Noname manuscript No.
(will be inserted by the editor)

Mark Holmes · Yevhen Mohylevskyy · Charles M. Newman

The voter model chordal interface in two dimensions

Received: date / Accepted: date

Abstract Consider the voter model on a box of side length L (in the triangular lattice) with boundary votes fixed forever as type 0 or type 1 on two different halves of the boundary. Motivated by analogous questions in percolation, we study several geometric objects at stationarity, as $L \to \infty$. One is the interface between the (large – i.e., boundary connected) 0-cluster and 1-cluster. Another is the set of large "coalescing classes" determined by the coalescing walk process dual to the voter model.

Keywords Voter model · coalescing random walks · interface · percolation

Mathematics Subject Classification (2000) 60K35 · 82B41 · 82C22 · 82C24

1 Introduction

In this section we motivate our study of the (two-dimensional) voter model and its dual coalescing walks through their connection with a number of

M. Holmes
Statistics Dept., The University of Auckland
Tel.: +64 9 923 8679
Fax: + 64 9 373 7018
E-mail: m.holmes@auckland.ac.nz

Y. Mohylevskyy
Statistics Dept., The University of Auckland
E-mail: y.mohylevskyy@auckland.ac.nz

C. Newman
Courant Institute of Mathematical Sciences and NYU-Shanghai
E-mail: newman@cims.nyu.edu

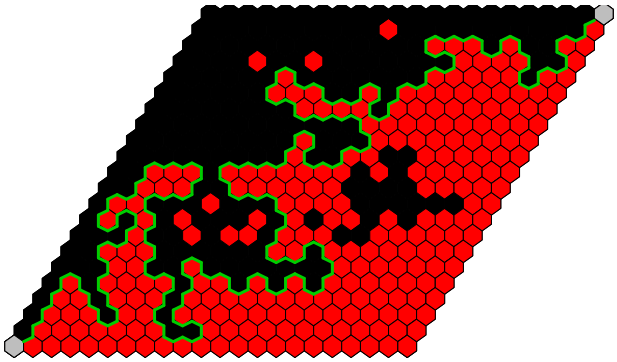

Fig. 1 A possible percolation configuration on B_L with $L = 22$, with the exploration path shown in green.

percolation models. In Section 2, we report on numerical results for the dimension of a natural "chordal interface" of the voter model. In Section 3 we give rigorous (and a few numerical) results on the large coalescing classes for coalescing walks (where vertices x and y in a box are in the same class if their walks coalesce before hitting the boundary), which can be seen as preliminary results for understanding the limiting behaviour of the interface. In the appendix, more details about our numerical results are provided.

Among the most important breakthroughs in statistical physics and probability in the last two decades is the work by Schramm and coauthors [8, 14, 15] and Smirnov [16–18] identifying (or conjecturing) members of the Schramm-Loewner Evolution family of random curves as the scaling limits of various random walks and interfaces in two-dimensional spin systems. In particular Smirnov [16, 17] (see also Camia and Newman's paper [2]) has shown that the scaling limit of critical site-percolation on the triangular lattice $\mathbb{T}$ is SLE_6. To give a rough description of one version of this statement, take a rhombic box B_L (containing $L \times L$ vertices) in the triangular lattice in two dimensions. Label the sides clockwise starting from the southwest corner as $\partial_1, \partial_2, \partial_3, \partial_4$. A percolation configuration on B_L is an element $\omega = (\omega_x)_{x \in B_L}$ of $\Omega_L = \{0,1\}^{B_L}$ defined as follows. Fix the vertices in ∂_1 and ∂_2 to have value 0 (or black or closed) and those in ∂_3 and ∂_4 to be 1 (or red or open). In the interior $\mathring{B}_L$, set each vertex to be (independently) 0 or 1, with probability 1/2 each – see Figure 1. There exists a unique simple path $Z_L(\omega)$ of length $|Z_L|$ from the southwest corner following edges in the dual hexagonal lattice to the opposite corner that keeps black/closed vertices on the left and red/open vertices on the right. Z_L is often referred to as the exploration path; we will also call it the chordal interface. As $L \nearrow \infty$ the law of Z_L, after rescaling,

converges weakly to a probability measure on continuous paths that is the law of *chordal* SLE_6 [2,16] in a rhombic domain. One can use this to prove (see [13, Prop. 2]) that $\mathbb{E}[|Z_L|] \approx L^{7/4}$, where

$$f(L) \approx g(L) \qquad \Longleftrightarrow \qquad \frac{\log(f(L))}{\log(g(L))} \to 1, \quad \text{as } L \to \infty. \tag{1}$$

In general for $\beta > 0$ we have that $f(L) \approx L^\beta$ if and only if

$$f(L) = L^\beta \ell(L), \quad \text{where} \quad \frac{\log \ell(L)}{\log L} \to 0, \quad \text{as } L \to \infty. \tag{2}$$

Note that $\omega \in \Omega_L$ uniquely determines the path $Z_L(\omega)$. One can therefore ask about the limiting behavior of Z_L when the configurations ω are generated by some other process (i.e., not i.i.d. critical site percolation) in the interior $\mathring{B}_L$. In the case of the Ising model (where the states at two sites are not independent) at the critical temperature, Smirnov and coauthors [18,3] have identified that the limiting probability measure is instead chordal SLE_3. We are interested in the limiting behavior of Z_L when the law of the configuration ω is the stationary distribution of the voter model (or related models) on $\mathring{B}_L$.

1.1 The voter model

In this section we define our primary model of interest, on B_L as described above, with boundary states set as 0 on one pair of adjacent sides and 1 on the other pair, while the law of the interior states $\mathring{\omega} = (\omega_x)_{x \in \mathring{B}_L}$ is the stationary measure for the voter model $\{V_t\}_{t \geq 0}$ on $\mathring{B}_L$, as follows.

Each $v \in \mathring{B}_L$ has its own independent Poisson clock (a Poisson process Γ_v) of rate 1. When the clock of a vertex v rings we update the state $V_t(v)$ of v by choosing one of its six neighbors uniformly at random and adopting the state of the chosen neighbor. Note that the neighbor may be one of the vertices in the boundary $\partial B_L = B_L \setminus \mathring{B}_L$ whose state is fixed. Defined this way, V_t is an irreducible Markov process with finite state space $\mathring{\Omega} = \{0,1\}^{\mathring{B}_L}$, and therefore it has a unique invariant distribution. We will write V_∞^L for a random configuration sampled from this invariant distribution.

The process admits a well-known graphical representation (due to T.E. Harris [7]) which we now review. For each $v \in B_L$, we draw a positive half line (representing time) in the third dimension, and on it we mark the times of Poisson clock rings of that vertex. Each mark on a time line represents a state update event which also has an arrow from v to the uniformly chosen neighbor whose state is adopted. The lines of the boundary vertices have arrow marks to them, but not from them, as those states are fixed.

Fix an initial configuration $V_0 = \{V_0(x)\}_{x \in \mathring{B}_L}$. To determine the state of a vertex $v \in \mathring{B}_L$ at time t we start at height/time t on the time line corresponding to v and follow it down until we reach height/time 0 or we encounter an outgoing arrow (whichever comes first) at height $t' \in (0,t)$. If we meet an outgoing arrow we follow it to the time line of a neighboring vertex

v' and proceed as before, following this time line down from height t' until reaching height/time 0 or an outgoing arrow. We stop this procedure when we reach a boundary vertex or height 0 on some time line. Thus from any $v \in \mathring{B}_L$ and $t > 0$ the path followed corresponds to a continuous time nearest neighbor simple random walk on B_L stopped upon reaching a boundary vertex or height 0. In either case the state a at the terminal vertex is known and we set $V_t(v) = a$.

Such a system of "state genealogy walks" from all the vertices at time t following backward in time is a dual model and is distributed as a system of *coalescing* simple symmetric continuous time random walks on the triangular lattice – see for example [6]. Since B_L is finite, if t is large enough all the walks starting then will with high probability hit the boundary before reaching height 0. Indeed, if we continue the time lines and Poisson clocks below height 0 (and do not terminate the walks at height 0) then almost surely from any height t there will be a random height $T_t \in (-\infty, t)$ at which the walks started from all vertices $v \in \mathring{B}_L$ at height t will have reached boundary vertices. What happens on the time lines below height T_t does not affect $\{V_s\}_{s \geq t}$ since the states of the boundary vertices are fixed for all time. This is equivalent to saying that the voter model itself reaches stationarity by a random finite time (distributed as $t - T_t$).

Therefore to sample from V_∞^L it is enough to follow a system of coalescing continuous time simple random walks from each vertex v of $\mathring{B}_L$ until they hit a boundary vertex x_v, and set $V_\infty^L(v) = V_\infty^L(x_v)$, i.e. $V_\infty^L(v) = 0$ if $x_v \in \partial_1 \cup \partial_2$, and $V_\infty^L(v) = 1$ otherwise. One could instead sample from V_∞^L by setting $V_0(v) = 2$ for every $v \in \mathring{B}_L$ (so the state space would become $\{0, 1, 2\}^{\mathring{B}_L}$) and simulating the voter model dynamics until there is no vertex with state 2.

Figure 2 shows a simulation of V_∞^L with $L = 1026$, obtained by simulating coalescing random walks from each vertex in the interior, until each one has reached the boundary.

1.2 Harmonic percolation and related models

The duality discussed in the previous section tells us that $\mathbb{P}(V_\infty(x) = 1)$ (we drop the superscript L when there is no ambiguity) is the probability that a simple random walk started at x first hits the boundary at a 1-site. In other words, the one-dimensional distributions of our voter model on B_L are equal to those of a model we would like to call *harmonic percolation*. This is a model under which the states $\{\omega_v\}_{v \in \mathring{B}_L}$ are independent of each other, and as we have already suggested, $\mathbb{P}(\omega_v = 1)$ is equal to the probability that a simple random walk started from v first hits the boundary ∂B_L at a 1-site (i.e., in $\partial_3 \cup \partial_4$). Harmonic percolation on an infinite strip of thickness L coincides with an independent percolation model called gradient percolation [10–12]. In the case of gradient percolation the probability $p(x)$ of a site x being open changes linearly from one boundary where it is 0 to the other boundary where it is 1. Thus the function $p(x)$ is harmonic inside the strip (with specified boundary conditions). The difference between the voter and

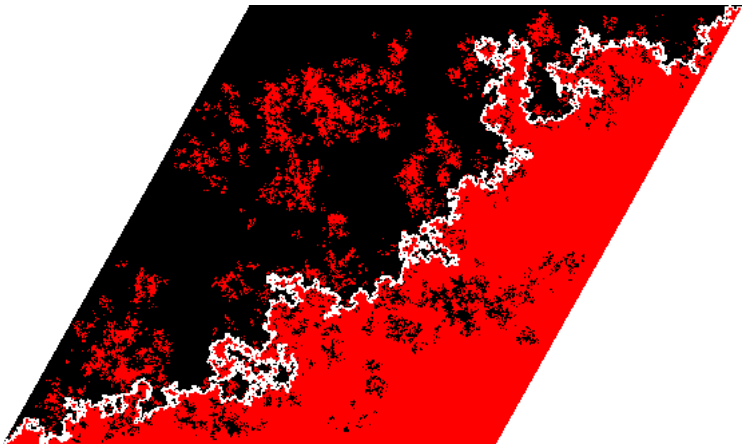

Fig. 2 A single realization of V_∞^L with $L = 1026$, simulated using C++. The white curve is the exploration path or chordal interface.

harmonic percolation models arises from the fact that the walks in the former are coalescing, whereas in the latter they are independent. To be more explicit, coalescence in the voter model leads to non-zero correlations as in the following simple lemma.

Lemma 11. *For any $n > 1$, and any $x_1, \ldots, x_n$ in the interior of B_L,*

$$\mathbb{P}\left(\cap_{i=1}^n \{V_\infty^L(x_i) = 1\}\right) > \prod_{i=1}^n \mathbb{P}\left(V_\infty^L(x_i) = 1\right). \tag{3}$$

Proof Fix L, x_1 and x_2, and let ∂^0 and ∂^1 denote the elements of ∂B_L with fixed states 0 and 1 respectively. Let S_1 and S_2 be two independent random walks starting from x_1 and x_2 respectively. Let $\tau^{x_1x_2} = \inf\{t : S_1(t) = S_2(t)\}$ be the first time that S_1 and S_2 meet each other. Let $S_1' = S_1$ for all times and define

$$S_2'(t) = \begin{cases} S_2(t), & \text{if } t \le \tau^{x_1x_2} \\ S_1(t), & \text{if } t > \tau^{x_1x_2}, \end{cases} \tag{4}$$

so that S_1' and S_2' are coalescing walks started from x_1 and x_2 respectively.

Let $\tau_L^i = \inf\{t : S_i(t) \in \partial B_L\}$ and $\tau'^i_L = \inf\{t : S'_i(t) \in \partial B_L\}$ denote the respective hitting times of the boundary, and note that $\tau'^1_L = \tau_L^1$. Then

$$\begin{aligned}
&\mathbb{P}(V_\infty(x_1) = 1, V_\infty(x_2) = 1) \\
&= \mathbb{P}(S'_1(\tau'^1_L) \in \partial^1,\ S'_2(\tau'^2_L) \in \partial^1) \\
&= \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^1, \tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \\
&\quad + \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^0, \tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \\
&\quad + \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^1, \tau^{x_1x_2} \geq \tau_L^1 \wedge \tau_L^2) \\
&= \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^1) + \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^0, \tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \\
&= \mathbb{P}(V_\infty(x_1) = 1)\mathbb{P}(V_\infty(x_2) = 1) + \mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^0, \tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \qquad (5) \\
&> \mathbb{P}(V_\infty(x_1) = 1)\mathbb{P}(V_\infty(x_2) = 1).
\end{aligned}$$

This proves the result for $n = 2$. A similar coupling argument can be made for any number n of walkers starting from vertices $x_1, \ldots, x_n \in B_L$ (choosing the lower indexed random walker to continue when any two meet), establishing the claim. ∎

For any $\epsilon > 0$, if $x_1(L)$ and $x_2(L)$ are distance at least ϵL from each other and the boundary ∂B_L then there exist $c_\epsilon > 0$ and $C_\epsilon < 1$ such that $\mathbb{P}(V_\infty(x_i) = 1) \in (c_\epsilon, C_\epsilon)$ for $i = 1, 2$ and all L, while

$$\begin{aligned}
&\mathbb{P}(S_1(\tau_L^1) \in \partial^1,\ S_2(\tau_L^2) \in \partial^0, \tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \\
&\qquad \leq \mathbb{P}(\tau^{x_1x_2} < \tau_L^1 \wedge \tau_L^2) \leq \mathbb{P}(\tau_o^\Delta < \tau_{\partial B_{2L}}^\Delta) = O\left(\frac{1}{\log L}\right),
\end{aligned}$$

where τ_o^Δ and $\tau_{\partial B_{2L}}^\Delta$ are times when the difference random walk $S_1(t) - S_2(t)$ started at $x_1 - x_2$ first hits the origin and the boundary of the box B_{2L} respectively, and the last equality follows from Proposition 6.4.3 of [9]. Then (5) implies that the correlation $\rho(V_\infty(x_1), V_\infty(x_2))$ between the votes at x_1 and x_2 goes to zero as per the following.

Lemma 12. *Let $\epsilon > 0$, and $x_1(L)$ and $x_2(L)$ be distance at least ϵL from each other and the boundary ∂B_L. Then $\rho(V_\infty(x_1), V_\infty(x_2)) \to 0$ as $L \to \infty$.*

One can consider i.i.d. percolation, harmonic percolation, and the stationary voter model on B_L as special cases of a general 2-parameter family of models as follows. Start a continuous-time walker from each site. Each walker initially wears a hat. Two walkers wearing hats coalesce when they meet, and instantly become a single walker wearing a hat. Walkers not wearing hats do not coalesce with any other walkers. In addition a Poisson clock is assigned to each walker. When such a clock rings, the walker takes a random walk step, but before doing so removes her coalescence hat with probability q. If a walker wearing a hat steps into a site with another walker with a hat on, the walker that just made a step becomes part of the coalescence set of the walker that was already at the site. Upon hitting a boundary site, with probability $1-p$ a walker (and her entire coalescence set) is assigned the vote

of the boundary vertex she hit, and otherwise (i.e., with probability p) her entire coalescence set attains an independently and uniformly chosen vote. Varying the boundary and coalescence noise parameters p and q between 0 and 1 allows us to interpolate between the four corner models: the voter model $(p,q) = (0,0)$; harmonic percolation $(p,q) = (0,1)$; i.i.d. percolation $(p,q) = (1,1)$; and the case $p = 1,\ q = 0$ corresponds to a model we would like to call cow (coalescing walk) percolation.

2 Interface length

Recall that $|Z_L|$ denotes the length of the interface. Since this path is a nearest neighbor simple path, there exist $c, C > 0$ such that $cL \le |Z_L| \le CL^2$ almost surely. We conjecture that

$$H_L \equiv \mathbb{E}[|Z_L|] \approx L^d \tag{6}$$

for some $d \in [1,2]$. In the case of critical i.i.d. percolation, (6) holds with $d = 7/4 = 1.75$ which is also the Hausdorff dimension of the limiting law (i.e., of SLE_6, see [1]).

For gradient percolation on an infinite strip, the interface curve between the occupied cluster and empty cluster is a.s. unique and has expected length approximately $L^{3/7} l_L$, where l_L is the horizontal length of the piece of strip in which we measure boundary length [10, Proposition 11]. So, for any $\epsilon > 0$, for all sufficiently large L, if we take a piece of strip which is L long (and L thick), the expected length of the interface curve H_L satisfies $L^{10/7-\epsilon} \le H_L \le L^{10/7+\epsilon}$. For any $\delta > 0$, with probability going to 1 with L, the curve stays in the central band (around the central $L/2$ line where $p = \frac{1}{2}$) of width $L^{4/7+\delta}$ [10, Theorem 6]. Thus, as $L \to \infty$, unless we appropriately zoom in around the central line, we expect to see the rescaled interface curve converge to a straight line in the center. Since the harmonic function inside a rhombic area with our boundary condition looks almost linear along the diagonal that connects the middle corner of the 1 valued boundary to the middle corner of the 0 valued boundary (or indeed along any parallel line), we expect that the interface curve for harmonic percolation inside our rhombus should scale to a straight line as well.

Writing $H_L = L^d \ell(L)$ for some function $\ell(L)$ which makes the equality true we have that

$$d = \frac{\log(H_L)}{\log(L)} - \frac{\log(\ell(L))}{\log(L)} = \log_2\left(\frac{H_{2L}}{H_L}\right) - \log_2\left(\frac{\ell(2L)}{\ell(L)}\right). \tag{7}$$

Computing the average interface curve length $\bar{Z}_m(L)$ from m independent realizations of V_∞^L we obtain the following estimators for d based on (7)

$$\tilde{d} = \tilde{d}_{m,L} = \frac{\log(\bar{Z}_m(L))}{\log(L)}, \tag{8}$$

$$\hat{d} = \hat{d}_{m,L} = \log_2\left(\frac{\bar{Z}_m(2L)}{\bar{Z}_m(L)}\right). \tag{9}$$

We say that an estimator $\hat{\beta}$ (more precisely a family of estimators $\{\hat{\beta}_{m,L} : m, L \in \mathbb{N}\}$) is a *consistent* estimator of some quantity β if

$$\lim_{L\to\infty} \lim_{m\to\infty} \hat{\beta}_{m,L} = \beta, \quad \text{almost surely.} \tag{10}$$

It is easy to show that $\tilde{d}$ is a consistent estimator of d if and only if $\log(\ell(L))/\log(L) \to 0$ as $L \to \infty$ (i.e., if and only if $H_L \approx L^d$), while $\hat{d}$ is a consistent estimator for d if and only if $\ell(2L)/\ell(L) \to 1$ as $L \to \infty$. Thus both estimators are consistent if ℓ is slowly varying at ∞.

If we are willing to assume that the random interface length $|Z_L|$ in a box of size L satisfies $|Z_L| = CL^d e^{\varepsilon}$, where ε is independent of L and $\mathbb{E}[\varepsilon] = 0$, then it is natural to consider the ordinary least squares estimator d^* for the slope coefficient d of the simple linear regression model

$$\log(|Z_i|) = d\log(L_i) + a + \varepsilon_i, \tag{11}$$

where $\{|Z_i|\}_{i\le n}$ are interface lengths on boxes of side lengths $\{L_i\}_{i\le n}$, and the ε_i are random variables (independent of L_i) with mean 0. Note that $\ell(L)$ is constant under this assumption. In the cases of the voter model and cow percolation, where we might expect log corrections to appear (see for example Theorem 31) if one assumes that $|Z_L| = CL^d(\log L)^{\zeta} e^{\varepsilon}$ then it is natural to consider the estimator d^{**} for the slope coefficient d of the modified simple linear regression model

$$\log(|Z_i|) = d\log(L_i) + \zeta\log(\log(L_i)) + a + \varepsilon_i. \tag{12}$$

In fitting this model we introduce a significant new problem due to the fact that $\log\log L$ varies very little (so is difficult to distinguish from the constant term a) while at the same time it is severely correlated with the main explanatory term $\log L$. This will result in poor estimates for ζ. We will let the data decide which of d^* and d^{**} is appropriate, based on the explanatory power of the variable $\log\log L$ in the regression model (12).

Estimates $\hat{d}, \tilde{d}, d^*, d^{**}$ of d, each based on $m = 10000$ independent simulations appear in Table 1. Since we make no assumptions about the distribution of the interface lengths (other than what we have already discussed), we have not constructed confidence intervals for the true value of the parameter, but instead have used bootstrapping methods to estimate the variability in our point estimates for $\hat{d}$ and $\tilde{d}$. For d^* (resp. d^{**}) we simply note that the standard error output from fitting a simple linear regression model is at most 0.002 (resp. 0.04) in each case. Each interval in Table 1 is the result of taking 10000 bootstrap samples (with replacement) each of size 10000 from the data, computing the relevant statistic for each bootstrap sample, and removing the smallest and largest 2.5% of values. See Section 4.1 for further details.

For ordinary percolation and harmonic percolation the true values are known (or expected) to be $7/4 = 1.75$ and $10/7 \approx 1.4286$, so for these models $\hat{d}$ (with $L = 512$ and $2L = 1024$) performs the best, while the linear regression estimator d^* also performs quite well. The estimator d^{**} is not given because for these two models the additional variable $\log(\log(L))$ does not have significant explanatory power (and results in a poorer estimate for d). For both

	$\hat{d}$			$\tilde{d}$				d^*
L	128	256	512	128	256	512	1024	(d^{**})
voter	1.450	1.461	1.463	1.595	1.577	1.564	1.554	1.459
	(1.442,1.458)	(1.453,1.469)	(1.455,1.470)					(1.514)
cow.	1.481	1.488	1.495	1.669	1.645	1.628	1.615	1.488
	(1.470,1.493)	(1.476,1.499)	(1.484,1.507)					(1.553)
harm.	1.423	1.422	1.429	1.625	1.600	1.580	1.565	1.426
	(1.419,1.427)	(1.418,1.426)	(1.426,1.432)					
perc.	1.740	1.746	1.751	1.849	1.836	1.826	1.818	1.746
	(1.729,1.751)	(1.735,1.757)	(1.740,1.762)					

Table 1 Estimates of d with $m = 10000$ and with $L = 128, 256, 512$ (and 1024), rounded to 3 decimal places. The brackets are "bootstrap 95% confidence intervals" obtained from the quantiles of 10000 bootstrap estimates. For $\tilde{d}$ the intervals all have width at most 0.002 (similarly the standard error given by the fitted linear model in estimating d^* is at most 0.002).

of these models (and for small values of L), $\hat{d}$ and $\tilde{d}$ appear to systematically underestimate and overestimate d respectively, with bias typically decreasing with L.

While the above discussion says very little about the relative performance of the estimators in any other setting (so in particular in the case of our dependent models, voter and cow), based on the fact that the $\hat{d}$ and $\tilde{d}$ estimates in these settings are also increasing and decreasing in L, one can hope that the true value lies between them (in fact the d^{**} estimates do lie between them). If so then the values of d for percolation and cow percolation would be different, and then one might also expect the voter model and harmonic percolation to have different d.

3 The sizes of coalescing classes and related questions

The interface curve cannot pass through any connected cluster of common votes. The difference between the voter and harmonic percolation models is that the states are determined by coalescing random walks rather than independent random walks (started at each site). If the coalescing classes in the voter model are negligible as $L \uparrow \infty$, both in terms of size and the correlation between votes in different classes, then perhaps some kind of rescaling argument would allow one to compare the voter model to the harmonic percolation model. One expects that the rescaled interface curve for harmonic percolation on B_L converges to a straight line (Pierre Nolin has proved this on the strip [10]), so one might expect the same to be true for the voter model, if the coalescing classes are indeed negligible as $L \uparrow \infty$.

Clustering behaviour for the 2-dimensional voter model has been well studied in the probability literature (see e.g. [5]), but (as far as we know) not in the current setting of a finite domain with unflinching boundary. As a small step in the direction of understanding the correlation between votes in different classes, let us verify that any two sites x and y are less likely to share a common vote (than they otherwise would be) if they are not in the same coalescing class. Fix L and start coalescing walks from every site

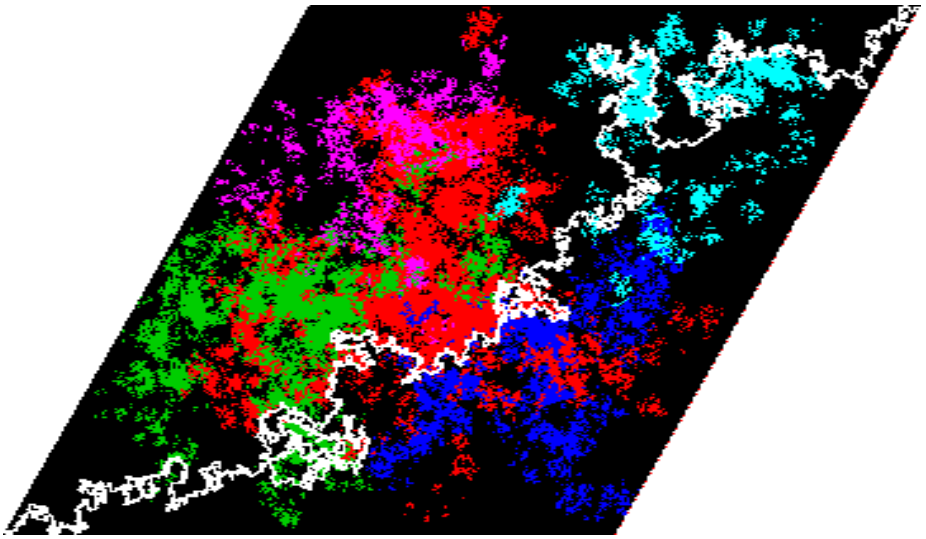

Fig. 3 A single realization of the 5 largest coalescing classes (colors other than black) for $L = 1026$. The white curve is the chordal interface.

in $\mathring{B}_L$. The walks define an equivalence relation on $\mathring{B}_L$ in the sense that $x \sim y$ if and only if the walks started from x and y coalesce before hitting the boundary. Let $C_L(x)$ denote the (random) equivalence class of x. Let $A = A_{xy} = \{V_\infty(x) = V_\infty(y)\}$ and $B = B_{xy} = \{C_L(x) = C_L(y)\} = \{x \sim y\}$. Then $\mathbb{P}(B) \in (0,1)$ and

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) = \mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) > \mathbb{P}(A|B^c),$$

as claimed. On the other hand, as we have seen earlier, if $x = x(L)$ and $y = y(L)$ are distance at least ϵL apart then $\mathbb{P}(B^c) \to 1$ and $\mathbb{P}(A|B^c) - \mathbb{P}(A) \to 0$ as $L \nearrow \infty$, with $\mathbb{P}(A)$ being bounded away from zero as $L \nearrow \infty$ if x and y are also at least ϵL distance from the boundary.

We are hereafter interested in the behavior of the expected size of the class of the centre of the rhombus B_L, which we will for convenience take to be the origin (if L is not odd we consider the centre/origin to be any one of the closest vertices to the centre) and the expected size of the largest class $\mathbb{E}[|M_L|]$ as $L \to \infty$, where

$$M_L = C_L(x'), \text{ where } x' \text{ is chosen such that } |C_L(x')| = \max_{x \in \mathring{B}_L} |C_L(x)|, \quad (13)$$

where some tie-breaking rule is used to choose x, if necessary. In particular, we ask what proportion of all vertices in the box are in the largest class, as $L \to \infty$? Since $|\mathring{B}_L| \approx cL^2$, we are interested in $\lim_{L\to\infty} \mathbb{E}\left[|M_L|/L^2\right]$. Figure 3 shows a single realization of the 5 largest classes for $L = 1026$.

The following is our main rigorous result, which shows that coalescing classes have (on average) small (but only logarithmically small) volume compared to the whole box. We note that the lower bound can be improved slightly with a little more effort.

Theorem 31. *There are constants C' and C'' in $(0,\infty)$, such that*

$$\frac{C'}{(\log L)^{713}} \le \mathbb{E}\left[\frac{|C_L(o)|}{|\mathring{B}_L|}\right] \le \mathbb{E}\left[\frac{|M_L|}{|\mathring{B}_L|}\right] \le \left(\frac{C''}{\log L}\right)^{1/2}. \tag{14}$$

Proof First note that to have B_L centered at the origin and to have the smaller rhombi used in the proof to be consistent with the lattice we need L to be such that $L-1$ is divisible by 12, and then for rhombi of fractional side lengths such as $L/2$ we should use $B_{\frac{L-1}{2}+1}$ instead of $B_{L/2}$. For notational convenience we will ignore these issues, but we note though that the same arguments would in any case work with trivial but messy modifications.

For both the upper and lower bounds in (14) we will use the fact that, for $x \in \mathring{B}_L$,

$$\mathbb{E}\left[|C_L(x)|\right] = \mathbb{E}\left[\sum_{y\in\mathring{B}_L} 1_{\{y\in C_L(x)\}}\right] = \sum_{y\in\mathring{B}_L} \mathbb{P}(y \in C_L(x)). \tag{15}$$

To verify the lower bound, let $\{S_x\}_{x\in\mathring{B}_L} = \{\{S_x(t)\}_{t\ge 0}\}_{x\in\mathring{B}_L}$ be independent continuous-time (with jump rate 1), nearest-neighbor random walks on the triangular lattice, with respective starting points $S_x(0) = x \in \mathring{B}_L$. For $x, y \in \mathring{B}_L$, let $\tau^{xy} = \inf\{t : S_x(t) = S_y(t)\}$, and $\tau_L^x = \inf\{t : S_x(t) \in \partial B_L\}$ denote the meeting times and boundary hitting times respectively. Then (15) with $x = o$ can be written as

$$\mathbb{E}\left[|C_L(o)|\right] = \sum_{x\in\mathring{B}_L} \mathbb{P}\left(\tau^{ox} < \tau_L^o \wedge \tau_L^x\right). \tag{16}$$

By Lemma 32 below there is a positive constant c' such that for each $x \in B_{L/2} \setminus B_{L/4}$,

$$\mathbb{P}\left(\tau^{ox} < \tau_L^o \wedge \tau_L^x\right) \ge \frac{c'}{(\log L)^{713}}. \tag{17}$$

Combining (16) and (17) we obtain

$$\mathbb{E}\left[\frac{|C_L(o)|}{|\mathring{B}_L|}\right] \ge \frac{1}{cL^2} \sum_{x\in B_{L/2}\setminus B_{L/4}} \mathbb{P}\left(\tau^{ox} < \tau_L^o \wedge \tau_L^x\right) \ge \frac{C'}{(\log L)^{713}}, \tag{18}$$

for another positive constant C', which verifies the lower bound in (14).

To establish the upper bound, note that for any $x, y \in \mathring{B}_L$ the difference walk $S^{\Delta}_{x-y}(t) = S_x(t) - S_y(t)$ is also a simple symmetric random walk started at $x - y$ but with jump rate 2. Let $\tau_o^{\Delta_x} = \inf\{t : S_x^{\Delta}(t) = o\}$ and $\tau_L^{\Delta_x} = \inf\{t : S_x^{\Delta}(t) \in \partial B_L\}$ be the first hitting times of the origin and the boundary ∂B_L

by the difference walk S_x^{Δ}, and let $\tau_o^x = \inf\{t : S_x(t) = o\}$ be the first time S_x hits the origin. Then

$$\begin{aligned}\frac{\mathbb{E}[|C_L(x)|]}{cL^2} &= \frac{1}{cL^2}\sum_{y\in\mathring{B}_L}\mathbb{P}(\tau^{xy} < \tau_L^x \wedge \tau_L^y)\\ &\le \frac{1}{cL^2}\sum_{y\in\mathring{B}_L}\mathbb{P}(\tau_o^{\Delta_{x-y}} < \tau_{2L}^{\Delta_{x-y}})\\ &\le \frac{1}{cL^2}\sum_{y\in B_{2L}\setminus o}\mathbb{P}(\tau_o^y < \tau_{2L}^y) + \frac{\mathbb{P}(\tau_o^o < \tau_{2L}^o)}{cL^2}. \end{aligned} \tag{19}$$

Using Theorem 6.4.3 of [9] on the summation, (19) is bounded above by

$$\frac{C}{L^2}\sum_{y\in B_{2L}\setminus o}\frac{\log(L/|y|)}{\log L} + \frac{1}{cL^2}. \tag{20}$$

Next, we split the sum into dyadic annuli (all but finitely many of which contain no vertices). Since B_L has been defined via the number of vertices on the boundary (so only for integer L), we let $\hat{B}_R$ denote the rhombic $R\times R$ box centered at the origin in $\mathbb{R}^2$ and let $A_{L,k}$ denote the (possibly empty) intersection of $(\hat{B}_{2L/2^{k-1}} \setminus \hat{B}_{2L/2^k})$ with the triangular lattice. Then the first term in (20) is equal to

$$\begin{aligned}\frac{C}{L^2}\sum_{k=1}^{\infty}\sum_{y\in A_{L,k}}\frac{\log(L/|y|)}{\log L} &\le \frac{C}{L^2}\sum_{k=1}^{\infty}\sum_{y\in A_{L,k}}\frac{\log(2^{k-1})}{\log L}\\ &\le \frac{C'}{L^2}\sum_{k=1}^{\infty}\left(\frac{2L}{2^k}\right)^2\cdot\frac{k-1}{\log L}\\ &\le \frac{C''}{\log L}.\end{aligned}$$

Therefore

$$\sup_{x\in\mathring{B}_L}\mathbb{E}\left[\frac{|C_L(x)|}{|\mathring{B}_L|}\right] \le \frac{C'''}{\log L}. \tag{21}$$

Markov's inequality gives $(\epsilon|\mathring{B}_L|)^{-1}\mathbb{E}\left[|C_L(x)|\right] \ge \mathbb{P}(|C_L(x)| > \epsilon|\mathring{B}_L|)$ for any $\epsilon = \epsilon(L) > 0$, so

$$\sup_{x\in\mathring{B}_L}\mathbb{P}(|C_L(x)| > \epsilon|\mathring{B}_L|) \le \frac{1}{\epsilon}\sup_{x\in\mathring{B}_L}\mathbb{E}\left[\frac{|C_L(x)|}{|\mathring{B}_L|}\right] \le \frac{1}{\epsilon}\cdot\frac{C'''}{\log L}. \tag{22}$$

It follows that

$$\mathbb{E}\left[\frac{|M_L|}{|\mathring{B}_L|}\right] \le \frac{1}{|\mathring{B}_L|}\mathbb{E}\left[|M_L|1_{\{|M_L|>\epsilon|\mathring{B}_L|\}}\right] + \epsilon = \frac{1}{|\mathring{B}_L|}\mathbb{E}\left[\sum_{x\in\mathring{B}_L} 1_{\{x\in M_L\}}1_{\{|M_L|>\epsilon|\mathring{B}_L|\}}\right] + \epsilon$$
$$\le \frac{1}{|\mathring{B}_L|}\mathbb{E}\left[\sum_{x\in\mathring{B}_L} 1_{\{|C_L(x)|>\epsilon|\mathring{B}_L|\}}\right] + \epsilon = \frac{1}{|\mathring{B}_L|}\sum_{x\in\mathring{B}_L}\mathbb{P}(|C_L(x)|>\epsilon|\mathring{B}_L|) + \epsilon \tag{23}$$
$$\le \frac{1}{|\mathring{B}_L|}\sum_{x\in\mathring{B}_L}\sup_{x\in\mathring{B}_L}\mathbb{P}(|C_L(x)|>\epsilon|\mathring{B}_L|) + \epsilon = \sup_{x\in\mathring{B}_L}\mathbb{P}(|C_L(x)|>\epsilon|\mathring{B}_L|) + \epsilon$$
$$\le \frac{1}{\epsilon}\cdot\frac{C''}{\log L} + \epsilon.$$

Choose $\epsilon(L) = (\frac{1}{\log L})^{1/2}$ to get the claimed upper bound. ∎

Lemma 32. *Fix some small $\epsilon > 0$. There exists $c' > 0$, such that for all $x \in B_{L/2} \setminus B_{\epsilon L}$,*

$$\mathbb{P}\left(\tau^{ox} < \tau_L^o \wedge \tau_L^x\right) \ge \frac{c'}{(\log L)^{713}}. \tag{24}$$

Proof The triangular lattice is constructed from 3 families of parallel lines (or "directions"), denoted by $D = \{\nearrow, \nwarrow, \leftrightarrow\}$, with each vertex being at the intersection of 3 such lines (one from each family), and having 6 nearest neighbors corresponding to moving "up" or "down" in any one of these directions. We define a system of two dependent *discrete-time* random walks $\hat{S}_o$ and $\hat{S}_x$ on the triangular lattice in the following way: $\hat{S}_o(0) = o$ and $\hat{S}_x(0) = x$; for each $t \in \mathbb{N}$ toss a fair coin to decide which of $\hat{S}_o(t)$ or $\hat{S}_x(t)$ makes an i.i.d. uniformly chosen nearest-neighbor step on the triangular lattice (while the other does not move). Let $\hat{\tau}^{ox}$, $\hat{\tau}_L^x$, and $\hat{\tau}_L^o$ be the corresponding meeting and boundary hitting times for $\hat{S}_o(t)$ and $\hat{S}_x(t)$. Since $\{(\hat{S}_o(t), \hat{S}_x(t))\}_{t\in\mathbb{Z}_+}$ has the same law as the jump process of $\{(S_o(t), S_x(t))\}_{t\in\mathbb{R}_+}$ and the event in (24) depends only on the relative sizes of the hitting times, we have

$$\mathbb{P}\left(\tau^{ox} < \tau_L^o \wedge \tau_L^x\right) = \mathbb{P}\left(\hat{\tau}^{ox} < \hat{\tau}_L^o \wedge \hat{\tau}_L^x\right). \tag{25}$$

Now we focus on the discrete time random walks $\hat{S}_o(t)$ and $\hat{S}_x(t)$. For $t \in \mathbb{N}$, let $\mathcal{R}_t$ denote the set of ordered partitions $r = (r_\nearrow, r_\nwarrow, r_\leftrightarrow)(t)$ of $\{1, 2, \ldots, t\}$ into three (possibly empty) sets. For $s \le t$, let $h(s) \in D$ denote the direction of the step taken by (one of) the pair $(\hat{S}_o, \hat{S}_x)$ at time s. For $r = (r_\nearrow, r_\nwarrow, r_\leftrightarrow) \in \mathcal{R}_t$ let A_r be the event that for each $\bullet \in \{\nearrow, \nwarrow, \leftrightarrow\}$ and $s \in r_\bullet$, $h(s) = \bullet$, i.e., that steps in direction $\nearrow$ are taken at times in $r_\nearrow$ etc. Conditioning on A_r we can rewrite the probability above as follows

$$\mathbb{P}\left(\hat{\tau}^{ox} < \hat{\tau}_L^o \wedge \hat{\tau}_L^x\right) = \sum_{t=0}^{\infty}\sum_{r\in\mathcal{R}_t}\mathbb{P}\left(\hat{\tau}^{ox} = t, \hat{\tau}_L^o \wedge \hat{\tau}_L^x > t \mid A_r\right)\mathbb{P}\left(A_r\right). \tag{26}$$

Let S^{Δ} and S^{Σ} denote the difference and sum walks starting at x, defined by $S^{\Delta}(t) = \hat{S}_x(t) - \hat{S}_o(t)$ and $S^{\Sigma}(t) = \hat{S}_x(t) + \hat{S}_o(t)$, and let τ_o^{Δ} and τ_L^{Δ}, and τ_o^{Σ} and τ_L^{Σ} be the hitting times of the origin and the boundary ∂B_L by the difference and sum walks respectively. Since

$$\hat{S}_x(t) = \frac{\hat{S}_x(t) - \hat{S}_o(t)}{2} + \frac{\hat{S}_x(t) + \hat{S}_o(t)}{2}, \tag{27}$$

we have

$$|\hat{S}_x(t)| \geq L \Rightarrow |\hat{S}_x(t) - \hat{S}_o(t)| \geq L \text{ or } |\hat{S}_x(t) + \hat{S}_o(t)| \geq L. \tag{28}$$

A similar statement holds for $\hat{S}_o(t)$. Thus we have

$$\{\hat{\tau}_L^o \wedge \hat{\tau}_L^x > t\} \supseteq \{\tau_L^{\Sigma} > t\} \cap \{\tau_L^{\Delta} > t\}. \tag{29}$$

Therefore (26) can be continued as follows

$$\begin{aligned}
&\sum_{t=0}^{\infty} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\hat{\tau}^{ox} = t, \hat{\tau}_L^o \wedge \hat{\tau}_L^x > t \mid A_r\right) \mathbb{P}\left(A_r\right) \\
&\quad \geq \sum_{t=0}^{\infty} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau_o^{\Delta} = t, \tau_L^{\Delta} > t, \tau_L^{\Sigma} > t \mid A_r\right) \mathbb{P}\left(A_r\right) \\
&\quad = \sum_{t=0}^{\infty} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau_o^{\Delta} = t, \tau_L^{\Delta} > t | A_r\right) \mathbb{P}\left(\tau_L^{\Sigma} > t \mid A_r\right) \mathbb{P}\left(A_r\right),
\end{aligned} \tag{30}$$

where the last equality follows from the fact that the sum and difference walks are conditionally independent given A_r (e.g., if we know that the sum walk makes a positive step in a specific direction, the difference walk is still equally likely to make either a positive or negative step in that direction).

Let $x \in B_{L/2}$. Truncating the infinite sum and using Lemma 33 below we have

$$\begin{aligned}
&\sum_{t=0}^{\infty} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau_o^{\Delta} = t, \tau_L^{\Delta} > t | A_r\right) \mathbb{P}\left(\tau_L^{\Sigma} > t \mid A_r\right) \mathbb{P}\left(A_r\right) \\
&\quad \geq \sum_{t=0}^{n} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau_o^{\Delta} = t, \tau_L^{\Delta} > t | A_r\right) \mathbb{P}\left(\tau_L^{\Sigma} > t \mid A_r\right) \mathbb{P}\left(A_r\right) \\
&\quad \geq \mathbb{P}\left(\check{\tau}_{\pm L/12}^o > n\right) \sum_{t=0}^{n} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau_o^{\Delta} = t, \tau_L^{\Delta} > t \mid A_r\right) \mathbb{P}\left(A_r\right),
\end{aligned} \tag{31}$$

where $\check{\tau}_{\pm K}^u$ denotes the (discrete) time a one-dimensional simple symmetric random walk started at $u \in [-K, K]$ first hits $+K$ or $-K$. Summarizing from

(26) until this point, and continuing we have

$$\begin{aligned}
\mathbb{P}\left(\hat{\tau}^{ox} < \hat{\tau}^o_L \wedge \hat{\tau}^x_L\right) &\geq \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \sum_{t=0}^{n} \sum_{r \in \mathcal{R}_t} \mathbb{P}\left(\tau^{\Delta}_o = t, \tau^{\Delta}_L > t \mid A_r\right) \mathbb{P}\left(A_r\right) \\
&= \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \mathbb{P}\left(\tau^{\Delta}_o \leq n, \tau^{\Delta}_o < \tau^{\Delta}_L\right) \\
&\geq \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \mathbb{P}\left(\tau^{\Delta}_o < \tau^{\Delta}_L \leq n\right) \\
&\geq \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \left[\mathbb{P}\left(\tau^{\Delta}_o < \tau^{\Delta}_L\right) - \mathbb{P}\left(\tau^{\Delta}_L > n\right)\right] \\
&\geq \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \left[\mathbb{P}\left(\tau^{\Delta}_o < \tau^{\Delta}_L\right) - \mathbb{P}\left(\check{\tau}^o_{\pm L/2} > \frac{n}{2}\right)\right], \quad (32)
\end{aligned}$$

where the last inequality follows from Lemma 33 below.

Let $R_n = \left|\{z \in \mathbb{Z} : S^o(t) = z \text{ for some } t \leq n\}\right|$ be the size of the range of a one-dimensional discrete-time random walk S^o (started at o) up to time n. Then

$$\mathbb{P}(R_n \leq L) \leq \mathbb{P}(\check{\tau}^o_{\pm L} > n) \leq \mathbb{P}(R_n \leq 2L). \quad (33)$$

According to Theorem 2 of [4], for any sequence $b_n = o(n)$ diverging to $+\infty$

$$\lim_{n\to\infty} \frac{1}{b_n} \log \mathbb{P}\left(R_n \leq \sqrt{\frac{n}{b_n}}\right) = -\frac{\pi^2 \sigma}{2}, \quad (34)$$

where $\sigma = \mathbb{E}\left[S^o(1)^2\right] = 1$. Letting $n = L^2 \log\log L$, and $b_n = (12)^2 \log\log L$, (33) and (34) yield (for $\delta \in (0, 1/711)$) that for large n (and L),

$$\mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \geq \mathbb{P}(R_n \leq \sqrt{n/b_n}) \geq \frac{1}{(\log L)^{\frac{(12\pi)^2(1+\delta)}{2}}} \geq \frac{1}{(\log L)^{712}}. \quad (35)$$

Similarly, with $n/2 = \frac{L^2}{2} \log\log L$ and $b_{n/2} = \frac{1}{2} \log\log L$, (33) and (34) yield (for $\delta \in (0, 1 - 8/\pi^2)$) that for large n (and L),

$$\mathbb{P}\left(\check{\tau}^o_{\pm L/2} > \frac{n}{2}\right) \leq \mathbb{P}\left(R_{n/2} \leq \sqrt{(n/2)/(b_{n/2})}\right) \leq \frac{1}{(\log L)^{\frac{\pi^2}{4}(1-\delta)}} \leq \frac{1}{(\log L)^2}. \quad (36)$$

According to Theorem 6.4.3 of [9] the term $\mathbb{P}\left(\tau^{\Delta}_o < \tau^{\Delta}_L\right)$ for $x \in B_{L/2} \setminus B_{\epsilon L}$ can be bounded uniformly from below and above by $\frac{c}{\log L}$ and $\frac{C}{\log L}$ for some positive constants c and C and L large enough. Inserting this estimate, (35) and (36) into (32) verifies that there exists a constant $c' > 0$ such that for all L, uniformly in $x \in B_{L/2} \setminus B_{\epsilon L}$,

$$\mathbb{P}\left(\hat{\tau}^{ox} < \hat{\tau}^o_L \wedge \hat{\tau}^x_L\right) \geq \frac{c'}{(\log L)^{713}}, \quad (37)$$

as required. ∎

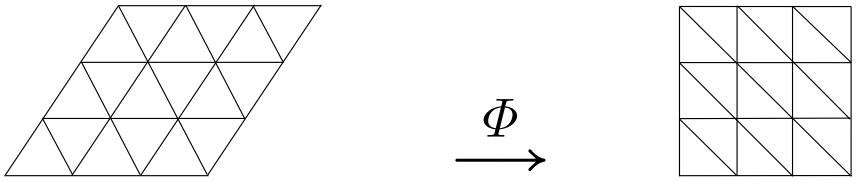

Fig. 4 A mapping between two planar embeddings of the triangular lattice.

Lemma 33. *With the definitions of* $\check{\tau}^o_{\pm L}$, τ^Σ_L, τ^Δ_L, *and* A_r *as in the proof of Lemma 32, for any* $r \in \mathcal{R}_t$ *and* $t \le n$ *we have*

$$\mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right) \le \mathbb{P}\left(\tau^\Sigma_L > t \mid A_r\right) \tag{38}$$

$$\mathbb{P}\left(\tau^\Delta_L > n\right) \le \mathbb{P}\left(\check{\tau}^o_{\pm L/2} > \frac{n}{2}\right). \tag{39}$$

Proof To verify the first claim, first recall the definition of $\mathcal{R}_t$ after (25). For each $r \in \mathcal{R}_t$ we construct a two-dimensional random walk on the triangular lattice (started at $x \in B_{L/2}$, with $i = |r_{\swarrow}|, j = |r_{\nwarrow}|$ and $k = |r_{\leftrightarrow}|$ steps along the three directions in D respectively) from a one-dimensional random walk of t steps in the following way: let the one-dimensional walk be $S^1(t) = \sum_{\ell=1}^t X_\ell$ with each $X_\ell \in \{\pm 1\}$; designate the first i steps to be in direction "$\swarrow$", the next j steps to be in direction "$\nwarrow$", and the final k steps to be in direction "$\leftrightarrow$"; construct the two-dimensional random walk starting at x by picking steps from each group according to the partition r (preserving the order of steps within each of the groups). If the one-dimensional walk started at the origin stays confined to the interval $(-L/12, L/12)$, then the first i steps, next j steps and next k steps have displacements from their respective starting points at most $L/12$, $2L/12$, and $2L/12$ respectively and the two-dimensional walk started at x stays confined to $\mathring{B}_L$. Therefore we have for $t \le n$ that

$$\mathbb{P}\left(\tau^\Sigma_L > t \mid A_r\right) \ge \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > t\right) \ge \mathbb{P}\left(\check{\tau}^o_{\pm L/12} > n\right). \tag{40}$$

This verifies the first claim.

For the second claim, we consider two one-dimensional random walks, $S^{1,\Delta}$ and $S^{2,\Delta}$, that are the following "projections" of the two-dimensional discrete-time (difference) random walk S^Δ starting at x onto the lines parallel to the two sides of the rhombic box B_L. Under the linear transformation Φ depicted in Figure 4 "projections" are simply standard orthogonal projections onto the two coordinate axes. Thus, the walk $S^{1,\Delta}$ makes no step when S^Δ steps in the direction $\swarrow$, makes a step -1 when the increment of S^Δ is either $\leftarrow$ or $\nwarrow$, and $+1$ when the increment of S^Δ is either $\rightarrow$ or $\searrow$. Similarly $S^{2,\Delta}$ does not move when S^Δ steps in the direction $\leftrightarrow$, while it makes an increment -1 (resp., $+1$) when the increment of S^Δ is $\swarrow$ or $\searrow$ (resp., $\nearrow$ or $\nwarrow$). Let $\tau^{i,\Delta}_{\pm L}$ be the hitting time of $\pm L$ by $S^{i,\Delta}$, and M^i_n be the number of steps made by

$S^{i,\Delta}$ by the time S^{Δ} makes n steps. Then we have

$$\begin{aligned}
\mathbb{P}\left(\tau_L^{\Delta} > n\right) &= \mathbb{P}\left(\tau_{\pm L/2}^{1,\Delta} > n, \tau_{\pm L/2}^{2,\Delta} > n\right) \qquad (41)\\
&= \mathbb{P}\left(\tau_{\pm L/2}^{1,\Delta} > n, \tau_{\pm L/2}^{2,\Delta} > n \mid M_n^1 \geq M_n^2\right) \times \mathbb{P}(M_n^1 \geq M_n^2)\\
&\quad + \mathbb{P}\left(\tau_{\pm L/2}^{1,\Delta} > n, \tau_{\pm L/2}^{2,\Delta} > n \mid M_n^1 < M_n^2\right) \times \mathbb{P}(M_n^1 < M_n^2)\\
&\leq \mathbb{P}\left(\tau_{\pm L/2}^{1,\Delta} > n \mid M_n^1 \geq M_n^2\right) \times \mathbb{P}(M_n^1 \geq M_n^2)\\
&\quad + \mathbb{P}\left(\tau_{\pm L/2}^{2,\Delta} > n \mid M_n^1 < M_n^2\right) \times \mathbb{P}(M_n^1 < M_n^2)\\
&\leq \mathbb{P}\left(\check{\tau}_{\pm L/2}^{o} > \frac{n}{2}\right).
\end{aligned}$$

■

Theorem 31 and the discussion preceding it lead us to make the following conjecture.

Conjecture 34. *The interface curve of the voter model in B_L converges to a straight line as $L \to \infty$.*

As in the case of harmonic percolation, the density of 1 (red) votes above any line ℓ_L (of distance $\epsilon\ell$ from the corners of the rhombus) parallel to the diagonal does not converge to 0 in the limit as $L \to \infty$. Indeed the expected proportion $\mathbb{E}[X_{\ell_L}]$ of 1's above such a line is the same for the two models, and is bounded below by a constant c. Moreover $\mathbb{P}(X_{\ell_L} \geq c/2) > c/2$ since

$$c \leq \mathbb{E}[X_{\ell_L}] = \mathbb{E}[X_{\ell_L} 1_{\{X_{\ell_L} \geq c/2\}}] + \mathbb{E}[X_{\ell_L} 1_{\{X_{\ell_L} < c/2\}}] \leq \mathbb{P}(X_{\ell_L} \geq c/2) + c/2.$$

Theorem 31 also provides us with a test of the quality of our numerical estimation techniques (which are of course for finite L). Having established that $\mathbb{E}[|C_L(o)|] \approx L^{\gamma}$ and $\mathbb{E}[|M_L|] \approx L^{\beta}$ with $\gamma = \beta = 2$, we estimated the exponents from simulation data with estimators as in (8),(9), (11), and (12) giving

$$\tilde{\gamma} = 1.579, \qquad \underset{(1.833,1.889)}{\hat{\gamma} = 1.861}, \qquad \gamma^{*} = 1.841, \qquad \gamma^{**} = 2.020 \qquad (42)$$

$$\tilde{\beta} = 1.657, \qquad \underset{(1.897,1.924)}{\hat{\beta} = 1.911}, \qquad \beta^{*} = 1.897, \qquad \beta^{**} = 2.048. \qquad (43)$$

Except for the $\bullet^{**}$ estimates, all of these underestimate the true value (the $\hat{\bullet}$ and $\tilde{\bullet}$ esimates are increasing with L). The regression estimates $\bullet^{**}$ are closest to the true value, with the $\hat{\bullet}$ estimates performing next best. These observations are perhaps not surprising since the true value also corresponds to an upper bound in that $|M_L| \leq CL^2$ almost surely (and in particular this guarantees that the $\tilde{\bullet}$ estimators will be smaller than 2).

Figure 3 suggests that the (largest) coalescing classes are rather disconnected and sparse, which poses a potential problem for a rescaling argument like that mentioned at the beginning of Section 3. This is because the coalescing classes will not scale to single points if their diameters are $\geq cL$

with non-vanishing probability. It is an open problem to prove that for some $c, c' \in (0,1)$, $\liminf_{L\to\infty} \mathbb{P}(\exists x, y \in B_L : |x-y| > cL, \tau^{xy} < \tau_x \wedge \tau_y) \geq c'$. This would imply that with positive probability there are coalescing classes with diameter at least cL. A very large proportion of our simulated curves cut through M_L in the sense that the interface curve has sites belonging to M_L on both sides. The proportion increases from 0.9819 for the boxes of size 128, to 0.9985 for the boxes of size 1024. Thus, the connected clusters/subsets $C_L^c(x)$ (containing $x \in B_L$) of coalescing classes $C_L(x)$, may be better candidates to use in rescaling arguments as the interface curve has to go around them. Assuming that $\mathbb{E}\left[|C_L^c(x)|\right] \approx L^{\gamma'}$ and $\mathbb{E}\left[\max_{x\in B_L} |C_L^c(x)|\right] \approx L^{\beta'}$, we obtain the particularly unreliable estimates (see Section 4.3) $\hat{\gamma}' = 1.548$ and $\hat{\beta}' = 1.741$.

Another piece of information that could in principle support Conjecture 34 is the displacement of the curve Z_L from its conjectured diagonal limit, D. Assuming that $\mathbb{E}\left[\max_{x\in Z_L} \min_{y\in D} |x-y|\right] \approx L^{\alpha}$, we obtained the point estimates $\hat{\alpha} = 0.969$ and $\tilde{\alpha} = 0.816$. The estimate $\hat{\alpha}$ (with $L = 512$ and $2L = 1024$) is disconcertingly close to 1 and moreover did not systematically decrease as we varied $L = 128, 256, 512$. The estimate $\tilde{\alpha}$ increased in L from a value of 0.750 when $L = 128$. Thus, estimates based on our displacement data provide little or no evidence in support of the conjecture.

Acknowledgements

We thank two referees for their suggestions, which led to significant improvements in the paper. MH thanks David Wilson for helpful discussions at the initial stages of this project. MH and YM thank Raghu Varadhan, Federico Camia, and Pierre Nolin for helpful discussions and the Centre for eResearch at U. Auckland for providing the computing resources and support. The work of MH and YM was supported by an FRDF grant from U. Auckland. The work of CMN was supported in part by US NSF grants OISE-0730136 and DMS-1007524.

References

1. V. Beffara. The dimension of the SLE curves. *Ann. Probab.* **36**:1421–1452, (2008).
2. F. Camia and C.M. Newman. Critical percolation exploration path and SLE_6: a proof of convergence. *Probab. Theory Related Fields* **139**:473–519, (2007).
3. D. Chelkak, H. Duminil-Copin, C. Hongler, A. Kemppainen, S. Smirnov. Convergence of Ising interfaces to Schramm's SLE curves *Comptes Rendus Mathematique*, **352**:157-161, (2014).
4. X. Chen. Moderate and small deviations for the ranges of one-dimensional random walks. *J. Theor. Probab.* **19**:721-739, (2006).
5. J.T. Cox and D. Griffeath. Diffusive clustering in the two-dimensional voter model. *Ann. Probab.* **14**:347–370, (1986).
6. D. Griffeath. Additive and Cancellative Interacting Particle Systems. Lecture Notes in Math. 724. *Springer, Berlin* (1979).
7. T.E. Harris. Additive set-valued Markov processes and graphical methods. *Ann. Probab.* **6**: 355–378, (1978).

8. G.F. Lawler, O. Schramm, and W. Werner. Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann. Probab.* **32**:939–995, (2004).
9. G.F. Lawler and V. Limic. Random Walk: A Modern Introduction. *Cambridge University Press*, (2010).
10. P. Nolin. Critical exponents of planar gradient percolation. *Ann. Probab.* **36**:1748–1776, (2008).
11. P. Nolin. SLE(6) and the geometry of diffusion fronts. *arXiv:0912.3770*
12. P. Nolin. Inhomogeneity and universality: off-critical behavior of interfaces. *arXiv:0907.1495*
13. P. Nolin and W. Werner. Asymmetry of near-critical percolation interfaces. *J. Amer. Math. Soc.* **22**:797–819, (2009).
14. O. Schramm. Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J. Math.* **118**:221–288, (2000).
15. O. Schramm and S. Sheffield. Harmonic explorer and its convergence to SLE_4. *Ann. Probab.* **33**:2127–2148, (2005).
16. S. Smirnov. Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *C. R. Acad. Sci. Paris Sér. I Math.* **333**:239-244, (2001).
17. S. Smirnov. Critical percolation in the plane. *arXiv:0909.4499* (2001).
18. S. Smirnov. Towards conformal invariance of 2D lattice models. *Proc. Int. Congr. Math.* **2**:1421-1451, (2006).

4 Appendix

As indicated earlier, each of our point estimates is based on 10000 independent simulations of the voter model for each value of L being considered. Although the simulations had a finite time horizon, in all cases all coalescing walks eventually reached the boundary. All simulations were conducted in C++ and all statistical analyses and plots were performed in R. The data is available on request, but at approximately 600GB, may be difficult to transfer.

Note also that all of our simulations actually took place on boxes of side length $L' = L + 2$, so our estimators were actually

$$\tilde{d} = \tilde{d}_{m,L} = \frac{\log(\bar{Z}_m(L+2))}{\log(L)}, \tag{44}$$

$$\hat{d} = \hat{d}_{m,L} = \log_2\left(\frac{\bar{Z}_m(2L+2)}{\bar{Z}_m(L+2)}\right). \tag{45}$$

Similarly our ordinary least squares estimator d^* is in fact an estimator for the slope coefficient d of the simple linear regression model

$$\log(|Z_i|) = d\log(L_i) + a + \varepsilon_i, \tag{46}$$

where $\{|Z_i|\}_{i\leq n}$ are interface lengths on boxes of side lengths $\{L_i + 2\}_{i\leq n}$, and the ε_i are random variables with mean 0. This does not change the consistency properties of the estimators, and e.g. results in an estimate $\tilde{d}$ differing in only the fourth decimal place when we are dividing by $\log(L)$ (instead of $\log(L+2)$).

4.1 Bootstrap intervals

In an attempt to quantify the variability of our estimators we computed 10000 standard (i.e. each of size 10000, sampled with replacement) bootstrap samples from our 10000 data points, calculated the value of our statistic in each case, and then constructed a "bootstrap 95% confidence interval" by removing the smallest and largest 2.5% of the values (so the interval is given by the min and max of the remaining values). Below we give an example of the R code used to produce the bootstrap intervals. Note that these are not confidence intervals for the true value, but rather measures of the variability of our estimators.

```
#load the standard bootstrapping package into R
library(boot)

#define the mean function for the bootstrap
meantest=function(x,indices){mean(x[indices])}

#read in the data of curve lengths and unlist (turn into a vector)
voterlengths1024=read.table(file.choose())
voterlengths1024=unlist(voterlengths1024)
voterlengths512=read.table(file.choose())
voterlengths512=unlist(voterlengths512)

#get 10000 bootstrap samples of mean curve lengths
voterboot1024=boot(voterlengths1024,meantest,R=10000)$t
voterboot512=boot(voterlengths512,meantest,R=10000)$t

#compute the hat estimators for each bootstrap sample
bootestimates1024_512=log(voterboot1024/voterboot512,2)
summary(bootestimates1024_512)
#       V1
# Min.   :1.448
# 1st Qu.:1.460
# Median :1.462
# Mean   :1.462
# 3rd Qu.:1.465
# Max.   :1.475

#sort the hat estimators
sorted1024_512=sort(bootestimates1024_512)

#delete the bottom and top 2.5% of the values to yield a bootstrap confidence interval
summary(sorted1024_512[251:9750])
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  1.455   1.460   1.462   1.462   1.465   1.470


#####now do the same thing for the dtilde estimates
```

```
dtilde1024=log(voterboot1024)/log(1024)

summary(dtilde1024)
#       V1
# Min.   :1.553
# 1st Qu.:1.553
# Median :1.554
# Mean   :1.554
# 3rd Qu.:1.554
# Max.   :1.555

sorted1024=sort(dtilde1024)
summary(sorted1024[251:9750])
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  1.553   1.553   1.554   1.554   1.554   1.554
```

4.2 The size of coalescing classes

Since the $\hat{\bullet}$ and $\tilde{\bullet}$ estimators are defined straightforwardly and have already been discussed somewhat at the end of Section 3, let us turn our attention here to the regression estimators $\bullet^*$ for the class sizes. Assume that the assumptions prior to (11) hold for $|C_L(o)|$ and $|M_L|$ with exponents γ and β respectively, so that e.g.

$$\log(|M_L|_i) = \beta \log(L) + a + \varepsilon_i. \tag{47}$$

We obtain an estimate β^* of β by fitting the simple linear model $\log(|M_L|) \sim \beta \log(L)$.

Fitting this linear model in R we obtain the following output:

```
Call:  lm(formula = log(largest_class) ~ log(L))

Residuals:
     Min       1Q   Median       3Q      Max
-0.93002 -0.22984 -0.02386  0.20920  1.34982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.726828   0.028139  -61.37   <2e-16 ***
log(L)       1.897190   0.004374  433.72   <2e-16 ***

Residual standard error: 0.3201 on 12998 degrees of freedom
Multiple R-squared: 0.9354,     Adjusted R-squared: 0.9354
F-statistic: 1.881e+05 on 1 and 12998 DF,  p-value: < 2.2e-16
```

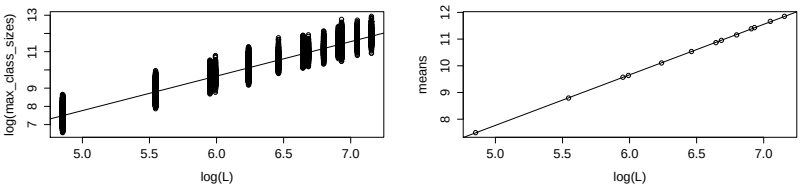

Fig. 5 The fitted line for (47) for the largest coalescing class, with background the raw data and the fixed L means respectively.

Figure 5 and the standard diagnostic tests suggest that the model fits quite well (although there is some skewness in the residuals). Despite this, our estimate $\beta^* = 1.897$ is more than 20 standard errors from the known (from Theorem 31) true value of $\beta = 2$, so this estimator seems to be doing a poor job of estimating the true limiting behaviour in L. However we also know from Theorem 31 that there are in fact log corrections to this model. Therefore we repeated the above analyses but included $\log\log L$ as an additional explanatory variable to get an estimator β^{**}.

```
Call:
lm(formula = log(largest_class) ~ log(Lvals) + log(log(Lvals)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.96107 -0.23202 -0.02224  0.21154  1.34380

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.98223    0.15032  -6.534 6.45e-11 ***
log(Lvals)       2.05204    0.03359  61.094  < 2e-16 ***
log(log(Lvals)) -0.93589    0.19694  -4.752 2.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3202 on 52997 degrees of freedom
Multiple R-squared:  0.9546,Adjusted R-squared:  0.9546
F-statistic: 5.576e+05 on 2 and 52997 DF,  p-value: < 2.2e-16
```

The estimate $\beta^{**} = 2.052$ is within two standard errors of the true value, and the estimate for ζ (the coefficient of $\log\log L$) has the right sign. However, due to the lack of orthogonality in the explanatory variables when $\log\log L$ is included, one cannot put a great deal of faith in these coefficients.

Unsurprisingly the linear regression estimator fits less well for the coalescing class of the center, and gives an estimate of $\gamma^* = 1.84$ (see also Figure 6).

```
Call:
  lm(formula = log(origin_class) ~ log(Lvals))
```

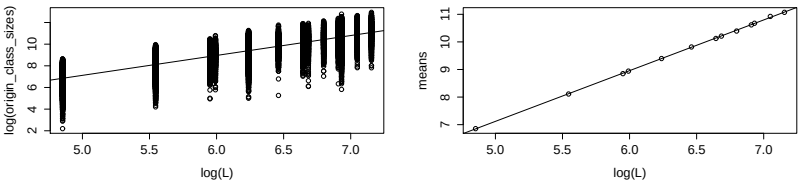

Fig. 6 The fitted line for (47) for the coalescing class of a central vertex, with background the raw data and the fixed L means respectively.

```
Residuals:
  Min      1Q  Median      3Q     Max
-4.8501 -0.4658  0.0884  0.5619  2.1032

Coefficients:
              Estimate  Std. Error t value  Pr(>|t|)
  (Intercept) -2.071629   0.026735  -77.49   <2e-16 ***
  log(Lvals)   1.838248   0.004407  417.12   <2e-16 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7874 on 52998 degrees of freedom
Multiple R-squared:  0.7665,  Adjusted R-squared:  0.7665
F-statistic: 1.74e+05 on 1 and 52998 DF,  p-value: < 2.2e-16
```

Including the extra explanatory variable $\log\log L$ yields an estimate of $\gamma^{**} = 1.97630$ with ζ having the correct sign, but again the estimators suffer due to a lack of orthogonality among the explanatory variables.

```
Call:
  lm(formula = log(origin_class) ~ log(Lvals) + log(log(Lvals)))

Residuals:
  Min      1Q  Median      3Q     Max
-4.8546 -0.4657  0.0883  0.5620  2.0986

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
  (Intercept)    -1.45452    0.36963  -3.935 8.33e-05 ***
  log(Lvals)      1.97630    0.08259  23.929  < 2e-16 ***
  log(log(Lvals)) -0.81064    0.48427  -1.674   0.0942 .
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7873 on 52997 degrees of freedom
Multiple R-squared:  0.7665,  Adjusted R-squared:  0.7665
F-statistic: 8.7e+04 on 2 and 52997 DF,  p-value: < 2.2e-16
```

In what follows, when we consider a linear regression estimator, we use $\bullet^{**}$ when the added explanatory variable has statistically significant (at the 10% level) explanatory power, and $\bullet^{*}$ otherwise.

4.3 Connected clusters of a coalescing class

Recall from the end of Section 3 that for each configuration of a voter model in B_L, $C_L^c(x)$ denotes the connected subset (containing x) of the coalescing class of x. Let M_L^c denote a largest such connected subset and assume

$$\mathbb{E}\left[|C_L^c(x)|\right] \approx L^{\gamma'} \text{ and } \mathbb{E}\left[|M_L^c|\right] \approx L^{\beta'}.$$

Recalling the discussion around (9), we obtain the following $\hat{\bullet}$ estimates, which indicate that this estimator converges more slowly with increasing L than the corresponding estimator for the expected curve length exponent.

L	128	256	512
$\hat{\beta'}_L$	1.693 (1.673,1.712)	1.718 (1.697,1.739)	1.741 (1.719,1.762)
$\hat{\gamma'}_L$	1.470 (1.416,1.526)	1.492 (1.428,1.554)	1.548 (1.477,1.618)

The linear regression estimates are quite far from those above. In particular $\gamma'^{*} \approx 1.02$ is very far off the $\hat{\gamma}$ estimates above and $\beta'^{**} \approx 1.89$ (with the coefficient of $\log\log L$ carrying a $-$ sign.

4.4 Maximum displacement of the curve from the diagonal

Recall the last paragraph of Section 3. Assuming that $\mathbb{E}\left[\max_{x\in Z_L} \min_{y\in D} |x-y|\right] \approx L^{\alpha}$ we have the following $\hat{\bullet}$ estimates for α.

L	128	256	512
$\hat{\alpha}_L$	0.964 (.9542,.9734)	0.973 (.9634,.9828)	0.969 (.9598,.9792)

Fitting a simple linear model to the data gives a very similar estimate of $\alpha^{*} \approx 0.969$ (approximately 18 standard errors from 1).

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.11443 -0.17013  0.01359  0.18474  0.59812

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.090206   0.009573  -113.9   <2e-16 ***
log(Lvec)    0.968864   0.001611   601.4   <2e-16 ***
```

```
Residual standard error: 0.2497 on 39998 degrees of freedom
Multiple R-squared: 0.9004,     Adjusted R-squared: 0.9004
F-statistic: 3.617e+05 on 1 and 39998 DF,  p-value: < 2.2e-16
```