Taylor & Francis
Taylor & Francis Group

Stochastics

An International Journal of Probability and Stochastic Processes

ISSN: 1744-2508 (Print) 1744-2516 (Online) Journal homepage: https://www.tandfonline.com/loi/gssr20

Parameter estimation in a subcritical percolation model with colouring

Felix Beck & Bence Mélykúti

To cite this article: Felix Beck & Bence Mélykúti (2019) Parameter estimation in a subcritical percolation model with colouring, Stochastics, 91:5, 657-694, DOI: 10.1080/17442508.2018.1539089

To link to this article: https://doi.org/10.1080/17442508.2018.1539089

Published online: 31 Oct 2018.

STOCHASTICS
2019, VOL. 91, NO. 5, 657–694
https://doi.org/10.1080/17442508.2018.1539089

Taylor & Francis
Taylor & Francis Group

Parameter estimation in a subcritical percolation model with colouring

Felix Beck[a,b] and Bence Mélykúti[b]

[a]Institute for Mathematics, University of Freiburg, Ernst-Zermelo-Straße 1, Freiburg, Germany; [b]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstraße 49, Freiburg, Germany

ABSTRACT

In the bond percolation model on a lattice, we colour vertices with n_c colours independently at random according to Bernoulli distributions. A vertex can receive multiple colours and each of these colours is individually observable. The colours colour the entire component into which they fall. Our goal is to estimate the $n_c + 1$ parameters of the model: the probabilities of colouring of single vertices and the probability with which an edge is open. The input data is the configuration of colours once the complete components have been coloured, without the information which vertices were originally coloured or which edges are open.

We use a Monte Carlo method, the method of simulated moments, to achieve this goal. Under the unproven assumption of identifiability, we show that this method is a strongly consistent estimator by proving a uniform strong law of large numbers for the vertices' weakly dependent colour values. Our proof method quantifies dependence among the spatially arranged random variables using percolation theory: the FKG and BK inequalities, and the exponential decay of the cluster size distribution. We evaluate the method in computer tests. We have made our software publicly available. The motivating application is cross-contamination rate estimation for digital PCR in lab-on-a-chip microfluidic devices.

ARTICLE HISTORY
Received 8 January 2018
Accepted 8 October 2018

KEYWORDS
Cross-contamination; method of simulated moments; microfluidics; parameter estimation; percolation; strong law of large numbers with dependence; uniform law of large numbers

1. Bond percolation with colouring

We consider bond percolation [15] on the triangular lattice, but our arguments hold for the square lattice as well. The vertex set of the infinite lattice is denoted by L. Edges are open (that is, included in the graph, alternatively, receive weight 1 as opposed to 0) independently at random with probability $\mu \in [0,1]$. There are $n_c \in \mathbb{N} \setminus \{0\}$ colours given, and for every colour $\ell \in \{1,2,\ldots,n_c\}$, a parameter $\lambda^\ell \in [0,1]$ is fixed. (Here ℓ is an index in superscript, not an exponent.) For every vertex $i \in L$, the vertex is coloured with colour $\ell \in \{1,2,\ldots,n_c\}$ according to a Bernoulli random variable with probability λ^ℓ. The colouring with different colours is independent in any one vertex, and it is also independent among different vertices. A vertex can receive multiple colours and each of these colours is

CONTACT B. Mélykúti melykuti@stochastik.uni-freiburg.de

individually observable. We call this colouring the *seeding*: $X_i^\ell \in \{0,1\}$ for every $i \in L$ and $\ell \in \{1,2,\dots,n_c\}$.

These colours propagate through open edges and colour ('*contaminate*') the entire component they are contained in. Let $i \leftrightarrow j$ mean that vertices $i,j \in L$ are connected by an open path. The observed colour configuration is

$$Y_i^\ell := X_i^\ell \vee \bigvee_{\substack{j\in L\\ j\leftrightarrow i}} X_j^\ell \in \{0,1\}$$

for every $i \in L$ and $\ell \in \{1,2,\dots,n_c\}$, where $\vee$ is the maximum operator. We write $i \sim j$ for adjacent lattice vertices $i,j \in L$ regardless of the state of the connecting edge.

We also consider this process on finite, connected subsets of the lattice $I \subset L$. (Here connected is meant with respect to the relation $\sim$, not only with respect to the open edges.) Picking the vertex set I implicitly fixes its edge set, the edges which connect vertices of I. We let $n_I := |I|$.

Often we consider nested sequences of such I where each successor is a superset of its predecessor and $n_I \to \infty$. We fix an ordering of the vertices of the infinite lattice L which is compatible with this sequence as $n_I \to \infty$, that is, each I comprises vertices labelled with $\{1,\dots,n_I\}$. We use $I_2 := \{(i,j) \in I \times I \mid i \sim j, i < j\}$ for the set of ordered pairs of adjacent vertices (independently of whether the connecting edge is open or closed) and $n_p := |I_2|$ for the total number of possible edges within I. We define the *exterior vertex boundary* of a subset I by

$$\Delta I := \{j \in L \mid j \notin I, \exists i \in I : i \sim j\}.$$

We always require that in our sequences, $|\Delta I|/|I| \to 0$. By the *degree sum formula*, this assumption also implies $n_p \sim 3n_I$ for the triangular lattice (asymptotic equality; $n_p \sim 2n_I$ is the corresponding condition for the square lattice).

For a fixed I, we define a variant of Y_i^ℓ that is determined exclusively by the seeding and edges in I:

$$\widetilde{Y}_i^\ell := X_i^\ell \vee \bigvee_{\substack{j\in I\\ j\overset{\sim}{\leftrightarrow} i}} X_j^\ell \in \{0,1\}$$

for every $i \in I$ and $\ell \in \{1,2,\dots,n_c\}$. Here $\overset{\sim}{\leftrightarrow}$ means connectedness by open edges in the edge set of I.

Our goal is to estimate the parameter $\theta = (\lambda^1,\dots,\lambda^{n_c},\mu)$ from the data $(\widetilde{Y}_i^\ell)_{i\in I,\ell\in\{1,2,\dots,n_c\}}$ (Figure 1). The spatial arrangement of $(\widetilde{Y}_i^\ell)$ within the lattice is known, but the seeding (X_i^ℓ) and the open or closed state of the edges are unavailable. We bring together four theoretical tools in this paper.

First, parameter estimation is conducted by the *method of simulated moments* (MSM; described in Section 2) [13,14]. This is a simulation-based, computationally intensive but parallelizable statistical method that yields a point estimate for θ which converges almost surely to the correct value as $n_I \to \infty$.

Second, as the first step towards proving the strong consistency of the estimator, we prove a strong law of large numbers (SLLN) with weakly dependent variables. We do this in Section 3 by adapting Theorem 1 of [9].

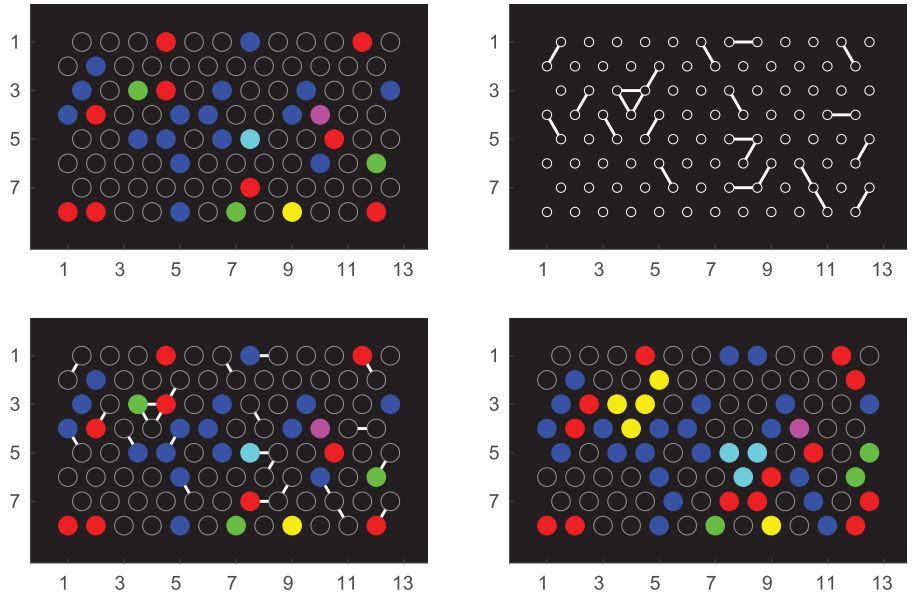

Figure 1. (*top left*) A realization of random seeding (X_i^ℓ) with $(\lambda^{\text{red}}, \lambda^{\text{green}}, \lambda^{\text{blue}}) = (0.1, 0.05, 0.2)$. Red and green together yield yellow, red and blue yield magenta, green and blue yield cyan, and red, green and blue together yield white. (*top right*) A realization of bond percolation on the triangular lattice with $\mu = 0.1$. (*bottom left*) The bond percolation overlaid with the seeding. (*bottom right*) The resulting configuration $(\widetilde{Y}_i^\ell)$ which serves as the data.

Third, the SLLN result requires some grasp of how small the dependence is between distant vertices of the lattice. The upper bounds on correlations are provided by the FKG and BK inequalities of percolation theory and the exponential decay of the cluster size distribution [1,5,10,15, Chapters 2 and 6,16,20] in Section 4.

Fourth, for the strong consistency of the estimator, we extend the SLLN to be uniform in the parameter vector. We verify in Section 5 that the conditions of a sufficient condition for the uniform law of large numbers (ULLN) hold [25, p. 8, Theorem 2], [28, p. 25, Lemma 3.1].

Our estimation method is tested on synthetic data with known parameter values in Section 6 and its performance is evaluated. Our software is publicly available [4]. In Section 7, the motivating problem is described, and the paper concludes with a discussion of possible improvements in modelling and methodology. Some readers will benefit from perusing Section 7 first, before progressing to Section 2. Readers interested only in applying the estimation method might choose to move from Section 2 directly to Section 6.

2. Method of simulated moments (MSM)

2.1. General description

The MSM is a modification of the classical method of moments for parameter estimation for the case when the moments of the sampling distribution cannot be computed from

the parameters in closed form. The MSM proposes to simulate n_s independent, identically distributed samples from the distribution, repeatedly with different parameter values θ (usually, but not strictly necessarily, with common random variables as θ is changed), and to choose the θ which gives the closest match between moments of the data and that of the simulated data. For its detailed description, we recommend perusing a combination of [13,14].

The data $\mathcal{Y} = (\mathcal{Y}_i)_{i\in I}$ originates from a distribution which is parameterized by the unknown $\theta_0 \in \Theta$. θ_0 is called the *true value of the parameter*. Normally, the $\mathcal{Y}_i$ are independent. A sample from this family of distributions with a general parameter is denoted by $Y = (Y_i)_{i\in I}$. Let K be some n_m-dimensional measurable function of the individual observations Y_i. Let $k(\theta)$ be the expectation of K when K is evaluated on a draw Y_i from the distribution with parameter $\theta \in \Theta$, $k(\theta) := \mathrm{E}_\theta[K(Y_i)]$. Thus k is a vector of n_m generalized moments of the distribution of Y_i. (E_θ is the expectation under the distribution with parameter θ. Similarly, P_θ is the probability of an event in that case.)

Let g be some multidimensional function that represents estimating constraints. In our case these are distances between observed moments and moments of the model with given parameter value θ:

$$g(\mathcal{Y}_i, \theta) = K(\mathcal{Y}_i) - k(\theta).$$

By introducing E_0 as a shorthand for E_{θ_0}, it is immediate that $\mathrm{E}_0[g(\mathcal{Y}_i, \theta_0)] = 0$. However, for the parameter estimation problem to be well posed, we require that

$$\mathrm{E}_0[g(\mathcal{Y}_i, \theta)] = 0 \iff \theta = \theta_0. \tag{1}$$

We call this condition *identifiability* (with respect to the criterion function). Implicit in this is that we have at least as many independent equations as parameters. For the MSM to work, identifiability is required in a broader sense (cf. [24, Corollary 3.2]): for all $\delta > 0$,

$$\inf_{\theta\in\Theta:|\theta-\theta_0|>\delta} \left|\mathrm{E}_0[g(\mathcal{Y}_i, \theta)]\right| > 0. \tag{2}$$

In our particular problem we assume (1) but, as we show, (2) follows from it.

The MSM is used when $k(\theta)$ is not available in analytical form but there exists an unbiased estimator $\widetilde{k}(U_i^s, \theta)$, and consequently an unbiased estimator for g, $\widetilde{g}(\mathcal{Y}_i, U_i^s, \theta) = K(\mathcal{Y}_i) - \widetilde{k}(U_i^s, \theta)$. Here $(U_i^s)_{i\in I, s\in\{1,\dots,n_s\}}$ is some source of randomness, typically vectors of independent, uniform random variables on $[0, 1]$ as provided by a pseudorandom number generator. The estimators satisfy $\mathrm{E}[\widetilde{k}(U_i^s, \theta)] = k(\theta)$ and $\mathrm{E}[\widetilde{g}(\mathcal{Y}_i, U_i^s, \theta) \mid \mathcal{Y}_i] = g(\mathcal{Y}_i, \theta)$.

We introduce a weighting by a symmetric, positive definite matrix $\Omega \in \mathbb{R}^{n_m\times n_m}$, and consider the quadratic form $\alpha(\eta) = \eta^{\mathrm{T}}\Omega\eta$. $\Omega = \Omega_{n_I}$ might even be a measurable function of the data. The broad principle of the MSM is the following.

The principle of the MSM: Let $\mathcal{Y} = (\mathcal{Y}_i)_{i\in I}$ be a finite, independent random sample from P_{θ_0} and $(U_i^s)_{i\in I, s\in\{1,\dots,n_s\}}$ an independent random sample from a fixed distribution. The MSM estimator is defined as

$$\hat{\theta}_{n_s,n_I} := \operatorname*{arg\,min}_{\theta\in\Theta} \alpha\left(\frac{1}{n_I}\sum_{i=1}^{n_I}\left(K(\mathcal{Y}_i) - \frac{1}{n_s}\sum_{s=1}^{n_s}\widetilde{k}(U_i^s, \theta)\right)\right).$$

If identifiability in the broader sense (2) holds, n_s is fixed and n_I tends to infinity, and the almost sure convergence guaranteed by the SLLN

$$\frac{1}{n_I}\sum_{i=1}^{n_I}\widetilde{k}(U_i^s,\theta)\underset{n_I\to\infty}{\longrightarrow}k(\theta) \tag{3}$$

is uniform in $\theta\in\Theta$ for every s, then $\hat{\theta}_{n_s,n_I}$ is strongly consistent (that is, $\hat{\theta}_{n_s,n_I}$ converges to θ_0 almost surely). In the case when Ω_{n_I} is dependent on the data, it must satisfy a further technical condition related to identifiability.

There are different ways of specifying the technical conditions to turn this into a theorem [13,14,22,24]. For example, if $k(\theta)=E_\theta[K(Y_i)]$ is continuous in θ, then $\mathrm{E}_0[g(\mathcal{Y}_i,\theta)]$ is also continuous in θ. If additionally Θ is compact, then (1) implies (2).

Notice that the number of simulations n_s can remain bounded, it is only n_I that must tend to infinity for consistency. For practical implementations, it is a crucial point that the (U_i^s) must be drawn at the beginning of the exploration of the parameter space and kept fixed afterwards while different parameter values are proposed, in order to avoid introducing an extra layer of fluctuation [14, p. 29]. This way, a gradient-based search of the parameter space is possible. At the theoretical level, in the limit $n_I\to\infty$, the estimator is strongly consistent even without using common random numbers.

For any $\theta\in\Theta$, the only source of randomness in (3) is the shared collection $(U_i^s)_{i\in I,s\in\{1,\dots,n_s\}}$ of random variables, which are not influenced by the value of θ. Through them we can define the probability measures P_θ in a consistent way on the same probability space, enabling us to talk about uniform almost sure convergence in θ.

If Ω_{n_I} is a function of the data, then one needs to rule out pathologies where identifiability is compromised in the limit: when Ω_{n_I} converges to a positive semidefinite but not definite matrix, or when some directions picked out by Ω_{n_I} become overly dominant. We suggest two alternative conditions on $\Omega_{n_I}\in\mathbb{R}^{n_m\times n_m}$. The first condition is that Ω_{n_I} has a positive definite limit Ω', continuously in θ_0, almost surely:

$$\lim_{n_I\to\infty}\Omega_{n_I}=\Omega'>0. \tag{4}$$

In the numerical example of Section 6, we specify Ω_{n_I} using sample moments such that its limit is a diagonal matrix of generalized moments which are positive. An alternative condition is that there exist constants $0<c_1<c_2$ such that almost surely for all sufficiently large n_I, for all $\eta\in\mathbb{R}^{n_m}$ and all $\theta_0\in\Theta$,

$$c_1|\eta|^2\le\eta^{\mathrm{T}}\Omega_{n_I}\eta\le c_2|\eta|^2 \tag{5}$$

holds. If Θ is compact, then condition (4) implies condition (5).

Under the additional condition that $\widetilde{g}(\mathcal{Y}_i,U_i^s,\theta)$ is twice differentiable with respect to θ, asymptotic normality of the estimator also holds and the asymptotic variance can be explicitly given [13,14,22,24].

2.2. The MSM specialized to our problem

For the MSM applied to our percolation model with colouring, the data points $\widetilde{Y}_i$ are neither identically distributed (because of boundary effects) nor independent, and the general

principle of the MSM as currently stated in its conventional form does not imply the validity of the method. The main theoretical result of this paper is the proof of the strong consistency of a particular MSM estimator for our estimation problem.

The generalized moment function K we propose contains, in addition to first moments Y_i^ℓ, products $Y_i^\ell Y_j^\ell$ for $i \sim j$ because these carry much information about open edges. We mention the consequence that the original formalism of K as a function of a single variable (that is, of an individual Y_i) no longer applies.

We assume without proof that for this generalized moment function, identifiability (1) holds. For supporting evidence, turn to Section 1 of the Appendix. This assumption is not true in some extreme cases which we exclude. If $(\lambda^1, \dots, \lambda^{n_c}) = h \in \{0,1\}^{n_c}$, then $\widetilde{Y}$ is almost surely identically h for any choice of μ (and so is Y). For an $h \in \{0,1\}^{n_c}$, the outcome $\widetilde{Y}$ is again h with high probability as $n_I \to \infty$ if $\mu = 1$ and $h_\ell = \chi_{\{\lambda^\ell > 0\}}$. ($\chi$ is the indicator function.)

The percolation parameter μ is allowed to take any value in the subcritical regime $[0, p_c[$. p_c is the *critical probability* of bond percolation. For the triangular lattice, its value is $p_c = 2\sin(\pi/18) \approx 0.3473$, while for the square lattice, it is $p_c = 1/2$ [15, Chapter 3,19, 27,29].

Section 3 details the steps leading to the SLLN result (3). Due to dependence between the Y_i, cross-correlations appear in the derivation in addition to variances. Section 4 deals with upper bounding these correlations using percolation theory. Section 5 describes the extension of the SLLN to the ULLN.

The observed colouring of the dataset is denoted by $\mathcal{Y}_i^\ell$ ($i \in I$, $\ell \in \{1, 2, \dots, n_c\}$), whereas in the simulated data it is $\widetilde{Y}_i^{\ell,s}$ ($s \in \{1, 2, \dots, n_s\}$). While it is clear that the simulated data must come from a finite I (or perhaps from some $I' : I \subsetneq I' \subsetneq L$), we leave flexibility whether the data is of type $(\widetilde{\mathcal{Y}}_i)_{i\in I}$, which is the case in our practical application, or of the theoretically appealing type $(\mathcal{Y}_i)_{i\in I}$. We let $(\mathcal{Y}_i)_{i\in I}$ denote both cases, to be interpreted as the context demands. Simulating on $I' \supsetneq I$ is an attractive option when the dataset $(\mathcal{Y}_i)_{i\in I}$ is generated by a process defined on a proper superset of I, perhaps on the infinite L. The reason is that by simulating on I', seeds and open edges outside I can contribute to the realization on I, better reflecting the origin of the dataset. Lastly, we introduce the following averages:

$$\bar{\mathcal{Y}}^\ell := \frac{1}{n_I}\sum_{i\in I} \mathcal{Y}_i^\ell, \quad \bar{Y}^{\ell,s} := \frac{1}{n_I}\sum_{i\in I} \widetilde{Y}_i^{\ell,s},$$

$$\bar{\mathcal{Z}}^\ell := \frac{1}{n_p}\sum_{(i,j)\in I_2} \mathcal{Y}_i^\ell \mathcal{Y}_j^\ell, \quad \bar{Z}^{\ell,s} := \frac{1}{n_p}\sum_{(i,j)\in I_2} \widetilde{Y}_i^{\ell,s}\widetilde{Y}_j^{\ell,s}.$$

Our main theorem is the following.

Theorem 2.1: *Let Θ be a compact subset of $([0,1]^{n_c} \setminus \{0,1\}^{n_c}) \times [0, p_c[$. (For the triangular lattice, $p_c = 2\sin(\pi/18) \approx 0.3473$, while in the square lattice case, $p_c = 1/2$.) Consider the bond percolation model with colouring and with the true parameter value $\theta_0 = (\lambda^1, \dots, \lambda^{n_c}, \mu) \in \Theta$. Let $\Omega \in \mathbb{R}^{2n_c \times 2n_c}$ be a symmetric, positive definite matrix, and write $\alpha(\eta) = \eta^{\mathrm{T}}\Omega\eta$ for the resulting quadratic form. $\Omega = \Omega_{n_I}$ might be a measurable function of the data, but in this case, one of the almost sure conditions* (4) *or* (5) *must hold. Under*

the assumption of identifiability (1)*, when n_s is fixed and n_I tends to infinity, the estimator*

$$\hat{\theta}_{n_s,n_I} := \arg\min_{\theta\in\Theta} \alpha \left(\begin{array}{c} \left(\frac{1}{n_I} \sum_{i\in I} \left(\mathcal{Y}_i^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \widetilde{Y}_i^{\ell,s} \right) \right)_{\ell\in\{1,\dots,n_c\}} \\ \left(\frac{1}{n_p} \sum_{(i,j)\in I_2} \left(\mathcal{Y}_i^\ell \mathcal{Y}_j^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \widetilde{Y}_i^{\ell,s} \widetilde{Y}_j^{\ell,s} \right) \right)_{\ell\in\{1,\dots,n_c\}} \end{array} \right)$$

$$= \arg\min_{\theta\in\Theta} \alpha \left(\begin{array}{c} \left(\bar{\mathcal{Y}}^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \bar{Y}^{\ell,s} \right)_{\ell\in\{1,\dots,n_c\}} \\ \left(\bar{\mathcal{Z}}^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \bar{Z}^{\ell,s} \right)_{\ell\in\{1,\dots,n_c\}} \end{array} \right)$$

is strongly consistent.

In order to prove the claim, we want to establish that for the arithmetic means generated under general θ, the following almost sure convergences hold as $n_I \to \infty$, uniformly in $\theta \in \Theta$:

$$\frac{1}{n_I} \sum_{i\in I} Y_i^\ell - \frac{1}{n_I} \sum_{i\in I} \mathrm{E}_\theta\, Y_i^\ell \longrightarrow 0$$

$$\text{and} \quad \frac{1}{n_p} \sum_{(i,j)\in I_2} Y_i^\ell Y_j^\ell - \frac{1}{n_p} \sum_{(i,j)\in I_2} \mathrm{E}_\theta \left[Y_i^\ell Y_j^\ell \right] \longrightarrow 0.$$

The relevant notion of uniform convergence is detailed in Section 5. The same proofs apply with $\widetilde{Y}$, too. This unusual formulation of the SLLN is needed because the random variables $\widetilde{Y}$ are not identically distributed due to boundary effects. At the end of Section 5, we will also see that these two SLLNs ultimately ensure that

$$\alpha \left(\begin{array}{c} \left(\bar{\mathcal{Y}}^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \bar{Y}^{\ell,s} \right)_{\ell\in\{1,\dots,n_c\}} \\ \left(\bar{\mathcal{Z}}^\ell - \frac{1}{n_s} \sum_{s=1}^{n_s} \bar{Z}^{\ell,s} \right)_{\ell\in\{1,\dots,n_c\}} \end{array} \right)$$

$$- \alpha \left(\begin{array}{c} \left(\frac{1}{n_I} \sum_{i\in I} \left(\mathrm{E}_0 Y_i^\ell - \mathrm{E}_\theta \widetilde{Y}_i^\ell \right) \right)_{\ell\in\{1,\dots,n_c\}} \\ \left(\frac{1}{n_p} \sum_{(i,j)\in I_2} \left(\mathrm{E}_0[Y_i^\ell Y_j^\ell] - \mathrm{E}_\theta[\widetilde{Y}_i^\ell \widetilde{Y}_j^\ell] \right) \right)_{\ell\in\{1,\dots,n_c\}} \end{array} \right) \underset{n_I\to\infty}{\longrightarrow} 0 \tag{6}$$

uniformly with probability 1. The second term is minimal when it is asymptotically zero (in the case when E_0 acts on Y_i^ℓ and $Y_i^\ell Y_j^\ell$; when it acts on $\widetilde{Y}_i^\ell$ and $\widetilde{Y}_i^\ell \widetilde{Y}_j^\ell$, then it is actually zero), and this is achieved only when $\theta = \theta_0$ under the assumption of identifiability (1). This gives the strong consistency for $\hat{\theta}_{n_s,n_I}$. The details are to be found in Section 5.

3. Strong law of large numbers with weak dependence

We adapt the proof of Theorem 1 of [9] in this section to suit our purposes. We write out the claims with Y, but they also hold for $\widetilde{Y}$ and the necessary proof updates are presented in Section 4.

Proposition 3.1: *Let $\theta \in [0,1]^{n_c} \times [0, p_c[$, where p_c is the critical probability of bond percolation. If Y is generated with parameter value θ, then*

$$\frac{1}{n_I}\left(\sum_{i\in I} Y_i^\ell - \sum_{i\in I} \mathrm{E}_\theta\, Y_i^\ell\right) \underset{n_I\to\infty}{\longrightarrow} 0$$

almost surely. The claim also holds for $\widetilde{Y}$.

Proposition 3.2: *Let $\theta \in [0,1]^{n_c} \times [0, p_c[$. If Y is generated with parameter value θ, then*

$$\frac{1}{n_p}\left(\sum_{(i,j)\in I_2} Y_i^\ell Y_j^\ell - \sum_{(i,j)\in I_2} \mathrm{E}_\theta[Y_i^\ell Y_j^\ell]\right) \underset{n_I\to\infty}{\longrightarrow} 0$$

almost surely. The claim also holds for $\widetilde{Y}$.

Proof of Proposition 3.1.: For the ease of notation, let $Y_i := Y_i^\ell$ for some fixed $\ell \in \{1,\dots,n_c\}$ $(i \in I)$, created by our percolation process with $\theta = (\lambda^1,\dots,\lambda^{n_c},\mu)$. Let $a > 1$ and define the lacunary sequence $k_n := [a^n]$. Let $S_k := \sum_{i=1}^k Y_i$.

By the application of Chebyshov's inequality, for every $\varepsilon > 0$,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathrm{P}\left(\left|\frac{S_{k_n} - \mathrm{E}S_{k_n}}{k_n}\right| > \varepsilon\right) &\le \sum_{n=1}^{\infty} \frac{\mathrm{Var}\, S_{k_n}}{\varepsilon^2 k_n^2} \\ &\le \frac{1}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{1}{k_n^2}\sum_{i=1}^{k_n}\mathrm{Var}\, Y_i \\ &\quad + \frac{1}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{1}{k_n^2}\sum_{1\le i\ne j\le k_n}(\mathrm{E}[Y_iY_j] - \mathrm{E}Y_i\,\mathrm{E}Y_j). \qquad (7)\end{aligned}$$

If we can prove that this is finite, then by the Borel–Cantelli lemma, as $n \to \infty$, for every $\theta \in \Theta$,

$$\left|\frac{S_{k_n} - \mathrm{E}S_{k_n}}{k_n}\right| \to 0 \text{ a.s.} \qquad (8)$$

We first notice that

$$\sum_{n=1}^{\infty}\frac{1}{k_n^2}\sum_{i=1}^{k_n}\mathrm{Var}\, Y_i < \infty$$

by recognizing that $\sup_{i\in I}\mathrm{Var}\, Y_i \le 1$, and that for $a > 1$,

$$\sum_{n=1}^{\infty}\frac{1}{k_n} < \infty. \qquad (9)$$

The upper bound (9) follows by the *limit comparison test* with $\sum_n (1/a^n)$ and $\sum_n (1/k_n)$, or by the *ratio test* for $\sum_n (1/k_n)$.

We prove in Section 4 that

$$\left| \sum_{1 \le i \ne j \le k_n} (\mathrm{E}[Y_i Y_j] - \mathrm{E}Y_i \, \mathrm{E}Y_j) \right| = \mathcal{O}(k_n), \tag{10}$$

so that by applying convergence (9) once again, we get that (7) is finite, as required.

In the case of a general $k := n_I$, k is sandwiched between some $k_n \le k < k_{n+1}$ and

$$\begin{aligned} \frac{S_k - \mathrm{E}S_k}{k} &\le \frac{S_{k_{n+1}} - \mathrm{E}S_{k_n}}{k} \\ &\le \left| \frac{S_{k_{n+1}} - \mathrm{E}S_{k_{n+1}}}{k_{n+1}} \right| \frac{k_{n+1}}{k_n} + \frac{\mathrm{E}S_{k_{n+1}} - \mathrm{E}S_{k_n}}{k_n}. \end{aligned} \tag{11}$$

Note that even for $S_{k_{n+1}} - \mathrm{E}S_{k_n} < 0$, one can change the denominator from k to k_n in the second inequality because the right-hand side is nonnegative. Here, for a fixed $a > 1$,

$$\frac{k_{n+1}}{k_n} = \frac{[a^{n+1}]}{[a^n]} \le \frac{a^{n+1}}{a^n - 1} = a + \frac{a}{a^n - 1}, \tag{12}$$

which in turn is arbitrarily close to a when n is sufficiently large. (An analogous lower bound can also be derived, giving $k_{n+1}/k_n = a + o(1)$.) Additionally,

$$\begin{aligned} \frac{\mathrm{E}S_{k_{n+1}} - \mathrm{E}S_{k_n}}{k_n} &\le \frac{(k_{n+1} - k_n) \sup_{i \in I} \mathrm{E}Y_i}{k_n} \\ &\le \left(a + \frac{a}{a^n - 1} - 1 \right) \sup_{i \in I} \mathrm{E}Y_i, \end{aligned}$$

and combining this with (8) yields

$$\limsup_{k \to \infty} \frac{S_k - \mathrm{E}S_k}{k} \le (a - 1) \sup_{i \in I} \mathrm{E}Y_i \le a - 1.$$

A similar lower bound can also be obtained. Since $a > 1$ can be chosen arbitrarily, the SLLN for Y_i (Proposition 3.1) holds once we prove the estimate (10). ■

Proof of Proposition 3.2.: This proof goes entirely analogously to that of Proposition 3.1. We keep using the notation $Y_i := Y_i^\ell$ for some fixed $\ell \in \{1, \ldots, n_c\}$ and fixed θ, and the lacunary sequence $k_n = [a^n]$ for $a > 1$. Let $T_k := \sum_{(i,j) \in I_2} Y_i Y_j$ for $I = I(k)$ composed of the first k vertices according to the fixed ordering. This sum has $n_p(k)$ terms. Then, by the

argument of (7), for every $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\left|\frac{T_{k_n} - \mathrm{E}T_{k_n}}{n_p(k_n)}\right| > \varepsilon\right) \le \sum_{n=1}^{\infty} \frac{\operatorname{Var} T_{k_n}}{\varepsilon^2 n_p(k_n)^2}$$
$$\le \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n_p(k_n)^2} \sum_{(i_1,i_2)\in I_2(k_n)} \operatorname{Var}[Y_{i_1} Y_{i_2}]$$
$$+ \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n_p(k_n)^2} \sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(k_n)\\ (i_1,i_2)\neq(j_1,j_2)}} \Big(\mathrm{E}[Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2}]$$
$$- \mathrm{E}[Y_{i_1} Y_{i_2}]\,\mathrm{E}[Y_{j_1} Y_{j_2}]\Big). \tag{13}$$

By

$$\sup_{(i_1,i_2)\in I_2(k_n)} \operatorname{Var}[Y_{i_1} Y_{i_2}] \le 1$$

and $|I_2(k_n)| = n_p(k_n) \sim 3n_I = 3k_n$, the convergence (9) gives

$$\frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n_p(k_n)^2} \sum_{(i_1,i_2)\in I_2(k_n)} \operatorname{Var}[Y_{i_1} Y_{i_2}] \le \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n_p(k_n)} < \infty.$$

In Section 4, it is shown that

$$\left| \sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(k_n)\\ (i_1,i_2)\neq(j_1,j_2)}} \Big(\mathrm{E}[Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2}] - \mathrm{E}[Y_{i_1} Y_{i_2}]\,\mathrm{E}[Y_{j_1} Y_{j_2}]\Big) \right| = \mathcal{O}(k_n), \tag{14}$$

and by (9), we get that the sum (13) is finite. By the Borel–Cantelli lemma,

$$\left|\frac{T_{k_n} - \mathrm{E}T_{k_n}}{n_p(k_n)}\right| \to 0 \text{ a.s.} \tag{15}$$

For a general $k = n_I$ with $k_n \le k < k_{n+1}$,

$$\frac{T_k - \mathrm{E}T_k}{n_p(k)} \le \left|\frac{T_{k_{n+1}} - \mathrm{E}T_{k_{n+1}}}{n_p(k_{n+1})}\right| \frac{n_p(k_{n+1})}{n_p(k_n)} + \frac{\mathrm{E}T_{k_{n+1}} - \mathrm{E}T_{k_n}}{n_p(k_n)}. \tag{16}$$

For a fixed $a > 1$, by using $n_p(k_n) \sim 3k_n$ and the remark after (12),

$$\frac{n_p(k_{n+1})}{n_p(k_n)} \sim \frac{3k_{n+1}}{3k_n} = a + o(1).$$

Additionally,

$$\frac{\mathrm{E}T_{k_{n+1}} - \mathrm{E}T_{k_n}}{n_p(k_n)} \le \frac{(n_p(k_{n+1}) - n_p(k_n)) \sup_{(i_1,i_2)\in I_2(k_n)} \mathrm{E}[Y_{i_1} Y_{i_2}]}{n_p(k_n)},$$

hence

$$\limsup_{n\to\infty} \frac{\mathrm{E}T_{k_{n+1}} - \mathrm{E}T_{k_n}}{n_p(k_n)} \le (a-1) \sup_{(i_1,i_2)\in I_2(k_n)} \mathrm{E}[Y_{i_1} Y_{i_2}].$$

Combining this with (15) and (16), we get

$$\limsup_{k\to\infty} \frac{T_k - \mathrm{E}T_k}{n_p(k)} \le (a-1) \sup_{(i_1,i_2)\in I_2(k_n)} \mathrm{E}[Y_{i_1} Y_{i_2}] \le a-1.$$

A similar lower bound can also be obtained. Since $a > 1$ can be chosen arbitrarily, the SLLN for $Y_i Y_j$, $i \sim j$ (Proposition 3.2) holds once we prove the estimate (14). ■

4. Upper bound on correlations

We prove the estimates (10) and (14) in greater generality, for every positive integer n. Let Θ be a compact subset of $[0,1] \times [0, p_c[$, where p_c is the critical probability of bond percolation.

Lemma 4.1: *As $n \to \infty$, we have*

$$\sup_{\theta\in\Theta} \left| \sum_{1\le i\ne j\le n} (\mathrm{E}[Y_i Y_j] - \mathrm{E}Y_i \, \mathrm{E}Y_j) \right| = \mathcal{O}(n).$$

Lemma 4.2: *As $n \to \infty$, we have*

$$\sup_{\theta\in\Theta} \left| \sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n) \\ (i_1,i_2)\ne(j_1,j_2)}} \Big(\mathrm{E}[Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2}] - \mathrm{E}[Y_{i_1} Y_{i_2}] \, \mathrm{E}[Y_{j_1} Y_{j_2}] \Big) \right| = \mathcal{O}(n).$$

For background, first we recapitulate from the fundamentals of percolation theory the meaning of increasing events, the FKG inequality, disjoint occurrence, the BK inequality and pivotality [15, Chapter 2]. It is well known that these concepts do not rely on the specific structure of the lattice graph and can be cast more generally in terms of functions of Boolean variables.

In this vein, one can consider a probability space $(\Gamma, \mathcal{F}, \mathrm{P})$ with sample space $\Gamma = \{0,1\}^S$ (S is finite or at most countably infinite) where the set of events $\mathcal{F}$ is the σ-algebra generated by the finite-dimensional cylinder sets and the measure is a product measure

$$\mathrm{P} = \prod_{s\in S} \nu_s$$

where ν_s is specified by some vector $(p(s))_{s\in S} \in [0,1]^S$ via

$$\nu_s(\omega(s) = 1) = p(s), \quad \nu_s(\omega(s) = 0) = 1 - p(s)$$

for sample vectors $(\omega(s))_{s\in S} \in \{0,1\}^S$ [15, Chapter 2, p. 33].

In our application, we have already fixed a colour $\ell \in \{1, \dots, n_c\}$ and look at colours independently. We extend the set of vertices L with an additional vertex that we call ∞^ℓ, or simply ∞ when the colour is fixed and unimportant: $L^* := L \cup \{\infty\}$. We also extend the edge set of the triangular lattice, L_2, with edges between each vertex and ∞, and the value assigned to such an edge indicates the presence or absence of seeding. We call these edges *source edges*. For the source edges, $p(s) = \lambda^\ell$, and for the edges of the lattice which represent contamination, $p(s) = \mu$. The interpretation is that $Y_i^\ell = 1$ if and only if $i \leftrightarrow^* \infty^\ell$, where the asterisk refers to connection in the extended graph.

An event $A \in \mathcal{F}$ of the σ-algebra is called *increasing*, if whenever $\omega \leq \omega', \omega \in A$ implies $\omega' \in A$.

Theorem 4.3 (FKG inequality [10,15, pp. 34–36,16]): *If A and B are increasing events in $\mathcal{F}$, then* $\mathrm{P}(A \cap B) \geq \mathrm{P}(A)\mathrm{P}(B)$.

Let $e_1, e_2, \dots, e_N$ be N distinct edges of the graph, and $A, B \in \mathcal{F}$ two increasing events which depend on the vector of the states of these N edges $\omega = (\omega(e_1), \dots, \omega(e_N))$ only. Such vectors ω are characterized uniquely by the set of edges with value 1: $J(\omega) = \{e_i \mid i \in \{1, \dots, N\}, \omega(e_i) = 1\}$.

For the increasing events A,B, the event $A \circ B$ (we say *A and B occur disjointly*) is the set of all $\omega \in \Gamma$ for which there exists an $H := H(\omega) \subseteq J(\omega)$ such that ω' determined by $J(\omega') = H$ belongs to A, and ω'' determined by $J(\omega'') = J(\omega) \setminus H$ belongs to B. In words, $A \circ B$ is the set of assignments of 0 and 1 to the edges for which there exist two disjoint sets of edges assigned the value 1 (*open* edges) such that the first such set ensures the occurrence of event A and the second set ensures the occurrence of B. It is easy to verify that $A \circ B$ is also increasing and $A \circ B \subseteq A \cap B$.

The classical example for disjoint occurrence is when A is the event that there is an open path joining i_1 to j_1 within the finite subgraph given by $\{e_1, \dots, e_N\}$ and B is the event that there is an open path between i_2 and j_2 within the same finite subgraph. Then $A \circ B$ is the event that there exist two edge-disjoint paths, the first between i_1 and j_1 and another one joining i_2 to j_2.

Theorem 4.4 (BK inequality [5,15, pp. 37–41]): *If A and B are increasing events in $\mathcal{F}$, then* $\mathrm{P}(A \circ B) \leq \mathrm{P}(A)\mathrm{P}(B)$.

The validity of the inequality extends to the existence of arbitrarily long (but finite-length) edge-disjoint open paths, which is what we need it for, by taking a sequence of growing, nested subsets of L [15, p. 38].

The notion of pivotality is not used until Section 5. For any event A an edge e is pivotal if its open or closed state is crucial to whether A occurs or not. In more detail, the edge e is pivotal for the pair (A, ω) if for the indicator function χ_A of A, $\chi_A(\omega) \neq \chi_A(\omega')$, where the configuration $\omega' \in \{0, 1\}^S$ is defined by $\omega'(e) = 1 - \omega(e)$, and $\omega'(f) = \omega(f)$ for every edge $f \neq e$. The event that *e is pivotal for A* is the set of ω for which e is pivotal for (A, ω).

Proof of Lemma 4.1.: In the extended lattice graph that has source edges with weight zero or one at every vertex for seeding, the event $\{Y_i = 1\}$ for $i \in I$ is increasing because it is

increasing in both seeding (source edges) and contamination edges. For any $i, j \in I$,

$$\mathrm{E}[Y_i Y_j] - \mathrm{E}Y_i \mathrm{E}Y_j = \mathrm{P}(Y_i Y_j = 1) - \mathrm{P}(Y_i = 1)\mathrm{P}(Y_j = 1) \geq 0$$

by the FKG inequality. Hence, for every $\theta \in \Theta$,

$$\sum_{1 \leq i \neq j \leq n} (\mathrm{E}[Y_i Y_j] - \mathrm{E}Y_i \, \mathrm{E}Y_j) \geq 0.$$

For the upper bound, consider that

$$\begin{aligned}
&\mathrm{P}(Y_i Y_j = 1) - \mathrm{P}(Y_i = 1)\,\mathrm{P}(Y_j = 1) \\
&\quad = \mathrm{P}\big(\{Y_i = 1\} \circ \{Y_j = 1\}\big) - \mathrm{P}(Y_i = 1)\,\mathrm{P}(Y_j = 1) \\
&\qquad + \mathrm{P}\Big(\{Y_i Y_j = 1\} \setminus \{Y_i = 1\} \circ \{Y_j = 1\}\Big) \\
&\quad \leq \mathrm{P}\Big(\{Y_i Y_j = 1\} \setminus \{Y_i = 1\} \circ \{Y_j = 1\}\Big) \qquad (17)
\end{aligned}$$

by the BK inequality. Cooccurrence of $\{Y_i = 1\}$ and $\{Y_j = 1\}$ which is not disjoint is one where i and j are in the same component in the edge set on the non-extended lattice:

$$\{Y_i Y_j = 1\} \setminus \{Y_i = 1\} \circ \{Y_j = 1\} \subseteq \{i \leftrightarrow j\}.$$

We show that

$$\sum_{1 \leq i \neq j \leq n} \mathrm{P}(i \leftrightarrow j) = \mathcal{O}(n) \qquad (18)$$

for $\mu < p_c$, and uniformly so for $\mu \in [0, p_c - \varepsilon]$ for every $\varepsilon \in]0, p_c[$. This follows from the exponential decay of the cluster size distribution and it will complete the proof of Lemma 4.1.

Let $C(i)$ denote the set of vertices in the component of $i \in L$ according to the non-extended edge set of L. Then

$$\sum_{1 \leq i \neq j \leq n} \mathrm{P}(i \leftrightarrow j) = \sum_{1 \leq i \leq n} \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \mathrm{E}\chi_{\{i \leftrightarrow j\}} = \sum_{1 \leq i \leq n} \mathrm{E}[|C(i)| - 1].$$

Theorem 4.5 (Exponential decay of the cluster size distribution [1,15, Theorem 6.75] and [20]): *For $\mu \in]0, p_c[$, there exists $g(\mu) > 0$ such that for all $k \geq 1$ and $i \in I$, for the bond percolation with parameter μ, it holds that* $\mathrm{P}(|C(i)| \geq k) \leq \mathrm{e}^{-kg(\mu)}$.

Take $\mu^* = p_c - \varepsilon$. As $\mathrm{P}(|C(i)| \geq k)$ is nondecreasing in μ, we get a uniform bound in $\theta \in \Theta$ if the bound is valid for μ^*:

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{E}[|C(i)| - 1] &= \sum_{i=1}^{n} \left(\left(\sum_{k=1}^{\infty} P(|C(i)| \geq k) \right) - 1 \right) \\
&\leq \sum_{i=1}^{n} \sum_{k=1}^{\infty} \mathrm{e}^{-kg(\mu^*)} = n \frac{1}{\mathrm{e}^{g(\mu^*)} - 1}. \qquad (19)
\end{aligned}$$

This proves Lemma 4.1, which in turn completes the proof of Proposition 3.1 for Y. ■

Proof of Proposition 3.1 for the case of $\widetilde{Y}$.: To go from the case of Y to $\widetilde{Y}$, first we couple the realizations of $(Y_i)_{i\in L}$ and $(\widetilde{Y}_i)_{i\in I}$ with varying lattices I (and later with varying parameter vectors) by defining them via shared random variables $(U_i^\ell)_{i\in L,\ell\in\{1,2,\dots,n_c\}}$ and $(V_{ij})_{(i,j)\in L_2}$ that are independent and all uniformly distributed on $[0,1]$. For $\theta = (\lambda^1,\dots,\lambda^{n_c},\mu)\in\Theta$, $i\in L$, $(i,j)\in L_2$ and $\ell\in\{1,2,\dots,n_c\}$, the seeding is defined by $X_i^\ell := \chi_{\{U_i^\ell<\lambda^\ell\}}$, and edges are open according to $\xi_{ij} := \chi_{\{V_{ij}<\mu\}}$.

Let us drop the superscript ℓ again. Notice that any $\widetilde{Y}_i$ can change in a nondecreasing manner when I is increased. In the proof of Proposition 3.1, the only occasion where Y_i from different I are compared is inequality (11). We mark the lattice size as a variable in the superscript of $\widetilde{Y}_i$. Observe that $\widetilde{Y}_i^{k_n}\le\widetilde{Y}_i^{k}\le\widetilde{Y}_i^{k_{n+1}}\le Y_i$ for $k_n\le k<k_{n+1}$ and $i\in I(k_n)$. With $\widetilde{S}_n^k := \sum_{i=1}^n \widetilde{Y}_i^k$, noting $\widetilde{S}_{k_n}^{k_n}\le\widetilde{S}_k^k\le\widetilde{S}_{k_{n+1}}^{k_{n+1}}$,

$$\begin{aligned}\frac{\widetilde{S}_k^k-\mathrm{E}\widetilde{S}_k^k}{k} &\le \frac{\widetilde{S}_{k_{n+1}}^{k_{n+1}}-\mathrm{E}\widetilde{S}_{k_n}^{k_n}}{k}\\ &\le \left|\frac{\widetilde{S}_{k_{n+1}}^{k_{n+1}}-\mathrm{E}\widetilde{S}_{k_{n+1}}^{k_{n+1}}}{k_{n+1}}\right|\frac{k_{n+1}}{k_n}+\frac{\mathrm{E}\widetilde{S}_{k_{n+1}}^{k_{n+1}}-\mathrm{E}\widetilde{S}_{k_n}^{k_n}}{k_n},\end{aligned}$$

where, similarly to inequality (11), the second inequality holds for different reasons when $\widetilde{S}_{k_{n+1}}^{k_{n+1}}-\mathrm{E}\widetilde{S}_{k_n}^{k_n}$ is negative and when not. The first term on the right-hand side is treated as in the original proof of Proposition 3.1. The second term is

$$\frac{\mathrm{E}\widetilde{S}_{k_{n+1}}^{k_{n+1}}-\mathrm{E}\widetilde{S}_{k_n}^{k_n}}{k_n} = \frac{1}{k_n}\sum_{i=k_n+1}^{k_{n+1}}\mathrm{E}\widetilde{Y}_i^{k_{n+1}}+\frac{1}{k_n}\sum_{i=1}^{k_n}\left(\mathrm{E}\widetilde{Y}_i^{k_{n+1}}-\mathrm{E}\widetilde{Y}_i^{k_n}\right). \tag{20}$$

Here

$$\frac{1}{k_n}\sum_{i=k_n+1}^{k_{n+1}}\mathrm{E}\widetilde{Y}_i^{k_{n+1}}\le\frac{k_{n+1}-k_n}{k_n}\sup_{i\in I(k_{n+1})}\mathrm{E}\widetilde{Y}_i^{k_{n+1}}$$

is dealt with as in the original proof of Proposition 3.1. For the other term of (20),

$$\frac{1}{k_n}\sum_{i=1}^{k_n}\left(\mathrm{E}\widetilde{Y}_i^{k_{n+1}}-\mathrm{E}\widetilde{Y}_i^{k_n}\right)\le\frac{1}{k_n}\sum_{i=1}^{k_n}\left(\mathrm{E}Y_i-\mathrm{E}\widetilde{Y}_i^{k_n}\right).$$

According to the next proposition, this vanishes in the limit, leaving us with

$$\limsup_{k\to\infty}\frac{\widetilde{S}_k^k-\mathrm{E}\widetilde{S}_k^k}{k}\le a-1,$$

as required.

Proposition 4.6: *For a compact subset $\Theta\subset[0,1]\times[0,p_c[$,*

$$\sup_{\theta\in\Theta}\frac{1}{n_I}\left|\sum_{i\in I}\mathrm{E}_\theta\widetilde{Y}_i-\sum_{i\in I}\mathrm{E}_\theta Y_i\right|\underset{n_I\to\infty}{\longrightarrow}0.$$

Proof: As $Y_i \geq \widetilde{Y}_i$ almost surely,

$$\begin{aligned}\mathrm{E}\left[Y_i - \widetilde{Y}_i\right] &= \mathrm{P}\left(Y_i = 1, \widetilde{Y}_i = 0\right) \\ &\leq \mathrm{P}\left(Y_i = 1 \text{ and } \exists j \in L \setminus I : X_j = 1,\ i \leftrightarrow j\right) \\ &\leq \mathrm{P}\left(i \leftrightarrow \Delta I\right),\end{aligned}$$

which expresses that Y_i and $\widetilde{Y}_i$ can differ only if $i \in I$ is connected to the exterior vertex boundary of I. Further,

$$\sum_{i \in I} \mathrm{P}\left(i \leftrightarrow \Delta I\right) = \mathrm{E}\left[\sum_{i \in I} \chi_{\{i \leftrightarrow \Delta I\}}\right].$$

But this is the expected size of the open component that is grown from all vertices of ΔI towards the inside of I. It is bounded from above by $|\Delta I| \times \mathrm{E}[|C(0)|]$. On the compact Θ, the mean size of the open component of any vertex has a universal finite upper bound by (19). Hence,

$$\sup_{\theta \in \Theta} \frac{1}{n_I} \left| \sum_{i \in I} \mathrm{E}_\theta \left[\widetilde{Y}_i - Y_i\right] \right| \leq \frac{|\Delta I|}{n_I} \sup_{\theta \in \Theta} \mathrm{E}_\theta[|C(0)|] \to 0$$

as $n_I \to \infty$, due to our assumption $|\Delta I|/|I| \to 0$ about the nested sequence of I. ■

As in the case of Y, a lower bound for $(\widetilde{S}_k^k - \mathrm{E}\widetilde{S}_k^k)/k$ does not pose any additional difficulty. In the proof of Lemma 4.1, the covariances cannot increase when we constrain the set of edges to those among the first n vertices. Concretely, $\mathrm{E}[|C(i)|]$ cannot increase. Therefore Proposition 3.1 stays true for $\widetilde{Y}$. ■

Proof of Lemma 4.2.: The proof follows closely that of Lemma 4.1. For any two pairs $(i_1, i_2), (j_1, j_2) \in I_2$,

$$\mathrm{E}[Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2}] - \mathrm{E}[Y_{i_1} Y_{i_2}]\,\mathrm{E}[Y_{j_1} Y_{j_2}] \geq 0$$

due to the FKG inequality applied to $\{Y_{i_1} Y_{i_2} = 1\}$ and $\{Y_{j_1} Y_{j_2} = 1\}$. Therefore, for any $\theta \in \Theta$,

$$\sum_{\substack{(i_1,i_2),(j_1,j_2) \in I_2(n) \\ (i_1,i_2) \neq (j_1,j_2)}} \Big(\mathrm{E}[Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2}] - \mathrm{E}[Y_{i_1} Y_{i_2}]\,\mathrm{E}[Y_{j_1} Y_{j_2}]\Big) \geq 0.$$

The first step towards the upper bound, similarly to (17), uses the BK inequality:

$$\begin{aligned}&\mathrm{P}(Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2} = 1) - \mathrm{P}(Y_{i_1} Y_{i_2} = 1)\,\mathrm{P}(Y_{j_1} Y_{j_2} = 1) \\ &\quad \leq \mathrm{P}\Big(\{Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2} = 1\} \setminus \{Y_{i_1} Y_{i_2} = 1\} \circ \{Y_{j_1} Y_{j_2} = 1\}\Big).\end{aligned}$$

Cooccurrence which is not disjoint is one where at least one of i_1 and i_2 is connected to at least one of j_1 and j_2 in the non-extended edge set, or in symbols for the case $i_1, i_2 \notin \{j_1, j_2\}$,

$$\begin{aligned}&\{Y_{i_1} Y_{i_2} Y_{j_1} Y_{j_2} = 1\} \setminus \{Y_{i_1} Y_{i_2} = 1\} \circ \{Y_{j_1} Y_{j_2} = 1\} \\ &\quad \subseteq \{i_1 \leftrightarrow j_1\} \cup \{i_1 \leftrightarrow j_2\} \cup \{i_2 \leftrightarrow j_1\} \cup \{i_2 \leftrightarrow j_2\}.\end{aligned}$$

So for every fixed $\theta \in \Theta$,

$$\begin{aligned}
&\sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\(i_1,i_2)\neq(j_1,j_2)}} \Big(\mathrm{E}[Y_{i_1}Y_{i_2}Y_{j_1}Y_{j_2}] - \mathrm{E}[Y_{i_1}Y_{i_2}]\,\mathrm{E}[Y_{j_1}Y_{j_2}]\Big)\\
&\quad\le \sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\ i_1,i_2\notin\{j_1,j_2\}}} \Big(\mathrm{P}\big(i_1\leftrightarrow j_1\big)+\mathrm{P}\big(i_1\leftrightarrow j_2\big)+\mathrm{P}\big(i_2\leftrightarrow j_1\big)+\mathrm{P}\big(i_2\leftrightarrow j_2\big)\Big)\\
&\qquad+\sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\ i_1\in\{j_1,j_2\}}} 1+\sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\ i_2\in\{j_1,j_2\}}} 1 \qquad (21)
\end{aligned}$$

The second and third terms are easier, and we show the estimate for the second only.

$$\begin{aligned}
&\sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\ i_1\in\{j_1,j_2\}}} 1\\
&\quad=\sum_{(j_1,j_2)\in I_2(n)}\ \sum_{\substack{(i_1,i_2)\in I_2(n)\\ i_1\in\{j_1,j_2\}}} 1\\
&\quad\le \sum_{j_1\in I(n)}\sum_{j_2\sim j_1}\left(\sum_{i_2\sim i_1=j_1} 1+\sum_{i_2\sim i_1=j_2} 1\right)\\
&\quad\le n\,6\,(6+6)=\mathcal{O}(n),
\end{aligned}$$

where in the first inequality, instead of (j_1, j_2), we sweep for j_1, and then for its neighbours j_2 separately, and similarly for (i_1, i_2). In the last inequality, we replace the sums for $j_2 \sim j_1$, $i_2 \sim j_1$ and $i_2 \sim j_2$ by factors of 6 (for the triangular lattice). Clearly, the goal with the first sum on the right-hand side of (21) is to reduce it to an already tackled sum. For example, for the first term in the sum,

$$\begin{aligned}
&\sum_{\substack{(i_1,i_2),(j_1,j_2)\in I_2(n)\\ i_1,i_2\notin\{j_1,j_2\}}} \mathrm{P}\big(i_1\leftrightarrow j_1\big)\\
&\quad\le \sum_{j_1\in I(n)}\sum_{j_2\sim j_1}\sum_{i_1\in I(n)\setminus\{j_1,j_2\}}\sum_{i_2\sim i_1}\mathrm{P}\big(i_1\leftrightarrow j_1\big)\\
&\quad\le 36\sum_{j_1\in I(n)}\sum_{i_1\in I(n)\setminus\{j_1\}}\mathrm{P}\big(i_1\leftrightarrow j_1\big)\\
&\quad\le 36\,n\,\frac{1}{\mathrm{e}^{g(\mu^*)}-1}=\mathcal{O}(n).
\end{aligned}$$

In the first two inequalities, we use manipulations from above and extend $i_1 \in I(n) \setminus \{j_1, j_2\}$ to $i_1 \in I(n) \setminus \{j_1\}$. In the last inequality, we use the previous case (19). This completes the proof of Lemma 4.2 and the proof of Proposition 3.2 for Y. ■

Proof of Proposition 3.2 for the case of $\widetilde{Y}$.: The proof of Proposition 3.2 can be adapted to $\widetilde{Y}$. For example, the following variant of Proposition 4.6 also holds:

$$\sup_{\theta\in\Theta}\frac{1}{n_p}\left|\sum_{(i,j)\in I_2}\mathrm{E}_\theta\left[\widetilde{Y}_i\widetilde{Y}_j\right]-\sum_{(i,j)\in I_2}\mathrm{E}_\theta\left[Y_iY_j\right]\right|\underset{n_I\to\infty}{\longrightarrow}0. \tag{22}$$

■

5. Uniform law of large numbers (ULLN) for our process

In the interests of conciseness, we continue assuming that there is only one colour: $n_c = 1$. This leads to no loss of generality. We prove that the SLLNs, Propositions 3.1 and 3.2, hold uniformly over the compact parameter set $\Theta \subset [0,1] \times [0, p_c[$. Similarly to the preceding, we write out everything for Y, but the result is also valid for $\widetilde{Y}$.

To prove the uniform version of Proposition 3.1, we check that the conditions of the following theorem hold, where we adapted [25, p. 8, Theorem 2] or [28, p. 25, Lemma 3.1] to our setting. For the rewriting of the theorem, we exploited that for a sample $((U_i)_{i\in L}, (V_{ij})_{(i,j)\in L_2})$ of the seeds and edges, any Y_i is nondecreasing in both λ and μ.

Theorem 5.1 (cf. [25, p. 8, Theorem 2], [28, p. 25, Lemma 3.1]): *Suppose that for every $\varepsilon > 0$ there exists a finite set of pairs of parameter vectors*

$$\mathcal{P}=\left\{\left(\theta_r^L,\theta_r^U\right)\in\left([0,1]\times[0,p_c[\right)^2\;\middle|\;r\in\{1,\dots,N(\varepsilon)\}\right\}$$

such that

(1) *for every $r \in \{1,\dots,N(\varepsilon)\}$, the SLLN holds for θ_r^L and θ_r^U; that is, if Y is generated with parameter value θ_r^L, then*

$$\frac{1}{n_I}\left(\sum_{i\in I}Y_i-\sum_{i\in I}\mathrm{E}_{\theta_r^L}Y_i\right)\underset{n_I\to\infty}{\longrightarrow}0$$

almost surely, and similarly for θ_r^U;

(2) *for every $\theta \in \Theta$, there is an $r \in \{1,\dots,N(\varepsilon)\}$ such that $\theta_r^L \le \theta \le \theta_r^U$ coordinatewise;*

(3) *for every $r \in \{1,\dots,N(\varepsilon)\}$ and $i \in I$, $\mathrm{E}_{\theta_r^U}Y_i - \mathrm{E}_{\theta_r^L}Y_i \le \varepsilon$.*

Then the ULLN holds, that is,

$$\sup_{\theta\in\Theta}\frac{1}{n_I}\left|\sum_{i\in I}Y_i-\sum_{i\in I}\mathrm{E}_\theta Y_i\right|\underset{n_I\to\infty}{\longrightarrow}0$$

almost surely, where Y is generated with parameter value θ.

We construct $\mathcal{P}$ such that the rectangles R_r spanned by $\theta_r^L = (\lambda_r^L, \mu_r^L)$ and $\theta_r^U = (\lambda_r^U, \mu_r^U)$, that is, the closed rectangles $[\lambda_r^L, \lambda_r^U] \times [\mu_r^L, \mu_r^U]$, cover Θ. By this construction, Condition (2) holds. No matter how we choose finitely many pairs (θ_r^L, θ_r^U), Condition (1)

holds for each by Proposition 3.1. We achieve Condition (3) by proving Lipschitz continuity of the expectation $\mathrm{E}_\theta Y_i$ in θ.

Lemma 5.2: *For any $\mu^* \in]0, p_c[$, the expectation $\mathrm{E}_\theta Y_i$ is Lipschitz continuous in θ over the set $[0,1] \times [0, \mu^*]$ with some Lipschitz constant L_0, which is universal for $i \in L$.*

For the ε required by Theorem 5.1 and specifically by Condition (3), we pick a $\delta \in]0, \varepsilon/L_0]$. We cover Θ with rectangles R_u of the form

$$\Big(]\lambda_u^L, \lambda_u^U[\times]\mu_u^L, \mu_u^U[\Big) \cap \Big([0,1] \times [0, p_c[\Big)$$

with $\mu_u^U < p_c$ (but negative λ_u^L and μ_u^L, and $1 < \lambda_u^U$ are possible) and $|\theta_u^U - \theta_u^L| = \delta$. This way the diameter of the rectangles is not greater than δ and they are relatively open in $[0,1] \times [0, p_c[$. Because of compactness, there is a finite subcover of Θ with such rectangles R_u: $\{R_{u_1} \dots, R_{u_{N(\varepsilon)}}\}$. We define $\mathcal{P}$ via the 'bottom left' and 'top right' vertices of these finitely many rectangles:

$$\mathcal{P} := \left\{ \Big((\lambda_{u_r}^L \vee 0, \mu_{u_r}^L \vee 0), (\lambda_{u_r}^U \wedge 1, \mu_{u_r}^U) \Big) \Big| r \in \{1, \dots, N(\varepsilon)\} \right\}.$$

These specify closed rectangles with diameter at most δ. Due to the Lipschitz condition with Lipschitz constant L_0, Condition (3) is satisfied and Theorem 5.1 applies. In conclusion, a proof of Lemma 5.2 proves the ULLN for Y.

Proof of Lemma 5.2.: Consider $\theta, \theta' \in [0,1] \times [0, \mu^*]$, $\theta = (\lambda, \mu)$ and $\theta' = (\lambda', \mu')$. We can assume that $\theta \le \theta'$ coordinatewise. If this were not the case, we would prove the inequality for $\theta^L := (\lambda \wedge \lambda', \mu \wedge \mu')$ and $\theta^U := (\lambda \vee \lambda', \mu \vee \mu')$. This suffices since $|\theta - \theta'| = |\theta^L - \theta^U|$, and both $\mathrm{E}_\theta Y_i$ and $\mathrm{E}_{\theta'} Y_i$ are contained in $[\mathrm{E}_{\theta^L} Y_i,\ \mathrm{E}_{\theta^U} Y_i]$ due to monotonicity.

We identify the vertices of L with the source edges, and fix an ordering of all source and contamination edges: $L \cup L_2 = \{e_0, e_1, e_2, \dots\}$. Let $\vartheta : \mathbb{N} \to [0,1]$ be such that $\vartheta_k = \lambda$ if e_k is a source edge, and $\vartheta_k = \mu$ if e_k is a contamination edge. Define ϑ' analogously with λ', μ' in place of λ, μ, respectively. Finally, let $\theta^k, \theta'^k : \mathbb{N} \to [0,1]$ be defined by

$$(\theta^k)_j := \begin{cases} \vartheta'_j & \text{if } j < k, \\ \vartheta_j & \text{if } j \ge k, \end{cases} \qquad (\theta'^k)_j := \begin{cases} \vartheta'_j & \text{if } j \le k, \\ \vartheta_j & \text{if } j > k, \end{cases}$$

for $k, j \in \mathbb{N}$. Let ω_θ be the configuration that is specified by $((U_i)_{i \in L}, (V_{ij})_{(i,j) \in L_2})$ and parameter θ via $(X_i)_{i \in L}$ and $(\xi_{ij})_{(i,j) \in L_2}$. Then

$$\begin{aligned} \mathrm{E}_{\theta'} Y_i - \mathrm{E}_\theta Y_i &= \mathrm{P}_{\theta'}(Y_i = 1) - \mathrm{P}_\theta(Y_i = 1) \\ &= \sum_{k=0}^{\infty} \mathrm{P}\big(Y_i(\omega_{\theta'^k}) = 1 \text{ and } Y_i(\omega_{\theta^k}) = 0\big) \\ &= \sum_{k=0}^{\infty} (\vartheta'_k - \vartheta_k) \mathrm{P}_{\theta^k}\big(e_k \text{ is pivotal for } Y_i = 1\big), \end{aligned}$$

where the second equality is just the law of total probability when we know that $\{Y_i = 1\}$ is an increasing event, and the third equality is elaborated in [15, pp. 41–43] as such a step

is used in the proof of Russo's formula. Note that the concerns in that derivation related to an infinite edge set do not apply here because we have always got only one edge e_k whose parameter differs between θ^k and θ'^k. (The price we pay is that each pivotality is with a different parameter vector θ^k.) If e_k is a source edge, then $\vartheta'_k - \vartheta_k = \lambda' - \lambda$, and if e_k is a contamination edge, then $\vartheta'_k - \vartheta_k = \mu' - \mu$. Further,

$$\begin{aligned}
&\mathrm{P}_{\theta^k}\big(e_k \text{ is pivotal for } Y_i = 1\big) \\
&\quad\le \begin{cases} 1, & \text{if } e_k \text{ is the source edge of vertex } i, \\ \mathrm{P}_{\theta^k}(j \leftrightarrow i), & \text{if } e_k \text{ is the source edge of vertex } j \neq i, \\ 1, & \text{if } e_k \text{ is an edge incident with } i, \\ \mathrm{P}_{\theta^k}(i_1 \leftrightarrow i) + \mathrm{P}_{\theta^k}(i_2 \leftrightarrow i), & \text{if } e_k \text{ is the edge } (i_1, i_2), i \neq i_1, i_2. \end{cases}
\end{aligned}$$

Then

$$\begin{aligned}
&\sum_{e_k \text{ source edge}} (\vartheta'_k - \vartheta_k)\mathrm{P}_{\theta^k}\big(e_k \text{ is pivotal for } Y_i = 1\big) \\
&\quad\le (\lambda' - \lambda)\left(1 + \sum_{j \in L\setminus\{i\}} \mathrm{P}_{\theta^k}(j \leftrightarrow i)\right) \\
&\quad\le (\lambda' - \lambda)\left(1 + \frac{1}{\mathrm{e}^{g(\mu^*)} - 1}\right)
\end{aligned}$$

by (19). Using tricks from the proof of Lemma 4.2,

$$\begin{aligned}
&\sum_{e_k \text{ contamination edge}} (\vartheta'_k - \vartheta_k)\mathrm{P}_{\theta^k}\big(e_k \text{ is pivotal for } Y_i = 1\big) \\
&\quad\le (\mu' - \mu)\left(6 + \sum_{(i_1,i_2) \in (L\setminus\{i\})_2} \big(\mathrm{P}_{\theta^k}(i_1 \leftrightarrow i) + \mathrm{P}_{\theta^k}(i_2 \leftrightarrow i)\big)\right) \\
&\quad\le (\mu' - \mu)\left(6 + 2 \times 6 \sum_{j \in L\setminus\{i\}} \mathrm{P}_{\theta^k}(j \leftrightarrow i)\right) \\
&\quad\le (\mu' - \mu)\left(6 + \frac{12}{\mathrm{e}^{g(\mu^*)} - 1}\right).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathrm{E}_{\theta'} Y_i - \mathrm{E}_{\theta} Y_i \\
&\quad\le (\lambda' - \lambda)\left(1 + \frac{1}{\mathrm{e}^{g(\mu^*)} - 1}\right) + (\mu' - \mu)\left(6 + \frac{12}{\mathrm{e}^{g(\mu^*)} - 1}\right) \\
&\quad\le (\lambda' - \lambda + \mu' - \mu)\left(6 + \frac{12}{\mathrm{e}^{g(\mu^*)} - 1}\right) \le L_0|\theta' - \theta|
\end{aligned}$$

for some $L_0 > 0$ because in finite dimensions, all norms are equivalent. ■

Lemma 5.2 for $\mathrm{E}_\theta[Y_iY_j]$ ($i \sim j$) can be shown by a now straightforward adjustment of the original proof. This then implies that the following modification of Theorem 5.1 holds.

Theorem 5.3: *Suppose that the conditions of Theorem* 5.1 *hold with the following updates to points* (1) *and* (3) :

(1') *for every* $r \in \{1, \dots, N(\varepsilon)\}$, *if* Y *is generated with parameter value* θ_r^L, *then*

$$\frac{1}{n_p}\left(\sum_{(i,j)\in I_2} Y_iY_j - \sum_{(i,j)\in I_2} \mathrm{E}_{\theta_r^L}[Y_iY_j]\right) \underset{n_I\to\infty}{\longrightarrow} 0$$

almost surely, and similarly for θ_r^U;

(3') *for every* $r \in \{1, \dots, N(\varepsilon)\}$ *and* $(i,j) \in I_2$, $\mathrm{E}_{\theta_r^U}[Y_iY_j] - \mathrm{E}_{\theta_r^L}[Y_iY_j] \le \varepsilon$.

Then the ULLN holds, that is

$$\sup_{\theta\in\Theta} \frac{1}{n_p}\left|\sum_{(i,j)\in I_2} Y_iY_j - \sum_{(i,j)\in I_2} \mathrm{E}_\theta[Y_iY_j]\right| \underset{n_I\to\infty}{\longrightarrow} 0$$

almost surely, where Y *is generated with parameter value* θ.

The derivations and results of this section until now hold with $\widetilde{Y}$, too. All elements of the proof of our main theorem, Theorem 2.1, are in place.

Proof of Theorem 2.1.: We introduce the following vectors as shorthand:

$$b_1 = \left(\left(\frac{1}{n_I}\sum_{i\in I}\left(\mathcal{Y}_i^\ell - \mathrm{E}_0 Y_i^\ell\right)\right)_{\ell=1}^{n_c}, \quad \left(\frac{1}{n_p}\sum_{(i,j)\in I_2}\left(\mathcal{Y}_i^\ell\mathcal{Y}_j^\ell - \mathrm{E}_0[Y_i^\ell Y_j^\ell]\right)\right)_{\ell=1}^{n_c}\right)$$

$$b_2 = \left(\left(\frac{1}{n_I}\sum_{i\in I}\left(\mathrm{E}_0 Y_i^\ell - \mathrm{E}_0\widetilde{Y}_i^\ell\right)\right)_{\ell=1}^{n_c}, \quad \left(\frac{1}{n_p}\sum_{(i,j)\in I_2}\left(\mathrm{E}_0[Y_i^\ell Y_j^\ell] - \mathrm{E}_0[\widetilde{Y}_i^\ell\widetilde{Y}_j^\ell]\right)\right)_{\ell=1}^{n_c}\right)$$

$$b_3 = \left(\left(\frac{1}{n_I}\sum_{i\in I}\left(\mathrm{E}_0\widetilde{Y}_i^\ell - \mathrm{E}_\theta\widetilde{Y}_i^\ell\right)\right)_{\ell=1}^{n_c}, \quad \left(\frac{1}{n_p}\sum_{(i,j)\in I_2}\left(\mathrm{E}_0[\widetilde{Y}_i^\ell\widetilde{Y}_j^\ell] - \mathrm{E}_\theta[\widetilde{Y}_i^\ell\widetilde{Y}_j^\ell]\right)\right)_{\ell=1}^{n_c}\right)$$

$$b_4 = \left(\left(\frac{1}{n_I}\sum_{i\in I}\left(\mathrm{E}_\theta\widetilde{Y}_i^\ell - \frac{1}{n_s}\sum_{s=1}^{n_s}\widetilde{Y}_i^{\ell,s}\right)\right)_{\ell=1}^{n_c}, \right.$$
$$\left.\left(\frac{1}{n_p}\sum_{(i,j)\in I_2}\left(\mathrm{E}_\theta[\widetilde{Y}_i^\ell\widetilde{Y}_j^\ell] - \frac{1}{n_s}\sum_{s=1}^{n_s}\widetilde{Y}_i^{\ell,s}\widetilde{Y}_j^{\ell,s}\right)\right)_{\ell=1}^{n_c}\right).$$

The left-hand side of (6) is just

$$\alpha\left(\sum_{j=1}^4 b_j\right) - \alpha\,(b_2+b_3) = b_1^{\mathrm{T}}\Omega b_1 + b_4^{\mathrm{T}}\Omega b_4 + \sum_{j\neq k} b_j^{\mathrm{T}}\Omega b_k - 2b_2^{\mathrm{T}}\Omega b_3 \qquad (23)$$

We write out the case when the dataset is of type $(\mathcal{Y}_i)_{i\in I}$. When the data is of type $(\widetilde{\mathcal{Y}}_i)_{i\in I}$, then $b_2 = 0$ and the calculations of the general case can be easily adapted.

One of conditions (4) and (5) is assumed to hold, hence there is some $c_2 > 0$ such that with probability 1 for large enough n_I, for all $\eta \in \mathbb{R}^{2n_c}$ and for $\Omega = \Omega_{n_I}$,

$$0 \le \eta^{\mathrm{T}}\Omega\eta \le \sup_{\theta\in\Theta} \eta^{\mathrm{T}}\Omega\eta \le c_2|\eta|^2. \tag{24}$$

It follows by the ULLN as implied by Theorems 5.1 and 5.3 that

$$\sup_{\theta\in\Theta} b_1^{\mathrm{T}}\Omega b_1 \le c_2 \sup_{\theta\in\Theta} |b_1|^2 \underset{n_I\to\infty}{\longrightarrow} 0,$$
$$\sup_{\theta\in\Theta} b_4^{\mathrm{T}}\Omega b_4 \le c_2 \sup_{\theta\in\Theta} |b_4|^2 \underset{n_I\to\infty}{\longrightarrow} 0.$$

The remaining terms on the right-hand side of (23),

$$\sum_{j\neq k} b_j^{\mathrm{T}}\Omega b_k - 2b_2^{\mathrm{T}}\Omega b_3,$$

consist of terms $b_j^{\mathrm{T}}\Omega b_k$ where at least one of j and k is either 1 or 4. For b_1 and b_4, the ULLN holds and they converge to zero almost surely, uniformly in θ. For b_2, we can apply Proposition 4.6 and (22), so it also converges to zero. In the case of b_3, Lipschitz continuity by Lemma 5.2 and its variant for $\mathrm{E}_\theta[Y_iY_j]$ guarantee boundedness on the compact Θ. Consequently, property (24) of Ω ensures that (23) (and (6)) converges to zero uniformly in θ almost surely. Of

$$\alpha\,(b_2 + b_3) = b_2^{\mathrm{T}}\Omega b_2 + b_3^{\mathrm{T}}\Omega b_3 + 2b_2^{\mathrm{T}}\Omega b_3,$$

$b_2^{\mathrm{T}}\Omega b_2$ and $2b_2^{\mathrm{T}}\Omega b_3$ converge to zero as $n_I \to \infty$ due to Proposition 4.6 and (22).

In analogy with (24), there exists a constant $c_1 > 0$ such that almost surely for all sufficiently large n_I and for all $\eta \in \mathbb{R}^{2n_c}$,

$$c_1|\eta|^2 \le \inf_{\theta\in\Theta} \eta^{\mathrm{T}}\Omega\eta. \tag{25}$$

By the continuity of expectations in θ on the compact set Θ assured by Lemma 5.2 and its variant for $\mathrm{E}_\theta[Y_iY_j]$, the assumption of identifiability (1) implies identifiability (2) (one can take $|\theta - \theta_0| \ge \delta/2$ to prove it for the infimum for $|\theta - \theta_0| > \delta$). By this stronger notion of identifiability, the coordinatewise infimum of b_3 for any θ with $|\theta - \theta_0| \ge \delta$ for some fixed $\delta > 0$ is positive. Let ε denote the least of these $2n_c$ positive infima. Now substitute $\eta = b_3$ and use (25): its right-hand side will be bounded below by $c_1\, 2n_c\, \varepsilon^2 > 0$.

At the heart of the MSM estimator is the minimization of $\alpha(\sum_{j=1}^4 b_j)$ to define $\hat{\theta}_{n_s,n_I}$. We have seen that (23) converges to zero uniformly in θ almost surely. For the reasons that $b_3^{\mathrm{T}}\Omega b_3 \ge 0$, that it is zero if and only if $\theta = \theta_0$, and it has the lower bound $c_1\, 2n_c\, \varepsilon^2 > 0$ for $|\theta - \theta_0| \ge \delta$, the minimization forces $b_3^{\mathrm{T}}\Omega b_3$ to converge to zero. We conclude that $\hat{\theta}_{n_s,n_I} \to \theta_0$ almost surely as $n_I \to \infty$. ■

We followed the philosophy that the dataset $\mathcal{Y}$ comes from the infinite lattice L although only a finite subset is observed. This is an idealized view that assumes the existence of a process on the infinite lattice. Otherwise, when the dataset is of type $\widetilde{\mathcal{Y}}$, the derivation is simpler because Proposition 4.6 and (22) are not needed.

6. Computer testing of the proposed method

6.1. Implementation

We implemented the proposed MSM parameter estimator in the MATLAB software (The MathWorks, Inc.) [4], and we report our findings in this section. See also [3] for an early version with $n_c = 3$ colours. For the objective function

$$\alpha\left(\begin{pmatrix}\bar{\mathcal{Y}}^\ell - \dfrac{1}{n_s}\displaystyle\sum_{s=1}^{n_s}\bar{Y}^{\ell,s}\end{pmatrix}_{\ell\in\{1,\ldots,n_c\}}\right), \tag{26}$$
$$\begin{pmatrix}\bar{\mathcal{Z}}^\ell - \dfrac{1}{n_s}\displaystyle\sum_{s=1}^{n_s}\bar{Z}^{\ell,s}\end{pmatrix}_{\ell\in\{1,\ldots,n_c\}}$$

we chose the quadratic form $\alpha(\eta) = \eta^{\mathrm{T}}\Omega\eta$ the following way:

$$\Omega = \operatorname{diag}\Big((\bar{\mathcal{Y}}^1)^{-2},\ldots,(\bar{\mathcal{Y}}^{n_c})^{-2},(\bar{\mathcal{Z}}^1)^{-2},\ldots,(\bar{\mathcal{Z}}^{n_c})^{-2}\Big). \tag{27}$$

In the unlikely case that a $\bar{\mathcal{Y}}^\ell$ or a $\bar{\mathcal{Z}}^\ell$ is zero, the corresponding diagonal element of Ω is set to 1. Through this normalization, we expect each coordinate to contribute roughly equally to the sum.

Common random numbers are used during the exploration of the parameter space. This removes an element of fluctuation as different $\theta = (\lambda^1,\ldots,\lambda^{n_c},\mu) \in \Theta$ are tested. We propose two alternative methods for sampling synthetic datasets. Method 1 is the canonical approach. We draw and fix independent random variables from the uniform distribution on $[0,1]$: $(U_i^{\ell,s})$ for $\ell \in \{1,\ldots,n_c\}$, $s \in \{1,\ldots,n_s\}$, $i \in I$, and (V_{ij}^s) for $s \in \{1,\ldots,n_s\}$, $(i,j) \in I_2$. Thereafter, for each parameter vector, seeding and the open or closed state of edges are defined by

$$X_i^{\ell,s} := \begin{cases}1 & \text{if } U_i^{\ell,s} < \lambda^\ell,\\ 0 & \text{otherwise,}\end{cases} \quad \text{for } \ell \in \{1,\ldots,n_c\},\ s \in \{1,\ldots,n_s\},\ i \in I;$$

$$\xi_{ij}^s := \begin{cases}1 & \text{if } V_{ij}^s < \mu,\\ 0 & \text{otherwise,}\end{cases} \quad \text{for } s \in \{1,\ldots,n_s\},\ (i,j) \in I_2.$$

This method gives a binomially distributed number of open edges and of seeded vertices for each colour ℓ.

We test whether it is beneficial for the parameter estimation to remove the randomness in the numbers of seeds and open edges, and to make exactly as many edges open as their expected number, $\zeta(\mu n_p)$, where ζ is the rounding to the nearest integer with some tie-breaking rule. The same is stipulated for seeds: $\zeta(\lambda^\ell n_I)$ random vertices shall be seeded with colour ℓ. This is what Method 2 does. We see this as a variance-reduction trick that achieves lower variance by introducing dependencies between random draws: for example, by knowing the state of all edges but one, we can infer the state of the remaining edge.

Let S_n denote the set of permutations of $\{1,\ldots,n\}$. In Method 2, we draw permutations $(\sigma^{\ell,s})$ from S_{n_I} independently, uniformly at random for $\ell \in \{1,\ldots,n_c\}, s \in \{1,\ldots,n_s\}$, and

independent permutations (τ^s) from S_{n_p} uniformly at random for $s \in \{1,\dots,n_s\}$. With these permutations fixed, for each $\theta \in \Theta$, we let

$$X_i^{\ell,s} := \begin{cases} 1 & \text{if } \sigma^{\ell,s}(i) \le \zeta(\lambda^\ell n_I), \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } \ell \in \{1,\dots,n_c\},\ s \in \{1,\dots,n_s\},\ i \in I;$$

$$\xi_{ij}^s := \begin{cases} 1 & \text{if } \tau^s\big((i,j)\big) \le \zeta(\mu n_p), \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } s \in \{1,\dots,n_s\},\ (i,j) \in I_2.$$

Minimization over the parameter space Θ is conducted with the MATLAB routine `fminsearchbnd` [8] for constrained optimization. $\lambda^\ell_{\max} := \bar{\mathcal{Y}}^\ell$ is certainly an upper bound on what λ^ℓ any point estimator might estimate $(\ell \in \{1,\dots,n_c\})$ as this is the moment estimate in case $\mu = 0$. The upper bound $\mu_{\max}$ on μ is left to the user's judgement.

The last user input in addition to n_s, Method and $\mu_{\max}$ is n_{opt} which specifies how many different initial states to try in the optimization runs. We expect an inverse relationship between seeding rates and the contamination rate, given the data. Thus the initial parameter values for $k \in \{1,\dots,n_{\text{opt}}\}$ are chosen as

$$\lambda^\ell_{\text{initial}}(k) = \left(1 - \frac{k-1}{n_{\text{opt}}}\right)\lambda^\ell_{\max}, \quad \text{for } \ell \in \{1,\dots,n_c\},$$

$$\mu_{\text{initial}}(k) = \chi_{\{n_{\text{opt}}>1\}} \frac{k-1}{n_{\text{opt}}-1}\,\mu_{\max}.$$

6.2. Results

In order to test the performance of the proposed estimation procedure, we created a number of synthetic datasets with $n_c = 3$ colours, different sizes and different, known parameter vectors using Method 1. Tables 1–4 report the results of estimating $\theta_0 = (\lambda^1, \lambda^2, \lambda^3, \mu)$ using different input settings $(n_s, n_{\text{opt}}, \mu_{\max})$.

Table 1. Six estimates for a synthetic dataset with $n_I = 25 \times 25 = 625$ vertices ($n_p = 1776$) and $\theta_0 = (0.1, 0.05, 0.07, 0.06)$.

n_s	n_{opt}	$\mu_{\max}$	θ_0	$\hat{\theta}^{(\text{M1})}_{n_s,n_I}$	$d^{(\text{M1})}$	$\alpha_{\hat{\theta}^{(\text{M1})}_{n_s,n_I}}$	$\hat{\theta}^{(\text{M2})}_{n_s,n_I}$	$d^{(\text{M2})}$	$\alpha_{\hat{\theta}^{(\text{M2})}_{n_s,n_I}}$
10	10	0.1	0.1	0.1292	29.16%	0.0124	0.1277	27.74%	0.0215
			0.05	0.0595	19.00%		0.0637	27.39%	
			0.07	0.0611	12.71%		0.0538	23.11%	
			0.06	0.0433	27.86%	(107 s)	0.0376	37.28%	(94 s)
50	10	0.1	0.1	0.1289	28.92%	0.0128	0.1267	26.75%	0.00875
			0.05	0.0641	28.26%		0.0611	22.18%	
			0.07	0.0603	13.79%		0.0588	15.94%	
			0.06	0.0394	34.37%	(523 s)	0.0422	29.65%	(467 s)
100	10	0.1	0.1	0.1350	35.01%	0.0108	0.1259	25.87%	0.0107
			0.05	0.0631	26.18%		0.0623	24.53%	
			0.07	0.0613	12.44%		0.0595	15.00%	
			0.06	0.0360	40.00%	(1.03e+03 s)	0.0403	32.84%	(922 s)

Notes: In this synthetic dataset, the relative frequency of the incidence of the three colours in the seeding is $\bar{\mathcal{X}} = (0.107, 0.0528, 0.064)$, while in the contamination-impacted observed data, it is $\bar{\mathcal{Y}} = (0.15, 0.0752, 0.08)$. The relative frequency of adjacent vertices having an open edge between them is 0.0574.

Table 2. Four estimates for a synthetic dataset with $n_I = 100 \times 100 = 10000$ vertices ($n_p = 29601$) and $\theta_0 = (0.07, 0.05, 0.04, 0.03)$.

n_s	n_{opt}	$\mu_{\max}$	θ_0	$\hat{\theta}^{(\text{M1})}_{n_s,n_I}$	$d^{(\text{M1})}$	$\alpha_{\hat{\theta}^{(\text{M1})}_{n_s,n_I}}$	$\hat{\theta}^{(\text{M2})}_{n_s,n_I}$	$d^{(\text{M2})}$	$\alpha_{\hat{\theta}^{(\text{M2})}_{n_s,n_I}}$
20	10	0.05	0.07	0.0734	4.79%	1.55e−05	0.0739	5.52%	0.00044
			0.05	0.0508	1.52%		0.0492	1.66%	
			0.04	0.0390	2.49%		0.0392	2.07%	
			0.03	0.0259	13.77%	(3.08e+03 s)	0.0262	12.79%	(3.01e+03 s)
40	10	0.05	0.07	0.0737	5.28%	5.96e−06	0.0733	4.78%	0.000108
			0.05	0.0505	0.93%		0.0500	0.06%	
			0.04	0.0392	2.05%		0.0393	1.82%	
			0.03	0.0253	15.54%	(5.9e+03 s)	0.0252	16.00%	(5.63e+03 s)

Notes: In this synthetic dataset, the relative frequency of the incidence of the three colours in the seeding is $\bar{\mathcal{X}} = (0.0717, 0.0497, 0.0387)$, while in the contamination-impacted observed data, it is $\bar{\mathcal{Y}} = (0.085, 0.0585, 0.0455)$. The relative frequency of adjacent vertices having an open edge between them is 0.0303.

Table 3. Eight estimates for a synthetic dataset with $n_I = 300 \times 300 = 90000$ vertices ($n_p = 268801$) and $\theta_0 = (0.05, 0.06, 0.03, 0.02)$.

n_s	n_{opt}	$\mu_{\max}$	θ_0	$\hat{\theta}^{(\text{M1})}_{n_s,n_I}$	$d^{(\text{M1})}$	$\alpha_{\hat{\theta}^{(\text{M1})}_{n_s,n_I}}$	$\hat{\theta}^{(\text{M2})}_{n_s,n_I}$	$d^{(\text{M2})}$	$\alpha_{\hat{\theta}^{(\text{M2})}_{n_s,n_I}}$
2	8	0.05	0.05	0.0472	5.57%	0.00241	0.0488	2.43%	0.000149
			0.06	0.0572	4.71%		0.0593	1.18%	
			0.03	0.0308	2.71%		0.0298	0.51%	
			0.02	0.0225	12.59%	(4e+03 s)	0.0220	9.86%	(3.95e+03 s)
4	4	0.05	0.05	0.0480	4.03%	0.000968	0.0489	2.12%	6.11e−05
			0.06	0.0578	3.59%		0.0589	1.90%	
			0.03	0.0305	1.65%		0.0298	0.83%	
			0.02	0.0223	11.64%	(3.8e+03 s)	0.0213	6.45%	(3.19e+03 s)
8	2	0.05	0.05	0.0483	3.32%	0.000904	0.0481	3.75%	0.000612
			0.06	0.0586	2.35%		0.0586	2.26%	
			0.03	0.0299	0.40%		0.0301	0.45%	
			0.02	0.0223	11.60%	(3.53e+03 s)	0.0220	10.11%	(3.83e+03 s)
5	5	0.05	0.05	0.0481	3.81%	0.000964	0.0485	2.97%	0.000495
			0.06	0.0580	3.29%		0.0588	1.99%	
			0.03	0.0302	0.79%		0.0303	1.04%	
			0.02	0.0221	10.69%	(6.08e+03 s)	0.0219	9.72%	(6.12e+03 s)

Notes: In this synthetic dataset, the relative frequency of the incidence of the three colours in the seeding is $\bar{\mathcal{X}} = (0.0498, 0.0595, 0.0299)$, while in the contamination-impacted observed data, it is $\bar{\mathcal{Y}} = (0.0558, 0.0667, 0.034)$. The relative frequency of adjacent vertices having an open edge between them is 0.0198.

Table 4. Four estimates for a synthetic dataset with $n_I = 500 \times 500 = 250000$ vertices ($n_p = 748001$) and $\theta_0 = (0.03, 0.04, 0.05, 0.02)$.

n_s	n_{opt}	$\mu_{\max}$	θ_0	$\hat{\theta}^{(\text{M1})}_{n_s,n_I}$	$d^{(\text{M1})}$	$\alpha_{\hat{\theta}^{(\text{M1})}_{n_s,n_I}}$	$\hat{\theta}^{(\text{M2})}_{n_s,n_I}$	$d^{(\text{M2})}$	$\alpha_{\hat{\theta}^{(\text{M2})}_{n_s,n_I}}$
1	1	0.04	0.03	0.0336	11.86%	0.0243	0.0331	10.47%	0.025
			0.04	0.0444	10.96%		0.0450	12.59%	
			0.05	0.0553	10.70%		0.0555	10.94%	
			0.02	0.0141	29.27%	(355 s)	0.0136	31.91%	(340 s)
5	5	0.04	0.03	0.0297	1.08%	0.000894	0.0297	1.04%	0.00106
			0.04	0.0394	1.43%		0.0395	1.13%	
			0.05	0.0518	3.57%		0.0519	3.86%	
			0.02	0.0196	1.94%	(2.8e+04 s)	0.0194	3.01%	(2.48e+04 s)

Notes: In this synthetic dataset, the relative frequency of the incidence of the three colours in the seeding is $\bar{\mathcal{X}} = (0.0299, 0.0402, 0.0503)$, while in the contamination-impacted observed data, it is $\bar{\mathcal{Y}} = (0.0336, 0.0451, 0.057)$. The relative frequency of adjacent vertices having an open edge between them is 0.01996.

The two estimators, which are based on Methods 1 and 2 of random number generation, are denoted by $\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}$ and $\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}$, respectively. We display the relative bias of the estimators in percentage terms:

$$d^{(\mathrm{M1})} = 100 \left| 1 - \hat{\theta}^{(\mathrm{M1})}_{n_s,n_I} / \theta_0 \right|$$

(the operations are coordinatewise), and analogously, $d^{(\mathrm{M2})}$ for $\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}$. Finally, we let $\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$ and $\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$ denote the value of the objective function α in (26) at $\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}$ and $\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}$, respectively.

The computations were conducted on a laptop computer equipped with a 2.8 GHz Intel Core i7-2640M dual-core processor and 8 GB RAM. Although it is clear that the n_{opt} parameter searches, as well as the n_s simulations in each step of each search lend themselves to parallelization, our implementation does not benefit from this insight. The columns of $\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$ and $\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$ display in brackets running times in seconds for completing the parameter estimation procedure. These times are indicative only and their use for comparisons is limited as less demanding other tasks were also running on the computer simultaneously. As far as we can tell, the parameter estimation ran in RAM without resorting to swap memory on disk.

We found no definitive answer as to whether Method 1 or 2 is preferable. Table 3 suggests Method 2, but Table 4 is as inconclusive as smaller-sized datasets.

Broadly, the relative bias of the estimates becomes smaller as n_I grows. From $n_I = 25 \times 25 = 625$ to $n_I = 500 \times 500 = 250000$, the relative bias of the μ estimate improves from about 35–40% to below 5%. We have also observed that as n_I grows, there is ever less need to try several initial states because the solutions tend to converge to the same estimator. In our experience, the existence of local optima that necessitate a greater n_{opt} were characteristic of the smaller lattice sizes only.

In the smallest dataset, Table 1, one can observe that λ^1 and λ^2 are consistently overestimated, whereas λ^3 and μ are underestimated in all six estimations. This turned out to be due to a quirk of the randomly generated dataset. While $(\lambda^1, \lambda^2, \lambda^3) = (0.1, 0.05, 0.07)$, in reality, the dataset had

$$\begin{aligned} \bar{\mathcal{X}} &= \frac{1}{n_I} \left(\sum_{i \in I} \mathcal{X}_i^1, \sum_{i \in I} \mathcal{X}_i^2, \sum_{i \in I} \mathcal{X}_i^3 \right) \\ &= (0.1072,\ 0.0528,\ 0.0640). \end{aligned}$$

One can notice that in Table 1, λ^1 and λ^2 are overestimated to a greater extent than how much λ^3 is underestimated. Then the observed systemic underestimation of μ is consistent with this in light of the expected inverse relationship between seeding and contamination described at the end of Section 6.1.

In Table 3, where the lattice size $n_I = 300 \times 300 = 90000$ is most relevant to our practical application in Section 7, $(n_s, n_{\mathrm{opt}}) \in \{(2, 8), (4, 4), (8, 2)\}$ allow a comparison of different input choices with approximately identical computational cost. $(n_s, n_{\mathrm{opt}}) = (4, 4)$ and right behind it $(8, 2)$ proved to be the best choices, beating $(n_s, n_{\mathrm{opt}}) = (2, 8)$. Against the expectations, $(n_s, n_{\mathrm{opt}}) = (5, 5)$ happened to not improve the estimate with input $(4, 4)$. On this lattice size, μ is estimated to 12% accuracy with 1–2 hours running time. In Table 4, we get better than 5% accuracy on a larger lattice with 7–8 hours running time.

In Tables 1–4, for fixed n_I, $\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$ and $\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$ tend to decrease for increasing n_s. This is reassuring, although not a necessity because it is possible that the synthetic dataset is atypical and more simulations (higher n_s) do not make it easier to approximate it. Instead, overfitting might yield the lowest α values.

For further analysis, we introduce two more symbols. One might consider a trivial estimator which assumes no contamination occurring: $\hat{\mu} = 0$, $\hat{\lambda}^{\ell} = \bar{\mathcal{Y}}^{\ell}$. The corresponding α_{triv} denotes a realization of α with parameters from this trivial estimator, computed from n_s simulations with Method 1 or 2. α_{θ_0} denotes a realization of α with the true parameter θ_0 and n_s simulations.

Table 5 compares $\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$ and $\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$, α_{triv} and α_{θ_0} for the four computer-generated datasets of Tables 1–4. Except for the smallest case, $n_I = 625$, α_{θ_0} is always smaller than α_{triv}, as expected. Whereas α_{θ_0} decreases with increasing n_I, α_{triv} stays roughly constant. $\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$ and $\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$ decrease only initially as n_I increases. One would expect them to be between α_{θ_0} and α_{triv}, which tends to hold for larger lattice sizes. In reality, their value is much lower than α_{θ_0}, but the ratio becomes ever less extreme as n_I grows. This is indicative of initially very strong, but later ever less pronounced overfitting.

To test the behaviour of the objective function α_{θ_0} as $n_I \to \infty$, we generated fresh synthetic datasets of different sizes with a common $\theta_0 = (0.03, 0.04, 0.05, 0.02)$. Just generating the single dataset of size 1000×1000 took 42 seconds. For this exercise, the single datasets were compared to simulations with common simulation count $n_s = 10$. Table 6 shows that

Table 5. A comparison of the values of the objective functions for the true value θ_0, for the trivial estimator and for the MSM estimator.

n_I	n_p	n_s	n_{opt}	$\alpha^{(\mathrm{M1})}_{\theta_0}$	$\alpha^{(\mathrm{M2})}_{\theta_0}$	$\alpha^{(\mathrm{M1})}_{\mathrm{triv}}$	$\alpha^{(\mathrm{M2})}_{\mathrm{triv}}$	$\alpha_{\hat{\theta}^{(\mathrm{M1})}_{n_s,n_I}}$	$\alpha_{\hat{\theta}^{(\mathrm{M2})}_{n_s,n_I}}$
25×25	1776	10	10	1.06	1.77	0.6	0.555	0.0124	0.0215
25×25	1776	50	10	0.82	0.919	0.61	0.618	0.0128	0.00875
25×25	1776	100	10	1.19	1.09	0.597	0.589	0.0108	0.0107
100×100	29601	20	10	0.042	0.0454	0.59	0.61	1.55e−05	0.00044
100×100	29601	40	10	0.0496	0.0576	0.598	0.601	5.96e−06	0.000108
300×300	268801	2	8	0.0179	0.0048	0.646	0.604	0.00241	0.000149
300×300	268801	4	4	0.00948	0.0014	0.653	0.624	0.000968	6.11e−05
300×300	268801	8	2	0.00792	0.00659	0.654	0.614	0.000904	0.000612
300×300	268801	5	5	0.0104	0.00786	0.651	0.622	0.000964	0.000495
500×500	748001	1	1	0.0223	0.0119	0.628	0.621	0.0243	0.025
500×500	748001	5	5	0.00626	0.00791	0.633	0.624	0.000894	0.00106

Note: The four synthetic datasets used are the same as in Tables 1–4.

Table 6. Realizations of the objective function α for the true parameter value θ_0 for different synthetic dataset sizes and of the not normalized variant of the objective function $\widetilde{\alpha}(\eta) = \eta^{\mathsf{T}}\eta$.

Size	n_I	n_p	n_s	$\alpha^{(\mathrm{M1})}_{\theta_0}$	$\alpha^{(\mathrm{M2})}_{\theta_0}$	$\widetilde{\alpha}^{(\mathrm{M1})}_{\theta_0}$	$\widetilde{\alpha}^{(\mathrm{M2})}_{\theta_0}$
25×25	625	1776	10	0.202	0.16	6.65e−05	4.21e−05
100×100	10000	29601	10	0.0427	0.0277	7.29e−06	2.57e−06
300×300	90000	268801	10	0.00624	0.00807	1.68e−06	1.57e−06
500×500	250000	748001	10	0.000799	0.0015	3.87e−07	3.97e−07
707×707	499849	1496720	10	0.000521	0.000365	5.08e−07	3.99e−07
1000×1000	1000000	2996001	10	0.00127	0.00117	9.78e−08	9.04e−08

Note: Here $\theta_0 = (0.03, 0.04, 0.05, 0.02)$ across fresh synthetic datasets.

both α_{θ_0} and $\widetilde{\alpha}(\eta) = \eta^{\mathrm{T}}\eta$ converge to zero, although α_{θ_0} has larger values because of the normalization by Ω in (27). This is numerical evidence in support of Propositions 3.1 and 3.2, even with fixed n_s.

7. Cross-contamination rate estimation for digital PCR in lab-on-a-chip microfluidic devices

Our motivation for investigating this problem is the need for quality control in parallelized biochemical experiments run in novel, lab-on-a-chip microfluidic devices for applications in basic research, biotechnology, medical diagnostics and rapid vaccine development. Our collaborators Dr Günter Roth and his group (Centre for Biological Systems Analysis [ZBSA], University of Freiburg) develop such microfluidic devices. The central element of their system is a rectangular well plate with 15 mm edge lengths, with more than 100,000 wells of 19 pℓ volume each. The wells on this chip are arranged in a hexagonal tiling pattern (honeycomb lattice).

Whereas the rival microfluidic technology uses an emulsion of water droplets flowing in an oil medium, this array-based setup fixes a spatial structure, allowing the otherwise neglected analysis of cross-contamination between reaction volumes. Our focus is on evaluating an experiment particularly well suited for this purpose, whose results generalize to other experiments conducted in this lab-on-a-chip device.

In the *digital PCR* experiment, a solution of DNA samples is injected onto the well plate, at such a low concentration that most wells receive 0 or 1 DNA molecule (hence the name *digital*). In the particular case, the solution is a mixture of three different DNA species. We call these template molecules *seeds*. The well plate is covered with a lid (a microscope slide) that is pre-coated with covalently bound DNA primers [17]. The well plate together with the lid serve to insulate the reaction volumes from each other. The DNA templates are amplified in each of the wells independently with a polymerase chain reaction (PCR). In more detail, the template molecules hybridize to the surface-bound primers and the PCR elongates these primers to form the complementary strand of the template. In the next heating step, the templates become resolved, whereas the generated complementary DNA strands stay covalently bound to the surface. The single-strand templates will bind to other surface-bound primers and turn them too into complementary strands via polymerization. The result of the PCR cycles is that the whole glass surface above the well gets covered with immobilized complementary DNA strands. They mirror the spatial arrangement of the initial seed pattern of the wells.

After the PCR, the three complementary DNA species on the slide are identified via three specifically binding fluorescent hybridization probes (fluorophores) and their presence or absence can be determined by imaging [18]. In the fluorescent image of the slide (Figure 2), we see either black background (where there was no seed), spots in one of the three primary colours indicating a single seed, and sometimes a mixture of two or three primary colours indicating heterogeneous seeding by multiple seeds. Sometimes we also see clusters of one colour, or an unusually high number of mixed colours, indicating cross-contamination between adjacent wells. This happens when the lid is not fitted tightly and during thermal cycling, liquid exchange occurs between reaction volumes around trapped air bubbles and dust particles. In the readout it remains unclear if two neighbours with the

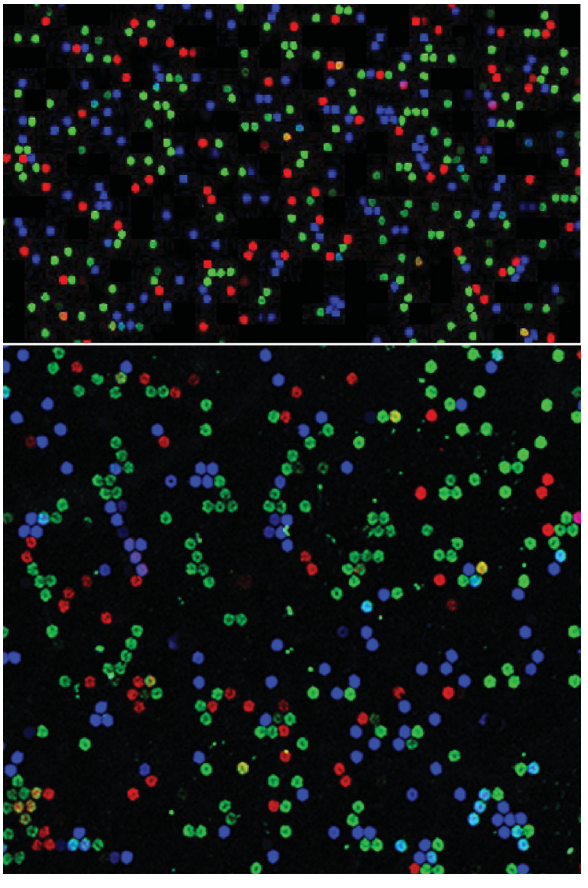

Figure 2. (*top*) Image of a glass slide from a digital PCR experiment with little sign of cross-contamination [26]. (*bottom*) Image of a slide with clustering fluorescent signals and a higher prevalence of cyan and yellow colours, suggesting higher cross-contamination rate.

same colour (or a single well with a mix of two colours, which has coloured neighbours) were initiated by two seeds or one contaminated the other (Fig. 2, bottom panel).

For cross-contamination rate estimation for this experimental setup it is necessary to define a mathematical model of the physical process. It has to involve the triangular lattice, which is the dual of the hexagonal tiling, and colouring of its vertices. The total numbers of DNA templates of each type $\ell \in \{1, \dots, n_c\}$ present in the chip are likely well approximated by n_c discretized normal random variables. We can safely assume that each well receives a Poisson distributed random number of DNA templates of type ℓ because then due to the superposition property, the total number of type ℓ templates in the chip is also Poisson distributed, which is close to a normal distribution. The Bernoulli distributed (X_i^ℓ) used in our model for seeding are really just a proxy to the either zero or positive value of the corresponding Poisson distribution. From a value λ^ℓ of the Bernoulli parameter, we can infer the parameter $\widetilde{\lambda}^\ell$ of the respective Poisson distribution through the identity $\lambda^\ell = 1 - \mathrm{e}^{-\widetilde{\lambda}^\ell}$.

It is also natural to model the possibility of contamination by open edges. It is a useful shortcut to draw the state of the edges independently of the seeding so that an open edge

means only the possibility of propagation, which is contingent on the presence of seeds. There are modelling choices to be made. Contamination might be

(i) unidirectional (there is the possibility of a pair of independent, oppositely oriented directed edges $\xi_{i\to j}$ and $\xi_{j\to i}$ between any two adjacent vertices $i \sim j$), or
(ii) symmetric (undirected edges ξ_{ij}).

Open edges might be best represented by

(1) independent Bernoulli variables, or by
(2) locally correlated 0–1 random variables.

Contamination might be

(A) confined to neighbours, or
(B) it might propagate via a series of open edges.

The choice of (ii,1,B) yields the model put forward in Section 1 (Figure 3). Its strength is that it can use standard percolation theory. Our MSM estimator was developed for this model.

For the quality certification of this lab-on-a-chip device, it is useful to estimate, in addition to μ, the total number of vertices which belong to a non-trivial component of the percolation graph. These vertices are the wells which were not insulated from their neighbours. Beyond the digital PCR paradigm, in experimental setups where most wells are expected to give some signal, vertices that are connected to any other are likely to give false signals.

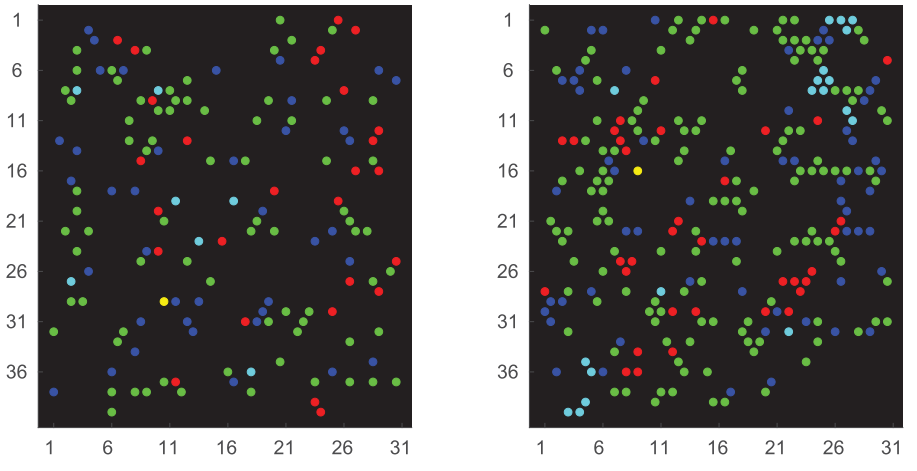

Figure 3. (*left*) Computer simulation of a glass slide from a digital PCR experiment under model (ii,1,B) with $\theta_0 = (\lambda^{\text{red}}, \lambda^{\text{green}}, \lambda^{\text{blue}}, \mu) = (0.02, 0.07, 0.05, 0.01)$ and relatively little sign of cross-contamination. (*right*) Computer-simulated slide with clustering fluorescent signals and a higher prevalence of cyan colour, suggesting higher cross-contamination rate. Here $\theta_0 = (0.02, 0.07, 0.05, 0.06)$.

An easy upper bound results from noticing that each additional edge makes at most two additional vertices connected. For small values of μ, edges are actually unlikely to share endpoints. The number of edges is distributed according to a binomial distribution with parameters n_p and μ. Therefore the mean number of potentially contaminated vertices can be estimated as

$$\mathrm{E}\left[\sum_{|C|\geq 2}|C|\right] \leq 2\mu n_p \sim 6\mu n_I$$

where the asymptotic equality holds under the assumption that the boundary of I is 'small'. For concrete examples, the conversion from n_p to n_I can be accurately determined.

Another approach results by noticing

$$\begin{aligned}\mathrm{E}\left[\sum_{|C|\geq 2}|C|\right] &= \mathrm{E}\left[n_I - \sum_{i\in I}\chi_{\{|C(i)|=1\}}\right]\\ &= n_I - n_I(1-\mu)^6 + e\\ &= \left(6\mu - 15\mu^2 + \sum_{k=3}^{6}\binom{6}{k}(-1)^{k+1}\mu^k\right)n_I + e,\end{aligned}$$

where e is the correction for boundary vertices.

Simpler cases are given by (i,1,A) and (ii,1,A) where the moments $\mathrm{E}[Y_i^\ell]$, $\mathrm{E}[Y_i^\ell Y_i^m]$ and $\mathrm{E}[Y_i^\ell Y_j^\ell]$ ($\ell \neq m$, $i \sim j$) can be computed explicitly. We used MATHEMATICA (Wolfram Research, Inc.) to deal with the many terms [4], and here we report truncations of the complete result for space considerations in the case (i,1,A). It is anticipated in the practical application that $\mu < \lambda^\ell$ for every ℓ. For non-boundary vertices, under this assumption on the anticipated magnitudes, the dominant terms of the moments of interest in decreasing order are given as

$$\begin{aligned}\mathrm{E}[Y_i^\ell] &= \mathrm{P}(X_i^\ell = 1) + \mathrm{P}(X_i^\ell = 0)\sum_{k=1}^{6}\binom{6}{k}\mu^k(1-\mu)^{6-k}\left(1-(1-\lambda^\ell)^k\right)\\ &= \lambda^\ell + 6\lambda^\ell\mu - 6(\lambda^\ell)^2\mu - 15(\lambda^\ell)^2\mu^2 + \mathcal{O}\big((\lambda^\ell)^5\big),\\ \mathrm{E}[Y_i^\ell Y_i^m] &= \mathrm{P}(X_i^\ell X_i^m = 1) + \mathrm{P}(X_i^\ell = 1, X_i^m = 0)\sum_{k=1}^{6}\binom{6}{k}\mu^k(1-\mu)^{6-k}\left(1-(1-\lambda^m)^k\right)\\ &\quad + \mathrm{P}(X_i^\ell = 0, X_i^m = 1)\sum_{k=1}^{6}\binom{6}{k}\mu^k(1-\mu)^{6-k}\left(1-(1-\lambda^\ell)^k\right)\\ &\quad + \mathrm{P}(X_i^\ell = X_i^m = 0)\sum_{k=1}^{6}\binom{6}{k}\mu^k(1-\mu)^{6-k}\left(1-(1-\lambda^\ell)^k\right)\left(1-(1-\lambda^m)^k\right)\\ &= \lambda^\ell\lambda^m + 18\lambda^\ell\lambda^m\mu - 12\Big((\lambda^\ell)^2\lambda^m + \lambda^\ell(\lambda^m)^2\Big)\mu\\ &\quad + 30\lambda^\ell\lambda^m\mu^2 + \mathcal{O}\big(\max\{\lambda^\ell,\lambda^m\}^5\big).\end{aligned}$$

For $\mathrm{E}[Y_i^\ell Y_j^\ell]$ ($i \sim j$), in the case $X_i^\ell + X_j^\ell = 1$, the empty vertex might have been contaminated by the seeded vertex, or it might have been contaminated from its five remaining neighbours. If $X_i^\ell = X_j^\ell = 0$, then one can separate cases according to the seeding status of the two shared neighbours of i and j. These considerations give

$$\mathrm{E}[Y_i^\ell Y_j^\ell] = (\lambda^\ell)^2 + 2\lambda^\ell\mu + 8(\lambda^\ell)^2\mu + 2\lambda^\ell\mu^2 - 10(\lambda^\ell)^3\mu + 9(\lambda^\ell)^2\mu^2 + \mathcal{O}\big((\lambda^\ell)^5\big).$$

These $n_c^2/2 + 3n_c/2$ moment equations provide the opportunity to estimate the $n_c + 1$ parameters via the method of moments. Of these, it is $\mathrm{E}[Y_i^\ell Y_j^\ell]$ where the first term with μ is highest up in the magnitude ranking, underpinning the physical intuition that the cooccurrence of a colour in two adjacent vertices is the most informative moment about the contamination rate μ.

Notably, the model (ii,1,A) gives exactly the above moment equations if for any $(i,j) \in I_2$,

$$\begin{aligned} &\mathrm{P}(\xi_{i\to j} = 1) = \mathrm{P}(\xi_{j\to i} = 1) = \mu \quad \text{in model (i,1,A), and} \\ &\mathrm{P}(\xi_{ij} = 1) = \mu \quad \text{in model (ii,1,A),} \end{aligned}$$

that is, if the numerical values on the right-hand sides are shared between the two models. The reason is that the propagation of colours is limited to neighbours, already second neighbours are ruled out. An edge between i and j makes a difference in any of the above three moments if and only if $X_i^\ell + X_j^\ell = 1$. Say, $X_j^\ell = 1 = 1 - X_i^\ell$. Then $\xi_{j\to i}$ has the same effect on these moments as ξ_{ij}, and also the same probability because one can marginalize over the state of $\xi_{i\to j}$. However, $\mathrm{E}[Y_i^\ell Y_j^\ell Y_i^m Y_j^m]$ would differ between the models (i,1,A) and (ii,1,A). See also [11, Appendix].

8. Discussion and open problems

This paper describes the solution of a statistical problem motivated by a concrete practical need. The mathematical modelling part is solved in one of multiple possible ways, and the choice of (ii,1,B) brings in bond percolation into the statistical model. We assume that the percolation is subcritical. The parameter estimation method we propose is the MSM, which gives a point estimate. We prove that it is strongly consistent in the limit as the sample size n_I tends to infinity (under the assumption of identifiability). It is an important point that the number of simulations per proposed parameter vector, n_s, can remain bounded to achieve this result.

What is unusual in our setting is that although the sample size is large, it is not independent (nor identically distributed). Introductory percolation theory provides upper bounds on long-range dependencies between the n_I samples.

We have implemented the method and its accuracy is tested on synthetic datasets in practically relevant parameter ranges. Estimates for wetlab data are to be published by our collaborators Günter Roth and his co-workers in the microfluidics literature.

Parameter estimation in connection with a (static) percolation model is not common in the literature, apart from the quest for the critical value. Dynamic percolation models and dynamic random graphs on a fixed vertex set provide a framework for the contact network in modelling the spread of epidemics. Gilligan and Gibson have been particularly active in

studying statistical problems for spatiotemporal models of plant epidemic spread [12,21]. Gilligan and co-workers also conducted experiments with the fungal pathogen *Rhizoctonia solani* grown in a Petri dish to test how infection probability between a pair of lattice points (that is, the parameter μ of percolation in the directed case (i)) depends on their distance and how invasive spread (percolation) probability depends on nutrient availability in lattice points and on the distance between lattice points [2]. They also demonstrated that the random removal (blocking) of sites can hinder and even stop disease spread by driving it subcritical [23].

Beyond the almost sure convergence and the numerical studies with synthetic data, we cannot predict the accuracy of our estimator for instance in terms of confidence intervals. It is known that under regularity conditions, especially that the estimator is continuously differentiable with respect to the parameter θ, $\sqrt{n_I}(\hat{\theta}_{n_s,n_I} - \theta_0)$ is asymptotically normal with known limiting variance [14, Section 2.3.1]. It is also possible to choose Ω optimally, that is, to minimize this asymptotic variance [14, Section 2.3.4]. However, our estimator is not even continuous in θ because we use what is called a frequency simulator. It is unknown to us whether it is possible to replace the frequency simulator with some importance sampling to achieve asymptotic normality.

Maximum likelihood estimation (MLE) would have the advantage over MSM that its output is reproducible. Its computational cost might also be lower. Consider the following. We know that black areas have no seeds but we have no information about contamination (edges) in them. We also know that at boundaries between different colours, there is no open edge. Therefore, for a MLE, one needs to establish the probabilities of patches with a fixed colour without knowing which vertices were seeded and which got contaminated only.

We wonder if it is possible by using a generating function that encodes the probabilities of seeding and open edges to compute the total probability that the particular patch was created: each vertex in a patch has been seeded or contaminated from a seed somewhere within the patch. We were only able to derive this generating function for patches that are a linear chain of vertices.

General finite, connected patch shapes (subgraphs) are called *(lattice) animals*. Mireille Bousquet-Mélou did much work on characterizing them via generating functions [6,7]. Our patches can arise as a disjoint union of adjacent connected components (animals). For our application, it would suffice to develop a recursion which allows one to compute generating functions of small patches (large patches are rare) with a computer algebra system. The difficulty is that the problem is two dimensional, and a patch must be split in all possible ways into two disjoint parts in the recursion. Any newly added vertex might have been seeded, or contaminated from the rest of the patch, but it might have itself contaminated other empty vertices of the patch.

Notably, the MSM estimator can be turned into an *approximate Bayesian computation* (ABC) method very easily. One needs to fix a prior distribution on Θ and a small $\varepsilon > 0$. The ABC rejection algorithm draws finitely many independent $\theta \in \Theta$ parameter values from the prior distribution. The objective function (26) is evaluated for each proposed θ. The simulations used for the evaluation should no longer use common random numbers but independent ones, and n_s can be set to one. If the value of the objective function is less than ε, then the proposed θ is accepted, otherwise it is rejected. This way the set of accepted θ is a good approximation of the posterior distribution.

We have not yet tested model fit due to the lack of experimental data. As contamination is caused by the imperfect fit of the glass lid, and trapped bubbles and dust, we anticipate that locally positively correlated open edges might be needed in the model. That is, case (ii,2,B) deserves close attention. One way of modelling positive correlations is to apply the Ising model to the edges. Let $\widetilde{\xi}_{ij} = 2\xi_{ij} - 1 \in \{-1, +1\}$. Then the energy or the Hamiltonian function of a configuration ξ of open edges is

$$H(\xi) = -J \sum_{i<j<k} (\widetilde{\xi}_{ij}\widetilde{\xi}_{ik} + \widetilde{\xi}_{ij}\widetilde{\xi}_{jk} + \widetilde{\xi}_{ik}\widetilde{\xi}_{jk}) - \widetilde{\mu} \sum_{(i,j)\in I_2} \widetilde{\xi}_{ij}$$

for some $J > 0$ and $\widetilde{\mu} < 0$, and in the first sum, out of the three terms those are missing where an adjacency condition is not met: $\widetilde{\xi}_{ij} = 0$ if $i \nsim j$, so that every pair of incident edges appears once. The probability of the system being in state ξ is proportional to $\mathrm{e}^{-\beta H(\xi)}$ for some $\beta > 0$. Although we have two new parameters J and the inverse temperature β in addition to $\widetilde{\mu}$, the increase in degrees of freedom is really just one, βJ and $\beta\widetilde{\mu}$ relative to μ.

Acknowledgements

The authors are grateful to Günter Roth and Christin Rath (ZBSA, University of Freiburg) for proposing the problem, for their relentless help in clarifying details of the experimental protocol and for providing sample images. The authors also thank Robin Ryder (Paris Dauphine University) for suggesting the method of simulated moments, and Ed Crane (University of Bristol) and Peter Pfaffelhuber (University of Freiburg) for insights.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

B. M. thanks the AXA Research Fund for their financial support in the form of a postdoctoral fellowship, and the Isaac Newton Institute for Mathematical Sciences (Cambridge, UK) for support and hospitality during the programme *Stochastic dynamical systems in biology: numerical methods and applications* when work on this paper was undertaken. This work was thereby supported by Engineering and Physical Sciences Research Council Grant Number EP/K032208/1.

ORCID

Bence Mélykúti http://orcid.org/0000-0002-9812-0240

References

[1] M. Aizenman and C.M. Newman, *Tree graph inequalities and critical behavior in percolation models*, J. Stat. Phys. 36 (1984), pp. 107–143.
[2] D.J. Bailey, W. Otten and C.A. Gilligan, *Saprotrophic invasion by the soil-borne fungal plant pathogen* Rhizoctonia solani *and percolation thresholds*, New Phytol. 146 (2000), pp. 535–544.
[3] F. Beck, *Parameter estimation in a percolation model with coloring*, Master's thesis, Institute for Mathematics, University of Freiburg, Germany, 2015.
[4] F. Beck and B. Mélykúti, *Parameter estimation of seeding and contamination rates on a triangular lattice using the method of simulated moments (MSM)*, 2017. Available at https://github.com/Melykuti/Parameterestimation_MSM, MATLAB software and MATHEMATICA notebook.

[5] J. van den Berg and H. Kesten, *Inequalities with applications to percolation and reliability*, J. Appl. Probab. 22 (1985), pp. 556–569. Available at http://www.jstor.org/stable/3213860.
[6] M. Bousquet-Mélou, *New enumerative results on two-dimensional directed animals*, Discrete Math. 180 (1998), pp. 73–106.
[7] M. Bousquet-Mélou and A. Rechnitzer, *Lattice animals and heaps of dimers*, Discrete Math. 258 (2002), pp. 235–274.
[8] J. D'Errico, `fminsearchbnd`, `fminsearchcon` MATLAB *files* (2012). Available at http://uk.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd–fminsearchcon.
[9] N. Etemadi, *On the laws of large numbers for nonnegative random variables*, J. Multivar. Anal. 13 (1983), pp. 187–193. Available at http://www.sciencedirect.com/science/article/pii/0047259X83900131.
[10] C.M. Fortuin, P.W. Kasteleyn and J. Ginibre, *Correlation inequalities on some partially ordered sets*, Commun. Math. Phys. 22 (1971), pp. 89–103.
[11] H.L. Frisch and J.M. Hammersley, *Percolation processes and related topics*, J. Soc. Indus. Appl. Math. 11 (1963), pp. 894–918. Available at http://www.jstor.org/stable/2946482.
[12] G.J. Gibson, W. Otten, J.A.N. Filipe, A. Cook, G. Marion and C.A. Gilligan, *Bayesian estimation for percolation models of disease spread in plant populations*, Stat. Comput. 16 (2006), pp. 391–402.
[13] C. Gouriéroux and A. Monfort, *Simulation based inference in models with heterogeneity*, Ann. Econ. Stat. 20–21 (1991), pp. 69–107. Available at http://www.jstor.org/stable/20075807.
[14] C. Gouriéroux and A. Monfort, Simulation-Based Econometric Methods, Oxford University Press, Oxford, 2002.
[15] G. Grimmett, *Percolation*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 1999.
[16] T.E. Harris, *A lower bound for the critical probability in a certain percolation process*, Math. Proc. Camb. Philos. Soc. 56 (1960), pp. 13–20.
[17] J. Hoffmann, S. Hin, F. von Stetten, R. Zengerle and G. Roth, *Universal protocol for grafting PCR primers onto various lab-on-a-chip substrates for solid-phase PCR*, RSC Adv. 2 (2012), pp. 3885–3889.
[18] J. Hoffmann, M. Trotter, F. von Stetten, R. Zengerle and G. Roth, *Solid-phase PCR in a picowell array for immobilizing and arraying 100 000 PCR products to a microscope slide*, Lab on a Chip 12 (2012), pp. 3049–3054.
[19] H. Kesten, *The critical probability of bond percolation on the square lattice equals 1/2*, Commun. Math. Phys. 74 (1980), pp. 41–59.
[20] H. Kesten, *Analyticity properties and power law estimates of functions in percolation theory*, J. Stat. Phys. 25 (1981), pp. 717–756.
[21] J.J. Ludlam, G.J. Gibson, W. Otten and C.A. Gilligan, *Applications of percolation theory to fungal spread with synergy*, J. R. Soc. Interface 9 (2012), pp. 949–956. Available at http://rsif.royalsocietypublishing.org/content/9/70/949.
[22] D. McFadden, *A method of simulated moments for estimation of discrete response models without numerical integration*, Econometrica 57 (1989), pp. 995–1026. Available at http://www.jstor.org/stable/1913621.
[23] W. Otten, D.J. Bailey and C.A. Gilligan, *Empirical evidence of spatial thresholds to control invasion of fungal parasites and saprotrophs*, New Phytol. 163 (2004), pp. 125–132.
[24] A. Pakes and D. Pollard, *Simulation and the asymptotics of optimization estimators*, Econometrica 57 (1989), pp. 1027–1057. Available at http://www.jstor.org/stable/1913622.
[25] D. Pollard, *Convergence of stochastic processes*, Springer Series in Statistics, Springer, New York, 1984.
[26] C. Rath, *DNA-Kopierprozess mit Thrombin-Aptamer Mikroarrays (in German)*, Master's thesis, Centre for Biological Systems Analysis (ZBSA), Faculty of Biology, University of Freiburg, Germany, 2014.
[27] M.F. Sykes and J.W. Essam, *Exact critical percolation probabilities for site and bond problems in two dimensions*, J. Math. Phys. 5 (1964), pp. 1117–1127. Available at http://scitation.aip.org/content/aip/journal/jmp/5/8/10.1063/1.1704215.

[28] S. van de Geer, *Applications of empirical process theory (Empirical processes in M-estimation)*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2010.
[29] J.C. Wierman, *Bond percolation on honeycomb and triangular lattices*, Adv. Appl. Probab. 13 (1981), pp. 298–313. Available at http://www.jstor.org/stable/1426685.

Appendix 1. Identifiability and numerical estimates of the selected moments

We outline why we conjecture that the parameter $\theta = (\lambda^1, \dots, \lambda^{n_c}, \mu)$ is identifiable from the moments $((\mathrm{E}Y_i^\ell)_{\ell\in\{1,2,\dots,n_c\}}, (\mathrm{E}[Y_i^\ell Y_j^\ell])_{\ell\in\{1,2,\dots,n_c\}})$ $(i \sim j)$. If we focus on just one colour ℓ, then the graph of the function $(\lambda^\ell, \mu) \mapsto \mathrm{E}Y_i^\ell$ on the domain $[0,1] \times [0, p_c]$ has level curves which go from high λ^ℓ and low μ to low λ^ℓ and high μ. In words, the density $\mathrm{E}Y_i^\ell$ of colour ℓ is constant if we compensate for a decreasing seeding rate λ^ℓ by an appropriately increasing contamination rate μ. The function $(\lambda^\ell, \mu) \mapsto \mathrm{E}[Y_i^\ell Y_j^\ell]$ $(i \sim j)$ has level curves with the same property.

However, we conjecture that the level curves of $\mathrm{E}Y_i^\ell$ and $\mathrm{E}[Y_i^\ell Y_j^\ell]$ do not coincide, instead any pair intersect in a single point. While either one of the two moments narrows down the possible value of the parameter vector to one of its level curves, the two moments jointly specify the intersection point of two level curves, which uniquely identifies the parameter value (λ^ℓ, μ).

We provide numerical evidence to back up this claim. For $n_c = 1$, we sampled $\mathrm{E}Y_i$ and $\mathrm{E}[Y_i Y_j]$ in 142 logarithmically spaced parameter vectors. We made an exception to the logarithmic rule to additionally sample along the line of critical μ (Figure A1). Dataset A contains a broader coverage of 100 parameter vectors. For each of these, we generated independently $n_s = 5$ realizations of the process on a lattice I' of size 300×300, and took its central 100×100 sublattice $I \subset I'$ as our data. $\mathrm{E}Y_i$ and $\mathrm{E}[Y_i Y_j]$ are estimated as averages over the central sublattice over $n_s = 5$ realizations.

In Dataset B, 56 parameter vectors are considered which have lower λ values in comparison with Dataset A, save for an overlap of 14 parameter vectors. For each vector, we generated independently $n_s = 5$ realizations of the process on a lattice I' of size 1500×1500, and its central 1000×1000 sublattice $I \subset I'$ serves as our data.

The sublattice sizes were selected such that in both datasets, the mean number of seeds is at least 5 in the central sublattice used for sampling, even for their respective lowest λ values ($\lambda = 5 \times 10^{-4}$ in Dataset A, and approximately 5.23×10^{-6} in Dataset B). At the larger lattice size used for Dataset

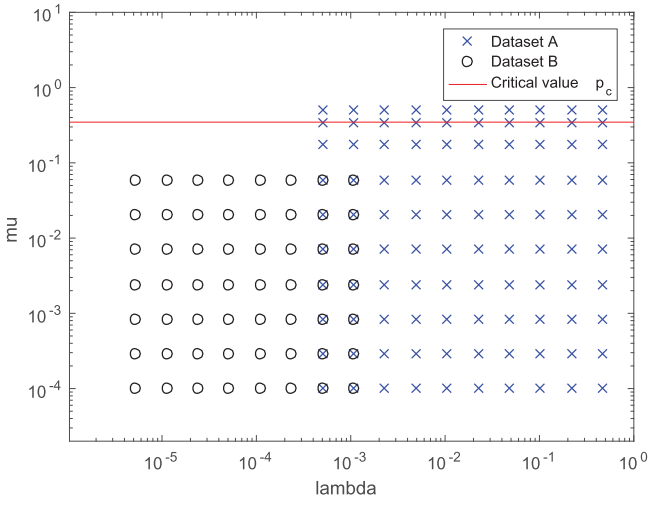

Figure A1. Sampled parameter values $\theta = (\lambda, \mu)$. Dataset A spans $[5 \times 10^{-4}, 0.4676] \times [10^{-4}, 0.5]$ and Dataset B spans $[5.23 \times 10^{-6}, 0.00107] \times [10^{-4}, 0.0595]$.

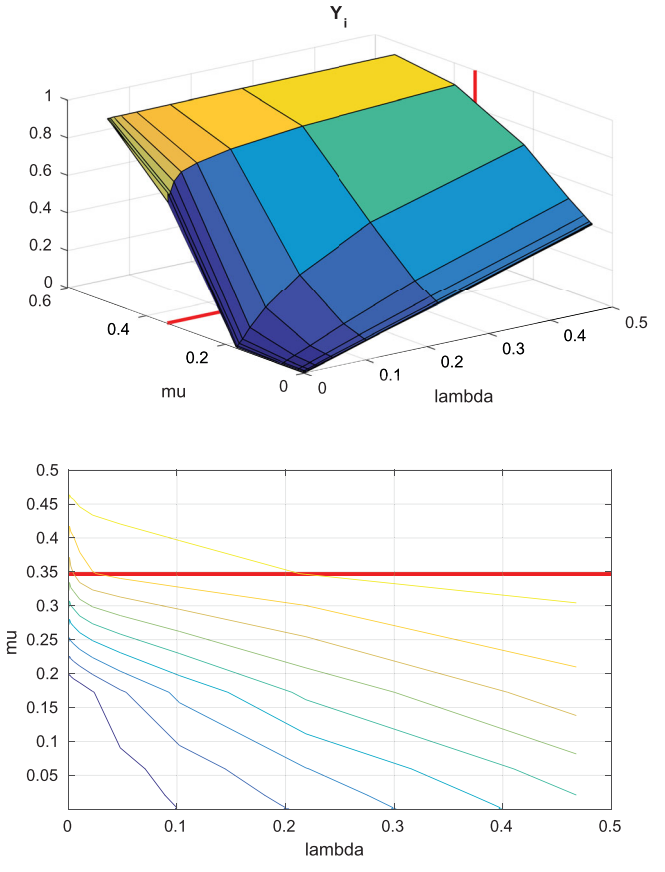

Figure A2. (*top*) Empirical means of Y_i for the various parameter vectors of Dataset A. (*bottom*) Level curves of this function. The red lines mark the critical value p_c.

B, for μ values larger than what we tested, the step of finding the connected open components to generate the data became prohibitively time-consuming.

Figures A2–A4 display graphs and level curves of the two coordinates of

$$(\lambda, \mu) \mapsto \left(\frac{1}{n_s n_I} \sum_{s=1}^{n_s} \sum_{i \in I} Y_i^s, \frac{1}{n_s n_p} \sum_{s=1}^{n_s} \sum_{(i,j) \in I_2} Y_i^s Y_j^s \right).$$

Close observation of the level curves seems to show that those in Figure A2 fan out with different slopes from a smaller region, while those in Figure A3 are closer to parallel. This supports our conjecture that level curves of one type intersect level curves of the other type in exactly one point, giving identifiability, except perhaps for a null set or otherwise small subset of Θ where the two types of level curves coincide.

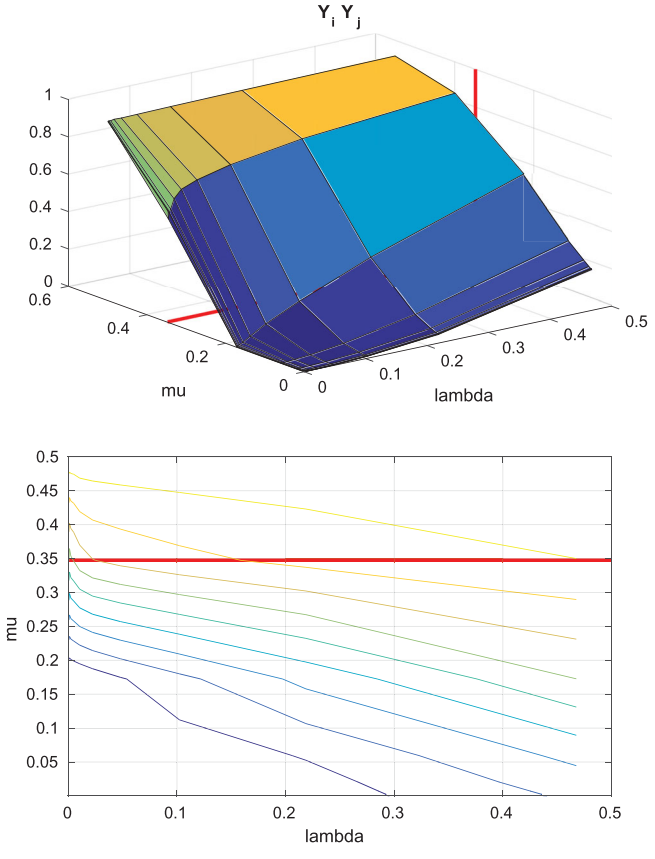

Figure A3. (*top*) Empirical means of Y_iY_j ($i \sim j$) for the various parameter vectors of Dataset A. (*bottom*) Level curves of this function. The red lines mark the critical value p_c.

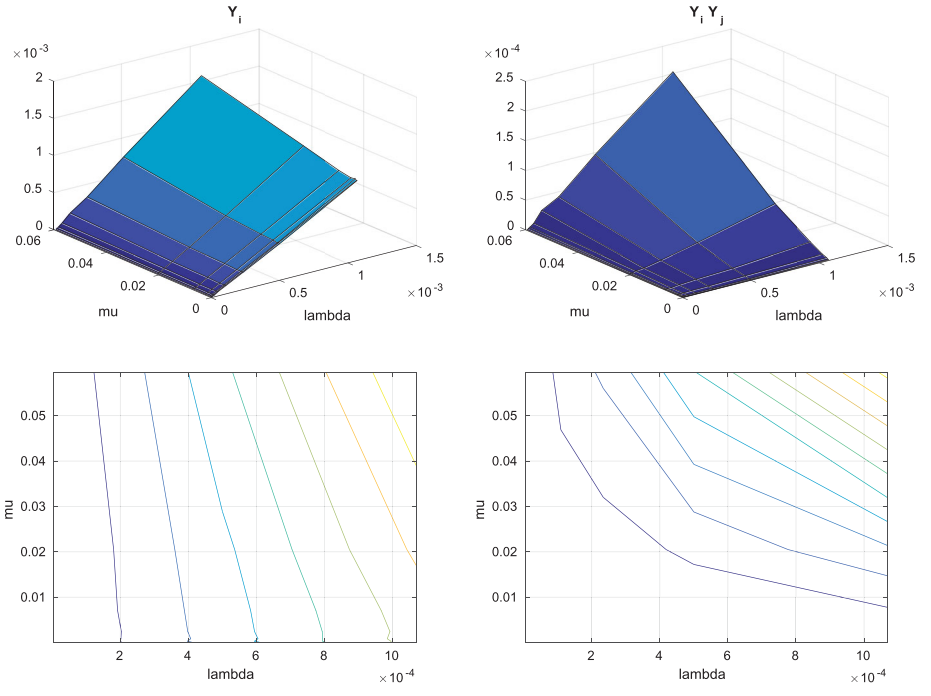

Figure A4. (*top left*) Empirical means of Y_i for the various parameter vectors of Dataset B. (*bottom left*) Level curves of this function. (*top right*) Empirical means of Y_iY_j ($i \sim j$) for the various parameter vectors of Dataset B. (*bottom right*) Level curves of this function.