Emergence of patterns in random processes. III. Clustering in higher dimensions

William I. Newman[*]
School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey 08540, USA
and Department of Earth & Space Sciences, Department of Physics & Astronomy, and Department of Mathematics,
University of California, Los Angeles, California 90095, USA

Philip Lu[†]
Department of Physics & Astronomy, University of California, Los Angeles, California 90095, USA

(Received 22 March 2019; published 8 July 2019)

Newman *et al.* [Phys. Rev. E **86**, 026103 (2012)] showed that points uniformly distributed as independent and identically distributed random variables with nearest-neighbor interactions form clusters with a mean number of three points in each. Here, we extend our analysis to higher dimensions, ultimately going to infinite dimensions, and we show that the mean number of points per cluster rises monotonically with a limiting value of four.

DOI: 10.1103/PhysRevE.100.012302

I. INTRODUCTION

The identification of patterns from observational data in the natural sciences and other disciplines remains a fundamental challenge. This is particularly true when there is no well-defined quantitative theory available for describing the underlying processes in complex environments. For example, the psychologist Gilovich [1] provides numerous and effective illustrations of the misperception and misinterpretation of random data, including what he calls the "clustering illusion" pertaining to random sequences of events. Psychologists have observed how we tend to "project" patterns onto familiar objects, the classic example being the so-called constellations of the zodiac as noted by the art historian Gombrich (Ref. [2], pp. 105–107). Psychologists have also coined the terms *Apophenia* to describe the experience of seeing patterns or connections in random or meaningless data, as well as *pareidolia* to describe the ability to see shapes or make pictures out of randomness. While the astrological example is almost universally regarded as an illusion, it points to a seeming inner need to associate some form of order to what we see. Is it possible, on the other hand, for random data to present the appearance of pattern?

As an illustration, consider situations in which measurements are made of some quantity that corresponds to a variable that itself is the sum of multiple random variables. Thanks to the central limit theorem [3], for example, the data will appear to be normally (i.e., Gaussian) distributed. However, there is no fundamental "physics" in that claim, and we observe a form of pattern lacking any causal connection. Instead, the central limit theorem is a statement that the observations belong to a class whose randomness by "mathematical necessity" conforms with a well-defined statistical distribution function. Could the impression that data are spatially or temporally clustered be an artifact of mathematical principles, of which we are unaware, and not of any underlying physical process?

Zaliapin *et al.* [4] introduced the concept that in two and presumably higher dimensions, points could provide the sense of forming clusters by virtue of their proximity to each other. In attempting to identify a pattern among earthquake epicenters, they sought to identify for each epicenter in an earthquake catalog the epicenter in the catalog that was "closest," i.e., the nearest-neighbor epicenter. In essence, the authors of Ref. [4] drew an arrow from each epicenter to its nearest neighbor and observed that the directed graph that was produced revealed sets of points that were mutually connected, calling these associations "clusters."

Newman *et al.* [5] considered spatial clustering in one dimension where the points are selected from a uniformly distributed uncorrelated distribution, i.e., they are independent and identically distributed random variables (i.i.d.). In their derivation, they also established a one-to-one correspondence between the behavior of such points and white-noise time series. In so doing, they showed that the spatial distribution of random points, where directed graphs are constructed from each point to its Euclidean nearest neighbor, produces disjoint "clusters" that do not share any points in common and contain an average of three points per cluster. This result is equivalent to white-noise time series producing a pattern of peak-to-peak sequences with, on average, three events per sequence.

In this paper, we return to the problem of spatial clustering and proceed to the more challenging questions that emerge in higher spatial dimension ultimately going to infinite dimension. The geometrical complexity present precludes the use of the kind of combinatoric analysis that Newman *et al.* [5] were able to employ in one dimension. We began by performing computational simulations to obtain insight. We observed there that the mean number of points per cluster increased with the dimension n. This is somewhat intuitive inasmuch as proceeding to a higher dimension introduces the possibility

[*] win@ucla.edu; https://epss.ucla.edu/people/faculty/570/
[†] philiplu11@gmail.com

of new viable linkages arising from the higher-dimensional space.

Nearest-neighbor interactions are ubiquitous in physics, so the implications of clustering could be highly significant, especially in the context of percolation phenomena. From our understanding of percolation processes that are possibly related, Kirkpatrick [6] has shown that there could exist a critical dimensionality where there is a qualitative change in the properties of such systems. Similar considerations could also be relevant in higher-dimensional situations seen in sociology, Milgram's [7] "small-world conjecture," and the famous "six degrees of separation problem" and much of contemporary network analysis [8]. Modern genome-scale genetics interactions are mapped out in an overarching genetic landscape [9]. Clustering phenomena present themselves in fractal geometries in the form of trees [10–12]. Fractal networks and clustering are pervasive in biology [13] as well as in geology [14]. The overarching issue is that high-dimensional clustering is omnipresent and merits investigation. In this paper, we will review briefly the role of directed graphs, how probability theory can be adapted to address such problems in two dimensions, and the emergence of conditional probabilities. We will then proceed to higher and ultimately infinite dimensions and consider the geometrical implications of the clusters that emerge.

II. PROBLEM DESCRIPTION EMPLOYING DIRECTED GRAPHS

As before, we consider a set of uniformly distributed i.i.d. points, and we are most interested in nearest-neighbor interactions between points. We construct *directed graphs* by drawing an arrow (*edge*) from each point (*vertex*) to its nearest-neighbor point. We observe that the ensemble of points could be regarded as a distinct set of *clusters* whose points are mutually connected, and that members of each cluster were distinct or disjoint from other clusters.

We will begin by briefly reviewing some aspects of graph theory and then proceed to a previously unrecognized feature of cluster formation: nearest-neighbor directed-graph structures always contain a *reflexive* pair of points, i.e., two points that are mutual nearest neighbors in whatever dimension space that is being considered. In Fig. 1, we illustrate this situation. We present an ensemble of random points, with no loss of generality, in two dimensions. From each point, we drew an arrow to its nearest neighbor, where we define "nearest" in a Euclidean sense. In this way, we have constructed a "directed graph" [15,16]. In graph theory, what we called points—or events in our illustrative example—are referred to as "vertices" or "nodes," while the arrows are designated as such or as "directed edges." We note, given the visually motivated rules that we established, that "loops" cannot occur. Moreover, we note that all such directed graphs must include one and only one "reflexive" pair of points, i.e., two nodes that point to each other that are mutual nearest neighbors. Each of the clusters present in Fig. 1 has one reflexive pair of points, as expected. Finally, in two or higher dimensions, the directed edges are related to the vertices insofar as they appear to be coming *in* to a vertex or coming *out* from a vertex. The number of arrows coming in to a vertex is referred to as the "indegree," while the number of arrows coming out of a vertex is referred to as the "outdegree" [16]. In our clustering description, the outdegree of all vertices is automatically 1, while the indegree can be 0, 1, 2, ... We now proceed to discuss some of the attendant probabilistic issues.

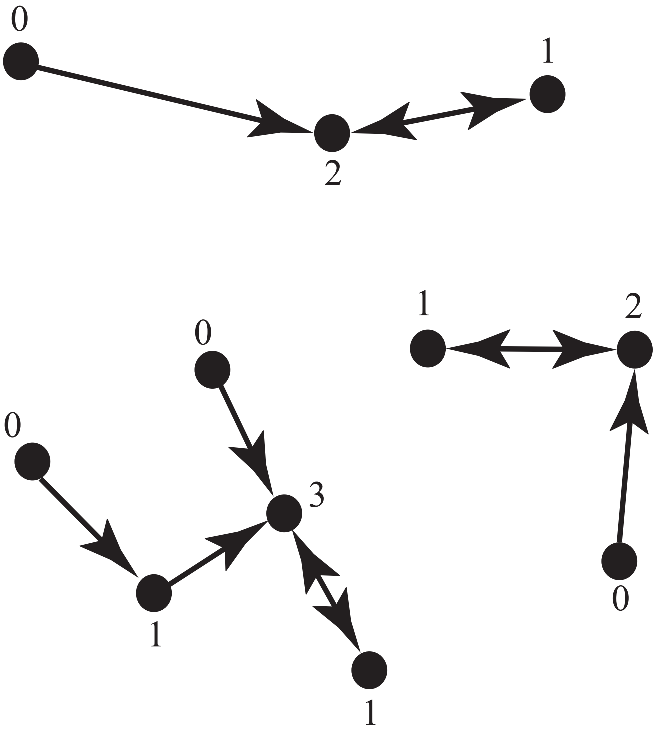

FIG. 1. Random ensemble of points in two dimensions with associated directed graphs and cluster formation. The integers show the "indegree" of each vertex, i.e., the number of other points that identify with it as a nearest neighbor; the "outdegree" of each vertex is, due to our rules of clustering, automatically one.

III. PROBABILISTIC CONSIDERATIONS

Newman *et al.* [5] performed a probabilistic calculation over a doubly infinite set of graphs resulting from a probabilistic treatment of cluster formation, obtaining not only the value of 3 for the mean number of points in a one-dimensional cluster, but also the likelihood of obtaining clusters with two points, three points, four points, and on to an infinite number of points. Geometrical complexity precludes performing that kind of analysis in higher dimension. We note that "loops" cannot occur in nearest-neighbor configurations, as a consequence of the triangle inequality in Euclidean geometry, thereby assuring that all clusters contain a single pair of points that are mutually nearest neighbors. (From another perspective, clusters can be viewed as the confluence of two trees, with the reflexive pair describing the two root nodes; the edges in each of the trees are ordered in length due to the nearest-neighbor progressions that emerge.) We turn our attention now to a probabilistic formulation predicated on the observation that all clusters contain one and only one reflexive pair of points.

The question we wish to formulate is how to calculate the probability that a given pair of points are mutual nearest neighbors. We begin exploring this problem in two dimensions. We will assume that $\mathcal{N}$ is the number density of points per unit area, points that are uniformly distributed i.i.d. random variables in an arbitrarily large domain. We begin by selecting one such point, and we will, without loss of generality, move our coordinate origin to that point, which we shall refer to as 1. We will employ the approach used by Feller [3] in addressing this question, and in a manner that can readily be generalized to arbitrarily high dimension.

Let $P(r)$ identify the probability that the nearest neighbor to point 1 is situated at a Euclidean distance greater than r from it. We note that P describes a cumulative distribution function and that the rate of reduction of P will be proportional to the infinitesimal increment in area multiplied by the number density of points $\mathcal{N}$. In differential form, we write

$$dP(r) = -2\pi r dr \mathcal{N} P(r), \tag{1}$$

where we note that the differential $2\pi r dr$ is the infinitesimal area enclosed in the annulus ranging from r to $r + dr$. Accordingly, the differential equation for the probability assumes the form

$$\frac{dP(r)}{dr} = -2\pi \mathcal{N} r P(r), \tag{2}$$

with the initial condition $P(0) = 1$, yielding the solution

$$P_2(r) = \exp(-\mathcal{N}\pi r^2). \tag{3}$$

We have now introduced a subscript 2 to the cumulative probability function P to identify that this result applies to $n = 2$ dimensions. Importantly, we note that the term πr^2 contained within the exponential is the area contained within a closed curve, namely a circle, whose radius is r. We will return to analogous considerations when we proceed to higher dimension.

IV. CONDITIONAL OR BAYESIAN ANALYSIS

We begin by assuming that the first point under consideration, designated by 1, is situated at the origin. We then ask, what is the differential probability that a second point, designated by 2, is situated in the infinitesimal thin annulus extending r to $r + dr$? It follows that this is $2\pi \mathcal{N} r P_2(r) dr$. For convenience, we will assume that we have rotated the coordinate axes so that point 2 is to the right and horizontal with respect to point 1. In Fig. 2, we illustrate this situation. We have drawn a solid circle at radius r from point 1, calling it A. We have also drawn a dashed line circle with radius r with its origin at 2, calling it B. We observe that, for point 1 to be point 2's nearest neighbor, all other members of the cluster must reside outside circles A *and* B. It follows that we can only consider locations such as the one labeled 3 that are outside *both* circles and not locations such as that labeled 4 that reside outside B but are inside A. Thus, the probability that a third point will not be too close to 2 must be established. It must be conditioned by the requirement that the area in B *not* include that *already* incorporated into circle A. The existence of a condition of this sort is the *sine qua non* of Bayesian probability.

We offer several salient observations evident from this figure. We show five radial lines associated with both circles, with four originating at the center of one circle making a 60° angle with the horizontal and ending at the intersection of that circle with the other circle. The horizontal radial line corresponds to the "edge" that is common to both points 1 and 2. We note that 60° or $\pi/3$ radian angles are ubiquitous—this feature will permeate the extension of clustering to higher dimensions. The two radial lines at 60° associated with point 1 establish a *sector* enclosing an area $\frac{1}{3}\pi r^2$. We can establish a similar sector with respect to point 2, as shown. We can draw a line from the upper intersection point of the two circles (call it a) to the lower intersection point (labeled b). This line is called a *chord*, and it divides each of the sectors into two parts. We refer to the area bounded by the chord and the circular arc lying between the chord's end points as its *segment*. Finally, we observe that the two segments together, sharing a common chord, constitute the region of intersection of the two circles, called for obvious reasons a *lens*. An elementary calculation reveals that the area $\mathcal{A}$ enclosed within the lens is

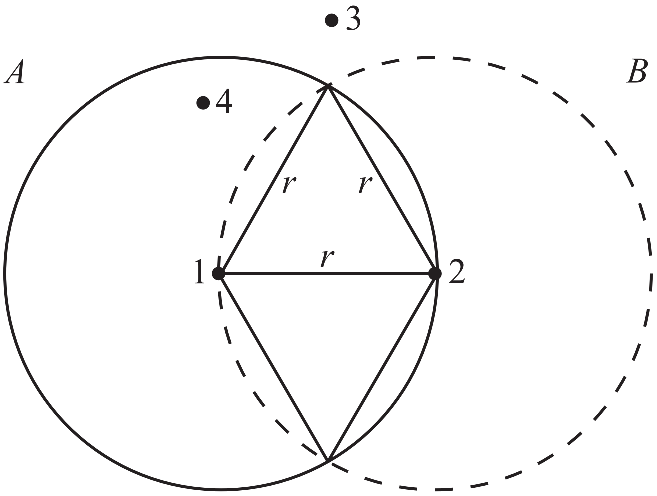

FIG. 2. Geometry associated with Bayesian probability calculation. First point at 1 showing radius r of second point at 2 and the circles A and B plus radial lines relevant to analysis.

$$\mathcal{A} = \left(\frac{2\pi}{3} - \frac{\sqrt{3}}{2}\right) r^2. \tag{4}$$

We will generalize this discussion shortly to accommodate higher-dimensional problems.

We must now calculate the probability that the nearest neighbor to point 2, apart from 1, is at a distance at least as great as r. In principle, we could begin with a differential form for the probability of the sort shown in Eq. (1). However, that equation assumed that the range in azimuthal angle over which the area integral was calculated was 2π. The situation is now complicated because of the probability's conditional nature, and a Bayesian argument is in order. We move our coordinate origin to point 2, and it follows that the radial component of the differential preserves its algebraic form but that the range in the azimuthal angle is now limited by the intersection of circle B with A. In Fig. 3, we elaborate on this situation. The probability $P(r)$ that we calculated in Eq. (3) corresponded to the likelihood that the nearest neighbor was situated outside the area enclosed by circle A,

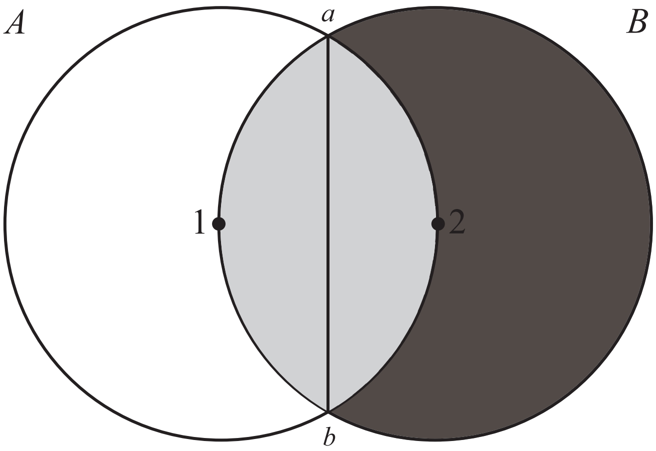

FIG. 3. Venn diagram illustrating areas over which integration is performed. In the first case, it is all the area enclosed by A, namely the component in white and in light gray. In the second case, it is only the area in dark gray corresponding to B less the area in the lens.

the white circle. However, in considering the probability that the nearest neighbor to point 2, apart from point 1, is more distant than r, the area that we must consider is solely the dark gray region shown in the figure, namely the area in B excluding the overlapping lens shown in light gray, namely $B\backslash A$, which reads as the content of B not contained in A. We can now utilize the areal calculation we have performed for the lens, namely Eq. (4), and note the probability $P_C(r)$ that the next nearest neighbor to point 1 and, due to the conditional nature of the problem, point 2 is

$$\begin{aligned} P_C(r) &= \exp\left\{-\mathcal{N}\left[\pi\, r^2 - \left(\frac{2\pi}{3} - \frac{\sqrt{3}}{2}\right) r^2\right]\right\} \\ &= \exp\left\{-\mathcal{N}\left[\left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right) r^2\right]\right\}. \end{aligned} \tag{5}$$

This again is the probability that a third point resides at a distance beyond the first point and excludes the region already established to be unoccupied by points. We evaluate the probability density that a pair of points separated by a distance r are reflexive and integrate over r from $r = 0$ to ∞ to obtain the probability $\mathcal{P}_2$, where the subscript "2" serves to remind us that this is in two dimensions, and that any two points are reflexive,

$$\begin{aligned} \mathcal{P}_2 &= \frac{1}{2}\int_{r=0}^{\infty} \exp\left\{-\mathcal{N}\left[\left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right) r^2\right]\right\} \\ &\quad \times 2\pi\mathcal{N} r P_2(r) dr \\ &= \frac{1}{2}\int_{r=0}^{\infty} \exp\left\{-\mathcal{N}\left[\left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right) r^2\right]\right\} \\ &\quad \times 2\pi\mathcal{N} r \exp\left(-\pi r^2 \mathcal{N}\right) dr \\ &= \frac{3\pi}{8\,\pi + 3\sqrt{3}} \approx 0.310\,752\,448\,5. \end{aligned} \tag{6}$$

We have explicitly reduced our probability $\mathcal{P}_2$ by a factor of 2 in order to avoid "double counting," i.e., beginning with particle 2 and ending with particle 1 in contrast with beginning with particle 1 and ending with particle 2. So, given that this is the probability that two nearest-neighbor points form a reflexive pair, its reciprocal is the mean number of points in each two-dimensional cluster, or

$$\frac{8\,\pi + 3\sqrt{3}}{3\pi} \approx 3.217\,995\,561, \tag{7}$$

confirming our intuition that the average number of points per cluster will modestly exceed the one-dimensional value of 3.

V. PROCEEDING TO HIGHER DIMENSION

The approach we have employed in the previous section was motivated by geometric and probabilistic considerations. All three figures are relevant in higher dimension, when we adjust our interpretation. For example, in three dimensions we emerge with intersecting spheres, segments are referred to as *caps*, and the chord is now replaced by a plane cutting across the two intersecting spheres. The two caps combined figuratively take on the more intuitive and recognizable aspect of a lens. As we proceed to still higher dimensions, we will employ the terms "hyperspheres," "hypercaps," and "hyperlenses," albeit sometimes omitting the prefix "hyper." The probability calculations proceed analogously. In place of an infinitesimally thin annulus in the calculation, we have an infinitesimally thin spherical shell, and the role of area is superseded by volume.

In higher dimension, extensions to surface area and to volume have been established via induction, and they are often a familiar exercise to undergraduate mathematics majors. We will now adapt the meaning of the number density of points $\mathcal{N}$ to represent the mean number of points per unit volume. For example, Dirichlet [17] was possibly the first to do so and obtained the now familiar formula for an n-dimensional volume $V_n(r)$ for an "n-sphere," namely

$$V_n(r) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)}\, r^n. \tag{8}$$

This expression, remarkably, is valid for $n = 1, 2, \ldots,$ which includes one dimension (length) and two dimensions (area) as well. Many other derivations exist—for example, Huber [18] exploited the properties of the Γ function to obtain this result. Wang [19] derived volume relevant to measures of distance using norms other than the Euclidean L_2, such as "diamonds" and "stars." We proceed from here to calculate the n-dimensional analog of $P_n(r)$, where r is the Euclidean radius.

Employing in n dimensions a derivation analogous to that employed in two dimensions, it follows that

$$P_n(r) = \exp\left[-\mathcal{N} V_n(r)\right]. \tag{9}$$

Similarly, higher-dimensional forms of segments—which we now identify as "caps" and denote symbolically by placing a "wide hat" over the variable—are also amenable to induction methods, as shown elegantly by Li [20],

$$\widehat{V_n}(r) = \int_0^{\theta_{\max}} V_{n-1}(r\,\sin\theta)\, dr\cos(\theta), \tag{10}$$

where the angle $\theta_{\max}$ is 60° or $\pi/3$ radians (as noted in Fig. 2). Following Li's derivation, we obtain

$$\widehat{V}_n(r) = \frac{2\pi^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}+1\right)}\, r^n \int_0^{\pi/3} \sin^n\theta\, d\theta, \tag{11}$$

with the volume of the corresponding lens being double this. For $n = 2$, this general formula recovers our explicit result for the lens area in two dimensions shown in Eq. (4).

We proceed now in precisely the same way as we did in deriving Eq. (6), and we write for the general $n \geqslant 2$ case

$$\begin{aligned}\mathcal{P}_n &= \frac{1}{2}\int_{r=0}^{\infty} \exp\{-\mathcal{N}[V_n(r) - 2\,\widehat{V}_n(r)]\}\\ &\quad\times \exp\,[-\mathcal{N}V_n(r)]\mathcal{N}\frac{dV_n(r)}{\mathrm{d}r}dr\\ &= \frac{1}{2}\int_{r=0}^{\infty} \exp\{-2\mathcal{N}[V_n(r) - \widehat{V}_n(r)]\}\\ &\quad\times \mathcal{N}\frac{dV_n(r)}{dr}dr\\ &= \frac{1}{4}\int_{r=0}^{\infty} \exp\left\{-2\mathcal{N}V_n(r)\left[1 - \frac{\widehat{V}_n}{V_n}\right]\right\}\\ &\quad\times 2\mathcal{N}\frac{dV_n(r)}{dr}dr. \end{aligned}\tag{12}$$

Before proceeding, we make two observations. First, the quantity $\widehat{V}_n(r)/V_n(r)$ is a constant that depends upon n but not upon the radial variable r whose dependence has canceled out, and we have consciously omitted showing the no longer relevant r dependence. Hence, we will regard the quantity $1 - \frac{\widehat{V}_n(r)}{V_n(r)}$ as a constant that depends only upon n for $n \geqslant 2$. Second, we make the substitution $x = 2\mathcal{N}V_n(r)$, where x intuitively refers to double the number of points expected to reside within an n-dimensional radius r. Amalgamating these results, the last integral in Eq. (12) becomes

$$\begin{aligned}\mathcal{P}_n &= \frac{1}{4}\int_{r=0}^{\infty} \exp\left\{-x\left[1 - \frac{\widehat{V}_n}{V_n}\right]\right\}dx\\ &= \frac{1}{4\left[1 - \frac{\widehat{V}_n}{V_n}\right]},\end{aligned}\tag{13}$$

where we no longer display the now ignorable dependence of this result upon r. As before, the reciprocal of this expression, which we shall call $\mathbb{N}_n$, is the mean number of points in a cluster in n-dimensional space, namely

$$\mathbb{N}_n = 4\left[1 - \frac{\widehat{V}_n}{V_n}\right]. \tag{14}$$

As further examples, we observe that

$$\mathbb{N}_3 = 3\frac{3}{8} = 3.375 \tag{15}$$

and

$$\begin{aligned}\mathbb{N}_4 &= \frac{20\,\pi - 3\,\sqrt{3}\;{}_2\mathrm{F}_1(1/2, 5/2; 7/2; 3/4)}{5\,\pi}\\ &\approx 3.493\,660\,010.\end{aligned}\tag{16}$$

which incorporated a hypergeometric function representation for the outcome of the integral over $\sin^n\theta$.

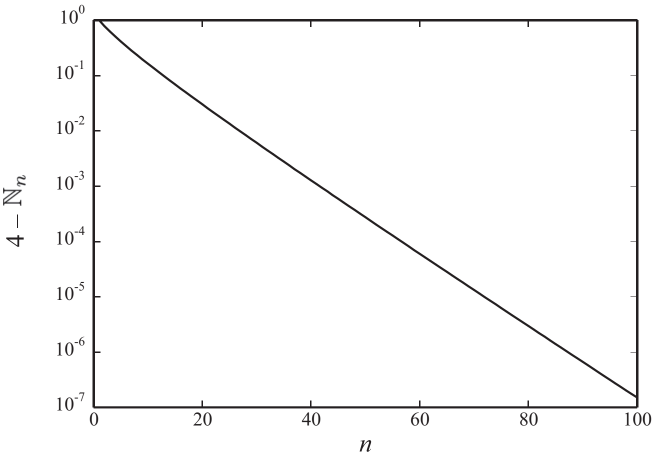

FIG. 4. Demonstration that $\widehat{V}_n/V_n$ converges geometrically to zero.

We confirmed the validity of these values for $\mathbb{N}_n$ for $n = 2, \ldots, 4$ using Monte Carlo simulations with $2^{28} \approx$ 268 435 456 points employing an algorithm we shall describe elsewhere. Our simulation results agreed to four significant figures, which is what you would expect from random-walk arguments based on a sample size of $O(10^8)$. We observed, nevertheless, that the sequence of $\mathbb{N}_n$ appeared to be monotonically increasing with $\widehat{V}_n/V_n$ monotonically decreasing. We verified this trend using computer algebra via MAPLE for $n = 1, \ldots, 100$ and observed that convergence was essentially geometric with a rate of reduction of $\approx\sqrt{3}/2$ each time we increased n by 1. We illustrate this in Fig. 4. We will now demonstrate this analytically for $\mathbb{N}_n$ as n varies from 1 to ∞, thereby demonstrating that in infinite-dimensional Euclidean space the mean cluster size is 4 having grown monotonically from 3 in one dimension.

VI. TO INFINITY... AND BEYOND

Our objective is to evaluate the ratio of $\widehat{V}_n/V_n$, again noting the cancellation of the r-dependence in both quantities, We combine Eqs. (8) and (10) and obtain the ratio

$$\begin{aligned}\frac{\widehat{V}_n}{V_n} &= \frac{\frac{2\pi^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}+1\right)}\int_0^{\pi/3}\sin^n\theta\, d\theta}{\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}}\\ &= \frac{2}{\sqrt{\pi}}\,\frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\left(\frac{n+1}{2}\right)}\int_0^{\pi/3}\sin^n\theta\, d\theta.\end{aligned}\tag{17}$$

Before proceeding with the evaluation of this expression, we will make use of the identity, which can be proven inductively, that

$$\int_0^{\pi/2}\sin^n\theta\, d\theta = \frac{\sqrt{\pi}}{2}\,\frac{\Gamma(1/2 + n/2)}{\Gamma(n/2+1)}, \tag{18}$$

whereupon we find

$$\begin{aligned}\frac{\widehat{V}_n}{V_n} &= \frac{\int_0^{\pi/3}\sin^n\theta\, d\theta}{\int_0^{\pi/2}\sin^n\theta\, d\theta}\\ &= \frac{\int_0^{\pi/3}\sin^n\theta\, d\theta}{\int_0^{\pi/3}\sin^n\theta\, d\theta + \int_{\pi/3}^{\pi/2}\sin^n\theta\, d\theta}.\end{aligned}\tag{19}$$

Using this latter expression, we can see that the ratio is monotonically decreasing by identifying what happens when n is increased by 1. Since the numerator is identical to the first term in the denominator, we observe that multiplying their integrands by an additional power of $\sin\theta$ causes *both* to decrease in precisely the same way. However, the second term in the denominator, whose integrand is also multiplied by $\sin\theta$, is not reduced as much because θ is in the range $(0, \pi/3)$ in the first integral and $(\pi/3, \pi/2)$ in the second where $\sin\theta$ is larger. Therefore, the extent of the diminution due to the $\sin\theta$ term is less significant in the second integral in the denominator. This establishes that the ratio is reduced as n increases.

To obtain explicitly the rate of reduction in the ratio, we return to Eq. (17) and evaluate it in terms of transcendental functions, namely

$$\begin{aligned}\frac{\widehat{V}_n}{V_n} &= \frac{2}{\sqrt{\pi}}\,\frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\left(\frac{n-1}{2}+1\right)}\int_0^{\pi/3}\sin^n\theta\,d\theta\\ &= \frac{1}{\sqrt{\pi}(n+1)}\left(\frac{3}{4}\right)^{\frac{n+1}{2}}\frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}\\ &\quad\times {}_2\mathrm{F}_1\left(\frac{1}{2},\frac{n+1}{2};\frac{3+n}{2};\frac{3}{4}\right).\end{aligned}\tag{20}$$

We can now dissect each of the terms as a function of n. Importantly, we note the appearance of the term 3/4, which is the sine of 60° or $\pi/3$ radians. The hypergeometric term is relatively constant, albeit monotone, with respect to n, and a bound can readily be obtained for it. The ratio of the two Γ functions, using Stirling's approximation, varies as $\sqrt{n}$ and is more than compensated by the $1/(n+1)$ term at the beginning of the equation, rendering their combined influence a slowly decreasing $1/\sqrt{n}$. Finally, we observe that the dominant term in the expression for the ratio is the term in $(\sqrt{3}/2)^n$, which conforms with our observation obtained by computer algebra. Hence, $\mathbb{N}_n$ converges to 4 as $n \to \infty$. We have completed our search for clustering in randomness, proceeding from one-dimensional i.i.d. uniformly distributed random points to infinite-dimensional space.

Before concluding this section, it is helpful to revisit the geometrical nature of the clusters that form as the dimensionality of the Euclidean space approaches infinity. It is quite remarkable that the mean number of points in the cluster is on average 4. This implies that clusters with relatively few points present low-dimensional structures. We illustrate this in the following figure. In the minimal case of two points in a cluster, they automatically form a reflexive pair and are intrinsically one-dimensional. In the case of three points in a cluster, two topologically equivalent structures occur including the one that we display as well as its mirror image. Importantly, the nonreflexive vertex and the edge associated with it can be at an angle, but that angle is restricted by the requirement that its vertex remains further away from the other (rightmost) vertex in the reflexive pair. Accordingly, three-point clusters are intrinsically two-dimensional. Finally, in the case of four points in a cluster, we present three subcases, labeled (a), (b), and (c), as well as their mirror-image equivalents. This situation presents a greater degree of geometric complexity inasmuch as the structures formed are fundamentally three-dimensional. It is conceptually simple but mathematically complex to construct clusters with five or more constituent points, which have the potential to generate four-dimensional objects. In contrast with what Newman *et al.* [5] were able to show in one dimension, the algebraic complexity of the available configurations of clusters that can form in two and higher dimensions makes it impossible to analytically calculate the probability of obtaining all of the available configurations such as the three that we show in Fig. 5. The geometrical implication of our calculation of the mean number of points in a cluster and their attendant structural character is clear: clusters formed via nearest-neighbor interactions are both small in number and are typically three-dimensional in character despite being derived from an infinite-dimensional space.

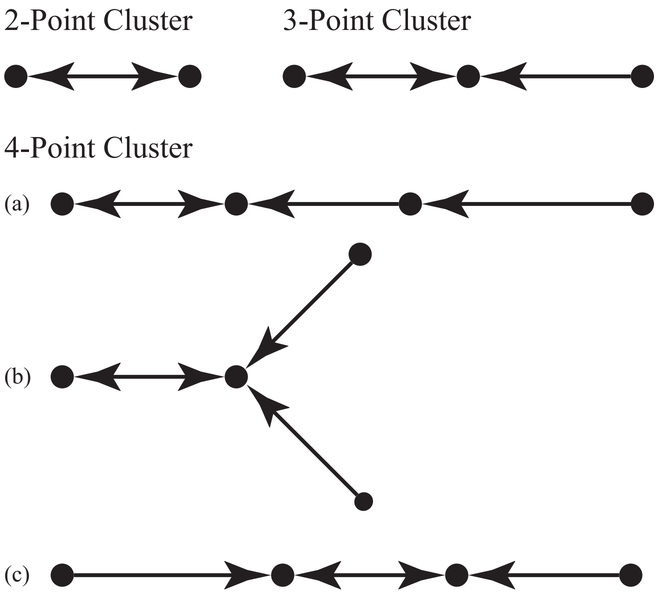

FIG. 5. Geometric configurations of clusters.

VII. CONCLUSIONS AND DISCUSSION

This investigation was stimulated by considerations of randomness that occurs in nature, and, in some sense, it can provide illusory evidence of pattern or structure. The search for spatial clustering based upon nearest-neighbor associations, for example in application to earthquake events, resulted in our desire to explore the *null hypothesis*: can uniformly distributed randomly distributed points that pass tests for being independently and identically distributed appear to project a pattern. Indeed, manifestations of clustering in high-dimensional data that do not conform with these i.i.d. results would be a tell-tale sign of some nonrandom link among points.

In earlier work, Newman *et al.* [5] proved in one spatial dimension that the mean number of points in a cluster would be 3. Furthermore, they showed that this result in one dimension is formally equivalent to peak-to-peak statistics in time series, apparently answering the time-honored maxim, "why do good things come in threes"—three, evidently, is an almost magic number that emerges from random processes and bears no additional meaning. Intrigued by the study of Zaliapin *et al.* [4] involving earthquake epicenter clustering, we investigated clustering in two and higher dimensions. We developed a simple Bayesian probability derivation for the mean number of points in a cluster $\mathbb{N}$ as a function of the dimension n, and we obtained steadily increasing values beginning in two dimensions of 3.218, 3.375, 3.494, etc. and speculated on the question of whether there could be an upper bound, for example 4. As a pattern evolved in the probabilistic calculations, we discovered—as we have shown here—that this is indeed the case, and that an infinite-dimensional space populated with i.i.d.uniformly distributed points with nearest-neighbor connectivity will establish clusters with a mean number of four points per cluster. Finally, these relatively small clusters establish low-dimensional structures, e.g., a four-point cluster conforms with a three-dimensional object.

ACKNOWLEDGMENTS

We wish to thank Todd Thompson and Gary Williams for helpful discussions and, especially, Dominic Reiss for significant computational insights into simulations of this problem. W.I.N. thanks the Institute for Advanced Study in Princeton for its hospitality during his sabbatical year away from the University of California, Los Angeles, and for funding as IBM Einstein Foundation Fellow at the Institute.

[1] T. Gilovich, *How We Know What Isn't So* (Simon and Schuster, New York, 2008).

[2] E. H. Gombrich, *Art and Illusion: A Study in the Psychology of Pictorial Representation* (Phaidon, London, 1977), Vol. 5.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 2008), Vol. 2.

[4] I. Zaliapin, A. Gabrielov, V. Keilis-Borok, and H. Wong, Phys. Rev. Lett. **101**, 018501 (2008).

[5] W. I. Newman, D. L. Turcotte, and B. D. Malamud, Phys. Rev. E **86**, 026103 (2012).

[6] S. Kirkpatrick, Phys. Rev. Lett. **36**, 69 (1976).

[7] S. Milgram, Psych. Today **2**, 60 (1967).

[8] M. Newman, *Networks* (Oxford University Press, Oxford, 2018).

[9] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi *et al.*, Science **327**, 425 (2010).

[10] W. Newman, D. Turcotte, and A. Gabrielov, Fractals **05**, 603 (1997).

[11] A. Gabrielov, W. I. Newman, and D. L. Turcotte, Phys. Rev. E **60**, 5293 (1999).

[12] G. Yakovlev, W. I. Newman, D. L. Turcotte, and A. Gabrielov, Geophys. J. Int. **163**, 433 (2005).

[13] D. Turcotte, J. Pelletier, and W. Newman, J. Theor. Biol. **193**, 577 (1998).

[14] D. L. Turcotte and W. I. Newman, Proc. Natl. Acad. Sci. (USA) **93**, 14295 (1996).

[15] B. Bollobas, *Graph Theory: An Introductory Course* (Springer Science & Business Media, New York, 2012).

[16] G. Chartrand, *Introductory Graph Theory* (Courier, New York, 1977).

[17] P. L. Dirichlet, J. Math. Pures Appl. **4**, 164 (1839).

[18] G. Huber, Am. Math. Mon. **89**, 301 (1982).

[19] X. Wang, Math. Mag. **78**, 390 (2005).

[20] S. Li, Asian J. Math. Statist. **4**, 66 (2011).