Metrika (2012) 75:761–793
DOI 10.1007/s00184-011-0351-y

The distribution of the relative arc density of a family of interval catch digraph based on uniform data

Elvan Ceyhan

Received: 21 March 2010 / Published online: 27 February 2011

Abstract We study a family of interval catch digraph called proportional-edge proximity catch digraph (PCD) which is also a special type of intersection digraphs parameterized with an expansion and a centrality parameter. PCDs are random catch digraphs that have been developed recently and have applications in classification and spatial pattern analysis. We investigate a graph invariant of the PCDs called relative arc density. We demonstrate that relative arc density of PCDs is a U-statistic and using the central limit theory of U-statistics, we derive the (asymptotic) distribution of the relative arc density of proportional-edge PCD for uniform data in one dimension. We also determine the parameters for which the rate of convergence to asymptotic normality is fastest.

Keywords Class cover catch digraph · Intersection digraph · Proximity catch digraph · Proximity map · Random graph · U-statistics

1 Introduction

The proximity catch digraphs (PCDs) were motivated by their applications in pattern classification and spatial pattern analysis, which have received considerable attention in the statistical literature. The proximity catch digraphs (PCDs) were motivated by their applications in these areas. In this article, the distribution of a graph invariant called *relative arc density* of the PCDs is investigated. The PCDs are vertex-random digraphs in which each vertex corresponds to a data point, and directed edges (i.e., arcs) are defined by some bivariate relation on the data. For example, nearest neighbor

E. Ceyhan (✉)
Department of Mathematics, Koç University, 34450 Sarıyer, Istanbul, Turkey
e-mail: elceyhan@ku.edu.tr

digraphs are defined by placing an arc between each vertex and its nearest neighbor. The PCDs are a special type of proximity graphs which were introduced by Toussaint (1980). Furthermore, the PCDs are closely related to the class cover problem of Cannon and Cowen (2000).

Priebe et al. (2001) introduced the class cover catch digraphs (CCCDs) in $\mathbb{R}$ and gave the exact and the asymptotic distribution of the domination number of the CCCDs based on data from two classes, say $\mathcal{X}$ and $\mathcal{Y}$, with uniform distribution on a bounded interval in $\mathbb{R}$. DeVinney et al. (2002), Marchette and Priebe (2003), Priebe et al. (2003a,b), and DeVinney and Priebe (2006) applied the concept in higher dimensions and demonstrated relatively good performance of CCCDs in classification.

The first PCD family is introduced by Ceyhan and Priebe (2003); the parameterized version of this PCD is later developed by Ceyhan et al. (2007) where the relative arc density of the PCD is calculated and used for testing spatial patterns in $\mathbb{R}^2$. Ceyhan and Priebe (2005) introduced another digraph family called *proportional-edge PCDs* and calculated the asymptotic distribution of its domination number in $\mathbb{R}^2$ and used it for the same purpose (Ceyhan and Priebe 2007; Ceyhan 2010). The relative arc density of this PCD family is also computed and used in spatial pattern analysis in $\mathbb{R}^2$ (Ceyhan et al. 2006).

Properly scaled, the relative arc density of the proportional-edge PCDs is also a U-statistic, which has asymptotic normality by the general central limit theory of U-statistics. In this article, we consider the proportional-edge PCDs for one dimensional data, where proportional-edge PCD has an expansion and a centrality parameter. We derive the explicit form of the asymptotic normal distribution of the relative arc density of the proportional-edge PCDs for uniform one dimensional $\mathcal{X}$ points whose support being partitioned by the class $\mathcal{Y}$ points. The asymptotic distribution of the relative arc density is derived for the entire ranges of the expansion and centrality parameters based on detailed calculations. The relative arc density of proportional-edge PCDs is first investigated for uniform data in one interval (in $\mathbb{R}$) and the analysis is generalized to uniform data in multiple intervals. These results will be of use in applying the relative arc density for testing interaction between classes of one dimensional data. Moreover, the behavior of the relative arc density in one dimensional case will form the foundations of our investigation and extension of the topic in higher dimensions.

We define the proximity catch digraphs and their relative arc density in Sect. 2, describe the proportional-edge PCD and provide preliminary results on their relative arc density in Sect. 3, provide the distribution of the relative arc density for uniform data in one interval in Sect. 4, in multiple intervals in Sect. 5, the extension of the proportional-edge PCD to higher dimensions in Sect. 6, and provide discussion and conclusions in Sect. 7. Shorter proofs are given in the main body of the article; while longer proofs are deferred to the Appendix sections.

2 Relative arc density of the proximity catch digraphs

Let $D_n = (\mathcal{V}, \mathcal{A})$ be a digraph with vertex set $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ and arc set $\mathcal{A}$ and let $|\cdot|$ stand for the set cardinality function. The relative arc density of the digraph D_n which is of order $|\mathcal{V}| = n \geq 2$, denoted $\rho(D_n)$, is defined as Janson et al. (2000)

$$\rho(D_n) = \frac{|\mathcal{A}|}{n(n-1)}.$$

Thus $\rho(D_n)$ represents the ratio of the number of arcs in the digraph D_n to the number of arcs in the complete symmetric digraph of order n, which is $n(n-1)$. For $n \leq 1$, we set $\rho(D_n) = 0$, since there are no arcs. If D_n is a random digraph in which arcs result from a random process, then the *arc probability* between vertices v_i, v_j is defined as $p_a(i, j) := P((v_i, v_j) \in \mathcal{A})$ for all $i \neq j, i, j = 1, 2, \ldots, n$.

Let $(\Omega, \mathcal{M})$ be a measurable space and $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, Y_2, \ldots, Y_m\}$ be two sets of Ω-valued random variables from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint probability distribution $F_{X,Y}$ and marginals F_X and F_Y, respectively. A PCD is comprised by a set $\mathcal{V}$ of vertices and a set $\mathcal{A}$ of arcs. For example, in the two class case, with classes $\mathcal{X}$ and $\mathcal{Y}$, we choose the $\mathcal{X}$ points to be the vertices and put an arc from $X_i \in \mathcal{X}_n$ to $X_j \in \mathcal{X}_n$, based on a binary relation which measures the relative allocation of X_i and X_j with respect to $\mathcal{Y}$ points. Consider the map $N : \Omega \to \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ represents the power set of Ω. Then given $\mathcal{Y}_m \subseteq \Omega$, the *proximity map* $N(\cdot)$ associates with each point $x \in \Omega$ a *proximity region* $N(x) \subseteq \Omega$. For $B \subseteq \Omega$, the Γ_1-region is the image of the map $\Gamma_1(\cdot, N) : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that associates the region $\Gamma_1(B, N) := \{z \in \Omega : B \subseteq N(z)\}$ with the set B. For a point $x \in \Omega$, we denote $\Gamma_1(\{x\}, N)$ as $\Gamma_1(x, N)$. Notice that while the proximity region is defined for one point, a Γ_1-region is defined for a point or set of points. The *vertex-random PCD* has the vertex set $\mathcal{V} = \mathcal{X}_n$ and arc set $\mathcal{A}$ defined by $(X_i, X_j) \in \mathcal{A}$ if $X_j \in N(X_i)$. Given $\mathcal{Y}_m = \{y_1, y_2, \ldots, y_m\}$, let $\mathcal{X}_n$ be a random sample from F_X. Then $N(X_i)$ are also iid and the same holds for $\Gamma_1(X_i, N)$. Hence $p_a(i, j) := P((X_i, X_j) \in \mathcal{A}) = p_a$ for all $i \neq j, i, j = 1, 2, \ldots, n$ for such $\mathcal{X}_n$.

Theorem 1 *Given $\mathcal{Y}_m = \{y_1, y_2, \ldots, y_m\}$, let $\mathcal{X}_n$ be a random sample from F_X and D_n be the PCD based on $N(\cdot)$ with vertices $\mathcal{X}_n$ and the arc set $\mathcal{A}$ is defined as $(X_i, X_j) \in \mathcal{A}$ if $X_j \in N(X_i)$. The relative arc density, $\rho(D_n)$, of D_n is a one-sample U-statistic of degree 2 and is an unbiased estimator of p_a. If, additionally, $\nu = \mathbf{Cov}\left[h_{ij}, h_{ik}\right] > 0$ for all $i \neq j \neq k, i, j, k \in \{1, 2, \ldots, n\}$, then $\sqrt{n}\,[\rho(D_n) - p_a] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\nu)$ as $n \to \infty$, where $2h_{ij} = \mathbf{I}((X_i, X_j) \in \mathcal{A}) + \mathbf{I}((X_j, X_i) \in \mathcal{A})$, and $\xrightarrow{\mathcal{L}}$ stands for convergence in law and $\mathcal{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2.*

Proof Let $g_{ij} = \mathbf{I}((X_i, X_j) \in \mathcal{A}) = \mathbf{I}(X_j \in N(X_i))$. The arcs $(X_i, X_j) \in \mathcal{A}$ and $(X_j, X_i) \in \mathcal{A}$ are different for $i \neq j$, so g_{ij} is not symmetric in i, j. But we can define a symmetric kernel as $h_{ij} = (g_{ij} + g_{ji})/2$. Then we have, $|\mathcal{A}| = \sum_{i<j} h_{ij}$. So

$$\rho(D_n) = \frac{1}{\binom{n}{2}} \sum \sum_{i<j} h_{ij} \tag{1}$$

Thus, $\rho(D_n)$ is a one-sample U-statistic of degree 2 with symmetric kernel h_{ij}. Moreover, $P((X_i, X_j) \in \mathcal{A}) = p_a$ for all $i \neq j, i, j = 1, 2, \ldots, n$. Then for $i \neq j$, we have

$$\begin{aligned}\mathbf{E}[h_{ij}] &= \mathbf{E}[(g_{ij}+g_{ji})/2] = (\mathbf{E}[g_{ij}]+\mathbf{E}[g_{ji}])/2 = \mathbf{E}[g_{ij}] = \mathbf{E}[g_{12}]\\ &= P((X_1,X_2)\in\mathcal{A}) = P(X_2\in N(X_1)) = p_a.\end{aligned}$$

Hence p_a is an estimable parameter of degree 2. Furthermore,

$$\begin{aligned}\mathbf{E}[\rho(D_n)] &= \frac{2}{n(n-1)}\mathbf{E}[|\mathcal{A}|] = \frac{2}{n(n-1)}\sum_{i<j}\mathbf{E}[h_{ij}] = \frac{2}{n(n-1)}\sum_{i<j}\mathbf{E}[g_{ij}]\\ &= \frac{2}{n(n-1)}\sum_{i<j}p_a = p_a. \qquad (2)\end{aligned}$$

Then, $\rho(D_n)$ is actually an unbiased estimator of the arc probability p_a.

For PCDs, the set of vertices $\mathcal{V} = \mathcal{X}_n$ is a random sample from a distribution F_X (i.e., the vertices directly result from a random process), and the arcs are defined based on the random sets (i.e., proximity regions) $N(X_i)$ as described before. Hence the set of arcs $\mathcal{A}$ (indirectly) results from a random process such that g_{ij} are identically distributed and g_{ij} and g_{kl} are independent for $i \neq k$ and $j \neq l$ and $\{i,j\} \neq \{k,l\}$.

Additionally, $\mathbf{Cov}(g_{ij}, g_{kl}) = \mathbf{E}[g_{ij}g_{kl}] - p_a^2 < \infty$, since $\mathbf{E}[g_{ij}g_{kl}] = P((g_{ij}, g_{kl}) = (1,1))$. Hence $\nu = \mathbf{Cov}(h_{ij}, h_{ik}) = \mathbf{Cov}((g_{ij}+g_{ji})/2, (g_{ik}+g_{ki})/2) = (\mathbf{Cov}[g_{ij}, g_{ik}] + \mathbf{Cov}[g_{ij}, g_{ki}] + \mathbf{Cov}[g_{ji}, g_{ik}] + \mathbf{Cov}[g_{ji}, g_{ki}])/4 < \infty$ as well. Then by Theorem 3.3.13 in Randles and Wolfe (1979), we have $\sqrt{n}\,[\rho(D_n) - p_a] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\nu)$ as $n \to \infty$, provided that $\nu > 0$. □

Recall that $2\,h_{ij} = (g_{ij}+g_{ji}) = \mathbf{I}(X_j \in N(X_i)) + \mathbf{I}(X_i \in N(X_j))$ is the number of arcs between X_i and X_j in D_n. For $X_i \overset{iid}{\sim} F_X$, $i = 1, 2, \ldots, n$, $\rho(D_n)$ is a random variable that depends on n, F, and $N(\cdot)$ (i.e., $\mathcal{Y}_m$). But $\mathbf{E}[\rho(D_n)] = \mathbf{E}[h_{12}] = p_a$ only depends on F and $N(\cdot)$. Furthermore,

$$0 \leq \mathbf{Var}[\rho(D_n)] = \frac{2}{n(n-1)}\mathbf{Var}[h_{12}] + \frac{4(n-2)}{n(n-1)}\mathbf{Cov}[h_{12}, h_{13}] \qquad (3)$$

where the variance is

$$\begin{aligned}\mathbf{Var}[h_{ij}] &= \mathbf{Var}[h_{12}] = \mathbf{E}[(h_{ij})^2] - (\mathbf{E}[h_{ij}])^2 = \mathbf{E}[(g_{ij}+g_{ji})^2/4] - p_a^2\\ &= (\mathbf{E}[g_{ij}] + 2\,\mathbf{E}[g_{ij}]\mathbf{E}[g_{ji}] + \mathbf{E}[g_{ji}])/4 - p_a^2 = (p_a + 2\,p_a + p_a)/4 - p_a^2\\ &= (p_a - p_a^2)/2 = p_a\,(1-p_a)\,/2\end{aligned}$$

and the covariance is

$$\mathbf{Cov}[h_{12}, h_{13}] = \mathbf{E}[h_{12}h_{13}] - \mathbf{E}[h_{12}]\,\mathbf{E}[h_{13}] = \mathbf{E}[h_{12}h_{13}] - p_a^2,$$

with

$$\begin{aligned}4\,\mathbf{E}\,[h_{12}h_{13}] &= \mathbf{E}[(g_{12}+g_{21})(g_{13}+g_{31})] = \mathbf{E}[g_{12}g_{13}+g_{12}g_{31}+g_{21}g_{13}+g_{21}g_{31}]\\ &= \mathbf{E}[\mathbf{I}(X_2 \in N(X_1))\,\mathbf{I}(X_3 \in N(X_1)) + \mathbf{I}(X_2 \in N(X_1))\,\mathbf{I}(X_1 \in N(X_3))\\ &\quad +\mathbf{I}(X_1 \in N(X_2))\,\mathbf{I}(X_3 \in N(X_1))] + \mathbf{I}(X_1 \in N(X_2))\,\mathbf{I}(X_1 \in N(X_3))]\\ &= \mathbf{E}[\mathbf{I}(\{X_2, X_3\} \subset N(X_1)) + \mathbf{I}(X_2 \in N(X_1))\,\mathbf{I}(X_3 \in \Gamma_1(X_3, N))\\ &\quad +\mathbf{I}(X_2 \in \Gamma_1(X_1))\,\mathbf{I}(X_3 \in N(X_1))]\\ &\quad +\mathbf{I}(X_2 \in \Gamma_1(X_1, N))\,\mathbf{I}(X_3 \in \Gamma_1(X_1, N))]\\ &= P(\{X_2, X_3\} \subset N(X_1)) + 2\,P(X_2 \in N(X_1), X_3 \in \Gamma_1(X_1, N))\\ &\quad +P(\{X_2, X_3\} \subset \Gamma_1(X_1, N)).\end{aligned}$$

Then $\nu = \mathbf{Cov}(h_{ij}, h_{ik}) = \mathbf{E}[h_{ij}h_{ik}] - \mathbf{E}[h_{ij}]\mathbf{E}[h_{ik}] = \mathbf{E}[h_{ij}h_{ik}] - p_a^2 = \mathbf{E}[h_{12}h_{13}] - p_a^2 > 0$ iff

$$\begin{aligned}P(\{X_2, X_3\} \subset N(X_1)) + 2\,P(X_2 \in N(X_1), X_3 \in \Gamma_1(X_1, N))\\ +P(\{X_2, X_3\} \subset \Gamma_1(X_1, N)) > 4p_a^2.\end{aligned}$$

Notice also that $\mathbf{E}[|h_{ij}|^3] < \infty$ since $\mathbf{E}[|h_{ij}|^3] \leq 1$. Then for $\nu > 0$, the sharpest rate of convergence in the asymptotic normality of $\rho(D_n)$ is

$$\sup_{t\in\mathbb{R}} \left| P\left(\frac{\sqrt{n}(\rho(D_n) - p_a)}{\sqrt{4\,\nu}} \leq t\right) - \Phi(t)\right| \leq 8\,K\,p_a\,(4\,\nu)^{-3/2}\,n^{-1/2} = K\,\frac{p_a}{\sqrt{n\,\nu^3}} \tag{4}$$

where K is a constant and $\Phi(t)$ is the cumulative distribution function for standard normal distribution (Callaert and Janssen 1978).

In general a random digraph, just like a random graph, can be obtained by starting with a set of n vertices and adding arcs between them at random. We can consider the counterpart of the Erdős–Rényi model for digraphs, denoted $D(n, p)$, in which every possible arc occurs independently with probability p (Erdős and Rényi 1959). Notice that the random digraph $D(n, p)$, satisfies the conditions of Theorem 1, so the relative arc density of $D(n, p)$ is a U-statistic; however, the asymptotic distribution of its relative arc density is degenerate (with $\rho(D(n, p)) \xrightarrow{\mathcal{L}} p$, as $n \to \infty$) since the covariance term is zero due to the independence between the arcs.

3 Proportional-edge PCDs for one dimensional data

Let $\Omega = \mathbb{R}$ and $Y_{(i)}$ be the ith order statistic of $\mathcal{Y}_m$ for $i = 1, 2, \ldots, m$. Assume $Y_{(i)}$ values are distinct (which happens with probability one for continuous distributions). Then $Y_{(i)}$ values partition $\mathbb{R}$ into $(m + 1)$ intervals. Let

$$-\infty =: Y_{(0)} < Y_{(1)} < \cdots < Y_{(m)} < Y_{(m+1)} := \infty.$$

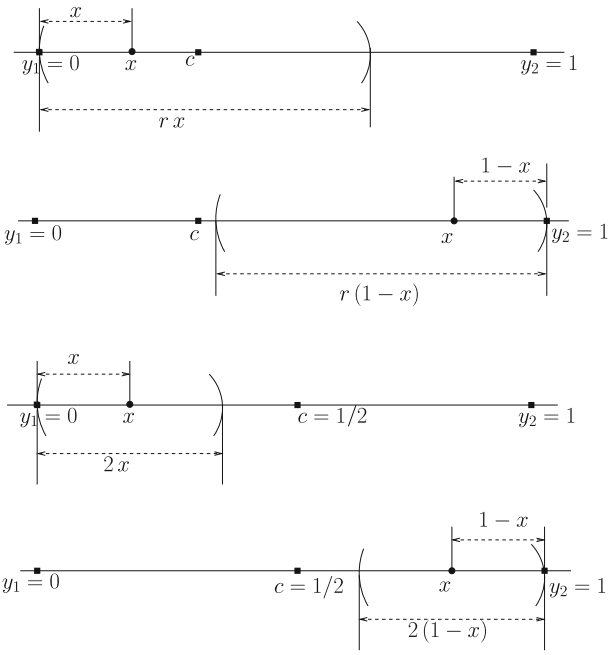

Fig. 1 Plotted in the *top two rows* are illustrations of the construction of proportional-edge proximity region, $N(x, r, c)$ for $\mathcal{Y}_2 = \{y_1, y_2\}$ with $y_1 = 0$ and $y_2 = 1$ (hence $M_c = c$) and $x \in (0, c)$ (*top*) and $x \in (c, 1)$ (*bottom*); and in the bottom two rows are for the proximity regions associated with CCCD, i.e., $N(x, r = 2, c = 1/2)$ for an $x \in (0, 1/2)$ (*top*) and $x \in (1/2, 1)$ (*bottom*)

We call intervals $\left(-\infty, Y_{(1)}\right)$ and $\left(Y_{(m)}, \infty\right)$ the *end intervals*, and intervals $\left(Y_{(i-1)}, Y_{(i)}\right)$ for $i = 2, \ldots, m$ the *middle intervals*. Then we define the proportional-edge PCD with the parameter $r \geq 1$ for two one dimensional data sets, $\mathcal{X}_n$ and $\mathcal{Y}_m$, from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively, as follows. For $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{2, \ldots, m\}$ (i.e., for x in a middle interval) and $M_c \in \left(Y_{(i-1)}, Y_{(i)}\right)$ such that $c \times 100\%$ of $(Y_{(i)} - Y_{(i-1)})$ is to the left of M_c (i.e., $M_c = Y_{(i-1)} + c\,(Y_{(i)} - Y_{(i-1)})$)

$$N(x, r, c) = \begin{cases} \left(Y_{(i-1)}, Y_{(i-1)} + r\,\left(x - Y_{(i-1)}\right)\right) \bigcap \left(Y_{(i-1)}, Y_{(i)}\right) & \text{if } x \in (Y_{(i-1)}, M_c), \\ \left(Y_{(i)} - r\,\left(Y_{(i)} - x\right), Y_{(i)}\right) \bigcap \left(Y_{(i-1)}, Y_{(i)}\right) & \text{if } x \in \left(M_c, Y_{(i)}\right). \end{cases} \tag{5}$$

For an illustration of $N(x, r, c)$ in the middle interval case, see also Fig. 1 where $\mathcal{Y}_2 = \{y_1, y_2\}$ with $y_1 = 0$ and $y_2 = 1$ (hence $M_c = c$).

Additionally, for $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{1, m + 1\}$ (i.e., for x in an end interval), the proportional-edge proximity region only has an expansion parameter, but not a centrality parameter. Hence we let $N_e(x, r)$ be the proportional-edge proximity region for an x in an end interval.

$$N_e(x, r) = \begin{cases} \left(Y_{(1)} - r\,\left(Y_{(1)} - x\right), Y_{(1)}\right) & \text{if } x < Y_{(1)}, \\ \left(Y_{(m)}, Y_{(m)} + r\,\left(x - Y_{(m)}\right)\right) & \text{if } x > Y_{(m)}. \end{cases} \tag{6}$$

If $x \in \mathcal{Y}_m$, then we define $N(x,r,c) = \{x\}$ and $N_e(x,r) = \{x\}$ for all $r \in [1,\infty]$, and if $x = M_c$, then in Eq. (5), we arbitrarily assign $N(x,r,c)$ to be one of $\left(Y_{(i-1)}, Y_{(i-1)} + r\left(x - Y_{(i-1)}\right)\right) \cap \left(Y_{(i-1)}, Y_{(i)}\right)$ or $\left(Y_{(i)} - r\left(Y_{(i)} - x\right), Y_{(i)}\right) \cap \left(Y_{(i-1)}, Y_{(i)}\right)$. For X from a continuous distribution, these special cases in the construction of proportional-edge proximity region—$x \in \mathcal{Y}_m$ and $x = M_c$—happen with probability zero. Notice that $r > 1$ implies $x \in N(x,r,c)$ for all $x \in \left[Y_{(i-1)}, Y_{(i)}\right]$ with $i \in \{2,\ldots,m\}$ and $x \in N_e(x,r)$ for all $x \in \left[Y_{(i-1)}, Y_{(i)}\right]$ with $i \in \{1, m+1\}$. Furthermore, $\lim_{r\to\infty} N(x,r,c) = \left(Y_{(i-1)}, Y_{(i)}\right)$ (and $\lim_{r\to\infty} N_e(x,r) = \left(Y_{(i-1)}, Y_{(i)}\right)$) for all $x \in \left(Y_{(i-1)}, Y_{(i)}\right)$ with $i \in \{2,\ldots,m\}$ (and $i \in \{1, m+1\}$), so we define $N(x,\infty,c) = \left(Y_{(i-1)}, Y_{(i)}\right)$ (and $N_e(x,\infty) = \left(Y_{(i-1)}, Y_{(i)}\right)$) for all such x.

The vertex-random proportional-edge PCD has the vertex set $\mathcal{X}_n$ and arc set $\mathcal{A}$ defined by $(X_i, X_j) \in \mathcal{A} \iff X_j \in N(X_i,r,c)$ for X_i, X_j in the middle intervals and $(X_i, X_j) \in \mathcal{A} \iff X_j \in N_e(X_i,r)$ for X_i, X_j in the end intervals. We denote such digraphs as $\mathscr{D}_{n,m}(r,c)$. A $\mathscr{D}_{n,m}(r,c)$-digraph is a *pseudo digraph* according some authors, if loops are allowed (see, e.g., Chartrand and Lesniak 1996). The $\mathscr{D}_{n,m}(r,c)$-digraphs are closely related to the *proximity graphs* of Jaromczyk and Toussaint (1992) and might be considered as a special case of *covering sets* of Tuza (1994) and *intersection digraphs* of Sen et al. (1989). Our vertex-random proximity digraph is not a standard random graph (see, e.g., Janson et al. 2000). The randomness of a $\mathscr{D}_{n,m}(r,c)$-digraph lies in the fact that the vertices are random with the joint distribution $F_{X,Y}$, but arcs (X_i, X_j) are deterministic functions of the random variable X_j and the random set $N(X_i,r,c)$ in the middle intervals and the random set $N_e(X_i,r)$ in the end intervals. In $\mathbb{R}$, the vertex-random PCD is a special case of *interval catch digraphs* (see, e.g., Sen et al. 1989; Prisner 1994). Furthermore, when $r = 2$ and $c = 1/2$ (i.e., $M_c = \left(Y_{(i-1)} + Y_{(i)}\right)/2$) we have $N(x,2,1/2) = B(x,r(x))$ for an x in a middle interval and $N_e(x,2) = B(x,r(x))$ for an x in an end interval where $r(x) = d(x,\mathcal{Y}_m) = \min_{y\in\mathcal{Y}_m} d(x,y)$ and the corresponding PCD is the CCCD of Priebe et al. (2001). See also Fig. 1.

3.1 Relative arc density of random $\mathscr{D}_{n,m}(r,c)$-digraphs

Let $\mathcal{F}(\mathbb{R}) := \{F_{X,Y}$ on $\mathbb{R}$ with $P(X = Y) = 0$ and the marginal distributions, F_X and F_Y, are non-atomic$\}$. In this article, we consider $\mathscr{D}_{n,m}(r,c)$-digraphs for which $\mathcal{X}_n$ and $\mathcal{Y}_m$ are random samples from F_X and F_Y, respectively, and the joint distribution of X, Y is $F_{X,Y} \in \mathcal{F}(\mathbb{R})$. We call such digraphs as $\mathcal{F}(\mathbb{R})$-*random* $\mathscr{D}_{n,m}(r,c)$-*digraphs* and focus on the random variable $\rho(\mathscr{D}_{n,m}(r,c))$. For notational brevity, we use $\rho_{n,m}(r,c)$ instead of $\rho(\mathscr{D}_{n,m}(r,c))$. It is trivial to see that $0 \le \rho_{n,m}(r,c) \le 1$, and $\rho_{n,m}(r,c) > 0$ for nontrivial digraphs.

3.2 The distribution of the relative arc density of $\mathcal{F}(\mathbb{R})$-random $\mathscr{D}_{n,m}(r,c)$-digraphs

For an $F_{X,Y} \in \mathcal{F}(\mathbb{R})$, since the marginals are non-atomic, the order statistics are distinct with probability one. Let $\mathcal{I}_i := \left(Y_{(i-1)}, Y_{(i)}\right)$, $\mathcal{X}_{[i]} := \mathcal{X}_n \cap \mathcal{I}_i$, and $\mathcal{Y}_{[i]} := \{Y_{(i-1)}, Y_{(i)}\}$ for $i = 1, 2, \ldots, (m+1)$. Let $D_{[i]}(r,c)$ be the component of the random $\mathscr{D}_{n,m}(r,c)$-digraph induced by the pair $\mathcal{X}_{[i]}$ and $\mathcal{Y}_{[i]}$. Then we have a disconnected

digraph with subdigraphs $D_{[i]}(r,c)$ for $i=1,2,\ldots,(m+1)$ each of which might be null or itself disconnected. Let $\mathcal{A}_{[i]}$ be the arc set of $D_{[i]}(r,c)$, and $\rho_{[i]}(r,c)$ denote the relative arc density of $D_{[i]}(r,c)$; $n_i := \left|\mathcal{X}_{[i]}\right|$, and F_i be the density F_X restricted to $\mathcal{I}_i$ for $i \in \{1,2,\ldots,m+1\}$. Furthermore, let $M_c^{[i]} \in \mathcal{I}_i$ be the point so that it divides the interval $\mathcal{I}_i$ in ratios c and $1-c$ (i.e., length of the subinterval to the left of $M_c^{[i]}$ is $c \times 100\%$ of the length of $\mathcal{I}_i$) for $i \in \{2,\ldots,m\}$. Notice that for $i \in \{2,\ldots,m\}$ (i.e., middle intervals), $D_{[i]}(r,c)$ is based on the proximity region $N(x,r,c)$ and for $i \in \{1,m+1\}$ (i.e., end intervals), $D_{[i]}(r,c)$ is based on the proximity region $N_e(x,r)$. Since we have at most $m+1$ subdigraphs that are disconnected, it follows that we have at most $n_T := \sum_{i=1}^{m+1} n_i(n_i-1)$ arcs in the digraph $\mathscr{D}_{n,m}(r,c)$. Then we define the relative arc density for the entire digraph as

$$\rho_{n,m}(r,c) := \frac{|\mathcal{A}|}{n_T} = \frac{\sum_{i=1}^{m+1}|\mathcal{A}_{[i]}|}{n_T} = \frac{1}{n_T}\sum_{i=1}^{m+1}(n_i(n_i-1))\rho_{[i]}(r,c). \tag{7}$$

Since $\frac{n_i(n_i-1)}{n_T} \geq 0$ for each i and $\sum_{i=1}^{m+1}\frac{n_i(n_i-1)}{n_T} = 1$, it follows that $\rho_{n,m}(r,c)$ is a mixture of the $\rho_{[i]}(r,c)$. We study the simpler random variable $\rho_{[i]}(r,c)$ first. In the remaining of this section, the almost sure (a.s.) results follow from the fact that the marginal distributions F_X and F_Y are non-atomic.

Lemma 1 *Let $D_{[i]}(r,c)$ be the digraph induced by $\mathcal{X}$ points in the end intervals (i.e., $i \in \{1,(m+1)\}$) and $\rho_{[i]}(r,c)$ be the corresponding relative arc density. For $r \geq 1$, if $n_i \leq 1$, then $\rho_{[i]}(r,c) = 0$. For $r = 1$, we have $\rho_{[i]}(1,c) = (1/2)\,\mathbf{I}(n_i > 1)$ a.s., where $\mathbf{I}(\cdot)$ stands for the indicator function.*

Proof Let $i = m+1$ (i.e., consider the right end interval). For all $r \geq 1$, if $n_{m+1} \leq 1$, then by definition $\rho_{[m+1]}(r,c) = 0$. So, we assume $n_{m+1} > 1$. Let $\mathcal{X}_{[m+1]} = \{Z_1, Z_2, \ldots, Z_{n_{m+1}}\}$ and $Z_{(j)}$ be the corresponding order statistics. There is an arc from $Z_{(j)}$ to each $Z_{(k)}$ for $k < j$, with $j,k \in \{1,2,\ldots,n_{m+1}\}$ (and possibly to some other Z_l), since $N\left(Z_{(j)},r,c\right) = (Y_{(m)}, Y_{(m)} + r\,(Z_{(j)} - Y_{(m)}))$ and so $Z_{(k)} \in N\left(Z_{(j)},r,c\right)$. So, there are at least $0+1+2+\cdots+n_{m+1}-1 = n_{m+1}(n_{m+1}-1)/2$ arcs in $D_{[m+1]}(r,c)$. Then $\rho_{[m+1]}(r,c) \geq (n_{m+1}(n_{m+1}-1)/2)/(n_{m+1}(n_{m+1}-1)) = 1/2$. By symmetry, the same results hold for $i = 1$. For $r = 1$ and $n_{m+1} > 1$, and $i = m+1$, there is an arc from $Z_{(j)}$ to each $Z_{(k)}$ for $k < j$, $j,k \in \{1,2,\ldots,m+1\}$ (and no arcs to other Z_l). So, there are $0+1+2+\cdots+n_{m+1}-1 = n_{m+1}(n_{m+1}-1)/2$ arcs in $D_{[m+1]}(1,c)$. Then $\rho_{[i]}(1,c) = (n_{m+1}(n_{m+1}-1)/2)/(n_{m+1}(n_{m+1}-1)) = 1/2$. By symmetry, the same results hold for $i = 1$. □

Using Lemma 1, we have the following lower bound for $\rho_{n,m}(r,c)$ and exact result for $\rho_{n,m}(1,c)$.

Theorem 2 *Let $D_{n,m}(r,c)$ be an $\mathcal{F}(\mathbb{R})$-random $\mathscr{D}_{n,m}(r,c)$-digraph with $n > 0$, $m > 0$ and k_1 and k_2 be two natural numbers defined as $k_1 := \sum_{i=2}^{m}(n_{i,1}(n_{i,1}-1)/2 + n_{i,2}(n_{i,2}-1)/2)$ and $k_2 := \sum_{i\in\{1,m+1\}} n_i(n_i-1)/2$, where $n_{i,1} := \left|\mathcal{X}_n \cap \left(Y_{(i)}, M_c^{[i]}\right)\right|$ and $n_{i,2} := \left|\mathcal{X}_n \cap \left(M_c^{[i]}, Y_{(i+1)}\right)\right|$. Then for $r \geq 1$, we have $(k_1+k_2)/n_T \leq \rho_{n,m}(r,c) \leq 1$ a.s. and for $r = 1$, we have $\rho_{n,m}(1,c) = (k_1+k_2)/n_T$ a.s.*

Proof For $i \in \{1, (m+1)\}$, we have k_2 as in Lemma 1. Let $i \in \{2, 3, \ldots, m\}$ and

$$\mathcal{X}_{i,1} := \mathcal{X}_{[i]} \cap \left(Y_{(i-1)}, M_c^{[i]}\right) = \{U_1, U_2, \ldots, U_{n_{i,1}}\},$$

and

$$\mathcal{X}_{i,2} := \mathcal{X}_{[i]} \cap \left(M_c^{[i]}, Y_{(i)}\right) = \{V_1, V_2, \ldots, V_{n_{i,2}}\}.$$

Furthermore, let $U_{(j)}$ and $V_{(k)}$ be the corresponding order statistics. There is an arc from $U_{(j)}$ to $U_{(k)}$ for $k < j$, $j, k \in \{1, 2, \ldots, n_{i,1}\}$ and possibly to some other U_l, and similarly there is an arc from $V_{(j)}$ to $V_{(k)}$ for $k > j$, $j, k \in \{1, 2, \ldots, n_{i,2}\}$ and possibly to some other V_l. Thus, there are at least $\frac{n_{i,1}(n_{i,1}-1)}{2} + \frac{n_{i,2}(n_{i,2}-1)}{2}$ arcs in $D_{[i]}(r, c)$. Hence $\rho_{n,m}(r, c) \geq (k_1 + k_2)/n_T$. For $r = 1$, we have $\frac{n_{i,1}(n_{i,1}-1)}{2} + \frac{n_{i,2}(n_{i,2}-1)}{2}$ many arcs in $D_{[i]}(1, c)$. Hence $\rho_{n,m}(1, c) = (k_1 + k_2)/n_T$. □

Theorem 3 *For $i = 1, 2, 3, \ldots, m+1$, $r = \infty$, and $n_i > 0$, we have $\rho_{[i]}(r = \infty, c) = \mathbf{I}(n_i > 1)$ and $\rho_{n,m}(r = \infty, c) = 1$ a.s.*

Proof For $r = \infty$, if $n_i \leq 1$, then $\rho_{[i]}(r = \infty, c) = 0$. So we assume $n_i > 1$ and let $i = m + 1$. Then $N_e(x, \infty) = \left(Y_{(m)}, \infty\right)$ for all $x \in \left(Y_{(m)}, \infty\right)$. Hence $D_{[m+1]}(\infty, c)$ is a complete symmetric digraph of order n_{m+1}, which implies $\rho_{[m+1]}(r = \infty, c) = 1$. By symmetry, the same holds for $i = 1$. For $i \in \{2, 3, \ldots, m\}$ and $n_i > 1$, we have $N(x, \infty, c) = \mathcal{I}_i$ for all $x \in \mathcal{I}_i$, hence $D_{[i]}(\infty, c)$ is a complete symmetric digraph of order n_i, which implies $\rho_{[i]}(\infty, c) = 1$. Then $\rho_{n,m}(\infty, c) = \sum \frac{n_i(n_i-1)\rho_{[i]}(\infty,c)}{n_T} = 1$, since when $n_i \leq 1$, n_i has no contribution to n_T, and when $n_i > 1$, $\rho_{[i]}(\infty, c) = 1$. □

4 The distribution of the relative arc density of proportional-edge PCDs for uniform data

Let $-\infty < \delta_1 < \delta_2 < \infty$, $\mathcal{Y}_m$ be a random sample from F_Y with support $\mathcal{S}(F_Y) \subseteq (\delta_1, \delta_2)$, and $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ be a random sample from $F_X = \mathcal{U}(\delta_1, \delta_2)$, the uniform distribution on (δ_1, δ_2) so that we have $F_{X,Y} \in \mathcal{F}(\mathbb{R})$. Assuming we have the realization of $\mathcal{Y}_m$ as $\mathcal{Y}_m = \{y_1, y_2, \ldots, y_m\} = \{y_{(1)}, y_{(2)}, \ldots, y_{(m)}\}$ with $\delta_1 < y_{(1)} < y_{(2)} < \cdots < y_{(m)} < \delta_2$, we let $y_{(0)} := \delta_1$ and $y_{(m+1)} := \delta_2$. Then it follows that the distribution of X_i restricted to $\mathcal{I}_i$ is $F_X|_{\mathcal{I}_i} = \mathcal{U}(\mathcal{I}_i)$. We call such digraphs as $\mathcal{U}(\delta_1, \delta_2)$-*random $\mathcal{D}_{n,m}(r, c)$-digraphs* and provide the distribution of their relative density for the whole ranges of r and c. We present a "scale invariance" result for proportional-edge PCDs. This invariance property will simplify the notation in our subsequent analysis by allowing us to consider the special case of the unit interval $(0, 1)$.

Theorem 4 (Scale Invariance Property) *Suppose $\mathcal{X}_n$ is a set of iid random variables from $\mathcal{U}(\delta_1, \delta_2)$ where $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ is set of m distinct $\mathcal{Y}$ points in (δ_1, δ_2). Then for any $r \geq 1$, the distribution of $\rho_{[i]}(r, c)$ is independent of $\mathcal{Y}_{[i]}$ (and hence of the restricted support interval $\mathcal{I}_i$) for all $i \in \{1, 2, \ldots, m+1\}$.*

Proof Let $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ be as in the hypothesis. Any $\mathcal{U}(\delta_1, \delta_2)$ random variable can be transformed into a $\mathcal{U}(0, 1)$ random variable by $\phi(x) = (x - \delta_1)/(\delta_2 - \delta_1)$, which maps intervals $(t_1, t_2) \subseteq (\delta_1, \delta_2)$ to intervals $(\phi(t_1), \phi(t_2)) \subseteq (0, 1)$. That is, if $X \sim \mathcal{U}(\delta_1, \delta_2)$, then we have $\phi(X) \sim \mathcal{U}(0, 1)$ and $P(X \in (t_1, t_2)) = P(\phi(X) \in (\phi(t_1), \phi(t_2))$ for all $(t_1, t_2) \subseteq (\delta_1, \delta_2)$. The distribution of $\rho_{[i]}(r, c)$ is obtained by calculating such probabilities. So, without loss of generality, we can assume $\mathcal{X}_{[i]}$ is a set of iid random variables from the $\mathcal{U}(0, 1)$ distribution. That is, the distribution of $\rho_{[i]}(r, c)$ does not depend on $\mathcal{Y}_{[i]}$ and hence does not depend on the restricted support interval $\mathcal{I}_i$. □

Note that scale invariance of $\rho_{[i]}(r = \infty, c)$ follows trivially for all $\mathcal{X}_n$ from any non-atomic F_X with support in (δ_1, δ_2) with $\delta_1 < \delta_2$, since for $r = \infty$, we have $\rho_{[i]}(r = \infty, c) = 1$ a.s.

Based on Theorem 4, we may assume each $\mathcal{I}_i$ as the unit interval $(0, 1)$ for uniform data. Then the proportional-edge proximity region for $x \in (0, 1)$ with parameters $c \in (0, 1)$ and $r \geq 1$ have the following forms. If $x \in \mathcal{I}_i$ for $i \in \{2, \ldots, m\}$ (i.e., in the middle intervals), when transformed under $\phi(\cdot)$ to $(0, 1)$, we have

$$N(x, r, c) = \begin{cases} (0, r\,x) \cap (0, 1) & \text{if } x \in (0, c), \\ (1 - r\,(1 - x), 1) \cap (0, 1) & \text{if } x \in (c, 1), \end{cases} \tag{8}$$

and $N(x = c, r, c)$ is arbitrarily taken to be one of $(0, r\,x) \cap (0, 1)$ or $(1 - r\,(1 - x), 1)$. This special case of "$X = c$" happens with probability zero for continuous X.

If $x \in \mathcal{I}_1$ (i.e., in the left end interval), when transformed under $\phi(\cdot)$ to $(0, 1)$, we have $N_e(x, r) = (\max(0, 1 - r\,(1 - x)), 1)$; and if $x \in \mathcal{I}_{m+1}$ (i.e., in the right end interval), when transformed under $\phi(\cdot)$ to $(0, 1)$, we have $N_e(x, r) = (0, \min(1, r\,x))$.

Notice that each subdigraph $D_{[i]}(r, c)$ is itself a $\mathcal{U}(\mathcal{I}_i)$-random $\mathscr{D}_{n,2}(r, c)$-digraph. The distribution of the relative arc density of $D_{[i]}(r, c)$ is given in the following result as a corollary to Theorem 1.

Corollary 1 *Let $\rho_{[i]}(r, c)$ be the relative density of subdigraph $D_{[i]}(r, c)$ of the proportional-edge PCD based on uniform data in (δ_1, δ_2) where $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in (δ_1, δ_2). Then for $r \in (1, \infty)$, as $n_i \to \infty$, we have*

(i) *for $i \in \{2, \ldots, m\}$, $\sqrt{n_i}\left[\rho_{[i]}(r, c) - \mu(r, c)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu(r, c))$, where $\mu(r, c)$ is the arc probability and $\nu(r, c) = \mathbf{Cov}[h_{12}, h_{12}]$ in the middle intervals, and*

(ii) *for $i \in \{1, m + 1\}$, $\sqrt{n_i}\left[\rho_{[i]}(r, c) - \mu_e(r)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu_e(r))$, where $\mu_e(r)$ is the arc probability and $\nu_e(r) = \mathbf{Cov}[h_{12}, h_{12}]$ in the end intervals.*

Proof (i) Let $i \in \{2, \ldots, m\}$ (i.e., $\mathcal{I}_i$ is a middle interval). By the scale invariance for uniform data (see Theorem 4), a middle interval can be assumed to be the unit interval $(0, 1)$. The mean of the asymptotic distribution $\rho_{[i]}(r, c)$ is computed as follows.

$$\mathbf{E}\left[\rho_{[i]}(r, c)\right] = \mathbf{E}[h_{12}] = P(X_2 \in N(X_1, r, c)) = \mu(r, c)$$

which is the arc probability. And the asymptotic variance of $\rho_{[i]}(r, c)$ is $\mathbf{Cov}[h_{12}, h_{13}] = 4\,\nu(r, c)$. For $r \in (1, \infty)$, since $2\,h_{12} = \mathbf{I}(X_2 \in N(X_1, r, c)) + \mathbf{I}(X_1 \in N(X_2, r, c))$

is the number of arcs between X_1 and X_2 in the PCD, h_{12} tends to be high if the proximity region $N(X_1, r, c)$ is large. In such a case, h_{13} tends to be high also. That is, h_{12} and h_{13} tend to be high and low together. So, for $r \in (1, \infty)$, we have $\nu(r, c) > 0$. Hence asymptotic normality follows.

(ii) In an end interval, the mean of the asymptotic distribution $\rho_{[i]}(r, c)$ is

$$\mathbf{E}\left[\rho_{[i]}(r, c)\right] = \mathbf{E}[h_{12}] = P(X_2 \in N_e(X_1, r)) = \mu_e(r)$$

the asymptotic variance of $\rho_{[i]}(r, c)$ is $\mathbf{Cov}[h_{12}, h_{13}] = 4\,\nu_e(r)$. For $r \in (1, \infty)$, as in (i), we have $\nu_e(r) > 0$. Hence asymptotic normality follows. □

Let $P_{2N} := P(\{X_2, X_3\} \subset N(X_1, r, c))$, $P_{NG} := P(X_2 \in N(X_1, r, c), X_3 \in \Gamma_1(X_1, r, c))$, and $P_{2G} := P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, c))$. Then

$$\begin{aligned}\mathbf{Cov}[h_{12}, h_{13}] &= \mathbf{E}[h_{12}h_{13}] - \mathbf{E}[h_{12}]\mathbf{E}[h_{13}] = \mathbf{E}[h_{12}h_{13}] - \mu(r, c)^2 \\ &= (P_{2N} + 2\,P_{NG} + P_{2G})/4 - \mu(r, c)^2,\end{aligned}$$

since

$$\begin{aligned}4\,\mathbf{E}[h_{12}h_{13}] &= P(\{X_2, X_3\} \subset N(X_1, r, c)) + 2\,P(X_2 \in N(X_1, r, c), \\ &\quad X_3 \in \Gamma_1(X_1, r, c)) + P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, c)) = P_{2N} + 2\,P_{NG} + P_{2G}.\end{aligned}$$

Similarly, let $P_{2N,e} := P(\{X_2, X_3\} \subset N_e(X_1, r))$, $P_{NG,e} := P(X_2 \in N_e(X_1, r), X_3 \in \Gamma_{1,e}(X_1, r))$, and $P_{2G,e} := P(\{X_2, X_3\} \subset \Gamma_{1,e}(X_1, r))$. Then

$$\mathbf{Cov}[h_{12}, h_{13}] = (P_{2N,e} + 2\,P_{NG,e} + P_{2G,e})/4 - \mu_e(r)^2.$$

Hence, we have $\nu(r, c) > 0$ (and $\nu_e(r) > 0$) iff $P_{2N} + 2\,P_{NG} + P_{2G} > 4\,\mu(r, c)^2$ (and $P_{2N,e} + 2\,P_{NG,e} + P_{2G,e} > 4\,\mu_e(r)^2$).

For $r = \infty$, we have $N(x, \infty, c) = \mathcal{I}_i$ for all $x \in \mathcal{I}_i$ with $i \in \{2, \ldots, m\}$ and $N_e(x, \infty) = \mathcal{I}_i$ for all $x \in \mathcal{I}_i$ with $i \in \{1, m + 1\}$. Then for $i \in \{2, \ldots, m\}$

$$\mathbf{E}\left[\rho_{[i]}(\infty, c)\right] = \mathbf{E}\left[h_{12}\right] = \mu(\infty, c) = P(X_2 \in N(X_1, \infty, c) = P(X_2 \in \mathcal{I}_i) = 1.$$

On the other hand, $4\,\mathbf{E}\left[h_{12}h_{13}\right] = P(\{X_2, X_3\} \subset N(X_1, \infty, c)) + 2\,P(X_2 \in N(X_1, \infty, c), X_3 \in \Gamma_1(X_1, \infty, c)) + P(\{X_2, X_3\} \subset \Gamma_1(X_1, \infty, c)) = (1 + 2 + 1)$. Hence $\mathbf{E}\left[h_{12}h_{13}\right] = 1$ and so $\nu(\infty, c) = 0$. Similarly, for $i \in \{1, m + 1\}$, we have $\mu_e(\infty) = 1$ and $\nu_e(\infty) = 0$. Therefore, the CLT result does not hold for $r = \infty$. Furthermore, $\rho_{[i]}(r = \infty, c) = 1$ a.s.

For $r = 1$, in a middle interval, we have $\rho_{[i]}(r, c) = [n_{i,1}(n_{i,1} - 1)/2 + n_{i,2}(n_{i,2} - 1)/2]/(n(n - 1))$ where $n_{i,1}$ and $n_{i,2}$ are as in Theorem 2. As $n_{i,1}$, and $n_{i,2}$ (and hence n_i) goes to ∞, we have $[n_{i,1}(n_{i,1} - 1)]/(n_i(n_i - 1)) \to c^2$ and $[n_{i,2}(n_{i,2} - 1)]/(n_i(n_i - 1)) \to (1 - c)^2$. Then $\rho_{[i]}(r, c) \xrightarrow{\mathcal{L}} (c^2 + (1 - c)^2)/2 = 1/2 + c\,(1 - c)$ which is degenerate. Likewise, in the end intervals, by Lemma 1, we have $\rho_{[i]}(r = 1, c) = 1/2$ for $n_i > 1$. Hence the CLT result does not hold for $r = 1$ either.

Remark 1 **The Joint Distribution of** (h_{12}, h_{13}): The pair (h_{12}, h_{13}) is a bivariate discrete random variable with nine possible values so that

$$(2\,h_{12}, 2\,h_{13}) \in \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)\}.$$

Then finding the joint distribution of (h_{12}, h_{13}) is equivalent to finding the joint probability mass function of (h_{12}, h_{13}). Hence the joint distribution of (h_{12}, h_{13}) can be found by calculating the probabilities such as $P((h_{12}, h_{13}) = (0,0)) = P(\{X_2, X_3\} \subset \mathcal{I}_i \setminus (N(X_1, r, c) \cup \Gamma_1(X_1, r, c)))$. □

4.1 The distribution of relative arc density of $\mathcal{U}(y_1, y_2)$-random $\mathscr{D}_{n,2}(r, c)$-digraphs

In the special case of $m = 2$ with $\mathcal{Y}_2 = \{y_1, y_2\}$ and $\delta_1 = y_1 < y_2 = \delta_2$, we have only one middle interval, and the two end intervals are empty. In this section, we consider the relative density of proportional-edge PCD based on uniform data in (y_1, y_2). By Theorem 4 and Corollary 1, the asymptotic distribution of any $\rho_{[i]}(r, c)$ for the middle intervals for $m \geq 2$ will be identical to the asymptotic distribution of $\mathcal{U}(y_1, y_2)$-random $\mathscr{D}_{n,2}(r, c)$-digraph.

First we consider the simplest case of $r = 2$ and $c = 1/2$. By Theorem 4, without loss of generality, we can assume (y_1, y_2) to be the unit interval $(0, 1)$. Then $N(x, 2, 1/2) = B(x, r(x))$ where $r(x) = \min(x, 1 - x)$ for $x \in (0, 1)$. Hence proportional-edge PCD based on $N(x, 2, 1/2)$ is equivalent to the CCCD of Priebe et al. (2001). Moreover, we have $\Gamma_1(X_1, 2, 1/2) = (X_1/2, (1 + X_1)/2)$.

Theorem 5 *As $n \to \infty$, we have*

$$\sqrt{n}\,[\rho_n(2, 1/2) - \mu(2, 1/2)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu(2, 1/2)),$$

where $\mu(2, 1/2) = 1/2$ and $4\,\nu(2, 1/2) = 1/12$.

Proof By symmetry, we only consider $X_1 \in (0, 1/2)$. Notice that for $x \in (0, 1/2)$, we have $N(x, 2, 1/2) = (0, 2\,x)$ and $\Gamma_1(x, 2, 1/2) = (x/2, (1 + x)/2)$. Hence

$$\mu(2, 1/2) = P(X_2 \in N(X_1, 2, 1/2)) = 2\,P(X_2 \in N(X_1, 2, 1/2), X_1 \in (0, 1/2))$$

by symmetry. Here

$$\begin{aligned} &P(X_2 \in N(X_1, 2, 1/2), X_1 \in (0, 1/2)) = P(X_2 \in (0, 2\,x_1), X_1 \in (0, 1/2)) \\ &= \int_0^{1/2} \int_0^{2x_1} f_{1,2}(x_1, x_2) dx_2 dx_1 = \int_0^{1/2} \int_0^{2x_1} 1 dx_2 dx_1 \\ &= \int_0^{1/2} 2\,x_1 dx_1 = x_1^2 |_0^{1/2} = 1/4. \end{aligned} \tag{9}$$

Then $\mu(2, 1/2) = 2\,(1/4) = 1/2$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability

$$\begin{aligned} P_{2N} &= P(\{X_2, X_3\} \subset N(X_1, 2, 1/2)) \\ &= 2\,P(\{X_2, X_3\} \subset N(X_1, 2, 1/2), X_1 \in (0, 1/2)) \end{aligned} \tag{10}$$

and

$$P(\{X_2, X_3\} \subset N(X_1, 2, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/2} (2\,x_1)^2 dx_1 = 1/6.$$

So $P_{2N} = 2\,(1/6) = 1/3$.

$$P_{NG} = 2\,P(X_2 \in N(X_1, 2, 1/2), X_3 \in \Gamma_1(X_1, 2, 1/2), X_1 \in (0, 1/2))$$

and

$$\begin{aligned} &P(X_2 \in N(X_1, 2, 1/2), X_3 \in \Gamma_1(X_1, 2, 1/2), X_1 \in (0, 1/2)) \\ &= \int_0^{1/2} (2\,x_1)(1/2) dx_1 = 1/8. \end{aligned} \tag{11}$$

Then $P_{NG} = 2\,(1/8) = 1/4$.

Finally, we have $P_{2G} = 2\,P(\{X_2, X_3\} \subset \Gamma_1(X_1, 2, 1/2), X_1 \in (0, 1/2))$ and

$$P(\{X_2, X_3\} \subset \Gamma_1(X_1, 2, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/2} (1/4) dx_1 = 1/8.$$

So $P_{2G} = 2\,(1/8) = 1/4$.

Therefore, $4\,\mathbf{E}[h_{12}h_{13}] = 1/3 + 2\,(1/4) + 1/4 = 13/12$. Hence $4\,\nu(2, 1/2) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = 13/12 - 4(1/2)^2 = 1/12$. □

The sharpest rate of convergence in Theorem 5 is $K\,\frac{\mu(2,1/2)}{\sqrt{n\,\nu(2,1/2)^3}} = 12\sqrt{3}\,\frac{K}{\sqrt{n}}$.

Next we consider the more general case of $r = 2$ and $c \in (0, 1)$. Without loss of generality, assume $0 < c < 1/2$. For $x \in (0, 1)$, the proximity region has the following form:

$$N(x, 2, c) = \begin{cases} (0, 2\,x) & \text{if } x \in (0, c), \\ (0, 1) & \text{if } x \in [c, 1/2), \\ (2\,x - 1, 1) & \text{if } x \in [1/2, 1) \end{cases} \tag{12}$$

and the Γ_1-region is $\Gamma_1(x, 2, c) = (\min(x/2, c), \max(1/2, (1+x)/2)) = (\min(x/2, c), (1+x)/2)$, since $(1+x)/2 > 1/2$.

Theorem 6 *As* $n \to \infty$, *for* $c \in (0,1)$, *we have* $\sqrt{n}\,[\rho_n(2,c) - \mu(2,c)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu(2,c))$, *where* $\mu(2,c) = \mu_1(2,c)\,\mathbf{I}(0 < c \le 1/2) + \mu_2(2,c)\,\mathbf{I}(1/2 \le c < 1)$ *and* $\nu(2,c) = \nu_1(2,c)\,\mathbf{I}(0 < c \le 1/2) + \nu_2(2,c)\,\mathbf{I}(1/2 \le c < 1)$ *with* $\mu_1(2,c) = c^2 - c + 3/4$ *and*

$$4\,\nu_1(2,c) = \begin{cases} 26\,c^3/3 + c + 1/24 - 4\,c^4 - 5\,c^2 & \text{if } 0 < c \le 1/4, \\ 6\,c^3 + c/2 + 1/12 - 4\,c^4 - 3\,c^2 & \text{if } 1/4 < c < 1/2, \end{cases} \tag{13}$$

and $\mu_2(2,c) = \mu_1(2, 1-c)$ *and* $\nu_2(2,c) = \nu_1(2, 1-c)$.

Proof is provided in Appendix 1. See Fig. 2 for the plots of the mean $\mu(2,c)$ and the asymptotic variance $4\,\nu(2,c)$. Notice that for $c = 1/2$, we have $\mu(2, c = 1/2) = 1/2$, and $4\,\nu(2, c = 1/2) = 1/12$, hence as $c \to 1/2$, the distribution of $\rho_n(2,c)$ converges to the one in Theorem 5. Furthermore, the sharpest rate of convergence in Theorem 6 is $K\,\frac{\mu(2,c)}{\sqrt{n\,\nu(2,c)^3}}$ which is, for $c \in (0, 1/2)$,

$$\frac{K}{\sqrt{n}} \begin{cases} \frac{72\,(4\,c - 4\,c^2 - 3)}{(96\,c^4 - 208\,c^3 + 120\,c^2 - 24\,c - 1)\sqrt{-576\,c^4 + 1248\,c^3 - 720\,c^2 + 144\,c + 6}} & \text{if } 0 < c \le 1/4, \\ \frac{18\,(4\,c - 4\,c^2 - 3)}{(48\,c^4 - 72\,c^3 + 36\,c^2 - 6\,c - 1)\sqrt{-144\,c^4 + 216\,c^3 - 108\,c^2 + 18\,c + 3}} & \text{if } 1/4 < c < 1/2, \end{cases}$$

and is minimized at $c \approx 0.19$ which is found by setting the first derivative of this rate with respect to c to zero and solving for c numerically. We also checked the plot of $\frac{\mu_1(2,c)}{\sqrt{\nu_1(2,c)^3}}$ (not presented) and verified that this is actually where the global minimum is attained. By symmetry, the same global minimum for $K\,\frac{\mu(2,c)}{\sqrt{n\,\nu(2,c)^3}}$ is also attained at $c \approx 0.81$.

Next we consider the case of $r \ge 1$ and $c = 1/2$. By symmetry, we only consider $X_1 \in (0, 1/2)$. For $x \in (0, 1/2)$, the proximity region is $N(x, r, c = 1/2) = (0, \min(r\,x, 1))$ and the Γ_1-region is $\Gamma_1(x, r, 1/2) = (\min(x/r, c), \max(1/2, 1 - (1 - x)/r))$.

Theorem 7 *For* $r \in (1, \infty)$, *as* $n \to \infty$, *we have* $\sqrt{n}\,[\rho_n(r, 1/2) - \mu(r, 1/2)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu(r, 1/2))$ *where*

$$\mu(r, 1/2) = \begin{cases} r/4 & \text{if } 1 \le r < 2, \\ 1 - 1/r & \text{if } r \ge 2, \end{cases} \tag{14}$$

and

$$4\,\nu(r, 1/2) = \begin{cases} \frac{4\,r^4 + 12\,r - r^5 - r^3 - 10\,r^2 - 4}{12\,r^2} & \text{if } 1 \le r < 2, \\ \frac{2\,r - 3}{3\,r^2} & \text{if } r \ge 2. \end{cases} \tag{15}$$

Proof is provided in Appendix 1. See Fig. 3 for the plots of the mean $\mu(r, 1/2)$ and the asymptotic variance $4\,\nu(r, 1/2)$. Notice that $\nu(r = 1, 1/2) = 0$ and

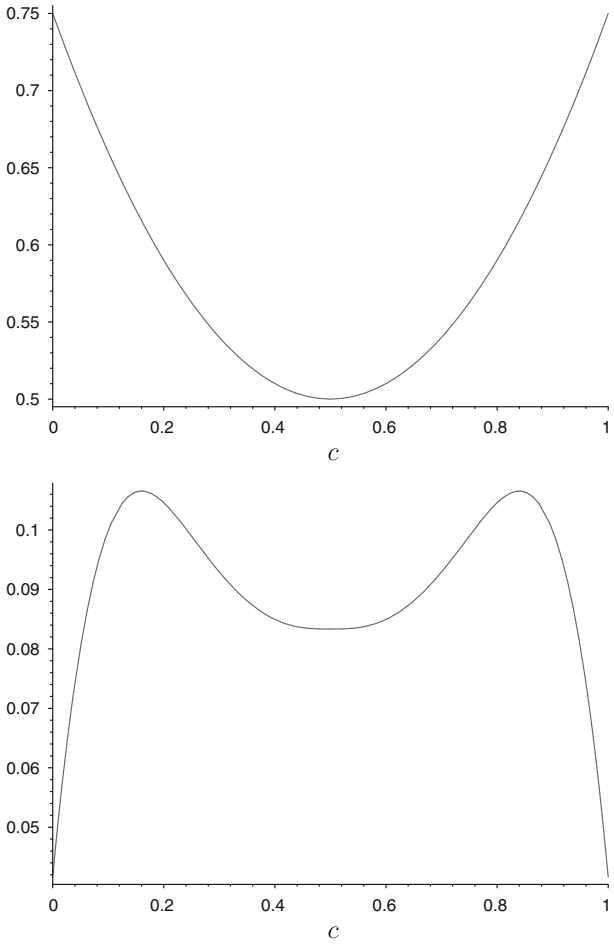

Fig. 2 The plots of the asymptotic mean $\mu(2, c)$ (*top*) and the variance $4\,\nu(2, c)$ (*bottom*) as a function of c for $c \in (0, 1)$

$\lim_{r\to\infty} \nu(r, 1/2) = 0$, so the CLT result fails for $r \in \{1, \infty\}$. For $r = 2$, we have $\mu(r = 2, c = 1/2) = 1/2$, and $4\,\nu(r = 2, c = 1/2) = 1/12$, hence as $r \to 2$, the distribution of $\rho_n(r, 1/2)$ converges to the one in Theorem 5. Furthermore, the sharpest rate of convergence in Theorem 7 is

$$K\,\frac{\mu(r, 1/2)}{\sqrt{n\,\nu(r, 1/2)^3}} = \frac{K}{\sqrt{n}} \begin{cases} \frac{6\sqrt{3}\,r^4}{(2r^2+4r-r^3-4)^{3/2}(r-1)^2} & \text{if } 1 \le r < 2, \\ \frac{3\sqrt{3}\,r^2(r-1)}{(2r-3)^{3/2}} & \text{if } r \ge 2. \end{cases} \tag{16}$$

and is minimized at $r \approx 2.55$ which is found by setting the first derivative of this rate with respect to r to zero and solving for r numerically. We also checked the plot

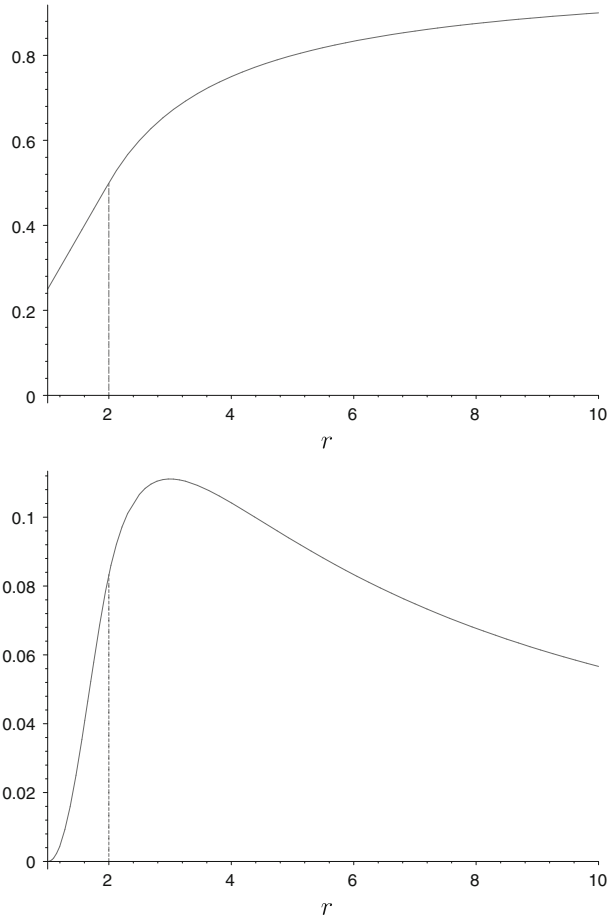

Fig. 3 The plots of the asymptotic mean $\mu(r, 1/2)$ (*top*) and the variance $4\,\nu(r, 1/2)$ (*bottom*) as a function of r for $r \in [1, 10]$

of $\mu(r, 1/2)/\sqrt{\nu(r, 1/2)^3}$ (not presented) and verified that this is where the global minimum is attained.

Finally, we consider the most general case of $r \geq 1$ and $c \in (0, 1)$. The proximity region has the following forms:

$$N(x, r, c) = \begin{cases} (0, \min(r\,x, 1)) & \text{if} \quad x \in (0, c), \\ (\max(0, 1 - r\,(1 - x)), 1) & \text{if} \quad x \in (c, 1), \end{cases} \tag{17}$$

and the Γ_1-region is $\Gamma_1(x, r, c) = (\min(x/r, c), \max(c, 1 - (1 - x)/r))$.

Theorem 8 *For $r \in [1, \infty)$, and $c \in (0, 1)$, we have $\sqrt{n}\,[\rho_n(r, c) - \mu(r, c)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,\nu(r, c))$, as $n \to \infty$, where $\mu(r, c) = \mu_1(r, c)\,\mathbf{I}(0 < c \leq 1/2) + \mu_2(r, c)\,\mathbf{I}(1/2 \leq c < 1)$ and $\nu(r, c) = \nu_1(r, c)\,\mathbf{I}(0 < c \leq 1/2) + \nu_2(r, c)\,\mathbf{I}(1/2 \leq c < 1)$. For $0 < c \leq 1/2$,*

$$\mu_1(r,c) = \begin{cases} c^2 r - c\,r + r/2 & \text{if } 1 \le r < 1/(1-c), \\ \frac{c^2 r^2 - 2c\,r + 2r - 1}{2r} & \text{if } 1/(1-c) \le r < 1/c, \\ 1 - 1/r & \text{if } r \ge 1/c, \end{cases} \tag{18}$$

and for $1/4 < c \le 1/2$,

$$\begin{aligned} \nu_1(r,c) = \kappa_1(r,c)\,\mathbf{I}(1 \le r < 1/(1-c)) + \kappa_2(r,c)\,\mathbf{I}(1/(1-c) \le r < 1/c) \\ + \kappa_3(r,c)\,\mathbf{I}(r \ge 1/c) \end{aligned} \tag{19}$$

where

$$\begin{aligned} 4\,\kappa_1(r,c) &= -[12\,c^4 r^4 - 24\,c^3 r^4 + 3\,c^2 r^5 + 15\,c^2 r^4 - 3\,c\,r^5 - 9\,c^2 r^3 - 3\,c\,r^4 + r^5 \\ &\quad + 6\,c^2 r^2 + 9\,c\,r^3 - r^4 - 6\,c\,r^2 - 2\,r^3 + 4\,r^2 - 3\,r + 1] \Big/ [3\,r^2], \\ 4\,\kappa_2(r,c) &= -[3\,c^4 r^4 + c^3 r^5 - c^3 r^4 - 11\,c^3 r^3 - 3\,c^2 r^4 + 6\,c^2 r^3 + 9\,c^2 r^2 + 3\,c\,r^3 \\ &\quad - 9\,c\,r^2 + 3\,c\,r - r^2 + 2\,r - 1] \Big/ [3\,r^2], \\ 4\,\kappa_3(r,c) &= \frac{2\,r - 3}{3\,r^2}, \end{aligned}$$

and for $0 < c \le 1/4$,

$$\begin{aligned} \nu_1(r,c) = \vartheta_1(r,c)\,\mathbf{I}\left(1 \le r < \frac{1}{1-c}\right) + \vartheta_2(r,c)\,\mathbf{I}\left(\frac{1}{1-c} \le r < \frac{1-\sqrt{1-4c}}{2c}\right) \\ + \vartheta_3(r,c)\,\mathbf{I}\left(\frac{1-\sqrt{1-4c}}{2c} \le r < \frac{1+\sqrt{1-4c}}{2c}\right) \\ + \vartheta_4(r,c)\,\mathbf{I}\left(\frac{1+\sqrt{1-4c}}{2c} \le r < \frac{1}{c}\right) + \vartheta_5(r,c)\,\mathbf{I}(r \ge 1/c) \end{aligned} \tag{20}$$

where $\vartheta_1(r,c) = \kappa_1(r,c)$, $\vartheta_2(r,c) = \vartheta_4(r,c) = \kappa_2(r,c)$, $\vartheta_5(r,c) = \kappa_3(r,c)$, *and*

$$4\,\vartheta_3(r,c) = \frac{3\,c^4 r^5 - c^3 r^5 - 11\,c^3 r^4 + 3\,c^2 r^4 + 9\,c^2 r^3 - 3\,c\,r^3 - r^2 + 2\,r - 1}{3\,r^3}.$$

And for $c \in (1/2, 1)$, *we have* $\mu_2(r,c) = \mu_1(r, 1-c)$ *and* $\nu_2(r,c) = \nu_1(r, 1-c)$.

Proof is provided in Appendix 1. See Fig. 4 for the plots of the mean $\mu(r,c)$ and the asymptotic variance $4\,\nu(r,c)$. Notice that $\lim_{c \to 1/2} \nu(r=1,c) = 0$ and $\lim_{r \to \infty} \nu(r,c) = 0$, so the CLT result fails for $(r,c) = (1, 1/2)$ and $r = \infty$. Furthermore, for $r = 2$ and $c = 1/2$, we have $\mu(r=2, c=1/2) = 1/2$, and $4\,\nu(r=2, c=1/2) = 1/12$, hence as $r \to 2$ and $c \to 1/2$, the distribution of $\rho_n(r,c)$ converges to the one in Theorem 5. The sharpest rate of convergence in Theorem 8 is $K\,\frac{\mu(r,c)}{\sqrt{n\,\nu(r,c)^3}}$

(the explicit form not presented) and is minimized at $r \approx 1.88$ and $c \approx 0.19$ (or $c \approx 0.81$) which is found by setting the first order partial derivatives of this rate with respect to r and c to zero and solving for r and c numerically. We also checked the surface plot of this rate (not presented) and observed that these points are actually where the global minimum is attained.

4.2 The case of end intervals: relative density for $\mathcal{U}\left(\delta_1, y_{(1)}\right)$ or $\mathcal{U}\left(y_{(m)}, \delta_2\right)$ data

Recall that with $m \geq 1$ for the end intervals, $\mathcal{I}_1 = \left(\delta_1, y_{(1)}\right)$ and $\mathcal{I}_{m+1} = \left(y_{(m)}, \delta_2\right)$, the proximity and Γ_1-regions were only dependent on x and r (but not on c). For $\mathcal{U}\left(\delta_1, y_{(1)}\right)$ and $\mathcal{U}\left(y_{(m)}, \delta_2\right)$ data, by symmetry, relative density has the same distribution. So we only consider $\left(y_{(m)}, \delta_2\right)$. Due to scale invariance from Theorem 4, we can assume that $\left(y_{(m)}, \delta_2\right)$ is $(0, 1)$. Let $\Gamma_{1,e}(x, r)$ be the Γ_1-region corresponding to $N_e(x, r)$ in the end interval case.

First we consider $r = 2$ and uniform data in the end intervals. Then for x in the right end interval, $N_e(x, 2) = (0, \min(1, 2x))$ for $x \in (0, 1)$ and the Γ_1-region is $\Gamma_{1,e}(x, 2) = (x/2, 1)$.

Theorem 9 *Let $D_{[i]}(2, c)$ be the subdigraph of the proportional-edge PCD based on uniform data in (δ_1, δ_2) where $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in (δ_1, δ_2). Then for $i \in \{1, m+1\}$ (i.e., in the end intervals), as $n_i \to \infty$, we have $\sqrt{n_i}\left[\rho_{[i]}(2, c) - \mu_e(2)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\nu_e(2))$, where $\mu_e(2) = 3/4$ and $4\nu_e(2) = 1/24$.*

Proof For $x_1 \in (0, 1)$, depending on the location of x_1, the following are the different types of the combinations of $N_e(x_1, 2)$ and $\Gamma_{1,e}(x_1, 2)$.

(i) for $0 < x_1 \leq 1/2$, $N_e(x_1, 2) = (0, 2x_1)$ and $\Gamma_{1,e}(x_1, 2) = (x_1/2, 1)$,
(ii) for $1/2 < x_1 < 1$, $N_e(x_1, 2) = (0, 1)$ and $\Gamma_{1,e}(x_1, 2) = (x_1/2, 1)$.

Then $\mu_e(2) = P(X_2 \in N_e(X_1, 2)) = \int_0^{1/2} 2x_1 dx_1 + \int_{1/2}^1 1 dx_1 = 3/4$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}.

$$P_{2N} = P(\{X_2, X_3\} \subset N_e(X_1, 2)) = \int_0^{1/2} (2x_1)^2 dx_1 + \int_{1/2}^1 1 dx_1 = 2/3.$$

$$\begin{aligned} P_{NG} &= P(X_2 \in N_e(X_1, 2), X_3 \in \Gamma_{1,e}(X_1, 2)) \\ &= \int_0^{1/2} (2x_1)(1 - x_1/2) dx_1 + \int_{1/2}^1 (1 - x_1/2) dx_1 = 25/48. \end{aligned}$$

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_{1,e}(X_1, 2)) = \int_0^1 (1 - x_1/2)^2 dx_1 = 7/12$.

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = 55/24$. Hence $4\nu_e(2) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = 1/24$. □

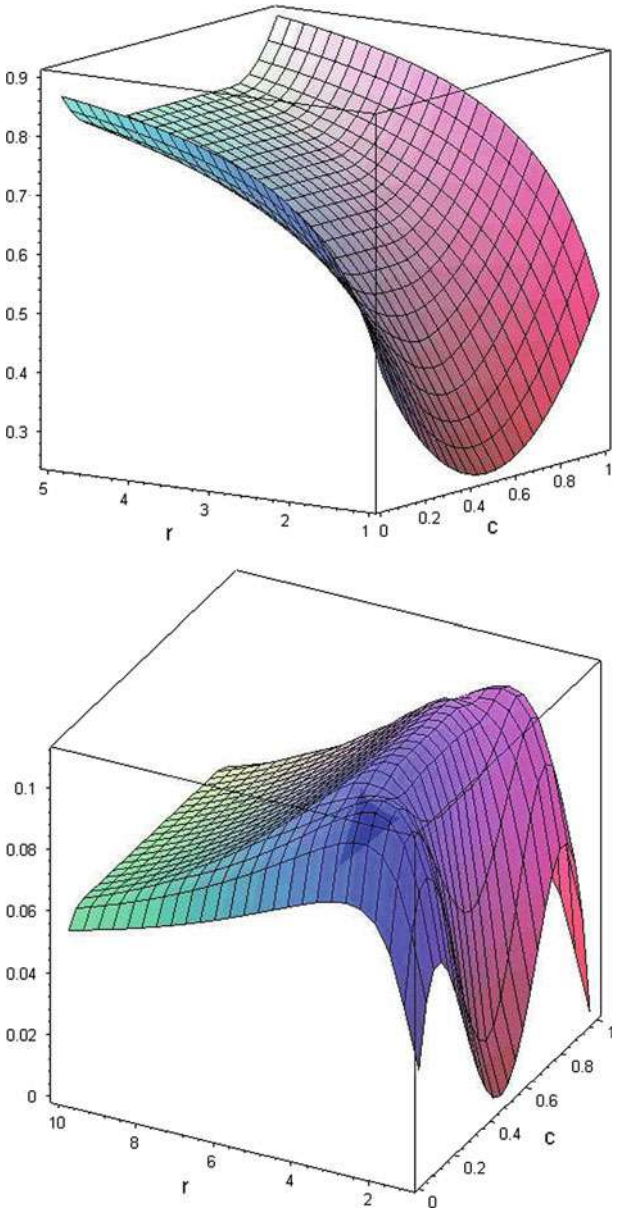

Fig. 4 The surface plots of the asymptotic mean $\mu(r, c)$ (*top*) and the variance $4\,\nu(r, c)$ (*bottom*) as a function of r and c for $r \in [1, 10]$ and $c \in (0, 1)$, respectively

The sharpest rate of convergence in Theorem 9 is $K\,\frac{\mu(2,1/2)}{\sqrt{n\,\nu(2,1/2)^3}} = 36\sqrt{6}\,\frac{K}{\sqrt{n}}$.

Next we consider the more general case of $r \geq 1$ for the end intervals. By Theorem 4, we can assume each end interval to be $(0, 1)$. For x in the right end interval,

the proximity region is $N_e(x,r) = (0, \min(1, r\,x))$ for $x \in (0,1)$ and the Γ_1-region is $\Gamma_{1,e}(x,r) = (x/r, 1)$.

Theorem 10 *Let $D_{[i]}(r,c)$ be the subdigraph of the proportional-edge PCD based on uniform data in (δ_1, δ_2) where $\delta_1 < \delta_2$ and $\mathcal{Y}_m$ be a set of m distinct $\mathcal{Y}$ points in (δ_1, δ_2). Then for $i \in \{1, m+1\}$ (i.e., in the end intervals), and $r \in (1, \infty)$, we have $\sqrt{n_i}\left[\rho_{[i]}(r,c) - \mu_e(r)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\,v_e(r))$, as $n_i \to \infty$, where $\mu_e(r) = 1 - 1/(2\,r)$ and $4\,v_e(r) = (r-1)^2/(3\,r^3)$.*

Proof For $x_1 \in (0,1)$, depending on the location of x_1, the following are the different types of the combinations of $N_e(x_1,r)$ and $\Gamma_{1,e}(x_1,r)$.

(i) for $0 < x_1 \le 1/r$, $N_e(x_1,r) = (0, r\,x_1)$ and $\Gamma_{1,e}(x_1,r) = (x_1/r, 1)$,
(ii) for $1/r < x_1 < 1$, $N_e(x_1,r) = (0,1)$ and $\Gamma_{1,e}(x_1,r) = (x_1/r, 1)$.

Then $\mu_e(r) = P(X_2 \in N_e(X_1,r)) = \int_0^{1/r} r\,x_1 dx_1 + \int_{1/r}^1 1 dx_1 = 1 - 1/(2\,r)$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate $P_{2N,e}$, $P_{NG,e}$, and $P_{2G,e}$. The probability

$$P_{2N,e} = P(\{X_2, X_3\} \subset N_e(X_1,r)) = \int_0^{1/r} (r\,x_1)^2 dx_1 + \int_{1/r}^1 1 dx_1 = 1 - 2/(2\,r).$$

$$\begin{aligned} P_{NG,e} &= P(X_2 \in N_e(X_1,r), X_3 \in \Gamma_{1,e}(X_1,r)) \\ &= \int_0^{1/r} (r\,x_1)(1 - x_1/r) dx_1 + \int_{1/r}^1 (1 - x_1/r) dx_1 = 1 - r^{-1} + r^{-3}/6. \end{aligned}$$

Finally,

$$P_{2G,e} = P(\{X_2, X_3\} \subset \Gamma_{1,e}(X_1,r)) = \int_0^1 (1 - x_1/r)^2 dx_1 = 1 - r^{-1} + r^{-2}/3.$$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N,e} + 2\,P_{NG,e} + P_{2G,e} = \frac{12\,r^3 - 11\,r^2 + r + 1}{3\,r^3}$. Hence $4\,v_e(r) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = (r-1)^2/(3\,r^3)$. □

See Fig. 5 for the plots of the mean $\mu_e(r)$ and the asymptotic variance $4\,v_e(r)$. Notice that $v_e(r=1) = 0$ and $\lim_{r\to\infty} v_e(r) = 0$, so the CLT result fails for $r \in \{1, \infty\}$. Furthermore, for $r = 2$, we have $\mu_e(r=2) = 3/4$, and $4\,v_e(r=2) = 1/24$, hence as $r \to 2$, the distribution of $\rho_n(r,c)$ converges to the one in Theorem 9. The sharpest rate of convergence in Theorem 10 is $K\,\frac{\mu(r,1/2)}{\sqrt{n\,v(r,1/2)^3}}$ (explicit form not presented) and is minimized at $r \approx 2.74$ which is found numerically as before. We also checked the plot of $\mu_e(r)/\sqrt{v_e(r)^3}$ (not presented) and verified that this is where the global minimum is attained.

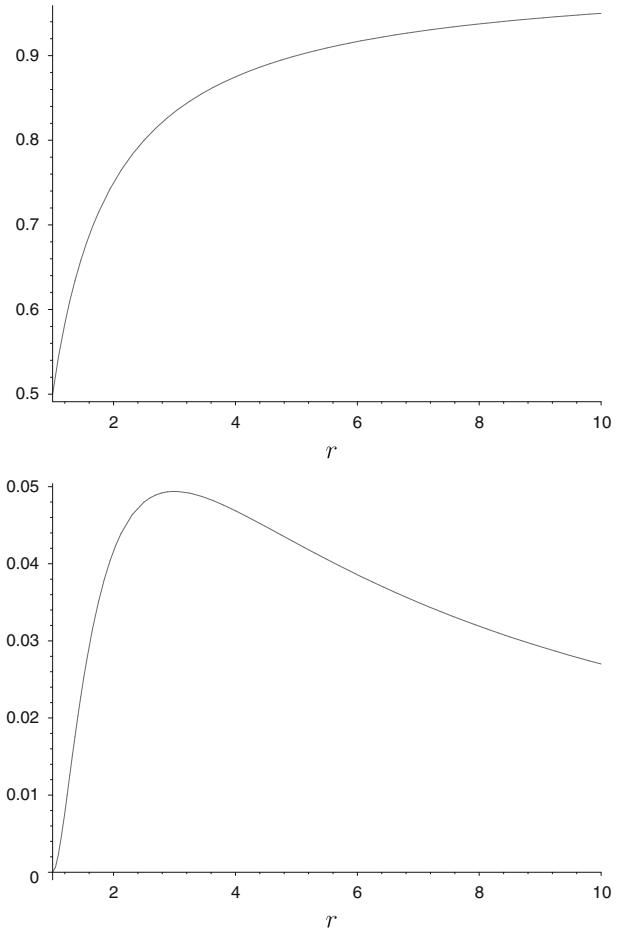

Fig. 5 The plots of the asymptotic mean $\mu_e(r)$ (*top*) and the variance $4\,v_e(r)$ (*bottom*) for the end intervals as a function of r for $r \in [1, 10]$

5 The distribution of the relative arc density of $\mathscr{D}_{n,m}(r, c)$-digraphs

In this section, we consider the more challenging case of $m \geq 2$.

5.1 First version of relative density in the case of $m \geq 2$

Recall that the relative density $\rho_{n,m}(r, c)$ is defined as in Eq. (7). Letting $w_i = (y_{(i)} - y_{(i-1)})/(\delta_2 - \delta_1)$, for $i = 1, 2, \ldots, m + 1$, we obtain the following as a result of Theorem 8.

Theorem 11 *Let $\mathcal{X}_n$ be a random sample from $\mathcal{U}(\delta_1, \delta_2)$ with $-\infty < \delta_1 < \delta_2 < \infty$ and $\mathcal{Y}_m$ be a set of m distinct points in (δ_1, δ_2). For $r \in (1, \infty)$ and $c \in (0, 1)$, the*

asymptotic distribution of $\rho_{n,m}(r,c)$ *conditional on* $\mathcal{Y}_m$ *is given by*

$$\sqrt{n}\left(\rho_{n,m}(r,c) - \breve{\mu}(m,r,c)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 4\,\breve{\nu}(m,r,c)\right), \tag{21}$$

as $n \to \infty$, *provided that* $\breve{\nu}(m,r,c) > 0$, *where*

$$\breve{\mu}(m,r,c) = \widetilde{\mu}(m,r,c) \Big/ \left(\sum_{i=1}^{m+1} w_i^2\right)$$

with $\widetilde{\mu}(m,r,c) = \mu(r,c)\sum_{i=2}^{m} w_i^2 + \mu_e(r)\sum_{i\in\{1,m+1\}} w_i^2$ *and* $\mu(r,c)$ *and* $\mu_e(r)$ *are as in Theorems 8 and 10, respectively. Furthermore,*

$$4\,\breve{\nu}(m,r,c) = 4\,\widetilde{\nu}(m,r,c) \Big/ \left(\sum_{i=1}^{m+1} w_i^2\right)^2$$

with

$$\begin{aligned} 4\,\widetilde{\nu}(m,r,c) = {} & [P_{2N} + 2\,P_{NG} + P_{2G}] \sum_{i=2}^{m} w_i^3 \\ & + [P_{2N,e} + 2\,P_{NG,e} + P_{2G,e}] \sum_{i\in\{1,m+1\}} w_i^3 - (\widetilde{\mu}(m,r,c))^2. \end{aligned}$$

Proof is provided in Appendix 2. Notice that if $y_{(1)} = \delta_1$ and $y_{(m)} = \delta_2$, there are only $m-1$ middle intervals formed by $y_{(i)}$ and the end intervals are empty. Hence in Theorem 11, $\breve{\mu}(m,r,c) = \mu(r,c)$ since $\widetilde{\mu}(m,r,c) = \mu(r,c)\sum_{i=2}^{m} w_i^2$. Furthermore, $4\,\breve{\nu}(m,r,c) = [P_{2N} + 2\,P_{NG} + P_{2G}]\sum_{i=2}^{m} w_i^3 - (\mu(r,c)\sum_{i=2}^{m} w_i^2)^2 = 4\,\nu(m,r,c) + \mu^2(r,c)\left(\sum_{i=2}^{m} w_i^3 - (\sum_{i=2}^{m} w_i^2)^2\right)$.

5.2 Second version of relative density in the case of $m \geq 2$

For $m \geq 2$, if we consider the entire data set $\mathcal{X}_n$, then we have n vertices. So we can also consider the relative density as $\widetilde{\rho}_{n,m}(r,c) = |\mathcal{A}| / (n\,(n-1))$.

Theorem 12 *Let* $\mathcal{X}_n$ *be a random sample from* $\mathcal{U}(\delta_1,\delta_2)$ *with* $-\infty < \delta_1 < \delta_2 < \infty$ *and* $\mathcal{Y}_m$ *be a set of* m *distinct points in* (δ_1,δ_2). *For* $r \in (1,\infty)$ *and* $c \in (0,1)$, *the asymptotic distribution for* $\widetilde{\rho}_{n,m}(r,c)$ *conditional on* $\mathcal{Y}_m$ *is given by*

$$\sqrt{n}\left(\widetilde{\rho}_{n,m}(r,c) - \widetilde{\mu}(m,r,c)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 4\,\widetilde{\nu}(m,r,c)\right), \tag{22}$$

as $n \to \infty$, *provided that* $\widetilde{\nu}(m,r,c) > 0$, *where* $\widetilde{\mu}(m,r,c)$ *and* $\widetilde{\nu}(m,r,c)$ *are as in Theorem 11.*

Proof is provided in Appendix 2. Notice that the relative arc densities, $\rho_{n,m}(r,c)$ and $\widetilde{\rho}_{n,m}(r,c)$ do not have the same distribution for neither finite nor infinite n. But we have $\rho_{n,m}(r,c) = \frac{n(n-1)}{n_T}\widetilde{\rho}_{n,m}(r,c)$ and since for large n_i and n, $\sum_{i=1}^{m+1} \frac{n_i(n_i-1)}{n(n-1)} \approx \sum_{i=1}^{m+1} w_i^2 < 1$, it follows that $\widetilde{\mu}(m,r,c) < \breve{\mu}(m,r,c)$ and $\widetilde{\nu}(m,r,c) < \breve{\nu}(m,r,c)$ for large n_i and n. Furthermore, the asymptotic normality holds for $\rho_{n,m}(r,c)$ iff it holds for $\widetilde{\rho}_{n,m}(r,c)$.

6 Extension of proportional-edge proximity regions to higher dimensions

Note that in $\mathbb{R}$ the proportional-edge PCDs are based on the intervals whose end points are from class $\mathcal{Y}$. This interval partitioning can be viewed as the *Delaunay tessellation* of $\mathbb{R}$ based on $\mathcal{Y}_m$. So in higher dimensions, we use the Delaunay triangulation based on $\mathcal{Y}_m$ to partition the space.

Let $\mathcal{Y}_m = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_m\}$ be m points in general position in $\mathbb{R}^d$ and T_i be the ith Delaunay cell for $i = 1, 2, \ldots, J_m$, where J_m is the number of Delaunay cells. Let $\mathcal{X}_n$ be a set of iid random variables from distribution F in $\mathbb{R}^d$ with support $\mathcal{S}(F) \subseteq \mathcal{C}_H(\mathcal{Y}_m)$ where $\mathcal{C}_H(\mathcal{Y}_m)$ stands for the convex hull of $\mathcal{Y}_m$.

6.1 Extension of proportional-edge proximity regions to $\mathbb{R}^2$

For illustrative purposes, we focus on $\mathbb{R}^2$ where a Delaunay tessellation is a *triangulation*, provided that no more than three points in $\mathcal{Y}_m$ are cocircular (i.e., lie on the same circle). Furthermore, for simplicity, we only consider the one Delaunay triangle case. Let $\mathcal{Y}_3 = \{\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3\}$ be three non-collinear points in $\mathbb{R}^2$ and $T(\mathcal{Y}_3) = T(\mathsf{y}_1, \mathsf{y}_2, \mathsf{y}_3)$ be the triangle with vertices $\mathcal{Y}_3$. Let $\mathcal{X}_n$ be a set of iid random variables from F with support $\mathcal{S}(F) \subseteq T(\mathcal{Y}_3)$.

For $r \in [1, \infty]$, define $N(\cdot, r, M)$ to be the *(parameterized) proportional-edge proximity map* with M-vertex regions as follows (see also Fig. 6 with $M = M_C$ which is the center of mass and $r = 2$). For $x \in T(\mathcal{Y}_3) \setminus \mathcal{Y}_3$, let $v(x) \in \mathcal{Y}_3$ be the vertex whose region contains x; i.e., $x \in R_M(v(x))$. In this article *M-vertex regions* are constructed by the lines joining any point $M \in \mathbb{R}^2 \setminus \mathcal{Y}_3$ to a point on each of the edges of $T(\mathcal{Y}_3)$. Preferably, M is selected to be in the interior of the triangle $T(\mathcal{Y}_3)^o$. For such an M, the corresponding vertex regions can be defined using the line segment joining M to e_j, which lies on the line joining y_j to M. With M_C, the lines joining M and $\mathcal{Y}_3$ are the *median lines*, that cross edges at M_j for $j = 1, 2, 3$. *M-vertex regions*, among many possibilities, can also be defined by the orthogonal projections from M to the edges. See Ceyhan (2005) for a more general definition. The vertex regions in Fig. 6 are center of mass vertex regions (i.e., M_C-vertex regions). If x falls on the boundary of two M-vertex regions, we assign $v(x)$ arbitrarily. Let $e(x)$ be the edge of $T(\mathcal{Y}_3)$ opposite of $v(x)$. Let $\ell(v(x), x)$ be the line parallel to $e(x)$ and passes through x. Let $d(v(x), \ell(v(x), x))$ be the Euclidean distance from $v(x)$ to $\ell(v(x), x)$. For $r \in [1, \infty)$, let $\ell_r(v(x), x)$ be the line parallel to $e(x)$ such that

$$d(v(x), \ell_r(v(x), x)) = r\, d(v(x), \ell(v(x), x))$$

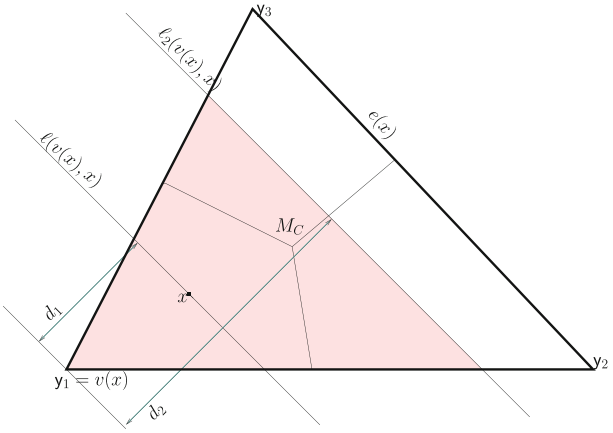

Fig. 6 Construction of proportional-edge proximity region, $N(x, r = 2, M_C)$ (*shaded region*) for an x in the M_C-vertex region for y_1, $R_{M_C}(\mathsf{y}_1)$

and

$$d(\ell(v(x), x), \ell_r(v(x), x)) < d(v(x), \ell_r(v(x), x)).$$

Let $T_r(x)$ be the triangle similar to and with the same orientation as $T(\mathcal{Y}_3)$ having $v(x)$ as a vertex and $\ell_r(v(x), x)$ as the opposite edge. Then the (parameterized) proportional-edge proximity region $N(x, r, M)$ is defined to be $T_r(x) \cap T(\mathcal{Y}_3)$. Notice that $\ell(v(x), x)$ divides the edges of $T_r(x)$ (other than the one lies on $\ell_r(v(x), x)$) proportionally with the factor r. Hence the name *proportional-edge proximity region*.

6.2 Extension of proportional-edge proximity regions to $\mathbb{R}^d$ with $d > 2$

The extension to $\mathbb{R}^d$ for $d > 2$ with $M = M_C$ is provided in Ceyhan and Priebe (2005), the extension for general M is similar: Let $\mathcal{Y}_{d+1} = \{\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_{d+1}\}$ be $d + 1$ non-coplanar points. Denote the simplex formed by these $d + 1$ points as $\mathcal{S}(\mathcal{Y}_{d+1})$. For $r \in [1, \infty]$, define the proportional-edge proximity map as follows. Given a point x in $\mathcal{S}(\mathcal{Y}_{d+1})$, let $Q_{\mathsf{y}}(M, x)$ be the polytope with vertices being the $d\,(d + 1)/2$ points on the edges, the vertex y and x so that the faces of $Q_{\mathsf{y}}(M, x)$ are formed by $d - 1$ line segments each of which joining one of $\mathcal{Y}$ points, say y_i, to M and that are between M and the face opposite y_i. That is, the vertex region for vertex v is the polytope with vertices given by v and such points on the edges. Let $v(x)$ be the vertex in whose region x falls. If x falls on the boundary of two vertex regions, we assign $v(x)$ arbitrarily. Let $\varphi(x)$ be the face opposite to vertex $v(x)$, and $\eta(v(x), x)$ be the hyperplane parallel to $\varphi(x)$ which contains x. Let $d(v(x), \eta(v(x), x))$ be the Euclidean distance from $v(x)$ to $\eta(v(x), x)$. For $r \in [1, \infty)$, let $\eta_r(v(x), x)$ be the hyperplane parallel to $\varphi(x)$ such that $d(v(x), \eta_r(v(x), x)) = r\, d(v(x), \eta(v(x), x))$ and $d(\eta(v(x), x), \eta_r(v(x), x)) < d(v(x), \eta_r(v(x), x))$. Let $\mathcal{S}_r(x)$ be the polytope similar to and with the same orientation

as $\mathcal{S}(\mathcal{Y}_{d+1})$ having $v(x)$ as a vertex and $\eta_r(v(x), x)$ as the opposite face. Then the proportional-edge proximity region $N(x, r, M) := \mathcal{S}_r(x) \cap \mathcal{S}(\mathcal{Y}_{d+1})$.

7 Discussion

In this article, we investigate a graph invariant of a random digraph family called proportional-edge proximity catch digraph (PCD) which is based on two classes of points in $\mathbb{R}$. The graph invariant of interest is the relative arc density (which is the number of arcs in a given digraph to the total number of arcs possible in a complete symmetric digraph with the same number of vertices). Points from one of the classes constitute the vertices of the PCDs and are a random sample from uniform distribution in compact intervals in $\mathbb{R}$. We demonstrate that the relative arc density of the PCDs is a U-statistic. Then, applying the central limit theory of the U-statistics, we derive the (asymptotic normal) distribution of the relative arc density.

The PCD we discuss here is based on a parameterized proximity map in which there is an expansion parameter, $r \geq 1$, and a centrality parameter, $c \in (0, 1)$. We provide the asymptotic distribution of the relative arc density for proportional-edge PCDs for uniform data for the entire ranges of r and c. We also determine the parameters r and c for which the rate of convergence to normality is the fastest. The PCD in this article can also be viewed as the one dimensional version of the PCD in Ceyhan and Priebe (2005, 2007) (see also Sect. 6). As in Ceyhan et al. (2006), we can use the relative arc density in testing one dimensional spatial point patterns and our results will help make the power comparisons possible for data from large families of distributions. In hypothesis testing, e.g., of spatial point patterns in the one dimensional case, the null hypothesis is some form of complete spatial randomness, which implies that distribution of $\mathcal{X}$ points has a uniform distribution in the support interval irrespective of the distribution of the $\mathcal{Y}$ points. The alternatives are segregation and association of $\mathcal{X}$ points with respect to the $\mathcal{Y}$ points. Under segregation, the points from the same class tend to cluster together, while under association, the points from the two different classes occur close to each other. In this context, under association, $\mathcal{X}$ points are clustered around $\mathcal{Y}$ points, while under segregation, $\mathcal{X}$ points are clustered away from the $\mathcal{Y}$ points. Notice that we can use the asymptotic distribution (i.e., the normal approximation) of the relative density for spatial pattern tests, so our methodology requires number of $\mathcal{X}$ points to be much larger compared to the number of $\mathcal{Y}$ points. Our results will make the power comparisons possible for data from large families of distributions. Moreover, one might determine the optimal (with respect to empirical size and power) parameter values against segregation and association alternatives.

Furthermore, a high dimensional data might be projected to one dimensional space (by some dimension reduction method) and then proportional-edge PCD might be used for classification as outlined in Priebe et al. (2003a). Here, one might also determine the optimal parameters (with respect to some penalty function in the classification procedure) for the best performance. This work will form the foundation of the generalizations and calculations for uniform and non-uniform cases in multiple dimensions. See Sect. 6 for the details of the extension to higher dimensions. For example, in $\mathbb{R}^2$, the expansion parameter is still r, but the centrality parameter is $M = (m_1, m_2)$, which is

two dimensional. The optimal parameters for testing spatial patterns and classification can also be determined, as in the one dimensional case.

Acknowledgments I would like to thank the anonymous referees whose constructive comments and suggestions greatly improved the presentation and flow of this article. This work was supported by TUBITAK Kariyer Project Grant 107T647.

Appendix 1: Proofs for the one interval case

Proof of Theorem 6 First we consider $0 < c \leq 1/2$, for which there are two cases, namely $0 < c \leq 1/4$ and $1/4 < c \leq 1/2$.
Case 1: $0 < c \leq 1/4$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, 2, c)$ and $\Gamma_1(x_1, 2, c)$.

(i) for $0 < x_1 \leq c$, $N(x_1, 2, c) = (0, 2x_1)$ and $\Gamma_1(x_1, 2, c) = (x_1/2, (1 + x_1)/2)$,
(ii) for $c < x_1 \leq 2c$, $N(x_1, 2, c) = (0, 1)$ and $\Gamma_1(x_1, 2, c) = (x_1/2, (1 + x_1)/2)$,
(iii) for $2c < x_1 \leq 1/2$, $N(x_1, 2, c) = (0, 1)$ and $\Gamma_1(x_1, 2, c) = (c, (1 + x_1)/2)$,
(iv) for $1/2 < x_1 < 1$, $N(x_1, 2, c) = (2x_1 - 1, 1)$ and $\Gamma_1(x_1, 2, c) = (c, (1 + x_1)/2)$.

Then $\mu_1(2, c) = P(X_2 \in N(X_1, 2, c)) = \int_0^c 2x_1 dx_1 + \int_c^{1/2} 1 dx_1 + \int_{1/2}^1 (2 - 2x_1) dx_1 = c^2 - c + 3/4$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, 2, c)) = \int_0^c (2x_1)^2 dx_1 + \int_c^{1/2} 1 dx_1 + \int_{1/2}^1 (2 - 2x_1)^2 dx_1 = 4c^3/3 - c + 2/3$.

$P_{NG} = P(X_2 \in N(X_1, 2, c), X_3 \in \Gamma_1(X_1, 2, c)) = \int_0^c (2x_1)(1/2) dx_1 + \int_c^{2c} (1/2) dx_1 + \int_{2c}^{1/2} ((1 + x_1)/2 - c) dx_1 + \int_{1/2}^1 (2 - 2x_1)((1 + x_1)/2 - c) dx_1 = 3c^2/2 - 5c/4 + 25/48$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, 2, c)) = \int_0^{2c} (1/2)^2 dx_1 + \int_{2c}^1 ((1 + x_1)/2 - c)^2 dx_1 = 2c^2 + 7/12 - 2c^3/3 - 3c/2$.

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2P_{NG} + P_{2G} = 2c^3/3 + 5c^2 - 5c + 55/24$. Hence $4\nu_1(2, c) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = 26c^3/3 + c + 1/24 - 4c^4 - 5c^2$.

Case 2: $1/4 < c \leq 1/2$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, 2, c)$ and $\Gamma_1(x_1, 2, c)$.

(i) for $0 < x_1 \leq c$, $N(x_1, 2, c) = (0, 2x_1)$ and $\Gamma_1(x_1, 2, c) = (x_1/2, (1 + x_1)/2)$,
(ii) for $c < x_1 \leq 1/2$, $N(x_1, 2, c) = (0, 1)$ and $\Gamma_1(x_1, 2, c) = (x_1/2, (1 + x_1)/2)$,
(iii) for $1/2 < x_1 \leq 2c$, $N(x_1, 2, c) = (2x_1 - 1, 1)$ and $\Gamma_1(x_1, 2, c) = (x_1/2, (1 + x_1)/2)$,
(iv) for $2c < x_1 < 1$, $N(x_1, 2, c) = (2x_1 - 1, 1)$ and $\Gamma_1(x_1, 2, c) = (c, (1 + x_1)/2)$.

Then $\mu_1(2, c) = P(X_2 \in N(X_1, 2, c)) = \int_0^c 2x_1 dx_1 + \int_c^{1/2} 1 dx_1 + \int_{1/2}^1 (2 - 2x_1) dx_1 = c^2 - c + 3/4$.

Next $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, 2, c)) = \int_0^c (2x_1)^2 dx_1 + \int_c^{1/2} 1 dx_1 + \int_{1/2}^1 (2 - 2x_1)^2 dx_1 = 4c^3/3 - c + 2/3$.

$P_{NG} = P(X_2 \in N(X_1, 2, c), X_3 \in \Gamma_1(X_1, 2, c)) = \int_0^c (2\,x_1)(1/2)dx_1 + \int_c^{1/2}(1/2)$ $dx_1 + \int_{1/2}^{2\,c}(2 - 2\,x_1)(1/2)dx_1 + \int_{2\,c}^1(2 - 2\,x_1)((1 + x_1)/2 - c)dx_1 = 5\,c^2/2 + 13/24 - 4\,c^3/3 - 3\,c/2$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, 2, c)) = \int_0^{2\,c}(1/2)^2 dx_1 + \int_{2\,c}^1((1+x_1)/2 - c)^2 dx_1 = -2\,c^3/3 + 2\,c^2 - 3\,c/2 + 7/12$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = 7\,c^2 + 7/3 - 2\,c^3 - 11\,c/2$. Hence $4\,\nu_1(2, c) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = 6\,c^3 + c/2 + 1/12 - 4\,c^4 - 3\,c^2$.

For $1/2 \le c < 1$, by symmetry we obtain $\mu_2(2, c) = \mu_1(2, 1 - c)$ and $\nu_2(2, c) = \nu_1(2, 1 - c)$. □

Proof of Theorem 7 There are two cases for r, namely $1 \le r < 2$ and $r \ge 2$.

Case 1: $1 \le r < 2$: In this case depending on the location of x_1, since $1 - r/2 < 1/2$ and $1/2 < 1/r$, the following are the different types of the combinations of $N(x_1, r, 1/2)$ and $\Gamma_1(x_1, r, 1/2)$.

(i) for $0 < x_1 \le 1 - r/2$, $N(x_1, r, 1/2) = (0, r\,x_1)$ and $\Gamma_1(x_1, r, 1/2) = (x_1/r, 1/2) = (a, 1/2)$ since $1/2 > 1 - (1 - x_1)/r$,
(ii) for $1 - r/2 < x_1 \le 1/2$, $N(x_1, r, 1/2) = (0, 1)$ and $\Gamma_1(x_1, r, 1/2) = (x_1/2, (1 + x_1)/2)$.

Let $a = x_1/r$ and $b = 1 - (1 - x)/r$. Then $\mu(r, 1/2) = P(X_2 \in N(X_1, r, 1/2)) = 2\,P(X_2 \in N(X_1, r, 1/2), X_1 \in (0, 1/2))$ by symmetry and $P(X_2 \in N(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/2} r\,x_1 dx_1 = r/8$. So $\mu(r, 1/2) = 2\,(r/8) = r/4$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, r, 1/2)) = 2\,P(\{X_2, X_3\} \subset N(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(\{X_2, X_3\} \subset N(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/2}(r\,x_1)^2 dx_1 = r^2/24$. So $P_{2N} = 2\,(r^2/24) = r^2/12$.

$P_{NG} = P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2)) = 2\,P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2) = \int_0^{1-r/2}(r\,x_1)(1/2 - a)dx_1 + \int_{1-r/2}^{1/2} r\,x_1(b - a)dx_1 = r^2/8 + 1/24 - r^3/48 - r/8$. So $P_{NG} = r^2/4 + 1/12 - r^3/24 - r/4$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2)) = 2\,P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1-r/2}(1/2 - a)^2 dx_1 + \int_{1-r/2}^{1/2}(b - a)^2 dx_1 = \frac{5r^3 - 12r^2 + 12r - 4}{24r^2}$. So $P_{2G} = \frac{5r^3 - 12r^2 + 12r - 4}{12r^2}$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = \frac{7r^4 + 12r - r^5 - r^3 - 10r^2 - 4}{12r^2}$. Hence $4\,\nu(r, 1/2) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = \frac{4r^4 + 12r - r^5 - r^3 - 10r^2 - 4}{12r^2}$.

Case 2: $r \ge 2$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, r, 1/2)$ and $\Gamma_1(x_1, r, 1/2)$.

(i) for $0 < x_1 \le 1/r$, $N(x_1, r, 1/2) = (0, r\,x_1)$ and $\Gamma_1(x_1, r, 1/2) = (a, b)$,
(ii) for $1/r < x_1 \le 1/2$, $N(x_1, r, 1/2) = (0, 1)$ and $\Gamma_1(x_1, r, 1/2) = (a, b)$,

Then $\mu(r, 1/2) = P(X_2 \in N(X_1, r, 1/2)) = 2\,P(X_2 \in N(X_1, r, 1/2), X_1 \in (0, 1/2))$ by symmetry and $P(X_2 \in N(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/r} r\,x_1 dx_1 + \int_{1/r}^{1/2} 1 dx_1 = (1 - 1/r)/2$. So $\mu(r, 1/2) = 1 - 1/r$.

Next $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, r, 1/2)) = 2\,P(\{X_2, X_3\} \subset N(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(\{X_2, X_3\} \subset N(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/r} (r\,x_1)^2 dx_1 + \int_{1/r}^{1/2} 1 dx_1 = (1 - 4/(3\,r))/2$. So $P_{2N} = 2\,(r^2/24) = 1 - 4/(3\,r)$.

$P_{NG} = P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2)) = 2\,P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(X_2 \in N(X_1, r, 1/2), X_3 \in \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2) = \int_0^{1/r} (r\,x_1)(b-a) dx_1 + \int_{1/r}^{1/2} (b-a) dx_1 = \frac{(r-1)^2}{2\,r^2}$. So $P_{NG} = \frac{(r-1)^2}{r^2}$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2)) = 2\,P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2))$ and $P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, 1/2), X_1 \in (0, 1/2)) = \int_0^{1/2} (b-a)^2 dx_1 + \int_{1-r/2}^{1/2} (b-a)^2 dx_1 = \frac{(r-1)^2}{2\,r^2}$. So $P_{2G} = \frac{(r-1)^2}{r^2}$

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = \frac{12\,r^2 - 22\,r + 9}{3\,r^2}$. Hence

$$4\,\nu(r, 1/2) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = \frac{2\,r - 3}{3\,r^2}.$$

□

Proof of Theorem 8 First we consider $0 < c \leq 1/2$, for which there are two cases, namely **Case 1:** $1/4 < c \leq 1/2$ and **Case 2:** $0 < c \leq 1/4$.

Case 1-I: $1 \leq r < 1/(1-c)$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1, r, c)$ and $\Gamma_1(x_1, r, c)$.

(i) for $0 < x_1 \leq 1 - r\,(1-c)$, $N(x_1, r, c) = (0, r\,x_1)$ and $\Gamma_1(x_1, r, c) = (a, c)$,
(ii) for $1 - r\,(1-c) < x_1 \leq c$, $N(x_1, r, c) = (0, r\,x_1)$ and $\Gamma_1(x_1, r, c) = (a, b)$,
(iii) for $c < x_1 \leq c\,r$, $N(x_1, r, c) = (1 - r\,(1 - x_1), 1)$ and $\Gamma_1(x_1, r, c) = (a, b)$,
(iv) for $c\,r < x_1 < 1$, $N(x_1, r, c) = (1 - r\,(1 - x_1), 1)$ and $\Gamma_1(x_1, r, c) = (c, b)$,

Then $\mu_1(r, c) = P(X_2 \in N(X_1, r, c)) = \int_0^c r\,x_1 dx_1 + \int_c^1 r\,(1 - x_1) dx_1 = c^2 r - c\,r + r/2$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, r, c)) = \int_0^c (r\,x_1)^2 dx_1 + \int_c^1 r^2 (1 - x_1)^2 dx_1 = c^2 r^2 - c\,r^2 + r^2/3$.

$P_{NG} = P(X_2 \in N(X_1, r, c), X_3 \in \Gamma_1(X_1, r, c)) = \int_0^{1-r\,(1-c)} (r\,x_1)(c-a) dx_1 + \int_{1-r\,(1-c)}^c r\,x_1 (b-a) dx_1 + \int_c^{c\,r} r\,(1 - x_1)(b-a) dx_1 + \int_{c\,r}^1 r\,(1 - x_1)(b-c) dx_1 = -c^2 r^3/2 + c^2 r^2 + c\,r^3/2 + c^2 r - c\,r^2 - r^3/6 - c^2 - c\,r + r^2/2 + c - 1/6$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, c)) = \int_0^{1-r\,(1-c)} (c-a)^2 dx_1 + \int_{1-r\,(1-c)}^{c\,r} (b-a)^2 dx_1 + \int_{c\,r}^1 (b-c)^2 dx_1 = \frac{3\,c^2 r^3 - 3\,c\,r^3 + 2\,r^3 - 3\,r^2 + 3\,r - 1}{3\,r^2}$.

Therefore

$$\begin{aligned} 4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = [&9\,c^2 r^4 - 3\,c^2 r^5 + 3\,c\,r^5 + 9\,c^2 r^3 - 9\,c\,r^4 \\ &- r^5 - 6\,c^2 r^2 - 9\,c\,r^3 + 4\,r^4 + 6\,c\,r^2 + 2\,r^3 - 4\,r^2 \\ &+ 3\,r - 1] \Big/ [3\,r^2]. \end{aligned}$$

Hence

$$4\,\kappa_1(r,c) = 4\,\mathbf{Cov}[h_{12},h_{13}] = [24\,c^3r^4 - 12\,c^4r^4 - 3\,c^2r^5 - 15\,c^2r^4 + 3\,c\,r^5 + 9\,c^2r^3 + 3\,c\,r^4 - r^5 - 6\,c^2r^2 - 9\,c\,r^3 + r^4 + 6\,c\,r^2 + 2\,r^3 - 4\,r^2 + 3\,r - 1] \Big/ [3\,r^2].$$

Case 1-II: $1/(1-c) \le r < 1/c$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1,r,c)$ and $\Gamma_1(x_1,r,c)$.

(i) for $0 < x_1 \le c$, $N(x_1,r,c) = (0, r\,x_1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(ii) for $c < x_1 \le 1 - 1/r$, $N(x_1,r,c) = (0,1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(iii) for $1 - 1/r < x_1 \le c\,r$, $N(x_1,r,c) = (1 - r\,(1-x_1), 1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(iv) for $c\,r < x_1 < 1$, $N(x_1,r,c) = (1 - r\,(1-x_1), 1)$ and $\Gamma_1(x_1,r,c) = (c,b)$,

Then $\mu_1(r,c) = P(X_2 \in N(X_1,r,c)) = \int_0^c r\,x_1 dx_1 + \int_c^{1-1/r} 1 dx_1 + \int_{1-1/r}^1 r\,(1-x_1)dx_1 = \frac{c^2r^2 - 2\,c\,r + 2\,r - 1}{2\,r}$.

Next $P_{2N} = P(\{X_2, X_3\} \subset N(X_1,r,c)) = \int_0^c (r\,x_1)^2 dx_1 + \int_c^{1-1/r} 1 dx_1 + \int_{1-1/r}^1 r^2(1-x_1)^2 dx_1 = \frac{c^3r^3 - 3\,c\,r + 3\,r - 2}{3\,r}$.

$P_{NG} = P(X_2 \in N(X_1,r,c), X_3 \in \Gamma_1(X_1,r,c)) = \int_0^c (r\,x_1)(b-a)dx_1 + \int_c^{1-1/r}(b-a)dx_1 + \int_{1-1/r}^{c\,r} r\,(1-x_1)(b-a)dx_1 + \int_{c\,r}^1 r\,(1-x_1)(b-c)dx_1 = \frac{3\,c^2r^4 - c^3r^5 + 3\,c^2r^3 - 3\,c^2r^2 - 3\,c\,r^3 - 6\,c\,r^2 + 6\,c\,r + 7\,r^2 - 9\,r + 3}{6\,r^2}$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1,r,c)) = \int_0^{c\,r} (b-a)^2 dx_1 + \int_{c\,r}^1 (b-c)^2 dx_1 = \frac{3\,c^2r^3 - c^3r^3 - 6\,c\,r^2 + 3\,c\,r + 3\,r^2 - 3\,r + 1}{3\,r^2}$.

Therefore

$$4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = \frac{c^3r^4 - c^3r^5 - c^3r^3 + 3\,c^2r^4 + 6\,c^2r^3 - 3\,c^2r^2 - 3\,c\,r^3 - 15\,c\,r^2 + 9\,c\,r + 13\,r^2 - 14\,r + 4}{3\,r^2}.$$

Hence

$$4\,\kappa_2(r,c) = 4\,\mathbf{Cov}[h_{12},h_{13}] = [c^3r^4 - 3\,c^4r^4 - c^3r^5 + 11\,c^3r^3 + 3\,c^2r^4 - 6\,c^2r^3 - 9\,c^2r^2 - 3\,c\,r^3 + 9\,c\,r^2 - 3\,c\,r + r^2 - 2\,r + 1] \Big/ [3\,r^2].$$

Case 1-III: $r \ge 1/c$: In this case depending on the location of x_1, the following are the different types of the combinations of $N(x_1,r,c)$ and $\Gamma_1(x_1,r,c)$.

(i) for $0 < x_1 \le 1/r$, $N(x_1,r,c) = (0, r\,x_1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(ii) for $1/r < x_1 \le c$, $N(x_1,r,c) = (0,1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(iii) for $c < x_1 \le 1 - 1/r$, $N(x_1,r,c) = (0,1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,
(iii) for $1-1/r < x_1 < 1$, $N(x_1,r,c) = (1-r\,(1-x_1), 1)$ and $\Gamma_1(x_1,r,c) = (a,b)$,

Then $\mu_1(r,c) = P(X_2 \in N(X_1,r,c)) = \int_0^{1/r} r\,x_1 dx_1 + \int_{1/r}^{1-1/r} 1 dx_1 + \int_{1-1/r}^1 r(1-x_1)dx_1 = 1 - 1/r$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, r, c)) = \int_0^{1/r} (r\,x_1)^2 dx_1 + \int_{1/r}^{1-1/r} 1 dx_1 + \int_{1-1/r}^{1} r^2(1-x_1)^2 dx_1 = 1 - 4/(3\,r)$.

$P_{NG} = P(X_2 \in N(X_1, r, c), X_3 \in \Gamma_1(X_1, r, c)) = \int_0^{1/r} (r\,x_1)(b-a) dx_1 + \int_{1/r}^{1-1/r} (b-a) dx_1 + \int_{1-1/r}^{1} r\,(1-x_1)(b-a) dx_1 = (r-1)^2/r^2$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, c)) = \int_0^1 (b-a)^2 dx_1 = (r-1)^2/r^2$.

Therefore $4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = \frac{12\,r^2-22\,r+9}{3\,r^2}$. Hence we obtain $4\,\kappa_3(r, c) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = \frac{2\,r-3}{3\,r^2}$.

Case 2: $0 < c \leq 1/4$: In this case, the calculations for $\mu_1(r, c)$ are as in **Case 1**.

Case 2-I: $1 \leq r < 1/(1-c)$ is same as **Case 1-I**, **Case 2-V:** $1 \leq r < 1/(1-c)$ is same as **Case 1-III**, **Case 2-II:** $1/(1-c) \leq r < (1-\sqrt{1-4c})/(2c)$ and **Case 2-IV:** $(1+\sqrt{1-4c})/(2c) \leq r < 1/c$ are same as **Case 1-II**. However, we have a new possible range of r in this case for the calculation of $\nu(r, c)$. **Case 2-III:** $(1-\sqrt{1-4c})/(2c) \leq r < (1+\sqrt{1-4c})/(2c)$: Depending on the location of x_1, the following are the different types of the combinations of $N(x_1, r, c)$ and $\Gamma_1(x_1, r, c)$.

(i) for $0 < x_1 \leq c$, $N(x_1, r, c) = (0, r\,x_1)$ and $\Gamma_1(x_1, r, c) = (a, b)$,
(ii) for $c < x_1 \leq c\,r$, $N(x_1, r, c) = (0, 1)$ and $\Gamma_1(x_1, r, c) = (a, b)$,
(iii) for $c\,r < x_1 \leq 1 - 1/r$, $N(x_1, r, c) = (0, 1)$ and $\Gamma_1(x_1, r, c) = (c, b)$,
(iv) for $1-1/r < x_1 < 1$, $N(x_1, r, c) = (1-r\,(1-x_1), 1)$ and $\Gamma_1(x_1, r, c) = (c, b)$,

Then $\mu_1(r, c) = P(X_2 \in N(X_1, r, c)) = \int_0^c r\,x_1 dx_1 + \int_c^{1-1/r} 1 dx_1 + \int_{1-1/r}^{1} r\,(1-x_1) dx_1 = \frac{c^2r^2-2\,c\,r+2\,r-1}{2\,r}$.

For $\mathbf{Cov}(h_{12}, h_{13})$, we need to calculate P_{2N}, P_{NG}, and P_{2G}. The probability $P_{2N} = P(\{X_2, X_3\} \subset N(X_1, r, c)) = \int_0^c (r\,x_1)^2 dx_1 + \int_c^{1-1/r} 1 dx_1 + \int_{1-1/r}^{1} r^2(1-x_1)^2 dx_1 = \frac{c^3r^3-3\,c\,r+3\,r-2}{3\,r}$.

$P_{NG} = P(X_2 \in N(X_1, r, c), X_3 \in \Gamma_1(X_1, r, c)) = \int_0^c (r\,x_1)(b-a) dx_1 + \int_c^{c\,r} (b-a) dx_1 + \int_{c\,r}^{1-1/r} (b-c) dx_1 + \int_{1-1/r}^{1} r\,(1-x_1)(b-c) dx_1 = [6\,c^2r^4 - 3\,c^2r^3 - 12\,c\,r^3 + 9\,c\,r^2 + 6\,r^3 - 6\,r^2 + 1] \big/ [6\,r^3]$.

Finally, $P_{2G} = P(\{X_2, X_3\} \subset \Gamma_1(X_1, r, c)) = \int_0^{c\,r} (b-a)^2 dx_1 + \int_{c\,r}^{1} (b-c)^2 dx_1 = \frac{3\,c^2r^3 - c^3r^3 - 6\,c\,r^2 + 3\,c\,r + 3\,r^2 - 3\,r + 1}{3\,r^2}$.

Therefore

$$4\,\mathbf{E}[h_{12}h_{13}] = P_{2N} + 2\,P_{NG} + P_{2G} = \frac{c^3r^5 - c^3r^4 + 9\,c^2r^4 - 3\,c^2r^3 - 21\,c\,r^3 + 12\,c\,r^2 + 12\,r^3 - 11\,r^2 + r + 1}{3\,r^3}.$$

Hence

$$4\,\vartheta_3(r, c) = 4\,\mathbf{Cov}[h_{12}, h_{13}] = \frac{3\,c^4r^5 - c^3r^5 - 11\,c^3r^4 + 3\,c^2r^4 + 9\,c^2r^3 - 3\,c\,r^3 - r^2 + 2\,r - 1}{3\,r^3}.$$

For $1/2 \leq c < 1$, by symmetry, it follows that $\mu_2(r, c) = \mu_1(r, 1-c)$ and $\nu_2(r, c) = \nu_1(r, 1-c)$. □

Appendix 2: Proofs for the multiple interval case

We give the proof of Theorem 12 first.

Proof of Theorem 12 Recall that $\widetilde{\rho}_{n,m}(r,c)$ is the relative arc density of the PCD for the $m \geq 2$ case. Then it follows that $\widetilde{\rho}_{n,m}(r,c)$ is a U-statistic of degree two, so we can write it as $\widetilde{\rho}_{n,m}(r,c) = \frac{2}{n(n-1)} \sum_{i<j} h_{ij}$ where $h_{ij} = (g_{ij} + g_{ji})/2$. Then the expectation of $\widetilde{\rho}_{n,m}(r,c)$ is

$$\begin{aligned}\mathbf{E}\left[\widetilde{\rho}_{n,m}(r,c)\right] &= \frac{2}{n\,(n-1)} \sum \sum_{i<j} \mathbf{E}\left[h_{ij}\right] = \mathbf{E}\left[h_{12}\right] = \mathbf{E}\left[g_{12}\right] \\ &= P((X_1, X_2) \in \mathcal{A}) = \widetilde{\mu}(m,r,c).\end{aligned}$$

But, by definition of $N(\cdot, r, c)$, if X_1 and X_2 are in different intervals, then $P((X_1, X_2) \in \mathcal{A}) = 0$. So, by the law of total probability, we have

$$\begin{aligned}\widetilde{\mu}(m,r,c) &:= P((X_1, X_2) \in \mathcal{A}) \\ &= \sum_{i=1}^{m+1} P((X_1, X_2) \in \mathcal{A} \,|\, \{X_1, X_2\} \subset \mathcal{I}_i)\, P(\{X_1, X_2\} \subset \mathcal{I}_i) \\ &= \sum_{i=2}^{m} \mu(r,c)\, P(\{X_1, X_2\} \subset \mathcal{I}_i) + \sum_{i \in \{1, m+1\}} \mu_e(r)\, P(\{X_1, X_2\} \subset \mathcal{I}_i) \\ &= \sum_{i=2}^{m} \mu(r,c)\, w_i^2 + \sum_{i \in \{1, m+1\}} \mu_e(r)\, w_i^2 \\ &= \mu(r,c) \sum_{i=2}^{m} w_i^2 + \mu_e(r) \sum_{i \in \{1, m+1\}} w_i^2.\end{aligned}$$

since $P(\{X_1, X_2\} \subset \mathcal{I}_i) = \left(\frac{y_{(i)} - y_{(i-1)}}{\delta_2 - \delta_1}\right)^2 = w_i^2$.

Furthermore, the asymptotic variance is

$$4\,\widetilde{v}(m,r,c) = 4\,\mathbf{E}\left[h_{12}h_{13}\right] - \mathbf{E}\left[h_{12}\right]\mathbf{E}\left[h_{13}\right] = 4\,\mathbf{E}\left[h_{12}h_{13}\right] - (\widetilde{\mu}(m,r,c))^2$$

where $4\,\mathbf{E}\left[h_{12}h_{13}\right] = \widetilde{P}_{2N} + 2\widetilde{P}_{NG} + \widetilde{P}_{2G}$ with,

$$\begin{aligned}\widetilde{P}_{2N} &= \sum_{i=2}^{m} P(\{X_2, X_3\} \subset N(X_1, r, c) \,|\, \{X_1, X_2, X_3\} \subset \mathcal{I}_i)\, P(\{X_1, X_2, X_3\} \subset \mathcal{I}_i) \\ &\quad + \sum_{i \in \{1, m+1\}} P(\{X_2, X_3\} \subset N_e(X_1, r) \,|\, \{X_1, X_2, X_3\} \subset \mathcal{I}_i) \\ &\quad \times P(\{X_1, X_2, X_3\} \subset \mathcal{I}_i) \\ &= \sum_{i=2}^{m} P_{2N}\, P(\{X_1, X_2, X_3\} \subset \mathcal{I}_i) + \sum_{i \in \{1, m+1\}} P_{2N,e}\, P(\{X_1, X_2, X_3\} \subset \mathcal{I}_i)\end{aligned}$$

$$\approx \sum_{i=2}^{m} P_{2N}\, w_i^3 + \sum_{i\in\{1,m+1\}} P_{2N,e}\, w_i^3 = P_{2N} \sum_{i=2}^{m} w_i^3 + P_{2N,e} \sum_{i\in\{1,m+1\}} w_i^3.$$

since $P(\{X_2, X_3\} \subset N(X_1, r, c) \,|\, \{X_1, X_2, X_3\} \subset \mathcal{I}_i)$ is P_{2N} for middle intervals and $P_{2N,e}$ for the end intervals and $P(\{X_1, X_2, X_3\} \subset \mathcal{I}_i) = \left(\frac{y_{(i)}-y_{(i-1)}}{\delta_2-\delta_1}\right)^3 = w_i^3$. Similarly,

$$\widetilde{P}_{NG} = P_{NG} \sum_{i=2}^{m} w_i^3 + P_{NG,e} \sum_{i\in\{1,m+1\}} w_i^3$$

and

$$\widetilde{P}_{2G} = P_{2G} \sum_{i=2}^{m} w_i^3 + P_{2G,e} \sum_{i\in\{1,m+1\}} w_i^3.$$

Therefore,

$$\begin{aligned} 4\,\widetilde{\nu}(m, r, c) = {} & (P_{2N} + 2\,P_{NG} + P_{2G}) \sum_{i=2}^{m} w_i^3 \\ & + (P_{2N,e} + 2\,P_{NG,e} + P_{2G,e}) \sum_{i\in\{1,m+1\}} w_i^3 - (\widetilde{\mu}(m, r, c))^2. \end{aligned}$$

Hence the desired result follows. □

Proof of Theorem 11 Recall that $\rho_{n,m}(r, c)$ is the version I of the relative arc density of the PCD for the $m > 2$ case. Moreover, $\rho_{n,m}(r, c) = \frac{n(n-1)}{n_T} \widetilde{\rho}_{n,m}(r, c)$. Then the expectation of $\rho_{n,m}(r, c)$, for large n_i and n, is

$$\mathbf{E}\left[\rho_{n,m}(r, c)\right] = \frac{n(n-1)}{n_T} \mathbf{E}[\widetilde{\rho}_{n,m}(r, c)] \approx \widetilde{\mu}(m, r, c) \left(\sum_{i=1}^{m+1} w_i^2\right)^{-1}$$

since $\frac{n(n-1)}{n_T} = \left(\sum_{i=1}^{m+1} n_i(n_i - 1)/(n(n-1))\right)^{-1} \approx \left(\sum_{i=1}^{m+1} w_i^2\right)^{-1}$ for large n_i and n. Here $\widetilde{\mu}(m, r, c)$ is as in Theorem 12.

Moreover, the asymptotic variance of $\rho_{n,m}(r, c)$, for large n_i and n, is

$$4\,\breve{\nu}(m, r, c) = \frac{n^2(n-1)^2}{n_T^2} 4\,\widetilde{\nu}(m, r, c) = 4\,\widetilde{\nu}(m, r, c) \left(\sum_{i=1}^{m+1} w_i^2\right)^{-2}$$

since

$$\frac{n^2(n-1)^2}{n_T^2} = \left(\sum_{i=1}^{m+1} n_i(n_i-1)/(n(n-1))\right)^{-2} \approx \left(\sum_{i=1}^{m+1} w_i^2\right)^{-2}$$

for large n_i and n. Here $\widetilde{\nu}(m,r,c)$ is as in Theorem 12. Hence the desired result follows. □

References

Callaert H, Janssen P (1978) The Berry-Esseen theorem for U-statistics. Ann Stat 6:417–421

Cannon A, Cowen L (2000) Approximation algorithms for the class cover problem. In: Proceedings of the 6th international symposium on artificial intelligence and mathematics

Ceyhan E (2005) An investigation of proximity catch digraphs in delaunay tessellations, also available as technical monograph titled "proximity catch digraphs: auxiliary tools, properties, and applications" by vdm verlag, isbn: 978-3-639-19063-2. PhD thesis, The Johns Hopkins University, Baltimore

Ceyhan E (2010) Spatial clustering tests based on domination number of a new random digraph family. Commun Stat Theory Methods, doi:101080/03610921003597211 (to appear)

Ceyhan E, Priebe C (2003) Central similarity proximity maps in Delaunay tessellations. In: Proceedings of the joint statistical meeting, statistical computing section, American statistical association

Ceyhan E, Priebe CE (2005) The use of domination number of a random proximity catch digraph for testing spatial patterns of segregation and association. Stat Probab Lett 73:37–50

Ceyhan E, Priebe CE (2007) On the distribution of the domination number of a new family of parametrized random digraphs. Model Assist Stat Appl 1(4):231–255

Ceyhan E, Priebe CE, Wierman JC (2006) Relative density of the random r-factor proximity catch digraphs for testing spatial patterns of segregation and association. Comput Stat Data Anal 50(8):1925–1964

Ceyhan E, Priebe CE, Marchette DJ (2007) A new family of random graphs for testing spatial segregation. Can J Stat 35(1):27–50

Chartrand G, Lesniak L (1996) Graphs & digraphs. Chapman & Hall/CRC Press LLC, Florida

DeVinney J, Priebe CE (2006) A new family of proximity graphs: class cover catch digraphs. Discr Appl Math 154(14):1975–1982

DeVinney J, Priebe CE, Marchette DJ, Socolinsky D (2002) Random walks and catch digraphs in classification. http://www.galaxy.gmu.edu/interface/I02/I2002/, Proceedings/DeVinney Jason/DeVinneyJason.paper.pdf, proceedings of the 34th symposium on the interface: computing science and statistics, vol 34

Erdős P, Rényi A (1959) On random graphs I. Publ Math (Debrecen) 6:290–297

Janson S, Łuczak T, Ruciński A (2000) Random graphs. Wiley-Interscience series in discrete mathematics and optimization. Wiley, New York

Jaromczyk JW, Toussaint GT (1992) Relative neighborhood graphs and their relatives. Proc IEEE 80:1502–1517

Marchette DJ, Priebe CE (2003) Characterizing the scale dimension of a high dimensional classification problem. Pattern Recogn 36(1):45–60

Priebe CE, DeVinney JG, Marchette DJ (2001) On the distribution of the domination number of random class cover catch digraphs. Stat Probab Lett 55:239–246

Priebe CE, Marchette DJ, DeVinney J, Socolinsky D (2003) Classification using class cover catch digraphs. J Classif 20(1):3–23

Priebe CE, Solka JL, Marchette DJ, Clark BT (2003) Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays. Comput Stat Data Anal Visual 43(4):621–632

Prisner E (1994) Algorithms for interval catch digraphs. Discr Appl Math 51:147–157

Randles RH, Wolfe DA (1979) Introduction to the theory of nonparametric statistics. Wiley, New York

Sen M, Das S, Roy A, West D (1989) Interval digraphs: an analogue of interval graphs. J Graph Theory 13:189–202

Toussaint GT (1980) The relative neighborhood graph of a finite planar set. Pattern Recogn 12(4):261–268

Tuza Z (1994) Inequalities for minimal covering sets in sets in set systems of given rank. Discr Appl Math 51:187–195