Percolation under noise: Detecting explosive percolation using the second-largest component

Wes Viles, Cedric E. Ginestet, Ariana Tang, Mark A. Kramer, and Eric D. Kolaczyk

Phys. Rev. E **93**, 052301 — Published 2 May 2016

DOI: 10.1103/PhysRevE.93.052301

Percolation under Noise: Detecting Explosive Percolation Using the Second Largest Component

Wes Viles,[1] Cedric E. Ginestet,[2] Ariana Tang,[1] Mark A. Kramer,[1] and Eric D. Kolaczyk[1]

[1] *Department of Mathematics and Statistics, Boston University*
[2] *Department of Biostatistics, Institute of Psychiatry, Psychology and Neuroscience, King's College London*

We consider the problem of distinguishing between different rates of percolation under noise. A statistical model of percolation is constructed allowing for the birth and death of edges as well as the presence of noise in the observations. This graph-valued stochastic process is composed of a latent and an observed non-stationary process, where the observed graph process is corrupted by Type I and Type II errors. This produces a hidden Markov graph model. We show that for certain choices of parameters controlling the noise, the classical (ER) percolation is visually indistinguishable from a more rapid form of percolation. In this setting, we compare two different criteria for discriminating between these two percolation models, based on the interquartile range (IQR) of the first component's size, and on the maximal size of the second-largest component (SLC). We show through data simulations that this second criterion outperforms the IQR of the first component's size, in terms of discriminatory power. The maximal size of the second component therefore provides a useful statistic for distinguishing between different rates of percolation, under physically motivated conditions for the birth and death of edges, and under noise. The potential application of the proposed criteria for the detection of clinically-relevant percolation in the context of applied neuroscience is also discussed.

I. INTRODUCTION

Understanding the emergence of organized structure in dynamic networks remains an active research area [1, 2]. In the study of random networks, percolation –the sudden emergence of a giant connected component (GCC)– is of critical importance from a theoretical, applied and statistical perspective. Percolation in the Erdős-Rényi (ER) model constitutes one of the first examples of a fully characterized mathematical phase transition [3, 4]. While the ER model of percolation is an example of a (second order) continuous phase transition, recent efforts have focused on identifying the conditions under which a random network process can yield a (first order) discontinuous percolation [5].

One of the most popular attempts to model discontinuous percolation has been the Achlioptas' process and its variants [6]. The Achlioptas' product rule (PR) slows down the growth of the GCC by favoring the creation of edges between small connected components. Although this particular percolation model has been shown to be, in fact, continuous and therefore of second order [5, 7, 8]; it nonetheless provides an interesting alternative to the ER model [9]. Achlioptas' processes have indeed generated a substantial amount of theoretical work, whereby authors have explored related strategies for producing explosive percolation in random networks [10–12]. In addition, Riordan and Warnke [5] have shown that genuine first order phase transitions can be realized by systematically adding, at every step of the process, the edge that joins the two smallest components in the *entire* network. In this manuscript, we use the phrase "explosive percolation" to refer to the sudden emergence of a large connected component in the Achlioptas process. However, we note that explosive percolation in this context is rapid, but continuous.

Interest in network percolation has been fueled by its relevance to several application domains. In clinical neuroscience, for instance, epileptic seizures have been associated with the sudden emergence of coupled activity across the brain [13–18]. The resulting functional networks –in which edges indicate strong enough coupling between brain regions [19]– are consistent with the notion of percolation. A better understanding of the type of phase transitions undergone at different stages of the seizure, may aid in the development of novel strategies for the treatment of epilepsy [20]. In this paper, we have therefore concentrated our attention on relatively small networks, with sizes ranging between 100 and 1,000 vertices. Such a number of vertices is comparable to the size of the networks often studied in neuroscience [17, 21–25]. However, note that percolation on such small networks can suffer from finite-size effects [26].

The rich theory on percolation, and its application to real world data, motivates the following question: How can we distinguish between different percolation regimes in practice? Previous theoretical work has concentrated on noise-free percolation, which constitutes an idealized perspective on percolation processes. In practice, however, the sampling of real-world networks is likely to be corrupted by measurement errors. Moreover, network growth has generally been conceived as a monotonic process, whereby only edge creations are allowed. However, this assumption may be too restrictive, since in real-world networks, the number of edges may increase and decrease over time, in a stochastic manner (see example in figure 1). Finally, to the best of our knowledge, there does not currently exist a statistical framework for distinguishing between different types of percolation regimes in the presence of edge birth and edge death, as well as noise.

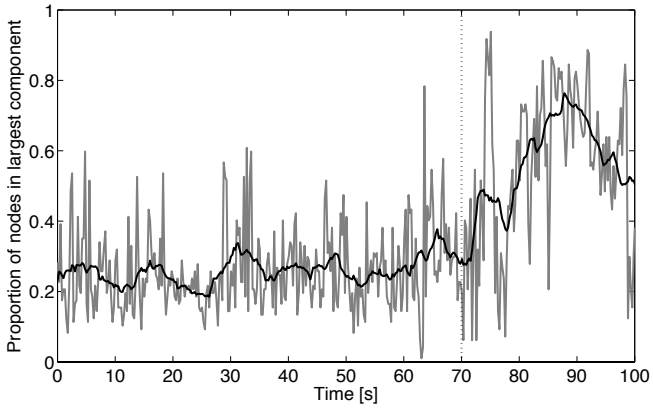


FIG. 1. Proportion of nodes in the largest component as a function of time for a functional network deduced from the electrocorticogram of a single patient with epilepsy during a seizure [see 27, for details]. The weighted functional networks have been binarized by conducting independent hypothesis tests on the maximum absolute values of the cross-correlations over $0.5s$ windows with 50% overlap and after correcting for multiple comparisons. The black trace is a smoothed version of this process. Seizure onset was clinically estimated to occur at the vertical dotted line.

In this paper, we propose a framework to distinguish between different percolation regimes in practice. To do so, we formulate the problem of recognizing a percolation regime from noisy observations as a question of statistical inference. Under this framework, we compare the discriminatory power of two potential percolation features deduced from the evolution of the first and second component of an observed dynamic network. We test this framework in simulation by constructing a hidden Markov graph model, which encompasses both a non-stationary latent process characterized by birth and death of edges, and an observed graph process that introduces both Type I and Type II errors. We show that edge death and noise renders the statistic deduced from the first component ineffective in distinguishing between the standard ER second-order percolation and Achlioptas' explosive percolation. However, a different detection criterion –based on the size of the second component– successfully discriminates between the two percolation regimes in the presence of edge death and noise. These results provide a framework for distinguishing percolation regimes in practice.

II. PERCOLATION MODELS

A. Birth and Death Erdős-Rényi (ER) Process

We first construct a graph-valued stochastic process that exhibits the Markov property. This provides a realistic model for generating noisy percolation processes, while maintaining a sufficient level of computational tractability. We will denote a sequence of graph-valued random variables on n vertices by

$$\big\{G_t = (V, E_t) : t = 0, \ldots, T\big\}, \tag{1}$$

where V denotes the vertex set of G_t, whereas E_t represents the edge set of G at time t. Observe that the vertex set does not vary with time. At each time step, a single edge is either added or deleted. Such a sequence will be said to be *Markov* if its edge sets are controlled by a Markov chain. We impose this dependence through the use of a binary random variable, denoted $\{Y_t : t = 0, \ldots, T\}$, whose state space is $\{0, 1\}$. This Markov chain is characterized by the following transition probability matrix P, for some choices of the birth and death rates, denoted respectively by p and q, and taking values in $[0, 1]$.

	$Y_{t+1} = 0$	$Y_{t+1} = 1$
$Y_t = 0$	$1 - p$	p
$Y_t = 1$	q	$1 - q$

Following customary notation, the entries of P will be denoted by $\mathbb{P}[Y_{t+1} = j | Y_t = i]$, with rows summing to one. The graph-valued Markov chain, G_t, is then obtained by associating Y_t with the addition or deletion of an edge in each edge set, E_t. Thus, provided that $p, q \neq 0$, it follows that the state space of this graph-valued Markov chain is the space of all simple graphs on n vertices, since every graph is reachable with positive probability.

In general, p and q are not required to sum to one. It will be of interest to let $p > q$ in order to study the large-scale behavior of the G_t's as the graph process accumulates edges. Moreover, observe that Y_t is a (time) *homogeneous* Markov chain, since $\mathbb{P}[Y_{t+1} = \omega | Y_t = \omega'] = \mathbb{P}[Y_1 = \omega | Y_0 = \omega']$, for any $\omega, \omega' \in \{0, 1\}$ and every t.

Now, suppose that there exist $m_t := |E_t|$ edges at time t in G_t and let $X_t(e)$ denotes the 'status' of edge e at time t, such that $X_t(e) = 1$, if that edge is present; and $X_t(e) = 0$, otherwise. Note that we have here two different sources of dependence. On one hand, the edges are dependent on each other, since no more than one edge can be added or deleted at every time step. On the other hand, the edges are also dependent over time, since the status of an edge at time $t + 1$ depends on the status of that same edge at time t.

In the sequel, we will concentrate on a special case of this birth and death process, where we will set $p = 1 - q$. This leads to simplified marginal distributions for the edges. Additional details of this birth and death model are provided in appendix A.

B. Birth and Death Product Rule (PR) Process

We extend the standard Achlioptas' framework of PR percolation to a birth and death process, by devising death steps. This model is analogous to the aforementioned ER birth and death model, except for the choice of the probability distribution of the latent $X_t(e)$'s. As for

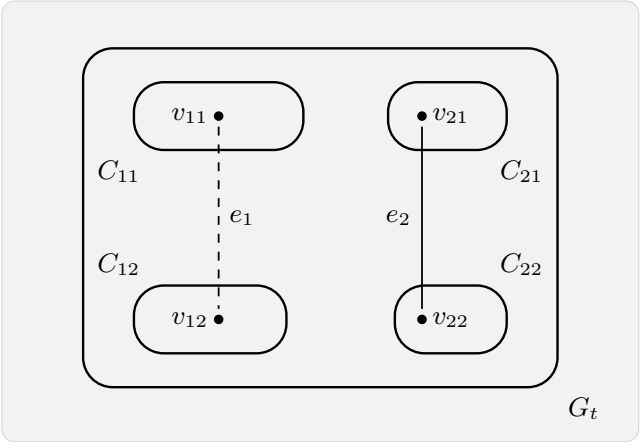


FIG. 2. Birth and death steps for the product rule (PR) in the Achlioptas' model of percolation. In this example, edge e_2 was born *before* edge e_1, since $|C_{11}||C_{12}| > |C_{21}||C_{22}|$. Therefore, when e_1 and e_2 are selected during a death step, e_2 is discarded *after* e_1. This death step specification ensures that births and deaths constitute genuine reverse PR operations.

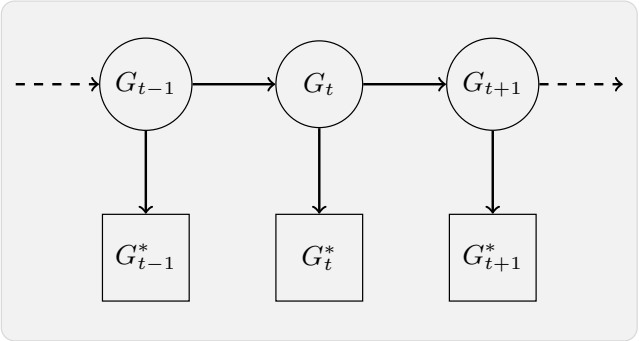


FIG. 3. Directed Acyclic Graph (DAG) representation of the hidden Markov process combining a *latent* stochastic graph process in the first row denoted by G_t, with an *observed* stochastic graph process contaminated by noise in the second row, denoted by G^*_t. Directed arrows indicate probabilistic dependence, such that the distribution of the observed G^*_t depends on the value taken by the latent graph, G_t.

the ER birth and death model, a binary random variable, Y_t, controls the addition or deletion of edges in each G_t. However, in the case of the PR model, the choice of the edge to be added or to be deleted is not uniform over the E_t's. Here, this choice depends on the modular structure of the graph at time t. Therefore, as for the ER model, we obtain a non-stationary stochastic process.

Assuming that $Y_t = 1$, the addition of a new edge is conducted by uniformly choosing two candidate vertex pairs among all the edges in E_t^C, the complement of the edge set, E_t. These two candidate edge pairs are denoted by $e_1 := (v_{11}, v_{12})$ and $e_2 := (v_{21}, v_{22})$, and satisfy $X_t(e_1) = 0$ and $X_t(e_2) = 0$, since $e_1, e_2 \in E_t^C$, as in figure 2. We then evaluate the size of the connected components to which v_{11}, v_{12}, v_{21} and v_{22} belong. These four connected components are denoted by C_{11}, C_{12}, C_{21}, and C_{22}, respectively. Then, following Achlioptas et al. [6], we apply the following product rule: If $|C_{11}||C_{12}| < |C_{21}||C_{22}|$, then $X_{t+1}(e_1) = 1$; otherwise, $X_{t+1}(e_2) = 1$.

Conversely, the death or deletion of an edge is handled in a symmetric manner. When $Y_t = 0$, we uniformly select two candidate edges from E_t. These vertex pairs are denoted $e_1 := (v_{11}, v_{12})$ and $e_2 := (v_{21}, v_{22})$ and satisfy $X_t(e_1) = 1$ and $X_t(e_2) = 1$, since $e_1, e_2 \in E_t$. Next, we set $X_t(e_1) = 0$ and $X_t(e_2) = 0$, in order to compute the size of the connected components to which v_{11}, v_{12}, v_{21} and v_{22} would belong to, if these edges were absent. This is done in order to ensure that the deletion of an edge exactly corresponds to the reverse operation of the addition of an edge under PR. Next, after having deleted these edges and computed the sizes of C_{11}, C_{12}, C_{21}, and C_{22}; we decide which edge should re-enter G_t, in order to produce G_{t+1}. Such a decision is also based on the PR, such that if $|C_{11}||C_{12}| < |C_{21}||C_{22}|$, then $X_{t+1}(e_2) = 0$; otherwise, $X_{t+1}(e_1) = 0$.

This choice of specification for the death step ensures that the *ordering* of the creation and deletion of edges are symmetrical. Given a sequence of two edges $\{e_1, e_2\}$ successively born during two time steps of G_t, if we encounter a death step, where both e_1 and e_2 are selected, we would then delete these edges in the reverse order, by eliminating e_2 before e_1. This order-preserving property is illustrated in figure 2. This constraint ensures that births and deaths are genuine reverse PR operations. In addition, observe that, as for the ER percolation process, this chain is *irreducible*, in the sense that there is positive probability of transitioning from any given edge configuration to any other in the space of the edge sets of G.

C. Hidden Markov Graph Model

Next, we assume that there exists a time-independent error process, which produces at each time point an *observed* edge status $X^*_t(e)$. This stochastic process is governed by two additional parameters α and β, whose behavior can be described using a traditional 'confusion matrix', such that for any $\alpha, \beta \in [0, 1]$, we have

	$X^*_t(e) = 0$	$X^*_t(e) = 1$
$X_t(e) = 0$	$1 - \alpha$	α
$X_t(e) = 1$	β	$1 - \beta$

The $X_t(e)$'s and $X^*_t(e)$'s are here treated as *latent* and *observed* stochastic processes, respectively, and α and β can therefore be interpreted as the Type I (false positive) and Type II (false negative) error probabilities. Combining the graph-valued Markov latent process with this time-independent error process, we obtain a graph-valued *hidden Markov process*, as described in figure 3. From

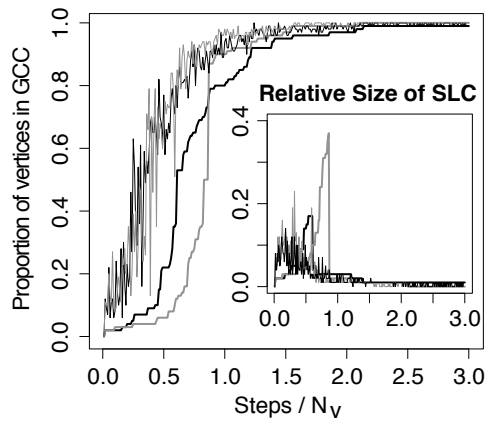


FIG. 4. Percentage of vertices in the giant connected component (GCC) in the main window, and the corresponding sizes of the second-largest component (SLC) in inset, for a birth rate of $p = 1$ and death rate of $q = 0$, for the ER (black) and PR (gray) percolation models, for $N_v = 100$; in which the x-axis denotes the number of steps in the process divided by N_v. In bold, these results are reported for a noise-free model, whereas the thin lines represent noisy simulations with Type I/II error rates of $\alpha = .0075$, $\beta = 0$. The details of the algebraic relationship between the Type I/II error rates and the graph process is described in appendix A, for the case of the ER model. Observe that the ER and PR models are nearly indistinguishable once a small amount of noise is added to these percolation processes.

this schematic representation, one can immediately see that the observed graphs denoted G^*_{t-1}, G^*_t and G^*_{t+1} are conditionally independent, given the latent graph process, G_t.

For the ER model, under the assumption that $p = 1-q$, these two stochastic graph processes can be combined by taking into account the time-dependence of the $X^*_t(e)$'s. In this case, the corresponding transition matrix linking the observed and latent processes is available in closed-form. Details of these derivations are provided in appendix A.

III. DETECTING EXPLOSIVE PERCOLATION

Explosive percolation is expected to produce a sharper phase transition than a typical ER percolation. When considering noisy observations, however, detecting such differences through visual inspection only, is hard. Figure 4 illustrates this problem, by comparing noisy and noise-free graph sequences for both explosive and ER models. Beyond visual inspection, the problem of discriminating between these two models of percolation can be formulated as a hypothesis-testing problem: The null hypothesis, denoted H_0, states that the observed process corresponds to an ER percolation, whereas the alternative hypothesis, H_1, is that the observed process does not correspond to this type of percolation model.

To proceed with this hypothesis-testing problem in practice, we specify a population parameter summarizing the percolation process, say θ^{ER} and θ, for the ER and target models, respectively. This leads to a hypothesis test of the form,

$$H_0 : \theta^{\text{ER}} = \theta, \qquad \text{and} \qquad H_1 : \theta^{\text{ER}} \neq \theta.$$

Several population parameters could be used for the purpose of discriminating between these two models of percolation. A natural candidate for such parameters would be a measure of the sharpness of the transition of the first component's size. The main panel of figure 4, however, suggests that this population parameter will not have sufficient discriminatory power, when confronted with a substantial amount of observational noise.

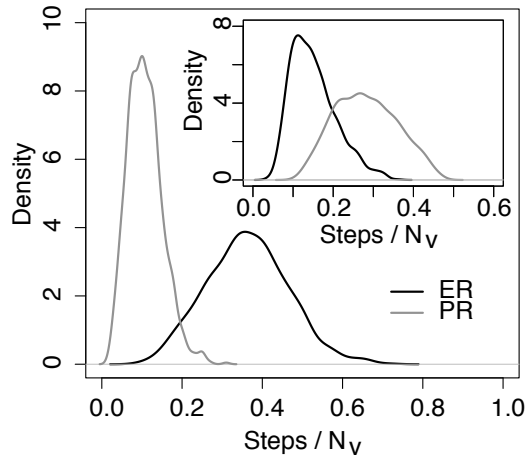
Therefore, as a second candidate population parameter, we consider the following natural extension: The size of the *second-largest* component (SLC). This provides a more sensitive marker of the sharp phase transition exhibited by explosive percolation models [28]. Several authors have considered the size of the second largest component as a useful marker. Margolina et al. [28] have investigated the ratio of the sizes of the first and second component in cubic and triangular lattices. They have shown that the number of nodes in the second largest component reaches a maximum at the percolation threshold. More recently, and also in the context of cubic and hypercubic lattices, da Silva et al. [29] have studied the scale invariance of the ratio of the sizes of the first and second components. The size of the second largest component has also been used in neuroimaging, in order to identify the percolation threshold [30], as well as in an effort to detect and prevent epileptic seizures [31]. However, to the best of our knowledge, it has not been previously used in a statistical context for the purpose of discriminating between different types of percolation regimes. In what follows, we will show that the use of the size of the SLC as a statistical marker to distinguish the two percolation regimes exhibits greater discriminatory power, than a statistic solely based on the first component.

The differences between the candidate percolation models are therefore quantified using two criteria: (i) the interquartile range (IQR) of the distribution of the size of the GCC, and (ii) the maximal size of the SLC over the entire time period. These two criteria are formally defined as follows. Given the graph process, $G_t = (V, E_t)$, and denoting the vertex subset of the largest component in G_t by $S_{1,t}$, we define the cumulative edge function as the *cardinality* of $S_{1,t}$, normalized by the maximal number of edges in the graph, such that

$$F(t) := \frac{|S_{1,t}|}{\binom{n}{2}}.$$

Although this function is not a cumulative distribution function (CDF), one can nonetheless uniquely define

(a)



(b)

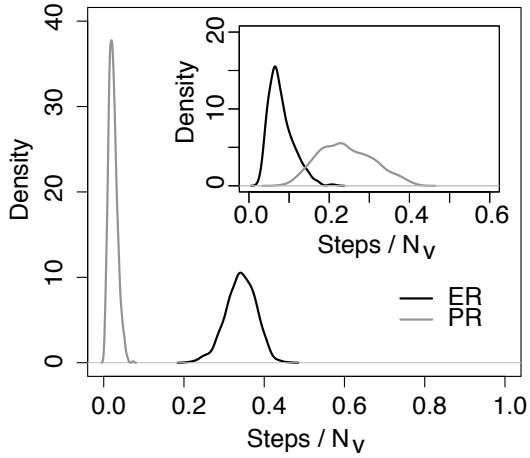


FIG. 5. Densities of the IQR of the GCC in the main windows, and densities of the maximal size of the SLC in the insets, are reported for networks of sizes 100 and 1,000 vertices in panels (a) and (b), respectively; for the ER (black) and PR (gray) percolation models, and based on 1,000 realizations from the distributions of these two (noise-free) models, specifying a birth rate of $p = 1$ and death rate of $q = 0$; in which the x-axis denotes the number of steps in the process divided by N_v. The IQR criterion has been described in equation (2). Note that the difference in scales of the density values of the y-axes in these two figures is due to the difference in scales of the x-axes, which represent the number of steps scaled by networks' sizes, t/N_v.

quantiles using the standard definition of quantiles for the CDFs of discrete random variables; such that for any $x \in [0, 1]$, we have

$$Q(x) := \min_{t=1,\ldots,T} \{t : F(t) \geq x\},$$

where T is the maximal number of time steps in the graph process. In this paper, we are especially interested in the classical interquartile range,

$$\text{IQR} := Q(0.75) - Q(0.25). \tag{2}$$

This parameter quantifies the steepness of the phase transition: the larger the IQR criterion, the longer the transition to a fully connected graph.

As a second criterion to distinguish the two percolation regimes, we consider the maximal size attained by the SLC over the entire time period of the dynamic network observation. If one defines the vertex set of the SLC at time t by $S_{2,t}$, this second criterion can be expressed as

$$\theta_{\text{SLC}} = \max \{|S_{2,t}| : t = 1, \ldots, T\}.$$

This quantity is known to constitute a good marker of the steepness of the phase transition, since it reflects the extent of separation of the graph process into large connected subgraphs [28, 31]. Indeed, a direct consequence of the Achlioptas' construction rule is that by inhibiting the growth of a single large component, we necessarily increase the production of several subcomponents.

Statistical inference on these two criteria is then drawn using a Monte Carlo hypothesis test. Letting the parameter $\theta := \theta_{\text{IQR}}$, and selecting the candidate percolation to be drawn from a PR process, we consider the following null and alternative hypotheses,

$$H_0 : \theta^{\text{ER}} = \theta^{\text{PR}}, \qquad \text{and} \qquad H_1 : \theta^{\text{ER}} > \theta^{\text{PR}},$$

respectively. The direction of this test is justified by the fact that we expect explosive percolation to occur rapidly, and thus to exhibit a smaller amount of variability in the size of its GCC, when transitioning to a fully connected graph. Our second criterion, by contrast, is tested in the opposite direction, since we naturally anticipate the PR process to be characterized by a *larger* maximal SLC. Thus, for $\theta := \theta_{\text{SLC}}$, the alternative hypothesis becomes $H_1 : \theta^{\text{ER}} < \theta^{\text{PR}}$.

In the results reported in this paper, the distributions of the ER and PR graph processes are known. It therefore suffices to simulate from these densities in order to construct the distribution of the two test statistics at hand. This procedure is illustrated in figure 5. We are especially interested in the discriminatory powers of these statistics, and we will therefore compare their respective merits, using the true positive and false positive rates, within a receiver operating characteristic (ROC) framework. The ROC curve illustrates the performance of a binary classifier by plotting the false positive rate against the true positive rate, as the discriminating threshold varies. For presentational convenience, the distributions of interest were smoothed using a normal density kernel, before computing the ROC curves and corresponding *areas under the curves* (AUCs). The computation of the AUCs allows us to summarize the differences between the models over the entire time period. (See figures 9 to 13.)

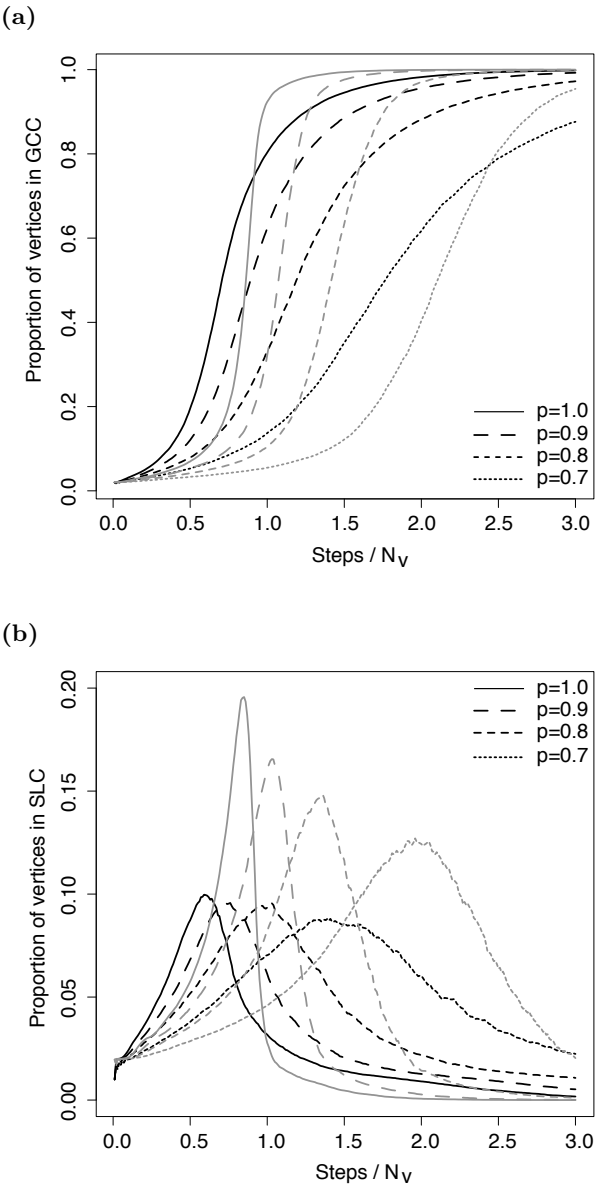


FIG. 6. Percentage of vertices in giant connected component (GCC), and in second largest component (SLC), in panels (a) and (b), respectively; with respect to scaled time (t/N_v), with $N_v = 100$; in which the x-axis denotes the number of steps in the process divided by N_v. These results are reported both the ER and PR models, in black and gray, respectively; for different choices of birth rates, p. Each curve represents the mean of 1,000 independent simulations.

IV. RESULTS

A. Birth and Death Processes

The ER and PR percolation models were simulated on graphs with $n = 100$ and $n = 1,000$ vertices. We first explored the effect of varying the birth and death rates (i.e. p) on the behavior of the two statistical criteria of interest, under the ER and PR models. The results of these Monte Carlo simulations are reported in figure 6, where each curve is the mean of 1,000 independent synthetic data sets. In these simulations, the death rate is set to $q = 1 - p$, and therefore the value of p controls both the birth and death rates.

The main effect of a change in p is to delay percolation, and to diminish the steepness of the phase transition. Observe that as p decreases, percolation tends to occur at a later time step in both the ER and PR models (see figure 6). In particular, for the lowest birth rate that we investigated ($p = 0.7$), the ER model did not produce a fully connected graph within the number of iterations considered, as can be seen from figure 6(a). When p was set to values equal to or less than 0.5, no phase transition could be observed, and these results are not reported.

The size of the second component was similarly affected by changes in p. Decreasing the birth rate delayed the time at which the size of the SLC attained its highest value. Moreover, lower values of p also yielded SLCs with smaller maximal sizes, under both the ER and PR models. Interestingly, we note that the time points at which the SLC reaches a maximal size tended to coincide in both models. Thus, it would be difficult to distinguish between these two percolation models on the sole basis of the *timing* of the occurrence of the maximal size of the SLCs. By contrast, the relative maximal *size* of the SLCs in the ER and PR models differ by approximately one order of magnitude, thereby providing a natural criterion for discriminating between these two types of percolation, as can be observed by comparing figures 6(a) and 6(b).

We formally quantified these differences in discriminatory powers by studying the ROC curves of these two criteria under different choices of p (see figure 11). The maximal size of the SLC substantially outperforms the relative size of the first component, for all values of p. The stark difference between these discriminatory criteria can be understood by considering the amount of overlap of the distributions of these two criteria in figure 5. Whereas the distributions of the IQR of the size of the GCC under the two models exhibit a large amount of overlap; the distributions of the maximal sizes of the SLC, by contrast, share very little common support. These differences in support account for the substantive gains in discriminatory power by the maximal size of the second component, as reported in figure 13.

In addition, we note that the IQR of the size of the first component was more sensitive to choices of p than the maximal size of the SLC. As p diminishes, it becomes increasingly more difficult to discriminate between the ER and PR percolation models, using the IQR of the size of the first component. This suggests that this criterion is more sensitive to a non-zero death rate, than the maximal size of the SLC, which provides further support for the use of this latter criterion, in practice.

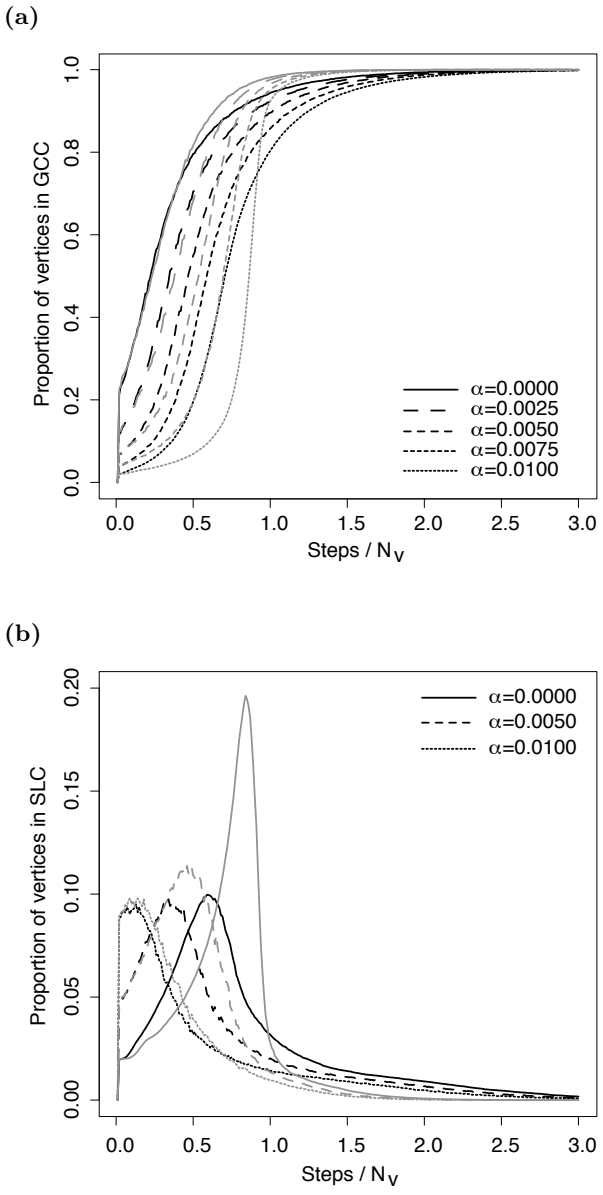


FIG. 7. Percentage of vertices in giant connected component (GCC), and in second largest component (SLC), in panels (a) and (b), respectively; with respect to scaled time (t/N_v), for $N_v = 100$; in which the x-axis denotes the number of steps in the process divided by N_v. These results are reported for both the ER and PR models, in black and gray, respectively; and for different choices of the error rate, α. Each curve represents the mean of 1,000 independent simulations.

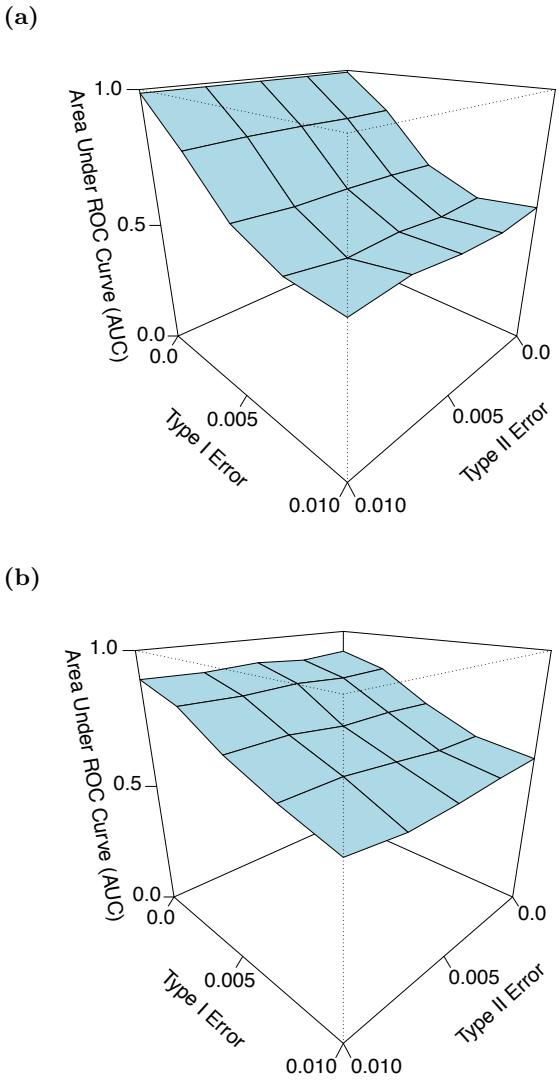


FIG. 8. Area under the Curve (AUC) surface plot for the Receiver Operating Characteristic (ROC) curves of the IQR of the GCC, and of the maximal size of the SLC, for different choices of the error rates α and β, and with $p = 1$ and $q = 0$; under the simulation settings of figure 7, and with $N_v = 100$.

B. Percolation under Noise

Secondly, we considered the effect of introducing noise in these models. The results reported in figure 8 were produced using our proposed hidden Markov graph model, and were averaged over 1,000 simulations. We were especially interested in the effect of Type I and Type II errors on our ability to discriminate between classical and explosive percolation, using the two criteria under scrutiny. Both the Type I and Type II error rates were made to vary between 0 and .01.

From figure 8, one can observe that the two types of errors had markedly different effects on the AUCs of the two discriminatory criteria. Introducing Type I errors led to a substantial diminution of the AUCs for both the IQR of the size of the first component in 8(a), and the maximal size of the SLC in 8(b). In particular, note that the two criteria reached equivalent levels of discriminatory power for $\alpha = 0.01$. Thus, although the maximal size of the SLC remains a more useful criterion for distinguishing between the ER and PR models than the IQR of the size of the first component, these two criteria exhibited comparable performance, under a moderate amount of

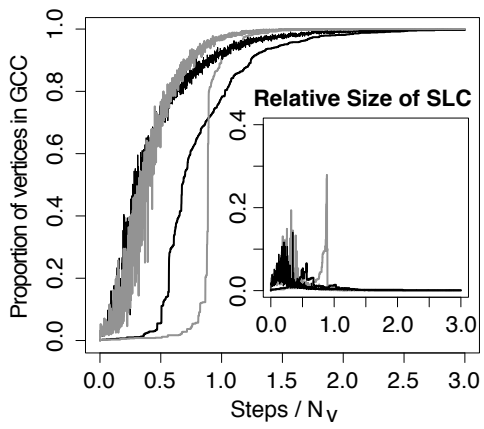


FIG. 9. Percentage of vertices in the giant connected component (GCC) in the main window, and the corresponding sizes of the second-largest component (SLC) in inset, for a birth rate of $p = 1$ and death rate of $q = 0$, for the ER (black) and the PR (gray) percolation models, for $N_v = 1000$; in which the x-axis denotes the number of steps in the process divided by N_v. In bold, these results are reported for a noise-free model, whereas the thin lines represent noisy simulations with Type I/II error rates of $\alpha = .0075$, $\beta = 0$. Observe that the ER and PR models are nearly indistinguishable once noise is added to these percolation processes.

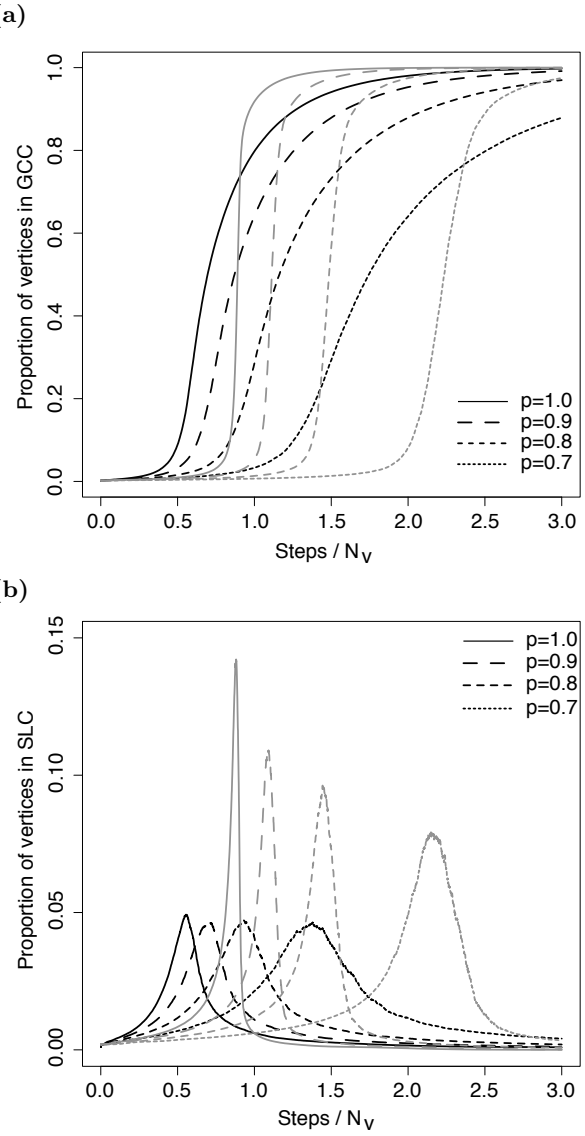


FIG. 10. Percentage of vertices in giant connected component (GCC), and in second largest component (SLC), in panels (a) and (b), respectively; with respect to scaled time (t/N_v), for $N_v = 1000$; in which the x-axis denotes the number of steps in the process divided by N_v. These results are reported for both the ER and PR models, in black and gray, respectively; and for different choices of the birth rate, p. Each curve represents the mean of 1,000 independent simulations.

Type I error.

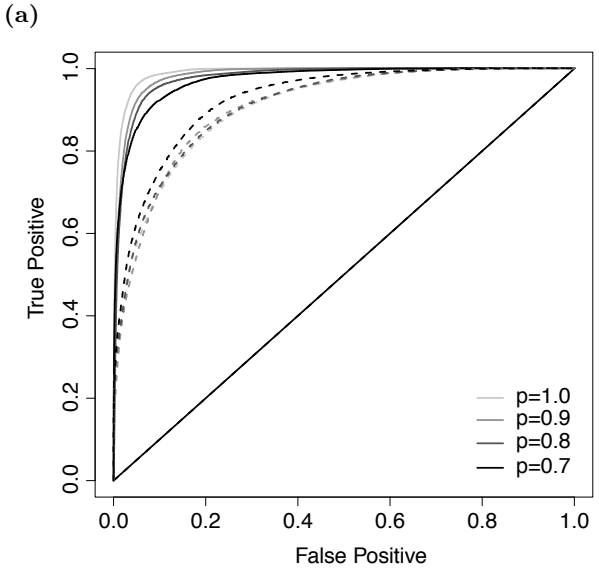
The impact of increasing the Type II error rate on the behavior of these two criteria was negligible. Introducing false negatives in the ER and PR models slightly increased the AUCs of both the IQR of the first component's size, and the maximal size of the second component. Thus, large Type II error rates may be marginally advantageous for discriminating between these two models of percolation, under the scenarios studied.

V. DISCUSSION AND CONCLUSIONS

In this paper, we have extended existing models of percolation by allowing for edge deletion steps and noisy observations. These modeling extensions have been articulated within a hidden Markov graph process, which builds links with the existing literature on the statistical properties of this family of models [32–34]. Moreover, we have compared different summary statistics for distinguishing between the ER and PR percolation models. Overall, for different birth and death rates, and for a range of noise levels, the maximal size of the SLC was found to have greater discriminatory power than the IQR of the size of the GCC.

Several methodological challenges remain before such models can be directly used for percolation detection on real-world data. Throughout this paper, we have considered the IQR of the size of the first component, using a particular choice of quantiles for this discriminatory criterion. In practice, an optimal choice of quantiles for quantifying the steepness of such phase transitions may be motivated by different factors, including (i) the range of the observations, and (ii) the need for early detection. We discuss these two practical aspects, in turn.

Firstly, note that when considering real-world applications, we rarely observe fully connected networks. In the data reported in figure 1, for instance, the size of the GCC encompasses at most 90% of the edges in the saturated network. The choice of the quantile interval



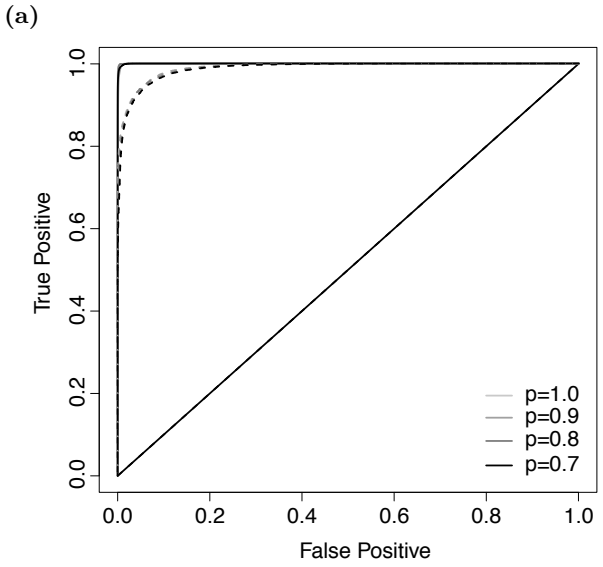
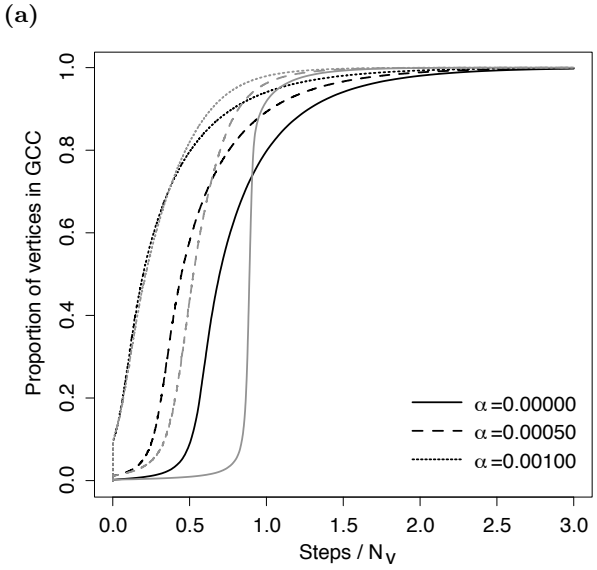


FIG. 11. Receiver Operating Characteristic (ROC) curves for tests based on the IQR and SLC statistics for networks of size $N_v = 100$ and $N_v = 1000$, in panels (a) and (b), respectively; for different choices of birth rate, p.



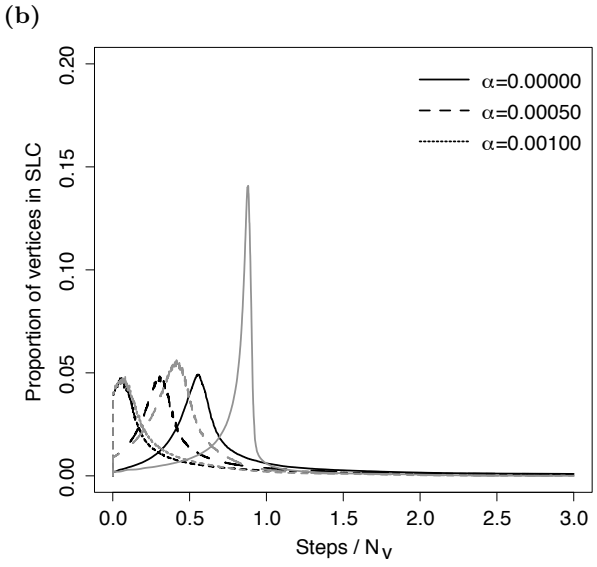


FIG. 12. Percentage of vertices in giant connected component (GCC), and in second largest component (SLC), in panels (a) and (b), respectively; with respect to scaled time (t/N_v), for $N_v = 1000$; in which the x-axis denotes the number of steps in the process divided by N_v. These results are reported for both the ER and PR models, in black and gray, respectively; and for different choices of the error rate, α. Each curve represents the mean of 1,000 independent simulations.

of interest for the first component will be therefore automatically constrained by the range of the observations in the data at hand. Therefore, as in sequential detection analysis, the statistical objective is to detect the outcome, on the basis of as little data as possible. Such constraints would naturally lead to a relatively narrow quantile range.

Secondly, in the context of clinical neuroscience and with particular emphasis on the prevention of a seizure; the detection of a percolation regime may be linked with patients' health and survival. In such cases, early detection will usually be favored, as this is likely to be associated with desirable clinical outcomes. Explosive percolation, such as the Achlioptas' PR process studied in this paper, is consistent with the sudden manifestation of a seizure as a highly synchronized event. Classifying models of percolation may then be utilized to deepen our understanding of seizures, and to gain a greater understanding of the mechanisms underlying epilepsy.

Further work in this area could be focused on estimating a percolation model from a given sequence of observed networks. In this sense, this work also contributes to the growing literature on time-indexed graph processes [35]. In such cases, the birth and death rates will need to be estimated, as well as the Type I and Type II error probabilities. These different parameters may not be fully identifiable from the data, and further assumptions are likely to be necessary, in order to discriminate

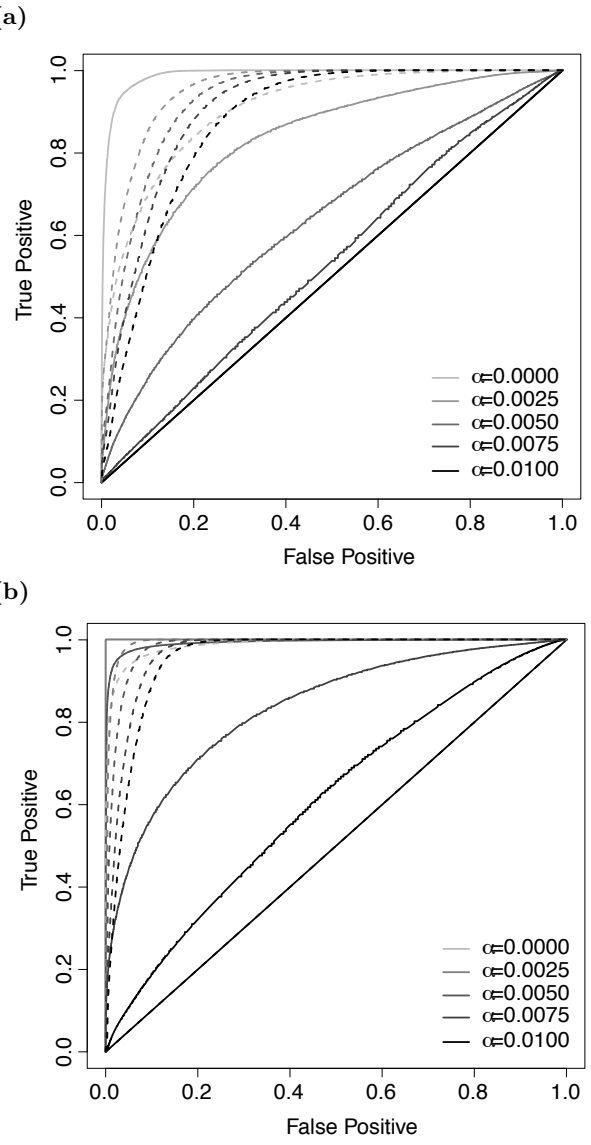


FIG. 13. Receiver Operating Characteristic (ROC) curves for tests based on the IQR and SLC statistics for networks of size $N_v = 100$ and $N_v = 1000$, in panels (a) and (b), respectively; for different choices of error rate, α.

between the two percolation models considered in this paper. Such estimation, however, may be amenable to a Bayesian formulation, as commonly implemented for hidden Markov models [36]. Note also that this work may be extended to some of the recently proposed generalized versions of the Achlioptas model [37–39].

Appendix A: Details of Birth and Death ER Process

Here, we describe the closed-form formulas of the probabilities of edge inclusion and edge deletion in the observed graph processes under the ER model. These analytic results are obtained by assuming that the birth and death probabilities are straightforwardly related, such that $p = 1 - q$. Such derivations may be useful for other authors, who may want to replicate these results, or extend the applications of the noisy model of percolation.

In this birth and death graph process, each edge is treated separately by integrating out the dependence of all other edges in the graph, and considering the marginal distribution of every $X_t(e)$. As before, we will here refer to $X_t(e)$ as the *latent* edge status, and $m_t := |E_t|$ will indicate the number of edges in the graph at time t. Given the Markov random variable Y_t, the conditional transition matrix for every $X_{t+1}(e)$, given some value of Y_t takes the following form,

	$X_{t+1}(e) = 0$	$X_{t+1}(e) = 1$
$X_t(e) = 0$	$\frac{\binom{n}{2} - m_t - I\{Y_t=1\}}{\binom{n}{2} - m_t}$	$\frac{I\{Y_t=1\}}{\binom{n}{2} - m_t}$
$X_t(e) = 1$	$\frac{I\{Y_t=0\}}{m_t}$	$\frac{m_t - I\{Y_t=0\}}{m_t}$

where $\binom{n}{2}$ denotes the number of edges in a saturated graph of size n, and where $I\{f(x)\}$ is the indicator function, which takes a value of 1 if $f(x)$ is true and 0, otherwise.

In this paper, we have concentrated on a special case of this birth and death process, where we have set $p = 1-q$. This choice of p and q leads to the following characterization of the Y_t process,

$$\begin{aligned} \mathbb{P}[Y_{t+1} = 1] &= \mathbb{P}[Y_{t+1} = 1 | Y_t = 0] \\ &= \mathbb{P}[Y_{t+1} = 1 | Y_t = 1] = p, \end{aligned} \tag{A1}$$

and similarly, $\mathbb{P}[Y_{t+1} = 0] = 1-p$. Under this simplifying assumption, the preceding conditional transition matrix becomes

	$X_{t+1}(e) = 0$	$X_{t+1}(e) = 1$
$X_t(e) = 0$	$\frac{\binom{n}{2} - m_t - p}{\binom{n}{2} - m_t}$	$\frac{p}{\binom{n}{2} - m_t}$
$X_t(e) = 1$	$\frac{1-p}{m_t}$	$\frac{m_t - 1 + p}{m_t}$

Each entry is obtained by taking the expectation with respect to Y_t. That is, $\mathbb{P}[X_{t+1}(e) = \omega | X_t(e) = \omega'] = \mathbb{E}[\mathbb{P}[X_{t+1}(e) = \omega | X_t(e) = \omega', Y_t]]$, for every $\omega, \omega' \in \{0, 1\}$, and where the marginal distribution of Y_t is known from equation (A1).

One can then combine the noise process described in section II C, with the birth and death stochastic process in order to link the latent and observed parts of the Markov hidden model. This gives the following table,

	$X^*_{t+1}(e) = 0$	$X^*_{t+1}(e) = 1$
$X_t(e) = 0$	$(1-\alpha)\left(\frac{\binom{n}{2} - m_t - p}{\binom{n}{2} - m_t}\right)$	$\alpha\left(\frac{p}{\binom{n}{2} - m_t}\right)$
$X_t(e) = 1$	$\beta\left(\frac{1-p}{m_t}\right)$	$(1-\beta)\left(\frac{m_t - 1 + p}{m_t}\right)$

Since this transition matrix links the latent and observed stochastic processes, one can immediately derive the marginal probabilities of the $X^*_t(e)$'s, such that

$$\mathbb{P}[X^*_{t+1}(e) = 0] = (1-\alpha)\left(\frac{\binom{n}{2} - m_t - p}{\binom{n}{2} - m_t}\right) + \beta\left(\frac{1-p}{m_t}\right),$$

and similarly for $\mathbb{P}[X_t^*(e) = 1]$. Observe that the resulting X_t process is *non-stationary*. Moreover, non-stationarity also holds when considering the case $p = 1 - q$. Indeed, since the probability of adding a new edge at time $t + 1$ is dependent on the number of existing edges, $m_t := |E_t|$, at time t; it follows that the resulting joint distribution of any subset of the X_t's depends on the choice of t.

[1] Lucas Daniel Valdez, Pablo Alejandro Macri, and Lidia Adriana Braunstein. Temporal percolation of the susceptible network in an epidemic spreading. *PloS one*, 7(9):e44188, 2012. ISSN 1932-6203.

[2] Joaquim Elias de Freitas, Liacir dos Santos Lucena, and Stphane Roux. Percolation as a dynamical phenomenon. *Physica A: Statistical Mechanics and its Applications*, 266(1):81–85, 1999. ISSN 0378-4371.

[3] P. Erdős and A. Rényi. The evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17, 1960.

[4] N. Alon and J.H. Spencer. *The Probabilistic Method*. Springer, London, 2004.

[5] Oliver Riordan and Lutz Warnke. Explosive percolation is continuous. *Science*, 333(6040):322–324, July 2011. URL `http://www.sciencemag.org/content/333/6040/322.abstract`.

[6] Dimitris Achlioptas, Raissa M. D'Souza, and Joel Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, March 2009.

[7] R. A. da Costa, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Explosive percolation transition is actually continuous. *Phys. Rev. Lett.*, 105(25):255701, December 2010. URL `http://link.aps.org/doi/10.1103/PhysRevLett.105.255701`.

[8] O. Riordan and L. Warnke. Achlioptas process phase transitions are continuous. *Annals of Applied Probability*, 4(22):1450–1464, 2012.

[9] R.M. D'Souza and J. Nagler. Anomalous critical and supercritical phenomena in explosive percolation. *Nat Phys*, 11(7):531–538, 2015. ISSN 1745-2473. URL `http://dx.doi.org/10.1038/nphys3378`.

[10] T. Bohman and A. Frieze. Avoiding a giant component. *Random Structures and Algorithms*, 19:75, 2001.

[11] A. Beveridge, T. Bohman, A. Frieze, and O. Pikhurko. Product rule wins a competitive game. *Proc. Amer. Math. Soc*, 135:3061–3071, 2007.

[12] Michael Krivelevich, Eyal Lubetzky, and Benny Sudakov. Hamiltonicity thresholds in Achlioptas processes. *Random Struct. Alg.*, 37(1):1–24, August 2010. ISSN 1098-2418. URL `http://dx.doi.org/10.1002/rsa.20302`.

[13] Maxime Guye, Jean Regis, Manabu Tamura, Fabrice Wendling, Aileen Mc Gonigal, Patrick Chauvel, and Fabrice Bartolomei. The role of corticothalamic coupling in human temporal lobe epilepsy. *Brain*, 129(7):1917–1928, July 2006. URL `http://brain.oxfordjournals.org/content/129/7/1917.abstract`.

[14] SC Ponten, F Bartolomei, and CJ Stam. Small-world networks and epilepsy: graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clinical neurophysiology*, 118(4):918–927, 2007. ISSN 1388-2457.

[15] Kaspar Schindler, Howan Leung, Christian E Elger, and Klaus Lehnertz. Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG. *Brain*, 130(1):65–77, 2007. ISSN 0006-8950.

[16] Kaspar Schindler, Christian E Elger, and Klaus Lehnertz. Increasing synchronization may promote seizure termination: evidence from status epilepticus. *Clinical neurophysiology*, 118(9):1955–1968, 2007. ISSN 1388-2457.

[17] M.A. Kramer, U.R. Eden, E.D. Kolaczyk, R. Zepeda, E.N. Eskandar, and S.S. Cash. Coalescence and fragmentation of cortical networks during focal seizures. *Journal of Neuroscience*, 30:10076–10085, 2010.

[18] Kaspar Schindler, Frederique Amor, Heidemarie Gast, Markus Muller, Alexander Stibal, Luigi Mariani, and Christian Rummel. Peri-ictal correlation dynamics of high-frequency (80–200Hz) intracranial EEG. *Epilepsy research*, 89(1):72–81, 2010. ISSN 0920-1211.

[19] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, 52:1059–1069, 2010.

[20] Mark A. Kramer and Sydney S. Cash. Epilepsy as a disorder of cortical network organization. *The Neuroscientist*, 18(4):360–372, August 2012. URL `http://nro.sagepub.com/content/18/4/360.abstract`.

[21] E. Bullmore and Olaf Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(1):1–13, 2009.

[22] Danielle S. Bassett, Edward T. Bullmore, Andreas Meyer-Lindenberg, Jos A. Apud, Daniel R. Weinberger, and Richard Coppola. Cognitive fitness of cost-efficient brain functional networks. *Proceedings of the National Academy of Sciences*, 106(28):11747–11752, July 2009. URL `http://www.pnas.org/content/106/28/11747.abstract`.

[23] Cedric E. Ginestet and A. Simmons. Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *NeuroImage*, 5(2):688–704, 2011. doi:doi:10.1016/j.neuroimage.2010.11.030.

[24] Raymond Salvador, John Suckling, Martin R. Coleman, John D. Pickard, David Menon, and Ed Bullmore. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cereb. Cortex*, 15(9):1332–1342, September 2005. URL `http://cercor.oxfordjournals.org/cgi/content/abstract/15/9/1332`.

[25] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, February 2009. URL `http://www.pnas.org/content/106/6/2035.abstract`.

[26] Zhenhua Wu, Cecilia Lagorio, Lidia A Braunstein, Reuven Cohen, Shlomo Havlin, and H Eugene Stanley. Numerical evaluation of the upper critical dimension of percolation in scale-free networks. *Physical Review E*, 75 (6):066110–, 2007.

[27] M.A. Kramer, U.T. Eden, S.S. Cash, and E.D. Kolaczyk. Network inference with confidence from multivariate time series. *Phys. Rev. E.*, 79:061916, 2009.

[28] A Margolina, HJ Herrmann, and D Stauffer. Size of

largest and second largest cluster in random percolation. *Physics Letters A*, 93(2):73–75, 1982. ISSN 0375-9601.

[29] CR da Silva, ML Lyra, and GM Viswanathan. Largest and second largest cluster statistics at the percolation threshold of hypercubic lattices. *Physical Review E*, 66 (5):056107, 2002.

[30] Ariel Haimovici, Enzo Tagliazucchi, Pablo Balenzuela, and Dante R Chialvo. Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Physical review letters*, 110(17): 178101–, 2013.

[31] Ana Ciurea, Ioana Mindruta, Mihai Dragos Maliia, Alexe Ciurea, Jean Ciurea, Andrei Barborica, Cristian Donos, Manuel F Casanova, and Ioan Opris. Modular signatures and neural avalanches in epileptic brain networks. In *Recent Advances on the Modular Organization of the Cortex*, pages 271–295. Springer, 2015.

[32] Paola Flocchini, Bernard Mans, and Nicola Santoro. On the exploration of time-varying networks. *Theoretical Computer Science*, 469:53–68, 2012. ISSN 0304-3975.

[33] Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012. ISSN 1744-5760.

[34] Hartmut H. K. Lentz, Thomas Selhorst, and Igor M. Sokolov. Unfolding accessibility provides a macroscopic approach to temporal networks. *Phys. Rev. Lett.*, 110 (11):118701, March 2013. URL `http://link.aps.org/doi/10.1103/PhysRevLett.110.118701`.

[35] P. Grindrod and D.J. Higham. Evolving graphs: Dynamical models, inverse problems and propagation. *Proceedings of the Royal Society, Series A*, 0456:1–18, 2009.

[36] Yariv Ephraim and Neri Merhav. Hidden markov processes. *IEEE Transactions on Information Theory*, 48 (6):1518–1569, 2002. ISSN 0018-9448.

[37] Paraskevas Giazitzidis and Panos Argyrakis. Generalized achlioptas process for the delay of criticality in the percolation process. *Physical Review E*, 88(2):024801, 2013.

[38] R. A. da Costa, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Critical exponents of the explosive percolation transition. *Phys. Rev. E*, 89(4): 042148, 2014. URL `http://link.aps.org/doi/10.1103/PhysRevE.89.042148`.

[39] R. A. da Costa, S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Inverting the Achlioptas rule for explosive percolation. *Phys. Rev. E*, 91 (4):042130, 2015. URL `http://link.aps.org/doi/10.1103/PhysRevE.91.042130`.