Thesis for the Degree of Licentiate of Philosophy

Inference in a Partially Observed Percolation Process

Oscar Hammar

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Göteborg, Sweden 2010

Inference in a Partially Observed Percolation Process
Oscar Hammar

ISSN 1652-9715
Report 2010:44

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology
and University of Gothenburg
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Author e-mail: oscham@chalmers.se

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Göteborg, Sweden 2010

Inference in a Partially Observed Percolation Process
Oscar Hammar

Abstract

In this licentiate thesis, inference in a partially oberved percolation process living on a graph, is considered. Each edge of the graph is declared open with probability θ and closed with probability $1-\theta$ independently of the states of all other edges. The inference problem is to draw inference about θ based on the information on whether or not particular pairs of vertices are connected by open paths.

Consistency results under certain conditions on the graph are given for both a Bayesian and a frequentist approach to the inference problem. Moreover, a simulation study is presented which in addition to illustrating the consistency results, also indicates that the consistency results might hold for percolation processes on more general graphs.

Keywords: Percolation, Bayesian inference, frequentistic inference, consistency, Markov chain Monte Carlo, Monte Carlo Expectation Maximization.

Acknowledgment

Firstly, I would like to thank my supervisors Olle Häggström and Anastassia Baxevani. Also, I would like to thank Emilio Bergroth, Erik Kristiansson, Marcus Isaksson, Janeli Sarv and Wengang Mao. Finally, I would like to thank my family and my fiancee Kristina Polby for all your support.

Oscar Hammar
Göteborg, September 24, 2010

1 Introduction

Percolation theory, introduced in 1957 by Broadbent and Hammersley [2], is a branch of probability theory that describes the behaviour of connected clusters in a random graph. We consider a particular class of percolation processes called bond percolation. Staring with a fixed graph (finite or infinite) each edge is declared open with probability θ and closed with probability $1-\theta$ independently of the states of all other edges. The random graph containing only open edges is a realisation of a percolation process.

The majority of research on percolation theory has centred around the so called critical phenomena of percolation processes, the main example being the sudden change of the probability of existence of an infinite open cluster from 0 to 1 as a function of θ.

From an applied point of view, a percolation process can model disordered media of various kinds. Applications of percolation theory can be found in a variety of fields, mainly in the area of statistical mechanics but also in as diverse applications as animal movement [12] and groundwater hydrology [1].

Despite the wide range of applications of percolation theory, there is little published on statistical inference in percolation processes. Anyhow, there are some results. For example, Meester and Steif [17] presented consistent estimators for quantities of fully observed percolation processes on a particular kind of graph. Also, Larson [15] considered an estimation problem on a model related to the one we study.

The statistical inference problem considered in the present paper is the following. Given a realisation of the bond percolation process we observe for any of a number of pairs of vertices whether or not the pair is connected by an open path. Based on these observations we want to draw inference about θ.

As an illustration of the statistical inference problem, we may think of a fractured rock mass. The percolation process models a system of cracks in the rock mass. Each edge in the graph corresponds to a crack and water can penetrate a particular crack if and only if the corresponding edge is open. Several boreholes are drilled into the fractured rock mass and water is pumped into one borehole at a time. For each of the boreholes we observe whether or not water can penetrate to the others. Based on these observations we want to estimate the proportion of penetrable cracks.

We consider both a Bayesian and a frequentist approach to the inference about θ and prove a consistency result in each setting under certain conditions on the graph. In the Bayesian setting the result states that asymptotically the posterior accumulates almost surely around the 'true' value of θ as the number of observations increase. In the frequentist setting, the consistency result states that the maximum likelihood esti-

mate of θ converges almost surely to the true value as the number of observations increase.

We also present algorithms to compute the relevant quantities in the Bayesian and frequentist approach. In the Bayesian approach we use a Markov chain Monte Carlo method and in the frequentist approach we use a stochastic version of the EM-algorithm. We evaluate the accuracy of the Bayesian and the frequentist approaches to the inference problem in a simulation study.

The paper is organized as follows. In Section 2 we introduce the percolation process an the data from the process. We also introduce the particular class of graphs for which we develop the theory in the following sections. In Section 3 we present consistency results for percolation processes on the particular class of graphs. The proofs of these results are given in Section 4.

In Sections 5 and 6 we give background on some relevant simulation methods. We also introduce the Block Updating MCMC which is an MCMC algorithm suitable for the problem we consider. In Section 7 we prove that the Block Updating MCMC has the necessary properties to converge in an appropriate way.

In Section 8 we introduce the simulation study. In Section 9 and 10 we specify starting values, stopping rules and choices of parameters of the algorithms we use in the simulation study. The simulation results are given in Section 11. Section 12 is a conclusion of the paper.

2 The percolation process and the data

We begin with some definitions. A graph $G = (V, E)$ is an ordered pair of sets where the first set, V, is a finite or countably infinite set of vertices and the second set, E, is a set of edges.

An edge is defined to be an unordered pair of vertices. We represent an edge between vertices v and w by $\langle v, w\rangle$. A path between vertices v_0 and v_n is defined to be an alternating sequence $v_0, e_0, v_1, e_1, \ldots, e_{n-1}, v_n$ of distinct vertices v_l and edges $e_l = \langle v_l, v_{l+1}\rangle$, $l = 0, 1, \ldots, n-1$.

Two vertices are connected if there is a path between the vertices. A cluster is a set of connected vertices. A graph is connected if any two of its edges are connected. A graph is said to be finite if its set of edges has finitely many elements.

We use $V(G)$ and $E(G)$ to denote the vertex set and the edge set of the graph G. Two graphs, G and H, are said to be isomorphic if there is a bijection $b : V(G) \to V(H)$ such that any two vertices v and w in G are connected by an edge in G if and only if $b(v)$ and $b(w)$ are connected by an edge in H.

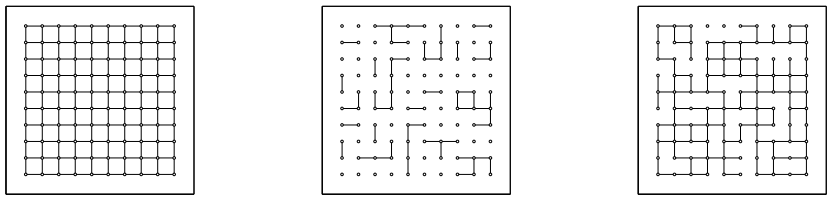

Figure 1: *Left:* A graph G on which a percolation process lives. *Middle:* A realisation $\mathbf{u}$ of the percolation process $\mathbf{U}$ on G generated by P_θ with $\theta = 0.3$. *Right:* A realisation $\mathbf{u}$ of the percolation process $\mathbf{U}$ on G generated by P_θ with $\theta = 0.7$.

2.1 The percolation process on a general finite graph

We now define a bond percolation process on a graph $G = (V, E)$. If $|E| = m$, then we define $\mathbb{U} = \{0, 1\}^m$. An element $\mathbf{u} = (u_1, \ldots, u_m)$ of $\mathbb{U}$ is thus an m-tuple of 0's and 1's. Each coordinate of $\mathbf{u}$ is associated with an edge $e \in E$ and a given edge $e \in E$ is said to be open if the corresponding coordinate of $\mathbf{u}$ is 1, and closed if the corresponding coordinate of $\mathbf{u}$ is 0. Consequently, each $\mathbf{u} \in \mathbb{U}$ corresponds to a graph obtained from G by removing all closed edges, i.e. removing those edges corresponding to a coordinate of $\mathbf{u}$ which is 0. We refer to an element $\mathbf{u}$ of $\mathbb{U}$ as a configuration and will not distinguish between a configuration $\mathbf{u}$ and the associated graph.

We now introduce a random variable $\mathbf{U}$ which we identify with the percolation process on G. Let $\mathbf{U}$ be defined on a sufficiently large sample space $(\Omega, \mathcal{F})$ and taking values in the set $\mathbb{U}$ of configurations. We let $o(\mathbf{u}) = \sum_{i=1}^{m} u_i$ denote the number of open edges in the configuration $\mathbf{u}$ and for each θ in the parameter space $\Theta = [0, 1] \subset \mathbb{R}$, we define the measure P_θ on $(\Omega, \mathcal{F})$ by

$$P_\theta(\mathbf{U} = \mathbf{u}) = \theta^{o(\mathbf{u})}(1 - \theta)^{m - o(\mathbf{u})}, \tag{2.1}$$

with m, as before denoting the total number of edges in the graph G. We also introduce the density (w.r.t counting measure) p_θ of the random variable $\mathbf{U}$:

$$p_\theta(\mathbf{u}) = P_\theta(\mathbf{U} = \mathbf{u}). \tag{2.2}$$

The right hand side of Equation (2.1) is a product of m terms; the number of terms equal to θ being $o(\mathbf{u})$ and the number of terms equal to $1 - \theta$ being $m - o(\mathbf{u})$. Thus from Equation (2.1) it follows that for a large value of θ, $p_\theta(\mathbf{u})$ is large if $o(\mathbf{u})$ - the number of open edges in $\mathbf{u}$ - is large, and if instead θ is small, $p_\theta(\mathbf{u})$ is large if $o(\mathbf{u})$ is small. Consequently, for a large value of θ, a typical realisation $\mathbf{u}$ of $\mathbf{U}$ has many 1's which

corresponds to a graph with many open edges and a typical realisation $\mathbf{u}$ of $\mathbf{U}$ has few 1's which corresponds to a graph with few open edges. Figure 1 gives an example of this for a percolation process on a tiny graph.

2.2 The data from a percolation process on a general finite graph $G = (V, E)$

In our set-up one does not observe the full realisation $\mathbf{u}$ of the percolation process $\mathbf{U}$. We consider the case when the data one observes is whether or not particular pairs of vertices are connected by an open path. We denote the event that there exists an open path between vertices v and w by $\{v \leftrightarrow w\}$.

Let $\mathcal{O}$ be any subset of the vertex set V with at least two elements and $\mathcal{D}$ be the set of unordered pairs of elements of $\mathcal{O}$. An element of $\mathcal{O}$ is referred to as an observation point and consequently an element of $\mathcal{D}$ is referred to as a pair of observation points.

We collect all indicator variables of connectedness of pair of observation points in a vector $\mathbf{X}$. Assume there are d pairs of observation points in $\mathcal{D}$ and fix an ordering of these pairs of observation points. For each $l \in \{1, \ldots, d\}$, we define the random variable X_l to be the indicator variable of connectedness of the l^{th} pair of observation points in $\mathcal{D}$. Thus, if (o_1, o_2) denotes the l^{th} pair of observation points of $\mathcal{D}$, then

$$X_l(\mathbf{U}) = I_{\{o_1 \leftrightarrow o_2\}}(\mathbf{U}).$$

The definition of X_l states that given a realisation $\mathbf{u}$ of $\mathbf{U}$, $X_l(\mathbf{u}) = 1$ if the l^{th} pair of vertices in $\mathcal{D}$ is connected by an open path in the graph corresponding to $\mathbf{u}$, and 0 otherwise.

We refer to the vector $\mathbf{X}(\mathbf{u}) = (X_1(\mathbf{u}), \ldots, X_d(\mathbf{u}))$ as the data from the percolation process on G and define the density

$$p_{\mathbf{X}|\theta}(\mathbf{x}) = P_\theta\{\mathbf{u} : \mathbf{X}(\mathbf{u}) = \mathbf{x}\}.$$

Assume a realisation $\mathbf{u}$ of the percolation process $\mathbf{U}$ has been generated according to P_θ, for some $\theta \in \Theta$. The observed values of the coordinates of $\mathbf{X}(\mathbf{u})$ carry some information about θ. We now explain this.

From the definition of P_θ, Equation (2.1), it follows that if the realisation $\mathbf{u}$ was generated by P_θ for a θ close to zero, then it is likely that the number of open edges, $o(\mathbf{u})$, is relatively small and the probability of these open edges creating an open path between two observation points o_1 and o_2 is small. In contrast, if the realisation $\mathbf{u}$ was generated by P_θ for a θ close to one, then it is likely the number of open edges, $o(\mathbf{u})$, is relatively large and the probability that there is an open path between the observation points o_1 and o_2 is close to 1. Consequently, observing

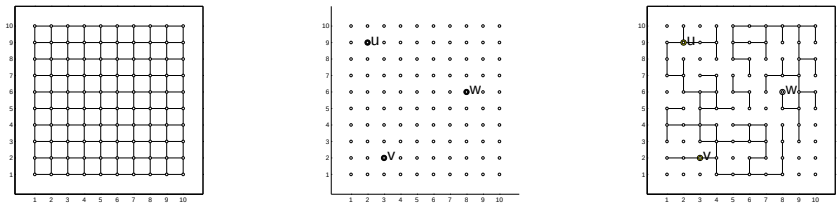

Figure 2: *Left:* A graph G on which a percolation process lives. *Middle:* An example of a of a set of observation points $\mathcal{O} = \{u, v, w\} \subset V$ is marked. *Right:* A realisation $\mathbf{u}$ of the percolation process $\mathbf{U}$ on G generated by P_θ with $\theta = 0.47$. This particular realisation gives $I_{\{u \leftrightarrow v\}}(\mathbf{u}) = 1$, $I_{\{u \leftrightarrow w\}}(\mathbf{u}) = 0$, $I_{\{v \leftrightarrow w\}}(\mathbf{u}) = 0$.

$I_{\{o_1 \leftrightarrow o_2\}}(\mathbf{u}) = 1$ makes large values of θ more plausible than small values of θ. This paper concerns the inference of $\theta \in \Theta$ based on the data $\mathbf{X}$ containing the indicator variables of connectedness of particular pairs of vertices.

The random variable $\mathbf{X}$ is well-defined as long as an ordering $\prec_{\mathcal{D}}$ of the elements of $\mathcal{D}$ is fixed. We now give an example of an ordering $\prec_{\mathcal{D}}$ in a special case and at the same time illustrate how the data $\mathbf{X}$ depends on the realisation $\mathbf{u}$ of $\mathbf{U}$. Keep in mind that the actual ordering is not of importance. What is important is the existence of an ordering.

Example 2.1. In this example we consider the percolation process on the graph G to the left in Figure 2. Each vertex of this graph is a vector with two coordinates, one for the x-direction and one for the y-direction.

Since an element of $\mathcal{D}$ is an unordered pair of elements of $\mathcal{O}$, it is natural to base an ordering $\prec_{\mathcal{D}}$ of the elements in $\mathcal{D}$ on an ordering $\prec_{\mathcal{O}}$ of the elements in $\mathcal{O}$. A natural ordering of the elements of $\mathcal{O}$ in this case is the lexicographic ordering: if (x_1, y_1) and (x_2, y_2) are two elements of $\mathcal{O}$, then $(x_1, y_1) \prec_{\mathcal{O}} (x_2, y_2)$ if and only if $x_1 < x_2$ or ($x_1 = x_2$ and $y_1 < y_2$).

Similarly, the ordering $\prec_{\mathcal{D}}$ defined from $\prec_{\mathcal{O}}$ can be given as follows. If (o_1, o_2) and (o_1', o_2') are two elements in $\mathcal{D}$, with $o_1 \prec_{\mathcal{O}} o_2$ and $o_1' \prec_{\mathcal{O}} o_2'$ then $(o_1, o_2) \prec_{\mathcal{D}} (o_1', o_2')$ if and only if $o_1 \prec_{\mathcal{O}} o_1'$ or ($o_1 = o_1'$ and $o_2 \prec_{\mathcal{O}} o_2'$).

For the set of observation points $\mathcal{O} = \{u, v, w\}$ with $u = (2, 9)$, $v = (3, 2)$ and $w = (8, 6)$ indicated in the middle picture of Figure 2, we have $u \prec_{\mathcal{O}} v \prec_{\mathcal{O}} w$ which implies the following ordering of the elements of $\mathcal{D}$: $(u, v) \prec_{\mathcal{D}} (u, w) \prec_{\mathcal{D}} (v, w)$.

Enumerating the elements of $\mathcal{D}$ in ascending order according to $\prec_{\mathcal{D}}$ gives for the particular realisation $\mathbf{u}$ of $\mathbf{U}$ in the graph to the right in Figure 2, $\mathbf{X}(\mathbf{u}) = (I_{u \leftrightarrow v}(\mathbf{u}), I_{u \leftrightarrow w}(\mathbf{u}), I_{v \leftrightarrow w}(\mathbf{u})) = (1, 0, 0)$. □

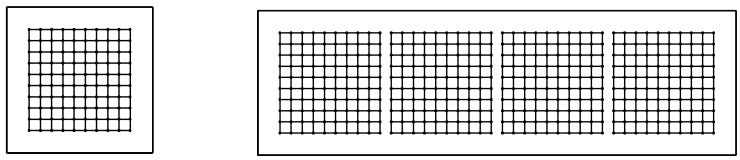

Figure 3: *Left:* A graph G. *Right:* The graph G^4, the union of four graphs isomorphic to G.

2.3 The graph G^n and the data $\mathbf{X}^n$

In Section 3 we present consistency results for the inference problem introduced in this section. The consistency results are shown for percolation processes on a particular class of graphs.

Let $G = (V, E)$ be a finite connected graph with vertex set $V = \{v_1, \ldots, v_N\}$. For each $k \in \mathbb{N}$, we define a graph $G_k = (V_k, E_k)$ with vertex set $V_k = \{v_{k,1}, \ldots, v_{k,N}\}$ isomorphic to G by means of the bijection $b_k(v_l) = v_{k,l}$.

The special class of graphs we consider in the next sections is, for $n \in \mathbb{N}$, $G^n = (V^n, E^n)$ with $V^n = \cup_{k=1}^n V_k$ and $E^n = \cup_{k=1}^n E_k$. Thus, G^n is the union of n copies of the graph G. We refer to G as the base graph of G^n and to G_k as the k^{th} primary subgraph of G^n. Figure 3 gives an example of the type of graph considered here.

For each $k \in \mathbb{N}$ we define $\mathcal{O}_k$, $\mathcal{D}_k$, $\mathbf{X}_k$, d_k and $p_{\mathbf{X}_k|\theta}$ analogously to the definitions of $\mathcal{O}$, $\mathcal{D}$, $\mathbf{X}$, d and $p_{\mathbf{X}|\theta}$ in Section 2.2. In particular, $\mathcal{O}_k$ is a subset of the vertices of the k^{th} primary subgraph and $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,d_k})$ contains all indicator variables on connectedness of pairs of elements in $\mathcal{O}_k$ and

$$p_{\mathbf{X}_k|\theta}(\mathbf{x}_k) = P_\theta\{\mathbf{u} : \mathbf{X}_k(\mathbf{u}) = \mathbf{x}_k\}.$$

Moreover, we define $\mathbf{X}^n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and the density

$$p_{\mathbf{X}^n|\theta}(\mathbf{x}^n) = P_\theta\{\mathbf{u} : \mathbf{X}^n(\mathbf{u}) = \mathbf{x}^n\}.$$

3 Two approaches to the inference problem

Given data from the percolation process we want to draw inference about θ. We consider both a Bayesian and a frequentist approach. In this section we introduce the two approaches and state a consistency result for a percolation process on the graph G^n for each approach.

3.1 The Bayesian approach

In the Bayesian approach, θ is viewed as a random variable. As before $\Theta = [0,1]$ denotes the parameter space of the percolation process and we let $\mathcal{B}(\Theta)$ denote the Borel σ-algebra on Θ. A probability distribution (the prior), here denoted by Π, reflecting the investigator's belief, is put on $(\Theta, \mathcal{B}(\Theta))$.

When data $\mathbf{X}$ are observed from the percolation process, the prior is updated to a posterior distribution, $\Pi(\cdot|\mathbf{X})$, given by, for $A \in \mathcal{B}(\Theta)$:

$$\Pi(A|\mathbf{X}) = \frac{\int_A p_{\mathbf{X}|\theta}(\mathbf{X})\Pi(d\theta)}{\int_\Theta p_{\mathbf{X}|\theta}(\mathbf{X})\Pi(d\theta)}.$$

A desirable property of a sequence of posterior distributions is that of consistency.

Definition 3.1. A sequence $\{\Pi_n\}_{n=1}^\infty$ of posterior distributions is said to be strongly consistent at θ if for each neighbourhood U of θ,

$$\lim_{n\to\infty} \Pi_n(U) = 1 \quad P_\theta\text{-a.s.}$$

Note that Definition 3.1 defines consistency at a particular θ. There are two types of consistency in the Bayesian setting. The weaker form is consistency for all parameter values θ in a set of prior measure 1. Thus, this form of Bayesian consistency does not guarantee consistency at a particular parameter value $\bar{\theta}$ of interest since this $\bar{\theta}$ may belong to a null set for which one does not have consistency.

The stronger form of Bayesian consistency is based on the frequentistic idea of a 'true' value of θ. This 'frequentistic' consistency of a Bayesian inference procedure means that if observations from the process would be generated under $\bar{\theta}$, then the posterior would accumulate in suitably defined neighbourhoods of $\bar{\theta}$. For a discussion of the two forms of Bayesian consistency see e.g. Ghosal [8].

We prove the stronger form, i.e. the 'frequentistic' consistency, of the Bayesian inference procedure. The proof of the following theorem is given in the next section.

Theorem 3.2. *Consider data $\mathbf{X}^n$ from a percolation process on the graph G^n. If the prior Π has support Θ, then the sequence of posterior distributions $\{\Pi(\cdot|\mathbf{X}^n)\}_{n=0}^\infty$ is consistent at all $\theta \in \Theta$.*

3.2 The frequentist approach

In the frequentist approach, when data $\mathbf{X}$ is observed from a percolation process, one wants to compute the maximum likelihood estimate (MLE), $\hat{\theta}(\mathbf{X})$, the maximizer of the likelihood function:

$$\hat{\theta}(\mathbf{X}) = \text{argmax}_{\theta\in\Theta} L(\theta, \mathbf{X}),$$

with $L(\theta, \mathbf{X}) = p_{\mathbf{X}|\theta}(\mathbf{X})$ viewed as a function of θ. As in the Bayesian approach, consistency is a desirable property.

Definition 3.3. A sequence of estimators h_n is strongly consistent for θ if it converges almost surely to θ, i.e. if

$$\lim_{n\to\infty} h_n = \theta \text{ a.s.}$$

The proof of the following theorem is also given in the next section.

Theorem 3.4. *Consider data $\mathbf{X}^n$ from a percolation process on the graph G^n. The sequence of maximum likelihood estimators $\{\hat{\theta}(\mathbf{X}^n)\}_{n=1}^{\infty}$ is strongly consistent for θ.*

3.3 Consistency for percolation processes on more general graphs

The results in this section states consistency for sequences of posteriors and MLE's from percolation processes on a particular type of graph, G^n, with a special structure.

The structure of G^n implies that all information on connectedness of pairs of observation points in the first n primary subgraphs is contained in $\mathbf{X}^n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and that $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent for $i \neq j$. This independence is important for the proofs of the consistency results given in Section 3.

It would be preferable to prove consistency also for percolation processes on more general graphs. Although we have not been able to prove such more general results, we present in the simulation study some numerical results, which indicate that the consistency theorems presented here may hold for more general graphs.

That one can not expect the consistency results to hold for inference from a percolation process on any graph is demonstrated in the following example.

Example 3.5. Consider a percolation process on the graph $G = (V, E)$ with $V = \mathbb{N}$ and $E = \{\langle k, k+1\rangle : k \in \mathbb{N}\}$ and with observation points $\mathcal{O} = \{2^k : k \in \mathbb{N}\}$. The special structure of G in this case implies that all information from the connectedness of the pairs of observation points in $\mathcal{D} = \{(o_1, o_2) \in \mathcal{O}^2 : o_1 < o_2\}$ is contained in the information of the connectedness of the observation points in the subset $\mathcal{D}' = \{(2^k, 2^{(k+1)}) : k \in \mathbb{N}\}$. Define for each pair of observation points in $\mathcal{D}'$ a random variable X'_k by $X'_k = I_{\{2^k \leftrightarrow 2^{k+1}\}}$ and let $\mathbf{X}'_n = (X'_1, \ldots, X'_n)$.

A necessary condition for infinitely many pairs of observation points in $\mathcal{D}'$ to be connected is that $P_\theta(\limsup_{k\to\infty} 2^k \leftrightarrow 2^{k+1}) \neq 0$. But for any $\theta \in \Theta \setminus \{1\}$ we have $\sum_{k\in\mathbb{N}} P_\theta(2^k \leftrightarrow 2^{k+1}) < \infty$, which by the first Borel-Cantelli Lemma [6] implies that $P_\theta(\limsup_{k\to\infty} 2^k \leftrightarrow 2^{k+1}) = 0$. We

conclude that, almost surely, the number of connected pairs of observation points in $\mathcal{D}'$ is finite. From this we conclude that there is a positive probability that no pair of $\mathcal{D}'$ is connected, i.e. $\mathbf{X}'_n = \mathbf{0}\ \forall n$.

Assume the event occurs that no pair of observation points in $\mathcal{D}'$ is connected. Then for any $\theta \in \Theta \setminus \{1\}$, the sequence of likelihoods $\{P_\theta(\mathbf{X}'_n = \mathbf{0})\}_{n=1}^{\infty}$ converges to a positive value. This means that the MLE is not consistent for any $\theta \in \Theta \setminus \{1\}$.

Likewise, for the Bayesian inference, since the posterior distribution is proportional to the likelihood function, the sequence of posterior distributions $\Pi(\cdot|\mathbf{X}_n)$ is not consistent for any $\theta \in \Theta \setminus \{1\}$. □

4 Proofs of Theorem 3.2 and Theorem 3.4

We commence the proofs of Theorems 3.2 and 3.4 with some definitions and preliminary results that are used in both proofs. The percolation process in Theorems 3.2 and 3.4 lives on the graph G^n, i.e. the union of n graphs, $G_1, \ldots, G_n$, that are isomorphic to a base graph G. The special structure of G^n implies that the elements $\mathbf{X}_i$ and $\mathbf{X}_j$ of the data $\mathbf{X}^n = (\mathbf{X}_1 \ldots, \mathbf{X}_n)$ from the percolation process on G^n are independent for $i \neq j$. The main idea of the proofs of Theorems 3.2 and 3.4 is to consider subsequences of $(\mathbf{X}_1 \ldots, \mathbf{X}_n)$ with independent and identically distributed elements.

Recall that V denotes the vertex set of the base graph G and let $\mathbb{O}$ denote the set of all subsets of V with at least two elements. Assume the cardinality of $\mathbb{O}$ is s and let $\mathcal{O}^{(1)}, \ldots, \mathcal{O}^{(s)}$ represent the elements of $\mathbb{O}$. Recall also the definition of the bijections b_k given in Section 2.3.

For $\Gamma \in \{1, \ldots, s\}$, we let $I_\Gamma = \{k \in \mathbb{N} : b_k^{-1}(\mathcal{O}_k) = \mathcal{O}^{(\Gamma)}\}$ denote the index set of primary subgraphs with observation points isomorphic to $\mathcal{O}^{(\Gamma)}$. We also define $I_\gamma^{(n)} = I_\gamma \cap \{1, \ldots, n\}$. Now, for each $\Gamma \in \{1, \ldots, s\}$, $(\mathbf{X}_\gamma)_{\gamma \in I_\Gamma^{(n)}}$ is a subsequence of $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$ with independent and identically distributed elements. If $\gamma \in I_\Gamma$, then we say that the primary subgraph G_γ generates a data vector of the Γ^{th} type.

We can now write the pmf $p_{\mathbf{X}^n|\theta}$ of the data $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$ as product of the pmf's of the iid subsequences,

$$p_{\mathbf{X}^n|\theta}(\mathbf{x}^n) = \prod_{\Gamma=1}^{s} \prod_{\gamma \in I_\Gamma^{(n)}} p_{\mathbf{X}_\gamma|\theta}(\mathbf{x}_\gamma).$$

Next, we give monotonicity results for two quantities, each used in the proof of one of Theorems 3.2 and 3.4. Recall that $\mathbf{X}_\gamma = (X_{\gamma,1}, \ldots, X_{\gamma,d_\gamma})$ is the data from the γ^{th} primary subgraph of G^n. In accordance with earlier notation, we let $p_{X_{\gamma,1}|\theta}$ denote the marginal pmf of the element $X_{\gamma,1}$ of $\mathbf{X}_\gamma$.

Lemma 4.1. *For γ in I_Γ for any $\Gamma \in \{1, \ldots, s\}$, the following holds.*

(a) The marginal pmf $p_{X_{\gamma,1}|\theta}(\cdot)$ of the first element of $\mathbf{X}_\gamma$ evaluated in 1 is strictly increasing in θ.

(b) The full pmf $p_{\mathbf{X}_\gamma|\theta}(\cdot)$ evaluated in $\mathbf{1} = (1, \ldots, 1)$ is strictly increasing in θ.

Proof. The assertions follow from a standard coupling argument. See e.g. Lindvall, page. 144 [16]. □

We let $B(\theta, \eta) = \{\theta' \in \Theta : |\theta - \theta'| < \eta\}$ denote the open ball in Θ with centre in θ and radius η and let $B^{\mathsf{c}}(\theta, \eta) = \Theta \setminus B(\theta, \eta)$ denote its complement.

4.1 Definitions and preliminary results for the proof of Theorem 3.2

Our proof of Theorem 3.2 is modelled after Choi and Ramamoorthi [4]. Before we proceed to the proof we need to fix some notation and some definitions.

Definition 4.2. Let f and g be two pmf's and $\mathcal{X}$ be the intersection of the support of f and g. The affinity between f and g is denoted $\mathrm{Aff}(f, g)$ and is given by

$$\mathrm{Aff}(f, g) = \sum_{x \in \mathcal{X}} \sqrt{f(x)g(x)}.$$

Note that $\mathrm{Aff}(f, g)$ for any two pmf's f and g satisfies $0 \leq \mathrm{Aff}(f, g) \leq 1$ and that $\mathrm{Aff}(f, g) = 1$ if and only if $f(x) \equiv g(x)$.

Definition 4.3. If f_θ is a pmf parametrized by $\theta \in \Theta$ and ν is a measure on Θ, then $q_\nu(f_\theta)$ is the pmf

$$q_\nu(f_\theta)(\cdot) = \int_\Theta f_\theta(\cdot)\nu(d\theta).$$

We state a property of the affinity between two marginal pmf's in relation to the affinity between the full pmf's which will be used several times in the proofs in the next section.

Lemma 4.4. *Let X_1 and X_2 be two discrete random vectors taking values in $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. If $f_{(X_1,X_2)}$ and $g_{(X_1,X_2)}$ are two joint pmf's of (X_1, X_2) and $f_{X_1}(\cdot) = \sum_{x_2 \in \mathcal{X}_2} f_{(X_1,X_2)}(\cdot, x_2)$ and $g_{X_1}(\cdot) = \sum_{x_2 \in \mathcal{X}_2} g_{(X_1,X_2)}(\cdot, x_2)$ are the corresponding marginal pmf's, then*

$$\mathit{Aff}(f_{(X_1,X_2)}, g_{(X_1,X_2)}) \leq \mathit{Aff}(f_{X_1}, g_{X_1}).$$

Proof. Let $f_{X_2|X_1}(\cdot|x_1) = f_{X_2}(\cdot)/f_{(X_1,X_2)}(x_1,\cdot)$ be the conditional pmf of $f_{(X_1,X_2)}$ and let $g_{X_1|X_2}(\cdot|x_2)$ be defined analogously. Then

$$\begin{aligned}
&\mathrm{Aff}(f_{(X_1,X_2)}, g_{(X_1,X_2)}) \\
&= \sum_{x_1\in\mathcal{X}_1}\sum_{x_2\in\mathcal{X}_2} \sqrt{f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1)g_{X_2|X_1}(x_2|x_1)g_{X_1}(x_1)} \\
&= \sum_{x_1\in\mathcal{X}_1} \sqrt{f_{X_1}(x_1)g_{X_1}(x_1)} \sum_{x_2\in\mathcal{X}_2} \sqrt{f_{X_2|X_1}(x_2|x_1)g_{X_2|X_1}(x_2|x_1)} \\
&= \sum_{x_1\in\mathcal{X}_1} \sqrt{f_{X_1}(x_1)g_{X_1}(x_1)}\mathrm{Aff}(f_{X_2|X_1}(\cdot|x_1)g_{X_2|X_1}(\cdot|x_1)) \\
&\le \sum_{x_1\in\mathcal{X}_1} \sqrt{f_{X_1}(x_1)g_{X_1}(x_1)} \\
&= \mathrm{Aff}(f_{X_1}, g_{X_1}).
\end{aligned}$$

where the inequality follows from the fact that $0 \le \mathrm{Aff}(f,g) \le 1$ for any pmf's f and g. □

4.2 Proof of Theorem 3.2

To prove Theorem 3.2 we need to show that for any $\bar{\theta} \in \Theta$ and any $\eta > 0$,

$$\Pi(B(\bar{\theta},\eta)|\mathbf{X}^n) \to 1 \quad P_{\bar{\theta}}\text{-a.s.}$$

or equivalently that for any $\bar{\theta} \in \Theta$ and any $\eta > 0$,

$$\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n) \to 0 \quad P_{\bar{\theta}}\text{-a.s.}$$

We write $\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n)$ as

$$\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n) = \frac{J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)}{J(\mathbf{X}^n)},$$

where

$$J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) = \int_{B^{\mathsf{c}}(\bar{\theta},\eta)} \frac{p_{\mathbf{X}^n|\theta}(\mathbf{X}^n)}{p_{\mathbf{X}^n|\bar{\theta}}(\mathbf{X}^n)}\Pi(d\theta)$$

and

$$J(\mathbf{X}^n) = \int_{\Theta} \frac{p_{\mathbf{X}^n|\theta}(\mathbf{X}^n)}{p_{\mathbf{X}^n|\bar{\theta}}(\mathbf{X}^n)}\Pi(d\theta).$$

We deal with the numerator and the denominator of $\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n)$ separately. Lemma 4.5 below takes care of the denominator of $\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n)$. It is a standard ingredient in proofs of consistency results in the Bayesian setting. For a proof see Ghosh and Ramamoorhi [9].

Lemma 4.5. *If the support of* Π *is* Θ*, then for all* $\beta > 0$*,*

$$\lim_{n\to\infty} e^{n\beta} J(\mathbf{X}^n) = \infty \quad P_{\bar{\theta}}\text{-a.s.}$$

Now fix $\bar{\theta} \in \Theta$ and note that Theorem 3.2 will follow from Lemma 4.5 if we also prove that for any $\eta > 0$ and some $\beta_0 > 0$

$$\lim_{n\to\infty} e^{n\beta_0} J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) = 0 \quad P_{\bar{\theta}}\text{-a.s.} \tag{4.1}$$

To prove the assertion in Equation (4.1) we use the affinity introduced above. We establish in Lemma 4.8 below that for any $\eta > 0$ and for any probability measure ν on $B^{\mathsf{c}}(\bar{\theta}, \eta)$ and for all sufficiently large n,

$$\mathrm{Aff}(p_{\mathbf{X}^n|\bar{\theta}}, q_\nu(p_{\mathbf{X}^n|\theta})) < e^{-\alpha n} \quad \text{for some } \alpha > 0. \tag{4.2}$$

Then, by relating the affinity, $\mathrm{Aff}(p_{\mathbf{X}^n|\bar{\theta}}, q_\nu(p_{\mathbf{X}^n|\theta}))$, for a particular choice of measure ν, to the expectation $\mathrm{E}_{\bar{\theta}}[J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)^{\frac{1}{2}}]$ we can in Lemma 4.9 below prove the convergence in Equation (4.1).

We now discuss how we prove Equation (4.2), which is done in a few steps. We first fix $\Gamma \in \{1, \ldots, s\}$ and consider the subsequence $(\mathbf{X}_\gamma)_{\gamma \in I_\Gamma^{(n)}}$ of $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$ of identically distributed elements and show in Lemma 4.6 below that for any $\eta > 0$ and any probability measure ν on $B^{\mathsf{c}}(\bar{\theta}, \eta)$

$$\mathrm{Aff}(p^{(2)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(2)}_{\mathbf{X}_\gamma|\theta})) < \delta \quad \text{for some } \delta > 0. \tag{4.3}$$

The next step is to show that for any $\eta > 0$ and any probability measure ν on $B^{\mathsf{c}}(\bar{\theta}, \eta)$ and for all $m \in \mathbb{N}$

$$\mathrm{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta})) < ke^{-m\beta} \quad \text{for some } \beta > 0, k > 0.$$

This is done in Lemma 4.7.

Then, since $(\mathbf{X}_\gamma)_{\gamma \in I_\Gamma^{(n)}}$ with pmf $p^{(m)}_{\mathbf{X}_\gamma|\bar{\theta}}$, where $m = |I_\Gamma^{(n)}|$, is a subsequence of $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$ with pmf $p_{\mathbf{X}^n|\bar{\theta}}$, the assertion in Equation (4.2) can be proven from the assertion in Equation (4.2) by using the relation between the affinity of the pmf's of a subsequence and the pmf's of the full sequence stated in Lemma 4.4.

The reason that we use the second power of $p_{\mathbf{X}_\gamma|\bar{\theta}}$ instead of the first power in Equation (4.3) is that the corresponding statement for the first power of $p_{\mathbf{X}_\gamma|\bar{\theta}}$ does not hold in general.

Lemma 4.6. *Fix* Γ *in* $\{1, \ldots, s\}$ *and let* $\gamma \in I_\Gamma$*. For any* $\eta > 0$ *there is a* $\delta > 0$ *such that for any probability measure* ν *on* $B^{\mathsf{c}}(\bar{\theta}, \eta)$ *we have*

$$\mathit{Aff}(p^{(2)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(2)}_{\mathbf{X}_\gamma|\theta})) < \delta. \tag{4.4}$$

Proof. Fix $\eta > 0$. Recall that $p_{X_{\gamma,1}|\theta}$ denotes the marginal pmf of the first element of $\mathbf{X}_\gamma = (X_{\gamma,1}, \ldots, X_{\gamma,d_\gamma})$. To prove Equation (4.4) it suffice, by Lemma 4.4, to show that there is a $\delta > 0$ such that for any probability measure ν on $B^{\mathsf{c}}(\bar{\theta}, \eta)$ we have

$$\mathrm{Aff}(p^{(2)}_{X_{\gamma,1}|\bar{\theta}}, q_\nu(p^{(2)}_{X_{\gamma,1}|\theta})) < \delta. \tag{4.5}$$

For notational simplicity we let in this proof $f_\theta = p_{X_{\gamma,1}|\theta}$ and rewrite Equation (4.5) as

$$\mathrm{Aff}(f^{(2)}_\theta, q_\nu(f^{(2)}_\theta)) < \delta.$$

We apply a proof of contradiction. Assume there is a probability measure μ on $B^{\mathsf{c}}(\bar{\theta}, \eta)$ such that $\mathrm{Aff}(f^{(2)}_\theta, q_\mu(f^{(2)}_\theta)) = 0$, or equivalently that $f^{(2)}_\theta$ and $q_\mu(f^{(2)}_\theta)$ are identical. In particular the probability measure μ satisfies

$$\left\{ \begin{array}{rcl} q_\mu(f^{(2)}_\theta)(1,1) & = & f^{(2)}_{\bar{\theta}}(1,1) \\ q_\mu(f^{(2)}_\theta)(1,0) & = & f^{(2)}_{\bar{\theta}}(1,0). \end{array} \right. \tag{4.6}$$

Partition $B^{\mathsf{c}}(\bar{\theta}, \eta)$ into $\Theta^- = [0, \bar{\theta} - \eta]$ and $\Theta^+ = [\bar{\theta} + \eta, 1]$. For any probability measure ν on $B^{\mathsf{c}}(\bar{\theta}, \eta)$ the corresponding pmf $q_\nu(f^{(2)}_\theta)$ can be written as a weighted average of two pmf's $f^{(2)}_{\theta^-}$ and $f^{(2)}_{\theta^+}$ for some $\theta^- \in \Theta^-$ and $\theta^+ \in \Theta^+$:

$$q_\nu(f^{(2)}_\theta)(x_1, x_2) = f^{(2)}_{\theta^-}(x_1, x_2)\nu(\Theta^-) + f^{(2)}_{\theta^+}(x_1, x_2)\nu(\Theta^+), \tag{4.7}$$

where $\theta^- = \int_{\Theta^-} \theta\nu(d\theta)$ and $\theta^+ = \int_{\Theta^+} \theta\nu(d\theta)$. In particular, the pmf $q_\mu(f^{(2)}_\theta)$ has the representation

$$q_\mu(f^{(2)}_\theta)(x_1, x_2) \quad = \quad f^{(2)}_{\theta^-}(x_1, x_2)\mu(\Theta^-) + f^{(2)}_{\theta^+}(x_1, x_2)\mu(\Theta^+) \tag{4.8}$$

for some $\theta^- \in \Theta^-$ and $\theta^+ \in \Theta^+$. Recall that $f^{(2)}_\theta(x_1, x_2)$ represents $f_\theta(x_1)f_\theta(x_2)$ and define $\mu^- = \mu(\Theta^-)$. Now Equation (4.8) can be rewritten as

$$q_\mu(f^{(2)}_\theta)(x_1, x_2) \quad = \quad f_{\theta^-}(x_1)f_{\theta^-}(x_2)\mu^- + f_{\theta^+}(x_1)f_{\theta^+}(x_2)(1 - \mu^-). \tag{4.9}$$

Now, let for $\theta \in \Theta$, $p(\theta)$ represent the success probability $p(\theta) = f_\theta(1)$ and define $p_0 = p(\bar{\theta})$, $p^- = p(\theta^-)$ and $p^+ = p(\theta^+)$. If we let $g(p) = p^2$ and $h(p) = p(1-p)$, then the system of equations (4.6) can by Equation (4.9) be expressed as

$$\left\{ \begin{array}{rcl} g(p^-)\mu^- + g(p^+)(1 - \mu^-) & = & g(p_0) \\ h(p^-)\mu^- + h(p^+)(1 - \mu^-) & = & h(p_0). \end{array} \right. \tag{4.10}$$

Solving these equations for μ^- gives

$$\left\{ \begin{array}{rcl} \mu^- & = & (g(p_0) - g(p^+))/(g(p^-) - g(p^+)) \\ \mu^- & = & (h(p_0) - h(p^+))/(h(p^-) - h(p^+)). \end{array} \right. \tag{4.11}$$

Recall that $f_\theta(1) = p_{X_{\gamma,1}|\theta}(1)$, which by Lemma 4.1(a) is monotonically increasing in θ. Therefore, we have $p^- < p_0 < p^+$ and we can thus write $p_0 = tp^- + (1-t)p^+$ for some $t \in (0,1)$. Noticing that $g(p) = p^2$ is a convex function, we have that the first equation of (4.11) gives

$$\begin{aligned} \mu^- &= \frac{g(tp^- + (1-t)p^+) - g(p^+)}{g(p^-) - g(p^+)} \\ &< \frac{tg(p^-) + (1-t)g(p^+) - g(p^+)}{g(p^-) - g(p^+)} \\ &= t. \end{aligned}$$

Noticing that $h(p) = p(1-p)$ is a concave function, a similar computation shows that the second equation of (4.11) gives $\mu^- > t$. This is a contradiction and the proof is completed. □

Lemma 4.7. *Fix Γ in $\{1,\ldots,s\}$ and let $\gamma \in I_\Gamma$. For any $\eta > 0$ and any probability measure ν on $B^{\mathsf{c}}(\bar{\theta},\eta)$ and for all $m \in \mathbb{N}$*

$$\mathit{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta})) < ke^{-m\beta} \quad \text{for some } \beta > 0, k > 0.$$

Proof. Fix $\eta > 0$. The first step of the proof is to show that for some $\beta > 0$

$$\begin{aligned} \mathrm{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta})) < e^{-m\beta} &\text{ for all } \textit{even } m \text{ and} \\ &\text{for all prob. measures } \nu \text{ on } B^{\mathsf{c}}(\bar{\theta},\eta). \end{aligned} \tag{4.12}$$

We show this by induction on even m. The base case, when $m = 2$,

$$\mathrm{Aff}(p^{(2)}_{\mathbf{X}_\gamma|\bar{\theta}}, q_\nu(p^{(2)}_{\mathbf{X}_\gamma|\theta})) < e^{-2m} \text{ for all probability measures } \nu \text{ on } B^{\mathsf{c}}(\bar{\theta},\eta), \tag{4.13}$$

follows from Lemma 4.6 with $\beta = -\frac{1}{2}\log(\delta)$. The induction hypothesis is

$$\mathrm{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\theta}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta})) < e^{-m\beta} \text{ for all probability measures } \nu \text{ on } B^{\mathsf{c}}(\bar{\theta},\eta). \tag{4.14}$$

Assuming the induction hypothesis is true we show that

$$\mathrm{Aff}(p^{(m+2)}_{\mathbf{X}_\gamma|\theta}, q_\nu(p^{(m+2)}_{\mathbf{X}_\gamma|\theta})) < e^{-(m+2)\beta} \text{ for all prob. measures } \nu \text{ on } B^{\mathsf{c}}(\bar{\theta},\eta). \tag{4.15}$$

For this, we write the affinity of the $(m+2)^{th}$ power of $p_{\mathbf{X}_\gamma|\theta}$ in terms of the affinity of the second and m^{th} powers of $p_{\mathbf{X}_\gamma|\theta}$. Let $\mathcal{X}$ denote the range of the random variable $\mathbf{X}_\gamma$ and write the affinity of the $(m+2)^{th}$

power of $p_{\mathbf{X}_\gamma|\theta}$ as

$$\begin{aligned}
&\mathrm{Aff}(p^{(m+2)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m+2)}_{\mathbf{X}_\gamma|\theta})) \\
&= \sum_{\mathbf{x}^m\in\mathcal{X}^m} p^{(m)}_{\mathbf{X}_\gamma|\bar\theta}(\mathbf{x}^m)^{\frac{1}{2}} \\
&\times \sum_{\mathbf{x}^2\in\mathcal{X}^2} p^{(2)}_{\mathbf{X}_\gamma|\bar\theta}(\mathbf{x}^2)^{\frac{1}{2}} \int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(2)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^2)^{\frac{1}{2}} p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}} \nu(d\theta).
\end{aligned} \tag{4.16}$$

Now, multiplying and dividing each term in the summation over $\mathcal{X}^m$ by $\int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}_m)^{\frac{1}{2}}\nu(d\theta)$ gives that

$$\begin{aligned}
&\mathrm{Aff}(p^{(m+2)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m+2)}_{\mathbf{X}_\gamma|\theta})) \\
&= \sum_{\mathbf{x}^m\in\mathcal{X}^m} p^{(m)}_{\mathbf{X}_\gamma|\bar\theta}(\mathbf{x}^m)^{\frac{1}{2}} \int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}} \nu(d\theta) \\
&\times \sum_{\mathbf{x}^2\in\mathcal{X}^2} p^{(2)}_{\mathbf{X}_\gamma|\bar\theta}(\mathbf{x}^2)^{\frac{1}{2}} \int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(2)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^2)^{\frac{1}{2}} \frac{p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}}}{\int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}}\nu(d\theta)} \nu(d\theta).
\end{aligned} \tag{4.17}$$

For each fixed outcome $\mathbf{x}^m$, the sum in the last row of Equation (4.17) is $\mathrm{Aff}(p^{(2)}_{\mathbf{X}_\gamma|\bar\theta}, q_{\nu_{\mathbf{x}^m}}(p^{(2)}_{\mathbf{X}_\gamma|\bar\theta}))$ where the measure $\nu_{\mathbf{x}^m}$ over Θ has density with respect to ν given by

$$\frac{p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}}}{\int_{B^{\mathsf{c}}(\bar\theta,\eta)} p^{(m)}_{\mathbf{X}_\gamma|\theta}(\mathbf{x}^m)^{\frac{1}{2}}\nu(d\theta)}.$$

Thus, by the base case, Equation (4.13), for each fixed outcome $\mathbf{x}_m$, the sum in the last row of Equation (4.17) is smaller than $e^{-2\beta}$. The sum in the second last row of (4.17) is $\mathrm{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta}))$ which by the induction hypothesis, Equation (4.14), is smaller than $e^{-m\beta}$. We conclude that $\mathrm{Aff}(p^{(m+2)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m+2)}_{\mathbf{X}_\gamma|\theta})) < e^{-(m+2)\beta}$ for all even m.

To complete the proof we need to consider also uneven powers of $p_{\mathbf{X}_\gamma|\theta}$. Let m be even and notice that by Lemma 4.4, with $k = e^\beta$, we also have

$$\mathrm{Aff}(p^{(m+1)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m+1)}_{\mathbf{X}_\gamma|\theta})) \le \mathrm{Aff}(p^{(m)}_{\mathbf{X}_\gamma|\bar\theta}, q_\nu(p^{(m)}_{\mathbf{X}_\gamma|\theta})) \le e^{-m\beta} = ke^{-(m+1)\beta}.$$

□

Lemma 4.8. *For any $\eta > 0$, for any probability measure ν on $B^{\mathsf{c}}(\bar\theta,\eta)$ and for all sufficiently large n,*

$$\mathit{Aff}(p_{\mathbf{X}^n|\bar\theta}, q_\nu(p_{\mathbf{X}^n|\theta})) < e^{-\alpha n} \quad \textit{for some } \alpha > 0.$$

Proof. As noted in the beginning of this section, the pmf of the data $\mathbf{X}^n$ can be expressed as

$$p_{\mathbf{X}^n|\theta}(\mathbf{x}^n) = \prod_{\Gamma=1}^{s} \prod_{\gamma \in I_\Gamma^{(n)}} p_{\mathbf{X}_\gamma|\theta}(\mathbf{x}_\gamma),$$

where for each $\Gamma \in \{1, \ldots, s\}$, $\prod_{\gamma \in I_\Gamma^{(n)}} p_{\mathbf{X}_\gamma|\theta}(\mathbf{x}_\gamma)$ is the pmf of a number of identically distributed vectors. Define c_Γ for $\Gamma \in \{1, \ldots, s\}$ by $c_\Gamma = \frac{1}{2} \liminf_{n\to\infty}(|I_\Gamma^{(n)}|/n)$. For some Γ we have $c_\Gamma > 0$. Assume for notational simplicity that $c_1 > 0$ and let $m_n = |I_1^{(n)}|$.

Consider now the subsequence $(\mathbf{X}_\gamma)_{\gamma \in I_1^{(n)}}$ of $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$ of identically distributed vectors with pmf $\prod_{\gamma \in I_1^{(n)}} p_{\mathbf{X}_\gamma|\theta} = p_{\mathbf{X}_1|\bar{\theta}}^{(m_n)}$. By Lemma 4.4 the affinity of two marginal pmf's is greater or equal to the affinity of the full pmf's, and hence

$$\mathrm{Aff}(p_{\mathbf{X}^n|\bar{\theta}}, q_\nu(p_{\mathbf{X}^n|\theta})) \quad \leq \quad \mathrm{Aff}(p_{\mathbf{X}_1|\bar{\theta}}^{(m_n)}, q_\nu(p_{\mathbf{X}_1|\theta}^{(m_n)})) \tag{4.18}$$

Moreover, by Lemma 4.7 we have that for all sufficiently large m_n

$$\mathrm{Aff}(p_{\mathbf{X}_1|\bar{\theta}}^{(m_n)}, q_\nu(p_{\mathbf{X}_1|\theta}^{(m_n)})) \quad \leq \quad k e^{-m_n \beta}. \tag{4.19}$$

Furthermore, for sufficiently large n, we have that $m_n = |I_1^{(n)}|$ is sufficiently much larger than nc_1 and for such n we have

$$k e^{-m_n \beta} \quad \leq \quad e^{-c_1 n \beta} \tag{4.20}$$

Thus, with $\alpha = c_1 \beta$ the lemma follows from equations (4.18), (4.19) and (4.20). □

Lemma 4.9. *For any $\eta > 0$, for some $\beta_0 > 0$*

$$\lim_{n\to\infty} e^{n\beta_0} J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) = 0 \quad P_{\bar{\theta}}\text{-a.s.}$$

Proof. Recall $J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) = \int_{B^{\mathsf{c}}(\bar{\theta},\eta)} \frac{p_{\mathbf{X}^n|\theta}(\mathbf{X}^n)}{p_{\mathbf{X}^n|\bar{\theta}}(\mathbf{X}^n)} d\Pi(\theta)$. Let $\Pi^\star$ denote the probability measure obtained by restricting Π to $B^{\mathsf{c}}(\bar{\theta}, \eta)$ and normalizing it. The key to the proof is to observe that

$$\mathrm{E}_{\bar{\theta}}[(J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n))^{\frac{1}{2}}] = \Pi(B^{\mathsf{c}}(\bar{\theta}, \eta))^{\frac{1}{2}} \mathrm{Aff}(p_{\mathbf{X}^n|\bar{\theta}}, q_{\Pi^\star}(p_{\mathbf{X}^n|\theta})), \tag{4.21}$$

which follows directly from expanding the right hand side by the definition of the affinity and using that

$$q_{\Pi^\star}(p_{\mathbf{X}^n|\theta})(\cdot) = \int_{B^{\mathsf{c}}(\bar{\theta},\eta)} p_{\mathbf{X}^n|\theta}(\cdot) \Pi(d\theta) / \Pi(B^{\mathsf{c}}(\bar{\theta}, \eta)). \tag{4.22}$$

We have

$$\begin{aligned} P_{\bar{\theta}}(J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)^{\frac{1}{2}} > e^{-n\gamma}) &\leq e^{n\gamma}\,\mathrm{E}_{\bar{\theta}}[J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)^{\frac{1}{2}}] \\ &= e^{n\gamma}\Pi(B^{\mathsf{c}}(\bar{\theta},\eta))^{\frac{1}{2}}\mathrm{Aff}(p_{\mathbf{X}^n|\bar{\theta}}, q_{\Pi^\star}(p_{\mathbf{X}^n|\theta})) \\ &\leq \Pi(B^{\mathsf{c}}(\bar{\theta},\eta))^{\frac{1}{2}} e^{n\gamma} e^{-n\alpha} \end{aligned} \tag{4.23}$$

for some $\alpha > 0$. The first inequality follows from the fact that the probability that a non-negative random variable is larger then 1 is less then the expectation of it. The equality follows from the key observation in Equation (4.21) and the last inequality follows from Lemma 4.8.

Now we chose γ so that $0 < \gamma < \alpha$ and conclude that

$$\sum_{n=1}^{\infty} P_{\bar{\theta}}(J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)^{\frac{1}{2}} > e^{-n\gamma}) < \Pi(B^{\mathsf{c}}(\bar{\theta},\eta))^{\frac{1}{2}} \sum_{n=1}^{\infty} e^{(\gamma-\alpha)n} < \infty.$$

By the first Borel-Cantelli Lemma (see Durrett [6]) we conclude that $P_{\bar{\theta}}(J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)^{\frac{1}{2}} > e^{-n\gamma}\, i.o.) = 0$ which implies that $P_{\bar{\theta}}(J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) > e^{-n\gamma}\, i.o.) = 0$, where i.o. stands for infinitely often. Thus, almost surely with respect to $P_{\bar{\theta}}$, for sufficiently large n, $J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) < e^{-n\gamma}$ and we conclude that, if $\beta_0 < \gamma$, then $\lim_{n\to\infty} e^{n\beta_0} J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n) = 0$ $P_{\bar{\theta}}$-a.s. □

We now prove Theorem 3.2

Proof. It suffice to prove that for any $\bar{\theta} \in \Theta$ and $\eta > 0$,

$$\lim_{n\to\infty} \Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n) = 0 \quad P_{\bar{\theta}}\text{-a.s.}$$

We have

$$\Pi(B^{\mathsf{c}}(\bar{\theta},\eta)|\mathbf{X}^n) = \frac{J_{B^{\mathsf{c}}(\bar{\theta},\eta)}(\mathbf{X}^n)}{J(\mathbf{X}^n)},$$

which, if the support of the prior Π is Θ, by Lemma 4.5 and Lemma 4.9 (by taking $\beta_0 = \beta$), converges to zero as $n \to \infty$ almost surely with respect to $P_{\bar{\theta}}$. □

4.3 Definitions and preliminary results for the proof of Theorem 3.4

Our proof of Theorem 3.4 is modelled after Lachout et. al. [13]. Before we proceed to the proof we need some definitions and preliminary results.

Definition 4.10. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{f_n\}_{n=1}^{\infty}$ a sequence of functions $f_n : \Theta \times \Omega \to \mathbb{R} \cup \{+\infty\}$, which can be regarded as random functions $\{f_n(\cdot)\}_{n=1}^{\infty}$, where $f_n(x)$ stands for $f_n(x, \theta)$. The

sequence $\{f_n\}_{n=1}^{\infty}$ is a lower semi-continuous approximation almost surely to $f : \Theta \to \mathbb{R} \cup \{+\infty\}$ on $\Theta' \subset \Theta$ if and only if

$$P\left\{ \sup_{V \in \mathcal{N}(\theta_0)} \liminf_{n\to\infty} \inf_{t \in V} f_n(t) \geq f(\theta_0) \, \forall \theta_0 \in \Theta' \right\} = 1,$$

where $\mathcal{N}(\theta_0)$ is the system of neighbourhoods of θ_0.

If $\{f_n\}_{n=1}^{\infty}$ is a lower semi-continuous approximation almost surely to $f : \Theta \to \mathbb{R} \cup \{+\infty\}$ on $\Theta' \subset \Theta$ we write

$$f_n \overset{\text{l. on } \Theta'}{\to} f \text{ a.s.}$$

The following Theorem is a rewritten form of Theorem 2.1 of Lachout et. al. [13].

Theorem 4.11. *Suppose $\hat{\theta}_n$, for each $n \in \mathbb{N}$, is the minimizer of f_n, i.e.,*

$$f_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} f_n(\theta). \tag{4.24}$$

and that Θ is compact. Furthermore, assume that there is a function $f : \Theta \to \mathbb{R} \cup \{+\infty\}$ with an unique minimum in $\theta = \bar{\theta}$ such that

$$f_n \overset{\text{l. on } \Theta\setminus\{\bar{\theta}\}}{\to} f \text{ a.s.} \tag{4.25}$$

and

$$\lim_{n\to\infty} f_n(\bar{\theta}) = f(\bar{\theta}) \text{ a.s.} \tag{4.26}$$

Then

$$\lim_{n\to\infty} \hat{\theta}_n = \bar{\theta} \text{ a.s.}$$

4.4 Proof of Theorem 3.4

As in the proof of Theorem 3.2 the strategy is to consider independent and identically distributed subsequences $(\mathbf{X}_\gamma)_{\gamma \in I_\Gamma^{(n)}}$, $\Gamma \in \{1, \ldots, s\}$, of the data $\mathbf{X}^n = (\mathbf{X}_i)_{i=1}^n$. Recall that d_Γ is the length of the data vector $\mathbf{X}_\gamma = (X_{\gamma,1}, \ldots, X_{\gamma,d_\Gamma})$, for $\gamma \in I_\Gamma$, and that each component $X_{\gamma,i}$ of $\mathbf{X}_\gamma$ can take the value 0 or 1. Let $m_\Gamma = 2^{d_\Gamma}$ denote the number of different outcomes of $\mathbf{X}_\gamma$, for $\gamma \in I_\gamma$, and let $\mathbf{b}_{\Gamma,1}, \ldots, \mathbf{b}_{\Gamma,m_\Gamma}$ represent these different vectors.

For each $\Gamma \in \{1, \ldots, s\}$, we define a probability vector $\mathbf{p}_\Gamma$ where the i^{th} element corresponds to the probability of the outcome $\mathbf{b}_{\Gamma,i}$ of the random variable $\mathbf{X}_\gamma$, for $\gamma \in I_\Gamma$:

$$\mathbf{p}_\Gamma(\theta) = (p_{\Gamma,1}(\theta), \ldots, p_{\Gamma,m_\Gamma}(\theta)) = (p_{\mathbf{X}_\gamma|\theta}(\mathbf{b}_{\Gamma,1}), \ldots, p_{\mathbf{X}_\gamma|\theta}(\mathbf{b}_{\Gamma,m_\Gamma})).$$

We also define for each $\Gamma \in \{1, \ldots, s\}$ and $n \in \mathbb{N}$ a vector $\mathbf{a}_\Gamma^{(n)}$ where the i^{th} element corresponds to the fraction of outcomes equal to $\mathbf{b}_{\Gamma,i}$ of the random variables $(\mathbf{X}_\gamma)_{\gamma \in I_\Gamma^{(n)}}$:

$$\begin{aligned}\mathbf{a}_\Gamma^{(n)} &= (a_{\Gamma,1}^{(n)}, \ldots, a_{\Gamma,m_\Gamma}^{(n)}) \\ &= \left(\frac{|\{\gamma \in I_\Gamma^{(n)} : \mathbf{X}_\gamma = \mathbf{b}_{\Gamma,1}\}|}{|I_\Gamma^{(n)}|}, \ldots, \frac{|\{\gamma \in I_\Gamma^{(n)} : \mathbf{X}_\gamma = \mathbf{b}_{\Gamma,m_\Gamma}\}|}{|I_\Gamma^{(n)}|} \right).\end{aligned}$$

Note that by the strong law of large numbers, if $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty$, then $\lim_{n\to\infty} \mathbf{a}_\Gamma^{(n)} = \mathbf{p}_\Gamma(\bar{\theta})$ $P_{\bar{\theta}}$-a.s.

Define $g_n(\theta, \omega) = (\frac{L(\bar{\theta}, \mathbf{X}^n)}{L(\theta, \mathbf{X}^n)})^{\frac{1}{n}}$ to be the $\left(\frac{1}{n}\right)^{th}$ power of the likelihood ratio of $\bar{\theta}$ and θ of data from the first n primary subgraphs from a percolation process on G^n. If we also define $g_n^{(\Gamma)}$ to be the likelihood ratio of the data vectors of the Γ^{th} type among the first n primary subgraphs,

$$g_n^{(\Gamma)}(\theta, \omega) = \prod_{i=1}^{m_\Gamma} \left(\frac{p_{\Gamma,i}(\bar{\theta})}{p_{\Gamma,i}(\theta)} \right)^{a_{\Gamma,i}^{(n)}},$$

then $g_n(\theta, \omega)$ can be written as

$$g_n(\theta, \omega) = \prod_{\Gamma=1}^{s} \left(g_n^{(\Gamma)}(\theta, \omega) \right)^{\frac{|I_\Gamma^{(n)}|}{n}}.$$

This is seen from the following computation:

$$\begin{aligned} g_n(\theta, \omega) &= \left(\frac{L(\bar{\theta}, \mathbf{X}^n)}{L(\theta, \mathbf{X}^n)} \right)^{\frac{1}{n}} \\ &= \left(\frac{p_{\mathbf{X}^n|\bar{\theta}}(\mathbf{X}^n)}{p_{\mathbf{X}^n|\theta}(\mathbf{X}^n)} \right)^{\frac{1}{n}} \\ &= \left(\prod_{\Gamma=1}^{s} \prod_{\gamma \in I_\Gamma^{(n)}} \frac{p_{\mathbf{X}|\bar{\theta}}(\mathbf{X}_\gamma)}{p_{\mathbf{X}|\theta}(\mathbf{X}_\gamma)} \right)^{\frac{1}{n}} \\ &= \left(\prod_{\Gamma=1}^{s} \prod_{i=1}^{m_\Gamma} \left(\frac{p_{\Gamma,i}(\bar{\theta})}{p_{\Gamma,i}(\theta)} \right)^{|\{\gamma \in I_\Gamma^{(n)} : \mathbf{X}_\gamma = \mathbf{b}_{\Gamma,i}\}|} \right)^{\frac{1}{n}} \\ &= \prod_{\Gamma=1}^{s} \left(\prod_{i=1}^{m_\Gamma} \left(\frac{p_{\Gamma,i}(\bar{\theta})}{p_{\Gamma,i}(\theta)} \right)^{a_{\Gamma,i}^{(n)}} \right)^{\frac{|I_\Gamma^{(n)}|}{n}} \\ &= \prod_{\Gamma=1}^{s} \left(g_n^{(\Gamma)}(\theta, \omega) \right)^{\frac{|I_\Gamma^{(n)}|}{n}}. \end{aligned} \tag{4.27}$$

For each Γ we let $\mathbf{b}_{\Gamma,1}$ represent the vector $\mathbf{1} = (1, \ldots, 1)$ containing d_Γ 1's and define

$$f^{(\Gamma)}(\theta) = \Big(\frac{p_{\Gamma,1}(\bar{\theta})}{p_{\Gamma,1}(\theta)}\Big)^{p_{\Gamma,1}(\bar{\theta})} \Big(\frac{1-p_{\Gamma,1}(\bar{\theta})}{1-p_{\Gamma,1}(\theta)}\Big)^{1-p_{\Gamma,1}(\bar{\theta})}$$

Moreover we define

$$f_{\min}(\theta) = \min_{\Gamma \in \{1,\ldots,s\}} f^{(\Gamma)}(\theta).$$

Below, we show that that

$$g_n \overset{\text{l. on } \Theta\setminus\{\bar{\theta}\}}{\to} \frac{1}{2} + \frac{1}{2} f_{\min} \; P_{\bar{\theta}}\text{-a.s.} \tag{4.28}$$

and that $\frac{1}{2} + \frac{1}{2} f_{\min}$ has a unique minimum in $\bar{\theta}$. Moreover we show that

$$\lim_{n\to\infty} g_n(\bar{\theta}) = \frac{1}{2} + \frac{1}{2} f_{\min}(\bar{\theta}) \; P_{\bar{\theta}}\text{-a.s.} \tag{4.29}$$

Since by definition, the MLE $\hat{\theta}_n$ is the maximizer of the likelihood function, and thus, the minimizer of g_n, the assertion of Theorem 3.4 then follows from Theorem 4.11. We first state a lemma about properties of the function $f_{\min}$.

Lemma 4.12. *The function $f_{min} : \Theta \to \mathbb{R} \cup +\infty$ has a unique minimum in $\theta = \bar{\theta}$. Moreover, f_{min} is monotonically decreasing for $\theta < \bar{\theta}$ and monotonically increasing for $\theta > \bar{\theta}$, and continuous. Furthermore, if we define*

$$g^{(\Gamma)}(\theta) = \prod_{i=1}^{m_\Gamma} \Big(\frac{p_{\Gamma,i}(\bar{\theta})}{p_{\Gamma,i}(\theta)}\Big)^{p_{\Gamma,i}(\bar{\theta})}.$$

then

$$f_{min}(\theta) \leq g^{(\Gamma)}(\theta) \; \forall\, \theta \; \forall\, \Gamma \tag{4.30}$$

Proof. Define for some constant $\bar{p}_1$,

$$\phi(p_1) = \frac{\bar{p}_1^{\bar{p}_1}(1-\bar{p}_1)^{1-\bar{p}_1}}{p_1^{\bar{p}_1}(1-p_1)^{1-\bar{p}_1}}, \tag{4.31}$$

and note that with $\bar{p}_1 = p_{\Gamma,1}(\bar{\theta})$,

$$f^{(\Gamma)}(\theta) = \phi(p_{\Gamma,1}(\theta)). \tag{4.32}$$

Taking derivatives of $\phi(p_1)$ shows that $\phi(p_1)$ is monotonically decreasing for $p_1 < \bar{p_1}$ and monotonically increasing for $p_1 > \bar{p_1}$. Since, by Lemma 4.1 (b), $p_{\Gamma,1}(\theta)$ is monotonically increasing, it follows that $f^{(\Gamma)}(\theta) = \phi(p_{\Gamma,1}(\theta))$, and thus $f_{\min}(\theta)$, is monotonically decreasing for $\theta < \bar{\theta}$ and

monotonically increasing for $\theta > \bar{\theta}$. The continuity of $f_{\min}(\theta)$ follows from the continuity of ϕ and $p_{\Gamma,1}(\theta)$ for all Γ.

We now prove Equation (4.30). Firstly, note that by definition, $f_{\min} \leq f^{(\Gamma)}$ for all Γ. We show $f^{(\Gamma)} \leq g^{(\Gamma)}$ for all Γ. Define, for some constant vector $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_m)$,

$$\delta(\mathbf{p}) = \frac{\prod_{i=1}^m \bar{p}_i^{\bar{p}_i}}{\prod_{i=1}^m p_i^{\bar{p}_i}}, \tag{4.33}$$

and note that with $\bar{\mathbf{p}} = \mathbf{p}_\Gamma(\bar{\theta})$

$$g^{(\Gamma)}(\theta) = \delta(\mathbf{p}_\Gamma(\theta)). \tag{4.34}$$

To compare $f^{(\Gamma)}(\theta) = \phi(p_{\Gamma,1}(\theta))$ and $g^{(\Gamma)}(\theta) = \delta(\mathbf{p}_\Gamma(\theta))$, we consider first the maximal possible value of the denominator of δ for a fixed value of p_1 under the constraints $\sum_{i=1}^m p_i = 1, p_i \in [0,1]$. By Lemma 4.13,

$$\operatorname*{arg\,max}_{(p_2,\ldots,p_m):\sum_{i=1}^m p_i = 1, p_i \in [0,1]} \prod_{i=1}^m p_i^{\bar{p}_i} = \frac{1-p_1}{1-\bar{p}_1}(\bar{p}_2, \ldots, \bar{p}_m)$$

and the maximum value is therefore

$$\begin{aligned}\max_{(p_2,\ldots,p_m):\sum_{i=1}^m p_i = 1, p_i \in [0,1]} \prod_{i=1}^m p_i^{\bar{p}_i} &= p_1^{\bar{p}_1} \prod_{i=2}^m \Big[\Big(\frac{1-p_1}{1-\bar{p}_1}\Big)\bar{p}_i\Big]^{\bar{p}_i} \\ &= p_1^{\bar{p}_1}\Big(\frac{1-p_1}{1-\bar{p}_1}\Big)^{1-\bar{p}_1} \prod_{i=2}^m \bar{p}_i^{\bar{p}_i}.\end{aligned}$$

We thus have, for any probability vector $\mathbf{p} = (p_1, \ldots, p_m)$,

$$\begin{aligned}\delta(\mathbf{p}) &= \frac{\prod_{i=1}^m \bar{p}_i^{\bar{p}_i}}{\prod_{i=1}^m p^{\bar{p}_i}} \\ &\geq \frac{\prod_{i=1}^m \bar{p}_i^{\bar{p}_i}}{p_1^{\bar{p}_1}\big(\frac{1-p_1}{1-\bar{p}_1}\big)^{1-\bar{p}_1} \prod_{i=2}^m \bar{p}_i^{\bar{p}_i}} \\ &= \Big(\frac{\bar{p}_1}{p_1}\Big)^{\bar{p}_1}\Big(\frac{1-\bar{p}_1}{1-p_1}\Big)^{1-\bar{p}_1} \\ &= \phi(p_1)\end{aligned}$$

which shows that for any θ, $g^{(\Gamma)}(\theta) = \delta(\mathbf{p}_\Gamma(\theta)) \geq \phi(p_{\Gamma,1}(\theta)) = f^{(\Gamma)}(\theta)$. □

Next, we state a lemma by which we, among other things, can derive a lower bound for g_n.

Lemma 4.13. *The function $h : [0,1]^n \to \mathbb{R}$ defined by $h(p_1, \ldots, p_n) = \prod_{i=1}^n p_i^{a_i}$, for non-negative constants a_i, subject to the constraint $\sum_{i=1}^n p_i = C$ is maximized for $\mathbf{p} = (p_1, \ldots, p_n)$ proportional to $\mathbf{a} = (a_1, \ldots, a_n)$, i.e., if $\sum_{i=1}^n a_i = K$, then*

$$\underset{\mathbf{p}:\sum_{i=1}^n p_i = C}{\arg\max} \; h(\mathbf{p}) = \frac{C}{K}\mathbf{a}.$$

Proof. We use Lagrange multipliers. We restrict the domain of to $(0,1]^n$. It is obvious that h does not take it's maximum value on the boundary $[0,1]^n \setminus (0,1]^n$. Let $g(\mathbf{p}) = \sum_{i=1}^n p_i$ and introduce the Lagrange function

$$\Lambda(\mathbf{p}, \lambda) = h(\mathbf{p}) + \lambda(g(\mathbf{p}) - C).$$

Setting all partial derivatives of Λ to zero we get

$$\begin{array}{rcl} \frac{a_1}{p_1} h(\mathbf{p}) & = & \lambda \\ \frac{a_2}{p_2} h(\mathbf{p}) & = & \lambda \\ & \vdots & \\ \frac{a_n}{p_n} h(\mathbf{p}) & = & \lambda \\ \sum_{i=1}^n p_i & = & C. \end{array} \tag{4.35}$$

Since we have restricted the domain of h to $(0,1]^n$ it is legitimate to divide each of the $n-1$ first equations by the second last equation in the system of Equations (4.35). This gives $p_i = \frac{a_i}{a_n} p_n$ for $i = 1 \ldots, n-1$. Insertion into the last equation of the system of Equations (4.35) gives

$$\begin{aligned} \frac{a_1}{a_n} p_n + \ldots + \frac{a_{n-1}}{a_n} p_n + p_n = C &\Rightarrow a_1 p_n + \ldots + a_n p_n = a_n C \\ &\Rightarrow p_n = a_n \frac{C}{K}. \end{aligned} \tag{4.36}$$

If we instead would have chosen to divide all the first n (except the j^{th}) equations of the system of Equations (4.35) by the j^{th} equation and inserted into the last equation we would have obtained $p_j = a_j \frac{C}{K}$. This shows that $\mathbf{p} = \frac{C}{K}\mathbf{a}$ is a stationary point of Λ. Inspection gives that it is in fact a maxima of Λ and thus a maxima of h under the given constraints. □

Lemma 4.14. *For $C \subset \Theta = [0,1]$, such that 0 or 1 are not contained in the closure of C,*

$$\liminf_{n\to\infty} \inf_{\theta \in C} \big(g_n(\theta, \omega) - f_{min}(\theta)\big) \geq 0 \; P_{\bar{\theta}}\text{-a.s.} \tag{4.37}$$

Proof. Recall that g_n is a product of (powers of) likelihood ratios of iid sequences:

$$g_n(\theta, \omega) = \prod_{\Gamma=1}^{s} \big(g_n^{(\Gamma)}(\theta, \omega)\big)^{\frac{|I_\Gamma^{(n)}|}{n}} \quad \text{where} \quad g_n^{(\Gamma)}(\theta, \omega) = \prod_{i=1}^{m_\Gamma} \Big(\frac{p_{\Gamma,i}(\bar{\theta})}{p_{\Gamma,i}(\theta)}\Big)^{a_{\Gamma,i}^{(n)}}.$$

We consider first the likelihood ratios of the iid sequences $g_n^{(\Gamma)}$ and define as in Lemma 4.12,

$$g^{(\Gamma)}(\theta) = \prod_{i=1}^{m_\Gamma} \Big(\frac{p_{\Gamma,i}(\bar\theta)}{p_{\Gamma,i}(\theta)}\Big)^{p_{\Gamma,i}(\bar\theta)},$$

and prove that for all Γ such that $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty$,

$$\lim_{n\to\infty} \sup_{\theta\in C} \big| g_n^{(\Gamma)}(\theta,\omega) - g^{(\Gamma)}(\theta)\big| = 0. \tag{4.38}$$

Having proved Equation (4.38) the assertion of the lemma, which concerns the likelihood ratio of all data, easily follows.

Since 0 and 1 are not in the closure of C we have that for some $\delta > 0$,

$$p_{\Gamma,i}(\theta) \geq \delta \ \forall\theta \in C\, \forall\Gamma\, \forall i,$$

which implies that

$$\frac{p_{\Gamma,i}(\bar\theta)}{p_{\Gamma,i}(\theta)} \leq \frac{1}{\delta} \ \forall\theta \in C\, \forall\Gamma\, \forall i \tag{4.39}$$

and we conclude that for all Γ such that $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty$,

$$\begin{aligned}
& \lim_{n\to\infty} \sup_{\theta\in C} \big| g_n^{(\Gamma)}(\theta,\omega) - g^{(\Gamma)}(\theta)\big| && (4.40)\\
= \ & \lim_{n\to\infty} \sup_{\theta\in C} \Big| \prod_{i=1}^{m_\Gamma} \Big(\frac{p_{\Gamma,i}(\bar\theta)}{p_{\Gamma,i}(\theta)}\Big)^{a_{\Gamma,i}^{(n)}} - \prod_{i=1}^{m_\Gamma} \Big(\frac{p_{\Gamma,i}(\bar\theta)}{p_{\Gamma,i}(\theta)}\Big)^{p_{\Gamma,i}(\bar\theta)} \Big| && (4.41)\\
\leq \ & \lim_{n\to\infty} \sup_{\theta\in C} \Big| \prod_{i=1}^{m_\Gamma} \Big(\frac{1}{\delta}\Big)^{a_{\Gamma,i}^{(n)}} - \prod_{i=1}^{m_\Gamma} \Big(\frac{1}{\delta}\Big)^{p_{\Gamma,i}(\bar\theta)} \Big| && (4.42)\\
= \ & 0\, P_{\bar\theta}\text{-a.s.} && (4.43)
\end{aligned}$$

where the inequality follows from Equation (4.39) and the last equality follows since, by the law of large numbers, $\lim_{n\to\infty} a_{\Gamma,i}^{(n)} = p_{\Gamma,i}(\bar\theta)\, \forall i$, for all Γ such that $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty\, P_{\bar\theta}$-a.s. Thus we have proved Equation (4.38).

We now prove Equation (4.37), the assertion of the lemma. From Equation (4.38), we conclude that for an arbitrary $\beta > 0$ and for $P_{\bar\theta}$-almost all $\omega \in \Omega$, there is an $n_0(\omega)$, such that for all $n > n_0$ and for all Γ such that $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty$,

$$\inf_{\theta\in C} g_n^{(\Gamma)}(\theta,\omega) \geq \inf_{\theta\in C} g^{(\Gamma)}(\theta) - \beta \tag{4.44}$$

Assume for notational simplicity that $\lim_{n\to\infty} I_\Gamma^{(n)} = \infty$ for $\Gamma \in \{1,\ldots,v\}$ and $\lim_{n\to\infty} I_\Gamma^{(n)} < \infty$ for $\Gamma \in \{v{+}1,\ldots,s\}$, where v obviously is positive,

and note that

$$\liminf_{n\to\infty}\inf_{\theta\in C}\prod_{\Gamma=v+1}^{s}\left(g_n^{(\Gamma)}(\theta,\omega)\right)^{\frac{|I_\Gamma^{(n)}|}{n}}=1. \tag{4.45}$$

We now have

$$\begin{aligned}
&\liminf_{n\to\infty}\inf_{\theta\in C}\left(g_n(\theta,\omega)-f_{\min}(\theta)\right)\\
&=\liminf_{n\to\infty}\inf_{\theta\in C}\Big(\prod_{\Gamma=1}^{v}\left(g_n^{(\Gamma)}(\theta,\omega)\right)^{\frac{|I_\Gamma^{(n)}|}{n}}\prod_{\Gamma=v+1}^{s}\left(g_n^{(\Gamma)}(\theta,\omega)\right)^{\frac{|I_\Gamma^{(n)}|}{n}}-f_{\min}(\theta)\Big)\\
&\geq\liminf_{n\to\infty}\inf_{\theta\in C}\Big(\prod_{\Gamma=1}^{v}\left(g^{(\Gamma)}(\theta)-\beta\right)^{\frac{|I_\Gamma^{(n)}|}{n}}-f_{\min}(\theta)\Big)\\
&=\liminf_{n\to\infty}\inf_{\theta\in C}\Big(\left(g^{(\Gamma)}(\theta)-\beta\right)^{\sum_{\Gamma=1}^{v}\frac{|I_\Gamma^{(n)}|}{n}}-f_{\min}(\theta)\Big)\\
&=\inf_{\theta\in C}\Big(g^{(\Gamma)}(\theta)-\beta-f_{\min}(\theta)\Big)\\
&>-\beta
\end{aligned} \tag{4.46}$$

where we have used Equations (4.44) and (4.45) for the first inequality and Equation (4.30) for the last. Since β can be chosen arbitrarily small the statement of the lemma follows. □

We now have the following result

Lemma 4.15. *It holds that*

$$g_n \overset{l.\ on\ \Theta\setminus\{\bar\theta\}}{\to} \frac{1}{2}+\frac{1}{2}f_{min}\ P_{\bar\theta}\text{-}a.s. \tag{4.47}$$

Proof. To prove Equation (4.47), we need to show that

$$P\left\{\sup_{V\in\mathcal{N}(\theta_0)}\liminf_{n\to\infty}\inf_{t\in V}g_n(t)\geq\frac{1}{2}+\frac{1}{2}f_{\min}(\theta_0)\,\forall\theta_0\in\Theta\setminus\{\bar\theta\}\right\}=1\ P_{\bar\theta}\text{-a.s.} \tag{4.48}$$

We consider the cases $\theta_0<\bar\theta$ and $\theta_0>\bar\theta$ separately and start with the case $\theta_0<\bar\theta$. By Lemma 4.14, for small $\rho>0$ (such that 0 or 1 are not contained in the closure of $B(\theta_0,\rho)$),

$$\liminf_{n\to\infty}\inf_{\theta\in B(\theta_0,\rho)}\left(g_n(\theta,\omega)-f_{\min}(\theta)\right)\geq 0\ P_{\bar\theta}\text{-a.s.}$$

which implies

$$\liminf_{n\to\infty}\inf_{\theta\in B(\theta_0,\rho)}g_n(\theta,\omega)\geq\inf_{\theta\in B(\theta_0,\rho)}f_{\min}(\theta)\ P_{\bar\theta}\text{-a.s.} \tag{4.49}$$

Equation (4.49) together with $f_{\min}$ being continuous, monotone, and strictly larger than 1 (by Lemma 4.12), imply that ($P_{\bar{\theta}}$-a.s.) for a sufficiently small $\rho > 0$

$$\liminf_{n\to\infty} \inf_{\theta\in B(\theta_0,\rho)} g_n(\theta) \geq \inf_{\theta\in B(\theta_0,\rho)} f_{\min}(\theta) = f_{\min}(\theta_0+\rho) \geq \frac{1}{2} + \frac{1}{2} f_{\min}(\theta_0),$$

which shows

$$P\left\{ \sup_{V\in\mathcal{N}(\theta_0)} \liminf_{n\to\infty} \inf_{t\in V} g_n(t) \geq \frac{1}{2} + \frac{1}{2} f_{\min}(\theta_0) \text{ for } \theta_0 < \bar{\theta} \right\} = 1 \; P_{\bar{\theta}}\text{-a.s.} \tag{4.50}$$

The proof of the corresponding statement for $\theta_0 > \bar{\theta}$ is analogous. □

We now prove Theorem 3.4

Proof. We verify that all conditions of Theorem 4.11 are satisfied. By definition, the MLE $\hat{\theta}_n = \hat{\theta}(\mathbf{X}^n)$ is the maximizer of the likelihood function $L(\theta, \mathbf{X}^n)$ and thus $\hat{\theta}_n$ is the minimizer of $g_n(\theta,\omega) = (\frac{L(\bar{\theta},\mathbf{X}^n)}{L(\theta,\mathbf{X}^n)})^{\frac{1}{n}}$.

By Lemma 4.12, $\theta = \bar{\theta}$ is the unique minimizer of $f_{\min}$ and thus also the unique minimizer of $\frac{1}{2} + \frac{1}{2} f_{\min}$ By Lemma 4.15,

$$g_n \overset{\text{l. on } \Theta\setminus\{\bar{\theta}\}}{\to} \frac{1}{2} + \frac{1}{2} f_{\min} \; P_{\bar{\theta}}\text{-a.s.}$$

Since $\frac{1}{2} + \frac{1}{2} f_{\min}(\bar{\theta}) = 1$ and $g_n(\bar{\theta}) = 1$ for all n we trivially have that

$$\lim_{n\to\infty} g_n(\bar{\theta}) = \frac{1}{2} + \frac{1}{2} f_{\min}(\bar{\theta}) \; P_{\bar{\theta}}\text{-a.s.}$$

Thus all conditions of Theorem 4.11 are satisfied and we conclude that $\lim_{n\to\infty} \hat{\theta}_n = \bar{\theta} \; P_{\bar{\theta}}$-a.s. □

5 Markov Chains

We have now proved two consistency results for inference in a percolation process. In Section 11 we present results for the inference procedures on simulated data. To compute the relevant quantities for this inference, we use Markov chain Monte Carlo algorithms. For completeness of presentation, we recall some Markov chain theory. One of our Markov chain Monte Carlo algorithms lives on an uncountable state space and we therefore recall theory for Markov chains on general state spaces.

Let S denote a general (i.e. not necessarily countable) state space. Throughout the presentation S will be an r-dimensional space, where $r > 1$ and we use boldface letters for vector-valued variables taking values in S.

From an informal perspective, a Markov chain on S is a sequence of random variables $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$, such that the distribution of $\mathbf{Y}_{n+1}$ is independent of $(\mathbf{Y}_0, \ldots, \mathbf{Y}_{n-1})$, given $\mathbf{Y}_n$.

In a more formal setting, assume that random variables $\mathbf{Y}_i$, $i = 0, 1, \ldots$ are defined on a sample space $(\Omega, \mathcal{F})$ with a probability measure P and take values in $(S, \mathcal{S})$, where $\mathcal{S}$ denotes a σ-algebra on S. A **transition kernel** is a function $Q : S \times \mathcal{S} \to [0, 1]$ such that the following two criteria are satisfied.

(i) For each $\mathbf{y} \in S, Q(\mathbf{y}, \cdot)$ is a probability measure on $(S, \mathcal{S})$.

(ii) For each $D \in \mathcal{S}, Q(\cdot, D)$ is a measurable function.

We define a sequence of S-valued random variables $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$ to be a Markov chain on $(S, \mathcal{S})$ with transition kernel Q if it satisfies

$$\mathsf{P}(\mathbf{Y}_{n+1} \in D | \mathbf{Y}_0, \ldots, \mathbf{Y}_n) = Q(\mathbf{Y}_n, D). \tag{5.1}$$

Equation (5.1) states that the probability of $\mathbf{Y}_{n+1}$ taking a value in D depends on $(\mathbf{Y}_0, \ldots, \mathbf{Y}_n)$ only through $\mathbf{Y}_n$.

The transition kernel $Q(\mathbf{Y}_n, D)$ of a Markov chain gives the probability of moving from a particular point $\mathbf{y} \in S$ to a set of points $D \in \mathcal{S}$ in one step. The m-step transition kernel is similarly defined to be the probability of moving from a particular point in S to a set of points in $\mathcal{S}$ in exactly m steps: $Q^m(\mathbf{Y}_n, D) = \mathsf{P}(\mathbf{Y}_{n+m} \in D | \mathbf{Y}_0, \ldots, \mathbf{Y}_n)$. In this notation, the transition kernel of the Markov chain is its 1-step transition kernel.

To be able to express probabilities concerning the whole sequence $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$, it is convenient to define a probability measure on the sequence space $(S^{\{0,1\ldots\}}, \mathcal{S}^{\{0,1\ldots\}})$. To specify probabilities of different outcomes of $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$ one needs, in addition to the transition kernel Q, an initial distribution.

For our application it is convenient to consider the case where the initial distribution is a point mass in some point $\mathbf{y_0}$. Denote by $P_{\mathbf{y_0}}$ the probability measure on the sequence space $(S^{\{0,1\ldots\}}, \mathcal{S}^{\{0,1\ldots\}})$ given by setting $\mathbf{Y}_0 = \mathbf{y_0}$ and setting the value $\mathbf{y}_i$ of $\mathbf{Y}_i$ according to $Q(\mathbf{y}_{i-1}, \cdot)$, for $i = 1, 2, \ldots$. The point $\mathbf{y_0}$ is called the **starting point** of the Markov chain.

A distribution Ψ on $(S, \mathcal{S})$ is called **stationary** for the transition kernel Q (or the corresponding Markov chain) if

$$\Psi(D) = \int \Psi(d\mathbf{y}) Q(\mathbf{y}, D).$$

Under certain conditions on the transition kernel Q of a Markov chain $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$, there exists a unique stationary distribution and $Q^n(\cdot, \mathbf{y_0})$, i.e. the distribution of $\mathbf{Y}_n$ conditioned on $\mathbf{Y}_0 = \mathbf{y_0}$, converges to this

stationary distribution, as n tends to infinity, for all starting points $\mathbf{y_0}$. We state some relevant properties of the transition kernel Q for this convergence to take place. The following definitions are from Roberts and Rosenthal [19].

For $D \in \mathcal{S}$, let $\tau_D = \inf\{n \geq 1 : \mathbf{Y}_n \in D\}$ be the first return time to D with $\tau_D = \infty$ if the chain never returns. For Markov chains on general state spaces irreducibility is defined with respect to a σ-finite measure. A Markov chain is **ϕ-irreducible** if there exists a non-zero σ-finite measure ϕ on $(S, \mathcal{S})$ such that $P_{\mathbf{y}}(\tau_D < \infty) > 0$ for all $\mathbf{y} \in S$ and $D \in \mathcal{S}$ with $\phi(D) > 0$.

The period of a ϕ-irreducible Markov chain with stationary distribution Ψ is the largest $n \in \mathbb{N} = \{1, 2, \ldots\}$ for which there exist disjoint subsets $S_1, S_2, \ldots, S_n \in \mathcal{S}$ with $\Psi(S_i) > 0$, such that $Q(\mathbf{y}, S_{i+1}) = 1$ for all $\mathbf{y} \in S_i \, (1 \leq i \leq n-1)$ and $Q(\mathbf{y}, S_1) = 1$ for all $\mathbf{y} \in S_n$. If $n = 1$, then the Markov chain is said to be **aperiodic**.

If a Markov chain on a general state space with Ψ as stationary distribution is aperiodic and ϕ-irreducible, then Ψ is the unique stationary distribution for the Markov chain [18]. However, the distribution $Q^n(\mathbf{y_0}, \cdot)$ of an aperiodic and ϕ-irreducible Markov chain $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$ on a general state space is not guaranteed to converge to Ψ for all starting points $\mathbf{y_0}$. The convergence is guaranteed only for a set of starting points with Ψ-measure 1. Instead, the property of Harris recurrence in combination with aperiodicity assures convergence from all starting points for a Markov chain on a general state space. For a general discussion of Harris recurrence see Roberts and Rosenthal [19].

Definition 5.1. A ϕ-irreducible Markov chain on a general state space $(S, \mathcal{S})$ with stationary distribution Ψ is Harris recurrent if for all $D \in \mathcal{S}$ with $\Psi(D) > 0$ and all $\mathbf{y_0} \in S$, it holds that $P_{\mathbf{y_0}}(\tau_D < \infty) = 1$.

In order to state the convergence results for aperiodic and Harris recurrent Markov chains on general state spaces, we introduce the total variation distance.

Definition 5.2. The total variation distance between two distributions Ψ_1 and Ψ_2 on $(S, \mathcal{S})$ is denoted by $||\Psi_1, \Psi_2||$ and given by

$$||\Psi_1, \Psi_2|| = \sup\{|\Psi_1(A) - \Psi_2(A)| : A \in \mathcal{S}\}.$$

Theorems 5.3 and 5.4 below are fundamental for our applications of Markov chains. See e.g. Tierney [21], (Theorem 1 and 3).

Theorem 5.3. *Consider a Markov chain on general state space $(S, \mathcal{S})$ with transition kernel Q and stationary distribution Ψ. If the Markov chain is aperiodic and Harris recurrent, then Ψ is the unique stationary distribution for the Markov chain and for all $\mathbf{y} \in S$ and $D \in \mathcal{S}$,*

$$\lim_{n \to \infty} ||Q^n(\mathbf{y}, D), \Psi(D)|| = 0.$$

We explain the usefulness of Theorem 5.3 for constructing Markov chain Monte Carlo (MCMC) algorithms. Assume we want to assign a value to a random variable $\mathbf{X}$ according to a distribution Ψ and that this cannot be done directly. Assume further that we can construct an aperiodic and Harris recurrent Markov chain, $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$, with Ψ as its stationary distribution. If we set $\mathbf{X} = \mathbf{Y}_n$ for a sufficiently large n, then by Theorem 5.3, $\mathbf{X}$ is distributed according to a distribution close to Ψ. This is the idea behind Markov chain Monte Carlo which we discuss further in the next section.

Theorem 5.4 below states a similar result as Theorem 5.3 above. Assume one wants to integrate an integrable function h with respect to a complex distribution Ψ. Assume further that one can construct an aperiodic and Harris recurrent Markov chain with Ψ as stationary distribution. Then averaging the function h over the states visited by the chain during a sufficiently long run gives an approximation of the integral which is sufficiently close to the value of the integral.

Theorem 5.4. *Consider an aperiodic and Harris recurrent Markov chain $(\mathbf{Y}_0, \mathbf{Y}_1, \ldots)$ on a general state space $(S, \mathcal{S})$ with stationary distribution Ψ. If $E_\Psi|h| = \int_S |h(\mathbf{y})|\Psi(d\mathbf{y}) < \infty$, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} h(\mathbf{Y}_i) = E_\Psi h \quad a.s.$$

for any initial distribution.

The following theorem (a rewritten form of Theorem 6 of Roberts and Rosenthal [19]) gives an easily checked criterion for Harris recurrence of a Markov chain.

Theorem 5.5. *Consider a ϕ-irreducible and aperiodic Markov chain on $(S, \mathcal{S})$ with stationary distribution Ψ. If for all $\mathbf{y} \in S$ and all $D \in \mathcal{S}$ with $\Psi(D) = 0$, $P_\mathbf{y}(\mathbf{Y}_n \in D \text{ for all } n) = 0$, then the Markov chain is Harris recurrent.*

5.1 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a sampling method based on a Markov chain. It is useful for sampling from high-dimensional distributions which are too complex to allow direct sampling. To sample from a target distribution Ψ one constructs an aperiodic and Harris recurrent Markov chain $(\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Y}_2, \ldots)$ with Ψ as its stationary distribution. By Theorem 5.3, the distribution of $\mathbf{Y}_n$ converges to Ψ, as n tends to infinity, irrespectively of the choice of starting point.

In the following sections we present two versions of MCMC algorithms. In Section 5.2 we present the single-site Gibbs sampler, which

is a simple variant and not suited for the target distributions related to our problem. However, it is used as a building block for a more elaborate MCMC algorithm that we construct, and which is called a Block Updating MCMC. This MCMC-algorithm is presented in Section 5.3.

We need to introduce some notation. Throughout the paper we use Ψ to represent the target distribution of an MCMC. We consider only the case when Ψ has a density denoted by ψ and referred to as the target density. We use S to denote the space on which the target distribution Ψ is defined and which, unless otherwise stated, is an r-dimensional product space.

For $\mathbf{Y} = (Y_1, \ldots, Y_r)$, a vector valued random variable taking values in an r-dimensional space S and $B \subset \{1, \ldots, r\}$, we let $\mathbf{Y}_B = \{Y_i : i \in B\}$ and S_B be the range of $\mathbf{Y}_B$. Moreover we denote $\mathbf{Y}_{-B} = \{Y_i : i \notin B\}$ and, for $i \in B$, $\mathbf{Y}_{-i} = \mathbf{Y}_{-\{i\}}$. Also, if Ψ is the distribution of the random variable $\mathbf{Y}$, then we let $\Psi(\mathbf{Y}_B|\mathbf{Y}_{-B})$ represent the conditional distribution of $\mathbf{Y}_B$, conditioned on $\mathbf{Y}_{-B}$. Likewise, if ψ is the density of $\mathbf{Y}$, then we let $\psi(\mathbf{Y}_B|\mathbf{Y}_{-B})$ represent the conditional density of $\mathbf{Y}_B$, conditioned on $\mathbf{Y}_{-B}$.

A transition kernel of an MCMC algorithm is typically built up by a series of sub-kernels. We call a transition by one of these sub-kernels a basic transition. In contrast, a transition according to the full transition kernel is referred to as the full transition.

5.2 The Gibbs sampler

The Gibbs sampler introduced by Geman and Geman [7] is one type of MCMC. An elementary version of this MCMC is the single-site Gibbs sampler. One full transition of a single-site Gibbs sampler consists of repeating a basic transition a number of times.

A basic transition from the current state, $\mathbf{y}$, to the next state, $\mathbf{y}'$, of a single-site Gibbs sampler on an r-dimensional space consists of two steps. In the first step an index $i \in \{1, \ldots, r\}$ is chosen in some way, then in the second step the new state $\mathbf{y}' = y_i' \cup \mathbf{y}_{-i}$ is obtained by drawing y_i' from $\Psi(\cdot|\mathbf{y}_{-i})$. Thus, a single-site Gibbs sampler evolves by changing one coordinate of $\mathbf{y}$ at a time.

If the index $i \in \{1, \ldots, r\}$ in the first step of this algorithm is chosen uniformly, then the algorithm is called a random-scan single-site Gibbs sampler. It is possible to prove by elementary techniques, that Ψ is a stationary distribution for the random-scan single-site Gibbs sampler. However, in light of Theorem 5.3 and 5.4, to be of practical use a Markov chain on a general state space must be Harris recurrent.

In Section 7 we give an example which demonstrates that the single-site Gibbs sampler is not ϕ-irreducible for all possible distributions related to our problem and thus not Harris recurrent. The lack of irreducibility of the single-site Gibbs sampler stems from the fact that only one element

is updated at a time. If $S' \subset S$ denotes the support of the target density ψ, then two elements $\mathbf{y}$ and $\mathbf{x}$ in S', may be such that it is impossible to move from $\mathbf{y}$ to $\mathbf{x}$ by changing only one element at a time without leaving S'. A single-site Gibbs sampler which visits only states in S' and updates one element at a time is obviously not ϕ-irreducible in this case.

The Block Gibbs sampler is an alternative to the single-site Gibbs. While a single-site Gibbs sampler updates one coordinate y_i of $\mathbf{y} = (y_1, \ldots, y_r)$ at a time, a Block Gibbs sampler updates more general subsets, $\mathbf{y}_B$ where $|B| \geq 1$, at a time.

A basic transition from the current state, $\mathbf{y}$, to the next state, $\mathbf{y}'$, of a Block Gibbs sampler consists of two steps. First a set $B \subseteq \{1, \ldots, r\}$ is is some way, then the new state $\mathbf{y}' = \mathbf{y}'_B \cup \mathbf{y}_{-B}$ is obtained by drawing $\mathbf{y}'_B$ from $\Psi(\cdot|\mathbf{y}_{-B})$. One disadvantage with the Block Gibbs sampler is that direct sampling from the possibly high-dimensional distribution $\Psi(\cdot|\mathbf{y}_{-B})$ may be very time consuming.

The Block Updating MCMC that we construct has the updates of the more general subsets $\mathbf{y}_B$ in common with the Block Gibbs Sampler, but circumvents direct sampling from the high-dimensional distributions. This Block Updating MCMC is inspired by Hurn [11].

5.3 The Block Updating MCMC

Like the Block Gibbs sampler, the Block Updating MCMC updates general subsets, $\mathbf{y}_B$, where $|B| \geq 1$, of coordinates of $\mathbf{y} = (y_1, \ldots, y_r)$ at a time. In contrast to the Block Gibbs sampler, the Block Updating MCMC avoids possible time consuming direct sampling from the conditional distributions $\Psi(\cdot|\mathbf{y}_{-B})$.

In the first step of a basic transition of a Block Updating MCMC, a set $B \in \{1, \ldots, r\}$ is chosen. In the second step, instead of generating a new state by direct sampling from $\Psi(\cdot|\mathbf{y}_{-B})$, the Block Updating MCMC employs a secondary Markov chain on S_B to generate a new state. Although this new state is not generated according to the conditional distribution $\Psi(\cdot|\mathbf{y}_{-B})$, we show in Theorem 5.7 below, that the Block Updating MCMC, in fact, has Ψ as its stationary distribution.

If the target density ψ has support on $S' \subset S$, then a state $\mathbf{y} \in S$ is said to be legal if $\mathbf{y} \in S'$ and illegal if $\mathbf{y} \in S \setminus S'$. Recall that the reason for not using the simpler single-site Gibbs sampler was the lack of irreducibility of that algorithm steaming from the updating of one element at a time. Two legal states, $\mathbf{y}, \mathbf{x} \in S'$, may be such that it is impossible to move from $\mathbf{y}$ to $\mathbf{x}$ by changing only one element at a time without leaving S'.

The key idea of the Block Updating MCMC is that the secondary Markov chain is constructed so that if $\mathbf{y}^\star_B$ denotes a state of S_B visited by the secondary chain, then $\mathbf{y}^\star_B \cup \mathbf{y}_{-B}$ can be illegal as well as legal. Of course, the output $\mathbf{y}'_B$ from the secondary chain is such that the generated

next state $\mathbf{y}'_B \cup \mathbf{y}_{-B}$ for the main chain is legal. This is achieved by defining a relaxed density $\psi^\star$ with support on the whole set S and let the secondary Markov chain update according to $\psi^\star(\cdot|\mathbf{y}_{-B})$. We show in Section 7 that this ensures irreducibility and Harris recurrence of the Block Updating MCMC for our application.

We use a random-scan single-site Gibbs sampler to implement the secondary Markov chain of the Block Updating MCMC.

Algorithm 5.6. ***The Block Updating MCMC for the (unnormalized) target density*** ψ

S	-	*an r-dimensional product space.*
ψ	-	*an (unnormalized) target density with support $S' \subset S$.*
$\psi^\star$	-	*a relaxed density with support on the whole set S satisfying $\psi(\mathbf{y}) \propto \psi^\star(\mathbf{y}) I_{S'}$.*
$\mathbb{B}$	-	*a set of subsets of components of S.*

One full transition of the Block Updating MCMC consists of a series of basic transitions. If the state of the chain before the i^{th} basic transition is $\mathbf{y}$, then a new state $\mathbf{y}'$ is generated by the i^{th} basic transition in two steps:

I. *A subset B of coordinates of $\mathbf{y}$ is drawn from $\mathbb{B}$ according to a distribution Δ_i (which is allowed to vary with i).*

II. *The new state $\mathbf{y}'$ is drawn from $Q_B(\mathbf{y}, \cdot)$, where Q_B is defined below in terms of the output from a secondary chain on S_B.*

Definition of Q_B. *Holding the elements $\mathbf{y}_{-B}$ fixed, a random-scan single-site Gibbs sampler on S_B, started in $\mathbf{y}_B$ and updated according to $\psi^\star(\mathbf{y}_B|\mathbf{y}_{-B})$ is initiated. If $\mathbf{y}_B, \mathbf{y}_B^{\star 1}, \mathbf{y}_B^{\star 2}, \ldots$ denotes the successive states of this chain, then the chain is terminated at the smallest k such that $\mathbf{y}_B^{\star k} \cup \mathbf{y}_{-B}$ is legal. If this $\mathbf{y}_B^{\star k}$ is denoted by $\mathbf{y}'_B$, then the new state $\mathbf{y}'$ of the Block Updating MCMC is $\mathbf{y}' = \mathbf{y}'_B \cup \mathbf{y}_{-B}$.*

We use the convention that $\mathbb{B}$ contains only subset of components of S with positive probability of being updated in some basic transition, i.e, B is in $\mathbb{B}$ if and only if B is in the support of Δ_i for some i.

The following theorem guarantees that the Block Updating MCMC in fact has ψ as stationary density.

Theorem 5.7. *Consider a Block Updating MCMC for the (unnormalized) density ψ. If Ψ denotes the distribution corresponding to ψ, then Ψ is a stationary distribution for the Block Updating MCMC.*

Proof. A distribution Ψ is said to be a reversible distribution for a transition kernel Q if

$$\Psi(d\mathbf{y})Q(\mathbf{y}, d\mathbf{x}) = \Psi(d\mathbf{x})Q(\mathbf{x}, d\mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y} \in S. \tag{5.2}$$

If Ψ is a reversible distribution for a Markov chain with transition kernel Q, then Ψ is also a stationary distribution for the Markov chain. This follows directly from the definition of reversibility by writing $\Psi(d\mathbf{x}) = \Psi(d\mathbf{x}) \int_S Q(\mathbf{x}, d\mathbf{y}) = \int_S \Psi(d\mathbf{x}) Q(\mathbf{x}, d\mathbf{y})$ and using Equation (5.2) in the last expression to get $\Psi(d\mathbf{x}) = \int_S \Psi(d\mathbf{y}) Q(\mathbf{y}, d\mathbf{x})$.

If Ψ denotes the distribution corresponding to ψ, then it follows as a special case of Equation (4) of Hurn et. al. [11] that the transition kernel corresponding to one basic transition of the Block Updating MCMC satisfies Equation (5.2). Thus Ψ is a stationary distribution for each basic transition and we conclude that Ψ is a stationary distribution for the full transition of the Block Updating MCMC. □

Of course, the idea to update according to a relaxed density could have been applied without restricting updates to elements of one subset B at a time. One could have constructed a single-site Gibbs sampler that updates according to the relaxed density and sampled it only at times it stayed in legal states. However, such sampler would suffer from serious time-inefficiency.

6 The Expectation Maximization (EM) algorithm

We now present the Expectation Maximization (EM) algorithm. A stochastic version of this algorithm, discussed in Section 6.1, is used to compute the Maximum likelihood estimate of θ from data from a percolation process. The EM algorithm introduced by Dempster et. al. [5] searches the maximum of the likelihood function in an iterative fashion. It is suited for problems that can be formulated in terms of unobserved data.

Let $\mathbf{x}$ represent a vector of observed data and $\mathbf{u}$ a vector of unobserved data. The observed and unobserved data will later coincide with the data $\mathbf{x}^n$ from the percolation process and the configuration $\mathbf{u}$, as defined in Section 2, but for the moment, $\mathbf{x}$ and $\mathbf{u}$ represent any observed and unobserved data. The vector $(\mathbf{x}, \mathbf{u})$ is referred to as the complete data. The vectors $\mathbf{x}$ and $\mathbf{u}$ are realisations of the random vectors $\mathbf{X}$ and $\mathbf{U}$.

For each θ in some parameter space Θ, let $f_{\mathbf{X},\mathbf{U}|\theta}(\cdot|\theta)$, $f_{\mathbf{X}|\theta}(\cdot|\theta)$ and $f_{\mathbf{U}|\mathbf{X},\theta}(\cdot|\mathbf{x},\theta)$ denote the density of the complete data, the density of the observed data and the conditional density of the unobserved data conditioned on the observed data, all with respect to some measure which we for each of these densities denote by ν.

The aim is to find the maximum likelihood estimate (MLE), $\hat{\theta}$, the maximizer of the likelihood function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta, \mathbf{x}),$$

where $L(\theta, \mathbf{x}) = f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$ is the density of the observed data.

The EM algorithm finds the MLE, $\hat{\theta}$, by an iterative procedure. Let in the t^{th} iteration of the EM algorithm the current value of θ be $\theta^{(t-1)}$. Each iteration of the EM algorithm consists of two steps, the E-step and the M-step. In the E-step an expectation of the complete data log-likelihood, $\log f_{\mathbf{X},\mathbf{U}|\theta}(\mathbf{x}, \mathbf{u}|\theta)$, is computed. The expectation is computed with respect to the density of the unobserved data, conditioned on the observed data and the current value $\theta^{(t-1)}$, i.e. under the density $f_{\mathbf{U}|\mathbf{X},\theta}(\mathbf{u}|\mathbf{x}, \theta^{(t-1)})$. This expectation, known as the Q-function, is thus given by

$$Q(\theta|\theta^{(t-1)}) = \mathrm{E}_{\theta^{(t-1)}}[\log f_{\mathbf{X},\mathbf{U}|\theta}|\mathbf{x}],$$

where

$$\mathrm{E}_{\theta^{(t-1)}}[\log f_{\mathbf{X},\mathbf{U}|\theta}|\mathbf{x}] = \int \log f_{\mathbf{X},\mathbf{U}|\theta}(\mathbf{x}, \mathbf{u}|\theta) f_{\mathbf{U}|\mathbf{X},\theta}(\mathbf{u}|\mathbf{x}, \theta^{(t-1)}) \nu(d\mathbf{u}).$$

Note the fundamental different roles played by θ and $\theta^{(t-1)}$ in the definition of the Q-function. In the M-step, the new value of θ, $\theta^{(t)}$, is set to the value that maximizes $Q(\theta|\theta^{(t-1)})$.

Theorem 6.1 below states that in each iteration of the EM algorithm the likelihood function is increased. Thus with a bounded likelihood function on an one-dimensional parameter space, convergence to a local maximum is guaranteed. The following property is fundamental for the EM-algorithm. For details see for example Lange [14].

Theorem 6.1. *The EM iterates satisfy*

$$L(\theta^{(t)}, \mathbf{x}) \geq L(\theta^{(t-1)}, \mathbf{x})$$

with strict inequality when

$$Q(\theta^{(t)}|\theta^{(t-1)}) > Q(\theta^{(t-1)}|\theta^{(t-1)}).$$

6.1 The Monte Carlo Expectation Maximization (MCEM) algorithm

The Monte Carlo Expectation Maximization (MCEM) algorithm introduced by Wei and Tanner [23] is an extension of the basic EM algorithm to situations where the Q-function is hard to compute. It is a stochastic version of the EM algorithm which instead of computing the Q-function analytically uses an MCMC method to approximate the Q-function.

Let $\tilde{\theta}^{(t-1)}$ denote the value of θ in the t^{th} iteration of the MCEM algorithm. Given a sample $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ from an MCMC algorithm with stationary distribution corresponding to the density $f_{\mathbf{U}|\mathbf{X},\theta}(\mathbf{u}|\mathbf{x}, \tilde{\theta}^{(t-1)})$, the Q-function is approximated by a Monte Carlo integration:

$$\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)}) = \frac{1}{n} \sum_{i=1}^{n} \log f_{\mathbf{X},\mathbf{U}}(\mathbf{x}, \mathbf{u}_i|\theta).$$

The theoretical base for the MCEM algorithm is Theorem 5.4. It states that if the MCMC algorithm that generates the sample $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is aperiodic and Harris recurrent, then, for a fixed θ, the approximation $\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)})$ converges almost surely to the value of the true Q-function, $Q(\theta|\tilde{\theta}^{(t-1)})$, as n tends to infinity.

Recall that n denotes the size of the sample used to approximate the Q-function and let $\tilde{\theta}^{(t,n)}$ denote the maximizer of the approximation $\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)})$ of the Q-function. Before $\tilde{\theta}^{(t,n)}$ is accepted, we want to convince ourselves that changing the value of θ from $\tilde{\theta}^{(t-1)}$ to $\tilde{\theta}^{(t,n)}$ will, with high probability, increase the true Q-function. We introduce some notation to answer this question.

Let $\Delta Q(\theta')$ be the increment in the Q-function when changing the value of θ from $\tilde{\theta}^{(t-1)}$ to some θ':

$$\Delta Q(\theta') = Q(\theta'|\tilde{\theta}^{(t-1)}) - Q(\tilde{\theta}^{(t-1)}|\tilde{\theta}^{(t-1)}).$$

This increment is estimated by the Monte Carlo integration analogue

$$\Delta\tilde{Q}_n(\theta') = \tilde{Q}_n(\theta'|\tilde{\theta}^{(t-1)}) - \tilde{Q}_n(\tilde{\theta}^{(t-1)}|\tilde{\theta}^{(t-1)}). \tag{6.1}$$

A large value of $\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)})$ indicates that changing the value of θ from $\tilde{\theta}^{(t-1)}$ to $\tilde{\theta}^{(t,n)}$ increases the true Q-function. However, to judge whether or not the suggested new value should be accepted, one also need to estimate the variance of $\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)})$. How to estimate this variance is discussed in Section 10.4.

If sufficiently strong evidence of an increase in the true Q-function is achieved, then $\tilde{\theta}^{(t,n)}$ is accepted and $\tilde{\theta}^{(t)}$ is set to $\tilde{\theta}^{(t,n)}$ and the algorithm continues to the next iteration. If $\tilde{\theta}^{(t,n)}$ is not accepted, then the sample size n is increased and the procedure is repeated. When producing the larger sample, the old sample is re-used and further realisations are appended to the old sample. This implementation of the MCEM algorithm is influenced by the MCEM algorithm of Caffo et. al. [3].

The MCMC algorithm we use to produce the sample for the Monte Carlo integration is the Block Updating MCMC introduced in Section 5.3. In the next section we show that the Block Updating MCMC for this purpose as well as the one used in the Bayesian inference are Harris recurrent under some conditions.

7 The Block Updating MCMC and data from a percolation process

In the last two sections we have presented necessary background on MCMC and the Monte Carlo EM algorithms. In this section we consider sampling from target distributions related to the percolation process. We show that while a single-site Gibbs sampler is not sufficient for

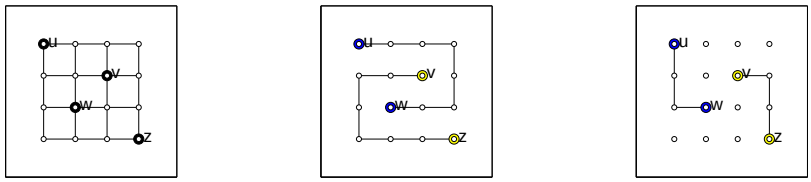

Figure 4: *Left:* A small graph with 4 observation points, u, v, w, z. Assume the observed data from this graph is that (u, v) and (w, z) are the only connected pairs. *Middle and right:* Two configurations agreeing with the data. There is no way to change the status of one single edge in the first configuration without entering an illegal state, thus a simple site Gibbs sampler visiting only legal states with respect to the data is not ϕ-irreducible.

sampling from these distributions, it is possible to construct Block Updating MCMC's which converges in the appropriate way to the desired target distributions.

The reason why the single-site Gibbs sampler is not sufficient for our problem is that it is not possible to create a ϕ-irreducible single-site Gibbs sampler for all possible data from a percolation process on a graph. Figure 4 gives an example of a tiny graph with four data points and data such that a single-site Gibbs sampler which visits only legal states with respect to the observed data is not ϕ-irreducible.

We fix some notation for distributions related to the percolation process on a finite graph. As before, given a graph with m edges, $\mathbb{U} = \{0, 1\}^m$ denotes the set of all configurations and p_θ denotes the probability mass function over $\mathbb{U}$ under the parameter value $\theta \in \Theta$. Given observed values $\mathbf{x}$ of the data vector $\mathbf{X}$ from the percolation process, we use $\mathbb{U}' = \{\mathbf{u} \in \mathbb{U} : \mathbf{X}(\mathbf{u}) = \mathbf{x}\}$ to denote the set of configurations agreeing with the data. The elements of $\mathbb{U}'$ are called the legal configurations with respect to $\mathbf{x}$.

For the Bayesian inference we define $\mathbb{W} = \mathbb{U} \times \Theta$ and $\mathbb{W}' = \mathbb{U}' \times \Theta$ and let $\mathbf{w} = (\mathbf{u}, \theta)$ represent a typical element of $\mathbb{W}$. We also need to specify a σ-algebra on $\mathbb{W}$. With $\mathcal{U}'$ denoting the set of all subsets of $\mathbb{U}'$, we define $\mathcal{W}' = \mathcal{U}' \times \Theta$. We assume the prior Π on $(\Theta, \mathcal{B}(\Theta))$ is absolutely continuous with respect to Lebesgue measure λ, and let π denote the density of Π with respect to λ. A natural prior to have in mind is the uniform prior, for which $\pi(\theta) \equiv 1$ for $\theta \in \Theta$.

7.1 Harris recurrence of the Block Updating MCMC used in the Bayesian inference

In the Bayesian setting, given data $\mathbf{x}$ from the percolation process on a finite graph, we want to compute the posterior distribution. With $\mathbb{U}'$ denoting the set of legal configurations with respect to $\mathbf{x}$, the density of the observed data can be expressed in terms of the density of the configurations $\mathbf{u}$ as

$$p_{\mathbf{X}|\theta}(\mathbf{x}) = \sum_{\mathbf{u}\in\mathbb{U}} p_\theta(\mathbf{u}) I_{\{\mathbf{u}\in\mathbb{U}'\}},$$

where we use I_E to denote the indicator function of an event E. Thus with π denoting the density of the prior Π with respect to Lebesgue measure λ, the posterior distribution $\Pi(\cdot|\mathbf{x})$ may be written as, for $A \in \mathcal{B}(\Theta)$,

$$\Pi(A|\mathbf{x}) \propto \int_A p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)\lambda(d\theta) = \int_A \sum_{\mathbf{u}\in\mathbb{U}} p_\theta(\mathbf{u}) I_{\{\mathbf{u}\in\mathbb{U}'\}} \pi(\theta)\lambda(d\theta). \tag{7.1}$$

Equation (7.1) states that sampling from the unnormalized target density $\psi(\mathbf{u},\theta) = p_\theta(\mathbf{u}) I_{\{\mathbf{u}\in\mathbb{U}'\}}\pi(\theta)$ and integrating over $\mathbb{U}$ with respect to Lebesgue measure λ produces a sample from the posterior distribution $\Pi(\cdot|\mathbf{x})$.

We use a Block Updating MCMC to sample from the unnormalized target density $\psi(\mathbf{u},\theta) = p_\theta(\mathbf{u}) I_{\{\mathbf{u}\in\mathbb{U}'\}}\pi(\theta)$. Theorem 7.1 below ensures that the Block Updating MCMC can be constructed to in fact converge in the appropriate way.

Theorem 7.1. *Consider a Block Updating MCMC for the unnormalized density $\psi(\mathbf{u},\theta) = p_\theta(\mathbf{u}) I_{\{\mathbf{u}\in\mathbb{U}'\}}\pi(\theta)$. Assume the Block Updating MCMC satisfies the following two conditions.*

1. *In each full transition every coordinate of $\mathbf{w} = (\mathbf{u},\theta)$ has positive probability of being updated.*
2. *All subsets of coordinates of $\mathbf{u}$ corresponding to a connected subgraph of G have positive probability of being chosen in step $\mathbf{I}$ of one basic transition of Algorithm 5.6.*

Then the Block Updating MCMC is aperiodic and Harris recurrent and if Q denotes the transition kernel of the MCMC, then for all $\mathbf{w} \in \mathbb{W}'$ and $D \in \mathcal{W}'$

$$\lim_{n\to\infty} ||Q^n(\mathbf{w}, D), \Psi(D)|| = 0.$$

Proof. The product space of this Block Updating MCMC is $\mathbb{W} = \mathbb{U} \times \Theta$. However, the actual state space of a Block Updating MCMC is the support of the target density, in this case $\mathbb{W}' = \mathbb{U}' \times \Theta$. Recall the choice

of σ-algebra on $\mathbb{W}'$, $\mathcal{W}' = \mathcal{U}' \times \Theta$, where $\mathcal{U}'$ denotes the set of all subsets of $\mathbb{U}'$

Now, let Q denote the transition kernel of a full transition of the Block Updating MCMC, $(\mathbf{W}_0, \mathbf{W}_1, \ldots)$, which satisfies the conditions of the theorem. By Theorem 5.7, ψ is a stationary density for this Block Updating MCMC. If we also prove that the MCMC is aperiodic and Harris recurrent, then the assertion of the theorem follows from Theorem 5.3

We first show that the MCMC is aperiodic and ϕ-irreducible. Then the Harris recurrence follows easily from Theorem 5.5. Both the aperiodicity and the ϕ-irreducibility are evident if we show that

$$\forall \text{ large } k: \; Q^k(\mathbf{w}, D) > 0 \;\; \forall \mathbf{w} \in \mathbb{W}' \text{ and } D \in \mathcal{W}' \text{ with } \Psi(D) > 0. \quad (7.2)$$

Instead of proving Equation (7.2) directly we prove that an alternative statement is true, which is sufficient for Equation (7.2) to hold. Firstly, note that each $D \in \mathcal{W}'$ with $\Psi(D) > 0$ contains a subset of the form $\mathbf{u} \times A$, where $\mathbf{u}$ is a legal configuration and A has Lebesgue measure $\lambda(A)$ greater than zero.

Secondly, note that after the first update of the θ-coordinate of $\mathbf{w} = (\theta, \mathbf{u})$, all following states visited by the Block Updating MCMC have θ-coordinate in the interior, $\text{int}(\Theta)$, of Θ (w.p.1). The proof of this statement does not enhance the understanding of the proof of the theorem and we therefore postpone it for the end.

Therefore, to show Equation (7.2) it suffices to show

$$\begin{aligned} \forall \text{ large } k: \quad Q^k((\mathbf{u}, \theta), \tilde{\mathbf{u}} \times A) > 0 \quad & \forall \mathbf{u}, \tilde{\mathbf{u}} \in \mathbb{U}', \\ & \forall \theta \in \text{int}(\Theta) \text{ and} \\ & \forall A \in \mathcal{B}(\Theta) \text{ s.t. } \lambda(A) > 0. \end{aligned} \quad (7.3)$$

To prove Equation (7.3) we introduce two new concepts, those of a critical element of a configuration and a critical configuration. An element u_j of a legal configuration $\mathbf{u} = (u_1, \ldots, u_m)$ with respect to the observed data $\mathbf{x}(\mathbf{u}) = (x_1(\mathbf{u}), \ldots, x_d(\mathbf{u}))$, is said to be **a critical element** for the observed data point $x_i(\mathbf{u}) = 1$, if $u_j = 1$ and $x_i(u_1, \ldots, u_{j-1}, 0, u_{l+1}, \ldots, u_m) = 0$. The meaning of a critical element is the following. Holding all other components of the configuration fixed and changing the critical element from 1 to 0 destroys an open path so that the new configuration does no longer satisfy the data.

A legal configuration $\mathbf{u} \in \mathbb{U}'$ with respect to observed data $\mathbf{x}$ is said to be **a critical configuration** if it satisfies that

(i) for each coordinate x_i of $\mathbf{x}$ with $x_i = 1$, there is a critical element in $\mathbf{u}$, and

(ii) all subsets of coordinates of $\mathbf{u}$ corresponding to some connected subgraph of G with no two observation points connected by an open path, are zero.

The first part, (i), states that a critical configuration has, for each pair of connected observation points, an edge such that if it is removed then the observation points are no longer connected by an open path.

The proof of Equation (7.3) consists of three parts. The idea is that from any legal configuration $\mathbf{u}$, it is possible to reach any other legal configuration $\tilde{\mathbf{u}}$, by first go to a critical configuration $\mathbf{u}^{\text{crit}}$ in finitely many transitions (part a), then in one transition go from $\mathbf{u}^{\text{crit}}$ to an other critical configuration $\tilde{\mathbf{u}}^{\text{crit}}$ (part c) from which it is possible to reach $\tilde{\mathbf{u}}$ in finitely many transitions (part b):

(a) For any $\mathbf{u} \in \mathbb{U}'$, there is a critical configuration $\mathbf{u}^{\text{crit}} \in \mathbb{U}'$, such that for any $\theta \in \text{int}(\Theta)$ and all large n:

$$Q^n((\mathbf{u}, \theta), \{\mathbf{u}^{\text{crit}}\} \times A) > 0, \text{ for all } A \in \mathcal{B}(\Theta) \text{ with } \lambda(A) > 0. \quad (7.4)$$

(b) For any $\tilde{\mathbf{u}} \in \mathbb{U}'$, there is a critical configuration $\tilde{\mathbf{u}}^{\text{crit}} \in \mathbb{U}'$, such that for any $\theta \in \text{int}(\Theta)$ and all large n:

$$Q^n((\tilde{\mathbf{u}}^{\text{crit}}, \theta), \{\tilde{\mathbf{u}}\} \times A) > 0, \text{ for all } A \in \mathcal{B}(\Theta) \text{ with } \lambda(A) > 0. \quad (7.5)$$

(c) For any two critical configuration $\mathbf{u}^{\text{crit}}$ and $\tilde{\mathbf{u}}^{\text{crit}}$ and for any $\theta \in \text{int}(\Theta)$:

$$Q((\mathbf{u}^{\text{crit}}, \theta), \{\tilde{\mathbf{u}}^{\text{crit}}\} \times A) > 0, \text{ for all } A \in \mathcal{B}(\Theta) \text{ with } \lambda(A) > 0. \quad (7.6)$$

We first show (a). For any legal configuration it is possible to reach a critical configuration by changing the value of q elements from 1 to 0. Formally, for any legal configuration $\mathbf{u}$ there is a critical configuration $\mathbf{u}^{\text{crit}}$ and a sequence of legal configurations $(\mathbf{u}^0, \mathbf{u}^1, \ldots, \mathbf{u}^q)$, with $\mathbf{u}^0 = \mathbf{u}$ and $\mathbf{u}^q = \mathbf{u}^{\text{crit}}$, such that for each $i \in \{0, 1, \ldots, q-1\}$ there is an $l' \in \{1, \ldots, m\}$ such that

$$\mathbf{u}^i_{l'} = 1, \, \mathbf{u}^{i+1}_{l'} = 0 \qquad \mathbf{u}^i_l = \mathbf{u}^{i+1}_l \text{ for } l \in \{1, \ldots, m\} \setminus \{l'\}.$$

At any basic transition of the Block Updating MCMC, the $\mathbf{u}$-coordinate can be unaltered and we thus have that for all $\theta \in \text{int}(\Theta)$,

$$\begin{aligned} P_{(\mathbf{u}^0, \theta)}(&\mathbf{W}_1 \in \{\mathbf{u}^1\} \times \text{int}(\Theta), \ldots, \\ &\mathbf{W}_{q-1} \in \{\mathbf{u}^{q-1}\} \times \text{int}(\Theta), \\ &\mathbf{W}_q \in \{\mathbf{u}^q\} \times A) > 0, \end{aligned} \quad (7.7)$$

which implies that for all $\theta \in \text{int}(\Theta)$,

$$Q^q((\mathbf{u}, \theta), \{\mathbf{u}^{\text{crit}}\} \times A) > 0 \text{ for all } A \in \mathcal{B}(\Theta) \text{ with } \lambda(A) > 0. \qquad (7.8)$$

Since q is obviously finite and the $\mathbf{u}$-coordinate can be unaltered by a full transition, Equation (7.4) follows from Equation (7.8). The proof of (b) is totally analogous.

We now prove (c). Assume the graph G has k connected subgraphs. Recall that each element of $\mathbf{u} \in \mathbb{U}$ corresponds to one particular edge of G. Let $\mathbb{U}_i$, for each $i = 1, \ldots, k$, denote the subspace of $\mathbb{U}$ where each element corresponds to an edge in the i^{th} connected subgraph of G. The elements of the two critical configurations $\mathbf{u}^{\text{crit}}$ and $\tilde{\mathbf{u}}^{\text{crit}}$ are partitioned in the obvious way

$$\mathbf{u}^{\text{crit}} = (\mathbf{u}_1^{\text{crit}}, \ldots, \mathbf{u}_k^{\text{crit}}) \quad \text{and} \quad \tilde{\mathbf{u}}^{\text{crit}} = (\tilde{\mathbf{u}}_1^{\text{crit}}, \ldots, \tilde{\mathbf{u}}_k^{\text{crit}}). \qquad (7.9)$$

Let $\mathbf{U}_i$ be a variable on $\mathbb{U}_i$ which denotes the state of the chain a some point. We show that if at some point $\mathbf{U}_i = \mathbf{u}_i^{\text{crit}}$, then within one full transition of the Block Updating MCMC, $\mathbf{U}_i = \tilde{\mathbf{u}}_i^{\text{crit}}$ happens with positive probability.

By definition, if the i^{th} connected subgraph of G has no two connected observation points then $\mathbf{u}^{\text{crit}}$ and $\tilde{\mathbf{u}}^{\text{crit}}$ are identical - all elements of both $\mathbf{u}^{\text{crit}}$ and $\tilde{\mathbf{u}}^{\text{crit}}$ are zero - and there is nothing to show.

Consider now the case when the i^{th} connected subgraph of G has at least two connected observation points. There is by assumption 2 of the theorem a positive probability that all components of $\mathbb{U}_i$ are chosen in step **I** of the algorithm. Given that all the components of $\mathbb{U}_i$ are chosen in step **I**, the secondary Markov chain in step **II**, can, with positive probability, evolve as follows:

In the first step of the secondary MCMC an element of $\mathbf{U}_i$ corresponding to a critical element of $\mathbf{u}_i^{\text{crit}}$ is set to zero. In the following steps all elements of $\mathbf{U}_i$ are set to zero. Then all elements of $\mathbf{U}_i$ which are 1 in the configuration $\tilde{\mathbf{u}}_i^{\text{crit}}$ are set to 1. The last element set to 1 is one corresponding to a critical element of $\tilde{\mathbf{u}}_i^{\text{crit}}$. This secondary chain does indeed not visit any legal configuration before it reaches $\tilde{\mathbf{u}}_i^{\text{crit}}$. We have shown (c). The statements of (a), (b) and (c) imply that the Block updating MCMC is ϕ-irreducible and aperiodic.

We use Theorem 5.5 to show that the ϕ-irreducible and aperiodic Block Updating MCMC on $(W', \mathcal{W}')$ is also Harris recurrent. We need to prove that for any $\mathbf{w} \in \mathbb{W}'$ and $D \in \mathcal{W}'$ with $\Psi(D) = 0$, we have $P_{\mathbf{w}}(\mathbf{W}_n \in D \text{ for all } n) = 0$.

If $D \in \mathcal{W}'$ is such that $\Psi(D) = 0$, then $D \subset \mathbb{U}' \times \xi$, where ξ has Lebesgue measure zero. By assumption 1 of the theorem, the θ-coordinate of $\mathbf{w} = (\theta, \mathbf{u})$ is updated within finitely many steps of the Block Updating MCMC. Since when θ is updated it leaves the null-set ξ the aperiodic

Block Updating MCMC is also Harris recurrent. The statement of the Theorem therefore follows from Theorem 5.3.

Left to prove is only the unproven statement in the beginning of the proof. It stated that after the first update of the θ-coordinate of $\mathbf{w} = (\theta, \mathbf{u})$, all following states visited by the Block Updating MCMC have θ-coordinate in the interior $\text{int}(\Theta)$ of Θ (w.p.1).

To validate this statement, recall that the secondary Markov chain of the Block Updating MCMC is a single-site Gibbs sampler which updates according to the conditional densities of the relaxed density $\psi^\star$ with support on the whole space $\mathbb{W} = \mathbb{U} \times \Theta$. When the θ-coordinate is updated by the single-site Gibbs sampler, a new θ-value is thus generated from the conditional density $\psi^\star(\cdot|\mathbf{u})$. Since $\psi^\star(\mathbf{u}, \theta)$ has support $\mathbb{W} = \mathbb{U} \times \Theta$, it follows that $\psi^\star(\theta|\mathbf{u})$ has support Θ for all $\mathbf{u} \in \mathbb{U}$.

We conclude that even if the Block Updating MCMC is started in a starting point with the value of the θ-coordinate on the boundary of Θ, after the first update, the θ-value will leave the boundary and never come back (w.p.1). □

7.2 Harris recurrence of the Block Updating MCMC used in the frequentist inference

In the frequentist setting, given data $\mathbf{x}$ from a percolation process, we want to compute the MLE, $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta, \mathbf{x})$, where $L(\theta, \mathbf{x}) = p_{\mathbf{X}|\theta}(\mathbf{x})$. We construct an MCEM algorithm to compute the MLE. We view the configuration $\mathbf{u}$ as missing data. The observed data are $\mathbf{x}$.

Within the t^{th} MCEM iteration we need to approximate the Q-function. Recall that the Q-function is the expectation of the complete data log-likelihood with respect to the density of the unobserved data conditioned on the observed data and the current value $\tilde{\theta}^{(t-1)}$. The complete data log-likelihood is in our case $\log p_\theta(\mathbf{u})$ and the density of the unobserved data $\mathbf{u}$ conditioned on the observed data $\mathbf{x}$, is proportional to $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u}) I_{\mathbb{U}'}$.

Given a sample $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ from $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u}) I_{\mathbb{U}'}$ we approximate the Q-function by

$$\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)}) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(\mathbf{u}_i).$$

We use a Block Updating MCMC to generate the sample $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$. Theorem 7.2 below ensures that it is possible to construct this Block Updating MCMC such that the approximation $\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)})$ converges to the true expectation, $Q(\theta|\tilde{\theta}^{(t-1)})$, irrespectively of the starting point of the Block Updating MCMC.

The Block Updating MCMC in Theorem 7.2 lives on a finite sample space $\mathbb{U}$ and we could have stated the result in Theorem 7.2 in terms of classical irreducibility for Markov chains on finite state spaces. However,

to keep the formulation of Theorems 7.1 and 7.2 uniform, we use the more general concept of Harris recurrence.

Theorem 7.2. *Consider a Block updating MCMC,* $(\mathbf{U}_0, \mathbf{U}_1, \ldots)$*, for the unnormalized target density* $\psi(\mathbf{u}) = p_\theta(\mathbf{u}) I_{\mathbb{U}'}$ *(for some* $\theta \in \Theta$*). Let* Ψ *be the distribution corresponding to* ψ*. Assume the Block Updating MCMC satisfies the following two conditions.*

1. *In each full transition every coordinate of* $\mathbf{u}$ *has positive probability of being updated.*

2. *All subsets of coordinates of* $\mathbf{u}$ *corresponding to a connected subgraph of* G *have positive probability of being chosen in step* $\mathbf{I}$ *of one basic transition of Algorithm 5.6.*

Then the Markov chain is aperiodic and Harris recurrent, and thus for any starting value $\mathbf{u}_0 \in \mathbb{U}'$ *and for any* $\theta \in \Theta$

$$\lim_{n\to\infty} \tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)}) = Q(\theta|\tilde{\theta}^{(t-1)}) \quad a.s.$$

Proof. We need to show that the Markov chain is Harris recurrent. Then, since for all $\theta \in \Theta$, $\log p_\theta$ in Equation 7.2 satisfies the integrability condition in Theorem 5.4, the convergence in 7.2 follows.

The Block Updating MCMC lives on the finite space $(\mathbb{U}', \mathcal{U}')$ where $\mathbb{U}'$ denotes the set of all legal configurations with respect to the data, and $\mathcal{U}'$ is the set of all subsets of $\mathbb{U}'$. We assume θ is not 0 and nor 1. If θ is 0 or 1 then there is only one configuration with positive probability and there is nothing to prove. If Q denotes a transition kernel on $(\mathbb{U}', \mathcal{U}')$ which satisfies the conditions of the theorem, then, in analogy with the proof of Theorem 7.1, to prove aperiodicity and irreducibility, we need to show that

$$\forall \text{ large } k: \; Q^k(\mathbf{u}, D) > 0 \;\; \forall \mathbf{u} \in \mathbb{U}' \text{ and } D \subset \mathbb{U}' \neq \emptyset. \tag{7.10}$$

The proof of Equation (7.10) consists of three parts.

(a') For any $\mathbf{u} \in \mathbb{U}'$, there is a critical configuration $\mathbf{u}^{\text{crit}}$ in $\mathbb{U}'$, such that for all large n:

$$Q^n(\mathbf{u}, \{\mathbf{u}^{\text{crit}}\}) > 0. \tag{7.11}$$

(b') For any $\tilde{\mathbf{u}} \in \mathbb{U}'$, there is a critical configuration $\tilde{\mathbf{u}}^{\text{crit}}$ in $\mathbb{U}'$, such that for all large n:

$$Q^n(\tilde{\mathbf{u}}^{\text{crit}}, \{\tilde{\mathbf{u}}\}) > 0. \tag{7.12}$$

(c') For any two critical configuration $\mathbf{u}^{\text{crit}}$ and $\tilde{\mathbf{u}}^{\text{crit}}$:

$$Q(\mathbf{u}^{\text{crit}}, \{\tilde{\mathbf{u}}^{\text{crit}}\}) > 0. \tag{7.13}$$

The proof of a', b' and c' are analogous to the proof of a, b and c in the proof of Theorem 7.1.

Classical irreducibility corresponds to ϕ-irreducibility with respect to counting measure. The conditions in Theorem 5.5, for an ϕ-irreducible and aperiodic Markov chain to be Harris recurrent are trivially satisfied for the Markov chain on the finite space $\mathbb{U}'$. □

8 A simulation study

We performed a simulation study to evaluate the inference procedures presented in Section 2. We start by describing the graphs on which the percolation processes lives in this simulation study and explain the choice of parameter values θ. Finally we comment on the use of burn-in and thinning of a sample from an MCMC algorithm.

In Section 9 we specify the Block Updating MCMC used in the Bayesian inference and in Section 10 the MCEM algorithm used in the frequentist approach. In Section 11 we present the results from the simulation study.

8.1 The graph used in the simulation study

To illustrate the consistency results of Theorems 3.2 and 3.4, we performed inference for a percolation process on a graph of the type introduced in Section 2, i.e. a graph which is the union of a number of graphs isomorphic to a base graph. We use L to denote the base graph and consequently L^n denotes the graph consisting of the n graphs, $L_1, \ldots, L_n$, that are isomorphic to L.

The choice of base graph L is made with respect to computational load. We decided to let L be a 60×60 subset of the square lattice. The vertex set $V(L)$ and edge set $E(L)$ of L are given by:

$$V(L) = \{1, \ldots, 60\}^2, \qquad E(L) = \{\langle u, v\rangle : u, v \in V(L), |u - v| = 1\}.$$

To illustrate the convergence results, we have performed the inference on L^n for an increasing value of n.

Recall that the subgraph L_k of L^n is referred to as the k^{th} primary subgraph of L^n. Although not required in Theorems 3.2 and 3.4, we use the same set of observation points in each primary subgraph L_k, which consists of 13 vertices. We denote this set by $\mathcal{O}^{(13)}$, see Figure 5. The choice of observation points is arbitrary.

We are also interested in the convergence of sequences of posterior distributions and MLE's based on data from more general graphs than those described above. Next, we turn to a type of graph L^n_{conn} which, in contrary to the graph L^n, is a connected graph and therefore generates data without the independence structure of the data from L^n.

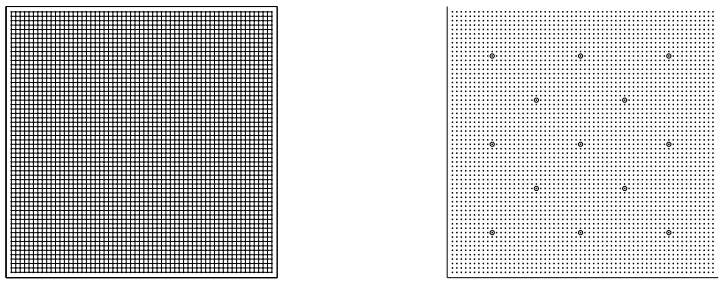

Figure 5: *Left:* The graph L which is a 60×60 subset of the square lattice. *Right:* The set of observation points $\mathcal{O}^{(13)}$.

Given L^n, L^n_{conn} is constructed by connecting the n primary subgraphs of L^n to obtain a single connected graph. Illustrations of the graphs L^n and L^n_{conn} for $n = 3$ are given in Figures 6 and 7.

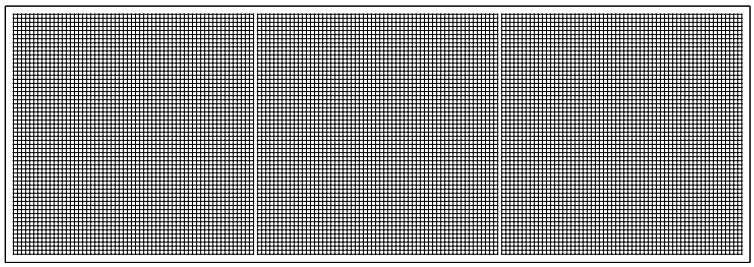

Figure 6: The graph L^3

8.2 The values of θ used in the simulation study

The existence of the so-called phase transitions is a principal result in percolation theory. For percolation processes on many different regular infinite lattices, there is a critical value θ_c of the parameter θ such that if $\theta < \theta_c$, then the probability that there exists an infinite open cluster is 0, and if $\theta > \theta_c$, then the probability that there exists an infinite open cluster is 1 [10]. This abrupt change in the systems behaviour resulting from a small change from a θ less than θ_c to a θ larger than θ_c is called a phase transition. The value θ_c at which the phase transition occurs is called the critical value. For a percolation process on the infinite square lattice the critical value is $\theta_c = 0.5$, see [10].

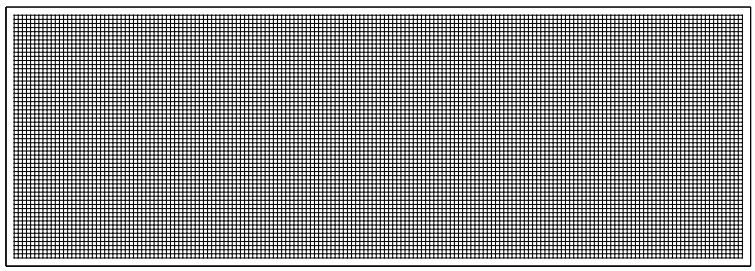

Figure 7: L^3_{conn}

The phase transition phenomenon for a percolation process on the infinite square lattice has implications also for the percolation process on the finite square lattice L. If the percolation process is generated with a θ that is much larger than $\theta_c = 0.5$, then it is very likely that any two vertices in L are connected by an open path. Likewise, if the percolation process is generated with a θ considerably smaller than $\theta_c = 0.5$, then it is highly unlikely that any two well-separated vertices in L are connected by an open path. Consequently, the most interesting values of θ are those in a interval around $\theta_c = 0.5$ for which some pairs are connected and other pairs are not. We have chosen to consider the parameter values $\theta = 0.47, 0.50$ and 0.53 in the simulation study.

8.3 Using a sample from an MCMC algorithm: burn-in and thinning

Given the output from an MCMC algorithm it is practice to disregard a number of initial iterations, which is called the **burn-in**. After the burn-in the Markov chain is assumed to be in stationarity.

An MCMC in stationarity provides a sample of correlated realisations from the desired target distribution. This correlation can be measured by means of the autocorrelation function. For a stationary process $(Y_0, Y_1, \ldots)$ the k**-lag autocorrelation** is given by

$$R(k) = \frac{E[(Y_t - \mu)(Y_{t+k} - \mu)]}{\sigma^2},$$

where $\mu = E[Y_t]$ and $\sigma^2 = Var(Y_t)$.

The autocorrelation can be reduced by thinning the Markov chain, which is done by retaining only every n^{th} iterate. The value n is called the **thinning interval**.

9 Specification of the Block Updating MCMC for the Bayesian inference

Assume data $\mathbf{x}^n$ has been observed from the percolation process on L^n. The goal is to compute the posterior distribution $\Pi(\cdot|\mathbf{x}^n)$ after choosing the prior distribution Π. Here we have chosen as prior the uniform with density $\pi(\theta) \equiv 1$, $\theta \in [0,1]$, with respect to Lebesgue measure λ.

With this choice of prior, we have from Section 7.1 that, if $\mathbb{U}'$ denotes the set of legal configurations with respect to the observed data $\mathbf{x}^n$, then integrating a sample from the unnormalized density $\psi(\mathbf{u},\theta) = p_\theta(\mathbf{u})I_{\{\mathbf{u}\in\mathbb{U}'\}}$ over $\mathbb{U}$ produces a sample from the posterior distribution $\Pi(\cdot|\mathbf{x}^n)$.

By Theorem 7.1, it is possible to construct a Block Updating MCMC on $\mathbb{W} = \mathbb{U} \times \Theta$ such that the distribution of the n^{th} iterate of the chain converges to $\psi(\mathbf{u},\theta) = p_\theta(\mathbf{u})I_{\{\mathbf{u}\in\mathbb{U}'\}}$ as n tends to infinity, for whatever choice of starting point.

Before we proceed any further we need to choose the values for the different parameters entering the algorithm. If the total number of edges in L^n is m, then the Block Updating MCMC lives on $\mathbb{W} = \mathbb{U} \times \Theta$, where $\mathbb{U} = \{0,1\}^m$. A typical element of $\mathbb{W}$ is $\mathbf{w} = (\mathbf{u},\theta)$, where $\mathbf{u}$ is a possible realisation of the percolation process on L^n.

Recall that the edge set of L^n is $E(L^n) = \cup_{k=1}^n E(L_k)$, where $E(L_k)$ is the edge set of the k^{th} primary subgraph of L^n. In order to specify the Block Updating MCMC we will occasionally write $\mathbf{u} = (\mathbf{u}_1,\ldots,\mathbf{u}_n)$ and let $\mathbf{u}_k$ contain the elements of $\mathbf{u}$ which corresponds to edges in the k^{th} primary subgraph L_k. Since each element of $\mathbf{u}$ corresponds to an edge we sometimes refer to these elements as edges.

9.1 Different basic transitions within a full transition

In the first basic transition of a full transition the θ-value of $\mathbf{w} = (\mathbf{u},\theta)$ is updated while in the rest of the basic transitions only the coordinates of $\mathbf{u} = (\mathbf{u}_1,\ldots,\mathbf{u}_k)$ are updated. For all k, $k = 1,\ldots,n$, R subsets of the $\mathbf{u}_k$'s are updated in each full transition. The value of R is chosen to achieve an efficient balance between updates of edges and updates of the parameter θ. We used different values of R in preliminary runs and evaluated the computational load and auto-correlation and found $R = |E(L_k)|$ to be suitable.

9.2 The set of subsets $\mathbb{B}$

Recall that $\mathbb{B}$ denotes, in the case of a Block Updating MCMC on $\mathbb{W} = \mathbb{U}\times\Theta$, the set of subsets of components of $\mathbb{W}$ that have positive probability of being updated simultaneously in one basic transition. From Section

9.1, it follows that the one-element set $\{\theta\}$ is included in $\mathbb{B}$ and that any other element of $\mathbb{B}$ is a subset of $E(L_k)$ for some $k \in \{1, \ldots, n\}$, i.e. a subset of edges of the k^{th} primary subgraph. We let $\mathbb{B}_k$ denote that subset of $\mathbb{B}$ that contains the subsets of $E(L_k)$ and is defined in the following way.

Let us denote by (i, j), for $i, j \in \{1, \ldots, 60\}$ an element of the vertex set $V(L_k)$. We define an ***s*-sized quadratic subset** V' of $V(L_k)$ to be a set of the form

$$V' = \{(i, j) : a \leq i \leq \min(a + s, 60), b \leq j \leq \min(b + s, 60)\} \tag{9.1}$$

for some $a, b \in \{1, \ldots, 60\}$ and an ***s*-sized quadratic subset** E' of $E(L_k)$ to be a set

$$E' = \{\langle u, v\rangle \in E(L_k) : u, v \in V'\} \tag{9.2}$$

for some s-sized quadratic subset V' of $V(L_k)$.

Now, let $\mathbb{B}_k^s$ be the set of s-sized quadratic subsets of $E(L_k)$ and for some appropriate minimal size $s_{\min}$, we define $\mathbb{B}_k = \cup_{s=s_{\min}}^{\infty} \mathbb{B}_k^s$.

There seems to be no point in including too small quadratic subsets of $E(L_k)$ in $\mathbb{B}_k$. On the other hand, since as seen in Example 4 in Section 7, a quadratic subsets of $E(L_k)$ of size only 4 can resolve some irreducibility problems, we chosen to take $s_{\min} = 4$.

9.3 The distribution Δ_i over $\mathbb{B}$

We now specify the distribution Δ_i over $\mathbb{B}$. Within the i^{th} basic transition of the Block Updating MCMC the subset of components of $\mathbb{W}$ to be updated simultaneously is chosen according to Δ_i.

The distribution Δ_i varies with the basic transition number i within a full transition. From Theorem 7.1 follows immediately, that for Harris recurrence, $\Delta_i(\mathbf{u}_k) > 0$ for some i, is sufficient.

Recall that $\mathbb{B}_k \subset \mathbb{B}$ denotes those subsets of edges of the k^{th} primary subgraph which have positive probability of being updated simultaneously. In each basic transition (except the first), from $\mathbb{B}_k$, $k = 1, \ldots, n$, a subset of edges is updated. Now, assume that in the i^{th} basic transition, we choose to update an element of $\mathbb{B}_k$. Then, we define Δ_i, by the following. First a size s is drawn from a geometrical distribution truncated at $s_{\min} = 4$. Then, given s, an element is chosen uniformly among the elements in $\mathbb{B}_k^s$, i.e. from the set containing all s-sized quadratic subgraphs of the k^{th} primary subgraph.

The truncated geometrical distribution is chosen in order to obtain a good balance between the updates of small and large subsets of edges. Updating many large subsets ensures that the chain can easily move between different regions of the state space. On the other hand, updating too many large subsets may slow down the algorithm. However, the choice of truncated geometrical distribution did not seem to influence

the performance of the algorithm very much. In this study, we have used the geometrical distribution with mean 4 truncated at $s_{\min} = 4$.

9.4 The relaxed density $\psi^\star$

By Theorem 7.1 we can choose the relaxed density $\psi^\star$ of the Block Updating MCMC freely as long as it is proportional to the unnormalized target density $\psi(\mathbf{u}, \theta) = p_\theta(\mathbf{u}) I_{\{\mathbf{u} \in \mathbb{U}'\}}$ on the set of legal states, $\mathbb{W}' = \mathbb{U}' \times \Theta$. The obvious choice of $\psi^\star = p_\theta(\mathbf{u})$ is far from optimal. Hence, we follow Hurn [11] in the choice of relaxed density $\psi^\star$.

We start by noting that there is a natural hierarchy among the illegal states. The more data points violated by an illegal state, the 'more illegal' it is. To prevent the Markov chain to move too far away from the legal states it is natural to choose a relaxed density $\psi^\star$ which penalizes the more illegal states.

For this, we define a function $C : \mathbb{U} \to [0, \infty)$ which takes the value 0 for a legal $\mathbf{u}$, and is increasing with respect to the 'degree of illegality'. Recall that the observed data $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,d_k})$ is a realisation of the data vector $\mathbf{X}^n$. Then let $C(\mathbf{u}) = \sum_{k=1}^{n} \sum_{l=1}^{d_k} I_{\{X_{k,l}(\mathbf{u}) \neq x_{k,l}\}}$ denote the number of data points not satisfied by the configuration $\mathbf{u}$. Finally, we choose $\psi^\star$ by

$$\psi^\star(\mathbf{u}, \theta) \propto p_\theta(\mathbf{u}) \exp\{-aC(\mathbf{u})\},$$

for some $a > 0$. The value of a determines how easily the secondary Markov chain of the Block Updating MCMC can move away from the legal states. We set $a = 0.8$.

9.5 Starting point for the Block Updating MCMC

We use a single-site Gibbs sampler started in $\theta = 0.5$ and a randomly generated (not necessarily legal) configuration to find a legal starting point for the Block Updating MCMC.

10 Specification of the Monte Carlo EM algorithm for the frequentist inference

Next, we specify the MCEM algorithm that we use in the frequentist inference. We need to specify several parameters in the algorithm. Moreover, we need to decide on appropriate starting values for the algorithm, stopping criteria and an estimation procedure for the Monte Carlo error of the approximation $\tilde{Q}_n$.

10.1 Starting value for the MCEM algorithm

The choice of starting value for θ in the MCEM algorithm is crucial. This value of θ is used to generate a starting point, $\mathbf{u}$, for the Block Updating MCMC in the first iteration of the EM algorithm. (In any other EM iterate the last configuration of the Block Updating MCMC in the previous EM iterate is used as starting point.)

As discussed in Section 8.2, the percolation process on the infinite square lattice exhibits a phase transition at $\theta_c = 0.5$. This implies that also on the finite square lattice L, the process generated by a θ somewhat smaller than $\theta_c = 0.5$ is qualitatively different from the process generated by a θ somewhat larger than $\theta_c = 0.5$. Thus, it is important to prevent the situation where the starting value and the true value of θ are on opposite sides of $\theta_c = 0.5$. Therefore, the only natural choice of starting value for the Monte Carlo EM algorithm is $\theta = 0.5$.

10.2 The scheme to increase the sample size in approximating the Q-function

Recall that $\tilde{\theta}^{(t,n)}$ denotes the maximizer of the approximation $\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)})$ of $Q(\theta|\tilde{\theta}^{(t-1)})$ based on a Monte Carlo sample of size n. Within each EM-iteration a sequence of suggested new values $\{\tilde{\theta}^{(t,n)}\}_n$ is computed until eventually a value is accepted. The sample size is increased by 5 at a time, i.e. $n = 5, 10, 15, \ldots$. Also, in generating the sample, we have used burn-in time 10 and thinning interval 1.

10.3 The Block Updating MCMC used within each iteration of the MCEM algorithm.

We turn now to the Block Updating MCMC used within each EM iteration to generate a sample for the Monte Carlo integration. This Block Updating MCMC is similar to the Block Updating MCMC used for the Bayesian inference presented in Section 9, the difference being that the Block Updating MCMC in the latter case lives on $\mathbb{W} = \mathbb{U} \times \Theta$, whereas the Block Updating MCMC in former lives on $\mathbb{U}$.

Recall that only in the first basic transition of the Block Updating MCMC used in the Bayesian inference, the element $\theta \in \Theta$ is updated, while in the rest we only update the coordinates of $\mathbf{u}$ of $\mathbb{U}$. The Block Updating MCMC used in the frequentist inference is defined by letting a full transition of this algorithm be equivalent to all but the first basic transition of a full transition in the Block Updating MCMC defined in Section 9.

The relaxed density $\psi^\star$ of this Block Updating MCMC on $\mathbb{U}$ is the obvious modification of the relaxed density for the Block Updating MCMC on $\mathbb{W} = \mathbb{U} \times \Theta$ obtained by fixing θ at the value of the current EM

iteration, $\tilde{\theta}^{(t-1)}$:

$$\psi^{\star}(\mathbf{u}) \propto p_{\tilde{\theta}^{(t-1)}}(\mathbf{u}) \exp\{-aC(\mathbf{u})\},$$

for some $a > 0$. As for the Block Updating MCMC used in the Bayesian setting, we chose $a = 0.8$.

10.4 The decision to accept or reject a suggested value $\tilde{\theta}^{(t,n)}$

Recall that the Q-function is the expectation of the log-likelihood of the complete data which in our case is $\log p_{\theta}(\mathbf{u})$. The expectation is computed with respect to the unnormalized density $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u})I_{\mathbb{U}'}$:

$$Q(\theta|\theta^{(t-1)}) = \mathrm{E}_{\theta^{(t-1)}}[\log p_{\theta}(\mathbf{U})|\mathbf{x}^n].$$

Recall also that $\tilde{Q}_n$ is a Monte Carlo integration analogue of Q:

$$\tilde{Q}_n(\theta|\tilde{\theta}^{(t-1)}) = \frac{1}{n}\sum_{i=1}^{n} \log p_{\theta}(\mathbf{u}_i), \tag{10.1}$$

where $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is a sample from an MCMC algorithm with stationary distribution corresponding to the unnormalized density $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u})I_{\mathbb{U}'}$.

In Section 6.1, $\Delta Q(\theta')$ was defined as the increment when the value of θ changes from the current value in the t^{th} EM iteration, $\theta^{(t-1)}$, to some new value θ':

$$\Delta Q(\theta') = Q(\theta'|\tilde{\theta}^{(t-1)}) - Q(\tilde{\theta}^{(t-1)}|\tilde{\theta}^{(t-1)}).$$

We also defined the Monte Carlo integration analogue $\Delta\tilde{Q}_n(\theta')$ of $\Delta Q(\theta')$:

$$\Delta\tilde{Q}_n(\theta') = \tilde{Q}_n(\theta'|\tilde{\theta}^{(t-1)}) - \tilde{Q}_n(\tilde{\theta}^{(t-1)}|\tilde{\theta}^{(t-1)}),$$

which for the suggested new value $\tilde{\theta}^{(t,n)}$, by Equation 10.1, can be written as

$$\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)}) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log \frac{p_{\tilde{\theta}^{(t,n)}}(\mathbf{u}_i)}{p_{\tilde{\theta}^{(t-1)}}(\mathbf{u}_i)}$$

where $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is a sample from an MCMC algorithm with stationary distribution corresponding to the unnormalized density $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u})I_{\mathbb{U}'}$.

With $\mathbf{U}$ distributed according to $p_{\tilde{\theta}^{(t-1)}}(\mathbf{u})I_{\mathbb{U}'}$, we define,

$$W(\mathbf{U}) = \log \frac{p_{\tilde{\theta}^{(t,n)}}(\mathbf{U})}{p_{\tilde{\theta}^{(t-1)}}(\mathbf{U})},$$

Hence, $\Delta Q(\tilde{\theta}^{(t,n)})$ is the expectation of W and by letting $W_i = W(\mathbf{U}_i)$, $\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)})$ is simply the sample average: $\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)}) = \frac{1}{n}\sum_{i=1}^{n} W_i$.

In order to judge whether to accept or reject $\tilde{\theta}^{(t,n)}$ we need to estimate the variance of $\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)})$. Since $W_1, W_2, \ldots$ are correlated, we use a (non-overlapping) batch means estimate. This is a way to reducing the correlation by grouping the observations.

Let a denote the number of batches and assume the sample size n is a multiple of an integer b. Then the k^{th} batch mean

$$\bar{W}_k^b = \frac{1}{b} \sum_{i=(k-1)b+1}^{kb} W_i$$

is the sample average based on the k^{th} batch of b observations from W. Let V denote the variance of W. We assume, somewhat inaccurately, that the batch means, $(\bar{W}_1^b, \ldots, \bar{W}_a^b)$, are independent and normally distributed with variance V/b. Under this assumption

$$\hat{V} = \frac{1}{b(a-1)} \sum_{k=1}^{a} (\bar{W}_k^b - \bar{\bar{W}})^2,$$

where $\bar{\bar{W}}$ is the mean of the batch means, is an unbiased estimator of V and we can form a t-statistic

$$T = \frac{\Delta\tilde{Q}_n(\tilde{\theta}^{(t,n)})}{\sqrt{\hat{V}/n}}.$$

The appropriate degree of freedom of T is $a-1$. (See Sherman [20].) If, for an appropriate choice of α, this t-statistic is in the upper α-quantile, then we accept $\tilde{\theta}^{(t,n)}$ and set $\tilde{\theta}^{(t)} = \tilde{\theta}^{(t,n)}$. We chose $\alpha = 0.05$.

10.5 A stopping rule for the MCEM algorithm

A well-known problem when using a regular non-stochastic EM algorithm is possible slow convergence [14], [22]. Slow convergence can be even more problematic when using a stochastic EM algorithm, such as the MCEM algorithm.

If the Monte Carlo error is large compared to the increase in the Q-function, then the time until a θ-value is accepted can be extremely long. Thus, we may have to stop the MCEM algorithm at values where it is suspected that we do not have yet convergence.

The randomness of a stochastic EM algorithm implies that the changes $|\tilde{\theta}^{(t)} - \tilde{\theta}^{(t-1)}|$ between consecutive iterates of the MCEM algorithm are not a smooth function of the iteration number. Therefore, it is not advisable to base a stopping rule for the MCEM algorithm on them.

Instead, we evaluate a stopping rule repeatedly within each step of the MCEM algorithm. Recall that a sequence of suggested new values

$\{\tilde{\theta}^{(t,n)}\}_{n\in\{5,10,15,\ldots\}}$ based on a Monte Carlo integration from an increasing sample are evaluated within each MCEM iteration. If $\tilde{\theta}^{(t,n)}$ is rejected and $|\tilde{\theta}^{(t,n)} - \tilde{\theta}^{(t-1)}|$ is less than some small number δ for a sequence of K consecutive suggested values, then the MCEM algorithm is terminated. We found $\delta = 0.5 \times 10^{-3}$ to be suitable.

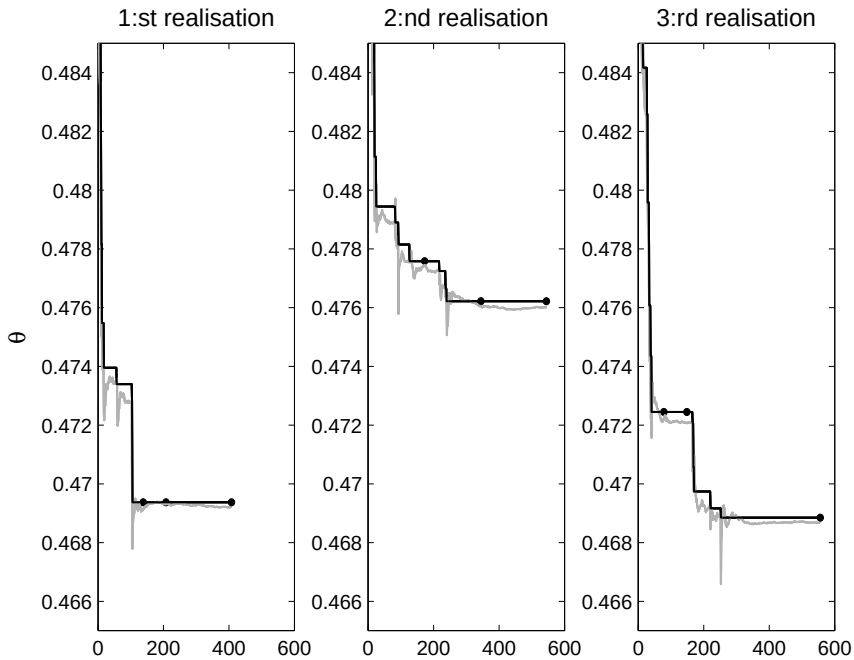

Figure 8: The evolution of $\theta^{(t)}$ in the MCEM algorithm for three different realisations. The dark line represents the value of $\theta^{(t)}$. The gray line represents the suggested new value $\theta^{(t,n)}$. Each jump in the value of $\theta^{(t)}$ corresponds to that a suggested new value $\theta^{(t,n)}$ has been accepted. Thus, each plateau of the value of $\theta^{(t)}$ corresponds to one iteration of the MCEM algorithm. The length of the plateau thus corresponds to n in the accepted value $\theta^{(t,n)}$, i.e. the size of the sample used to approximate the Q-function when the suggested new value is accepted. The three dots in each graph indicates the time of termination for three different rules for termination. All three rules use the same value $\delta = 0.5 \times 10^{-3}$ but different values of K. The three values of K are 30, 100 and 300.

The choice of value of K is crucial. It must be chosen so that the

algorithm terminates in a neighbourhood of the maximum of the likelihood function but at the same time is not too time-consuming. Because of the slow convergence of the MCEM algorithm this is a delicious problem. We illustrate the problem with the slow convergence of the MCEM algorithm for our particular application in Figure 8. The value of $\theta^{(t)}$ in each iteration of the MCEM algorithm for three realisations of the percolation process on L^8 with $\theta = 0.47$ is presented. The time of termination is indicated for three different choices of termination rule, all with $\delta = 0.5 \times 10^{-3}$ but with different values of K.

Obviously there is a trade of between the accuracy in the approximation of the MLE and the computation time. We decided to use $K = 100$ in our simulations.

11 Simulation results

In this section we present results of the implementation of the Block Updating MCMC for the Bayesian inference and the MCEM algorithm for the frequentist inference. The algorithms were implemented in C and executed on a 3 GHz computer.

11.1 Illustration of Theorem 3.2, the consistency result in the Bayesian inference

We first illustrate Theorem 3.2, the consistency result in the Bayesian approach to inference. We generated three different realisations of the percolation process on L^{64} with parameter value $\theta = 0.47$. Recall that the data from L^{64} are denoted $\mathbf{X}^{64}$ and that for $n < 64$, $\mathbf{X}^n$ denotes the data from the first n primary subgraphs of L^{64}.

For each of the three realisations we computed the sequence of posterior distributions corresponding to the data, $\mathbf{x}^2, \mathbf{x}^4, \mathbf{x}^8, \mathbf{x}^{16}, \mathbf{x}^{32}, \mathbf{x}^{64}$. The results are presented in Figures 9 and 10. Figure 9 illustrates how the posterior distribution accumulates close to the true value $\theta = 0.47$ as more data are observed. In Figure 10 we present the means and standard deviations of the posterior distributions.

The decrease in the standard deviations as more data are observed is visible for all three realisations. It is also clear that the means of the posteriors tends to be closer to 0.47 the more data the posteriors are based on. We note though that this trend is not clear when the number of primary subgraphs that the posteriors are based on, is changed from 32 to 64. An explanation for this is that two of the three posteriors based on data from 32 primary subgraphs, happens, by chance, to be very close to 0.47.

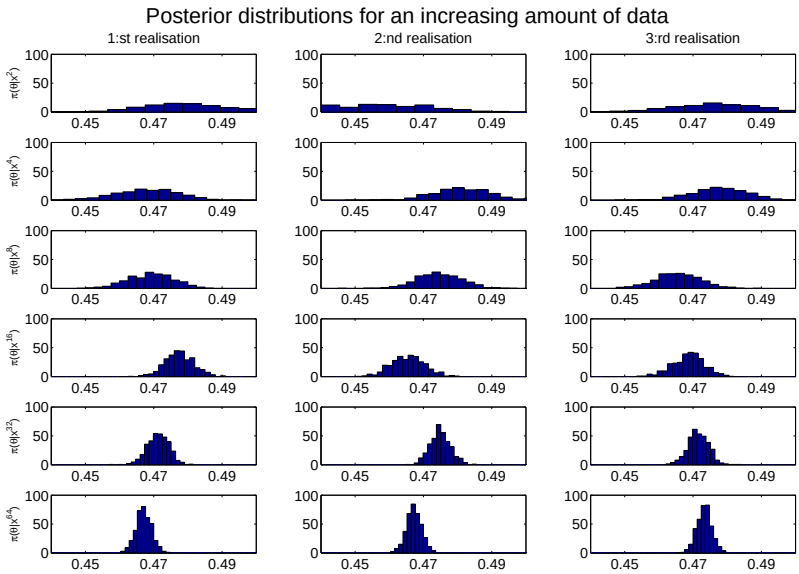

Figure 9: Illustration of the consistency result, Theorem 3.2. Each of the three columns corresponds to one realisation of the percolation process. Each row represents the posterior based on an increasing amount of data: $\pi(\theta|\mathbf{X}^2), \pi(\theta|\mathbf{X}^4), \pi(\theta|\mathbf{X}^8), \pi(\theta|\mathbf{X}^{16}), \pi(\theta|\mathbf{X}^{32}), \pi(\theta|\mathbf{X}^{64})$. The burn-in is 1000, the thinning interval is 5 and the sample size is 4200.

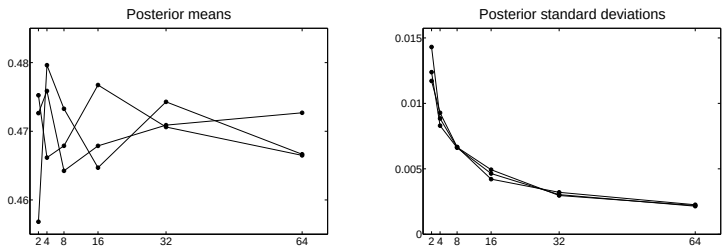

Figure 10: Summary of the posterior distributions in Figure 9. *Left:* The mean for each posterior distribution for an increasing amount of data, for three different realisations. *Right:* The standard deviation for each posterior distribution for an increasing amount of data, for three different realisations.

11.2 Illustration of Theorem 3.4, the consistency result in the frequentist inference

We now illustrate Theorem 3.2, the consistency result in the frequentist approach to inference. We generated 27 different realisations of the percolation process on L^{64} with parameter value $\theta = 0.47$. For each of the 27 realisations we computed the sequence of maximum likelihood estimates corresponding to the data, $\mathbf{x}^2, \mathbf{x}^4, \mathbf{x}^8, \mathbf{x}^{16}, \mathbf{x}^{32}, \mathbf{x}^{64}$. The results are presented in Figure 11 and Table 1.

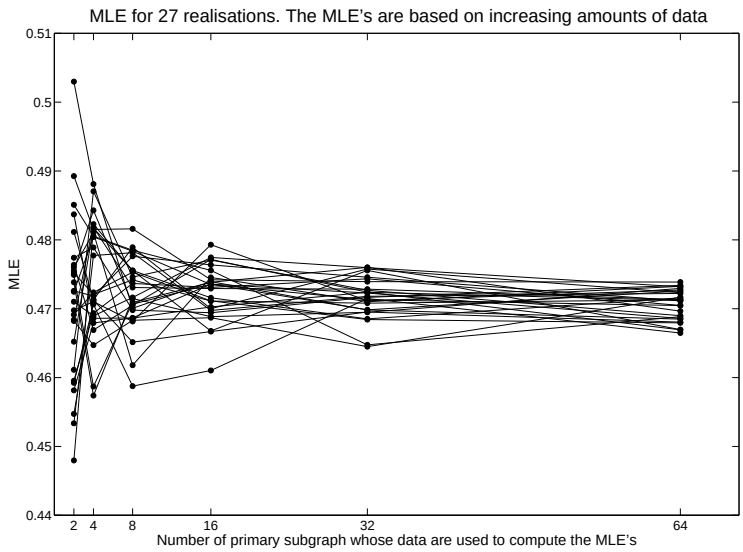

Figure 11: Illustration of the consistency result, Theorem 3.4. The MLE's $\hat{\theta}_2, \hat{\theta}_4, \hat{\theta}_8, \hat{\theta}_{16}, \hat{\theta}_{32}$ and $\hat{\theta}_{64}$ for 27 different realisations. Each line represents a particular realisation.

Estimate h	$\bar{h} = \sum_{i=1}^{27} h_i$	$\sqrt{\frac{1}{26}\sum_{i=1}^{27}(h_i - \bar{h})^2}$
$h = \hat{\theta}_2$	0.4712	0.0119
$h = \hat{\theta}_4$	0.4744	0.0081
$h = \hat{\theta}_8$	0.4722	0.0053
$h = \hat{\theta}_{16}$	0.4723	0.0040
$h = \hat{\theta}_{32}$	0.4716	0.0030
$h = \hat{\theta}_{64}$	0.4706	0.0022

Table 1: Summary of the MLE's in Figure 11.

11.3 Convergence results for more general graph

We now consider a percolation process on the more general graph L^n_{conn} introduced in Section 8.2. For $n = 2, 4, 8$ we generated a realisation of the percolation process L^n_{conn} from a realisation of the process on L^n, where L^n, for $n < 8$, denotes the subgraph of L^8 consisting of the first n primary subgraphs of L^8. The process on L^8 was generated with $\theta = 0.47$. For three different realisations and for each $n = 2, 4, 8$, we computed the posterior distributions and the MLE's based on the data from the percolation process on L^n_{conn}.

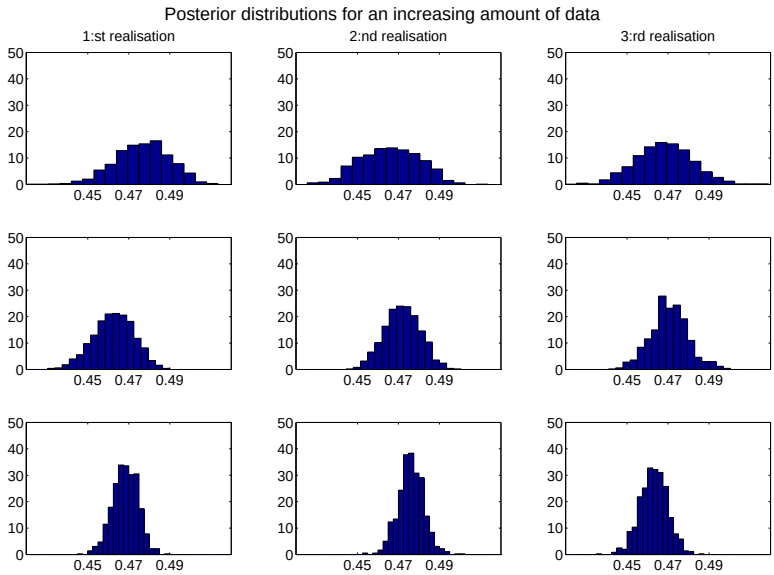

Figure 12: Each column corresponds to the posterior from L^2_{conn}, L^4_{conn} and L^8_{conn} generated from L^2, L^4 and L^8, where L^n for $n < 8$ is the subgraph of L^8containing the n first primary subgraphs of L^8. The burn-in is 1000, the thinning interval is 5 and the sample size is 5000.

The results for the Bayesian inference are presented in Figure 12. A comparison of the posteriors based on data from L^n and L^n_{conn} for $n = 2, 4, 8$ is presented in Table 2. A comparison of the MLE's based on the data from L^n and L^n_{conn} for $n = 2, 4, 8$ is presented in Table 3. For both aproaches the rate of convergence is similar for the data from L^n and L^n_{conn} for this particular case.

Posterior means from realisation 1		
n	L^n	L^n_{conn}
2	0.4752 (0.0117)	0.4740 (0.0135)
4	0.4662 (0.0093)	0.4614 (0.0097)
8	0.4679 (0.0067)	0.4670 (0.0061)
Posterior means from realisation 2		
n	L^n	L^n_{conn}
2	0.4568 (0.0143)	0.4630 (0.0140)
4	0.4796 (0.0088)	0.4700 (0.0086)
8	0.4733 (0.0066)	0.4753 (0.0060)
Posterior means from realisation 3		
n	L^n	L^n_{conn}
2	0.4726 (0.0124)	0.4640 (0.0137)
4	0.4759 (0.0083)	0.4681 (0.0091)
8	0.4642 (0.0066)	0.4618 (0.0068)

Table 2: Comparison of the posterior distributions in Figure 9 (from L^n) and the posterior distributions in Figure 12 (from L^n_{conn}).

	$\bar{h} = \sum_{i=1}^{27} h_i$		$\sqrt{\frac{1}{26}\sum_{i=1}^{27}(h_i - \bar{h})^2}$	
n	L^n	L^n_{conn}	L^n	L^n_{conn}
2	0.4712	0.4721	0.0119	0.0122
4	0.4744	0.4750	0.0081	0.0060
8	0.4722	0.4718	0.0053	0.0045

Table 3: Comparison of the MLE's in Figure 11 (from L^n) and the MLE's from L^n_{conn}

11.4 Performance of the inference procedures depending on the value of θ

We have also carried out simulations to evaluate the dependence of the performance of the inference procedures on the 'true' value of θ. We have considered both the accuracy of the inference procedures themselves and the performance of the algorithms for different values of θ.

For each parameter value $\theta = 0.47, 0.50$ and 0.53 we generated a number of realisations of the percolation process on L^4. The Block Updating MCMC algorithm for the Bayesian inference was run on data from three different realisations for each of the three parameter values. The posterior distributions are presented in Figure 13 and Table 4.

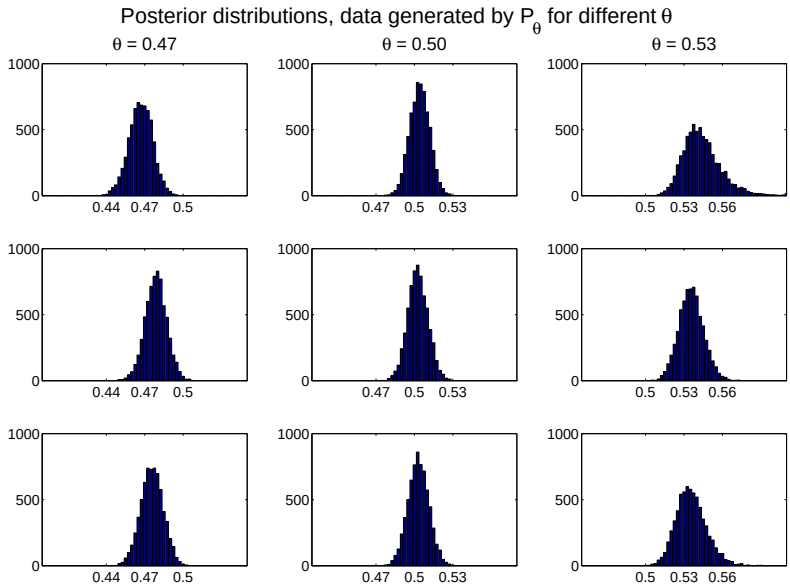

Figure 13: The posterior distribution for three realisations of the percolation process on L^4 for three different values of θ. The burn in is 10000, the thinning interval is 20 and the sample size is 6800. *Column 1:* $\theta = 0.47$, *Column 2:* $\theta = 0.5$, *Column 3:* $\theta = 0.53$.

The MCEM algorithm for the frequentist inference was run on 168 realisations for each of the three parameter values, $\theta = 0.47, 0.50$ and 0.53. The mean of the MLE's, $\hat{\theta}_4$, over the 168 realisations for the different values of θ are presented in Figure 14 and Table 5.

$\theta = 0.47$	$\theta = 0.50$	$\theta = 0.53$
0.4669 (0.0093)	0.5041 (0.0078)	0.5441 (0.0152)
0.4787 (0.0084)	0.5027 (0.0080)	0.5356 (0.0097)
0.4756 (0.0088)	0.5031 (0.0082)	0.5358 (0.0121)

Table 4: The means and standard deviations (in parenthesis) for the 9 posterior distributions in Figure 13.

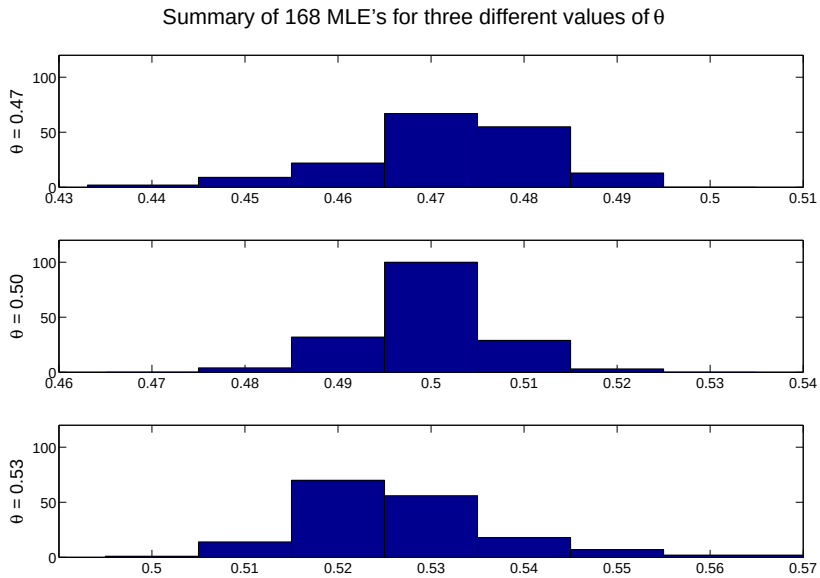

Figure 14: Summary of MLE's for different values of θ.

We note that for both approaches the accuracy seems to be higher for $\theta = 0.5$ than for $\theta = 0.47$ or 0.53 (See Tables 4 and 5), which seems to be natural, since this is the critical value of the percolation process. The derivative with respect to θ of the probability $P_\theta\{o_1 \leftrightarrow o_2\}$ of connectedness of two well-separated vertices o_1 and o_2 on a large proportion of the infinite square lattice is large at values of θ close to the critical probability $\theta_c = 0.5$. Consequently, given the information of connectedness for a set of pairs of vertices, two parameter values θ and $\theta + \epsilon$ (for a small $\epsilon > 0$) are easily distinguishable if θ is close to $\theta_c = 0.5$.

In Figure 14 we also note something that has been visible also in earlier graphs. There seems to be a bias in the approximated MLE towards 0.5. This bias is due to the starting value $\theta = 0.5$ for the MCEM algorithm.

As explained in Section 10.1, this choice was necessary. Due to the slow rate of convergence of the MCEM algorithm (See discussion in Section 10.5) it has to be stopped before it converges. A better implementation of the MCEM can probably decrease this bias.

θ	$\bar{h} = \sum_{i=1}^{168} h_i$	$\sqrt{\frac{1}{167}\sum_{i=1}^{168}(h_i - \bar{h})^2}$
0.47	0.4721	0.0097
0.50	0.4999	0.0070
0.53	0.5264	0.0104

Table 5: Summary of the 168 estimates $\hat{\theta}_4$ in Figure 14. Here h_i denotes the value of $\hat{\theta}_4$ from the i^{th} realisation of the percolation process on L^4.

We also considered the dependence of the auto-correlation and the computational time on θ. The sample auto-correlations and the times to compute the posterior distributions in Figure 13 are given in Table 6. Both the auto-correlation and the computational time seems to be dependent on θ, see Table 6.

$\theta = 0.47$		$\theta = 0.50$		$\theta = 0.53$	
$R(20)$	time	$R(20)$	time	$R(20)$	time
0.23	46	0.12	85	0.63	168
0.17	57	0.12	89	0.28	140
0.19	56	0.15	96	0.47	159

Table 6: The estimated 20-lag auto-correlation, $R(20)$, and the time (in hours) to generate each of the samples for the 9 posterior distributions in Figure 13.

12 Conclusions

We have presented an inference problem for a percolation process on a graph and shown a consistency result for a particular class of graphs in both the Bayesian and the frequentist approach to inference.

Moreover we have developed two algorithms in order to compute the relevant quantities. In the Bayesian inference we developed a block updating MCMC which was shown to converges to the posterior distribution, from any starting point. In the frequentist case, we developed a Monte Carlo EM algorithm. The algorithms were implemented in a simulation study.

The simulation results illustrated the already shown theoretical consistency results for the percolation process on the restricted class of graphs. Moreover, the simulation study indicated that the convergence of both the sequences of posterior distributions and MLE's might extend to more general graphs.

The simulation results also suggest that when we have the option of choosing the inference approach, the Bayesian should be preferred in this special case, since we have seen that the MCEM approach introduced some bias into the estimate due to the critical phenomenon of the percolation process.

References

[1] Berkowitz, B., Balberg, I. Percolation Theory and Its Application to Groundwater Hydrology *Water Resources Research, Vol 29, No 4, 775-794* 1993

[2] Broadbent, S. R., Hammarsley, J. M. Percolation processes I. Crystals and mazes *Proceedings of the Cambridge Philosophical Society 53, 629-641* 1957.

[3] Caffo, B. S., Jank, W., Jones, G. L. Ascent Based Monte Carlo EM *Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 67, 235-TODO*, 2005

[4] Choi, T., Ramamoorthi, R. V. Remarks on consistency of posterior distributions. *arXiv:0805.3248v1* 2008

[5] Dempster, A. P., Laird, N. M., Rubin, D. B. Maximum Likelihood From Incomplete Data via the EM algorithm *Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol 39, 1-38* 1977

[6] Durrett, R. *Probability: Theory and Examples, Duxbury advanced series, Third edition, Thomson Brooks/Cole*, 2005.

[7] Geman, D., Geman, S. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741*, 1984

[8] Ghosal, S. A Review of Consistency and Convergence of Posterior Distribution. *unpublished (TODO)*

[9] Ghosh, J. K. and Ramamoorthi, R. V. *Bayesian Nonparametrics, Springer, New York* 2003.

[10] Grimmett, G. *Percolation, Second edition, Springer* 1999.

[11] Hurn, M. A., Rue, H., Sheehan, N. A. Block updating in constrained Markov chain Monte Carlo sampling *Statistics & Probability Letters 41, 353-361*, 1999.

[12] Keitt, T. H., Urban D. L., Milne, B. T. Detecting critical scales in fragmented landscapes *Conservation Ecology [online] 1:4* 1997

[13] Lachout, P., Liebscher E., Vogel, S. Strong Concergence of Estimators as ϵ_n-Minimiser of Optimisation Problems *Annals of the Institute of Statistical Mathematics* 2005.

[14] Lange, K. *Numerical analysis for statisticians*, 2000.

[15] Larson, K. Estimation of the passage time distribution on a graph via the EM algorithm. *Research report in mathematical statistics, 1653-0829; 1* 2010.

[16] Lindvall, T. *Lectures on the Coupling Method, Wiley Series in Probability and Statistics*, 1992.

[17] Meester, R., Steif, J. Consistent Estimation of Percolation Quantities. *Statistica Neerlandica* 1998

[18] Meyn S.P., Tweedie R.L. *Markov Chains and Stochastic Stability*, 1993

[19] Roberts, G. O., Rosenthal, J. S. Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov Chains. *TODO* TODO

[20] Sherman, M. On Batch Mean in the Simulation and Statistics Communities. *Proceedings of the 1995 Winter Simulation Conference*, 1995

[21] Tierney, L. Markov Chains for Exploring Posterior Distributions *The Annals of Statistics, Vol. 22, No. 4, 1701-1728*, 1994.

[22] Watanabe, M., Yamaguchi K. *The EM Algorithm and Related Statistical Models, Marcel Dekker* 2004

[23] Wei, G. C. G., Tanner, M. A. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms *Journal of the American Statistical Association, Vol. 85, No. 411, 699-704*, 1990.