_i)$ and $n-1$. Further, it is worth noting that the extreme cases meet the expected outcome as we have

$$\lim_{a\to 0} \mathbb{E}^P \left[\mathbb{P}^G \left(\eta_i(G) = k \mid P\right) \Big| P_i = p_i\right] \overset{(4)}{=} \binom{n-1}{k} \cdot 0^k \cdot 1^{n-k-1} = \begin{cases} 1, & \text{if } k = 0 \\ 0, & \text{else} \end{cases},$$

$$\lim_{a\to 1} \mathbb{E}^P \left[\mathbb{P}^G \left(\eta_i(G) = k \mid P\right) \Big| P_i = p_i\right] \overset{(5)}{=} \binom{n-1}{k} \cdot \lambda^k \cdot (1-\lambda)^{n-k-1},$$

where the latter term, unsurprisingly, is equal to the probability for any agent to have k links in the BRG model with independent linking probability λ. Unfortunately, the calculation in case that the whole vector of characteristics p is unknown is analytically not tractable.

One major reason why random network models are used frequently is to match qualitative characteristics of real-world networks. The law of large numbers in this case yields that large networks indeed meet these characteristics with a high probability (see e.g. Jackson, 2008b, Chapter 4). A seminal contribution of Erdős and Rényi (1959) was to provide so called threshold theorems for the case of the BRG model. These results state that, if the network size n goes to infinity while the linking probability $\lambda(n)$ goes to zero slower than some threshold $t(n)$, then the limit network has a certain property with probability one. On the contrary, if $\lambda(n)$ goes to zero faster than $t(n)$, then the limit network has the same property only with probability zero.[13] It is clear that this kind of results can only be found for monotone properties, that is for those which yield that, if any network g has the property, then also any network $g' \supseteq g$ has it. One example is the property that a given agent has at least one link which we establish in the next proposition. For instance, regarding our application of the labor market (Section 6) this feature is of great importance. In that context, we assume this as a prerequisite as determining the expected characteristic of a given agent's contact is meaningful only if this agent is not isolated.

Proposition 3. *Assume a minimal level of homophily to be guaranteed as the network size becomes large. Then the function $t(n) = 1/(n-1)$ is a threshold for a given agent to be non-isolated in the following sense:*

$$\mathbb{E}^P \left[\mathbb{P}^G \left(\eta_i(G) \geq 1 \mid P\right) \Big| P_i = p_i\right] \to 1 \quad \forall\, p_i \in [0,1] \quad \textit{if} \quad \frac{-\lambda(n)/\ln(a(n))}{t(n)} \to \infty,$$

$$\mathbb{E}^P \left[\mathbb{P}^G \left(\eta_i(G) \geq 1 \mid P\right) \Big| P_i = p_i\right] \to 0 \quad \forall\, p_i \in [0,1] \quad \textit{if} \quad \frac{-\lambda(n)/\ln(a(n))}{t(n)} \to 0.$$

[13] For a more elaborate characterization of thresholds as well as several results see Bollobás (1998).

First, note that in Proposition 3 the right-hand side conditions are equivalent to $\varphi(\lambda(n), a(n), \hat{p})/t(n)$ converging to infinity or zero, respectively, for any arbitrary $\hat{p} \in [0, 1]$. For details refer to the proof in the appendix. What is surprising about this (as well as about other threshold theorems), is the sharp distinction made by the threshold $t(n)$, in the sense that if the growth of probability φ passes the threshold $t(n)$, then the probability of any agent to be isolated changes "directly" from zero to one. What is more, notice that the threshold $t(n) = 1/(n-1)$ is actually the same as in the BRG model. However, it has to hold for φ rather than just for λ since in this model both λ and a may vary with respect to the size of the network. Indeed, it does not seem farfetched to assume that homophily increases with the network size as the assortment of similar agents gets larger. Having understood this, one can directly deduce the cases where only one of the two parameters varies with n.

Corollary 3. *If $a \equiv a(n)$ depends on n but λ does not, one gets that if $a(n)$ goes toward zero faster than* $\exp(-n)$*, then any given agent is isolated with probability one in the limit while if $a(n)$ does not go toward zero or at least slower than* $\exp(-n)$*, then any given agent has at least one link with probability one in the limit.*
If $\lambda \equiv \lambda(n)$ depends on n but a does not, the condition collapses to the threshold of $t(n)$ for $\lambda(n)$ as in the BRG model where any given agent has at least one link if $\lambda(n)$ grows faster than $t(n)$ while if $\lambda(n)$ grows slower than $t(n)$, any given agent is isolated with probability one.

Both parts of the corollary follow directly from Proposition 3 such that a proof can be omitted.

4 Clustering

As mentioned in the introduction, a main criticism of the Bernoulli Random Graph (BRG) model is that the resulting networks do not exhibit clustering while most examples of real-world networks do so (see e.g. Watts and Strogatz, 1998; Newman, 2003, 2006). In this section, we show that our Homophilous Random Network model indeed exhibits clustering and one can use the homophily parameter a to calibrate it to a broad range of degrees of clustering.

The notion of clustering in general captures the extent to which connections in networks are transitive, that is the frequency with which two agents are linked to each other given that they have a common neighbor. Watts and Strogatz (1998), who introduced this concept, measure the transitivity of a network by a global clustering coefficient which denotes the average probability that two neighbors of a given agent are directly linked as well. A random graph model is said to exhibit clustering if the coefficient is larger than the general, unconditional linking probability of two agents (see Newman, 2006). Considering the set of networks that contain some link $ij \in g^N$, that is $\mathbb{G}_{ij} := \{g \subseteq g^N \mid ij \in g\} \subset \mathbb{G}$, this can be transferred to our model in the following way:

Definition 1 (Clustering). *For the Homophilous Random Network model with $\lambda \in [0,1]$ and $a \in (0,1)$ the* clustering coefficient *is defined as*

$$C(\lambda, a) := \mathbb{E}^P \left[\mathbb{P}^G \left(G \in \mathbb{G}_{jk} \mid P\right) \;\middle|\; G \in \mathbb{G}_{ij} \cap \mathbb{G}_{ik}\right]$$

where $i, j, k \in N$. The model is said to exhibit clustering *if we have $C(\lambda, a) > \Phi(\lambda, a)$.*

The choice of the agents i, j and k obviously cannot have an influence in this context since ex ante, i.e. before characteristics realize, all agents are assumed to be equal. Further, recall that Φ gives the probability of two agents to be connected, characteristics being unknown. The function C captures this probability as well, however, conditional on the existence of a common neighbor. It should be clear that the original BRG model does not exhibit clustering since every link is formed with the same independent probability. As a main result of this paper, we discover next that, apart from the limit case of no homophily, our Homophilous Random Network model has this feature and is insofar more realistic.

Theorem 1 (Clustering in Homophilous Random Networks). *In the Homophilous Random Network model the clustering coefficient is given by*

$$C(\lambda, a) = \lambda \frac{3\Big(\ln(a)a^2 + \ln(a) - a^2 + 1\Big)}{2\Big(2\ln(a)a + 4\ln(a) + a^2 - 8a + 7\Big)}.$$

Given a non-extreme homophily parameter, the model exhibits clustering, that is we have

$$C(\lambda, a) > \Phi(\lambda, a)$$

for all $\lambda \in (0,1]$, $a \in (0,1)$.

The intuition for the proof of this theorem (which is again presented in the appendix) is the following: If there is homophily to some degree and two agents have a common neighbor, then this fact contains additional information. The expected distance between these two agents is smaller than if there is no assumption about a common neighbor. Again due to homophily, it is therefore more likely that a link between these two agents forms. Also, Figure 3 might contribute to a better understanding of the situation. Note here that $C(\lambda, a)/\lambda \equiv C(1, a)$ and $\Phi(\lambda, a)/\lambda \equiv \Phi(1, a)$. One can additionally perceive that the difference $C(\lambda, a) - \Phi(\lambda, a)$ is strictly decreasing in $a \in (0, 1)$ for all $\lambda \in (0, 1]$, that is clustering is strictly increasing in the degree of homophily.

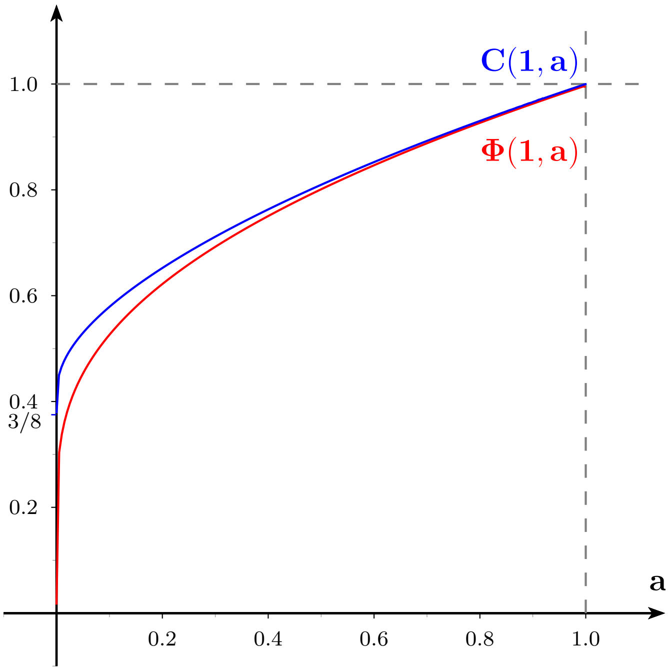

Figure 3: Clustering coefficient $C(1, a)$ and unconditional linking probability $\Phi(1, a)$ for all homophily parameters $a \in (0, 1)$

Again, it is of interest to consider the limit cases of maximal and no homophily which we do in the following corollary.

Corollary 4. *For maximal homophily, i.e. for $a \to 0$, we have*

$$\lim_{a \to 0} C(\lambda, a) = \lim_{a \to 0} [C(\lambda, a) - \Phi(\lambda, a)] = \frac{3}{8}\lambda.$$

In case of no homophily, i.e. for $a \to 1$, we get

$$\lim_{a \to 1} C(\lambda, a) = \lim_{a \to 1} \Phi(\lambda, a) = \lambda.$$

If there is no homophily, we are again back in the BRG model which we already know not to exhibit clustering. Insofar, the second part of the corollary is consistent. However, the more interesting case is the one of maximal homophily. Though in the limit no link forms with positive probability, one can deduce properties regarding the case of homophily being high, yet not maximal, due to continuity of the functional forms. Let us clarify the intuition why the clustering coefficient takes a value strictly between zero and λ if homophily is maximal. Recall first that we have $\lim_{a\to 0}\Phi(\lambda, a) = 0$ since for maximal homophily only agents with identical characteristics are linked with positive probability and such two agents exist with probability zero. However, the clustering coefficient is a probability conditioned on the existence of links to a common neighbor. This additional information implies that either characteristics are equal or links have formed despite differing characteristics. Though both events occur only with probability zero, this does not preclude them per se. Having understood this, it should be clear that in the former case the probability of the third link would indeed be λ while in the latter case it would still be zero. Taken together, this yields $\lim_{a\to 0} C(\lambda, a) \in (0, \lambda)$. It remains surprising, however, that the clustering coefficient takes the specific value $\frac{3}{8}\lambda$.

5 The Small-World Phenomenon

Besides the presence of homophily and clustering, another stylized fact is frequently observed in real-world networks which is widely known as the small-world phenomenon. It captures the finding that, even in large networks, there typically exist remarkably short paths between two individuals. The original BRG model is known to reproduce this feature (see e.g. Bollobás, 2001; Chung and Lu, 2002).

Thus, in this section, we aim to establish the small-world phenomenon to be preserved in our Homophilous Random Network (HRN) model even in case of homophily being high. For this purpose, we present and analyze simulations of homophilous random networks as this issue seems to be no longer analytically tractable. Our simulations provide a strong indication that also in cases of high homophily the small-world phenomenon remains present. Additionally, we apply two alternative statistical notions of clustering. It turns out that their values are not significantly different from the analytical measure given in Definition 1. In the following, Figure 4 may already provide a first intuition regarding the differences between cases of high and low homophily. In particular, while the total number of links is almost the same in both simulated 100-agent networks, one observes clustering merely in the first case.

The notion of the small-world phenomenon usually grounds on the average shortest path length between all pairs of agents belonging to a network and having a connecting path. We also refer to this as the "average distance" between agents.

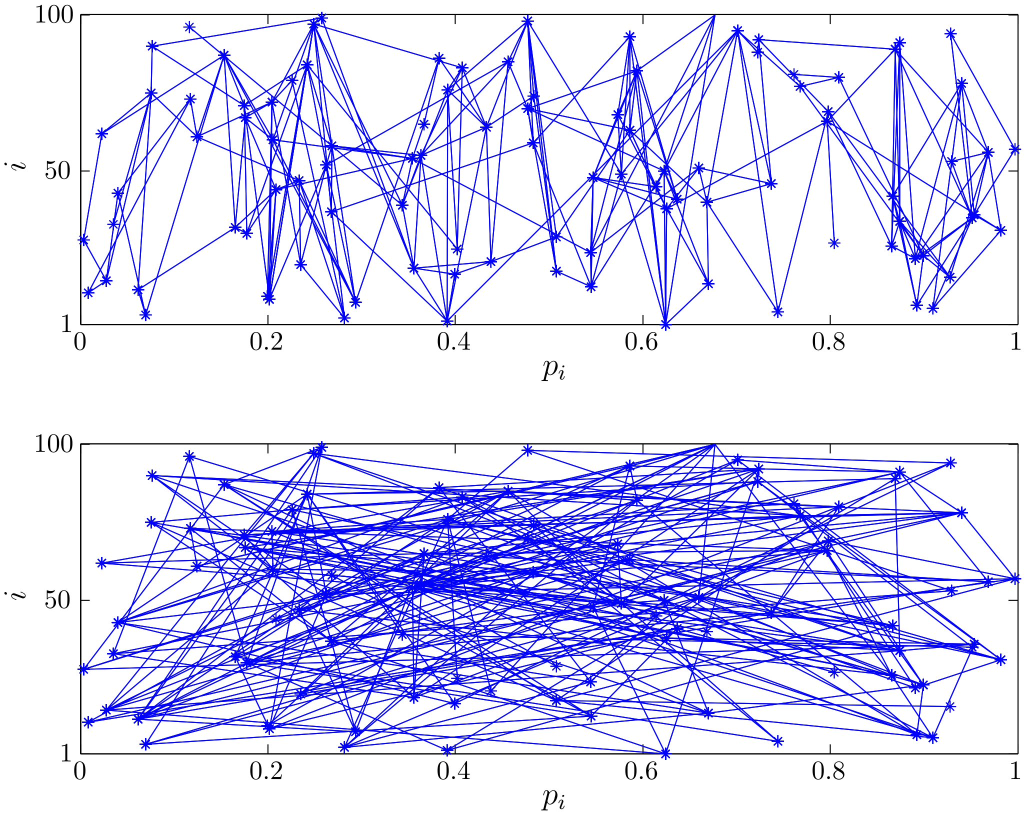

Figure 4: Top: HRN with $\lambda = 0.5, a = 10^{-8}$; $\#links = 484$
Bottom: BRG with linking probability $\Phi(0.5, 10^{-8}) = 0.0513$; $\#links = 496$
(created with MATLAB, 2014)

With regard to real-world networks the small-world phenomenon is a rather vague concept since it is typically based on subjective assessments of path lengths rather than on verifiable, definite criteria. However, most people will agree that the values for several real-world networks as for instance compiled by Watts and Strogatz (1998) and Newman (2003) are surprisingly low. Insofar, it could be said that most of these networks exhibit the small-world phenomenon. A formal definition of the small-world phenomenon applicable to most random network models is formulated by Newman (2003) and reads as follows:

Definition 2 (Small-World Phenomenon)**.** *A random network is said to exhibit the* small-world phenomenon *if the average distance $\bar{d}$ between agents scales logarithmically or slower with network size n while keeping agents' expected degree constant, that is if $\bar{d}/\ln(n)$ is non-increasing in n.*

As already mentioned, it has been established that the original BRG model exhibits the small-world phenomenon according to Definition 2 (see e.g. Bollobás, 2001; Chung and Lu, 2002). It is not clear, however, whether this still holds for our generalization, given a considerably high level of homophily, but the results of the following simulations provide some indication.

Prior to this, let us additionally introduce two statistical notions of clustering which are frequently used in the literature and closely related to the one given in Definition 1. The simulations allow to compare these for our model. Here, clustering is associated with an increased number of triangles in the network. More precisely, both alternative clustering measures are defined based on the ratio of the number of triangles and the number of connected triples. A triangle is a subnetwork of three agents all of whom being connected to each other while a connected triple is a subnetwork of three agents such that at least one of them is linked to the other two. Formally, this amounts to the following definition.

Definition 3 (Statistical Clustering)**.** *For a given network with set of agents* $N = \{1, ..., n\}$*, the* (statistical) clustering coefficients $C^{(1)}$ *and* $C^{(2)}$ *are determined by*

$$C^{(1)} := \frac{3 \times \textit{number of triangles in the network}}{\textit{number of connected triples in the network}} \qquad \textit{and}$$
$$C^{(2)} := \frac{1}{n} \sum_{i \in N} \frac{\textit{number of triangles containing agent } i}{\textit{number of connected triples centered on agent } i}.$$

The coefficient $C^{(1)}$ counts the overall number of triangles and relates it to the overall number of connected triples in the network. The factor of three accounts for the fact that each triangle contributes to three connected triples. The second one, $C^{(2)}$, which goes back to Watts and Strogatz (1998), first calculates an individual clustering coefficient for each agent and then averages these. Compared to the first one, $C^{(2)}$ gives more weight to low-degree agents.[14] Additionally, note that $C^{(2)}$ is only well-defined if there are no isolated or loose-end agents in the network.

To capture both the heuristic and the formal approach to the small-world phenomenon, we present the outcomes of two different simulations. In Simulation 1, we fix the number of agents $n = 500$ and the ex-ante expected degree of any agent to $\eta^{exp} = 15$. Recalling Remark 1, the latter is done by choosing $\lambda \equiv \bar{\lambda}(a, 500, 15) = \frac{15 \ln(a)^2}{998(a-1-\ln(a))}$. We then select several homophily levels ranging from no homophily, i.e. the limit case of the BRG model, to very high homophily, represented by $a = 10^{-8}$. For each parameter value of a, we then simulate a homophilous random network $R = 1000$ times and assess the averaged network statistics. The parameters and network statistics of this simulation are stated in Table 1. Note that fixing the ex-ante

[14] Referring to $C^{(2)}$, Newman (2003, p. 184) states "This definition effectively reverses the order of the operations of taking the ratio of triangles to triples and of averaging over vertices – one here calculates the mean of the ratio, rather than the ratio of the means."

expected degree enables us to compare our results for different homophily levels as this implies identical values for $\Phi(\lambda, a)$ in all cases. Recall that Φ captures the ex-ante expected probability of two agents to be connected, that is for characteristics being unknown (recall Proposition 1 and Corollary 2).

Parameter / Statistics	$a = 1$	$a = 10^{-2}$	$a = 10^{-4}$	$a = 10^{-6}$	$a = 10^{-8}$
n	500				
R	1000				
Exp. Degree η^{exp}	15				
Exp. Linking Prob. Φ	0.0301				
$\bar{\lambda}(a, n, \eta^{exp})$	0.0301	0.0882	0.1553	0.2239	0.2928
Avg. Degree $\bar{\eta}$	14.9990	15.0074	15.0098	14.9899	15.0037
	(0.2475)	(0.3064)	(0.2986)	(0.2925)	(0.2839)
Avg. Distance $\bar{d}$	2.5944	2.6288	2.8086	3.0806	3.3939
	(0.0113)	(0.0164)	(0.0277)	(0.0429)	(0.0611)
$\bar{d}/\ln(n)$	0.4175	0.4230	0.4519	0.4957	0.5461
	(0.0018)	(0.0026)	(0.0045)	(0.0069)	(0.0098)
Clustering Coeff. C	0.0301	0.0411	0.0641	0.0892	0.1147
Clustering Coeff. $C^{(1)}$	0.0301	0.0411	0.0642	0.0891	0.1147
	(0.0013)	(0.0016)	(0.0023)	(0.0029)	(0.0035)
Clustering Coeff. $C^{(2)}$	0.0301	0.0411	0.0642	0.0892	0.1148
	(0.0015)	(0.0019)	(0.0026)	(0.0032)	(0.0039)

Table 1: Results of Simulation 1 comparing network statistics for different homophily levels ranging from no homophily (BRG) to extreme homophily; Standard errors stated in parentheses (carried out with MATLAB, 2014)

Regarding the results of Simulation 1, we find that the average distance increases in homophily. This is in line with intuition as agents with (widely) differing characteristics are increasingly likely to be distant in the network. Moreover, in Section 3 we revealed that (and how) the expected degree of an agent with given characteristic varies in homophily (see again Remark 1 and Figure 2). To be more precise, the increase in average distance seems to be due to the fact that there are typically agents with extreme characteristics whose expected degrees decrease as homophily increases.[15] However, it increases by less than one link from no to highest homophily. Also, an average distance of less than 3.4 between two agents can still be considered relatively small in a network of 500 agents with about 15 links on average. Thus, regarding the heuristic approach, it seems reasonable to accept the small-world phe-

[15] The results of Jackson (2008a) can be regarded as a confirmation thereof as in his model, where agents' expected degrees are exogenously given and therefore invariant with respect to homophily, the average distance remains unchanged if homophily increases.

nomenon to be exhibited for all homophily levels.[16]

Furthermore, we observe an increasing level of clustering for the simulated homophilous random networks. This is in line with the findings in Section 4. If homophily is highest, the probability that two agents are linked, given they have a common neighbor, is about four times as high as in the case of the Bernoulli Random Graphs where this probability coincides with the unconditional linking probability $\Phi(\lambda, a)$. Another expectable, yet important observation is that there are no significant differences between the expected clustering coefficient C (recall Definition 1) and the values we determined for the statistical coefficients $C^{(1)}$ and $C^{(2)}$ (recall Definition 3).[17] To sum up, Simulation 1 indicates that the Homophilous Random Network model exhibits the small-world phenomenon and clustering at the same time for all $a \in (0, 1)$. In what follows, we consider the most interesting case of highest homophily captured by $a = 10^{-8}$ in more detail.

Simulation 2 focuses on the formal Definition 2 of the small-world phenomenon. For this purpose, we simulate a collection of $R = 100$ networks for each size $n = 150, 200, 250, ..., 1000$, again keeping agents' ex-ante expected degree fixed, and compute the respective averages of the relevant network statistics. To this end, we consider the parameter of highest homophily that is regarded in Simulation 1. The precise data is stated in Table 2. Note that, for each network size, this second simulation is structurally the same as the first one, merely a smaller number of iterations is chosen due to computational restrictions. However, as can be seen in Table 1, all standard errors and especially the one of the ratio $\bar{d}/\ln(n)$ are very low. Thus, 100 iterations should be sufficient to generate a precise estimate.

In Figure 5, we plot the ratio of the average distance and the logarithm of the network size $\bar{d}/\ln(n)$ for the different network sizes n. This ratio is decreasing in n as the illustration reveals. From this, we deduce that the average distance $\bar{d}$ increases slower in n than $\ln(n)$ does. Thus, the homophilous random networks exhibit the small-world phenomenon according to Definition 2.

Finally, one would expect to observe the less triangles of links between agents, that is the less clustering, the larger the network. This is because we keep agents' ex-ante expected degrees fixed while increasing the number of possible neighbors. Indeed, the statistical clustering coefficients $C^{(1)}$ and $C^{(2)}$ are decreasing in the network size n (see Table 2). By increasing the network size even further, our simulation indicates

[16] To calculate the average distance, one commonly restricts to agents having a connecting path if the network has more than one component. However, such a network realized extremely rarely in this simulation, namely only in 0.06% of all cases.

[17] Note that isolated and loose-end agents never appeared in the simulation, guaranteeing that $C^{(2)}$ was steadily well-defined.

Parameter / Statistics	$n = 150$	200	250	300	350	400
R	100					
a	10^{-8}					
Expected Degree η^{exp}	15					
$\bar{\lambda}(a, n, \eta^{exp})$	0.980	0.734	0.587	0.489	0.419	0.366
Average Degree $\bar{\eta}$	15.10	14.98	14.96	15.00	14.95	14.98
Average Distance $\bar{d}$	3.027	3.109	3.189	3.235	3.284	3.323
$\bar{d}/\ln(n)$	0.6042	0.5868	0.5776	0.5671	0.5606	0.5546
Clustering Coeff. $C^{(1)}$	0.385	0.287	0.229	0.191	0.164	0.143
Clustering Coeff. $C^{(2)}$	0.386	0.288	0.229	0.191	0.164	0.143

Parameter / Statistics	$n = 450$	500	550	600	650	700
R	100					
a	10^{-8}					
Expected Degree η^{exp}	15					
$\bar{\lambda}(a, n, \eta^{exp})$	0.325	0.293	0.266	0.244	0.225	0.209
Average Degree $\bar{\eta}$	15.04	15.00	14.98	15.02	15.02	14.97
Average Distance $\bar{d}$	3.356	3.401	3.417	3.459	3.466	3.501
$\bar{d}/\ln(n)$	0.5493	0.5472	0.5416	0.5408	0.5352	0.5345
Clustering Coeff. $C^{(1)}$	0.128	0.115	0.104	0.095	0.088	0.082
Clustering Coeff. $C^{(2)}$	0.128	0.115	0.104	0.095	0.088	0.082

Parameter / Statistics	$n = 750$	800	850	900	950	1000
R	100					
a	10^{-8}					
Expected Degree η^{exp}	15					
$\bar{\lambda}(a, n, \eta^{exp})$	0.195	0.183	0.172	0.162	0.154	0.146
Average Degree $\bar{\eta}$	15.02	14.98	14.98	14.99	15.01	15.00
Average Distance $\bar{d}$	3.519	3.541	3.546	3.575	3.599	3.605
$\bar{d}/\ln(n)$	0.5315	0.5297	0.5257	0.5255	0.5249	0.5218
Clustering Coeff. $C^{(1)}$	0.076	0.072	0.068	0.064	0.060	0.057
Clustering Coeff. $C^{(2)}$	0.076	0.072	0.068	0.064	0.061	0.057

Table 2: Results of Simulation 2 computing average degrees, distances and small world ratios of the HRN model for a growing network size (carried out with MATLAB, 2014)

that both clustering coefficients will approach zero as the network becomes infinitely large. Note that this is in line with the behavior of the expected clustering coefficient C where we simply have a factor of $(n-1)$ in the denominator if we insert $\lambda = \bar{\lambda}(a, n, \eta^{exp})$ (see Theorem 1 and Remark 1).

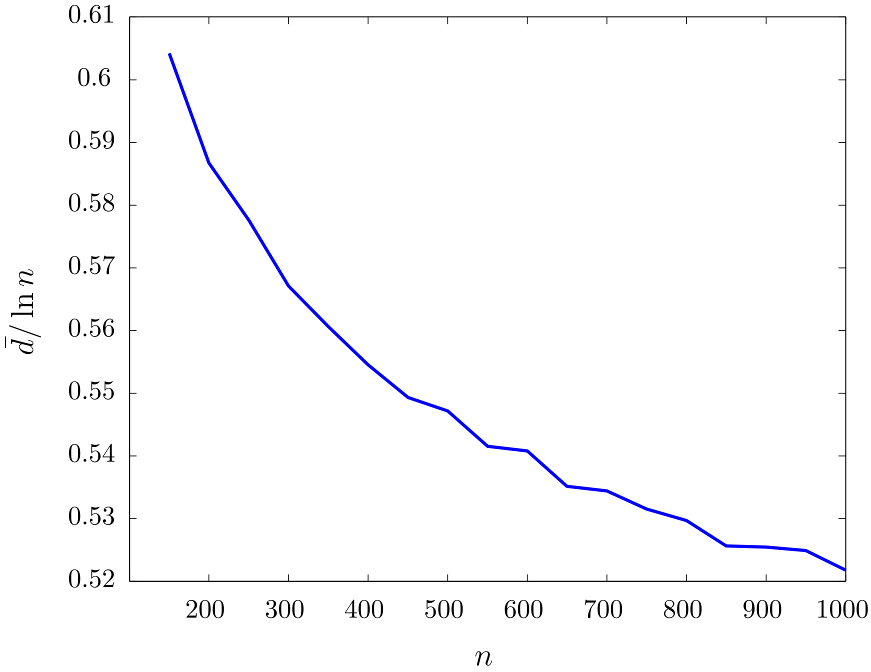

Figure 5: Small World of HRN as indicated by Simulation 2 (created with MATLAB, 2014)

6 An Example of the Labor Market

While in the previous sections, a theoretical analysis of the suggested Homophilous Random Network model is presented, we now provide one possible economic application. In recent years, more and more research in the field of labor economics has been dedicated to understanding the mechanisms of different hiring channels. One of these channels, which is commonly used in reality, relies on the contacts of current employees. Starting with the seminal contribution of Montgomery (1991), a lot of researchers decided to model connections between workers as a social network (see e.g. Calvó-Armengol, 2004; Calvó-Armengol and Jackson, 2007; Dawid and Gemkow, 2014).[18] As known from the extensive sociological literature (see Section 1), in these social networks, one should expect to observe homophily with respect to skills or competence, performance, education, level of income, and geographical distance. While there are lots of empirical studies confirming the existence of homophily in workers' social contacts and analyzing the implications thereof (see e.g. Mayer and Puller, 2008; Rees, 1966), only few work has yet been dedicated to developing theoretical models capturing this effect.[19]

[18] For an extensive survey including both empirical and theoretic literature from sociology and economics see Ioannides and Loury (2004).

[19] Exceptions are Horváth (2014), van der Leij and Buhai (2008) and Zaharieva (2013), however, all using binary notions of homophily.

In our application, we consider a risk-neutral firm that plans to fill an open vacancy. Two possible hiring channels are available. On the one hand, there is the formal job market and, on the other hand, the possibility to hire a contact of its current employee. Based on the model introduced in Section 2, we consider n workers and a vector of characteristics p capturing the ability of each worker to do the vacant job. W.l.o.g. we assume that agent 1 is the current employee of the firm while all other agents $2, ..., n$ are supposed to be available on the job market. While we fix p_1 as a parameter of the model, meaning that the firm knows the ability of its current employee, $p_{-1} = (p_2, .., p_n)$ is again considered as a realization of the $(n-1)$-dimensional random variable P_{-1}. Given this situation and based on individual linking probabilities (1) for parameters $\lambda, a \in (0,1)$, we assume that a homophilous random network forms.

Knowing the distribution function of the random variable P_{-1} and the conditional linking probabilities but not the realization, the firm has to decide on one hiring channel. For this purpose, the expected characteristic of a contact of agent 1 is the crucial statistic. It can be calculated as follows.[20]

Proposition 4. *Given some homophily parameter $a \in (0,1)$, the expected characteristic of a neighbor $j \in \{2, ..., n\}$ of agent 1 with given characteristic $p_1 \in [0,1]$ is*

$$\mathbb{E}^P \left[P_j \mid G \in \mathbb{G}_{1j} \right] = \frac{1}{2} + \frac{(a^{p_1} - a^{1-p_1})(\frac{1}{2} - \frac{1}{\ln(a)}) + 2p_1 - 1}{2 - a^{p_1} - a^{1-p_1}}. \tag{10}$$

In Figure 6, the expected characteristic of an agent's neighbor as in (10) is plotted as a function of $p_1 \in [0,1]$ and $a \in (0,1)$. However, an analytical investigation reveals some intuitive properties, at least for some special cases. These might contribute to a better understanding of the rather complicated functional form and its appearance. We collect these properties in the following corollary. Note that all of them can be detected in Figure 6.

Corollary 5. *Function (10) in Proposition 4 yields:*

(i) $\mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}]\Big|_{p_1 = \frac{1}{2}} = \frac{1}{2} \quad \forall a \in (0,1)$,

(ii) $\lim_{a \to 0} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] = p_1 \quad \forall p_1 \in [0,1]$, *and*

(iii) $\lim_{a \to 1} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] = \frac{1}{2} \quad \forall p_1 \in [0,1]$.

[20] Note that this probability is meaningful only if agent 1 has at least one link. For large networks, however, this is guaranteed whenever the corresponding condition of the threshold theorem (recall Proposition 3) is fulfilled.

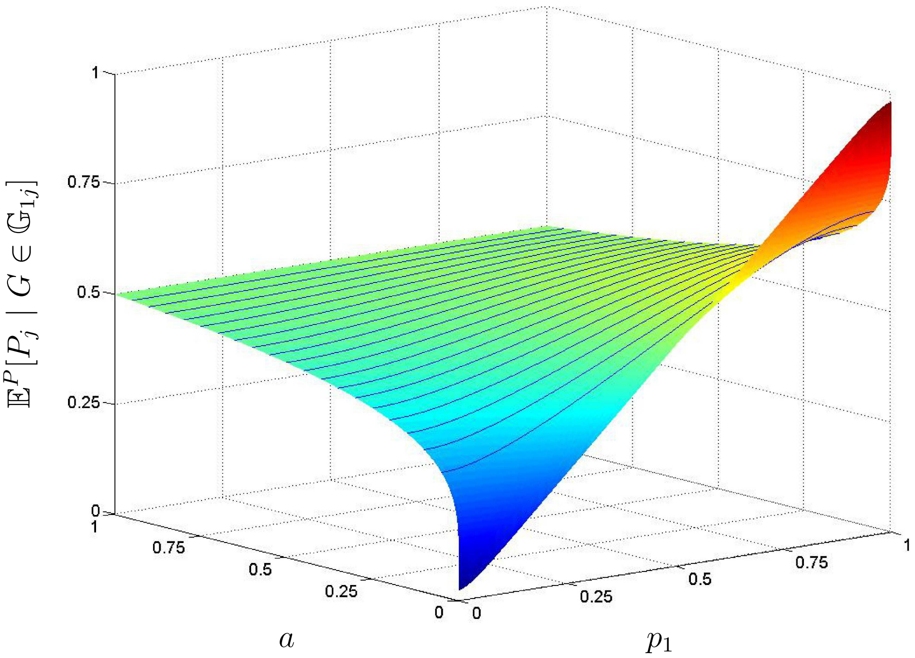

Figure 6: Expected characteristic of a neighbor of agent 1 with realized characteristic $p_1 \in [0, 1]$, given homophily level $a \in (0, 1)$ (created with MATLAB, 2014)

If, for simplicity, one assumes that the expected characteristic or rather ability of a worker hired via the formal job market is some value $\bar{p} \in (0, 1)$ which is independent of the homophily parameter a and the ability of the current employee p_1. Given this situation, the firm faces a simple decision rule when to hire via the social network. We have that, for sufficiently high p_1 and low a, respectively, the expected ability of the current employee's contact exceeds any ability level $\bar{p}$. More precisely, for any parameter value $a \in (0, 1)$, solving the equation $\mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] = \bar{p}$ yields a minimum ability level p_1 (if existing at this homophily level) that has to be reached for the expected ability of the current employee's contact to exceed $\bar{p}$. Similarly, given $p_1 \in [0, 1]$, we obtain a maximum level of a, that is a minimum level of homophily. Thus, the decision rule is that the firm should hire a randomly chosen contact instead of recruiting via the formal job market if and only if the respective calculated minimum level is exceeded (or at least reached).

Finally, note that this would still hold, at least qualitatively, if one would consider the best contact of agent 1, that is the neighbor j with maximal p_j instead of the neighbors' average ability. Certainly, the adapted minimum levels of homophily and current employee's ability (see above) would be smaller in this case, meaning that it would be optimal for the firm to hire via the social network for an even broader range of parameter combinations (a, p_1).

7 Discussion and Conclusion

In this paper, we set up a novel Homophilous Random Network model incorporating heterogeneity of agents. In a two-stage random process, first each agent (or vertex) is assigned a one-dimensional characteristic. Second, based on these realized characteristics, the links of a random network form whilst taking into account a continuous notion of homophily. This captures the frequently observed propensity of individuals to connect with similar others. Exploiting this continuous formalization of homophily, our approach allows for a broad range of homophily levels ranging from the extreme case of maximal homophily where only equal agents get linked with positive probability up to the case where there is no homophily at all. The latter case corresponds to the Bernoulli Random Graph (BRG) model, often referred to as the Erdős-Rényi model. Insofar, our model can also be regarded as a generalization thereof. Most importantly, unlike the vast majority of related economic models, we indeed capture homophily as it is defined and used in the sociological literature, namely in terms of similarity rather than equality.

In our work, we first reveal some basic properties and network statistics of the Homophilous Random Network model and establish a threshold theorem. The comparison with the BRG model provides additional insight. To derive one of our main results, we focus on another stylized fact of real-world networks, namely the occurrence of clustering. Although homophily and clustering are frequently observed in reality, both phenomena are not captured by the original BRG model. While revealing by simulations that the small-world phenomenon is apparently preserved, we are able to show analytically that homophily induces clustering in our model. This gives rise to the conjecture that also in reality there might be a considerable causality between the two. It might be worthwhile for future research to pursue this question. Finally, we provide an easily accessible application of our model for labor economics. Assuming homophily with respect to abilities to do a certain job, we consider workers being connected through a homophilous random network. We determine the expected ability of a given worker's random contact depending on the level of homophily and the given worker's own ability. This yields a simple decision rule for a firm which intends to fill an open vacancy and needs to decide whether to hire through a current employee's contacts or the formal job market.

Our Homophilous Random Network model is now available as a tool which can be used to understand and predict diffusion processes in social networks. As it complies with those important stylized facts which we frequently observe in social networks, it might yield meaningful results, for instance, regarding the spread of information or a disease.

Beyond that, there are certainly several further questions which remain open for future research. Although our simulation results yield a strong indication in this direction, one task would be to show analytically that the small-world phenomenon is generally preserved in our model. As a second point, it could be of interest to expand our considerations about threshold theorems and to establish those for different properties such as connectedness in our model. Also, a calibration of the model to real-world data is yet to be done. Performing this in a meaningful way is most certainly a challenge, especially as the level of homophily in a given network is not clearly observable. However, one way to deal with this could be to calibrate the model to the observable degree of clustering which we showed to be directly connected to homophily in our model.

Further, it would be a natural, yet analytically challenging extension to check the qualitative robustness of our findings for different distributions of characteristics. For many applications, a distribution that puts more weight on intermediate characteristics might represent reality more accurately. For instance, this could be captured in our model by drawing agents' characteristics from an appropriate beta distribution. In fact, we found that replacing the uniform distribution by a beta distribution with different shape parameters, does not change the results of our simulations qualitatively. In particular, choosing a combination of shape parameters such that it is scarcer for agents to have extreme characteristics, entails that the average distance is still increasing in homophily, however less strongly. This is interesting as it confirms our explanation that the increase of the average distance in homophily is due to the fact that, in our model, agents with extreme characteristics become relatively less connected as homophily increases. Furthermore, an extension of our model to multi-dimensional characteristics would be valuable, in particular if one would succeed to combine characteristics of both continuous and binary nature.

Finally, within our labor market application, one could pursue the idea as it is outlined at the end of Section 6, that is to calculate the expected maximum characteristic of a given agent's neighbor. In this way, one could determine the firm's gains from giving bonuses to its current employee for recommending her best instead of a random neighbor.

A Proofs

A.1 Proof of Proposition 1

We calculate the expected probability:

$$\begin{aligned}
\mathbb{E}^P\left[\mathbb{P}^G\left(ij \in G \mid P\right) \mid P_i = p_i\right] &= \mathbb{E}^P\left[\lambda a^{|P_i - P_j|} \mid P_i = p_i\right] \\
&= \lambda\Big(\int_0^1 \underbrace{f_{P_j}(p_j)}_{1} a^{|p_i - p_j|} dp_j\Big) \\
&= \lambda\left(\int_0^{p_i} a^{p_i - p_j} dp_j + \int_{p_i}^1 a^{p_j - p_i} dp_j\right) \\
&= \lambda\left(a^{p_i}\int_0^{p_i} a^{-p_j} dp_j + a^{-p_i}\int_{p_i}^1 a^{p_j} dp_j\right) \\
&= \lambda\Big(a^{p_i}\frac{1 - a^{-p_i}}{\ln(a)} + a^{-p_i}\frac{a - a^{p_i}}{\ln(a)}\Big) \\
&= \frac{\lambda}{\ln(a)}\left(a^{p_i} + a^{1-p_i} - 2\right). \qquad (11)
\end{aligned}$$

Moreover, by integrating equation (11) with respect to p_i, we get the expected probability if p is unknown:

$$\begin{aligned}
\mathbb{E}^P\left[\mathbb{P}^G\left[ij \in G \mid P\right]\right] &= \mathbb{E}^P\left[\lambda a^{|P_i - P_j|}\right] \\
&= \lambda\Big(\int_{[0,1]^2} \underbrace{f_{P_i,P_j}(p_i, p_j)}_{=f_{P_i}(p_i) f_{P_j}(p_j)=1} a^{|p_i - p_j|} d(p_i, p_j)\Big) \\
&\overset{(11)}{=} \lambda\Big(\int_0^1 \frac{(a^{p_i} + a^{1-p_i} - 2)}{\ln(a)} dp_i\Big) \\
&= \frac{\lambda}{\ln(a)}\left[\frac{a^{p_i} - a^{1-p_i} - 2p_i\ln(a)}{\ln(a)}\right]\Bigg|_{p_i=0}^{p_i=1} \\
&= \frac{\lambda}{\ln(a)^2}\left[a - 1 - 2\ln(a) - 1 + a\right] \\
&= \frac{2\lambda}{\ln(a)^2}\left[a - 1 - \ln(a)\right].
\end{aligned}$$

□

A.2 Proof of Corollary 1

Using l'Hôpital's rule, we calculate the limit of φ as

$$\begin{aligned}\lim_{a\to 0}\varphi(\lambda,a,p_i) = \lim_{a\to 0}\frac{\lambda(a^{p_i}+a^{1-p_i}-2)}{\ln(a)} &= \lim_{a\to 0}\frac{\lambda(p_i a^{p_i-1}+(1-p_i)a^{-p_i})}{1/a}\\ &= \lim_{a\to 0}\lambda(p_i a^{p_i}+(1-p_i)a^{1-p_i}) = 0.\end{aligned}$$

Similarly, we get

$$\begin{aligned}\lim_{a\to 1}\varphi(\lambda,a,p_i) = \lim_{a\to 1}\frac{\lambda(a^{p_i}+a^{1-p_i}-2)}{\ln(a)} &= \lim_{a\to 1}\frac{\lambda(p_i a^{p_i-1}+(1-p_i)a^{-p_i})}{1/a}\\ &= \lim_{a\to 1}\lambda(p_i a^{p_i}+(1-p_i)a^{1-p_i}) = \lambda.\end{aligned}$$

For the case of Φ, by now using l'Hôpital's rule twice, we get

$$\lim_{a\to 0}\Phi(\lambda,a) = \lim_{a\to 0}2\lambda\frac{a-1-\ln(a)}{\ln(a)^2} = \lim_{a\to 0}2\lambda\frac{1-1/a}{2\ln(a)/a} = \lim_{a\to 0}\lambda\frac{a-1}{\ln(a)} = 0$$

as well as

$$\lim_{a\to 1}\Phi(\lambda,a) = \lim_{a\to 1}2\lambda\frac{a-1-\ln(a)}{\ln(a)^2} = \lim_{a\to 1}2\lambda\frac{a-1}{2\ln(a)} = \lim_{a\to 1}\lambda\frac{1}{1/a} = \lambda.$$

□

A.3 Proof of Proposition 2

Taking into account equation (2), we calculate

$$\begin{aligned}&\mathbb{E}^P\left[\mathbb{P}^G\left(\eta_i(G)=k \mid P\right) \mid P_i = p_i\right]\\ &= \mathbb{E}^P\left[\sum_{K\subseteq N\setminus\{i\}:|K|=k}\left(\prod_{j\in K}(q(P_i,P_j))\cdot\prod_{l\in N\setminus K\setminus\{i\}}(1-q(P_i,P_l))\right) \mid P_i = p_i\right]\\ &= \sum_{K\subseteq N\setminus\{i\}:|K|=k}\left(\mathbb{E}^P\left[\prod_{j\in K}(q(P_i,P_j))\cdot\prod_{l\in N\setminus K\setminus\{i\}}(1-q(P_i,P_l)) \mid P_i = p_i\right]\right)\\ &= \sum_{K\subseteq N\setminus\{i\}:|K|=k}\left(\int_{[0,1]^{n-1}}\left(\underbrace{f_{P_{-i}}(p_{-i})}_{=1}\cdot\prod_{j\in K}(q(p_i,p_j))\cdot\prod_{l\in N\setminus K\setminus\{i\}}(1-q(p_i,p_l))\right)dp_{-i}\right)\\ &= \sum_{K\subseteq N\setminus\{i\}:|K|=k}\left(\prod_{j\in K}\left(\int_0^1(q(p_i,p_j))\,dp_j\right)\cdot\prod_{l\in N\setminus K\setminus\{i\}}\left(\int_0^1(1-q(p_i,p_l))\,dp_l\right)\right)\end{aligned}$$

$$\overset{(2)}{=} \sum_{K\subseteq N\setminus\{i\}:|K|=k} \left(\left(\frac{\lambda}{\ln(a)} \left(a^{p_i} + a^{1-p_i} - 2 \right) \right)^k \cdot \left(1 - \frac{\lambda}{\ln(a)} \left(a^{p_i} + a^{1-p_i} - 2 \right) \right)^{n-k-1} \right)$$
$$\overset{(2)}{=} \binom{n-1}{k} \cdot (\varphi(\lambda, a, p_i))^k \cdot (1 - \varphi(\lambda, a, p_i))^{n-k-1}.$$

□

A.4 Proof of Proposition 3

The probability that an agent i with given characteristic p_i is isolated is

$$\mathbb{E}^P \left[\mathbb{P}^G \left(\eta_i(G) = 0 \mid P \right) \mid P_i = p_i \right] \overset{(9)}{=} (1 - \varphi(\lambda(n), a(n), p_i))^{n-1}.$$

If we assume that there is at least some homophily as the size of the network becomes large, that is formally

$$\exists\ \tilde{\epsilon} > 0,\ \bar{n} \in \mathbb{N} :\ a(n) \leq 1 - \tilde{\epsilon} \quad \forall\ n \geq \bar{n},$$

then we have that

$$\exists\ \epsilon > 0 :\ 2 - a(n)^{\hat{p}} - a(n)^{1-\hat{p}} \in [\epsilon, 2] \quad \forall\ n \geq \bar{n}.$$

Now it holds that if $\lim_{n\to\infty}[-\lambda(n)/(\ln(a(n))t(n))] = \infty$, then we have

$$\begin{aligned}
&\lim_{n\to\infty} (1 - \varphi(\lambda(n), a(n), p_i))^{n-1} \\
&= \lim_{n\to\infty} \left(1 - \frac{\varphi(\lambda(n), a(n), p_i)/t(n)}{n-1} \right)^{n-1} \\
&\overset{(2)}{=} \lim_{n\to\infty} \left(1 - \frac{\frac{\lambda(n)(n-1)}{\ln(a(n))}(a(n)^{p_i} + a(n)^{1-p_i} - 2)}{n-1} \right)^{n-1} \\
&= \lim_{n\to\infty} \exp \left(\underbrace{-\frac{\lambda(n)(n-1)}{\ln(a(n))}}_{\to\infty} \underbrace{(a(n)^{p_i} + a(n)^{1-p_i} - 2)}_{\in[-2,-\epsilon]} \right) = 0.
\end{aligned}$$

On the contrary, if $\lim_{n\to\infty}[-\lambda(n)/(\ln(a(n))t(n))] = 0$, then we get

$$\begin{aligned}
&\lim_{n\to\infty} (1-\varphi(\lambda(n), a(n), p_i))^{n-1} \\
&= \lim_{n\to\infty} \exp\left(\underbrace{-\frac{\lambda(n)(n-1)}{\ln(a(n))}}_{\to 0} \underbrace{(a(n)^{p_i} + a(n)^{1-p_i} - 2)}_{\in[-2,-\epsilon]} \right) = 1.
\end{aligned}$$

□

A.5 Proof of Theorem 1

We calculate the clustering coefficient

$$\begin{aligned}
&C(\lambda, a) \\
&= \mathbb{E}^P \left[\lambda a^{|P_j - P_k|} \;\middle|\; G \in \mathbb{G}_{ij} \cap \mathbb{G}_{ik} \right] \\
&= \lambda \int_{[0,1]^n} a^{|p_j - p_k|} f_P(p \mid G \in \mathbb{G}_{ij} \cap \mathbb{G}_{ik}) dp \\
&= \lambda \int_{[0,1]^n} a^{|p_j - p_k|} \frac{f_{P,G}(p, \mathbb{G}_{ij} \cap \mathbb{G}_{ik})}{f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik})} dp \\
&= \frac{\lambda}{f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik})} \int_{[0,1]^n} a^{|p_j - p_k|} f_{P,G}(p, \mathbb{G}_{ij} \cap \mathbb{G}_{ik}) dp \\
&= \frac{\lambda}{f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik})} \int_{[0,1]^n} a^{|p_j - p_k|} f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik} \mid P = p) \overbrace{f_P(p)}^{=1} dp \\
&= \frac{\lambda}{\int_{[0,1]^n} \underbrace{f_P(x)}_{=1} f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik} \mid P = x) dx} \int_{[0,1]^n} a^{|p_j - p_k|} f_G(\mathbb{G}_{ij} \cap \mathbb{G}_{ik} \mid P = p) dp \\
&= \frac{\lambda}{\int_{[0,1]^n} \mathbb{P}^G(G \in \mathbb{G}_{ij} \cap \mathbb{G}_{ik} \mid P = x) dx} \int_{[0,1]^n} a^{|p_j - p_k|} \mathbb{P}^G(G \in \mathbb{G}_{ij} \cap \mathbb{G}_{ik} \mid P = p) dp \\
&= \frac{\lambda}{\int_{[0,1]^n} \lambda a^{|x_i - x_j|} \lambda a^{|x_i - x_k|} dx} \int_{[0,1]^n} a^{|p_j - p_k|} \lambda a^{|p_i - p_j|} \lambda a^{|p_i - p_k|} dp \\
&= \lambda \frac{\int_{[0,1]^n} a^{|p_j - p_k| + |p_i - p_j| + |p_i - p_k|} dp}{\int_{[0,1]^n} a^{|x_i - x_j| + |x_i - x_k|} dx} = \lambda \frac{\int_{[0,1]^3} a^{|p_j - p_k| + |p_i - p_j| + |p_i - p_k|} d(p_i, p_j, p_k)}{\int_{[0,1]^3} a^{|x_i - x_j| + |x_i - x_k|} d(x_i, x_j, x_k)}. \qquad (12)
\end{aligned}$$

Let us solve the integral in the denominator first. For the sake of readability denote $x = (x_i, x_j, x_k)$. We have

$$\begin{aligned}
\int_{[0,1]^3} a^{|x_i - x_j| + |x_i - x_k|} dx &= \int_{\substack{x \in [0,1]^3: \\ x_j, x_k \leq x_i}} a^{2x_i - x_j - x_k} dx + \int_{\substack{x \in [0,1]^3: \\ x_i \leq x_j, x_k}} a^{x_j + x_k - 2x_i} dx \\
&\quad + \int_{\substack{x \in [0,1]^3: \\ x_j \leq x_i \leq x_k}} a^{x_k - x_j} dx + \int_{\substack{x \in [0,1]^3: \\ x_k \leq x_i \leq x_j}} a^{x_j - x_k} dx \\
&= \frac{2\ln(a) - 4a + a^2 + 3}{2(\ln(a))^3} + \frac{2\ln(a) - 4a + a^2 + 3}{2(\ln(a))^3} \\
&\quad + \frac{2\ln(a) - 4a + 2a\ln(a) + 4}{2(\ln(a))^3} + \frac{2\ln(a) - 4a + 2a\ln(a) + 4}{2(\ln(a))^3} \\
&= \frac{1}{2(\ln(a))^3} \Big[8\ln(a) - 16a + 2a^2 + 4\ln(a)a + 14 \Big].
\end{aligned}$$

Next, we solve the integral in the numerator of (12), substituting x for p in order to use the same notation as above. This yields

$$\begin{aligned}
&\int_{[0,1]^3} a^{|x_j - x_k| + |x_i - x_j| + |x_i - x_k|} dx \\
&= \int_{\substack{x \in [0,1]^3: \\ x_i \leq x_j \leq x_k}} a^{2x_k - 2x_i} dx + \int_{\substack{x \in [0,1]^3: \\ x_i \leq x_k \leq x_j}} a^{2x_j - 2x_i} dx + \int_{\substack{x \in [0,1]^3: \\ x_j \leq x_i \leq x_k}} a^{2x_k - 2x_j} dx \\
&\quad + \int_{\substack{x \in [0,1]^3: \\ x_j \leq x_k \leq x_i}} a^{2x_i - 2x_j} dx + \int_{\substack{x \in [0,1]^3: \\ x_k \leq x_i \leq x_j}} a^{2x_j - 2x_k} dx + \int_{\substack{x \in [0,1]^3: \\ x_k \leq x_j \leq x_i}} a^{2x_i - 2x_k} dx \\
&= 6 \frac{\ln(a) - a^2 + a^2\ln(a) + 1}{4(\ln(a))^3} = \frac{1}{2(\ln(a))^3} \Big[3\ln(a) - 3a^2 + 3a^2\ln(a) + 3 \Big].
\end{aligned}$$

Taken together, this gives

$$C(\lambda, a) = \lambda \frac{3\ln(a) - 3a^2 + 3a^2\ln(a) + 3}{8\ln(a) - 16a + 2a^2 + 4\ln(a)a + 14}.$$

By using this, we can now start with the actual proof. We have

$$\begin{aligned}
&C(\lambda, a) - \Phi(\lambda, a) \\
&= \lambda \left(\frac{3\Big(\ln(a)a^2 + \ln(a) - a^2 + 1\Big)}{2\Big(2\ln(a)a + 4\ln(a) + a^2 - 8a + 7\Big)} + \frac{2\Big(\ln(a) - a + 1\Big)}{\ln(a)^2} \right) \\
&= \lambda \tfrac{3\ln(a)^3(a^2+1) + \ln(a)^2(-3a^2+8a+19) + \ln(a)(-4a^2-40a+44) + (-4a^3+36a^2-60a+28)}{2\ln(a)^2(2\ln(a)a + 4\ln(a) + a^2 - 8a + 7)}.
\end{aligned} \tag{13}$$

In what follows, we use that for $a \in (0,1)$ we have

$$\ln(a) = -\sum_{m=0}^{\infty} \frac{(1-a)^{m+1}}{m+1}$$

which implies that $\ln(a) < -\sum_{m=0}^{M} \frac{(1-a)^{m+1}}{m+1} < 0$ for all $M \in \mathbb{N}$. The first and easier part is to show that the denominator of the term on the right-hand side of equation (13) is negative for all $a \in (0,1)$. We calculate

$$\begin{aligned}
&2\ln(a)a + 4\ln(a) + a^2 - 8a + 7 \\
=\ & 2(a+2)\ln(a) + a^2 - 8a + 7 \\
<\ & -2(a+2)\Big(1 - a + \frac{1}{2}(1-a)^2 + \frac{1}{3}(1-a)^3\Big) + a^2 - 8a + 7 \\
=\ & \frac{1}{3}(a+2)\Big(2a^3 - 9a^2 + 18a - 11\Big) + a^2 - 8a + 7 \\
=\ & \frac{1}{3}\Big(2a^4 - 5a^3 + 3a^2 + a - 1\Big) = -\frac{1}{3}(1-a)^3(2a+1) < 0.
\end{aligned}$$

Further, we define

$$\begin{aligned}
g(a) := {} & 3\ln(a)^3(a^2+1) + \ln(a)^2(-3a^2 + 8a + 19) + \ln(a)(-4a^2 - 40a + 44) \\
& + (-4a^3 + 36a^2 - 60a + 28).
\end{aligned}$$

Then $\lambda g(a)$ is the numerator of the term on the right-hand side of equation (13). We calculate the derivatives

$$\begin{aligned}
\tfrac{dg}{da}(a) &= \tfrac{1}{a}\ [6\ln(a)^3a^2 + \ln(a)^2(3a^2 + 8a + 9) + 2\ln(a)(-7a^2 - 12a + 19) \\
&\quad +4(-3a^3 + 17a^2 - 25a + 11)], \\
\tfrac{d^2g}{da^2}(a) &= \tfrac{1}{a^2}[6\ln(a)^3a^2 + 3\ln(a)^2(7a^2 - 3) + 4\ln(a)(-2a^2 + 4a - 5) \\
&\quad +6(-4a^3 + 9a^2 - 4a - 1)], \\
\tfrac{d^3g}{da^3}(a) &= \tfrac{1}{a^3}[18\ln(a)^2(a^2 + 1) + 2\ln(a)(21a^2 - 8a + 11) + 8(-3a^3 - a^2 + 5a - 1)], \\
\tfrac{d^4g}{da^4}(a) &= \tfrac{1}{a^4}[18\ln(a)^2(-a^2 - 3) + 2\ln(a)(-3a^2 + 16a - 15) + 2(25a^2 - 48a + 23)], \\
\tfrac{d^5g}{da^5}(a) &= \tfrac{1}{a^5}[36\ln(a)^2(a^2 + 6) + 12\ln(a)(-2a^2 - 8a + 1) + 2(-53a^2 + 160a - 107)], \\
\tfrac{d^6g}{da^6}(a) &= \tfrac{1}{a^6}[108\ln(a)^2(-a^2 - 10) + 12\ln(a)(12a^2 + 32a + 31) + 2(147a^2 - 688a + 541)].
\end{aligned}$$

Notice here that

$$g(1) = \frac{dg}{da}(1) = \frac{d^2g}{da^2}(1) = \frac{d^3g}{da^3}(1) = \frac{d^4g}{da^4}(1) = \frac{d^5g}{da^5}(1) = 0.$$

Moreover, we have

$$\begin{aligned}\frac{d^6 g}{da^6}(a) &= \frac{1}{a^6}\Big[108\ln(a)^2\underbrace{(-a^2-10)}_{<0}+12\ln(a)\underbrace{(12a^2+32a+31)}_{>0}\\ &\qquad +2(147a^2-688a+541)\Big]\\ &< \frac{1}{a^6}\Big[108(1-a)^2(-a^2-10)-12(1-a)(12a^2+32a+31)\\ &\qquad +2(147a^2-688a+541)\Big]\\ &= \frac{2}{a^6}\Big[-54a^4+180a^3-327a^2+386a-185\Big]\\ &= \frac{2}{a^6}(1-a)\Big[54(a-\tfrac{7}{9})^3+103(a-\tfrac{7}{9})-\tfrac{2146}{27}\Big]\\ &< \frac{2}{a^6}(1-a)\Big[54\cdot(\tfrac{2}{9})^3+103\cdot\tfrac{2}{9}-\tfrac{2146}{27}\Big]=-\frac{112}{a^6}(1-a)<0.\end{aligned}$$

Combining this, it follows for all $a \in (0,1)$ that

$$\begin{aligned}\frac{d^5 g}{da^5}(a)>0 \Rightarrow \frac{d^4 g}{da^4}(a)<0 \Rightarrow \frac{d^3 g}{da^3}(a)>0 \Rightarrow \frac{d^2 g}{da^2}(a)<0 \Rightarrow \frac{dg}{da}(a)>0\\ \Rightarrow\ g(a)<0.\end{aligned}$$

Taken together, we have indeed that

$$C(\lambda,a)-\Phi(\lambda,a)=\lambda\frac{g(a)}{2\ln(a)^2\Big(2\ln(a)a+4\ln(a)+a^2-8a+7\Big)}>0$$

which concludes the proof of the theorem. □

A.6 Proof of Corollary 4

By applying l'Hôpital's rule three times, we calculate

$$\begin{aligned}
\lim_{a\to 0} C(\lambda, a) &= \lambda \lim_{a\to 0} \frac{3\ln(a) - 3a^2 + 3a^2\ln(a) + 3}{8\ln(a) - 16a + 2a^2 + 4\ln(a)a + 14} \\
&= \lambda \lim_{a\to 0} \frac{3/a - 6a + 6a\ln(a) + 3a}{8/a - 16 + 4a + 4\ln(a) + 4} \\
&= \frac{3\lambda}{4} \lim_{a\to 0} \frac{1 - a^2 + 2a^2\ln(a)}{2 - 3a + a^2 + a\ln(a)} \\
&= \frac{3\lambda}{4} \frac{\lim_{a\to 0}[1 - a^2 + 2a^2\ln(a)]}{\lim_{a\to 0}[2 - 3a + a^2 + a\ln(a)]} \\
&= \frac{3\lambda}{4} \frac{\lim_{a\to 0}[1] - \lim_{a\to 0}[a^2] + \lim_{a\to 0}[2a^2\ln(a)]}{\lim_{a\to 0}[2] - \lim_{a\to 0}[3a] + \lim_{a\to 0}[a^2] + \lim_{a\to 0}[a\ln(a)]} \\
&= \frac{3\lambda}{4} \frac{1 - 0 + \lim_{x\to\infty}[2\ln(1/x)/x^2]}{2 - 0 + 0 + \lim_{x\to\infty}[\ln(1/x)/x]} \\
&= \frac{3\lambda}{4} \frac{1 + \lim_{x\to\infty}[-2x(1/x^2)/2x]}{2 + \lim_{x\to\infty}[-x(1/x^2)/1]} = \frac{3\lambda}{4} \frac{1 + \lim_{x\to\infty}[-1/x^2]}{2 + \lim_{x\to\infty}[-1/x]} = \frac{3\lambda}{8}.
\end{aligned}$$

The stated result follows immediately since we established in Corollary 1 that $\lim_{a\to 0} \Phi(\lambda, a) = 0$. On the contrary, by again using l'Hôpital's rule three times, we get

$$\begin{aligned}
\lim_{a\to 1} C(\lambda, a) &= \lambda \lim_{a\to 1} \frac{3\ln(a) - 3a^2 + 3a^2\ln(a) + 3}{8\ln(a) - 16a + 2a^2 + 4\ln(a)a + 14} \\
&= \lambda \lim_{a\to 1} \frac{3/a - 6a + 6a\ln(a) + 3a}{8/a - 16 + 4a + 4\ln(a) + 4} \\
&= \lambda \lim_{a\to 1} \frac{3 - 3a^2 + 6a^2\ln(a)}{8 - 12a + 4a^2 + 4a\ln(a)} \\
&= \lambda \lim_{a\to 1} \frac{-6a + 12a\ln(a) + 6a}{-12 + 8a + 4\ln(a) + 4} = \lambda \lim_{a\to 1} \frac{12\ln(a) + 12}{8 + 4/a} = \lambda.
\end{aligned}$$

According to Corollary 1, we have $\lim_{a\to 1} \Phi(\lambda, a) = \lambda$ which concludes the proof.

□

A.7 Proof of Proposition 4

We calculate

$$\begin{aligned}
\mathbb{E}^P\left[P_j \mid G \in \mathbb{G}_{1j}\right] = \int_0^1 p_j f_{P_j|G}(p_j, \mathbb{G}_{1j}) dp_j &= \int_0^1 p_j f_{P_j}(p_j \mid G \in \mathbb{G}_{1j}) dp_j \\
&= \int_0^1 p_j \frac{f_{P_j,G}(p_j, \mathbb{G}_{1j})}{f_G(\mathbb{G}_{1j})} dp_j \\
&= \int_0^1 p_j \frac{f_G(\mathbb{G}_{1j} \mid P_j = p_j) \overbrace{f_{P_j}(p_j)}^{1}}{f_G(\mathbb{G}_{1j})} dp_j \\
&= \int_0^1 p_j \frac{f_G(\mathbb{G}_{1j} \mid P_j = p_j)}{\int_0^1 \underbrace{f_{P_j}(x)}_{1} \underbrace{f_G(\mathbb{G}_{1j} \mid P_j = x)}_{\mathbb{P}(G \in \mathbb{G}_{1j} \ \mid \ P_j = x)} dx} dp_j \\
&= \int_0^1 p_j \frac{\overbrace{f_G(\mathbb{G}_{1j} \mid P_j = p_j)}^{\lambda a^{|p_1 - p_j|}}}{\underbrace{\int\limits_0^1 \lambda a^{|p_1 - x|} dx}_{\frac{\lambda}{\ln(a)}(a^{p_1} + a^{1-p_1} - 2)}} dp_j \\
&= \frac{\ln(a)}{a^{p_1} + a^{1-p_1} - 2} \int_0^1 p_j a^{|p_1 - p_j|} dp_j .
\end{aligned}$$

Focusing on the integral first gives

$$\begin{aligned}
\int_0^1 p_j a^{|p_1 - p_j|} dp_j &= \int_0^{p_1} p_j a^{(p_1 - p_j)} dp_j + \int_{p_1}^1 p_j a^{(p_j - p_1)} dp_j \\
&= \frac{a^{p_1} - p_1 \ln(a) - 1}{\ln(a)^2} + \frac{a^{1-p_1}(\ln(a) - 1) - p_1 \ln(a) + 1}{\ln(a)^2} .
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}^P(P_j \mid G \in \mathbb{G}_{1j}) &= \frac{a^{p_1} + a^{1-p_1}(\ln(a) - 1) - 2p_1 \ln(a)}{\ln(a)(a^{p_1} + a^{1-p_1} - 2)} && \text{(14a)} \\
&= \frac{1}{2} + \frac{(a^{p_1} - a^{1-p_1})(\frac{1}{2} - \frac{1}{\ln(a)}) + 2p_1 - 1}{2 - a^{p_1} - a^{1-p_1}} . && \text{(14b)}
\end{aligned}$$

□

A.8 Proof of Corollary 5

Considering the functional form (10), we prove the properties in question one after the other. Regarding Part (i), by using equation (14b) we calculate for $a \in (0, 1)$

that

$$\mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}]\Big|_{p_1=\frac{1}{2}} = \frac{1}{2} + \frac{(\sqrt{a}-\sqrt{a})(\frac{1}{2}-\frac{1}{\ln(a)})+1-1}{2-\sqrt{a}-\sqrt{a}} = \frac{1}{2}.$$

Next, we consider Part (ii). Again applying equation (14b), we get

$$\lim_{a\to 0} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] = \frac{1}{2} + \frac{(0-0)(\frac{1}{2}+0)+2p_1-1}{2-0-0} = p_1$$

for $p_1 \in (0,1)$ and for the marginals we have

$$\begin{aligned}\lim_{a\to 0} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}]\Big|_{p_1=0} &= \frac{1}{2} + \frac{(1-0)(\frac{1}{2}+0)+0-1}{2-1-0} = 0,\\ \lim_{a\to 0} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}]\Big|_{p_1=1} &= \frac{1}{2} + \frac{(0-1)(\frac{1}{2}+0)+2-1}{2-0-1} = 1.\end{aligned}$$

To establish Part (iii), we have to apply l'Hôpital's rule. For $p_1 \in [0,1]$ we get

$$\begin{aligned}\lim_{a\to 1} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] &\overset{(14a)}{=} \lim_{a\to 1} \frac{a^{p_1}+a^{1-p_1}(\ln(a)-1)-2p_1\ln(a)}{\ln(a)(a^{p_1}+a^{1-p_1}-2)}\\ &= \lim_{a\to 1} \frac{p_1a^{p_1-1}+(1-p_1)a^{-p_1}(\ln(a)-1)+a^{-p_1}-\frac{2p_1}{a}}{\frac{1}{a}(a^{p_1}+a^{1-p_1}-2)+\ln(a)(p_1a^{p_1-1}+(1-p_1)a^{-p_1})} \end{aligned} \tag{15}$$

while using l'Hôpital's rule once. However, we obviously need to apply it a second time. For this purpose, we calculate the derivatives of the numerator and denominator of the term on the right-hand side in equation (15). We get

$$\begin{aligned}&\frac{\partial}{\partial a}\Big[p_1a^{p_1-1}+(1-p_1)a^{-p_1}(\ln(a)-1)+a^{-p_1}-\frac{2p_1}{a}\Big]\\ &=p_1(p_1-1)a^{p_1-2}+p_1(p_1-1)a^{-p_1-1}(\ln(a)-1)+(1-p_1)a^{-p_1-1}-p_1a^{-p_1-1}+\frac{2p_1}{a^2}\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial}{\partial a}\Big[\frac{1}{a}(a^{p_1}+a^{1-p_1}-2)+\ln(a)(p_1a^{p_1-1}+(1-p_1)a^{-p_1})\Big]\\ &=-\frac{1}{a^2}(a^{p_1}+a^{1-p_1}-2)+\frac{2}{a}(p_1a^{p_1-1}+(1-p_1)a^{-p_1})\\ &\quad+\ln(a)(p_1(p_1-1)a^{p_1-2}+p_1(p_1-1)a^{-p_1-1}).\end{aligned}$$

By recalling equation (15) and using l'Hôpital's rule the second time, this gives

$$\lim_{a\to 1} \mathbb{E}^P[P_j \mid G \in \mathbb{G}_{1j}] = \frac{p_1(p_1-1)+p_1(p_1-1)(0-1)+(1-p_1)-p_1+2p_1}{-(1+1-2)+2(p_1+(1-p_1))+0} = \frac{1}{2}$$

which concludes the proof. □

Acknowledgements

The authors would like to thank Michael Günther, Tim Hellmann, Christoph Kuzmics, Dunia López-Pintado, Agnieszka Rusinowska, Emily Tanimura, Fernando Vega-Redondo, and anonymous reviewers, as well as the participants of various seminars at Bielefeld University and University Paris 1 for valuable comments and suggestions. This research was carried out within the International Research Training Group "Economic Behavior and Interaction Models" (EBIM) financed by the German Research Foundation (DFG) under contract GRK 1134/2.

References

Baerveldt, C., Van Duijn, M. A., Vermeij, L., and Van Hemert, D. A. (2004). Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. *Social Networks*, 26(1):55–74.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Bloznelis, M. (2013). Degree and clustering coefficient in sparse random intersection graphs. *The Annals of Applied Probability*, 23(3):1254–1289.

Bollobás, B. (1998). *Modern Graph Theory.* Springer, New York, NY.

Bollobás, B. (2001). *Random Graphs.* Cambridge University Press, Cambridge, second edition.

Bramoullé, Y., Currarini, S., Jackson, M. O. ., Pin, P., and Rogers, B. W. (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory*, 147(5):1754–1786.

Burt, R. S. (1991). Measuring age as a structural concept. *Social Networks*, 13(1):1–34.

Calvó-Armengol, A. (2004). Job contact networks. *Journal of Economic Theory*, 115(1):191–206.

Calvó-Armengol, A. and Jackson, M. O. (2007). Networks in labor markets: Wage and employment dynamics and inequality. *Journal of Economic Theory*, 132(1):27–46.

Campbell, K. E. (1990). Networks past: A 1939 Bloomington neighborhood. *Social Forces*, 69(1):139–155.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882.

Currarini, S., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045.

Dawid, H. and Gemkow, S. (2014). How do social networks contribute to wage inequality? Insights from an agent-based analysis. *Industrial and Corporate Change*, 23(5):1171–1200.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6(1):290–297.

Gilles, R. P. and Johnson, C. (2000). Spatial social networks. *Review of Economic Design*, 5(3):273–299.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338.

Horváth, G. (2014). Occupational mismatch and social networks. *Journal of Economic Behavior & Organization*, 106:442–468.

Ibarra, H. (1995). Race, opportunity, and diversity of social circles in managerial networks. *Academy of Management Journal*, 38(3):673–703.

Iijima, R. and Kamada, Y. (2014). Social distance and network structures. Working Paper.

Ioannides, Y. M. (1997). Evolution of trading structures. In Arthur, W. B., Durlauf, S. N., and Lane, D. A., editors, *The Economy as an Evolving Complex System II*, pages 129–167. Addison-Wesley, Boston, MA.

Ioannides, Y. M. and Loury, L. D. (2004). Job information networks, neighborhood effects, and inequality. *Journal of Economic Literature*, 42(4):1056–1093.

Jackson, M. O. (2006). The economics of social networks. In Blundell, R., Newey, W. K., and Persson, T., editors, *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Volume I*, pages 1–56. Cambridge University Press, Cambridge.

Jackson, M. O. (2008a). Average distance, diameter, and clustering in social networks with homophily. In Papadimitriou, C. and Zhang, S., editors, *Internet and Network Economics*, pages 4–11. Springer, Berlin Heidelberg.

Jackson, M. O. (2008b). *Social and Economic Networks.* Princeton University Press, Princeton, NJ.

Kalmijn, M. (2006). Educational inequality and family relationships: Influences on contact and proximity. *European Sociological Review*, 22(1):1–16.

Karonski, M., Scheinerman, E. R., and Singer-Cohen, K. B. (1999). On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8(1):131–159.

Laumann, E. O. (1966). *Prestige and Association in an Urban Community: An Analysis of an Urban Stratification System.* Bobbs-Merrill, Indianapolis, IN.

Laumann, E. O. (1973). *Bonds of Pluralism: The Form and Substance of Urban Social Networks.* John Wiley & Sons, New York, NY.

Lazarsfeld, P. F. and Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In Berger, M., editor, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York, NY.

Marsden, P. V. (1987). Core discussion networks of Americans. *American Sociological Review*, 52(1):122–131.

Marsden, P. V. (1988). Homogeneity in confiding relations. *Social Networks*, 10(1):57–76.

MATLAB (2014). *Version 8.3 (R2014a).* The MathWorks Inc., Natick, MA.

Mayer, A. and Puller, S. L. (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 92(1):329–347.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1):60–67.

Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *American Economic Review*, 81(5):1408–1418.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. (2006). Random graphs as models of networks. In Bornholdt, S. and Schuster, H. G., editors, *Handbook of Graphs and Networks: From the Genome to the Internet*, pages 35–68. Wiley-VCH, Weinheim.

Rees, A. (1966). Information networks in labor markets. *American Economic Review*, 56(1):559–566.

van der Leij, M. and Buhai, S. (2008). A social network analysis of occupational segregation. Working Paper, Available at SSRN 1117949.

Verbrugge, L. M. (1977). The structure of adult friendship choices. *Social Forces*, 56(2):576–597.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika*, 61(3):401–425.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Wellman, B. (1996). Are personal communities local? A Dumptarian reconsideration. *Social Networks*, 18(4):347–354.

Zaharieva, A. (2013). Social welfare and wage inequality in search equilibrium with personal contacts. *Labour Economics*, 23(1):107–121.