Pair correlation functions for identifying spatial correlation in discrete domains

Enrico Gavagnin,[*,†] Jennifer P. Owen,[†,‡] and Christian A. Yates
Centre for Mathematical Biology, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom

(Received 12 January 2018; published 4 June 2018)

Identifying and quantifying spatial correlation are important aspects of studying the collective behavior of multiagent systems. Pair correlation functions (PCFs) are powerful statistical tools that can provide qualitative and quantitative information about correlation between pairs of agents. Despite the numerous PCFs defined for off-lattice domains, only a few recent studies have considered a PCF for discrete domains. Our work extends the study of spatial correlation in discrete domains by defining a new set of PCFs using two natural and intuitive definitions of distance for a square lattice: the taxicab and uniform metric. We show how these PCFs improve upon previous attempts and compare between the quantitative data acquired. We also extend our definitions of the PCF to other types of regular tessellation that have not been studied before, including hexagonal, triangular, and cuboidal. Finally, we provide a comprehensive PCF for any tessellation and metric, allowing investigation of spatial correlation in irregular lattices for which recognizing correlation is less intuitive.

DOI: 10.1103/PhysRevE.97.062104

I. INTRODUCTION

A system of agents is considered in a state of spatial correlation if, given any agent in the system, the likelihood that there are other agents at a certain, close distance, is either increased or decreased with respect to the situation in which the agents are distributed uniformly at random. Spatial correlation is a dominant feature of many biological and physical systems [1–12]. For example, in cell biology, spatial correlation can be seen in the form of patterns on animal fur or fish skin [1,3,13]. In a clinical setting, cell aggregation is a characteristic feature of melanoma and its identification is essential for early diagnosis and effective therapy [14,15]. Resource competition in ecology can lead to spatial correlation in the form of segregation, for example, in ant nest displacement in a competitive environment [16]. In epidemiology, spatial correlation can be observed in the occurrence of disease across different geographical regions [17].

The same spatial configuration can have different origins. For example, spatial aggregation in cell biology can be caused by a result of cell-to-cell adhesion [8], external signals, as in *chemotaxis* [5,6], or even slime following [7]. Alternatively, cells may form clusters during development due to a combination of a high proliferation rate and a low movement rate [9]. Given a system exhibiting spatial correlation, one may hypothesize an underlying mechanism responsible for these properties. These assumptions may form the basis of a mathematical model that can be simulated for the purpose of testing. Quantifying spatial correlation in both the simulation and observed experimental data can be a way to connect these studies and to validate or disprove such a theory. As a result, a great number of statistical tools have been developed in the past decade to analyze and measure spatial correlation [12,18–23]. Among the most popular are pair correlation functions (PCFs) [11,12,20–22,24–28] and the fast Fourier transform (FFT) [29,30]. In this paper we focus our attention on the study of spatial correlation using PCFs.

Given a system of agents, a PCF determines whether pairs of agents are more or less likely to be found with a given separation than in the situation in which the agents are positioned uniformly at random in the domain. A PCF is considered effective if it fits two main criteria. First, the PCF distance metric should be well-defined, but most importantly be readily interpretable in the context of the system considered. This criteria is essential so that in the case of correlation (aggregation or segregation), the PCF can be used to obtain more details about the spatial configuration. For example, if the system exhibits aggregation, the PCF should be able to provide a measure for the average size of the clusters and their pairwise separation. Second, the PCF should be correctly calibrated. The PCF should be able to distinguish between three basic types of configurations: spatial randomness, aggregation, and segregation. For this, the PCF should be normalized correctly; i.e., the PCF should return the value unity at all pairwise distances (no correlation) when applied to a uniformly distributed set of agents. If the PCF is not normalized correctly, a spatially random set of agents may be incorrectly identified as a correlated system. This inconsistency makes PCF profiles hard to interpret.

Depending on the type of investigation, the mathematical framework can either be continuous (off-lattice) or discrete (on-lattice). The corresponding PCF has to be defined in accordance with the given framework. Despite the abundance of PCFs defined for off-lattice domains [11,26–28], only a few recent studies have defined a PCF for domains partitioned with a lattice [12,20–22]. On-lattice PCFs often assume exclusion properties, that is, that each lattice site in a domain can be occupied by at most one agent at any given time. This is consistent with typical on-lattice correlation studies, such as

[*]e.gavagnin@bath.ac.uk
[†]These authors contributed equally to this work.
[‡]j.owen@bath.ac.uk

those designed to quantify correlation in binary pixelated images, or to determine spatial correlation in exclusion processes simulated using a discrete domain.

Currently, there are two PCFs defined on-lattice. The first is a naive approach consisting of applying the classic off-lattice PCF to lattice-based systems. We refer to this from now on as the *annular* PCF. In the annular PCF, given some small positive δ, the number of agents at a distance m from a focal agent is defined as the number of agents whose centres lie in the annulus $(m-\delta, m]$, where distance is defined by the Euclidean metric. A limitation of this method is that, while the normalization is a good approximation for a continuous domain (see Sec. II for more details), it is poor in the case of a discrete domain, thus the PCF is not correctly calibrated. In more recent work, Binder and Simpson [12,21] defined a PCF specifically designed for a two-dimensional square on-lattice exclusion process, which we will refer to as the *rectilinear* PCF (see Sec. II for more details). While their approach correctly identifies the spatial correlation in many examples, due to an anisotropic definition of distance, spatial structures that are biased in either Cartesian directions can remain unidentified by this PCF. To summarize, to the best of our knowledge, a discrete isotropic PCF with correct normalization does not currently exist in the literature.

In this paper, we extend the study of pairwise spatial correlation for on-lattice exclusion processes which tackles the flaws of previous PCFs. We define isotropic PCFs for a square lattice on which distance is defined using two of the most natural and intuitive metrics for a discrete domain: the taxicab and uniform metric. We call these the *square taxicab* PCF and *square uniform* PCF, respectively, after the square lattice set up and metric type. We define them in both the nonperiodic and periodic boundary cases. Using synthetically generated data, we demonstrate that our PCF can correctly distinguish between spatial randomness, aggregation, and segregation. Furthermore, we show that it can also provide quantitative information about the structure of the system, such as approximate aggregate size or segregation distance both in the short and long scales. Moreover, we investigate how the choice of metric, uniform, or taxicab can affect this quantitative information. We demonstrate that our PCFs represent a significant improvement on previous on-lattice PCFs by showing that, first, our method is correctly calibrated (unlike the annular PCF) and, second, that it can identify anisotropic patterns of the type that are routinely missed by the rectilinear PCF. As a natural extension, we define PCFs for higher dimensions and other types of tessellations (cubic, triangular, and hexagonal) that have not been considered previously. We name these the *triangle* PCF, the *hexagon* PCF, the *cube taxicab* PCF, and the *cube uniform* PCF after the lattice set up and metric type.

Finally, we extend the concept of a PCF by introducing the *general* PCF. This PCF can be defined using any metric on any discrete domain type, with the caveat that it is more computationally expensive. We give an example of how we can use this PCF on a discrete irregular lattice (both tessellation and domain shape), where we define adjacent sites to be at unit distance from one another. We show how our PCF can identify aggregation and segregation on an irregular domain using some synthetic examples.

The paper is organized as follows. In Sec. II we discuss the successes and limitations of previous on-lattice PCFs. In Sec. III we introduce our square taxicab and square uniform PCFs. We apply our square taxicab and square uniform PCFs to some relevant examples and make comparisons with previous on-lattice PCFs from the literature in Sec. IV. In Sec. V A we define the triangle PCF, the hexagon PCF, the cube taxicab PCF, and the cube uniform PCF. We extend our PCF to more generic and possibly irregular lattices by defining the general PCF in Sec. VI. For reference, in Sec. S.4 of the Supplemental Material, we supply a table summarizing all the formulae for the normalizations of our PCFs [31]. Finally, we conclude in Sec. VII by summarising the relevance of our results and discussing potential avenues for future work.

II. EXISTING ON-LATTICE PAIR CORRELATION FUNCTIONS

In this section we provide a summary of the only two existing PCFs defined for discrete domains: the annular PCF and the rectilinear PCF. For each, we describe their strengths and limitations.

First, consider a system of agents on a two-dimensional square lattice of size $L_x \times L_y$, with lattice step Δ, and with the exclusion property that, at any given time, each lattice site can be occupied by at most one agent. If N agents occupy the domain, then the occupancy of the lattice can be represented by a matrix M:

$$M_{xy} = \begin{cases} 0 & \text{if lattice site } (x,y) \text{ is vacant,} \\ 1 & \text{if lattice site } (x,y) \text{ is occupied,} \end{cases} \tag{1}$$

where

$$N = \sum_{x=1}^{L_x} \sum_{y=1}^{L_y} M_{xy} \leqslant L_x L_y. \tag{2}$$

Let ψ^M be the set of all agent pairs in the lattice defined by matrix M, i.e.,

$$\psi^M = \big\{(\boldsymbol{a},\boldsymbol{b}) \in \mathbb{L} \times \mathbb{L} | \boldsymbol{a} = (x_a, y_a), \\ \boldsymbol{b} = (x_b, y_b),\ a \neq b,\ M_{x_a,y_a} = M_{x_b,y_b} = 1\big\}, \tag{3}$$

where $\mathbb{L} = \{1,\dots,L_x\} \times \{1,\dots,L_y\}$ is the set of all sites in the lattice. With agents in configuration M, let us define the subset of agent pairs separated by distance m according to some (as yet unspecified) definition of distance, denoted by d, as

$$C_d(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M | \|\boldsymbol{a}-\boldsymbol{b}\|_d = m\}, \text{ for } m \in \mathcal{D}_d, \tag{4}$$

where $\mathcal{D}_d$ is the set of possible distances under the metric d. We define the total number of pairs of agents for each value of distance $m \in \mathcal{D}_d$ as

$$c_d(m) = |C_d(m)|. \tag{5}$$

Similarly, we define the set of pairs of sites (regardless of their occupancy) which are separated by distance m according to the metric d as

$$S_d(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \mathbb{L} \times \mathbb{L} \big| \|\boldsymbol{a}-\boldsymbol{b}\|_d = m\}, \text{ for } m \in \mathcal{D}_d, \tag{6}$$

hence the total number of pairs of sites at distance m is given by

$$s_d(m) = |S_d(m)|. \tag{7}$$

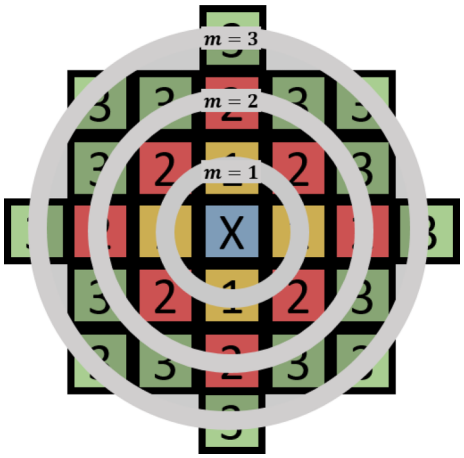


FIG. 1. Example agent pairs in the sets $C_A(m)$ with $m = 1,2,3$ and bandwidth $\delta = 1$. Concentric annuli, $(m-\delta,m]$, are superposed on top of the square lattice and the sites whose centres fall into each annulus are colored differently. Sites in yellow, red, and green (labeled 1, 2, and 3, respectively) are defined to be distance one, two, and three from the blue site, respectively, labeled X.

To produce a PCF, we aim to normalize the index $c_d(m)$ with the number of pairs we would expect at distance m if the system had no spatial correlation. That is, we consider the case in which the same number of agents are displaced uniformly at random on the lattice and compute the expected number of pairs at distance m. Let U be a random matrix such that $U_{xy} = 0$ for all sites $(x,y) \in \mathbb{L}$, apart from N sites chosen uniformly at random without replacement, for which $U_{xy} = 1$. Then we define

$$\bar{C}_d(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^U \,|\, \|\boldsymbol{a}-\boldsymbol{b}\|_d = m\}, \quad \text{for} \quad m \in \mathcal{D}_d. \tag{8}$$

Hence, for each value of $m \in \mathcal{D}_d$, the PCF at distance m is defined as

$$f_d(m) := \frac{c_d(m)}{\mathbb{E}[\bar{c}_d(m)]}, \tag{9}$$

where $\bar{c}_d(m) = |\bar{C}_d(m)|$ and $\mathbb{E}$ represents the expectation operator.

A. The annular PCF

The annular PCF was originally designed for two-dimensional off-lattice systems [19] but can be extended to on-lattice systems with periodic boundary conditions (BC). For the annular PCF, the set C_A, where A denotes the annular metric, is defined as follows:

$$C_A(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \sqrt{(x_a-x_b)^2+(y_a-y_b)^2} \in (m-\delta,m]\}, \tag{10}$$

where $m \in \{\delta k \,|\, k \in \mathbb{N}^+\}$ and δ is a small real number which determines the bandwidth of the PCF. The schematics in Fig. 1 show a representation of some elements in the sets $C_A(1)$, $C_A(2)$, and $C_A(3)$ with $\delta = 1$.

The normalization factor is given by

$$\mathbb{E}[\bar{c}_A(m)] \approx \frac{N(N-1)(2\pi m\delta)}{L_x L_y}, \tag{11}$$

where $2\pi m\delta$ approximates the area of the mth annulus (assuming small δ) from any given agent. The annular PCF, f_A, follows from definition Eq. (9).

The normalization in Eq. (11) is a good approximation for a continuous domain with a small δ. However, when the agents are positioned on a lattice, this approach is no longer appropriate. The main issue is that the counts of agents in each annulus vary in an unpredictable manner with the distance, m, and the annular width, δ. For example, consider a square lattice with spacing Δ. The only possible distances two agents can be separated by are in the countable set

$$\begin{aligned}\mathcal{D}_A &= \{\Delta\sqrt{x^2+y^2} \,|\, (x,y) \in \mathbb{N}^2 \backslash \{0,0\}\} \\ &= \{\Delta, \sqrt{2}\Delta, 2\Delta, \sqrt{5}\Delta, \dots\}.\end{aligned} \tag{12}$$

Partitioning these distances into regularly spaced intervals, as it is required by the Euclidean distance metric, we can see that the number of agent pairs does not increase smoothly with the distance, m. Depending on the value of δ it may not even increase monotonically. However, the definition of the normalization factor Eq. (11) suggests that the expected number of pairs increases smoothly and monotonically with both m and δ. This disparity means the on-lattice annular PCF will not be properly normalized and will either be an over- or under- approximation, making results hard to interpret (see Fig. 7(b) as an example).

B. The Rectilinear PCF

In more recent work, Binder and Simpson [12,21] define the Rectilinear PCF specifically for two-dimensional, on-lattice exclusion processes with nonperiodic BC. Their definition is easily extendible to periodic BC. They define two PCFs for the two Cartesian directions. In each case the distance is defined by the number of columns (or rows) separating two agents.

Thus, the set of pairs of agents separated by integer distance $m \in \mathbb{N}^+$ are defined in the x direction and y direction respectively as

$$C_{R_x}(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, |x_a - x_b| = m\}, \tag{13a}$$

$$C^M_{R_y}(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, |y_a - y_b| = m\}, \tag{13b}$$

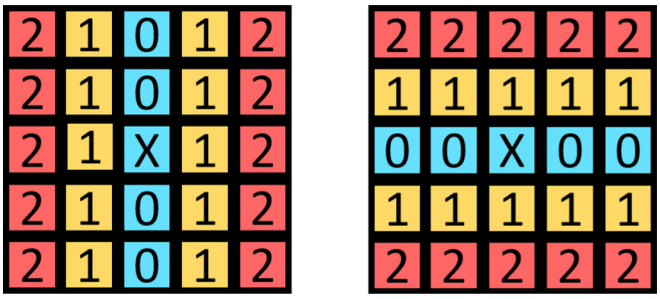
where subscripts R_x, R_y refer to the metrics defined by the Rectilinear PCFs. The schematics in Fig. 2 represent examples of sites separated by distances $m = 0$, $m = 1$ and $m = 2$ for metrics R_x and R_y.

The counts are then normalized by the expected number of pairs of agents at distance m assuming N uniformly distributed agents:

$$\mathbb{E}\big[\bar{c}_{R_x}(m)\big] = \frac{N-1}{L_x L_y - 1}\frac{N}{L_x L_y} L_y^2 (L_x - m), \tag{14a}$$

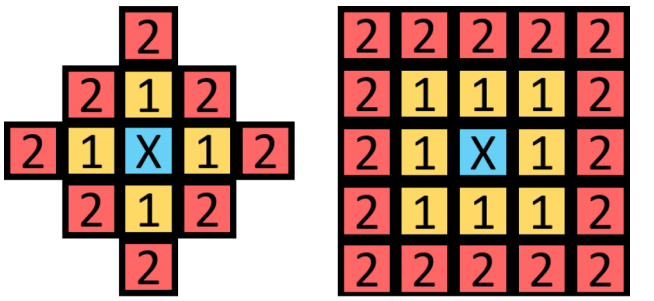
$$\mathbb{E}\big[\bar{c}_{R_y}(m)\big] = \frac{N-1}{L_x L_y - 1}\frac{N}{L_x L_y} L_x^2 (L_y - m), \tag{14b}$$

respectively. For details of the derivation of these factors, see Ref. [12]. The final rectilinear PCF is defined as the arithmetic



(a) (b)

FIG. 2. Schematic of agent pairs using (a) the R_x metric, (b) the R_y metric. Sites in yellow and red (labelled 1, 2 respectively) are defined to be distance-one and -two neighbors, respectively, from the blue site labeled X.



(a) (b)

FIG. 3. Schematic of agent pairs using (a) the taxicab metric (b) the uniform metric. Sites in yellow and red, labeled with numbers 1 and 2, respectively, are defined to be distance-one and -two neighbors from the site marked in blue (labeled with X).

average of the two orthogonal PCFs, i.e.,

$$f_R^M(m) = \frac{1}{2}\big[f_{R_x}(m) + f_{R_y}(m)\big],$$

where

$$f_{R_x}^M(m) = \frac{c_{R_x}}{\mathbb{E}\big[\bar{c}_{R_x}(m)\big]},$$
$$f_{R_y}^M = \frac{c_{R_y}}{\mathbb{E}\big[\bar{c}_{R_y}(m)\big]}. \tag{15}$$

The rectilinear PCF correctly identifies spatial correlation in many examples and, unlike the annular PCF, is normalized correctly. However, one major issue that the rectilinear PCF suffers from is that, due to the inherent anisotropy of its definition, spatial structures that are biased in either Cartesian direction may be missed. For such patterns, the PCF given by the R_x and R_y metrics are approximately constant functions of distance, because the averaged row and column densities are constant along the axes, despite the fact that clustering can still be present. Examples of these spatial patterns include many biologically and chemically relevant cases, such as diagonal stripes and chessboard patterns [32,33] (see Fig. 8). We note that when the pattern structure is biased in only one Cartesian direction, the preaveraging rectilinear PCFs f_{R_x} and f_{R_y} will identify further information about the direction of the spatial pattern. Another limitation of the rectilinear PCF is that it applies only to regular square lattices and a generalisation to other forms of tessellations would be challenging.

III. THE SQUARE TAXICAB AND SQUARE UNIFORM PCFS

In this section we define two discrete PCFs for a square lattice: the square taxicab and square uniform PCF, using the taxicab and uniform metric, respectively, under both periodic and nonperiodic BC. Using the same notation as in Sec. II, we define the subsets of agent pairs separated by distance m under nonperiodic BC as

$$C_1^n(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \|\boldsymbol{a}-\boldsymbol{b}\|_1 = m\}, \quad m \in \mathcal{D}_1^n, \tag{16a}$$
$$C_\infty^n(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \|\boldsymbol{a}-\boldsymbol{b}\|_\infty = m\}, \quad m \in \mathcal{D}_\infty^n, \tag{16b}$$

for the taxicab and uniform metric, respectively. Here, $\mathcal{D}_1^n = \mathcal{D}_\infty^n = \{1,2,\ldots \min\{L_x,L_y\}-1\}$ and the superscript n refers to the fact we are considering *nonperiodic* BC. Using the definitions of the uniform and taxicab metrics, we can express these sets as

$$C_1^n(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, |x_a - x_b| + |y_a - y_b| = m\}, \tag{17a}$$
$$C_\infty^n(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \max\{|x_a - x_b|, |y_a - y_b|\} = m\}. \tag{17b}$$

Similarly, we define the subsets of agent pairs separated by distance m under periodic BC as

$$C_1^p(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \min\{|x_a - x_b|, L_x - |x_a - x_b|\} + \min\{(|y_a - y_b|, L_y - |y_a - y_b|\} = m\}, \quad m \in \mathcal{D}_1^p, \tag{18a}$$
$$C_\infty^p(m) = \{(\boldsymbol{a},\boldsymbol{b}) \in \psi^M \,|\, \max\{\min\{|x_a - x_b|, L_x - |x_a - x_b|\}, + \min\{(|y_a - y_b|, L_y - |y_a - y_b|\}\} = m\}, \quad m \in \mathcal{D}_\infty^p, \tag{18b}$$

where $\mathcal{D}_1^p = \mathcal{D}_\infty^p = \{1,2,\ldots \min\{\lfloor \frac{L_x}{2} \rfloor, \lfloor \frac{L_y}{2} \rfloor\}\}$. The corresponding definitions of S_1^n, S_∞^n, S_1^p, and S_∞^p can be obtained similarly by using Eq. (6). Here the superscript p refers to the fact we are considering *periodic* BC. Notice that we restrict the largest m to be $\min\{\lfloor \frac{L_x}{2} \rfloor, \lfloor \frac{L_y}{2} \rfloor\}$ to simplify the computation of the normalization factor (see Sec. III A). However, with some work this restriction could be relaxed. The schematics in Fig. 3 represent examples of sites separated by distance $m = 1$ and $m = 2$ using the taxicab (a) and uniform (b) metrics.

For the normalization factors, Binder and Simpson [12] use

$$\mathbb{E}[\bar{c}_d(m)] = \left(\frac{N}{L_x L_y}\right)\left(\frac{N-1}{L_x L_y - 1}\right) s_d(m), \tag{19}$$

where $s_d(m)$ is defined as in Eq. (7) and d refers to the metric used. In other words, the expected number of pairs of agents at distance m on a lattice with N uniformly distributed agents can

be written as the probability that two different sites at distance m are simultaneously occupied, multiplied by the total number of pairs of sites at distance m in the domain.

To complete the definitions of our square lattice PCFs we need to provide an expression for s_1 and s_∞. We address this in the next two sections, distinguishing between the cases of periodic and nonperiodic BC.

A. Normalization of the square taxicab and square uniform PCF under periodic boundary conditions

As the derivation for the normalization is simple under periodic BC and more complicated under nonperiodic BC, we first consider the system with periodic BC and determine s_1^p, s_∞^p, where p denotes periodic BC. Let us define the number of sites separated by a distance m from any given reference site on a lattice as $t_1(m)$, $t_\infty(m)$ under the taxicab and uniform metric, respectively. These read

$$t_1(m) = 4m, \tag{20a}$$

$$t_\infty(m) = 8m. \tag{20b}$$

The proofs of Eqs. (20) are omitted, but they can be obtained easily by induction on m. Examples for $m = 1,2$ can be seen in Fig. 3. Notice that for $m \leqslant \min\{\lfloor \frac{L_x}{2} \rfloor, \lfloor \frac{L_y}{2} \rfloor\}$, given any site on the lattice, the number of sites at distance m from this reference site in the case of periodic BC is exactly $t(m)$. Consider the lattice of size $L_x \times L_y$ with $L_x, L_y > 2$. If we multiply the total number of lattice sites by $t(m)$, we count each pair of sites separated by distance m exactly twice. Hence, we conclude that

$$s^p_{d=1,\infty}(m) = \frac{t(m)L_xL_y}{2}, \tag{21}$$

using the taxicab and uniform metrics. Substituting values for $t_1(m)$ and $t_\infty(m)$ from Eqs. (20) we deduce that

$$s_1^p(m) = 2mL_xL_y, \tag{22a}$$

$$s_\infty^p(m) = 4mL_xL_y. \tag{22b}$$

Therefore, by substituting Eqs. (22) into Eq. (19), the normalization factors under periodic BC are

$$\mathbb{E}\big[\bar{c}_1^p(m)\big] = \frac{2mN(N-1)}{L_xL_y - 1}, \tag{23a}$$

$$\mathbb{E}\big[\bar{c}_\infty^p(m)\big] = \frac{4mN(N-1)}{L_xL_y - 1}. \tag{23b}$$

B. Normalization of the square taxicab and square uniform PCF under nonperiodic BC

In this section we derive expressions for $s_1^n(m)$ and $s_\infty^n(m)$, where n denotes nonperiodic BC. Notice that, for all $m \in \mathcal{D}$, we have that $s^p(m) > s^n(m)$ since $s^p(m)$ includes pairs that cross the domain boundary, whereas $s^n(m)$ does not. Therefore, to find a formula for $s^n(m)$, it is enough to determine a formula for the remainders defined by

$$r_1(m) = s_1^p(m) - s_1^n(m), \tag{24a}$$

$$r_\infty(m) = s_\infty^p(m) - s_\infty^n(m). \tag{24b}$$

The remainders count the number of pairs of sites that cross a boundary under the periodic BC. For simplicity, throughout this section, we will only derive the normalization for the taxicab metric; however, the derivation for the uniform metric is similar and can be found in Supplemental Material Sec. S.1 for reference [31]. Let us define the set of pairs of sites separated by distance $m \in \mathcal{D}_1^n$ that cross the x boundary (horizontal axis) or y boundary (vertical axis), respectively, as

$$P_1^x(m) = \big\{(\boldsymbol{a},\boldsymbol{b}) \in S_1^p(m) \big| |y_a - y_b| > L_y - |y_a - y_b|\big\}, \tag{25a}$$

$$P_1^y(m) = \big\{(\boldsymbol{a},\boldsymbol{b}) \in S_1^p(m) \big| |x_a - x_b| > L_x - |x_a - x_b|\big\}. \tag{25b}$$

Of these pairs, let us consider those pairs that are at distance $k \in \{1,\ldots m\}$ rows or columns from each other, respectively. We define these subsets as

$$P_1^x(m,k) = \big\{(\boldsymbol{a},\boldsymbol{b}) \in P_1^x(m) \big| L_y - |y_a - y_b| = k\big\}, \tag{26a}$$

$$P_1^y(m,k) = \big\{(\boldsymbol{a},\boldsymbol{b}) \in P_1^y(m) \big| L_x - |x_a - x_b| = k\big\}. \tag{26b}$$

Notice that $P_1^x(m) = \bigcup_{k=1}^m P_1^x(m,k)$ and $P_1^y(m) = \bigcup_{k=1}^m P_1^y(m,k)$. Figures 4(a) and 4(b) give visualizations of pairs of sites within $P_1^x(m,m)$. Figure 4(c) gives examples of distances between pairs of sites in $P_1^x(m,k)$, for $k = 1,\ldots,m$. By definition Eqs. (16a) and (18a) we have that

$$S_1^p(m) \backslash S_1^n(m) = P_1^x(m) \cup P_1^y(m). \tag{27}$$

Hence, by combining Eqs. (24a) and (7), we obtain

$$\begin{aligned} r_1(m) &= \big|P_1^x(m) \cup P_1^y(m)\big| \\ &= \sum_{k=1}^m \big|P_1^x(m,k)\big| + \sum_{k=1}^m \big|P_1^y(m,k)\big| - \big|P_1^x(m) \cap P_1^y(m)\big|. \end{aligned} \tag{28}$$

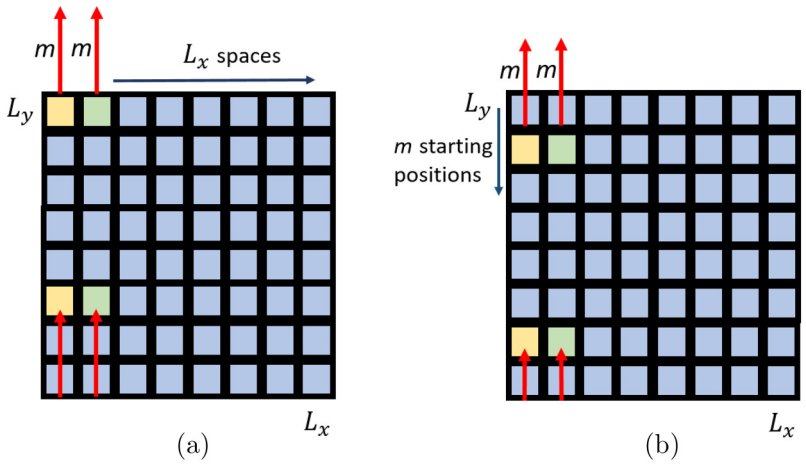
To conclude the computation we derive an expression for the two sums in Eq. (28) and the corresponding equation for the size of the intersection. By counting the contribution of each type of pair (see Fig. 4 for a visualization), one can write down the following expressions for the two sums in Eq. (28):

$$\sum_{k=1}^m \big|P_1^x(m,k)\big| = 2[L_x + 2L_x + \cdots + L_x(m-1)] + L_xm, \tag{29a}$$

$$\sum_{k=1}^m \big|P_1^y(m,k)\big| = 2[L_y + 2L_y + \cdots + L_y(m-1)] + L_ym\ . \tag{29b}$$

Hence,

$$\begin{aligned} &\sum_{k=1}^m \big|P_1^x(m,k)\big| + \sum_{k=1}^m \big|P_1^y(m,k)\big| \\ &\quad = 2[L_x + 2L_x + \cdots + L_x(m-1)] + L_xm \\ &\qquad + 2[L_y + 2L_y + \cdots + L_y(m-1)] + L_ym \end{aligned}$$



k	1	2	...	m-1	m
Illustration of pair types in $P_1^x(m,k)$	m-1 1 m-1	m-2 2 m-2		1 m-1 1	m
$\lvert P_1^x(m,k)\rvert$	$2(L_x)$	$2(2L_x)$	...	$2((m-1)L_x)$	mL_x

(c)

FIG. 4. A visualization of the pairs of sites in $P_1^x(m)$. Panels (a) and (b) show two different site pairs in $P_1^x(m,m)$. For each of the L_x columns, each site in the rows $\{L_y-m+1, L_y-m+2, \ldots, L_y\}$ has a single corresponding site at distance m separated by m rows and 0 columns and reached by crossing the horizontal boundary. Therefore $|P_1^x(m,m)| = mL_x$. Panel (c) shows all the different types of pairs in $P_1^x(m,k)$ for $k = 1,\ldots,m$ with the corresponding value of $|P_1^x(m,k)|$.

$$
\begin{aligned}
&= (L_x + L_y)\left(m + 2\sum_{i=1}^{m-1} i\right) \\
&= (L_x + L_y)\left[m + 2\frac{(m-1)m}{2}\right] \\
&= (L_x + L_y)m^2. \qquad (30)
\end{aligned}
$$

We now focus on deriving an expression for the size of intersection, $|P_1^x(m) \cap P_1^y(m)|$, in Eq. (28). The set $P_1^x(m) \cap P_1^y(m)$ consists of pairs of sites separated by distance m that cross both the x and y boundaries simultaneously.

There are two regions of the domain where site pairs cross two boundaries. These are any two consecutive corners of the four corners of the domain. Examples of these regions and site pairs within these regions are visualized in Fig. 5. In Fig. 6 we give an illustrative example in which we count the number of these pairs for $m = 5$. All sites inside the boundaries of the domain colored in orange, purple, green, and yellow are distance $m = 5$ from other sites of the same color outside the boundaries of the domain. Notice that the yellow site in the corner at (L_x, L_y) is distance five from a total of four sites, reached by crossing the x and y boundaries, denoted by a 4 in the site. Similarly, the two green sites at $(L_x - 1, L_y)$ and $(L_x, L_y - 1)$ are distance m from three sites, reached by crossing the x and y boundary, denoted by a 3 in the two sites. $|P_1^x(5) \cap P_1^y(5)|$ is the sum of all the numbers in the colored sites multiplied by two to account for the second corner region. Extrapolating, for any value of m, the number of pairs of sites

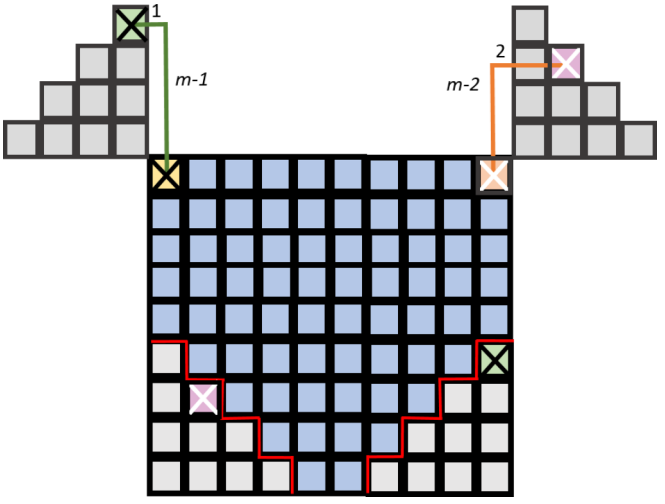


FIG. 5. Examples of pairs of sites separated by distance $m = 5$ that cross both the x and y boundaries, i.e., pairs of sites in $P_1^x(m) \cap P_1^y(m)$. The gray sites outside the domain correspond to the gray sites inside the domain in the diametrically opposite corner. As illustrations of site pairs at a distance m which cross both boundaries, the orange site containing a white cross is distance m from the pink site containing a white cross. Similarly, the yellow site containing a black cross is distance m from the green site containing a black cross.

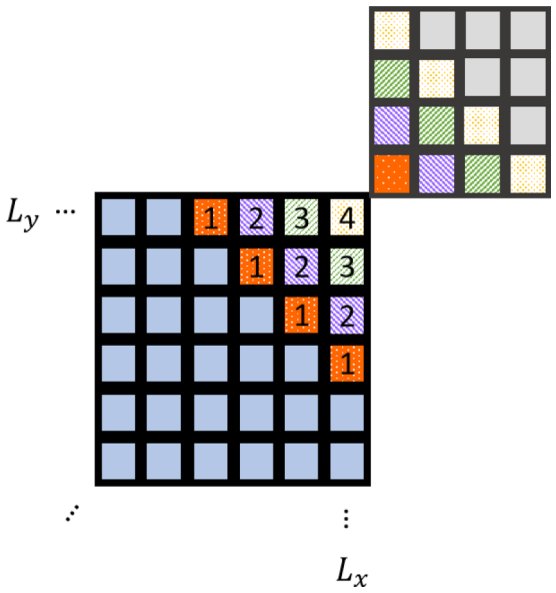


FIG. 6. Examples of pairs in $P_1^x(5) \cap P_1^y(5)$ on a zoomed-in corner of a larger domain. Sites with a given pattern (not plain) are distance $m = 5$ away from sites with the same pattern that can be reached by crossing the x and y boundaries. The number in each given site corresponds to the number of sites at distance $m = 5$, reached by crossing the x and y boundaries.

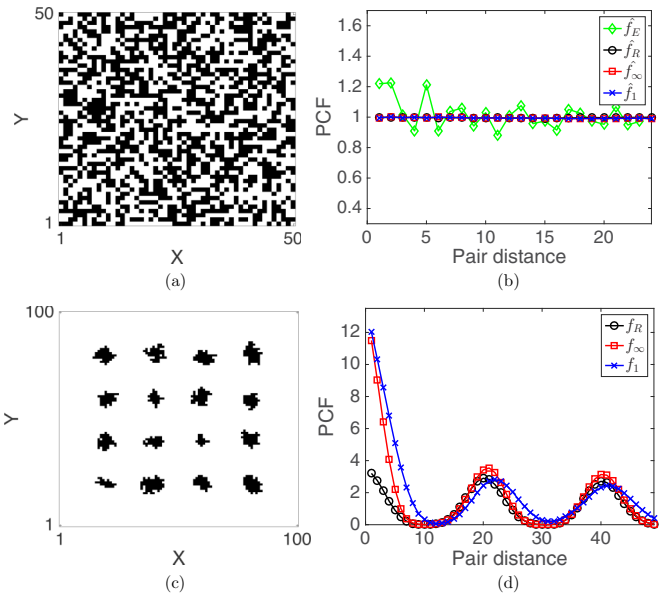


FIG. 7. Examples of spatial structure analysis. In panels (a) and (b) a case with no spatial correlation is considered. Panel (a) is an example visualization of an occupancy matrix uniformly populated with density 0.5. Occupied sites are colored in black and are white otherwise. Panel (b) shows the four PCFs each averaged over 50 uniformly populated matrices. Panels (c) and (d) refer to a discrete simulation of the agent-based model described in the text, at time $t = 10$. Panel (c) is a visualization of the occupancy matrix and in panel (d) our PCFs are compared with the Rectilinear PCF.

that cross the two boundaries is exactly

$$\begin{aligned}\left|P_1^x(m) \cap P_1^y(m)\right| &= 2[(m-1) + 2(m-2) \\ &\quad + 3(m-3) + \cdots + m - 1] \\ &= 2\sum_{i=1}^{m-1}(m-i)i = \frac{m^3 - m}{3}. \end{aligned} \tag{31}$$

By substituting Eqs. (31) and (30) into Eq. (28) we gain an expression for the remainder $r_1(m)$. By rearranging Eq. (24a) we determine $s_1^n(m)$, which we then substitute into Eq. (19) to obtain the exact expression for the normalization in the nonperiodic case. This is given by

$$\begin{aligned}\mathbb{E}\left[\bar{c}_1^n(m)\right] &= \left(\frac{N}{L_x L_y}\right)\left(\frac{N-1}{L_x L_y - 1}\right) \\ &\quad \left[2mL_xL_y - (L_x + L_y)m^2 + \frac{m^3 - m}{3}\right]. \end{aligned} \tag{32}$$

A similar approach can be used to obtain the normalization factor for the uniform metric under nonperiodic BC:

$$\begin{aligned}\mathbb{E}\left[\bar{c}_\infty^n(m)\right] &= \left(\frac{N}{L_x L_y}\right)\left(\frac{N-1}{L_x L_y - 1}\right) \\ &\quad [4mL_xL_y - 3(L_x + L_y)m^2 + 2m^3]. \end{aligned} \tag{33}$$

For more details on the derivation of Eq. (33), see Supplemental Material Sec. S.1.

IV. RESULTS

In this section we use the square taxicab and square uniform PCFs defined in Sec. III to analyze the spatial correlation in some examples. We compare our results with previously suggested on-lattice PCFs.

We start by computing the PCFs for a system without any spatial correlation. We consider 50 independent occupancy matrices, U_i, $i = 1, \ldots, 50$, populated uniformly at random with density 0.5 [see Fig. 7(a) for example]. For each realisation, U_i, we compute the corresponding PCF, $f_d^{U_i}$, and then we average the results over the 50 realisations which we denote $\hat{f}_d$. If the normalization is correct, $\hat{f}_d(m)$ should return the value unity for every pair distance, m, meaning that no spatial correlation is found. In Fig. 7(b) all four aforementioned averaged PCFs are plotted: $\hat{f}_A, \hat{f}_R, \hat{f}_1$, and $\hat{f}_\infty$. The results show that both the averaged square uniform PCF, $\hat{f}_1$, and square taxicab PCF, $\hat{f}_\infty$, correctly predict that there is no spatial correlation. The averaged rectilinear PCF, $\hat{f}_R$, also correctly predicts no spatial correlation. However, the averaged annular PCF, $\hat{f}_A$, has clear peaks, suggesting, incorrectly, the presence of spatial correlation. Since the results are averaged over multiple repeats, such a discrepancy can not be attributed to stochasticity, but is due to incorrect normalization as explained in Sec. II. Note that the annular PCF can still correctly identify spatial correlation in many examples; however, the incorrect normalization often makes the results hard to interpret. This is because the annular PCF makes it difficult to distinguish between genuine correlation and systematic error. For this reason, for the rest of this section, we omit the results of the annular PCF and continue to compare between our PCFs and the rectilinear PCF using nonperiodic BC.

Next we consider examples of strong spatial correlation. Figure 7(c) shows an example of aggregation driven by a

proliferation mechanism. The occupancy matrix is obtained by simulating an on-lattice agent-based model with periodic BC as described in Binder and Simpson [12], which we summarize as follows. The model is initialized with 16 agents, located at coordinates given by $\{(x,y)\,|\,x,y \in \{20,40,60,80\}\}$ on a regular square lattice with $L_x = 100$, $L_y = 100$. Time is discretized with a time step $\tau = 1$ and the number of agents at time t is denoted by $n(t)$. At each time step the configuration at time $t+\tau$ is obtained from the configuration at time t, by repeating the following steps $n(t)$ times. (1) An agent is chosen uniformly at random from the $n(t)$ agents present at the end of the previous time-step; (2) one of its four von Neumann neighbors is selected at random with equal probability; (3) if the selected site is empty, then a new agent is placed in this site and $n(t+\tau) = n(t)+1$; otherwise, the configuration is left unchanged.

Figure 7(c) shows a single realisation after 10 time steps and Fig. 7(d) shows the corresponding PCFs: f_R, f_1, and f_∞. The results indicate that all of the PCFs correctly identify aggregation. However, the quantitative information about aggregate sizes at different length scales provide by each PCF varies. For example, we see that all PCFs in Fig. 7(d) exhibit three peaks; at $m = 1$, $m \approx 20$, and $m \approx 40$. The different peaks and troughs of the PCF profiles have different qualitative meanings related to the correlation type. Due to the local approach of the square taxicab PCF and square uniform PCF, the first peak at $m = 1$ is three times higher than the peaks at larger values of distance. These differences in amplitude highlight the different peak origins. Specifically, the first and highest peak distinguishes the individual cluster aggregate and the later peaks indicate correlation between different clusters. In contrast, all three peaks in the rectilinear PCF are the same amplitude. Note that, in the case of aggregation, the average diameter of the aggregate corresponds to the first value of distance which achieves the minimum of the PCF. The rectilinear, square uniform, and square taxicab PCFs estimate the aggregate diameter to be 9, 9 and 11, respectively. Importantly, the PCFs capture the fact that this diameter depends on the metric used. In particular the distance between two sites measured using the uniform metric is always less than or equal to the taxicab distance. This phenomenon is seen more clearly in later examples.

We now consider a series of examples with spatial correlation constructed artificially to compare and evaluate the different PCFs. In Fig. 8 we compare our square uniform and square taxicab PCF with the rectilinear PCF for three different patterns with strong spatial correlation. All three examples (Fig. 8(a), diagonal stripes; Fig. 8(c), chessboard pattern; and Fig. 8(e), concentric circles) are chosen so that the column- and row-averaged densities are constant and hence the spatial structure is not recognised by the rectilinear PCF, as shown in Figs. 8(b), 8(d), and 8(f). This is in contrast to the approach of our PCFs (both square uniform and square taxicab), which successfully recognize the spatial structure in all three examples. In addition, these examples uncover other interesting differences between the taxicab and the uniform approaches. Consider the PCF for the case of diagonal stripes and the chessboard pattern [Figs. 8(b) and 8(d)]. Here the square uniform PCF quickly converges to unity (no spatial correlation) for large distance m, while in both cases, the square taxicab PCF still shows a strong oscillatory behavior for large distance m, suggesting spatial correlation. To give an intuitive

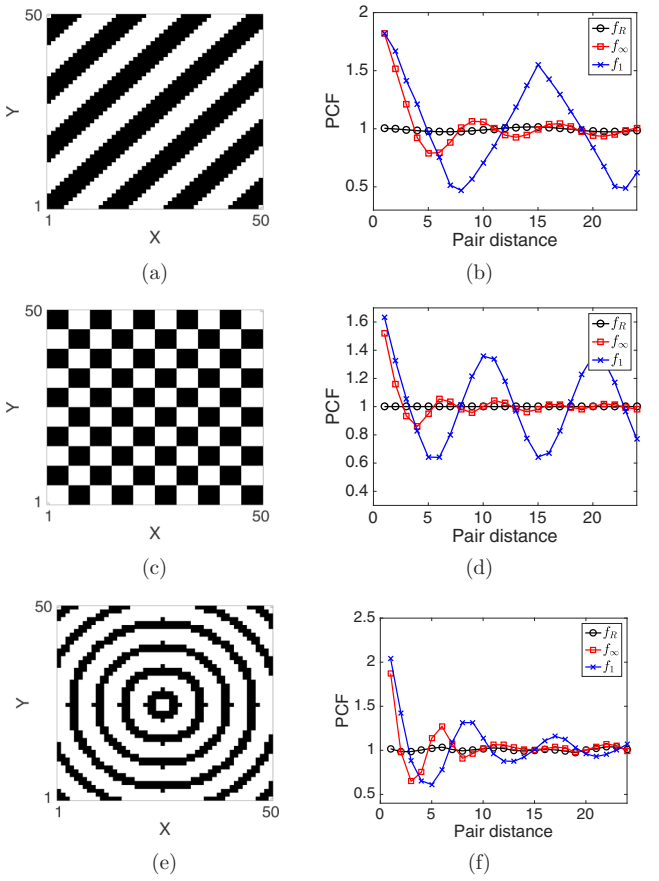


FIG. 8. Examples of pattern analysis. Panels (a), (c), and (e) visualize three constructed spatial patterns. Panels (b), (d), and (f) displays the corresponding square taxicab, square uniform and rectilinear PCF using the occupancy matrices for (a), (c), and (e), respectively.

explanation of this phenomenon, let us consider the shapes of balls of size m centered at a given site $\boldsymbol{a}$ under the two metrics. These balls are defined as $B_m(\boldsymbol{a}) = \{\boldsymbol{b} \in \mathbb{L} \,|\, \|\boldsymbol{a}-\boldsymbol{b}\|_d \leqslant m\}$ with $d = 1,\infty$, respectively (see Fig. 3). The ball corresponding to the uniform metric [Fig. 3(b)] has a square shape with the sides aligned with the directions of the axis. This implies that when distance m becomes close to either L_x or L_y in size, the ball corresponding to the uniform metric of distance m contains most of the sites in the corresponding row or column at distance m. For large m, therefore, the uniform metric begins to work in a similar way to the Rectilinear PCF and thus fails to recognize anisotropic patterns biased in the Cartesian directions. The ball of the taxicab metric [see Fig. 3(a)], however, has a diamond shape. Consequently, the long-distance correlations appear clear even for patterns in which both the average column and row densities are constant, as in Fig. 8.

The examples in Fig. 8 were constructed specifically to underline the main differences between the three PCFs. Nevertheless, similar patterns also arise in many biologically and mathematically relevant applications [32–34]. We conclude this section by comparing the three PCF approaches applied to some real-world examples taken from the literature. In Fig. 9 we analyze three images representing examples of Turing patterns. A corresponding occupancy matrix is obtained by representing each pixel of the image as a value in a matrix

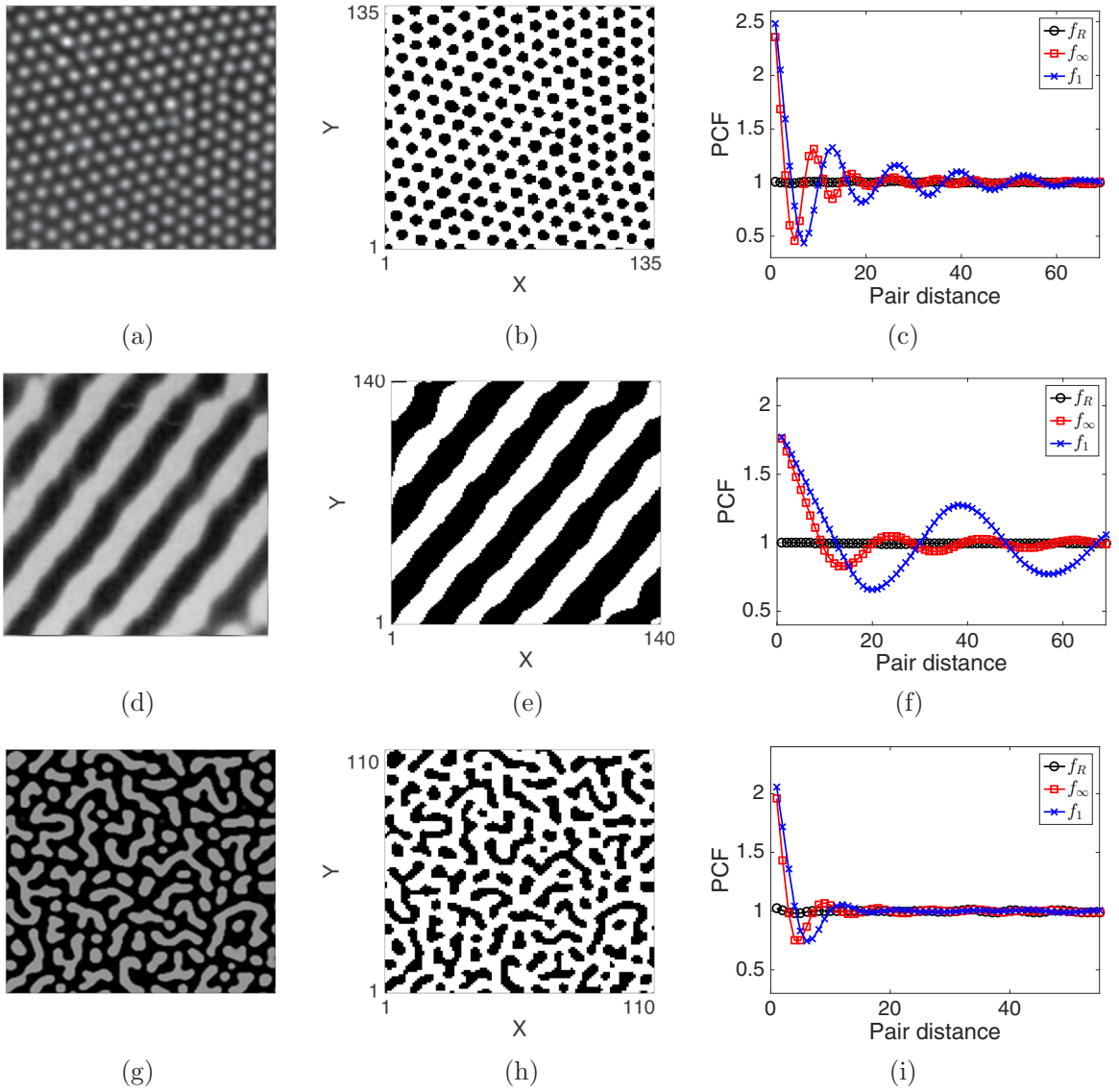


FIG. 9. Spatial analysis of Turing patterns. Panels (a), (d), and (g) show original images representing the results of a reaction-diffusion mechanism between two chemical substances, (a) is reprinted from Ref. [35], (d) from Ref. [32], and (g) from Ref. [36]. Panels (b), (e), and (h) visualize the occupancy matrices corresponding to the original images (described in text). In panels (c), (f), and (i) we compare the square taxicab, square uniform, and rectilinear PCFs for each of the examples.

which is 1 (i.e., occupied) if the three values of the RGB colorization of the pixel are above a certain threshold (80) and 0 otherwise. In all cases the column and row densities are almost constant, hence the spatial structure again remains largely undetected by the rectilinear PCF, while our square uniform and square taxicab PCF correctly identify the patterns. As already observed in the previous examples, we note that the estimated diameter of the aggregate, wavelength, and amplitude of the oscillations differ according to the metric used.

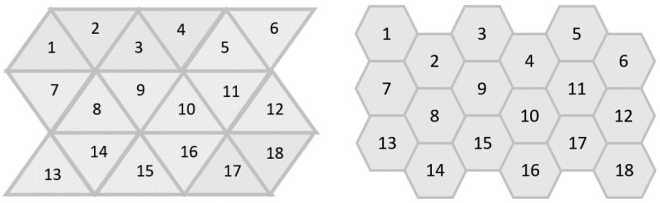
V. THE TRIANGLE, HEXAGON, AND CUBE PCFS

Despite the square lattice being the most popular set up for spatially discrete models [13,37–40], in some situations other types of tessellation, either regular or irregular, can be more suitable [13,17,41].

In the following subsections we extend our definition of the PCFs in Sec. III to more general types of tessellations. We define the triangle, hexagon, cube uniform, and cube taxicab PCFs under nonperiodic and periodic BC for triangular, hexagonal, and cuboidal tessellations, respectively. The following subsections represent qualitative discussions of the different cases. We refer the reader to the Supplemental Material Secs. S.2 and S.3 for the full details of the derivation of the PCF formulas [31].

A. Triangle and hexagon PCF

First, we define triangularly and hexagonally tessellated domains of size $L_x \times L_y$. These comprise an array of L_y rows of L_x regular triangles or hexagons, respectively. Examples for which $L_x = 6$ and $L_y = 3$ for each of the two cases are given in Figs. 10(a) and 10(b), respectively. Notice that, for a



(a) (b)

FIG. 10. Example domains for (a) triangular and (b) hexagonal tessellations in which $L_x = 6$ and $L_y = 3$.

periodic BC to be meaningful in these domain definitions, L_x must be even. Therefore, we enforce this as a condition in what follows.

In the context of triangular and hexagonal tessellations we focus our attention on the taxicab metric, both for simplicity and as the most natural metric on this domain type. Using the taxicab metric, the number of sites of distance m from any given reference site is given by

$$t_{\text{tri}}(m) = 3m, \tag{34a}$$

$$t_{\text{hex}}(m) = 6m. \tag{34b}$$

The proofs of Eqs. (34) are omitted, but they can be obtained easily by induction on m. Examples for $m = 1,2,3$ are visualized in Fig. 11. Using the same reasoning as in Sec. III under periodic BC:

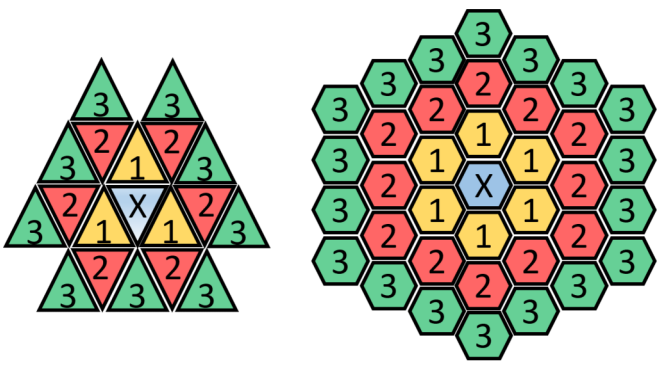
$$s^p_{\text{tri}}(m) = 3m\frac{L_x L_y}{2}, \tag{35a}$$

$$s^p_{\text{hex}}(m) = 3mL_x L_y. \tag{35b}$$

Substituting Eqs. (35) into Eq. (19) we obtain the normalizations for the triangle and hexagon PCF, respectively, under periodic BCs, namely,

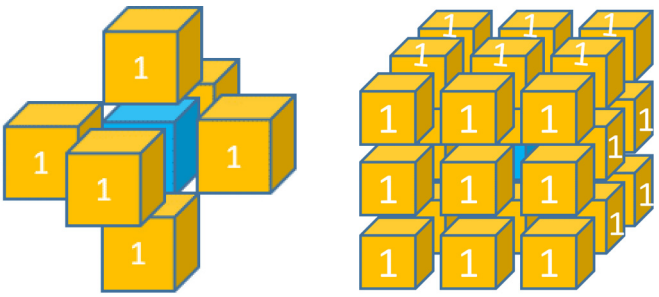
$$\mathbb{E}\left[\bar{c}^p_{\text{tri}}(m)\right] = \frac{3mN(N-1)}{2(L_x L_y - 1)}, \tag{36a}$$

$$\mathbb{E}\left[\bar{c}^p_{\text{hex}}(m)\right] = \frac{3mN(N-1)}{L_x L_y - 1}. \tag{36b}$$



(a) (b)

FIG. 11. Schematic of agent pairs with the (a) triangular tessellation and (b) hexagonal tessellation using the taxicab metric. Sites in yellow, red, and green, labeled 1, 2, and 3, respectively, are distance-one, -two, and -three neighbors from the blue site (labeled with X), respectively.



(a) (b)

FIG. 12. Schematic of agent pairs using (a) the taxicab metric and (b) the uniform metric. Sites in yellow are defined to be distance-one neighbors from the site marked in blue.

From these expressions one can obtain the formulas of f_{tri} and f_{hex} under periodic BC by using the definition Eq. (9). The normalizations in the case of nonperiodic BC are given in the Supplemental Material Sec. S.2 [31].

B. Uniform cube and taxicab cube PCF

We define a three-dimensional $L_x \times L_y \times L_z$ cuboidal lattice with unit spacing. Using the taxicab and uniform metric, respectively, the number of sites of distance m from any given reference site is given by

$$t_{\text{cube}_1}(m) = 2(2m^2 + 1), \tag{37a}$$

$$t_{\text{cube}_\infty}(m) = 2(12m^2 + 1). \tag{37b}$$

The proofs of Eqs. (37) are omitted, but they can be obtained easily by induction on m. Examples of agent pairs for $m = 1$ are given in Figs. 12(a) and 12(b) for the taxicab and uniform metrics, respectively. Using the same reasoning as in Sec. III, under periodic BCs, the normalizations for taxicab and uniform cube PCFs, respectively, are as follows:

$$s^p_{\text{cube}_1}(m) = (2m^2 + 1)L_x L_y L_z, \tag{38a}$$

$$s^p_{\text{cube}_\infty}(m) = (12m^2 + 1)L_x L_y L_z. \tag{38b}$$

For simplicity we refer the reader to Sec. S.3 of the Supplemental Material for the normalization factors for the cases with nonperiodic BC [31].

VI. THE GENERAL PCF

In this section we provide a comprehensive method for generating a PCF for any tessellation type, BC, and metric but with the caveat of having a high computational cost.

This PCF is a valuable tool for irregular domain shapes and partitions although it can be used for any tessellation of any domain.

First, we consider a two-dimensional domain partitioned into Z regions (or sites) with arbitrary shapes and sizes, each labeled with a number from 1 to Z. Figure 13(a) shows an example of an irregularly shaped domain partitioned in $Z = 17$ regions. Given the domain, we choose a suitable metric. For the irregular lattice, which we consider in the following example, we consider the taxicab metric. This means that we define the

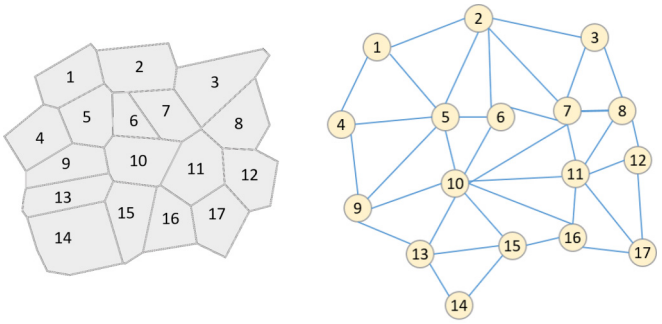


FIG. 13. An example of an irregular domain partition with its corresponding connectivity graph under the taxicab metric. Panel (a) shows A size 17 irregular lattice domain. Panel (b) shows corresponding connectivity graph for the tessellation in (a) under nonperiodic BC using the taxicab metric for distances.

distance, m, between two sites to be the minimum number of sites visited when starting at one site and moving consecutively through adjacent sites to the other. For example, in Fig. 13(a) the sites 4 and 7 are at distance three. Similarly, adjacent sites are defined to be at distance one.

Having chosen and defined a suitable metric, we may now represent the connections between lattice sites as an undirected connectivity graph $G(V,E)$, where each vertex represents a lattice site and each edge connects vertices whose corresponding sites are distance-one neighbors. Using the taxicab metric, edges connect vertices whose corresponding sites are adjacent. Figure 13(b) shows an example of such an association [applied to the irregular lattice in Fig. 13(a)] using the taxicab metric.

The corresponding adjacency matrix of graph G is a $Z \times Z$ matrix defined as follows:

$$A^G_{i,j} = \begin{cases} 1 & \text{for } (i,j) \in E, \\ 0 & \text{for } (i,j) \notin E. \end{cases} \tag{39}$$

We use properties of the adjacency matrix to determine the number of sites at a given distance. In particular, we can compute $(A^G)^m$ whose entries $(A^G)^m_{i,j}$ are the number of walks of length m from vertex i to vertex j. To compute the minimum walk between two sites (and hence the distance between them) we produce the *distance matrix* D^G. This is a $Z \times Z$ matrix defined as

$$D^G_{i,j} = \begin{cases} \min\left\{m \in \mathbb{N}^+ \middle| (A^G)^m_{i,j} \neq 0\right\} & \text{for } i \neq j, \\ 0 & \text{for } i = j. \end{cases} \tag{40}$$

Notice that, each entry, $D^G_{i,j}$, denotes the distance between the vertices i and j on graph G and hence on the original lattice.

Given the distance matrix of the domain, D^G, and the set of the occupied sites $M \subseteq V$, the PCF of the system can be computed as follow. The number of pairs of agents at distance m for a general metric d is given by

$$c_d(m) = \tfrac{1}{2}\left|\left\{(i,j) \in M \times M \middle| D^G_{i,j} = m\right\}\right|. \tag{41}$$

Similarly, we can express the number of pairs of sites at distance m as

$$s_d(m) = \tfrac{1}{2}\left|\left\{(i,j) \in V \times V \middle| D^G_{i,j} = m\right\}\right|. \tag{42}$$

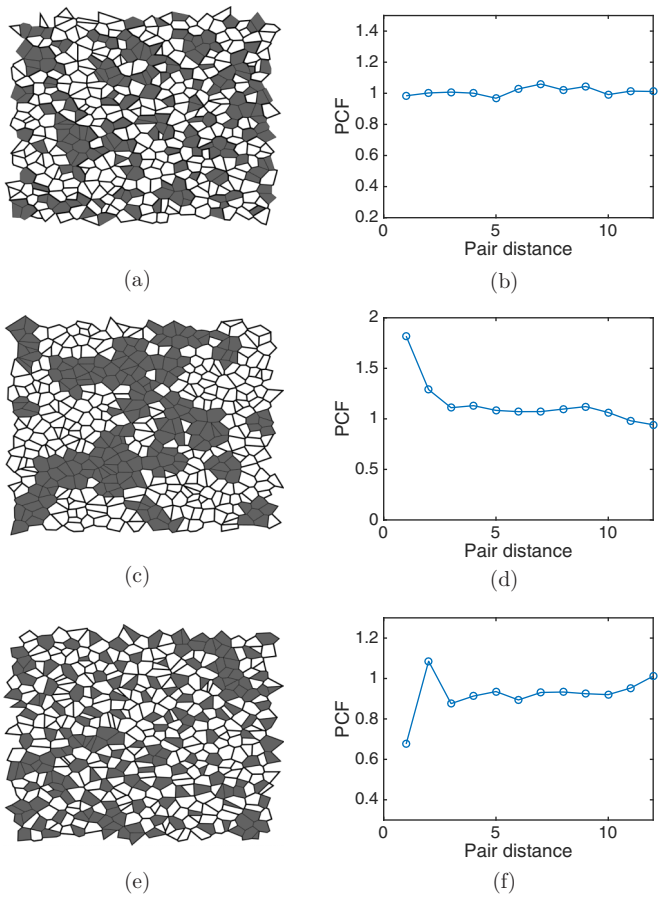


FIG. 14. Examples of spatial correlation analysis on an irregular domain. Panels (a), (c), and (e) show three examples of irregular lattices populated with density 0.4 with agents (gray sites). In panel (a), agents are displaced uniformly at random (no spatial correlation). In panel (c), agents are in a strong form of aggregation, while in panel (e) agents are displaced in a segregate manner. Panels (b), (d), and (f) are the corresponding general PCF evaluations for panels (a), (c), and (e), respectively.

To compute the normalization factor, denote the total number of agents as $N = |M|$, and hence using the same argument as in Sec. III we can write

$$\mathbb{E}[\bar{c}_d(m)] = \left(\frac{N}{Z}\right)\left(\frac{N-1}{Z-1}\right)s_d(m). \tag{43}$$

The general PCF is then defined by combining Eqs. (42) and (43) in Eq. (9). Notice that the computation of the normalization for the general PCF can be computationally expensive. This is because the computation of D^G involves calculating powers of matrices of size $Z \times Z$, where Z is often large. In particular, the cost of computing the normalization of the general PCF is $O(Z^3 m_{\max})$ in which $m_{\max}$ is the maximum value m for which the PCF is computed. For this reason, the general PCF is better reserved for cases in which the expression for the normalization factor cannot be computed analytically, unlike in Secs. III, V A, and V B, although it can, of course, be used even if an analytical formula is available.

In Fig. 14 we apply the general PCF to three examples of agent-based systems on an irregular lattice. In all three examples, the irregular tessellation is the Voronoi partition

based on a set of randomly distributed points. The randomized points are obtained by starting with a square lattice, with lattice size Δ, and perturbing the coordinates of each point (x_i, y_i) to $(x_i + \delta_i^x, y + \delta_i^y)$ with each δ_i^x, δ_i^y chosen uniformly at random in the interval $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$.

In the first example, Fig. 14(a), N lattice sites selected uniformly at random to be occupied (gray sites). By eye, several larger clusters of occupied and unoccupied lattice sites are evident, indicating that there may be spatial correlation. Figure 14(b) shows the corresponding general PCF. The values are close to unity, correctly identifying that there is, indeed, no spatial correlation in the system. This highlights the importance of accurate quantitative methods for determining spatial correlation rather than a reliance on ad hoc judgements.

In the other two cases we test our PCF with examples of strong spatial correlation. In Fig. 14(c) we consider a system in an aggregated state. To generate such a configuration, we start with an empty domain, and select an empty site uniformly at random. We then place an agent in this site and all of its neighboring sites (if they are not already occupied). We repeat this process until we reach density 0.4. The process leads to a strong form of aggregation which the general PCF, in Fig. 14(d), correctly identifies. Finally, in Fig. 14(e) we consider a system in a segregated state. To generate such a configuration, we start with a fully populated domain, and repeatedly select at random an occupied site. We remove all agents occupying adjacent sites but leave the initially selected site occupied. The process ends when density 0.4 is reached. This mechanism generates a system which is unlikely to have adjacent sites occupied and more likely to have agents displaced at distance two. The corresponding PCF, shown in Fig. 14(f), correctly captures both of these features: the PCF has value 0.68 at pair distance one, which implies negative correlation at the shortest distance, and value 1.1 at pair distance two, highlighting the positive correlation at slightly larger distances.

Note that, since the normalization in the general PCF can give an exact value for the number of sites at certain pair distances, we used this method to check the normalization factors for all of our previously proposed PCFs. In all cases the results confirmed the analytical expressions given in Secs. III, V A, and V B and in the Supplemental Material [31].

VII. CONCLUSIONS

In this paper we have developed a set of tools to study spatial correlation on discrete domains. We derived two discrete pair-correlation functions for an exclusion process on a square lattice: the square uniform and square taxicab PCF. We applied our PCFs to patterns observed in nature and to computational simulations and showed that our PCF can not only distinguish and quantify different types of correlation but also that it improves upon previous on-lattice PCFs. For example, we showed that our PCF was normalized correctly, unlike the annular PCF, and was able to identify and quantify anisotropic patterns such as the chessboard or the diagonal stripes that the rectilinear PCF [12] missed. Furthermore, we highlighted how different measures of distance, taxicab and uniform, can lead to different quantifications of spatial correlation.

We extended the calculation of appropriate PCFs to deal with exclusion processes on the other regular spatial tessellations in two dimensions as well as the cubic lattice in three dimensions. We derived the triangle PCF, hexagon PCF, cube taxicab PCF, and cube uniform PCF. These are the first PCFs defined specifically for these discrete lattice types. Finally, we derived a comprehensive PCF for any kind of discrete domain, BC, and metric, which we referred to as the general PCF. The method can be computationally expensive, however, it allows complete freedom in defining a suitable PCF for more complex cases, including those for which recognizing spatial correlation by eye becomes less intuitive.

All of our PCFs are designed for a single species of agents. However, in many applications the agents are divided into multiple species and it can be important to distinguish between different types of spatial correlations: either within agents of the same species (autocorrelation) or comparing the position of agents of different species (cross-correlation). Dini *et al.* [42] have recently investigated correlation in multiple species by using the Rectilinear PCF. We believe a similar approach to that of Dini *et al.* [42] can be applied to our isotropic PCFs to quantify heterogenous correlations, yet it lies beyond the scope of this paper and, as such, we will tackle it in a future publication. The isotropic PCFs that we defined in this paper will be important in further studies and applications. In particular, our functions can improve previous studies (see Ref. [43], for example) which have used PCF as an efficient summary statistics to infer model parameters.

ACKNOWLEDGMENT

The authors would like to thank the CMB-CNCB preprint club for constructive and helpful comments on a preprint of this paper. J.P.O. acknowledges support from the SWBio DTP.

[1] M. S. Steinberg, Adhesion in development: An historical overview, Dev. Biol. **180**, 377 (1996).

[2] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, Scale-free correlations in starling flocks, Proc. Natl. Acad. Sci. U.S.A. **107**, 11865 (2010).

[3] J. E. F. Green, S. L. Waters, J. P. Whiteley, L. Edelstein-Keshet, K. M. Shakesheff, and H. M. Byrne, Nonlocal models for the formation of hepatocyte–stellate cell aggregates, J. Theor. Biol. **267**, 106 (2010).

[4] R. J. Thomas, A. Bennett, B. Thomson, and K. M. Shakesheff, Hepatic stellate cells on poly (dl-lactic acid) surfaces control the formation of 3D hepatocyte coculture aggregates *in vitro*, eCells Mater. **11**, 16 (2005).

[5] F. Chalub, Y. Dolak-Struss, P. Markowich, D. Oelz, C. Schmeiser, and A. Soreff, Model hierarchies for cell aggregation by chemotaxis, Math. Models Methods Appl. Sci. **16**, 1173 (2006).

[6] H. G. Othmer and A. Stevens, Aggregation, blowup, and collapse: The ABC's of taxis in reinforced random walks, SIAM J. Appl. Math. **57**, 1044 (1997).

[7] A. Stevens, A stochastic cellular automaton modeling gliding and aggregation of myxobacteria, SIAM J. Appl. Math. **61**, 172 (2000).

[8] H. Murakawa and H. Togashi, Continuous models for cell-cell adhesion, J. Theor. Biol. **374**, 1 (2015).

[9] M. J. Simpson, B. J. Binder, P. Haridas, B. K. Wood, K. K. Treloar, D. McElwain, and R. E. Baker, Experimental and modeling investigation of monolayer development with clustering, Bull. Math. Biol. **75**, 871 (2013).

[10] K. K. Treloar, M. J. Simpson, P. Haridas, K. J. Manton, D. I. Leavesley, D. S. McElwain, and R. E. Baker, Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies, BMC Syst. Biol. **7**, 137 (2013).

[11] E. Hinde, F. Cardarelli, M. A. Digman, and E. Gratton, *In vivo* pair correlation analysis of EGFP intranuclear diffusion reveals DNA-dependent molecular flow, Proc. Natl. Acad. Sci. U.S.A. **107**, 16560 (2010).

[12] B. J. Binder and M. J. Simpson, Quantifying spatial structure in experimental observations and agent-based simulations using pair-correlation functions, Phys. Rev. E **88**, 022705 (2013).

[13] A. Deutsch and S. Dormann, *Cellular Automaton Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis* (Springer Science & Business Media, Berlin, 2007).

[14] R. J. Friedman, D. S. Rigel, and A. W. Kopf, Early detection of malignant melanoma: The role of physician examination and self-examination of the skin, CA. Cancer J. Clin. **35**, 130 (1985).

[15] M. A. Weinstock, Early detection of melanoma, JAMA **284**, 886 (2000).

[16] A. F. G. Bourke and N. R. Franks, *Social Evolution in Ants* (Princeton University Press, Princeton, NJ, 1995).

[17] M. J. Keeling, The effects of local spatial structure on epidemiological invasions, Proc. R. Soc., Ser. B, Biol. Sc., Lond. **266**, 859 (1999).

[18] P. J. Diggle, J. Besag, and T. J. Gleaves, Statistical analysis of spatial point patterns by means of distance methods, Biometrics **32**, 659 (1976).

[19] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modeling of Spatial Point Patterns* (John Wiley & Sons, New York, 2008).

[20] D. J. G. Agnew, J. E. F. Green, T. M. Brown, M. J. Simpson, and B. J. Binder, Distinguishing between mechanisms of cell aggregation using pair-correlation functions, J. Theor. Biol. **352**, 16 (2014).

[21] B. J. Binder and M. J. Simpson, Spectral analysis of pair-correlation bandwidth: Application to cell biology images, R. Soc. Open Sci. **2**, 140494 (2015).

[22] B. J. Binder, J. F. Sundstrom, J. M. Gardner, V. Jiranek, and S. G. Oliver, Quantifying two-dimensional filamentous and invasive growth spatial patterns in yeast colonies, Plos Comput. Biol. **11**, e1004070 (2015).

[23] E. J. Hackett-Jones, K. J. Davies, B. J. Binder, and K. A. Landman, Generalized index for spatial data sets as a measure of complete spatial randomness, Phys. Rev. E **85**, 061908 (2012).

[24] N. A. Bahcall and R. M. Soneira, The spatial correlation function of rich clusters of galaxies, Astrophys. J. **270**, 20 (1983).

[25] A. Donev, S. Torquato, and F. H. Stillinger, Pair correlation function characteristics of nearly jammed disordered and ordered hard-sphere packings, Phys. Rev. E **71**, 011105 (2005).

[26] W. R. Young, A. J. Roberts, and G. Stuhne, Reproductive pair correlations and the clustering of organisms, Nature **412**, 328 (2001).

[27] R. N. Binny, A. James, and M. J. Plank, Collective cell behavior with neighbor-dependent proliferation, death, and directional bias, Bull. Math. Biol. **78**, 2277 (2016).

[28] M. Raghib, N. A. Hill, and U. Dieckmann, A multiscale maximum entropy moment closure for locally regulated space–time point process models of population dynamics, J. Math. Biol. **62**, 605 (2011).

[29] D. J. Bone, H. A. Bachor, and R. J. Sandeman, Fringe-pattern analysis using a 2D Fourier transform, Appl. Optics. **25**, 1653 (1986).

[30] M. Takeda, H. Ina, and S. Kobayashi, Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry, J. Opt. Soc. Am. **72**, 156 (1982).

[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.97.062104 for the normalizations of the pair correlation functions that were omitted in the main text, as well as a summary of the normalization formulas. All the MATLAB codes which accompany the paper can be found as online supplementary material.

[32] Q. I. Ouyang and H. L. Swinney, Transition from a uniform state to hexagonal and striped turing patterns, Nature **352**, 610 (1991).

[33] S. Y. Bhide and S. Yashonath, Dependence of the self-diffusion coefficient on the sorbate concentration: A two-dimensional lattice gas model with and without confinement, J. Chem. Phys. **111**, 1658 (1999).

[34] J. D. Murray, *Mathematical Biology: I. An Introduction*, Vol. 17 (Springer Science & Business Media, Berlin, 2007).

[35] P. Ball, Forging patterns and making waves from biology to geology: A commentary on Turing (1952); The chemical basis of morphogenesis, Phil. Trans. R. Soc. B **370**, 20140218 (2015).

[36] S. Kondo, An updated kernel-based turing model for studying the mechanisms of biological pattern formation, J. Theor. Biol. **414**, 120 (2017).

[37] M. J. Simpson, K. A. Landman, and B. D. Hughes, Multi-species simple exclusion processes, Phys. A **388**, 399 (2009).

[38] C. A. Yates, R. E. Baker, R. Erban, and P. K. Maini, Going from microscopic to macroscopic on nonuniform growing domains, Phys. Rev. E **86**, 021921 (2012).

[39] R. J. H. Ross, C. A. Yates, and R. E. Baker, Inference of cell–cell interactions from population density characteristics and cell trajectories on static and growing domains, Math. Biosci. **264**, 108 (2015).

[40] R. E. Baker, C. A. Yates, and R. Erban, From microscopic to macroscopic descriptions of cell migration on growing domains, Bull. Math. Biol. **72**, 719 (2010).

[41] A. P. Browning, S. McCue, and M. J. Simpson, A Bayesian computational approach to explore the optimal duration of a cell proliferation assay, Bull. Math. Biol. **79**, 1888 (2017).

[42] S. Dini, B. J. Binder, and J. E. F. Green, Understanding interactions between populations: Individual-based modeling and quantification using pair correlation functions, J. Theor. Biol. **439**, 50 (2018).

[43] S. T. Johnston, M. J. Simpson, D. L. S. McElwain, B. J. Binder, and J. V. Ross, Interpreting scratch assays using pair density dynamics and approximate Bayesian computation, Op. Bio. **4**, 140097 (2014).