



Domination number of an interval catch digraph family and its use for testing uniformity

Elvan Ceyhan

Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA

ABSTRACT

We consider a special type of interval catch digraph (ICD) for one-dimensional data in a randomized setting and propose its use for testing uniformity. These ICDs are defined with expansion and centrality parameters, hence are called parameterized ICDs (PICDs). We derive the exact (and asymptotic) distribution of the domination number of this PICD when its vertices are from a uniform (and non-uniform) distribution in one dimension for the entire parameter ranges; thereby determine the parameters for which the asymptotic distribution is non-degenerate. We use the domination number for testing uniformity of one-dimensional data, prove its consistency against certain alternatives, and compare it with commonly used and recently proposed tests in literature and also arc density of two ICD families in terms of size and power. Based on our Monte Carlo simulations, we demonstrate that PICD domination number has higher power for certain types of alternatives compared to other tests.

ARTICLE HISTORY

Received 6 January 2019
Accepted 7 January 2020

KEYWORDS

Arc density; class cover catch digraph; consistency; proximity catch digraph; uniform distribution


AMS 2000 SUBJECT CLASSIFICATIONS

05C80; 05C20; 60D05; 60C05; 62E20

1. Introduction

Graphs and digraphs for one dimensional points as vertices have been extensively studied and have far-reaching applications despite their simplicity. In this article, we introduce an interval catch digraph (ICD) family, provide the distribution of its domination number for random vertices, and employ the domination number in testing uniformity of one-dimensional data. Interval graphs and digraphs have applications in many fields such as chronological ordering of artifacts in archeology, modelling traffic lights in transportation, food web models in ecology, document localization, classification of RNA structures and so on (see [1–4]). ICDs were introduced as a special type of interval digraphs and found applications in various fields (see [5,6] for a characterization and detailed discussion of ICDs). The new digraph family we consider in this article is parameterized by an expansion parameter and a centrality parameter. We demonstrate that this digraph family is actually an ICD family, hence it is referred to as parameterized ICD (PICD). A *digraph* is a directed graph with vertex set \mathcal{V} and arcs (directed edges) each of which is from one vertex to another based on a binary relation. The pair $(p, q) \in \mathcal{V} \times \mathcal{V}$ is an ordered pair which stands for an *arc* from vertex p to vertex q in \mathcal{V} .

CONTACT Elvan Ceyhan  ceyhan@auburn.edu

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02331888.2020.1720020>

© 2020 Informa UK Limited, trading as Taylor & Francis Group

The PICDs are closely related to the *class cover problem* (CCP) of [7] which is motivated by applications in statistical classification. To properly describe the CCP problem, let (Ω, d) be a metric space with a dissimilarity function $d: \Omega \times \Omega \rightarrow \mathbb{R}$ such that $d(a, b) = d(b, a) \geq d(a, a) = 0$ for all $a, b \in \Omega$. Let $\mathcal{X}_n = \{X_1, X_2, \dots, X_n\}$ and $\mathcal{Y}_m = \{Y_1, Y_2, \dots, Y_m\}$ be two sets of i.i.d. Ω -valued random variables from classes \mathcal{X} and \mathcal{Y} , with class-conditional distributions F_X and F_Y , respectively. We also assume that each X_i is independent of each Y_j and all $X_i \in \mathcal{X}_n$ and all $Y_j \in \mathcal{Y}_m$ are distinct with probability one, and $(X_i, Y_j) \sim F_{X,Y}$ (i.e. (X_i, Y_j) has joint distribution $F_{X,Y}$ with the marginal distributions F_X for X_i and F_Y for Y_j). The CCP for a target class refers to finding a collection of neighbourhoods, \mathcal{N} around X_i , denoted $N(X_i) \in \mathcal{N}$, such that (i) $\mathcal{X}_n \subseteq (\cup_i N(X_i))$ and (ii) $\mathcal{Y}_m \cap (\cup_i N(X_i)) = \emptyset$. The neighbourhood $N(X_i)$ is a subset of Ω , containing X_i , and is defined based on the dissimilarity d (between X_i and \mathcal{Y}_m). A collection of neighbourhoods satisfying both conditions is called a *class cover*. Clearly, it follows by condition (i) that the set of all covering regions (i.e. neighbourhoods $N(X_i)$ around X_i) is a class cover; however, the goal is to have a class cover for \mathcal{X}_n that has as few points as possible. Thus, e.g. in statistical learning, the classification will be less complex while most of the relevant information being kept. Hence, the CCP considered here is a *minimum-cardinality class cover*. One can convert the CCP to the graph theoretical problem of finding dominating sets. In particular, our ICD is the digraph $D = (\mathcal{V}, \mathcal{A})$ with vertex set $\mathcal{V} = \mathcal{X}_n$ and arc set \mathcal{A} such that there is an arc $(X_i, X_j) \in \mathcal{A}$ iff $X_j \in N(X_i)$. It is easy to see that solving the CCP is equivalent to finding a minimum domination set of the corresponding PICD, hence *cardinality of a solution to CCP is equal to the domination number of the associated digraph* (see [8]). Hence the tool introduced in this article can be seen as a parameterized extension to the original CCP problem of [7]. That is, the cardinality of the smallest cover (i.e. the domination number) is investigated when the cover(ing) regions, $N(X_i)$, depend on two parameters and the distribution of this cardinality is based on $N(X_i)$ (hence the parameters) and $F_{X,Y}$.

Our PICDs are *random digraphs* (according to the digraph version of classification of [9]) in which each vertex corresponds to a data point and arcs are defined in terms of some bivariate relation on the data, and are also related to the class cover catch digraph (CCCD) introduced by Priebe et al. [10] who derived the exact distribution of its domination number for uniform data from two classes in \mathbb{R} . A CCCD consists of a vertex set in \mathbb{R}^d and arcs (u, v) if v is inside the ball centred at u with a radius based on spatial proximity of the points. CCCDs were also extended to higher dimensions and were demonstrated to be a competitive alternative to the existing methods in classification (see [11] and references therein) and to be robust to the class imbalance problem [12]. Furthermore, a CLT result for CCCD based on one-dimensional data is proved [13] and the distribution of the domination number of CCCDs is also derived for non-uniform data [14].

We investigate the distribution of domination number of the PICDs for data in $\Omega = \mathbb{R}$. The domination in graphs has been studied extensively in recent decades (see, e.g. [15] and the references therein and [16]), and domination in digraphs has received comparatively less attention but is also studied in literature (see, e.g. [17–19]). We provide the exact and asymptotic distributions of the domination number of PICDs with vertices from uniform (and non-uniform) one-dimensional distributions. Some special cases and bounds for the domination number of PICDs are handled first, then the domination number is investigated for uniform data in one interval (in \mathbb{R}) and the analysis is generalized to uniform data in multiple intervals and to non-uniform data in one and multiple intervals.

We use domination number in testing uniformity of one-dimensional data. Testing uniformity is important in its own right in numerous fields, e.g. in assessing the quality of random number generators [20] and in chemical processes [21]. Furthermore, testing that data come from a particular distribution can be reduced to testing uniformity, hence uniformity tests are of great importance for goodness-of-fit tests (see [22] and references therein). Some graph theoretical tools are employed (although not so commonly) in two-sample testing [23] and in testing uniformity; for example, Jain et al. [24] use minimum spanning trees and Ceyhan [25] use the arc density of another family of ICDs for this purpose. Moreover, Ceyhan [26] provide the probabilistic investigation of the arc density for the PICD of this article, but it is not applied for uniformity testing previously. In [14], the distribution of the domination number of CCCDs is studied when vertices are from a non-uniform one-dimensional distribution, but the domination number of the PICD introduced here is not studied previously. To the author's knowledge *domination number is not used in literature for testing uniformity*. We compare the size and power performance of our test with two well known competitors, namely Kolmogorov–Smirnov (KS) test and Pearson's χ^2 goodness-of-fit test, and the arc density of PICDs and of another ICD family, and also a uniformity test which is based on Too-Lin characterization of the uniform distribution due to [22], and two entropy-based tests due to [27]. We demonstrate that the test based on the domination number has higher power for certain types of deviations from uniformity. Furthermore, this article forms the foundation of the extensions of the methodology to higher dimensions. The domination number has other applications, e.g. in testing spatial point patterns (see, e.g. [28]) and our results can help make the power comparisons possible for a large family of alternative patterns in such a setting. Some trivial proofs regarding PICDs are omitted, while others are mostly deferred to the Supplementary File.

We define the PICDs and their domination number in Section 2, provide the exact and asymptotic distributions of the domination number of PICDs for uniform data in one interval in Section 3, discuss the distribution of the domination number for data from a general distribution in Section 4. We extend these results to multiple intervals in Section 5, use domination number in testing uniformity in Section 6, prove consistency of the domination number tests under certain alternatives in Section 7, and provide discussion and conclusions in Section 8.

2. A parameterized random interval catch digraph family

Let $N : \Omega \rightarrow \wp(\Omega)$ be a map where $\wp(\Omega)$ represents the power set of Ω . Then 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 $\Gamma_1(\cdot, N) : \wp(\Omega) \rightarrow \wp(\Omega)$ that associates the region $\Gamma_1(B, N) := \{z \in \Omega : B \subseteq N(z)\}$ with the set B . For a point $x \in \Omega$, for convenience, 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 can be defined for a set of points. The PICD has the vertex set $\mathcal{V} = \mathcal{X}_n$ and arc set \mathcal{A} defined by $(X_i, X_j) \in \mathcal{A}$ iff $X_j \in N(X_i)$.

Although the above definition of the proximity region does not require multiple classes, in this article, we will define proximity regions in a two-class setting based on relative allocation of points from one class (say \mathcal{X}) with respect to points from the other class (say \mathcal{Y}). We now get more specific and restrict our attention to $\Omega = \mathbb{R}$ and define N explicitly. Let \mathcal{Y}_m consist of m distinct points from class \mathcal{Y} and $Y_{(i)}$ be the i th order statistic (i.e. i th

smallest value) of \mathcal{Y}_m for $i = 1, 2, \dots, m$ with the additional notation for $i \in \{0, m + 1\}$ as

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

Then $Y_{(i)}$ values partition \mathbb{R} into $(m + 1)$ intervals which is called the *intervalization* of \mathbb{R} by \mathcal{Y}_m . Let also that $\mathcal{I}_i := (Y_{(i)}, Y_{(i+1)})$ for $i \in \{0, 1, 2, \dots, m\}$ and $M_{c,i} := Y_{(i)} + c(Y_{(i+1)} - Y_{(i)})$ (i.e. $M_{c,i} \in \mathcal{I}_i$ such that $c \times 100\%$ of length of \mathcal{I}_i is to the left of $M_{c,i}$). We define the parameterized proximity region with the expansion parameter $r \geq 1$ and centrality parameter $c \in [0, 1]$ for two one-dimensional data sets, \mathcal{X}_n and \mathcal{Y}_m , from classes \mathcal{X} and \mathcal{Y} , respectively, as follows (see also Figure 1). For $x \in \mathcal{I}_i$ with $i \in \{1, 2, \dots, m - 1\}$ (i.e. for x in the middle intervals)

$$N(x, r, c) = \begin{cases} (Y_{(i)}, \min(Y_{(i+1)}, Y_{(i)} + r(x - Y_{(i)}))) & \text{if } x \in (Y_{(i)}, M_{c,i}), \\ (\max(Y_{(i)}, Y_{(i+1)} - r(Y_{(i+1)} - x)), Y_{(i+1)}) & \text{if } x \in (M_{c,i}, Y_{(i+1)}). \end{cases} \quad (1)$$

Additionally, for $x \in \mathcal{I}_i$ with $i \in \{0, m\}$ (i.e. for x in the end intervals)

$$N(x, r, c) = \begin{cases} (Y_{(1)} - r(Y_{(1)} - x), Y_{(1)}) & \text{if } x < Y_{(1)}, \\ (Y_{(m)}, Y_{(m)} + r(x - Y_{(m)})) & \text{if } x > Y_{(m)}. \end{cases} \quad (2)$$

Notice that for $i \in \{0, m\}$, the proximity region does not have a centrality parameter c . For $x \in \mathcal{Y}_m$, we define $N(x, r, c) = \{x\}$ for all $r \geq 1$ and $c \in [0, 1]$. If $x = M_{c,i}$, then in Equation (1), we arbitrarily assign $N(x, r, c)$ to be one of the defining intervals. For $c = 0$, we have $(M_{c,i}, Y_{(i+1)}) = \mathcal{I}_i$ and for $c = 1$, we have $(Y_{(i)}, M_{c,i}) = \mathcal{I}_i$. So, we set $N(x, r, 0) := (\max(Y_{(i)}, Y_{(i+1)} - r(Y_{(i+1)} - x)), Y_{(i+1)})$ and $N(x, r, 1) := (Y_{(i)}, \min(Y_{(i+1)}, Y_{(i)} + r(x - Y_{(i)})))$. For $r > 1$, we have $x \in N(x, r, c)$ for all $x \in \mathcal{I}_i$. Furthermore, $\lim_{r \rightarrow \infty} N(x, r, c) = \mathcal{I}_i$ for all $x \in \mathcal{I}_i$, so we define $N(x, \infty, c) = \mathcal{I}_i$ for all such x .

The PICD has the vertex set \mathcal{X}_n and arc set \mathcal{A} defined by $(X_i, X_j) \in \mathcal{A}$ iff $X_j \in N(X_i, r, c)$. We denote such PICDs as $\mathbf{D}_{n,m}(F_{X,Y}, r, c)$. The randomness of the PICD lies in the fact that the vertices are randomly generated from the distribution F_X and proximity regions

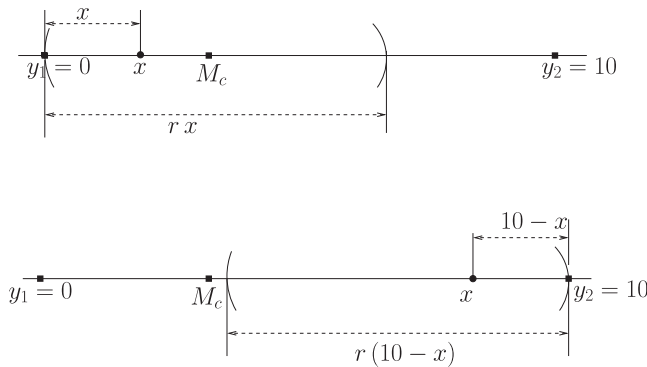


Figure 1. Illustrations of the construction of the parameterized proximity region, $N(x, r, c)$ with $c \in (0, 1/2)$ for $\mathcal{Y}_2 = \{y_1, y_2\}$ with $y_1 = 0$ and $y_2 = 10$ (hence $M_c = 10c$) and $x \in (0, M_c)$ (top) and $x \in (M_c, 10)$ (bottom).

are random depending on $F_{X,Y}$, but arcs (X_i, X_j) are deterministic functions of the random variable X_j and the random set $N(X_i)$. Notice that although N depends on \mathcal{Y}_m , we omit \mathcal{Y}_m for brevity in notation of proximity region $N(x, r, c)$.

2.1. Relation of PICDs with other graph families

Interval graphs are a special type of intersection graphs, which have emerged from a problem in genetics called Benzer problem (see [4] for details) and they have been extensively studied in graph theory since their introduction [2,29]. On the other hand, interval digraphs have recently gained attention after their introduction in [30] (see, e.g. [31]). Let \mathcal{V} be a set of n index points in some arbitrary space; for simplicity take $\mathcal{V} = \{1, 2, \dots, n\}$. Consider a set of ‘source’ intervals S_v and a set of ‘target’ intervals T_v in \mathbb{R} associated with $v \in \mathcal{V}$. The family of ordered pairs of these intervals $(S_v, U_v)_{v \in \mathcal{V}}$ such that $U_v \in S_v$ for each v is called a *nest representation* [6]. The digraph $D = (\mathcal{V}, \mathcal{A})$ is called an *interval nest digraph*, if there exists a nest representation with the index set \mathcal{V} such that $(i, j) \in \mathcal{A}$ iff $S_i \cap U_j \neq \emptyset$. *Interval catch digraphs* (ICDs) are interval nest digraphs with each T_v containing just one element [6]. In fact, for catch digraphs the nest representation constitutes a family of sets with points (or *pointed sets*) $(S_v, p_v)_{v \in \mathcal{V}}$ where each set S_v is associated with a base point $p_v \in S_v$. Then $D = (\mathcal{V}, \mathcal{A})$ is a catch digraph with $(i, j) \in \mathcal{A}$ iff $p_j \in S_i$. Such a catch digraph is called an *interval catch digraph*, if there is a totally ordered set (T, \leq) such that D is the catch digraph of a family of pointed intervals in T . Here, $I \subset T$ is an interval if, for all $x, y, z \in T$, $x \leq y \leq z$ and $x, z \in I$ imply that $y \in I$. For finite ICDs, T can always be taken as the real line (see, e.g. [5] who also provides a characterization of ICDs).

The PICDs are closely related to the *proximity graphs* of [32] and might be considered as one-dimensional versions of proportional-edge proximity catch digraphs of [28]. Furthermore, when $r = 2$ and $c = 1/2$ (i.e. $M_{c,i} = (Y_{(i)} + Y_{(i+1)})/2$) we have $N(x, r, c) = B(x, r(x))$ where $B(x, r(x))$ is the ball centred at x with radius $r(x) = d(x, \mathcal{Y}_m) = \min_{y \in \mathcal{Y}_m} d(x, y)$. The region $N(x, 2, 1/2)$ corresponds to the proximity region which gives rise to the CCCD of [10]. Note also that, $N(x, r, c)$ can be viewed as a *homothetic transformation (enlargement)* with $r \geq 1$ applied on a translation of the region $N(x, 1, c)$. Furthermore, this transformation is also an *affine similarity transformation*. Since (\mathbb{R}, \leq) is a total order, by the characterization theorem of [33], our random digraph is clearly an interval catch digraph, since there exists a total order ‘ \leq ’ on $\mathcal{X}_n \subset \mathbb{R}$ such that for $x < y < z \in \mathcal{X}_n$, $(x, z) \in \mathcal{A}$ implies $(x, y) \in \mathcal{A}$ and $(z, x) \in \mathcal{A}$ implies $(z, y) \in \mathcal{A}$. Our ICD is based on two parameters, so we call it *parameterized interval catch digraph* (PICD).

2.2. Domination number of PICDs

In a digraph $D = (\mathcal{V}, \mathcal{A})$ of order $|\mathcal{V}| = n$, a vertex u *dominates* itself and all vertices of the form $\{v : (u, v) \in \mathcal{A}\}$. A *dominating set*, S_D , for the digraph D is a subset of \mathcal{V} such that each vertex $v \in \mathcal{V}$ is dominated by a vertex in S_D . A *minimum dominating set*, S_D^* , is a dominating set of minimum cardinality; and the *domination number*, denoted $\gamma(D)$, is defined as $\gamma(D) := |S_D^*|$, where $|\cdot|$ stands for set cardinality [34]. Chartrand et al. [35] distinguish domination in digraphs as out- and in-domination and provide definitions for out- and in-domination numbers for digraphs. Domination in this article refers to the *out-domination* in PICDs. If a minimum dominating set consists of only one vertex, we call

that vertex a *dominating vertex*. Clearly, the vertex set \mathcal{V} itself is always a dominating set, so we have $\gamma(D) \leq n$ in general, and $1 \leq \gamma(D) < n$ for non-trivial digraphs.

Let

$$\mathcal{F}(\mathbb{R}^d) := \{F_{X,Y} \text{ on } \mathbb{R}^d \text{ with } (X, Y) \sim F_{X,Y}, \text{ and random variables } X \text{ and } Y \text{ do not collide}\}.$$

That is, if \mathcal{X}_n and \mathcal{Y}_m are two samples from F_X and F_Y , respectively, with $(X, Y) \sim F_{X,Y}$ and the marginal distributions of X and Y are F_X and F_Y , respectively. Furthermore, ‘no collision of X and Y ’ condition is equivalent to $P(X_i = Y_j) = 0$ for all $i = 1, \dots, n$ and $j = 1, \dots, m$. Notice that if $F_{X,Y}$ continuous, then $F_{X,Y} \in \mathcal{F}(\mathbb{R}^d)$ follows. Furthermore, if the probability distributions F_X and F_Y respectively have probability measures \mathcal{M}_X and \mathcal{M}_Y which are non-atomic, then the associated joint distribution would be in $\mathcal{F}(\mathbb{R}^d)$ as well. If \mathcal{M}_Y contains an atom, Y_j points might collide, but without loss of generality we can assume that there are m distinct \mathcal{Y} points. We restrict our attention to one dimensional data (i.e. $d = 1$), so we consider the random digraph for which \mathcal{X}_n and \mathcal{Y}_m are samples from F_X and F_Y , respectively, with the joint distribution of X, Y being $F_{X,Y} \in \mathcal{F}(\mathbb{R})$. We focus on the random variable $\gamma(\mathbf{D}_{n,m}(F_{X,Y}, r, c))$, the domination number of the digraph $\mathbf{D}_{n,m}(F_{X,Y}, r, c)$. To make the notation simpler, we will use $\gamma_{n,m}(F_{X,Y}, r, c)$ instead of $\gamma(\mathbf{D}_{n,m}(F_{X,Y}, r, c))$. For $n \geq 1$ and $m \geq 1$, it is immediate to see that $1 \leq \gamma_{n,m}(F_{X,Y}, r, c) \leq n$.

Let $\mathcal{X}_{[i]} := \mathcal{X}_n \cap \mathcal{I}_i$, and $\mathcal{Y}_{[i]} := \{Y_{(i)}, Y_{(i+1)}\}$ for $i = 0, 1, 2, \dots, m$. This yields a disconnected digraph with subdigraphs each of which might be null or itself disconnected. Let $\mathbf{D}_{[i]}$ be the component of $\mathbf{D}_{n,m}(F_{X,Y}, r, c)$ induced by $\mathcal{X}_{[i]}$ for $i = 0, 1, 2, \dots, m$, $n_i := |\mathcal{X}_{[i]}|$ (provided that $n_i > 0$), and F_i be the density F_X restricted to \mathcal{I}_i (note that \mathcal{I}_i is also random here), and $\gamma_{[i]}(F_i, r, c)$ be the domination number of $\mathbf{D}_{[i]}$. Let also that $M_{c,i} \in \mathcal{I}_i$ be the internal point that divides the interval \mathcal{I}_i in ratios $c/(1-c)$ (i.e. length of the subinterval to the left of $M_{c,i}$ is $c \times 100\%$ of the length of \mathcal{I}_i). Then $\gamma_{n,m}(F_{X,Y}, r, c) = \sum_{i=0}^m \gamma_{[i]}(F_i, r, c)$.

A summary of results in this article is as follows:

- In the middle intervals (i.e. for $i = 1, 2, \dots, m-1$), we show that $\gamma_{[i]}(F_i, r, c) - 1$ has a Bernoulli distribution with the parameter depending on $F_{X,Y}$. In the end intervals (i.e. $i \in \{0, m\}$) where the domination number $\gamma_{[i]}(F_i, r, c)$ is $\mathbf{I}(n_i > 0)$.
- Conditional on \mathcal{Y}_m (i.e. \mathcal{Y}_m is given), randomness in the digraph (hence in the domination number) stem from F_X . So if \mathcal{Y}_m is given, we write the corresponding domination number as $\gamma_{n,m}(F_X, r, c)$. In this case, we modify our notations as $\mathbf{D}_{n,m}(F, r, c)$ and $\gamma_{n,m}(F, r, c)$ for the PICD and the associated domination number, where $F = F_X$.
 - (i) Then we show that $\gamma_{n,2}(F, r, c)$ is scale invariant for $\mathcal{Y}_2 = \{a, b\}$, $F = \mathcal{U}(a, b)$ with $-\infty < a < b < \infty$, where $\mathcal{U}(a, b)$ stands for uniform distribution on (a, b) , hence (without loss of generality) we can consider $\mathcal{U}(0, 1)$.
 - (ii) We find the exact (and hence the asymptotic) distribution of $\gamma_{n,2}(\mathcal{U}, r, c)$ for $r \geq 1, c \in [0, 1]$ (which is the most general case for these parameters).
 - (iii) We extend the result in (ii) by considering the general non-uniform F satisfying mild regularity conditions, thereby find the asymptotic distribution of $\gamma_{n,2}(F, r, c)$.

- (iv) Finally, we provide the more general form (in terms of n and m) of $\gamma_{n,m}(F, r, c)$ by considering general m (i.e. $m > 2$) and find the asymptotic distribution of $\gamma_{n,m}(F, r, c)$.
- Domination number is employed as a test statistic for testing uniformity of one-dimensional data, is consistent and exhibits a good performance for certain types of alternatives.

2.3. Special cases for the distribution of $\gamma_{n,m}(F_{X,Y}, r, c)$

We study the simpler random variable $\gamma_{[i]}(F_i, r, c)$ first. The following lemma follows trivially.

Lemma 2.1: For $i \in \{0, m\}$, we have $\gamma_{[i]}(F_i, r, c) = \mathbf{I}(n_i > 0)$ for all $r \geq 1$. For $i = 1, 2, 3, \dots, (m-1)$, if $n_i = 1$, then $\gamma_{[i]}(F_i, r, c) = 1$.

Let $\Gamma_1(B, r, c)$ be the Γ_1 -region for set B associated with the proximity map $N(\cdot, r, c)$.

Lemma 2.2: The Γ_1 -region for $\mathcal{X}_{[i]}$ in \mathcal{I}_i with $r \geq 1$ and $c \in [0, 1]$ is

$$\Gamma_1(\mathcal{X}_{[i]}, r, c) = \left(\frac{\max(\mathcal{X}_{[i]}) + Y_{(i)}(r-1)}{r}, M_{c,i} \right] \cup \left[M_{c,i}, \frac{\min(\mathcal{X}_{[i]}) + Y_{(i+1)}(r-1)}{r} \right)$$

with the understanding that the intervals (a, b) , $(a, b]$, and $[a, b)$ are empty if $a \geq b$.

Notice that if $\mathcal{X}_{[i]} \cap \Gamma_1(\mathcal{X}_{[i]}, r, c) \neq \emptyset$, we have $\gamma_{[i]}(F_i, r, c) = 1$, hence the name Γ_1 -region and the notation $\Gamma_1(\cdot)$. For $i = 1, 2, 3, \dots, (m-1)$ and $n_i > 1$, we prove that $\gamma_{[i]}(F_i, r, c) = 1$ or 2 with distribution dependent probabilities. Hence, to find the distribution of $\gamma_{[i]}(F_i, r, c)$, it suffices to find the probability of $\gamma_{[i]}(F_i, r, c)$ is 1 or 2. For computational convenience, we employ the latter in our calculations henceforth and denote it as $p(F_i, r, c) := P(\gamma_{[i]}(F_i, r, c) = 2) = P(\mathcal{X}_{[i]} \cap \Gamma_1(\mathcal{X}_{[i]}, r, c) = \emptyset)$.

Furthermore, let $\text{BER}(p)$ and $\text{BIN}(n', p)$, respectively, denote the Bernoulli and Binomial distributions where p is the probability of success with $p \in [0, 1]$ and $n' > 0$ is the number of trials.

Lemma 2.3: For $i = 1, 2, 3, \dots, (m-1)$, let the support of F_i have positive Lebesgue measure. Then for $n_i > 1$, $r \in (1, \infty)$, and $c \in (0, 1)$, we have $\gamma_{[i]}(F_i, r, c) - 1 \sim \text{BER}(p(F_i, r, c))$. Furthermore, $\gamma_{1,2}(F_i, r, c) = 1$ for all $r \geq 1$ and $c \in [0, 1]$; $\gamma_{[i]}(F_i, r, 0) = \gamma_{[i]}(F_i, r, 1) = 1$ for all $n_i \geq 1$ and $r \geq 1$; and $\gamma_{[i]}(F_i, \infty, c) = 1$ for all $n_i \geq 1$ and $c \in [0, 1]$.

The probability $p(F_i, r, c)$ depends on the distribution $F_{X,Y}$ and the interval $\Gamma_1(\mathcal{X}_{[i]}, r, c)$, which, if known, will make the computation of $p(F_i, r, c)$ possible. We can bound the domination number with some crude bounds in this general case (see the Supplementary File).

Based on Proposition S2.2, we have $P(\gamma_{[i]}(F_i, 1, c) = 1) = P(\mathcal{X}_{[i]} \subset (Y_{(i)}, M_{c,i})) + P(\mathcal{X}_{[i]} \subset (M_{c,i}, Y_{(i+1)}))$ and $P(\gamma_{[i]}(F_i, 1, c) = 2) = P(\mathcal{X}_{[i]} \cap (Y_{(i)}, M_{c,i}) \neq \emptyset, \mathcal{X}_{[i]} \cap (M_{c,i}, Y_{(i+1)}) \neq \emptyset)$.

Remark 2.4: Restrictions on the Joint and Marginal Distributions for the Rest of the Article: The only restriction we imposed on $F_{X,Y}$ thus far was that $P(X = Y) = 0$ and collisions

were not allowed (i.e. $P(X_i = Y_j) = 0$ for all $i = 1, \dots, n$ and $j = 1, \dots, m$). Note that \mathcal{X}_n and \mathcal{Y}_m need not be independent of each other; collisions would be avoided if X has a continuous distribution. But in general X and Y can both be continuous, discrete or mixed. Although we define in this very general setting, *in the rest of the article we will condition on a realization of \mathcal{Y}_m* . Henceforth for brevity in notation, we write $F = F_X$ and $\mathcal{M} = \mathcal{M}_X$ and we also assume that \mathcal{X}_n is a random sample from F (i.e. $X_j \stackrel{\text{iid}}{\sim} F$ for $j = 1, \dots, n$). For $X_j \stackrel{\text{iid}}{\sim} F$, with the additional assumption that support $\mathcal{S}(F_i) \subseteq \mathcal{I}_i$ and F is absolutely continuous around $M_{c,i}$ and around the end points of \mathcal{I}_i , it follows that the special cases in the construction of $N(\cdot, r, c) - X$ falls at $M_{c,i}$ or the end points of \mathcal{I}_i – occurs with probability zero. Notice that X_j having a non-degenerate one-dimensional probability density function (pdf) f which is continuous around $M_{c,i}$ and around the end points of \mathcal{I}_i is a special case of this (additional) assumption. Furthermore, for such an F , the region $N(X_i, r, c)$ is an interval a.s.

The results so far have been straightforward so far. The more interesting cases are presented in the subsequent sections.

3. The distribution of the domination number of PICDs for uniform data in one interval

We first consider the simplest case of $m = 2$ with $\mathcal{Y}_2 = \{y_1, y_2\}$ with $-\infty < y_1 < y_2 < \infty$ and $\mathcal{X}_n = \{X_1, X_2, \dots, X_n\}$ a random sample from $\mathcal{U}(y_1, y_2)$, we have the PICD with vertices from $F = \mathcal{U}(y_1, y_2)$. The special case of $m = 2$ is important in deriving the distribution of the domination number in the general case of $m > 2$, because the domination number in multiple interval case is the sum of the domination numbers for the intervals. We denote such digraphs as $\mathbf{D}_{n,2}(\mathcal{U}(y_1, y_2), r, c)$ and provide the exact distribution of their domination number for the entire range of r and c . Let $\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c)$ be the domination number of the PICD based on $N(\cdot, r, c)$ and \mathcal{X}_n and $p_n(\mathcal{U}(y_1, y_2), r, c) := P(\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c) = 2)$, and $p(\mathcal{U}(y_1, y_2), r, c) := \lim_{n \rightarrow \infty} p_n(\mathcal{U}(y_1, y_2), r, c)$. We first present a ‘scale invariance’ result for $\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c)$.

Theorem 3.1 (Scale Invariance Property): *Suppose \mathcal{X}_n is a random sample from $\mathcal{U}(y_1, y_2)$ with $-\infty < y_1 < y_2 < \infty$. Then for any $r \in [1, \infty]$ the distribution of $\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c)$ is independent of \mathcal{Y}_2 and hence independent of the support interval (y_1, y_2) .*

Proof: Let \mathcal{X}_n be a random sample from $\mathcal{U}(y_1, y_2)$ distribution. Any $\mathcal{U}(y_1, y_2)$ random variable can be transformed into a $\mathcal{U}(0, 1)$ random variable by the transformation $\phi(x) = (x - y_1)/(y_2 - y_1)$, which maps intervals $(t_1, t_2) \subseteq (y_1, y_2)$ to intervals $(\phi(t_1), \phi(t_2)) \subseteq (0, 1)$. That is, if $X \sim \mathcal{U}(y_1, y_2)$, then we have $\phi(X) \sim \mathcal{U}(0, 1)$ and $P_1(X \in (t_1, t_2)) = P_2(\phi(X) \in (\phi(t_1), \phi(t_2)))$ for all $(t_1, t_2) \subseteq (y_1, y_2)$ where P_1 is the probability measure with respect to $\mathcal{U}(y_1, y_2)$ and P_2 is with respect to $\mathcal{U}(0, 1)$. So, the distribution of $\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c)$ does not depend on the support interval (y_1, y_2) , i.e. it is scale invariant. ■

Note that scale invariance of $\gamma_{n,2}(F, \infty, c)$ follows trivially for all \mathcal{X}_n from any F with support in (y_1, y_2) , since for $r = \infty$, we have $\gamma_{n,2}(F, \infty, c) = 1$ a.s. for all $n > 1$ and $c \in (0, 1)$. The scale invariance of $\gamma_{1,2}(F, r, c)$ holds for all $r \geq 1$ and $c \in [0, 1]$, and scale invariance of

$\gamma_{n,2}(F, r, c)$ with $c \in \{0, 1\}$ holds for all $n \geq 1$ and $r \geq 1$ as well. The scale invariance property in Theorem 3.1 will *simplify the notation and calculations in our subsequent analysis of $\gamma_{n,2}(\mathcal{U}(y_1, y_2), r, c)$ by allowing us to consider the special case of the unit interval, $(0, 1)$.* Hence we drop the interval end points y_1 and y_2 in our notation and write $\gamma_{n,2}(\mathcal{U}, r, c)$ and $p_u(r, c, n)$, and $p_u(r, c)$ for $p_n(\mathcal{U}, r, c)$ and $p(\mathcal{U}, r, c)$ henceforth when vertices are from uniform distribution. Then the proximity region for $x \in (0, 1)$ with parameters $r \geq 1$ and $c \in [0, 1]$ simplifies to

$$N(x, r, c) = \begin{cases} (0, \min(1, rx)) & \text{if } x \in (0, c), \\ (\max(0, 1 - r(1 - x)), 1) & \text{if } x \in (c, 1) \end{cases} \quad (3)$$

with the comments below Equation (2) applying to $N(x, r, c)$ as well. ■

Remark 3.2: Given $X_{(1)} = x_1$ and $X_{(n)} = x_n$, let $\Gamma_1(\mathcal{X}_n, r, c) = (\delta_1, \delta_2)$. Then the probability of $\gamma_{n,2}(F, r, c) = 2$ (i.e. the quantity $p_n(F, r, c)$) is $(1 - [F(\delta_2) - F(\delta_1)]/[F(x_n) - F(x_1)])^{(n-2)}$ provided that $\delta_1 < \delta_2$ (i.e. $\Gamma_1(\mathcal{X}_n, r, c) \neq \emptyset$); if $\Gamma_1(\mathcal{X}_n, r, c) = \emptyset$, then we would have $\gamma_{n,2}(F, r, c) = 2$. That is, $P(\gamma_{n,2}(F, r, c) = 2) = P(\gamma_{n,2}(F, r, c) = 2, \Gamma_1(\mathcal{X}_n, r, c) \neq \emptyset) + P(\gamma_{n,2}(F, r, c) = 2, \Gamma_1(\mathcal{X}_n, r, c) = \emptyset)$. Then

$$P(\gamma_{n,2}(F, r, c) = 2, \Gamma_1(\mathcal{X}_n, r, c) \neq \emptyset) = \int \int_{S_1} f_{1n}(x_1, x_n) \left(1 - \frac{F(\delta_2) - F(\delta_1)}{F(x_n) - F(x_1)}\right)^{(n-2)} dx_n dx_1, \quad (4)$$

where $S_1 = \{0 < x_1 < x_n < 1 : (x_1, x_n) \notin \Gamma_1(\mathcal{X}_n, r, c), \text{ and } \Gamma_1(\mathcal{X}_n, r, c) \neq \emptyset\}$ and $f_{1n}(x_1, x_n) = n(n - 1)f(x_1)f(x_n)(F(x_n) - F(x_1))^{(n-2)}\mathbf{I}(0 < x_1 < x_n < 1)$ is the joint pdf of $X_{(1)}, X_{(n)}$. The integral in (4) becomes

$$P(\gamma_{n,2}(F, r, c) = 2, \Gamma_1(\mathcal{X}_n, r, c) \neq \emptyset) = \int \int_{S_1} H(x_1, x_n) dx_n dx_1, \quad (5)$$

where

$$H(x_1, x_n) := n(n - 1)f(x_1)f(x_n)(F(x_n) - F(x_1) + F(\delta_1) - F(\delta_2))^{n-2}. \quad (6)$$

If $\Gamma_1(\mathcal{X}_n, r, c) = \emptyset$, then $\gamma_{n,2}(F, r, c) = 2$. So

$$P(\gamma_{n,2}(F, r, c) = 2, \Gamma_1(\mathcal{X}_n, r, c) = \emptyset) = P(\Gamma_1(\mathcal{X}_n, r, c) = \emptyset) = \int \int_{S_2} f_{1n}(x_1, x_n) dx_n dx_1 \quad (7)$$

where $S_2 = \{0 < x_1 < x_n < 1 : \Gamma_1(\mathcal{X}_n, r, c) = \emptyset\}$.

3.1. Exact distribution of $\gamma_{n,2}(\mathcal{U}, r, c)$

We first consider the case of $\mathcal{U}(y_1, y_2)$ data with $r \geq 1$ and $c \in [0, 1]$ and $n = 1, 2, \dots$. That is, we derive the distribution of $\gamma_{n,2}(\mathcal{U}, r, c)$ for the entire range of the parameters r and c . For $r \geq 1$ and $c \in (0, 1)$, the Γ_1 -region is $\Gamma_1(\mathcal{X}_n, r, c) = (X_{(n)}/r, c] \cup [c, (X_{(1)} + r - 1)/r)$ where $(X_{(n)}/r, c]$ or $[c, (X_{(1)} + r - 1)/r)$ or both could be empty.

Theorem 3.3 (Main Result 1): Let \mathcal{X}_n be a random sample from $\mathcal{U}(y_1, y_2)$ distribution with $n \geq 1$, $r \geq 1$, and $c \in (0, 1)$. Then we have

$$\gamma_{n,2}(\mathcal{U}, r, c) - 1 \sim \text{BER}(p_u(r, c, n))$$

with

$$p_u(r, c, n) = \begin{cases} p_{u,a}(r, c, n) & \text{for } c \in [(3 - \sqrt{5})/2, 1/2], \\ p_{u,b}(r, c, n) & \text{for } c \in [1/4, (3 - \sqrt{5})/2], \\ p_{u,c}(r, c, n) & \text{for } c \in (0, 1/4), \end{cases}$$

where explicit forms of $p_{u,a}(r, c, n)$, $p_{u,b}(r, c, n)$, and $p_{u,c}(r, c, n)$ are provided in Section S3.1 of the Supplementary File. By symmetry, for $c \in (1/2, (\sqrt{5} - 1)/2]$, we have $p_u(r, c, n) = p_{u,a}(r, 1 - c, n)$, for $c \in ((\sqrt{5} - 1)/2, 3/4]$, $p_u(r, c, n) = p_{u,b}(r, 1 - c, n)$, and for $c \in (3/4, 1)$, $p_u(r, c, n) = p_{u,c}(r, 1 - c, n)$ with the understanding that the transformation $c \rightarrow 1 - c$ is also applied in the interval endpoints in the piecewise definitions of $p_{u,a}(r, c, n)$, $p_{u,b}(r, c, n)$ and $p_{u,c}(r, c, n)$, respectively.

Furthermore, we have $\gamma_{n,2}(\mathcal{U}, r, 0) = \gamma_{n,2}(\mathcal{U}, r, 1) = 1$ for all $n \geq 1$.

Some remarks are in order for Main Result 1. The partitioning of $c \in (0, 1/2)$ as $c \in (0, 1/4)$, $c \in [1/4, (3 - \sqrt{5})/2)$, and $c \in [(3 - \sqrt{5})/2, 1/2)$ is due to the relative positions of $1/(1 - c)$ and $(1 - c)/c$ and the restrictions arising from various cases in the probability computations (see the Supplementary File). For example, for $c \in ((3 - \sqrt{5})/2, 1/2)$, we have $1/(1 - c) > (1 - c)/c$ and for $c \in (0, (3 - \sqrt{5})/2)$, we have $1/(1 - c) < (1 - c)/c$.

We present the (three-dimensional) surface plots of $p_u(r, c, n)$ for $n = 10$ and $n = 100$ in Figure 2. As expected $\lim_{r \rightarrow 1} p_u(r, c, n) = 0$. For finite $n \geq 1$, the probability $p_u(r, c, n)$ is continuous in $(r, c) \in \{(r, c) \in \mathbb{R}^2 : r \geq 1, 0 \leq c \leq 1\}$. For fixed $c \in (0, 1)$ and fixed n , $p_u(r, c, n)$ is decreasing as r is increasing, while for fixed $r \in (1, \infty)$ and fixed n , $p_u(r, c, n)$ is increasing as c is approaching to $1/2$. In particular, as $(r, c) \rightarrow (2, 1/2)$ the distribution of $\gamma_{n,2}(\mathcal{U}, r, c) - 1$ converges to $\text{BER}(p_u(2, 1/2, n))$, where $p_u(2, 1/2, n) = 4/9 - (16/9)4^{-n}$ as in [10]. In the special cases of $c = 1/2$ or $r = 2$ or $(r, c) = (2, 1/2)$, the probability $p_u(r, c, n)$ reduces to much simpler forms. See Section S3.3 in the Supplementary Materials file.

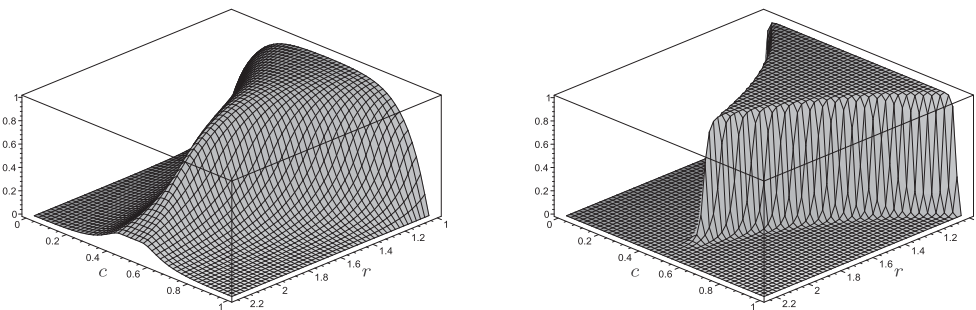


Figure 2. Surface plots of $p_u(r, c, n)$ with $n = 10$ (left) and $n = 100$ (right).

3.1.1. Asymptotic distribution of $\gamma_{n,2}(\mathcal{U}, r, c)$

Theorem 3.4 (Main Result 2): For the PICD, $\mathbf{D}_{n,2}(\mathcal{U}, r, c)$, with $c \in (0, 1)$ and $r^* = 1/\max(c, 1 - c)$, the domination number $\gamma_{n,2}(\mathcal{U}, r, c)$ has the following asymptotic distribution. As $n \rightarrow \infty$, for $c \in (0, 1)$,

$$\gamma_{n,2}(\mathcal{U}, r, c) - 1 \xrightarrow{\mathcal{L}} \begin{cases} 0, & \text{for } r > r^*, \\ \text{BER}(p_r), & \text{for } r = r^*, \\ 1, & \text{for } 1 \leq r < r^*. \end{cases} \quad (8)$$

where

$$p_r = \begin{cases} \frac{r^*}{r^* + 1}, & \text{for } c \neq 1/2, \\ \frac{4}{9}, & \text{for } c = 1/2. \end{cases} \quad (9)$$

Notice the interesting behaviour of the asymptotic distribution of $\gamma_{n,2}(\mathcal{U}, r, c)$ around $r = r^*$ for any given $c \in (0, 1)$. The asymptotic distribution is non-degenerate only for $r = r^*$. For $r > r^*$, $\lim_{n \rightarrow \infty} \gamma_{n,2}(\mathcal{U}, r, c) = 1$ w.p. 1, and for $1 \leq r < r^*$, $\lim_{n \rightarrow \infty} \gamma_{n,2}(\mathcal{U}, r, 1/2) = 2$ w.p. 1. The critical value $r = r^*$ corresponds to $c = (r - 1)/r$, if $c \in (0, 1/2)$ (i.e. $r^* = 1/(1 - c)$) and $c = 1/r$, if $c \in (1/2, 1)$ (i.e. $r^* = 1/c$) and $r = r^*$ only possible for $r \in (1, 2)$. The probability $p_u(r, c)$ is continuous in r and c for $r \neq r^*$ and there is a jump (hence discontinuity) in the probability $p_u(r, c)$ at $r = r^*$, since $p_u(r^*, c) = r^*/(r^* + 1)$ for $c \neq 1/2$ (see also Figure 3). Therefore, given a centrality parameter $c \in (0, 1)$, we can choose the expansion parameter r for which the asymptotic distribution is non-degenerate, and vice versa. There is yet another interesting behaviour of the asymptotic distribution around $(r, c) = (2, 1/2)$. The probability $p_u(r^*, c)$ has jumps at $(r, c) = (r^*, c)$ for $r \in [1, 2]$ with $p_u(r^*, c) = r^*/(r^* + 1)$ for $c \neq 1/2$. That is, for fixed $(r, c) \in \mathcal{S}$, $\lim_{n \rightarrow \infty} p_u(r^*, c, n) = r^*/(r^* + 1)$ for $c \neq 1/2$. Letting $(r, c) \rightarrow (2, 1/2)$, we get $p_u(r^*, c) \rightarrow 2/3$, but $p_u(2, 1/2) = 4/9$. Hence for $(r, c) \neq (2, 1/2)$ the distribution of $\gamma_{n,2}(\mathcal{U}, r^*, c) - 1$ converges to $\text{BER}(r^*/(r^* + 1))$, but the distribution of $\gamma_{n,2}(\mathcal{U}, 2, 1/2) - 1$ converges to $\text{BER}(4/9)$ as $n \rightarrow \infty$ (rather than $\text{BER}(2/3)$). In other words, $p_u(r^*, c)$ has another jump at $(r, c) = (2, 1/2)$. This interesting behaviour occurs due to the symmetry around $c = 1/2$. Because for $c \in (0, 1/2)$, with $r = 1/(1 - c)$, for sufficiently large n , a point X_i in $(c, 1)$ can dominate all the points in \mathcal{X}_n (implying $\gamma_{n,2}(\mathcal{U}, 1/(1 - c), c) = 1$), but no point in $(0, c)$ can dominate all points a.s. Likewise, for $c \in (1/2, 1)$ with $r = 1/c$, for sufficiently large n , a point X_i in $(0, c)$ can dominate all the points in \mathcal{X}_n (implying $\gamma_{n,2}(\mathcal{U}, 1/c, c) = 1$), but no point in $(c, 1)$ can dominate all points a.s. However, for $c = 1/2$ and $r = 2$, for sufficiently large n , points to the left or right of c can dominate all other points in \mathcal{X}_n .

4. Distribution of $\gamma_{n,2}(F, r, c)$

We now relax the assumption of uniformity for the vertices of our PICD (i.e. for \mathcal{X} points). Let $\mathcal{F}(y_1, y_2)$ be a family of continuous distributions with support in $\mathcal{S}_F \subseteq (y_1, y_2)$. Consider a distribution function $F \in \mathcal{F}(y_1, y_2)$. For simplicity, assume $y_1 = 0$ and $y_2 = 1$. Let \mathcal{X}_n be a random sample from F , Γ_1 -region $\Gamma_1(\mathcal{X}_n, r, c) = (\delta_1, \delta_2)$, and $p_n(F, r, c) :=$

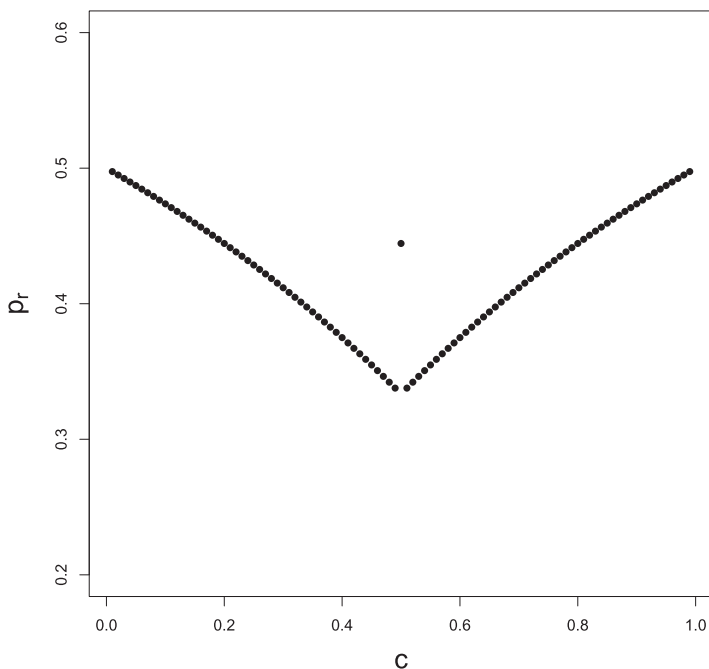


Figure 3. Plot of the limiting probability $p_r := \lim_{n \rightarrow \infty} \gamma_{n,2}(\mathcal{U}, r, c)$ for $r = r^* = 1/\max(c, 1 - c)$ (see also Equation (9)).

$P(\gamma_{n,2}(F, r, c) = 2)$, $p(F, r, c) := \lim_{n \rightarrow \infty} P(\gamma_{n,2}(F, r, c) = 2)$. The exact and asymptotic distributions of $\gamma_{n,2}(F, r, c) - 1$ are $\text{BER}(p_n(F, r, c))$ and $\text{BER}(p(F, r, c))$, respectively. That is, for finite $n > 1$, $r \in [1, \infty)$, and $c \in (0, 1)$, we have

$$\gamma_{n,2}(F, r, c) = \begin{cases} 1 & \text{w.p. } 1 - p_n(F, r, c), \\ 2 & \text{w.p. } p_n(F, r, c). \end{cases} \quad (10)$$

Moreover, $\gamma_{1,2}(F, r, c) = 1$ for all $r \geq 1$ and $c \in [0, 1]$, $\gamma_{n,2}(F, r, 0) = \gamma_{n,2}(F, r, 1) = 1$ for all $n \geq 1$ and $r \geq 1$, $\gamma_{n,2}(F, \infty, c) = 1$ for all $n \geq 1$ and $c \in [0, 1]$, and $\gamma_{n,2}(F, 1, c) = k_4$ for all $n \geq 1$ and $c \in (0, 1)$ where k_4 is as in Proposition S2.2 with $m = 2$. The asymptotic distribution is similar with $p_n(F, r, c)$ being replaced with $p(F, r, c)$. The special cases are similar in the asymptotics with the exception that $p(F, 1, c) = 1$ for all $c \in (0, 1)$. The finite sample mean and variance of $\gamma_{n,2}(F, r, c) - 1$ are $p_n(F, r, c)$ and $p_n(F, r, c)(1 - p_n(F, r, c))$, respectively; and similarly the asymptotic mean and variance of $\gamma_{n,2}(F, r, c) - 1$ are $p(F, r, c)$ and $p(F, r, c)(1 - p(F, r, c))$, respectively.

For $\mathcal{Y}_2 = \{y_1, y_2\} \subset \mathbb{R}$ with $-\infty < y_1 < y_2 < \infty$, a quick investigation shows that, by Lemma 2.2, the Γ_1 -region is $\Gamma_1(\mathcal{X}_n, r, c) = (\frac{X_{(n)} + y_1(r-1)}{r}, M_c] \cup [M_c, \frac{X_{(1)} + y_2(r-1)}{r})$. Notice that for a given $c \in [0, 1]$, the corresponding $M_c \in [y_1, y_2]$ is $M_c = y_1 + c(y_2 - y_1)$. Let F be a continuous distribution with support $\mathcal{S}(F) \subseteq (0, 1)$. The simplest of such distributions is $\mathcal{U}(0, 1)$, which yields the simplest exact distribution for $\gamma_{n,2}(F, r, c)$ with $(r, c) = (2, 1/2)$. If $X \sim F$, then by probability integral transform, $F(X) \sim \mathcal{U}(0, 1)$. So for any continuous F , we can construct a proximity map depending on F for which the distribution of the domination number of the associated digraph has the same distribution as that of $\gamma_{n,2}(\mathcal{U}, r, c)$,

which is explicated in the below proposition whose proof is provided in the Supplementary File.

Proposition 4.1: Let $X_i \stackrel{\text{iid}}{\sim} F$ which is an absolutely continuous distribution with support $\mathcal{S}(F) = (0, 1)$ and let $\mathcal{X}_n := \{X_1, X_2, \dots, X_n\}$. Define the proximity map $N_F(x, r, c) := F^{-1}(N(F(x), r, c))$. That is,

$$N_F(x, r, c) = \begin{cases} (0, \min(1, F^{-1}(rF(x)))) & \text{if } x \in (0, F^{-1}(c)), \\ (\max(0, F^{-1}(1 - r(1 - F(x))), 1) & \text{if } x \in (F^{-1}(c), 1). \end{cases} \quad (11)$$

Then the domination number of the digraph based on N_F , \mathcal{X}_n , and $\mathcal{Y}_2 = \{0, 1\}$ has the same distribution as $\gamma_{n,2}(\mathcal{U}, r, c)$.

The result in Proposition 4.1 can easily be generalized for a distribution F with $\mathcal{S}(F) = (a, b)$ with finite $a < b$. For $X \sim F$, the transformed random variable $W = \frac{X-a}{b-a}$ would have cdf $F_W(w) = F_X(a + w(b - a))$ which has support $\mathcal{S}(F_W) = (0, 1)$. Then one can apply Proposition 4.1 to $W_i \stackrel{\text{iid}}{\sim} F_W$. There is also a stochastic ordering between $\gamma_{n,2}(F, r, c)$ and $\gamma_{n,2}(\mathcal{U}, r, c)$ provided that F satisfies some regularity conditions, which are provided in Proposition S4.1 in the Supplementary File. We can also find the exact distribution of $\gamma_{n,2}(F, r, c)$ for F whose pdf is piecewise constant with support in $(0, 1)$, see Remark S4.2 in the Supplementary File for more details.

Recall the PICD, $\mathbf{D}_{n,m}(F, r, c)$. We denote the digraph which is obtained in the special case of $\mathcal{Y}_2 = \{y_1, y_2\}$ and support of F_X in (y_1, y_2) as $\mathbf{D}_{n,2}(F, r, c)$. Below, we provide asymptotic results pertaining to the distribution of domination number of such digraphs.

4.1. Asymptotic distribution of $\gamma_{n,2}(F, r, c)$

Although the exact distribution of $\gamma_{n,2}(F, r, c)$ may not be analytically available in a simple closed form for F whose density is not piecewise constant, the asymptotic distribution of $\gamma_{n,2}(F, r, c)$ is available for larger families of distributions. First, we present the asymptotic distribution of $\gamma_{n,2}(F, r, c)$ for $\mathbf{D}_{n,2}(F, r, c)$ with $\mathcal{Y}_2 = \{y_1, y_2\} \subset \mathbb{R}$ with $-\infty < y_1 < y_2 < \infty$ for general F with support $\mathcal{S}(F) \subseteq (y_1, y_2)$. Then we will extend this to the case with $\mathcal{Y}_m \subset \mathbb{R}$ with $m > 2$.

Let $c \in (0, 1/2]$ and $r \in (1, 2]$. Then for $(r, c) = (1/(1 - c), c)$, we define the family of distributions

$$\begin{aligned} \mathcal{F}_1(y_1, y_2) &:= \{F : (y_1, y_1 + \varepsilon) \cup (M_c, M_c + \varepsilon) \subseteq \mathcal{S}(F) \\ &\subseteq (y_1, y_2) \text{ for some } \varepsilon \in (0, c) \text{ with } c = (0, 1/2]\}. \end{aligned}$$

Similarly, let $c \in [1/2, 1)$ and $r \in (1, 2]$. Then for $(r, c) = (1/c, c)$, we define

$$\begin{aligned} \mathcal{F}_2(y_1, y_2) &:= \{F : (y_2 - \varepsilon, y_2) \cup (M_c - \varepsilon, M_c) \subseteq \mathcal{S}(F) \\ &\subseteq (y_1, y_2) \text{ for some } \varepsilon \in (0, 1 - c) \text{ with } c = [1/2, 1)\}. \end{aligned}$$

Let k th order right (directed) derivative at x be defined as $f^{(k)}(x^+) := \lim_{h \rightarrow 0^+} \frac{f^{(k-1)}(x+h) - f^{(k-1)}(x)}{h}$ for all $k \geq 1$ and the right limit at u be defined as $f(u^+) :=$

$\lim_{h \rightarrow 0^+} f(u+h)$. Let the left derivatives and limits be defined similarly with $+$'s being replaced by $-$'s.

Theorem 4.2 (Main Result 3): Suppose $\mathcal{Y}_2 = \{y_1, y_2\} \subset \mathbb{R}$ with $-\infty < y_1 < y_2 < \infty$, $\mathcal{X}_n = \{X_1, X_2, \dots, X_n\}$ with $X_i \stackrel{iid}{\sim} F$ with $\mathcal{S}(F) \subseteq (y_1, y_2)$, and $c \in (0, 1)$ and $r^* = 1/\max(c, 1-c)$. Let $\mathbf{D}_{n,2}(F, r, c)$ be the PICD based on \mathcal{X}_n and \mathcal{Y}_2 .

- (i) Then for $n > 1$, $r \in (1, \infty)$, we have $\gamma_{n,2}(F, r^*, c) - 1 \sim \text{BER}(p_n(F, r^*, c))$. Note also that $\gamma_{1,2}(F, r, c) = 1$ for all $r \geq 1$ and $c \in [0, 1]$; for $r = 1$, we have $\gamma_{n,2}(F, 1, 0) = \gamma_{n,2}(F, 1, 1) = 1$ for all $n \geq 1$ and for $r = \infty$, we have $\gamma_{n,2}(F, \infty, c) = 1$ for all $n \geq 1$ and $c \in [0, 1]$.
- (ii) Suppose $c \in (0, 1/2)$ and $r = r^* = 1/(1-c)$, $F \in \mathcal{F}_1(y_1, y_2)$ with pdf f , and $k \geq 0$ is the smallest integer for which $F(\cdot)$ has continuous right derivatives up to order $(k+1)$ at y_1 , M_c , and $f^{(k)}(y_1^+) + r^{-(k+1)}f^{(k)}(M_c^+) \neq 0$ and $f^{(i)}(y_1^+) = f^{(i)}(M_c^+) = 0$ for all $i = 0, 1, 2, \dots, (k-1)$ and suppose also that $F(\cdot)$ has a continuous left derivative at y_2 . Then for bounded $f^{(k)}(\cdot)$, we have the following limit

$$p(F, 1/(1-c), c) = \lim_{n \rightarrow \infty} p_n(F, 1/(1-c), c) = \frac{f^{(k)}(y_1^+)}{f^{(k)}(y_1^+) + (1-c)^{(k+1)}f^{(k)}(M_c^+)}.$$

- (iii) Suppose $c \in (1/2, 1)$ and $r = r^* = 1/c$, $F \in \mathcal{F}_2(y_1, y_2)$ with pdf f , and $\ell \geq 0$ is the smallest integer for which $F(\cdot)$ has continuous left derivatives up to order $(\ell+1)$ at y_2 , and M_c , and $f^{(\ell)}(y_2^-) + r^{-(\ell+1)}f^{(\ell)}(M_c^-) \neq 0$ and $f^{(i)}(y_2^-) = f^{(i)}(M_c^-) = 0$ for all $i = 0, 1, 2, \dots, (\ell-1)$ and suppose also that $F(\cdot)$ has a continuous right derivative at y_1 . Then for bounded $f^{(\ell)}(\cdot)$, we have the following limit

$$p(F, 1/c, c) = \lim_{n \rightarrow \infty} p_n(F, 1/c, c) = \frac{f^{(\ell)}(y_2^-)}{f^{(\ell)}(y_2^-) + c^{(\ell+1)}f^{(\ell)}(M_c^-)}.$$

- (iv) Suppose $(M_c - \varepsilon, M_c + \varepsilon) \cup (y_1, y_1 + \varepsilon) \cup (y_2 - \varepsilon, y_2) \subset \mathcal{S}(F)$ for some $\varepsilon > 0$, then

$$p(F, r, c) = \begin{cases} 1 & \text{if } r > r^*, \\ 0 & \text{if } r < r^*. \end{cases}$$

The asymptotic distribution of $\gamma_{n,2}(F, r, c)$ for $r = 2$ and $c = 1/2$ is provided in Theorem S4.3 in the Supplementary File.

In Theorem 4.2 parts (ii) and (iii), we assume that $f^{(k)}(\cdot)$ and $f^{(\ell)}(\cdot)$ are bounded on (y_1, y_2) , respectively. The extension to the unbounded derivatives is provided in Remark S4.4 in the Supplementary File. The rates of convergence in Theorem 4.2 parts (ii) and (iii) depend on f and are provided in Remark S4.5 in the Supplementary File. The conditions of the Theorems 4.2 and S4.3 might seem a bit esoteric. However, most of the well known functions that are scaled and properly transformed to be pdf of some random variable with support in (y_1, y_2) satisfy the conditions for some k or ℓ , hence one can compute the corresponding limiting probability $p(F, r^*, c)$.

Examples: (a) With $F = \mathcal{U}(y_1, y_2)$, in Theorem 4.2 (ii), we have $k = 0$ and $f(y_1^+) = f(M_c^+) = 1/(y_2 - y_1)$, and in Theorem 4.2 (iii), we have $\ell = 0$ and $f(y_2^-) = f(M_c^-) =$

$1/(y_2 - y_1)$. Then $\lim_{n \rightarrow \infty} p_n(\mathcal{U}, r^*, c) = r^*/(r^* + 1)$ for $c \neq 1/2$, which agrees with the result given in Equation (8) and $\lim_{n \rightarrow \infty} p_u(2, 1/2, n) = 4/9$.

(b) For F with pdf $f(x) = (x + 1/2)\mathbf{I}(0 < x < 1)$, we have $k = 0$, $f(0^+) = 1/2$, and $f(c^+) = c + 1/2$ in Theorem 4.2 (ii). Then $p(F, 1/(1 - c), c) = \frac{1}{2+c-2c^2}$ for $c \neq 1/2$. In Theorem 4.2 (iii), we have $\ell = 0$, $f(1^-) = 3/2$ and $f(c^-) = c + 1/2$, then $p(F, 1/c, c) = \frac{3}{3+c+2c^2}$ for $c \neq 1/2$. Based on Theorem S4.3, $p(F, 2, 1/2) = 3/8$.

(c) For F with pdf $f(x) = (\pi/2)|\sin(2\pi x)|\mathbf{I}(0 < x < 1) = (\pi/2)(\sin(2\pi x)\mathbf{I}(0 < x \leq 1/2) - \sin(2\pi x)\mathbf{I}(1/2 < x < 1))$, we have $k = 0$, $f(0^+) = 0$, and $f(c^+) = (\pi/2)(\sin(2\pi c))$ in Theorem 4.2 (ii). Then $p(F, 1/(1 - c), c) = 0$ for $c \neq 1/2$. As for Theorem 4.2 (iii), we have $\ell = 0$, $f(1^-) = 0$ and $f(c^-) = -(\pi/2)(\sin(2\pi c))$. Then $p(F, 1/c, c) = 0$ for $c \neq 1/2$. Moreover, by Theorem S4.3, $p(F, 2, 1/2) = 0$ as well.

For more examples, see Supplementary File. In Theorem 4.2 (ii), if we have $f^{(k)}(0^+) = f^{(k)}(c^+)$, then $\lim_{n \rightarrow \infty} p_n(F, 1/(1 - c), c) = \frac{1}{1+(1-c)^{(k+1)}}$. In particular, if $k = 0$, then $\lim_{n \rightarrow \infty} p_n(F, 1/(1 - c), c) = 1/(2 - c)$. Hence $\gamma_{n,2}(F, 1/(1 - c), c)$ and $\gamma_{n,2}(\mathcal{U}, 1/(1 - c), c)$ would have the same limiting distribution. Likewise, in Theorem 4.2 (iii), if we have $f^{(\ell)}(1^-) = f^{(\ell)}(c^-)$, then $\lim_{n \rightarrow \infty} p_n(F, 1/c, c) = \frac{1}{1+c^{(\ell+1)}}$. In particular, if $\ell = 0$, then $\lim_{n \rightarrow \infty} p_n(F, 1/c, c) = 1/(1 + c)$. Hence $\gamma_{n,2}(F, 1/c, c)$ and $\gamma_{n,2}(\mathcal{U}, 1/c, c)$ would have the same limiting distribution.

5. Distribution of $\gamma_{n,m}(F_{X,Y}, r, c)$

We now consider the more challenging case of $m > 2$. For $\omega_1 < \omega_2$ in \mathbb{R} , define the family of distributions

$$\begin{aligned} \mathcal{H}(\mathbb{R}) &:= \{F_{X,Y} : (X_i, Y_i) \sim F_{X,Y} \text{ with support } \mathcal{S}(F_{X,Y}) = (\omega_1, \omega_2)^2 \\ &\subseteq \mathbb{R}^2, X_i \sim F_X \text{ and } Y_i \stackrel{\text{iid}}{\sim} F_Y\}. \end{aligned}$$

We provide the exact distribution of $\gamma_{n,m}(F_{X,Y}, r, c)$ for the PICD, $\mathbf{D}_{n,m}(F_{X,Y}, r, c)$, with $F_{X,Y} \in \mathcal{H}(\mathbb{R})$ in Theorem S5.1 in the Supplementary File.

This exact distribution for finite n and m has a simpler form when \mathcal{X} and \mathcal{Y} points are both uniformly distributed in a bounded interval in \mathbb{R} . Define $\mathcal{U}(\mathbb{R})$ as follows

$$\begin{aligned} \mathcal{U}(\mathbb{R}) &:= \{F_{X,Y} : X \text{ and } Y \text{ are independent } X_i \stackrel{\text{iid}}{\sim} \mathcal{U}(\omega_1, \omega_2) \text{ and } Y_i \stackrel{\text{iid}}{\sim} \mathcal{U}(\omega_1, \omega_2), \\ &\text{with } -\infty < \omega_1 < \omega_2 < \infty\}. \end{aligned}$$

Clearly, $\mathcal{U}(\mathbb{R}) \subsetneq \mathcal{H}(\mathbb{R})$. Then we have Corollary S5.2 to Theorem S5.1 (see the Supplementary File).

For $n, m < \infty$, the expected value of domination number is

$$\begin{aligned} \mathbf{E}[\gamma_{n,m}(F_{X,Y}, r, c)] &= P(X_{(1)} < Y_{(1)}) + P(X_{(n)} > Y_{(m)}) \\ &\quad + \sum_{i=1}^{m-1} \sum_{k=1}^n P(N_i = k) \mathbf{E}[\gamma_{[i]}(F_i, r, c)] \end{aligned} \quad (12)$$

see Supplementary File for details and its limit as $n \rightarrow \infty$.

Theorem 5.1 (Main Result 4): Let $\mathbf{D}_{n,m}(F_{X,Y}, r, c)$ be the PICD with $F_{X,Y} \in \mathcal{H}(\mathbb{R})$. Then

- (i) for fixed $n < \infty$, $\lim_{m \rightarrow \infty} \gamma_{n,m}(F_{X,Y}, r, c) = n$ a.s. for all $r \geq 1$ and $c \in [0, 1]$.
For fixed $m < \infty$, and
- (ii) for $r = 1$ and $c \in (0, 1)$, $\lim_{n \rightarrow \infty} P(\gamma_{n,m}(F_{X,Y}, 1, c) = 2m) = 1$ and $\lim_{n \rightarrow \infty} P(\gamma_{n,m}(F_{X,Y}, 1, 0) = m + 1) = \lim_{n \rightarrow \infty} P(\gamma_{n,m}(F_{X,Y}, 1, 1) = m + 1) = 1$,
- (iii) for $r > 2$ and $c \in (0, 1)$, $\lim_{n \rightarrow \infty} P(\gamma_{n,m}(F_{X,Y}, r, c) = m + 1) = 1$,
- (iv) for $r = 2$, if $c \neq 1/2$, then $\lim_{n \rightarrow \infty} P(\gamma_{n,m}(F_{X,Y}, 2, c) = m + 1) = 1$; if $c = 1/2$, then $\lim_{n \rightarrow \infty} \gamma_{n,m}(F_{X,Y}, 2, 1/2) \stackrel{d}{=} m + 1 + \sum_{i=1}^m B_i$ with $B_i \sim \text{BER}(p(F_i, 2, 1/2))$,
- (v) for $r \in [1, 2)$, if $r \neq r^* = 1/\max(c, 1 - c)$, then $\lim_{n \rightarrow \infty} \gamma_{n,m}(F_{X,Y}, r, c)$ is degenerate; otherwise, it is non-degenerate. That is, for $r \in [1, 2)$, as $n \rightarrow \infty$,

$$\gamma_{n,m}(F_{X,Y}, r, c) \xrightarrow{\mathcal{L}} \begin{cases} m + 1, & \text{for } r > r^*, \\ m + 1 + \sum_{i=1}^m B_i, & \text{for } r = r^*, \\ 2m, & \text{for } r < r^*, \end{cases} \quad (13)$$

where $B_i \sim \text{BER}(p(F_i, r, c))$.

Proof: Part (i) is trivial. Part (ii) follows from Propositions S2.1 and S2.2, since as $n_i \rightarrow \infty$, we have $\mathcal{X}_{[i]} \neq \emptyset$ a.s. for all i .

Part (iii) follows from Theorem 3.4, since for $c \in (0, 1)$, it follows that $r > r^*$ implies $r > 2$ and as $n_i \rightarrow \infty$, we have $\gamma_{[i]}(F_i, r, c) \rightarrow 1$ in probability for all i .

In part (iv), for $r = 2$ and $c \neq 1/2$, based on Corollary S3.2, as $n_i \rightarrow \infty$, we have $\gamma_{[i]}(F_i, r, c) \rightarrow 1$ in probability for all i . The result for $r = 2$ and $c = 1/2$ is proved in [14].

Part (v) follows from Theorem 3.4. ■

The PICD discussed in this article can be viewed as the one-dimensional version of proportional-edge proximity catch digraphs introduced in [28] for two-dimensional data. The extension to higher dimensions \mathbb{R}^d with $d > 2$ is also provided in [28,36].

6. Practical application: testing uniformity with domination number of PICDs

Let X_i , $i = 1, 2, \dots, n$, be iid random variables from a distribution F with finite support. We will employ domination number of the PICD to test for uniformity of one-dimensional data in a bounded interval, say $(0, 1)$; i.e. our null hypothesis is $H_0 : F = \mathcal{U}(0, 1)$. For this purpose, we consider three approaches:

approach (i) In Theorem 3.3, we derived the $P(\gamma_{n,2}(\mathcal{U}, r, c) = 2)$ for all $n \geq 2$, $c \in (0, 1)$ and $r \geq 1$ for uniform data on $(0, 1)$. In this approach, we will use $\gamma_{n,m}(\mathcal{U}, r, c)$ as an *approximate binomial test statistic* for testing uniformity of data in $(0, 1)$ (by Theorem 3.1, the results would also be valid for uniform data on any bounded interval (a_1, a_2) with $-\infty < a_1 < a_2 < \infty$). Here, the approximation is not the large sample

convergence to binomial distribution, but in estimating the probability of success (i.e. $P(\gamma_{n,2}(\mathcal{U}, r, c) = 2)$) as we are using the expected number of observations for n_i for each subinterval i under uniformity assumption.

approach (ii) In Theorem S5.1 in the Supplementary File, we have the exact distribution of $\gamma_{n,m}(F, r^*, c)$. One could use this distribution in an exact testing procedure, but for convenience, we estimate the Monte Carlo critical values of $\gamma_{n,m}(F, r^*, c)$ and use it in our tests.

approach (iii) In Theorem 3.4, we have the asymptotic distribution of $\gamma_{n,m}(F, r^*, c)$. We will use this distribution in an approximate testing procedure, where the asymptotic value of the probability of success (i.e. $\lim_{n \rightarrow \infty} P(\gamma_{n,2}(\mathcal{U}, r, c) = 2)$) is used in the binomial test (i.e. large sample approximation is used for the probability of success).

In approaches (i)–(iii), we divide the interval $(0, 1)$ into m subintervals, and treat the interval endpoints to be the \mathcal{Y} points, i.e. we set $\mathcal{Y}_m = \{0, 1/(m-1), 2/(m-1), \dots, 1\}$. This can be done without loss of generality in this context, because we are testing uniformity of points from one class in a bounded interval, and the proximity regions are constructed using arbitrarily chosen \mathcal{Y} points.

In both approaches, we compute the domination number for each subinterval and use $G_n := \gamma_{n,m}(r, c) - m$ as our test statistic. However in approach (i), we use an approximate binomial test with G_n approximately having $\text{BIN}(m, p_u(r, c, n_i))$ with $n_i = \lfloor n/m \rfloor$. This is an approximate procedure since $\mathbf{E}[N_i] = n/m$, i.e. $n_i = n/m$ on the average. Furthermore, if $G_n < 0$, then we set the corresponding p -value to 0 for this test, since this is already evidence of severe deviation from uniformity. In approach (ii), we use the exact distribution provided in Theorem S5.1. However, *for convenience, we estimate the critical value by Monte Carlo simulations*. In particular, we generate 10,000 samples for each (r, c) combination considered and compute the domination number $\gamma_{n,m}(r, c)$ for each sample. Then for the left-sided (right-sided) alternative, 5th percentile (95th percentile) of the test statistic constitutes the empirical critical value at $\alpha = 0.05$ level.

For comparative purposes, we employ Kolmogorov–Smirnov (KS) test for uniform distribution and Pearson's χ^2 goodness of fit test, since these are the most well known and commonly used tests for checking the goodness of distributional fit. We also consider three recently proposed tests, namely, a uniformity test based on Too-Lin characterization of the uniform distribution [22], and two entropy-based tests, denoted as TB1 and TB2 in [27]. The entropy tests due to [27] reject the null hypothesis of uniformity for small values of TB1 and TB2. On the other hand, the uniformity test denoted as $T_n^{(m)}$ in [22], uses $m = 2$ and k th order statistic Too-Lin characterization rejects for large absolute values of the test statistic and we take $k = 1$ in $T_n^{(2)}$. For all these tests TB1, TB2 and $T_n^{(2)}$, the critical values are obtained by Monte Carlo simulations.

We also compare the performance of PICD domination number test with that of the arc density of two ICDs: (i) PICD and (ii) Central ICD (CICD) which is based on central similarity (CS) proximity region. For a digraph $D_n = (\mathcal{V}, \mathcal{A})$ with vertex set \mathcal{V} and arc set \mathcal{A} , the arc density of D_n which is of order $|\mathcal{V}| = n \geq 2$, denoted $\rho(D_n)$, is defined as $\rho(D_n) = \frac{|\mathcal{A}|}{n(n-1)}$ where $|\cdot|$ stands for the set cardinality function [37]. So $\rho(D_n)$ is 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 $\rho(D_n) = 0$. Arc density of ICDs is shown to be a U -statistic, and hence its asymptotic distribution is a normal distribution,

provided that its asymptotic variance is positive [26]. Arc density of PICDs is studied in [26] and but not used in testing uniformity before. Likewise, CICDs were introduced in [25] and its arc density was employed for testing uniformity in the same article as well. CS proximity region is defined as follows [25]: For $\tau > 0$, $c \in (0, 1)$ and $x \in \mathcal{I}_i$

$$N_{CS}(x, \tau, c) = \begin{cases} \left(x - \tau(x - Y_{(i-1)}), x + \frac{\tau(1-c)}{c}(x - Y_{(i-1)}) \right) & \text{if } x \in (Y_{(i-1)}, M_{c,i}), \\ \cap (Y_{(i-1)}, Y_{(i)}) & \\ \left(x - \frac{c\tau}{1-c}(Y_{(i)} - x), x + \tau(Y_{(i)} - x) \right) & \text{if } x \in (M_{c,i}, Y_{(i)}). \\ \cap (Y_{(i-1)}, Y_{(i)}) & \end{cases} \quad (14)$$

6.1. Empirical size analysis

We perform a size analysis to determine whether the tests have the appropriate size in testing $H_0 : F = \mathcal{U}(0, 1)$. Along this line, we partition the domain of $p_u(r, c, n)$ for r and c as follows. We take $c = .01, .02, \dots, .99$ and $r = 1.00, 1.01, \dots, 2.10$, and consider each (r, c) combination on a 99×210 grid with $n = 20, 50, 100$. For each (r, c) combination, we generate $N_{mc} = 10,000$ samples each of size n iid from $\mathcal{U}(0, 1)$ distribution. We also partition the interval $(0, 1)$ into m equal subintervals where m equals \sqrt{n} (rounded to the nearest integer) whose choice is inspired by the choice of windows size in entropy-based goodness-of-fit tests [38]. This choice is not to justify the use of binomial distribution, as the distribution of the domination number is available for any $r > 1$, $c \in (0, 1)$ and finite $n \geq 2$. That is, the binomial distribution would hold regardless of the size of m , but it is preferable that it is large enough to give enough resolution for the discrete binomial test. The reason we use the (r^*, c) combination that renders the asymptotic distribution non-degenerate is that other choices of (r, c) could make the distribution close to being degenerate for large n , whose rate of convergence to 0 or 1 depends on the values of r and c . Then for each subinterval, we compute the domination number (which is either 0, 1, or 2), and sum the domination numbers over the m subintervals and thus obtain $\gamma_{n,m}(r, c)$. We use this summed domination number minus m , i.e. G_n , in an approximate binomial test statistic (i.e. we follow approach (i) above). Under H_0 , G_n approximately has $\text{BIN}(m, p_u(r, c, \lfloor n/m \rfloor))$ distribution, so we compute the p -value based on the binomial test with m trials and probability of success being $p = p_u(r, c, \lfloor n/m \rfloor)$ for the two-sided alternative. For each of the 10,000 samples generated, we also compute the arc density of the ICDs for the parameters of choice and appeal to the asymptotic normality of the arc density of these ICDs. We compute size estimates based on the corresponding normal critical values for the arc density for each (r, c) (resp. (τ, c)) combination for PICD (resp. CICD). For each sample, we also compute KS, χ^2 , TB1 and TB2 and $T_n^{(2)}$ tests as well. In the χ^2 test, we use the same partition of $(0, 1)$ with m subintervals, and compare the observed and expected frequencies of data points in these subintervals under uniformity. Empirical size is estimated as the frequency of number of times p -value is significant at $\alpha = .05$ level divided by $N_{mc} = 10,000$. With $N_{mc} = 10,000$, empirical size estimates larger than .0536 are deemed liberal, while those less than .0464 are deemed conservative. These bounds

are also based on binomial test for the proportions for $N_{mc} = 10,000$ trials at .05 level. Since the entropy tests TB1 and TB2 and $T_n^{(2)}$ test and PICD domination number test with approach (ii) are using critical values based on Monte Carlo simulations, we exclude them in the empirical size comparison, as they, by construction, attain the nominal size. However, we find the empirical critical values for these tests as the sample 100α th percentile of the TB1 and TB2 values computed in our simulations, and $100(1 - \alpha)$ th percentile of the $|T_n^{(2)}|$ values computed in our simulations.

We present the empirical size estimates of the tests based on the *domination number of PICD with approach (i)* as two-level image plots (with empirical sizes not significantly different from 0.05 in black dots, and others are blanked out in white) with $n = 20$, $c = .01, .02, \dots, .99$ and $r = 1.00, 1.01, \dots, 2.10$ in Figure 4 (the plots for $n = 50$ and 100 have the similar trend, hence not presented). Notice that the sizes for the right-sided alternatives are at about the nominal level for (r, c) around $(1, 0)$ or $(1, 1)$, while the sizes for the left-sided alternatives are about the nominal level of 0.05 at the asymptotically non-degenerate $(r, c) = (r^*, c)$ pairs for $c \in (.25, .75)$. The reason for the asymmetric performance for the left-sided versus right-sided alternatives is that $p_u(r, c, n)$ values are higher (i.e. close to 1) around $(r, c) = (1, 0)$ or $(1, 1)$, and lower for other values, but away from 1 or 0 for $(r, c) = (r^*, c)$ pairs. Therefore, for the power analysis, we only consider $(r, c) = (r^*, c)$ pairs, as empirical size is closer to the nominal level for these parameters in approach (i).

In approach (ii), by construction the size estimates should be around the nominal level of .05. But due to the discrete nature of $\gamma_{n,m}$ with very few atoms for small n and m , the exact test is liberal or conservative depending on whether we include the critical value in our size estimation. In particular, let $\gamma_{n,m,i}$ be the domination number for sample i and $\gamma_{.05}$ be the 5th percentile for the exact distribution of $\gamma_{n,m}(r, c)$ (as in Theorem S5.1). Also let $\alpha_{inc} := \sum_{i=1}^{N_{mc}} \gamma_{n,m,i} \leq \gamma_{.05}$ and $\alpha_{exc} := \sum_{i=1}^{N_{mc}} \gamma_{n,m,i} < \gamma_{.05}$. Then for testing the left-sided alternative, α_{inc} tends to be much larger than .05 (implying the procedure is liberal) and α_{exc} tends to be much smaller than .05 (implying the procedure is conservative). In our power computations with approach (ii), we adjust for this discrepancy.

The size estimates in approach (iii) depend on the sample size n , and the parameters r and c , i.e. they tend to be liberal for some values of (r, c) , and conservative for others,

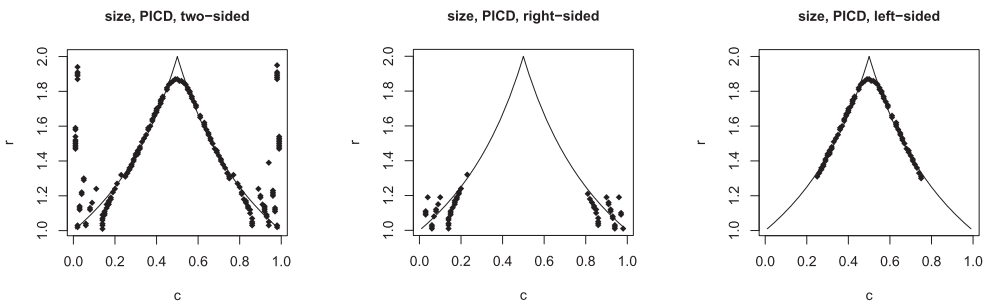


Figure 4. The empirical size estimates of the tests based on *domination number of PICD with approach (i)* for $n = 20$ and $N_{mc} = 10,000$ for $r = 1.01, 1.02, \dots, 2.10$ and $c = .01, .02, \dots, .99$ for the two-sided, right-sided and left-sided alternatives (left to right); size estimates significantly different from .05 are blanked out, while size estimates within .0536 and .0464 are plotted as black dots. The solid lines in the bottom row plots indicate the case of $(r, c) = (r^*, c)$ which yields the asymptotically non-degenerate distribution for the domination number.

especially when n is not large enough. Our simulations suggest that large sample sizes are needed (about 30 or more per each subinterval seems to work), where the required sample size would also depend on r and c as well. Hence we do not present approach (iii) except for the large sample simulation cases (in the cases with $n = 1000$ here).

We estimate the empirical sizes of the tests based on the arc density of the PICDs and CICDs for $n = 20, 50$ and 100 and $c = .01, .02, \dots, .99$ with $r = 1.1, 1.2, \dots, 10.0$ for PICDs and $\tau = .1, .2, \dots, 10.0$ for CICDs. For the one-sided alternatives, the regions at which size estimates are about the nominal level of 0.05 are somewhat complementary, in the sense that, the sizes are appropriate for the parameter combinations in one region for left-sided alternative and mostly in its complement for the right-sided alternative. We also observe that arc density of PICD has appropriate size for the two-sided alternative for more parameter combinations, and arc density of CICD has appropriate size for the left-sided alternative for more parameter combinations. See Figure S3 in the Supplementary File for the related image plots of the empirical size estimates.

6.2. Empirical power analysis

We perform a power analysis to determine which tests have better performance in detecting deviations from uniformity. For the alternatives (i.e. deviations from uniformity), we consider five types of non-uniform distributions with support in $(0, 1)$:

- (I) $f_1(x, \delta) = (2\delta x + 1 - \delta)\mathbf{I}(0 < x < 1)$,
- (II) $f_2(x, \sigma) = \phi(x, 1/2, \sigma)/(\Phi(1, 1/2, \sigma) - \Phi(0, 1/2, \sigma))\mathbf{I}(0 < x < 1)$ where $\phi(x, 1/2, \sigma)$ is the pdf for normal distribution with mean $\mu = 1/2$ and standard deviation σ , (i.e. normal distribution with $\mu = 1/2$ restricted to $(0, 1)$),
- (III) $f_3(x, \delta) = (\delta(x - 1/2)^2 + 1 - \delta/12)\mathbf{I}(0 < x < 1)$,
- (IV) $f_4(x, \varepsilon) = (1/(1 - 2m\varepsilon))\mathbf{I}(x \in (0, 1) \setminus \cup_{i=0}^m (i/m - \varepsilon, i/m + \varepsilon))$, that is, $f_4(x, \varepsilon)$ is a pdf so that $\varepsilon \times 100$ % of the regions around the m subinterval end points are prohibited, and the data is uniform in the remaining regions.
- (V) $f_5(x, \varepsilon') = (1/2m\varepsilon')\mathbf{I}(x \in (0, 1) \cap (\cup_{i=0}^m (i/m - \varepsilon', i/m + \varepsilon')))$, that is, $f_5(x, \varepsilon')$ is a distribution so that data is uniform over the $\varepsilon' \times 100$ % of the regions around the m subinterval end points are prohibited, and the remaining regions are prohibited. Notice that the supports of $f_4(x, \varepsilon)$ and $f_5(x, \varepsilon')$ are complimentary in $(0, 1)$.

That is,

$$H_a^I : f = f_1(x, \delta) \text{ with } \delta \in (0, 1) H_a^{II} : f = f_2(x, \sigma) \text{ with } \sigma > 0$$

$$H_a^{III} : f = f_3(x, \delta) \text{ with } \delta \in (0, 12]$$

$$H_a^{IV} : f = f_4(x, \varepsilon) \text{ with } \varepsilon \in (0, 1/2) \text{ and } H_a^V : f = f_5(x, \varepsilon') \text{ for } \varepsilon' \in (0, 1/2)$$

In type I alternatives, $\delta = 0$ corresponds to $\mathcal{U}(0, 1)$ distribution, and with increasing $\delta > 0$, the density of the distribution is more clustered around 1 and less clustered around 0; in type II alternatives, with decreasing σ , the density of the distribution gets more clustered around 1/2 (and less clustered around the end points, 0 and 1); and in type III alternatives, $\delta = 0$ corresponds to $\mathcal{U}(0, 1)$ distribution, and with increasing $\delta > 0$, the density of the distribution is more clustered around the end points, 0 and 1, and less clustered around

1/2. Types IV and V alternatives are actually motivated from two-class one-dimensional spatial point patterns called segregation and association. Roughly defined, segregation is the pattern in which points from the same class tend to cluster, while under association, points from one class is clustered around the points from the other class and vice versa. In one-dimensional case, the segregation alternative is as in H_a^{IV} , where X points are distributed according to f_4 and Y points constitute the end points of the interval partition of $(0, 1)$ (i.e. $\{0, 1/(m-1), 2/(m-1), \dots, 1\}$). Hence, X points tend to stay away from Y points, which suggests segregation between the classes X and Y . Furthermore, $\varepsilon = 0$ in type IV alternative corresponds to the null case (i.e. uniform distribution). The association alternative is as in H_a^V . The pdf under type I alternative is skewed left for $\delta > 0$, while pdfs under other alternatives are symmetric around 1/2. See Figure S6 in the Supplementary File for sample plots of the pdfs with various parameters for alternative types I–III.

Under each alternative, we generate n points according to the specified alternatives with various parameters. In particular, for $H_a^I : F = F_1(x, \delta)$, we consider $\delta = .2, .4, .6, .8$, for $H_a^{II} : F = F_2(x, \sigma)$, we consider $\sigma = .1, .2, .3, .4$, for $H_a^{III} : F = F_3(x, \delta)$, we consider $\delta = 2, 4, 6, 8$, and for $H_a^{IV} : F = F_4(x, \varepsilon)$, we consider $\varepsilon = .1, .2, .3, .4$ (also called H_a^{IV} -case (1)). For the domination number of PICDs, we replicate each case N_{mc} times for $(r, c) = (r^*, c)$ with $c = .01, .02, \dots, .99$ (i.e. for (r, c) values that make $\gamma_{n,2}(\mathcal{U}, r, c)$ non-degenerate in the limit (see Theorem 3.4)). We compute the power using the critical values based on $\text{BIN}(m, p_u(r, c, \lfloor n/m \rfloor))$ distribution (i.e. approach (i)) and based on the empirical critical values (i.e. approach (ii)). For types I–IV alternatives, we take $n = 20, 50, 100$ and $N_{mc} = 10,000$. By construction, our domination number test is more sensitive for segregation/association type alternatives which also implies the same direction for each subinterval considered hence, the sum of domination number over the subintervals detects such deviations from uniformity better. In fact, we have consistency results for the domination number test under H_a^{IV} and H_a^V type alternatives (see Section 7). These consistency results suggest that domination number test gets very sensitive under very mild forms of H_a^{IV} and H_a^V when n gets large. Along this line, we consider two more cases for the type IV alternative in addition to case (1). More specifically, we consider H_a^{IV} -case (1): $\varepsilon = .1, .2, .3, .4$ $n = 50$, $m = 7 \approx \sqrt{n}$ and $N_{mc} = 10,000$, H_a^{IV} -case (2): $n = 1000$, $m = 32 \approx \sqrt{n}$ and $N_{mc} = 1000$; and H_a^{IV} -case (3): $n = 1000$, $m = 20$ and $N_{mc} = 1000$ where in cases (2) and (3) we take $\varepsilon = .01, .02, .03, .04$.

For the arc density of the ICDs, we generate n points according to the specified alternatives with various parameters (where n is taken as in the simulations for the domination number for the null case and each alternative). With CICDs, we use (τ, c) for $\tau = .1, .2, \dots, 10.0$ and $c = .01, .02, \dots, .99$ and with PICDs, we use (r, c) for $r = 1.1, 1.2, \dots, 10.0$ and $c = .01, .02, \dots, .99$. With CICDs, for each (τ, c) and δ combination, and with PICDs, for each (r, c) and δ combination, we replicate the sample generation N_{mc} times. We compute the power using the asymptotic critical values based on the normal approximation. We also keep the parameter combinations $((r, c)$ for PICDs and (τ, c) for CICDs) at which the tests have the appropriate level (of .05), i.e. if the test is conservative or liberal for the one-sided version in question, we ignore that parameter combination in our power estimation, as they would yield unreliable results which might have a substantial effect on the power values. We call this procedure the ‘size adjustment’ for power estimation. For the arc density of PICDs and CICDs, we only report the maximum power estimates under each alternative.

The power comparisons between PICD domination number test, KS, χ^2 , TB1, TB2 and $T_n^{(2)}$ tests are presented in Figure 5 for alternatives $H_a^I - H_a^{III}$, and in Figure 6 for alternatives H_a^{IV} -cases (1)-(3). The power estimates based on asymptotic critical values of the tests (i.e. the power estimates for the test based on domination number of PICD with approach (i), Kolmogorov–Smirnov test, and Chi-square test) are provided in the top row and those based on Monte Carlo critical values (for the test based on domination number of PICD with approach (ii), $T_n^{(2)}$ test based on the uniformity characterization, two versions of the entropy-based tests) are provided in the bottom row in these figures. The power estimates under alternatives $H_a^I - H_a^{III}$ and H_a^{IV} -case (1) are presented in Table 1, and those under alternative H_a^{IV} -cases (2) and (3) in Table 2; in both tables the power estimates are rounded to two decimal places. In Figures 5 and 6, we do not present the power estimates for ICD arc density tests, due to the difficulty in presentation since ICD arc density tests depend on two parameters. For the domination number test, the power estimates based on asymptotic critical values are provided in the top row and those based on Monte Carlo critical values are provided in the bottom row in these figures. *In Tables 1 and 2, we only present the maximum power estimates for the ICD arc density tests for the two-sided alternative and for the PICD domination number tests.* Considering Figures 5 and 6 and Tables 1 and 2, we observe that power estimates increase as the departure from uniformity gets more severe. In particular, power estimate increases as δ increases in H_a^I or H_a^{II} , as ε increases in H_a^{IV} and as σ decreases in H_a^{II} . Under $H_a^I - H_a^{III}$ and H_a^{IV} -case(1), arc density of PICD and CICD

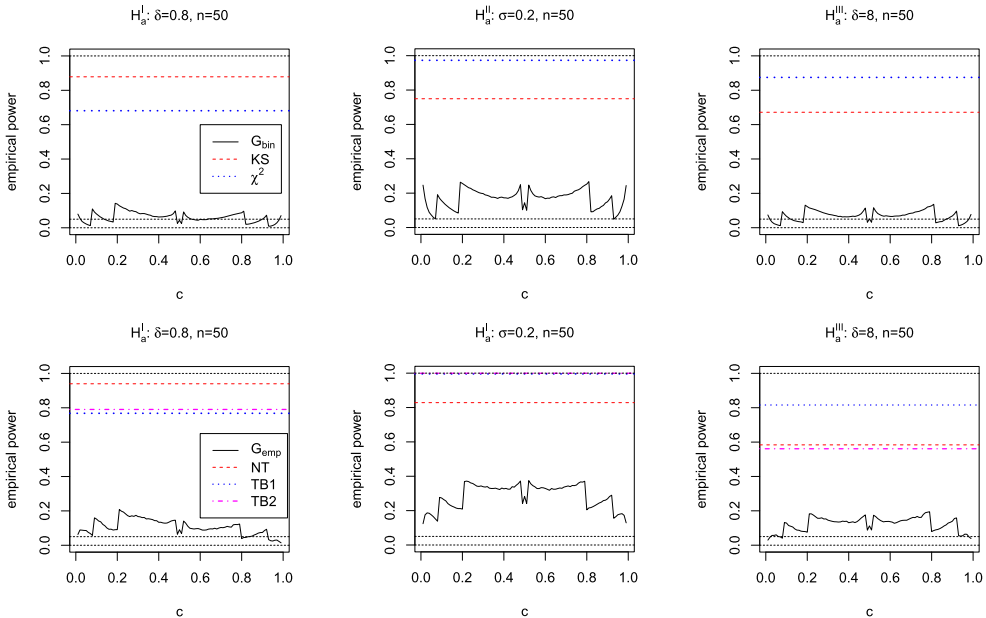


Figure 5. Power estimates under under $H_a^I : F = F_1(x, \delta = .8)$, $H_a^{II} : F = F_2(x, \sigma = 0.2)$, $H_a^{III} : F = F_3(x, \delta = 8)$, with $n = 50$ and $N_{mc} = 10000$. G_{bin} and G_{emp} : tests based on domination number of PICD with approaches (i) and (ii), respectively, KS: Kolmogorov–Smirnov test, χ^2 : Chi-square test, NT: $T_n^{(2)}$ test based on the uniformity characterization, TB1 and TB2: two versions of the entropy-based tests. Tests presented in each row are indicated in the legend in that row.

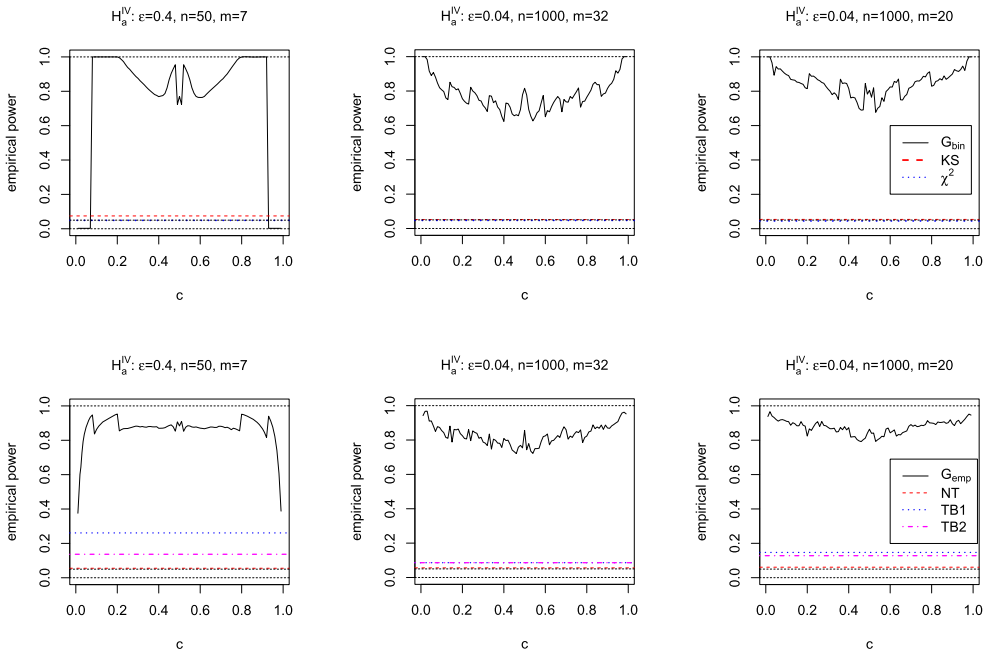


Figure 6. Power estimates under under $H_a^{IV} : F = F_1(x, \varepsilon)$, case (1) with $\varepsilon = .4, n = 50, m = 7$ and $N_{mc} = 10,000$ (left column), case (2) $\varepsilon = .04, n = 1000, m = 32$ and $N_{mc} = 1000$ (middle column), and case (3) $\varepsilon = .04, n = 1000, m = 20$ and $N_{mc} = 1000$ (right column). Labelling of the tests are as in Figure 5. Tests presented in each row are indicated in the legend in that row.

has the highest power estimates, where PICD arc density tends to perform better (worse) than CICD arc density under $H_a^I - H_a^{III}$ (under H_a^{IV} -case (1)). Under H_a^I , ICD arc density tests are followed by $T_n^{(2)}$; under H_a^{II} , ICD arc density tests are followed by TB1 and TB2; under H_a^{III} , ICD arc density tests are followed by χ^2 test; and under H_a^{IV} -case (1), ICD domination number test is followed by CICD arc density test. Under H_a^{IV} -cases (2) and (3) ICD domination number tests have the highest power estimates, where under case (1) PICD domination number test with approach (i) and under case (2) domination number test with approach (ii) has better performance, and power estimates for the other tests are just above .05 or at about .05. In these large sample cases, approach (iii) also works, and has higher power estimates than the other two approaches (corresponding estimates not presented to be consistent with the presentations of the other alternatives). Moreover, PICD domination number test performs better when the support is partitioned by $m \approx \sqrt{n}$. We omit the power performance under H_a^V as it is the opposite pattern to the one under H_a^{IV} . More simulation results for the arc density of ICDs are presented in in the Supplementary File, where we observe that the power estimates are symmetric around $c = 1/2$ under types II-IV alternatives, which is in agreement with the symmetry in the corresponding pdfs (around $c = 1/2$).

We also considered the power comparisons under $H_a^I - H_a^{III}$ and H_a^{IV} case (1) at the same alternative parameters with $n = N_{mc} = 1000$, to see the effect of the large samples on the power estimates. The results are similar to those in the smaller sample cases, with higher power for each test (hence not presented). In particular, under $H_a^I - H_a^{III}$ all tests

Table 1. The power estimates under the alternatives H_a^I to H_a^{IV} with all four parameter values considered and $n = 50$ and $N_{mc} = 10,000$ for the tests we employed.

	$H_a^I, n = 50$				$H_a^{II}, n = 50$				
	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.6$	$\delta = 0.8$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	
PICD	.33	.51	.76	.94	PICD	1.00	1.00	.79	.39
	.15	.35	.64	.88		1.00	1.00	.79	.37
CICD	.19	.43	.71	.92	CICD	1.00	1.00	.81	.42
	.13	.33	.62	.88		1.00	1.00	.81	.41
G_{bin}	.06	.08	.09	.14	G_{bin}	.99	.25	.07	.06
G_{emp}	.06	.09	.12	.21	G_{emp}	.95	.37	.11	.07
KS	.11	.32	.62	.88	KS	1.00	.75	.17	.07
χ^2	.07	.17	.38	.68	χ^2	1.00	.97	.39	.14
$T_n^{(2)}$.13	.37	.70	.94	$T_n^{(2)}$	1.00	.83	.17	.07
TB1	.08	.20	.43	.77	TB1	1.00	.99	.54	.22
TB2	.08	.20	.43	.79	TB2	1.00	1.00	.68	.31

	$H_a^{III}, n = 50$				$H_a^{IV}, n = 50$				
	$\delta = 2$	$\delta = 4$	$\delta = 6$	$\delta = 8$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.3$	$\varepsilon = 0.4$	
PICD	.41	.66	.92	.99	PICD	.27	.27	.34	.50
	.28	.66	.92	.99		.06	.08	.10	.14
CICD	.19	.53	.86	.98	CICD	.11	.15	.28	.52
	.19	.52	.86	.98		.06	.07	.08	.24
G_{bin}	.07	.07	.09	.14	G_{bin}	.28	1.00	1.00	1.00
G_{emp}	.06	.07	.11	.19	G_{emp}	.88	.95	.95	.95
KS	.09	.19	.40	.67	KS	.05	.06	.06	.07
χ^2	.09	.27	.38	.87	χ^2	.05	.05	.05	.05
$T_n^{(2)}$.07	.14	.31	.58	$T_n^{(2)}$.05	.05	.05	.06
TB1	.07	.19	.48	.82	TB1	.07	.10	.15	.26
TB2	.03	.06	.21	.56	TB2	.06	.07	.09	.14

Notes: PICD and CICD represent the arc densities of the ICD tests, and for each, top row is without size adjustment and bottom row is with size adjustment (see the text for the description of size adjustment), G_{bin} and G_{emp} : tests based on domination number of PICD with approaches (i) and (ii), KS: Kolmogorov–Smirnov test, χ^2 : Chi-square test, NT: $T_n^{(2)}$ test based on the uniformity characterization, TB1 and TB2: two versions of the entropy-based tests.

have much higher power, with most having power virtually 1.00, but domination number tests with approaches (i) and (ii) exhibit mild improvement, while under H_a^{IV} case (1), PICD domination number tests attain the highest power estimates, virtually 1.00, while there is mild improvement in the performance of other tests, except for TB1 and TB2, which show moderate improvement. We also observe that in the large sample case, PICD domination number with approach (iii) attains very high power under each alternative.

The above methodology can easily be extended for testing non-uniform distributions (see Remark S5.4 in the Supplementary File).

7. Consistency of the tests based on domination number of PICDs under H_a^{IV} and H_a^V

Let b_α be the $\alpha \times 100$ th percentile of the binomial distribution $BIN(m, p_u(r, c, \lfloor n/m \rfloor))$.

Theorem 7.1 (Consistency – Type I): Let $\gamma_{n,m}(F, r, c)$ be the domination number under segregation and association alternatives, H_a^{IV} and H_a^V , respectively, in the multiple interval case with m intervals. The test against segregation with $F = F_4(x, \varepsilon)$ which rejects for

Table 2. The power estimates under the alternatives H_a^{IV} with all four ε values considered and $n = 1000$ and $N_{mc} = 1000$ for the tests we employed.

$H_a^{IV}, n = 1000, m = 32, N_{mc} = 1000$				
	$\varepsilon = 0.01$	$\varepsilon = 0.02$	$\varepsilon = 0.03$	$\varepsilon = 0.04$
PICD	.08	.08	.08	.08
	.08	.08	.08	.08
CICD	.08	.08	.08	.08
	.07	.08	.08	.08
DN	.49	1.00	1.00	1.00
	.66	.95	.96	.97
KS	.04	.04	.04	.05
χ^2	.05	.05	.06	.05
$T_n^{(2)}$.04	.04	.04	.06
TB1	.04	.07	.08	.09
TB2	.04	.06	.07	.09
$H_a^{IV}, n = 1000, m = 20, N_{mc} = 1000$				
	$\varepsilon = 0.01$	$\varepsilon = 0.02$	$\varepsilon = 0.03$	$\varepsilon = 0.04$
PICD	.08	.08	.08	.08
	.08	.08	.08	.08
CICD	.08	.08	.08	.07
	.08	.08	.08	.07
G_{bin}	.40	1.00	1.00	1.00
G_{emp}	.61	.95	.95	.97
KS	.06	.04	.05	.05
χ^2	.05	.05	.04	.05
$T_n^{(2)}$.06	.05	.05	.06
TB1	.04	.08	.09	.15
TB2	.04	.07	.09	.13

Labelling of the tests is as in Table 1.

$G_n < b_\alpha$ and the test against association with $F = F_5(x, \varepsilon')$ which rejects for $G_n > b_{1-\alpha}$ are consistent.

Proof: Given $F = F_4(x, \varepsilon)$. Let $\gamma_{n,m}(\mathcal{U}, r, c)$ be the domination number for \mathcal{X}_n being a random sample from $\mathcal{U}(0, 1)$. Then $P(\gamma_{n,m}(F, r, c) = 1) \geq P(\gamma_{n,m}(\mathcal{U}, r, c) = 1)$; and $P(\gamma_{n,m}(F, r, c) = 2) \leq P(\gamma_{n,m}(\mathcal{U}, r, c) = 2)$. Hence $G_n < mp_u(r, c, \lfloor n/m \rfloor)$ with probability 1, as $n \gg m \rightarrow \infty$. Furthermore, $\text{BIN}(m, p_u(r, c, \lfloor n/m \rfloor))$ distribution converges to normal distribution with mean $mp_u(r, c, \lfloor n/m \rfloor)$ and variance $mp_u(r, c, \lfloor n/m \rfloor)(1 - p_u(r, c, \lfloor n/m \rfloor))$. Hence consistency follows from the consistency of tests which have asymptotic normality. The consistency against the association alternative can be proved similarly. ■

Below we provide a result which is stronger, in the sense that it will hold for finite m as $n \rightarrow \infty$. Let $\bar{G}_n := G_n/m$ (i.e. domination number averaged over the number of subintervals) and z_α be the $\alpha \times 100$ th percentile of the standard normal distribution.

Theorem 7.2 (Consistency – Type II): Let $\gamma_{n,m}(F, r, c)$ be the domination numbers under segregation and association alternatives H_a^{IV} and H_a^V , respectively, in the multiple interval case with m intervals where $m < \infty$ is fixed. Let $m^*(\alpha, \varepsilon) := \lceil (\frac{\sigma \cdot z_\alpha}{\bar{G}_n(r, c) - \mu})^2 \rceil$ where $\lceil \cdot \rceil$ is

the ceiling function and ε -dependence is through $\overline{G}_{n,m}(r, c)$ under a given alternative. Then the test against H_a^{IV} which rejects for $S_{n,m} < z_\alpha$ is consistent for all $\varepsilon \in (0, \min(c, 1 - c))$ and $m \geq m^*(\alpha, \varepsilon)$, and the test against H_a^V which rejects for $S_{n,m} > z_{1-\alpha}$ is consistent for all $\varepsilon \in (0, \min(c, 1 - c))$ and $m \geq m^*(1 - \alpha, \varepsilon)$.

Proof: Let $\varepsilon \in (0, \min(c, 1 - c))$. Under H_a^{IV} , $\gamma_n(F, r, c)$ is degenerate in the limit as $n \rightarrow \infty$, which implies $\overline{G}_n(r, c)$ is a constant a.s. In particular, for $\varepsilon \in (0, \min(c, 1 - c))$, $\overline{G}_n(r, c) = 1$ a.s. as $n \rightarrow \infty$. Then the test statistic $S_{n,m} = \sqrt{m}(\overline{G}_n(r, c) - \mu)/\sigma$ is a constant a.s. and $m \geq m^*(\alpha, \varepsilon)$ implies that $S_{n,m} < z_\alpha$ a.s. Furthermore, $S_{n,m} \xrightarrow{\mathcal{L}} N(0, 1)$ as $n \rightarrow \infty$. Hence consistency follows for segregation.

Under H_a^V , as $n \rightarrow \infty$, $\overline{G}_n(r, c) = 2$ for all $\varepsilon \in (0, \min(c, 1 - c))$ a.s. Then $m \geq m^*(1 - \alpha, \varepsilon)$ implies that $S_{n,m} > z_{1-\alpha}$ a.s., hence consistency follows for association. ■

Notice that in Theorem 7.2 we actually have more than what consistency requires. In particular, we show that the power of the test reaches 1 for m greater than a threshold as $n \rightarrow \infty$.

8. Discussion and conclusions

In this article, we derive the distribution of the domination number of a random digraph family called *parameterized interval catch digraph* (PICD) which is based on two classes of points, say \mathcal{X} and \mathcal{Y} . Points from one of the classes (say, class \mathcal{X}), denoted \mathcal{X}_n , constitute the vertices of the PICDs, while the points from the other class (say, class \mathcal{Y}), denoted \mathcal{Y}_m , are used in the binary relation that assigns the arcs of the PICDs. Our PICD is based on a parameterized proximity map which has an expansion parameter r and a centrality parameter c . We provide the exact and asymptotic distributions of the domination number of the PICDs for uniform data and compute the asymptotic distribution for non-uniform data for the entire range of (r, c) .

We demonstrate an interesting behaviour of the domination number of the PICD for one-dimensional data. For uniform data or data from a distribution which satisfies some regularity conditions and fixed finite sample size $n > 1$, the distribution of the domination number restricted to any interval is a translated form of Bernoulli distribution, $\text{BER}(p)$, where p is the probability that the domination number being 2. In the case of $\mathcal{Y}_2 = \{y_1, y_2\}$ with $\mathcal{U}(y_1, y_2)$ data, for finite $n \geq 1$, the parameter of the asymptotic distribution of the domination number of the PICD based on uniform data (i.e. probability of domination number being 2, denoted $p_u(r, c)$) is continuous in r and c for all $r \geq 1$ and $c \in (0, 1)$. For fixed $(r, c) \in [1, \infty) \times (0, 1)$, $p_u(r, c)$ exhibits some discontinuities. The asymptotic distribution of the domination number is degenerate for the expansion parameter $r > 2$ regardless of the value of c . For $c \in (0, 1)$ the asymptotic distribution is non-degenerate when the expansion parameter r equals $r^* = 1/\max(c, 1 - c)$. For $r = r^*$, the asymptotic distribution of the domination number is a translated form of $\text{BER}(p_u(r^*, c))$ where $p_u(r^*, c)$ is continuous in c . For $r > r^*$ the domination number converges in probability to 1, and for $r < r^*$ the domination number converges in probability to 2. On the other hand, at $(r, c) = (2, 1/2)$, the asymptotic distribution is again a translated form of $\text{BER}(p_u(2, 1/2))$, but there is yet another jump at $(r, c) = (2, 1/2)$, as $p_u(2, 1/2) = 4/9$ while $\lim_{(r,c) \rightarrow (2,1/2)} p_u(r^*, c) = 2/3$. This second jump is due

to the symmetry for the domination number at $c = 1/2$ (see the discussion at the end of Section 3.1.1).

We employ domination number for testing uniformity of one-dimensional data. In this application, we have n \mathcal{X} points and we take m \mathcal{Y} points to be the equidistant points in the support of \mathcal{X} points. For example, if the support of \mathcal{X} points is $(0, 1)$, we take \mathcal{Y} points to be $\mathcal{Y}_m = \{0, 1/(m-1), 2/(m-1), \dots, 1\}$. Since under H_0 , H_a^{IV} and H_a^{V} the data is uniform with different support regions, we can extend the methodology to the random \mathcal{Y}_m case, but currently the method is only applicable given \mathcal{Y}_m as above.

We compare the size and power performance of PICD domination number with two well known tests, namely, Kolmogorov–Smirnov (KS) test and Pearson’s χ^2 goodness-of-fit test, three recently introduced tests, the uniformity test based on Too-Lin characterization, denoted as $T_n^{(2)}$ [22], and two entropy-based tests, denoted as TB1 and TB2 in [27], and also the arc density of PICDs and of another ICD family called central ICD (CICD), by Monte Carlo simulations. Based on the simulation results, we see that ICDs have better performance than their competitors (in terms of size and power). Arc density of ICDs perform better than others under most alternatives for some of the parameter values and the domination number outperforms others under certain types of alternatives. In particular, under the alternatives $H_a^{\text{I}} - H_a^{\text{III}}$, ICD arc density tests outperform other tests, and under H_a^{IV} -cases (1)–(3), PICD domination number tests outperform other tests. For the ICD arc density tests, we use the asymptotic critical values based on normal approximation. For the PICD domination number test, we use the binomial critical values with an approximate probability of success (i.e. approach (i)) and also the empirical critical values based on Monte Carlo simulations (i.e. approach (ii)). For $T_n^{(2)}$, TB1 and TB2 tests, the critical values are also based on Monte Carlo simulations.

We recommend using the PICD domination number test for uniformity in the following scenario. If we are testing uniformity of data in multiple intervals (by hypothesis or one can partition the support of the data), and the deviation from uniformity is in the same direction at each interval, then, by construction, domination number tends to be more sensitive to detect such alternatives (even if they are very mild deviations from uniformity). Among the types of critical value computations, we recommend the use of the exact distribution provided in Theorem S5.1 (with Monte Carlo critical values as an approximation in practice), i.e. approach (ii) for small samples (this approach could be used provided running time is feasible), and the approximate Binomial test for any n , i.e. approach (i) (see Section 6.1). For large samples, binomial test with asymptotic probability of success (i.e. approach (iii)) could also be employed. Our simulations suggest that about 30 or more for each subinterval seems to work for most (r^*, c) combination, however, the sample size requirements for approach (iii) have not been studied thoroughly in this article. The relevant functions for these tests are PEdom1D and TSDomPEBin1D which are available in the R package pcds which is available on github and can be installed using the command `devtools::install_github("elvanceyhan/pcds")` in an R session. The function PEdom1D computes the domination number when one or two one-dimensional data sets are provided, and the function TSDomPEBin1D uses the finite sample binomial approximation (i.e. approach (i)) by default or can use the asymptotic binomial version (i.e. approach (iii)) for very large samples when `asy.bin = TRUE` option is employed. Monte Carlo critical values can also be computed using PEdom1D with sampling from the

uniform distribution of the data sets (i.e. approach (ii)). See the help pages for PE_{dom}1D and TS_{Dom}PE_{Bin}1D for more details. The domination number approach is easily adaptable to testing non-uniform distributions as well (see Remark S5.4 for more detail). PICDs have other applications, e.g. as in [28], we can use the domination number in testing one-dimensional spatial point patterns and our results can help make the power comparisons possible for a large family of distributions (see, e.g. Section 6.2 for a brief treatment of this issue). PICDs can also be employed in pattern classification as well (see, e.g. [12,39]). Furthermore, this article may form the foundation of the generalizations and calculations for uniform and non-uniform distributions in multiple dimensions.

In our calculations, we extensively make use of the ordering of points in \mathbb{R} . The order statistics of \mathcal{Y}_m partition the support of X points into disjoint intervals. This nice structure in \mathbb{R} allows us to find a minimum dominating set and hence the domination number, both in polynomial time. Furthermore, the components of the digraph restricted to intervals (see Section 2.3) are not connected to each other, since the defining proximity regions $N(x_i, r, M) \cap N(x_j, r, M) = \emptyset$ for x_i, x_j in distinct intervals. Extension of this approach to higher dimensions is a challenging problem, since there is no such ordering for point in \mathbb{R}^d with $d > 1$. However, we can use the Delaunay tessellation based on \mathcal{Y}_m to partition the space as in [28]. Furthermore, for most of the article and for all non-trivial results (i.e. for the exact and asymptotic distributions of the domination number), we assumed \mathcal{Y}_m is given; removing the conditioning on \mathcal{Y}_m is a topic of ongoing research along various directions, namely: (i) X and Y both have uniform distribution, (ii) X and Y both have the same (absolutely) continuous distribution, and (iii) X is distributed as F_X and Y is distributed as F_Y (where $F_X \neq F_Y$ and both F_X and F_Y are absolutely continuous).

Acknowledgments

I would like to thank the anonymous referees, whose constructive comments and suggestions greatly improved the presentation and flow of this article. I also would like to thank Prof B. Milošević and Prof E. Zamanzade for providing the R code for their tests upon request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the European Commission under the Marie Curie International Outgoing Fellowship Programme via Project # 329370 titled PRinHDD.

References

- [1] Arlazarov VV, Zhukovsky AE, Krivtsov VE, et al. Using intersection graphs for smartphone-based document localization. *Sci Tech Inform Process.* 2017;44(5):365–372.
- [2] Drachenberg R. Interval digraphs: a generalization of interval graphs [Master's thesis]. Denver (CO): University of Colorado; 1994. p. 80202.
- [3] Quadrini M, Culmone R, Merelli E. Topological Classification of RNA Structures via Intersection Graph. In: Martín-Vide C, Neruda R, Vega-Rodríguez M. editors. *Theory and Practice of Natural Computing, TPNC 2017. Lecture Notes in Computer Science*, vol. 10687. Cham: Springer; 2017. https://link.springer.com/chapter/10.1007/978-3-319-71069-3_16#citeas.
- [4] Roberts FS. *Discrete mathematical models*. Upper Saddle River (NJ): Prentice-Hall; 1976.

- [5] Prisner E. A characterization of interval catch digraphs. *Discrete Math.* **1989**;73:285–289.
- [6] Prisner E. Algorithms for interval catch digraphs. *Discrete Appl Math.* **1994**;51:147–157.
- [7] Cannon A, Cowen L. Approximation algorithms for the class cover problem. Proceedings of the 6th International Symposium on Artificial Intelligence and Mathematics; 2000 Jan 5–7; Fort Lauderdale, Florida; 2000.
- [8] Marchette DM. Random graphs for statistical pattern recognition. Hoboken (NJ): Wiley-Interscience; **2004**.
- [9] Beer E, Fill JA, Janson S, et al. On vertex, edge, and vertex-edge random graphs. *Electron J Comb.* **2011**;18:#P110.
- [10] Priebe C, DeVinney JG, Marchette DJ. On the distribution of the domination number of random class cover catch digraphs. *Stat Probab Lett.* **2001**;55:239–246.
- [11] DeVinney J, Priebe C. A new family of proximity graphs: class cover catch digraphs. *Discrete Appl Math.* **2006**;154(14):1975–1982.
- [12] Manukyan A, Ceyhan E. Classification of imbalanced data with a geometric digraph family. *J Mach Learn Res.* **2016**;17(189):1–40.
- [13] Xiang P, Wierman JC. A CLT for a one-dimensional class cover problem. *Stat Probab Lett.* **2009**;79(2):223–233.
- [14] Ceyhan E. The distribution of the domination number of class cover catch digraphs for non-uniform one-dimensional data. *Discrete Math.* **2008**;308(23):5376–5393.
- [15] Hedetniemi ST, Laskar RC. Bibliography on domination in graphs and some basic definitions of domination parameters. *Discrete Math.* **1990**;86(1-3):257–277.
- [16] Henning MA, Yeo A. Total domination in graphs. New York: Springer; **2013**.
- [17] Hao G. Total domination in digraphs. *Quaest Math.* **2017**;40(3):333–346.
- [18] Lee C. Domination in digraphs. *J Korean Math Soc.* **1998**;4:843–853.
- [19] Niepel L, Knor M. Domination in a digraph and in its reverse. *Discrete Appl Math.* **2009**;157(13):2973–2977.
- [20] L'Ecuyer P. Software for uniform random number generation: distinguishing the good and the bad. In Peters BA, Smith JS, Medeiros DJ, Rohrec MW, editors. Proceedings of the 2001 winter simulation conference, Arlington, VA, USA; 2001.
- [21] Fahidy TZ. On the potential application of uniformity tests in circular statistics to chemical processes. *Int J Chem.* **2013**;5(1):31–38.
- [22] Milošević B. Asymptotic efficiency of goodness-of-fit tests based on too-lin characterization. *Commun Stat-Simul Comput.* **2018**;1–20. doi:10.1080/03610918.2018.1511805
- [23] Chen H, Friedman JH. A new graph-based two-sample test for multivariate and object data. *J Amer Statist Assoc.* **2017**;112(517):397–409.
- [24] Jain AK, Xu X, Ho TK, et al. Uniformity testing using minimal spanning tree. Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02); Vol. 4, 2002. p. 40281.
- [25] Ceyhan E. Density of a random interval catch digraph family and its use for testing uniformity. *REVSTAT.* **2016**;14(4):349–394.
- [26] Ceyhan E. The distribution of the relative arc density of a family of interval catch digraph based on uniform data. *Metrika.* **2012**;75(6):761–793.
- [27] Zamanzade E. Testing uniformity based on new entropy estimators. *J Stat Comput Simul.* **2015**;85(16):3191–3205.
- [28] Ceyhan E, Priebe C. The use of domination number of a random proximity catch digraph for testing spatial patterns of segregation and association. *Stat Probab Lett.* **2005**;73(1):37–50.
- [29] Francis MC, Jacob D, Jana S. Uniquely restricted matchings in interval graphs. *SIAM J Discrete Math.* **2018**;32(1):148–172.
- [30] Sen M, Das S, Roy A, et al. Interval digraphs: an analogue of interval graphs. *J Graph Theory.* **1989**;13:189–202.
- [31] Das AK, Das S, Sen M. Forbidden substructure for interval digraphs/bigraphs. *Discrete Math.* **2016**;339(2):1028–1051.
- [32] Jaromczyk JW, Toussaint GT. Relative neighborhood graphs and their relatives. *Proc IEEE.* **1992**;80:1502–1517.

- [33] Maehara H. A digraph represented by a family of boxes or spheres. *J Graph Theory*. [1984](#);8(3):431–439.
- [34] West DB. *Introduction to graph theory*. 2nd ed. Upper Saddle River: Springer; [2001](#).
- [35] Chartrand G, Harary F, Bill QY. On the out-domination and in-domination numbers of a digraph. *Discrete Math*. [1999](#);197-198:179–183.
- [36] Ceyhan E, Priebe C. On the distribution of the domination number of a new family of parametrized random digraphs. *Model Assist Stat Appl*. [2007](#);1(4):231–255.
- [37] Janson S, Łuczak T, Ruciński A. *Random graphs*. New York: John Wiley & Sons, Inc.; 2000. (Wiley-Interscience series in discrete mathematics and optimization).
- [38] Grzegorzewski P, Wiczorkowski R. Entropy-based goodness of fit test for exponentiality. *Commun Stat Theory Methods*. [1999](#);28:1183–1202.
- [39] Priebe C, Marchette DJ, DeVinney J, et al. Classification using class cover catch digraphs. *J Classif*. [2003](#);20(1):3–23.