THE ORDER-CHAOS PHASE TRANSITION FOR A GENERAL CLASS OF COMPLEX BOOLEAN NETWORKS

By Shirshendu Chatterjee*

New York University*

We consider a model for heterogeneous gene regulatory networks that is a generalization of the model proposed by Chatterjee and Durrett [6] as an “annealed approximation” of Kauffmann’s [16] random Boolean networks. In this model, genes are represented by the vertices of a random directed graph \( G_n \) on \( n \) vertices having specified in-degree distribution \( p_{\text{in}} \) (resp. joint distribution \( p_{\text{in,out}} \) of in-degree and out-degree), and the expression bias (the expected fraction of 1’s in the Boolean functions) \( p \) is same for all vertices. Following [6] and a standard practice in the physics literature, we use a discrete-time threshold contact process with parameter \( q = 2p(1-p) \) (in which vertices are either occupied or vacant, a vertex with at least one (resp. no) occupied input at time \( t \) will be occupied at time \( t+1 \) with probability \( q \) (resp. 0)) on \( G_n \) to approximate the dynamics of the Boolean network. We find a parameter \( r \), which has an explicit expression in terms of certain moments of \( p_{\text{in}} \) (resp. \( p_{\text{in,out}} \)), such that, with probability tending to 1 as \( n \) goes to infinity, if \( r \cdot 2p(1-p) > 1 \), then starting from all occupied sites the threshold contact process on \( G_n \) maintains a positive (quasi-stationary) density \( \pi(p_{\text{in}}) \) (resp. \( \pi(p_{\text{in,out}}) \)) of occupied sites till time which is exponential in \( n \), whereas if \( r \cdot 2p(1-p) < 1 \), then the persistence time of the threshold contact process on \( G_n \) is at most logarithmic in \( n \). These two phases correspond to the chaotic and ordered behavior of the gene networks respectively.

1. Introduction. Experimental evidence [2, 25, 29] suggests that in various biological systems, the complex kinetics of genetic control is reasonably well approximated by Boolean network models. This kind of models were first formulated by Kauffman [16]. Over the last few years, these models have received significant attention, both at the level of model formulations and numerical simulations (see e.g. the surveys [14, 15, 17, 23, 27] and the references therein) and at the rigorous level (see e.g. [6, 11]). The class of models can be described as follows. Genes are represented by the vertices of a directed graph \( G_n = (\mathbb{N}_n, \mathcal{E}_n) \) on \( n \) vertices, where \( \mathbb{N}_n := \{1, 2, \ldots, n\} \) denotes the vertex set and \( \mathcal{E}_n \) denotes the edge set of \( G_n \). The state \( \eta_{t}(x) \) of a vertex \( x \in [n] \) at time \( t = 0, 1, 2, \ldots \) is either 1 (‘on’) or 0 (‘off’), and each vertex \( x \) receives input from the vertices which point to it in \( G_n \), namely

\[
Y^x := \{ y \in [n] : \langle y, x \rangle \in \mathcal{E}_n \}.
\]

\( Y^x \) is called the input set and its members are called the input vertices for \( x \). The states \( \{\eta_{t}(x)\}_{t \geq 0, x \in [n]} \) evolve according to the update rule

\[
\eta_{t+1}(x) = f_x((\eta_{t}(y), y \in Y^x)), \quad x \in [n],
\]

AMS 2000 subject classifications: Primary 60K35, 05C80; secondary 06E30, 82B26.

Keywords and phrases: random graphs, inhomogeneous graphs, directed networks, Boolean networks, threshold contact process, phase transition, gene regulatory networks.
where each \( {f_x : \{0, 1\}^{[Y^x]} \to \{0, 1\}} \) is some time-independent Boolean function defined on the set of states of the input vertices for \( x \). Here and later we use \( |A| \) to denote the size of a set \( A \).

In order to understand general properties of such dynamical systems, various random Boolean network models have been formulated, which form an important subfamily of these models. The simplest such model with parameters \( r \) (number of inputs per vertex) and \( p \) (expression bias), which we denote by RBN\(_{r,p}^0\), consists of the following specification of the constructs in the above general model. The ground graph \( \mathcal{G}_n \) is constructed by letting the input set \( Y^x \) consist of \( r \) distinct vertices \( Y_1(x), \ldots, Y_r(x) \), which are uniformly from \( [n] \setminus \{x\} \). The values \( f_x(v), x \in [n], v \in \{0, 1\}^r \), are assigned independently and each equals 1 with probability \( p \). The dynamics then proceeds as in (1.1) from a specified starting configuration \( \{\eta_0(x) : x \in [n]\} \) at time \( t = 0 \). Note that the functions \( f_x \) and the graph \( \mathcal{G}_n \) are fixed at time 0 and then the dynamics of this system is deterministic.

Kauffman introduced RBN\(_{r,1/2}^0\) in [16], and that model has been analyzed in detail for \( r = 1 \) [11]. The general model RBN\(_{r,p}^0\) has been studied extensively via simulations (see e.g. [15, 24]) and using heuristics from Statistical Physics (see e.g., [8, 9]). It has been argued in [7] that the behavior of RBN\(_{r,p}^0\) undergo a phase transition for \( r \geq 3 \), and

\[ (1.2) \quad \text{the order-chaos phase transition curve for RBN}_{r,p}^0 \text{ is given by } 2p(1-p) \cdot r = 1 \]

in the sense that RBN\(_{r,p}^0\) is “ordered” (the configuration of zeros and ones rapidly converges to a fixed point or attractor) for \( 2p(1-p) \cdot r < 1 \), whereas RBN\(_{r,p}^0\) is “chaotic” (the configuration keeps on changing for an exponentially long time) when \( 2p(1-p) \cdot r > 1 \). This phase transition picture has recently been proved for an “annealed approximation” of the deterministic dynamical system [6, 20], which we shall describe soon.

Although RBN\(_{r,p}^0\) has been studied extensively, the model deals with an idealized setting of homogeneous graphs, where every vertex has the same in-degree. This assumption constrains the model in the context of biological applications where such networks have quite heterogeneous degrees [1]. So, the obvious natural questions are that if the update rule is similar to that for RBN\(_{r,p}^0\) (namely, given the input sets \( \{Y^x : x \in [n]\} \), (1.1) holds for the values \( \{f_x(v) : v \in \{0, 1\}^r, x \in [n]\} \) chosen via independent coin flips such that each value is 1 with probability \( p \)), then whether similar phase transition occurs in case of heterogeneous and more complex ground graphs, and if it occurs, how do the corresponding phase transition curves depend on the underlying parameters describing the properties of the ground graphs.

In the context of gene networks, heterogeneous graphs have been considered in the physics and biology literature. Mainly two classes of such graph models have been formulated.

(i) **Graphs with prescribed in-degree:** A number of authors (see e.g., [3, 12, 19, 26]) have considered directed graph models with prescribed in-degree distribution. Here, one starts with a probability mass function \( p^\text{in} := \{p_k^\text{in}\}_{k \geq 1} \) and chooses the ground graph uniformly at random from the collection of all simple directed graphs for which the proportion of vertices having in-degree \( k \) is asymptotically \( p_k^\text{in} \) for all \( k \geq 1 \). The precise method of construction is described in Section 1.2. We denote the associated random Boolean network model, where the ground graph has in-degree distribution \( p^\text{in} \) and the update rule is similar to that of RBN\(_0^0\) with expression bias \( p \), by RBN\(_1^0(p^\text{in}, p)\). It has been argued in the above
papers that

\[
\text{the order-chaos phase transition curve for } \text{RBN}^1(p^\text{in}, p) \text{ is } 2p(1 - p) \cdot r^\text{in} = 1,
\]

(1.3) where \( r^\text{in} := \sum_k k \cdot p^\text{in}_k \) is the average in-degree.

\text{Remark 1.1.} An analogous model can be built, where one starts with the out-degree distribution \( p^\text{out} = \{p^\text{out}_k\}_{k \geq 0} \) and the ground graph is chosen uniformly at random from the collection of all directed graphs for which the proportion of vertices having out-degree \( k \) is asymptotically \( p^\text{out}_k \) for all \( k \geq 0 \). In that case, the associated in-degree distribution turns out to be asymptotically Poisson with mean \( \sum_k k p^\text{in}_k \), which means that a positive fraction of vertices will have no input vertex for them. For these vertices the update rule cannot be defined. So we avoid this particular model. We didn’t find occurrence of this model in the physics and biology model as well.

(ii) \textbf{Graphs with prescribed joint distribution of in-degree and out-degree:} Another class of models have been considered in order to incorporate correlation between the in-degree and out-degree via prescribing their joint distribution [18]. These models are more general and complex in nature. Here, one starts with a bivariate probability mass function \( p^{\text{in, out}} = \{p^{\text{in, out}}_{k,l}\}_{k \geq 0, l \geq 0} \) and chooses a graph uniformly at random from the collection of all graphs for which the asymptotic density of vertices having in-degree \( k \) and out-degree \( l \) is \( p^{\text{in, out}}_{k,l} \). We defer a complete description of the construction to Section 1.2. We denote the analogue of RBN\(^1\) with joint distribution \( p^{\text{in, out}} \) and expression bias \( p \) by RBN\(^2\)(\( p^{\text{in, out}}, p \)). It is argued non-rigorously in [18] that for this model

\[
\text{the order-chaos phase transition curve for } \text{RBN}^2(p^{\text{in, out}}, p) \text{ is } 2p(1 - p) \cdot \frac{r^{\text{in, out}}}{r^\text{in}} = 1,
\]

(1.4) where \( r^{\text{in, out}} := \sum_{k,l} k \cdot l \cdot p^{\text{in, out}}_{k,l} \), \( r^\text{in} := \sum k \cdot p^{\text{in}}_k \).

One of our aims is to prove (1.3) and (1.4) for a certain approximation of the dynamics.

1.1. \textbf{Annealed approximations to Boolean networks.} Proving rigorous results about the formulated discrete dynamical systems turns out to be quite hard. In order to understand the conjectured phase transitions in (1.3) and (1.4) rigorously, we consider a different process called the \textit{threshold contact process}. To motivate this process and explain its connection to the Boolean network model \( \{\eta_t(x) : x \in \mathcal{G}_n, t \geq 0\} \), first consider the process \( \{\zeta_t(x), x \in [n], t \geq 1\} \), where \( \zeta_t(x) = 1 \) if \( \eta_t(x) \neq \eta_{t-1}(x) \) and \( \zeta_t(x) = 0 \) otherwise. Fix vertex \( x \in [n] \). Suppose at least one of the inputs \( y \in Y^x \) changes its state between time epochs \( t - 1 \) and \( t \) so that \( \eta_t(y) \neq \eta_{t-1}(y) \). Then the state of vertex \( x \) at time \( t+1 \) is computed by looking at a different entry of \( f_x \). Ignoring the fact that we may have used this entry before, one approximately has

\[
P(\zeta_{t+1}(x) = 1|\zeta_t(y) = 1 \text{ for at least one } y \in Y^x) = 2p(1 - p).
\]

If \( \zeta_t(y) = 0 \) for all \( y \in Y^x \), then obviously \( \zeta_{t+1}(x) = 0 \). This dynamics motivates to consider the following process, which will be the main subject of this study.
The discrete-time threshold contact process with parameter \( q \in (0, 1) \) on a directed graph having vertex set \([n]\), where \( Y^x \) denotes the set of input vertices for \( x \in [n] \), is the Markov process \( \{\xi_t(x) : x \in [n], t \geq 0\} \) on \([0,1]^n\) for which the evolution dynamics is

\[
P(\xi_{t+1}(x) = 1 | \xi_t(y) = 1 \text{ for at least one } y \in Y^x) = q, \text{ and }
P(\xi_{t+1}(x) = 0 | \xi_t(y) = 0 \text{ for all } y \in Y^x) = 1.
\]

(1.5)

Conditional on the state at time \( t \), the decisions on the values of \( \xi_{t+1}(x), x \in [n] \), are independent.

This process has been called the annealed approximation to the random Boolean network model [7]. For the remainder of the paper we will write \( q = 2p(1-p) \). For the threshold contact process, we shall prove that the conjectures in (1.3) and (1.4) do represent the order-chaos phase transition for the dynamics.

1.2. Construction of random directed ground graphs. In this section, we provide precise mathematical formulation of the ground graph models which are used for defining RBN\(^1\) and RBN\(^2\).

1.2.1. Construction of the graph for RBN\(^1\). For RBN\(^1\), we start with a prescribed in-degree distribution \( p^{in}_k = \{p^{in}_k\}_{k \geq 1} \) having finite mean \( r^{in} := \sum k \cdot p^{in}_k < \infty \). Following [6], we construct the (directed) ground graph \( \mathcal{G}_n = \mathcal{G}_n([n], \mathcal{E}_n) \) having in-degree distribution \( p^{in} \) as follows. First, we choose the in-degrees \( I_1, I_2, \ldots, I_n \) independently with common distribution \( p^{in} \). Then, for each vertex \( x \in [n] \) we construct the corresponding input set \( Y^x := \{Y_1(x), Y_2(x), \ldots, Y_{I_x}(x)\} \) by choosing \( |I_x| \) many distinct vertices uniformly from \([n]\) \( \setminus \{x\}\). Finally we place oriented edges from these chosen vertices to \( x \) to obtain the edge set

\[
\mathcal{E}_n := \{(Y_i(x), x) : x \in [n], 1 \leq i \leq I_x\}
\]

of the graph \( \mathcal{G}_n \). In other words, if we write \( \mathbb{P}_{1,I} \) for the conditional ("quenched") distribution of \( \mathcal{G}_n \) given the in-degree sequence \( I = (I_1, I_2, \ldots, I_n) \) and \( \mathbb{P}_{1,n} \) for the unconditional ("annealed") distribution of \( \mathcal{G}_n \), and if \( E_{zx} \) denotes the number of directed edges from vertex \( z \) to \( x \), then

\[
\mathbb{P}_{1,n}(\cdot) := \sum_{I \in \mathbb{N}^n} \mathbb{P}_{1,I}(\cdot) \cdot p^{in}_\otimes(I),
\]

where \( p^{in}_\otimes \) is the product measure on \( \mathbb{N}^n \) with marginal \( p^{in} \), and

\[
\mathbb{P}_{1,I}(E_{zx} = e_{zx}, z, x \in [n]) = 1 \prod_{i=1}^n \left( \begin{array}{c} n-1 \\ I_i \end{array} \right) \text{ if } e_{x,x} \in \{0, 1\}, e_{x,z} = 0, \sum_{z=1}^n e_{x,z} = I_x,
\]

(1.6)

for all \( x \in [n] \) and \( \mathbb{P}_{1,I}(E_{x,z} = e_{z,x}, 1 \leq z, x \leq n) = 0 \) otherwise. Thus, under \( \mathbb{P}_{1,I} \), \( \mathcal{G}_n \) is distributed uniformly over all simple directed graphs having in-degree sequence \( I \).

1.2.2. Construction of the graph for RBN\(^2\). We follow the construction procedure of Newman, Strogatz and Watts [21, 22] to obtain the configuration model. We start with a prescribed joint distribution for in-degree and out-degree \( p^{in, out} = \{p^{in, out}_{k,l}\}_{k,l \geq 1} \) such that \( r^{in, out} := \sum k \cdot l \cdot p^{in, out}_{k,l} < \infty \) and \( r^{in} := \sum k \cdot p^{in}_{k,l} = \sum k \cdot l \cdot p^{out}_{k,l} =: r^{out} \). It is easy to see that the later condition is necessary for \( p^{in, out} \) to be eligible for our purpose. Let \( \{(I_i, O_i)\}_{i=1}^n \)
be i.i.d. with common distribution \( \mathbf{p}_{\text{in, out}} \); here \( I_x \) and \( O_x \) denote the in-degree and out-degree of vertex \( x \) respectively. We need to condition on the event

\[
E_n := \left\{ \sum_{i=1}^{n} I_i = \sum_{i=1}^{n} O_i \right\}
\]

in order to have a valid degree sequence. Having chosen the degree sequence \( (\mathbf{I}, \mathbf{O}) = ((I_1, O_1), \ldots, (I_n, O_n)) \), we allocate \( I_x \) many “inward arrows” and \( O_x \) many “outward arrows” for vertex \( x \). Then we pick a uniform random matching between the sets of inward and outward arrows. If one of the inward arrows of \( x \) is matched with one of the outward arrows of \( z \), then we let \( (z, x) \in E_n \). We will use \( \mathbb{P}_{2,1,0} \) to denote the conditional ("quenched") distribution of \( \mathcal{G}_n \) given the in-degree and out-degree sequence \( (\mathbf{I}, \mathbf{O}) \). We also condition on the event

\[
F_n := \{ \mathcal{G}_n \text{ is simple} \},
\]

i.e., it neither contains any self-loop at some vertex, nor contains multiple edges between two vertices. So if \( \mathbb{P}_{2,n} \) denotes the unconditional ("annealed") distribution of \( \mathcal{G}_n \), then

\[
\mathbb{P}_{2,n}(\cdot) = \sum_{\mathbf{I}, \mathbf{O} \in \mathbb{N}^n} \mathbb{P}_{2,1,0}(\cdot|F_n) \cdot \mathbf{p}_{\text{in, out}}((\mathbf{I}, \mathbf{O})|E_n),
\]

where \( \mathbf{p}_{\text{in, out}} \) is the product measure on \( (\mathbb{N}^2)^n \) with marginal \( \mathbf{p}_{\text{in, out}} \).

**Remark 1.3.** Given the prescribed in-degree (resp. out-degree) distribution \( \mathbf{p}_{\text{in}} \) (resp. \( \mathbf{p}_{\text{out}} \)) having finite mean, one can construct a graph \( \mathcal{G}_n \) by following the construction of \( \mathcal{G}_n \) for \( \text{RBN}^2 \) (see Section 1.2.2) corresponding to any \( \mathbf{p}_{\text{in, out}} \) with marginal \( \mathbf{p}_{\text{in}} \) (resp. \( \mathbf{p}_{\text{out}} \)). But the distribution of \( \mathcal{G}_n \) will no longer be uniform over all candidates.

1.3. **Dynamics.** Once the ground graph has been fixed via one of the above construction procedures, we will be interested in properties of the threshold contact process \( \{\xi_t(x), x \in [n]\}_{t \geq 0} \) as defined in Definition 1.2 with parameter \( q = 2p(1-p) \in [0, 1/2] \). We will often view this as a set valued process \( \{\xi_t\}_{t \geq 0} \), where \( \xi_t := \{x \in [n] : \xi_t(x) = 1\} \). We will sometimes refer to \( \xi_t \) as the set of occupied sites and \( [n] \setminus \xi_t \) as the set of vacant sites at time \( t \). We will write \( \mathbb{P}_{\mathcal{G}_n, q} \) for the distribution of the threshold contact process \( \{\xi_t\}_{t \geq 0} \) with parameter \( q \), conditioned on the ground graph \( \mathcal{G}_n \). For a fixed set \( A \subseteq [n] \), we will write \( \{\xi_t^A\}_{t \geq 0} \) for the process started with \( \xi_0^A = A \).

1.4. **Main results.** As described in Section 1.2 and 1.3, there are two layers of randomness: one corresponds to the distribution of the ground graph \( \mathcal{G}_n \), and the other corresponds to the distribution of the threshold contact process conditioned on the ground graph \( \mathcal{G}_n \).

We need another ingredient to present our main result. For a probability distribution \( \mathbf{\mu} \) on \( \{0,1,\ldots\} \) and \( q \in [0, 1] \) let \( \mathbf{\pi}(\mathbf{\mu}, q) \in [0, 1] \) denote the survival probability for the branching process with offspring distribution \((1-q)\mathbf{d}_0 + q\mathbf{\mu}\) starting from one individual. From the branching process theory,

\[
\mathbf{\pi}(\mathbf{\mu}, q) = 1 - \theta,
\]

where \( \theta \in [0,1] \) is the minimum value satisfying \( \theta = 1 - q + \sum_{k \geq 0} q \cdot \mathbf{\mu}\{k\} \cdot \theta^k \). Using the above ingredients we now present our main result.
Theorem 1.4. For any probability distribution $p^m = \{p^{m}_{k}\}_{k \geq 1}$ on $\mathbb{N}$, let $P_{1,n}$ be the probability distribution (as defined in (1.6)) on the set of random directed simple graphs on $n$ vertices having in-degree distribution $p^m$. Suppose $p^m$ has finite second moment, $p^m_k = 0$ and mean $r^m > 2$. Let $1/r^m < q \leq 1/2$ and $\pi := \pi(p^m, q) > 0$ be the branching process survival probability as defined in (1.11). For any $\varepsilon > 0$ there is a constant $\Delta(\varepsilon) > 0$ and a ‘good’ set of graphs $G_n$ satisfying $P_{1,n}(G_n) = 1 - o(1)$ such that if $G_n \in G_n$ and $\{t_n\}$ is any sequence with $\lim_{n \to \infty} t_n = \infty$, then

$$P_{G_n,q} \left( \inf_{t \leq \exp(\Delta n)} \frac{1}{n} |\xi^m_t| \geq \pi - \varepsilon, \sup_{t_n \leq t \leq \exp(\Delta n)} \frac{1}{n} |\xi^m_t| \leq \pi + \varepsilon \right) = 1 - o(1).$$

Moreover, if $q$ satisfies $q \cdot r^m < 1$, then there is a constant $C(q, r^m) > 0$ such that $P_{G_n,q}(\xi^m_C \not= \emptyset) = o(1)$ for all $G_n \in G_n$.

The above theorem proves (1.3) for the threshold-contact process. It also proves exponential persistence in the supercritical regime. Next we move to RBN$^2$.

Theorem 1.5. For any bivariate probability distribution $p^{in,out} = \{p^{in,out}_{k,l}\}_{k \geq 1, l \geq 1}$ on $\mathbb{N}^2$, let $P_{2,n}$ be the probability distribution (as defined in (1.10)) on the set of random directed simple graphs on $n$ vertices for which the joint distribution of in-degree and out-degree is $p^{in,out}$. Suppose $p^{in,out}_{k,l} = 0$ whenever $k \leq 1$, and the marginal distributions corresponding to $p^{in,out}$ have equal mean $r^*$ and finite second moment. Let $\tilde{p}^m = (\tilde{p}^m_k := (r^m)^{-1} \sum_l l \cdot p^{in,out}_{k,l})_{k \geq 2}$ be the size-biased in-degree distribution with mean $\tilde{r}^m$, $1/\tilde{r}^m < q \leq 1/2$ and $\pi := \pi(\tilde{p}^m, q) > 0$ be the branching process survival probability as defined in (1.11). For any $\varepsilon > 0$ there is a constant $\Delta(\varepsilon) > 0$ and a ‘good’ set of graphs $G_n$ satisfying $P_{2,n}(G_n) = 1 - o(1)$ such that if $G_n \in G_n$ and $\{t_n\}$ is any sequence satisfying $t_n \leq \exp(\Delta n)$ and $\lim_{n \to \infty} t_n = \infty$, then

$$P_{G_n,q} \left( \inf_{t \leq \exp(\Delta n)} \frac{1}{n} |\xi^m_t| \geq \pi - \varepsilon, \sup_{t_n \leq t \leq \exp(\Delta n)} \frac{1}{n} |\xi^m_t| \leq \pi + \varepsilon \right) = 1 - o(1).$$

Moreover, if $q$ satisfies $q \cdot \tilde{r}^m < 1$, then there is a constant $C(q, \tilde{r}^m) > 0$ such that $P_{G_n,q}(\xi^{in}_C \not= \emptyset) = o(1)$ for all $G_n \in G_n$.

The above theorem proves (1.4) for the threshold-contact process. It also proves exponential persistence for the supercritical regime in RBN$^2$.

1.5. Proof ideas. The subcritical case is relatively easier, as we can couple with a subcritical branching process. Although there are technical challenges because of heterogeneity of the ground graph, these have been tackled using a suitable second moment argument.

In the supercritical regime, there are two main aspects: estimate the probability of survival of the dual starting from (i) a single vertex (ii) large initial sets. To tackle the first, we have approximated the small neighborhoods of vertices within distance $O(\log \log n)$ by appropriate trees, and hence the survival probability becomes approximately that of a suitable branching process (see Proposition 4.2).

The remaining part is the most challenging. Here we have shown that there is a ‘good’ set of graphs (see Proposition 4.1) such that whenever the size of the occupied set in the dual process (irrespective of its location) drops below $\varepsilon_n \cdot n$, where $\varepsilon_n \cdot n \in [(\log n)^{\beta}, \varepsilon n]$, then comparing
with the behavior of the dual process on a suitably truncated forest (see Lemma 5.1) the size reaches above \( \varepsilon_n \cdot n \) within time \( O(\log \log(1/\varepsilon_n)) \) with exponentially small error probability (see Proposition 5.2). This result enables us to prove our main results.

1.6. Discussion. It is needless to say that one of the main challenges to prove prolonged persistence in the supercritical regime for the random Boolean network models that we consider is the heterogeneity of the ground graphs. The techniques used in [6], which are mostly based on “isoperimetric inequalities” for uniformly chosen directed graphs with constant in-degree, is not helpful for us. The approach adopted in [20] are also not applicable, as they need the same constraint too. So, a new approach would be required to deal with the heterogeneity.

On the other hand, not much is understood about the roles of different initial subsets of occupied sites in the prognosis of the dynamics beyond the level of the “first moment method”. It is also not understood whether the dynamics avoids certain ‘bad’ configurations of occupied sites from which survival of the dynamics is difficult when \( q \) is near the critical value, but stays in the supercritical regime. So, we need to have a very good control (more than exponential) over the probabilities of unlikely events, as the number of possible initial subsets of size \( m \) is super-exponentially large in \( m \).

Keeping these in mind, the technique used in this paper, namely coupling the dual process with a suitably truncated branching process up to a certain time, seems to be the only effective way to obtain exponential persistence in the supercritical regime.

However, this technique does not work that well in the critical regime where the parameters \( p, r^{in} \) and \( r^{in, out} \) satisfy the equality in (1.3) and (1.4). Based on the behavior of critical contact process on \( d \)-dimensional torus, one expects to have persistence of activity till some polynomial time in this case.

1.7. Organization of the paper. The remainder of the paper is organized as follows. In Section 2, we describe the dual process for the threshold contact process, which will play a crucial role in our argument, and give quantitative estimates for the error involved in approximating the local neighborhoods of certain small subsets of vertices in the dual (edge-reversed) graph. Then, in Section 3 we mention some more ingredient lemmas, which are used later. Section 4 contains description of the set of ‘good’ graphs that appears in the theorems and proof of the facts that their probabilities are \( 1 - o(1) \). Finally, in Section 5 we put all the ingredients together to have the proofs of the main theorems.

2. Preliminaries. Before jumping into the core of the proof we need some preliminary facts. We begin this section with the definition of the dual process for the threshold contact process, which will play a major role in proving the main results. We also collect asymptotic properties of local neighborhoods of the random graph models in \( \{RBN^i : i = 1, 2\} \).

2.1. Dual coalescing branching process. For a given directed graph \( \mathcal{G}_n = ([n], E_n) \), let \( \tilde{\mathcal{G}}_n = ([n], \tilde{E}_n) \) be the dual (directed) graph obtained by reversing the edges, i.e., \( \tilde{E}_n := \{\langle y, x \rangle : \langle x, y \rangle \in E_n\} \). We write \( x \to z \) if \( \langle x, z \rangle \in \tilde{E}_n \), and in that case we will occasionally refer to \( z \) as a child of \( x \) in \( \tilde{\mathcal{G}}_n \). Now for the threshold contact process \( \{\xi_t : t \geq 0\} \) on the graph \( \mathcal{G}_n \), the dual process \( \{\tilde{\xi}_t : t \geq 0\} \) is the following coalescing branching process on the dual graph \( \tilde{\mathcal{G}}_n \). For any \( t \geq 0 \), each site of \( \tilde{\xi}_t \) gives birth at time \( t \) independently with probability \( q \). If \( x \in \tilde{\xi}_t \) gives birth, all of its children are included in \( \tilde{\xi}_{t+1} \). In other words, every vertex \( \tilde{\xi}_t \)
gives birth with probability $q$ independently across vertices, and $z \in [n]$ is included in $\xi_{t+1}$ if there exists $x \in \xi_t$ which gives birth at time $t$ and $x \to z$. We write $\{\xi_t^B : t \geq 0\}$ for the coalescing branching process starting from $\xi_0^B = B$.

It is easy to check [13] that the following duality relation holds. For any $t \geq 0$ and for any two sets $A, B \subseteq [n]$ we have

$$\Pr \left( \xi_t^A \cap B \neq \emptyset \right) = \Pr \left( \xi_t^B \cap A \neq \emptyset \right).$$

This will enable us to do suitable analysis for the dual process, and then carry the implications forward to prove our main results about $\{\xi_t\}$.

We will need the following notation in the sequel. For the graph $\ Annex$ and $U \subset [n]$ define

$$U^{*1} := \{ z \in [n] : x \to z \text{ for some } x \in U \}.$$

2.2. Local neighborhoods. Next, we need to understand the structure of the neighborhood of a small set of vertices of $\ Annex$. The goal is to see whether the oriented neighborhood of a typical small vertex set contains an oriented forest having offspring distribution is close to the out-degree distribution for $\ Annex$.

For $A \subset [n]$ we let $\ Annex = A$ and for $l \geq 1$ let

$$\ Annex := \{ z \in [n] : \text{there is an oriented path in } \ Annex \text{ of length } l \text{ from some } x \in A \text{ to } z \}.$$

Let $\{K_m\}_{m=1}^n$ be a sequence of numbers, which will be specified later (see (4.4)). Using $\cup$ to denote disjoint union, we introduce the following coupling between the directed subgraph of $\ Annex$ induced by $\cup_{l=0}^{K[A]} \ Annex$ and a forest $\{Z_t^A, 0 \leq t \leq K[A]\}$ along with partitions $Z_t^A = C_t^A \cup O_t^A \cup R_t^A$, where $C_t^A$, $O_t^A$ and $R_t^A$ represent `closed', `open' and `removed' sites at level $t$ respectively. Let $A = \{u_1, \ldots, u_{|A|}\}$. For the root level of the forest we choose $Z_0^A = A$ and $C_0^A = R_0^A = \emptyset$, so $O_0^A = A$. The sites in $Z_0^A$ are labeled $u_1, \ldots, u_{|A|}$. For each $t \geq 0$, every site of $O_t^A$ mimics the corresponding vertex in $\ Annex$ with same label, so a site of $O_t^A$ having label $u$ gives birth to $I_u$ many children at level $t + 1$. The new born sites at level $t + 1$ are assigned the same labels following those of $\ Annex$. Writing $r = r^\inf$ (resp. $r^\sup$) in case of RBN$^1$ (resp. RBN$^2$), and letting

$$\Pi(l, A)$$

denote the subset of $A$ consisting of $l \land |A|$ elements with minimum indices,

we scan the sites of $Z_{t+1}^A$ in an increasing order of labels, and define

$$R_{t+1}^A := Z_{t+1}^A \setminus \Pi(2r|O_{t+1}^A|, Z_{t+1}^A).$$

For a site in $Z_{t+1}^A \setminus R_{t+1}^A$, we say that a “collision” has occurred if its label either matches with that of a site in $\cup_{z=0} Z_t^A$, or has already been found while scanning the sites of level $t + 1$. We include all of these sites in $C_{t+1}^A$. If collision does not occur at a site, we include that in $O_{t+1}^A$.

For $u \in Z_t^A$, let $\ Annex \subseteq A$ denote the label of the unique ancestor of $u$ having level 0. For any subset $B \subset A$ and $t \geq 1$ let $Z_{t+1}^{A,B} = O_{t+1}^{A,B} = B$ and

$$Z_{t+1}^{A,B} := \{ u \in Z_t^A : \ Annex \subseteq B \}$$

$$C_{t+1}^{A,B} := \Pi \left( 2r \left| O_{t-1}^{A,B} \right|, Z_{t+1}^{A,B} \right) \cap C_t^A,$$

$$O_{t+1}^{A,B} := \Pi \left( 2r \left| O_{t-1}^{A,B} \right|, Z_{t+1}^{A,B} \right) \cap O_t^A.$$
SO, each site in $Q^A$ corresponds to a unique vertex in $\tilde{Z}^A_t$ with the same label. Note that this map from $Q^A_t$ to $\tilde{Z}^A_t$ may not be onto because of collisions and removal of sites.

The law of $\mathcal{G}_n$ induces the law of $\{Z^A_t\}$ along with its partitions. We identify these two laws. Now our aim is to estimate the probability of collision, and then understand the offspring distribution in the above forest. We write

\[(2.2) \quad \vartheta(m) := m + m \sum_{t=1}^{K_m} (2r)^t, \quad \text{and} \]

\[I_n := \frac{1}{n} \sum_{z=1}^{n} I_z, \quad \bar{T}^2_n := \frac{1}{n} \sum_{z=1}^{n} f_z^2, \quad \bar{O}_n := \frac{1}{n} \sum_{z=1}^{n} O_z, \quad \bar{O}^2_n := \frac{1}{n} \sum_{z=1}^{n} O_z^2, \quad \bar{I}_O_n := \frac{1}{n} \sum_{z=1}^{n} I_z O_z. \]

For $\eta \in (0, 1)$ and any probability distribution $\mu$, we define

\[(2.3) \quad \Gamma(\eta, \mu) := \int_0^{\eta} \mu^{-t} (1 - t) \, dt, \quad \text{where} \quad \mu^{-t}(t) := \inf \{ y \in \mathbb{R} : \mu((\infty, y]) \geq t \} \quad \text{for} \quad t \in [0, 1]. \]

Recall from Section 1.2 that $I_x$ denotes the in-degree of $x$ and $\{Y^{x}_1, \ldots, Y^{x}_{I_x}\}$ denotes the set of input vertices for $x$ in $\mathcal{G}_n$.

**Lemma 2.1 (For RBN$^1$).** If $p^m$ has finite second moment, then there is a constant $C_{2.1} > 0$ and a set of in-degree sequences $\mathcal{J} \subset \mathbb{N}^n$ such that $\mathbb{P}^m_{\mathcal{G}_n}(\mathcal{J}_n) = 1 - o(1)$ and $I \in \mathcal{J}_n$ implies

1. $\mathbb{P}_{1,1}(x \in C^A_t) \leq 2\vartheta(|A|)/n \quad \text{for any} \quad A \subset [n], x \in Z^A_t \setminus R^A_t$ and $1 \leq t \leq K_1 |A|$.
2. $\mathbb{P}_{1,1} \left( (\tilde{Y}^x_i, x \in [n], 1 \leq i \leq I_x) \in \cdot \right) \leq C_{2.1} \mathbb{P}_{1,1} \left( (\tilde{Y}^z_i, z \in [n], 1 \leq i \leq I_z) \in \cdot \right)$,

where $\{\tilde{Y}^x_i\}_{x \in [n], i \leq I_x}$ are i.i.d. with common distribution $\text{Uniform}([n])$.

**Proof.** We take

\[\mathcal{J}_n := \left\{ \max_z I_z \leq n^{3/4}, T^z_n < c_1, \bar{T}^2_n < c_2 \right\}, \]

where $c_i := 2 \sum_i k^i p^m_i$. Obviously $\mathbb{P}^m_{\mathcal{G}_n}(\mathcal{J}_n) = 1 - o(1)$.

(1) Note that $\sum_{t=0}^{K_1 |A|} |Z^A_t \setminus R^A_t| \leq \vartheta(|A|)$. So if $I \in \mathcal{J}_n$, then it is easy to see from the construction of $\mathcal{G}_n$ under the law $\mathbb{P}_{1,1}$ that for any $1 \leq t \leq K_1 |A|$ and $x \in Z^A_t \setminus R^A_t$,

\[\mathbb{P}_{1,1}(x \in C^A_t) \leq \frac{\vartheta(|A|)}{n - \max_z I_z} \leq 2\vartheta(|A|)/n. \]

(2) It is easy to see that

\[(Y^x_i, 1 \leq i \leq I_x, x \in [n]) \overset{d}{=} \left( \tilde{Y}^x_i, 1 \leq i \leq I_x, x \in [n] \left| \tilde{Y}^z_i \neq \tilde{Y}^z_j \neq \forall z \in [n], i \neq j \right. \right), \]

so it suffices to show that $I \in \mathcal{J}_n$ implies $\mathbb{P}_{1,1}(\tilde{Y}^z_i \neq \tilde{Y}^z_j \neq z \forall z \in [n], 1 \leq i \neq j \leq I_z) \geq c$ for some constant. Using the inequality $1 - x \geq e^{-2x}$ for $x > 0$ small enough, if $I \in \mathcal{J}_n$, then

\[\mathbb{P}_{1,1}(\tilde{Y}^z_i \neq \tilde{Y}^z_j \neq z \forall z \in [n], 1 \leq i \neq j \leq I_z) = \prod_{z=1}^{n} \prod_{i=1}^{I_z} (1 - i/n) \geq \exp \left( -2 \sum_{z=1}^{n} \sum_{i=1}^{I_z} (i/n) \right) = \exp(-T^z_n - \bar{T}^2_n) \geq e^{c_1 - c_2} \]

for large enough $n$.  

\[\blacksquare\]
Lemma 2.2 (For RBN^2). If the marginal distributions \( p^{in} \) and \( p^{out} \) corresponding to \( p^{in, out} \) have finite second moment, then there is a constant \( C_{2.2} > 0 \) and a set of degree sequences \( \mathcal{A} \subset (N^2)^n \) such that \( p^{in, out}_{\mathcal{A}_n}(\mathcal{A}|E_n) = 1 - o(1) \) and for all \( (I, O) \in \mathcal{A} \),

\[
\begin{align*}
(1) & \quad P_{2.1,0}(x \in C^A) \leq 4\Gamma(2\varepsilon, p^{out}(I))/r^{in} + 2\varepsilon \quad \text{whenever} \quad A \subset [n] \quad \text{satisfies} \quad \vartheta(|A|) \leq \varepsilon n \\
(2) & \quad P_{2.1,0}(\{Y^x_i, 1 \leq i \leq I_x, x \in [n]\} \in \cdot | F_n) \\
& \leq C_{2.2} P_{2.1,0} \left( \left\{ Y^x_i, 1 \leq i \leq I_x, x \in [n] \right\} \in \cdot \left\{ \sum_{x=1}^n I_x \prod_{i=1}^n 1_{\{Y^x_i = z\}} = O_2 \forall z \in [n] \right\} \right),
\end{align*}
\]

where \( \{Y^x_i\} \) are i.i.d. with common distribution \( \sum_{z \in [n]} O_2 \delta_z/ \sum_{x \in [n]} O_x \). Moreover, if \( p^{out}_k \sim ck^{-\alpha} \) for some \( \alpha > 2 \), then \( \Gamma(\eta, p^{out}) \sim c\eta^{(\alpha - 2)/2} \) for some constant \( c > 0 \) which depends on \( \alpha \) only.

If \( N = rn \) for some constant \( r \) and \( \{Y^x_i\}_{i=1}^N \) are i.i.d. with common distribution Multinomial \( (1; \alpha_1, \alpha_2, \ldots, \alpha_n) \), then for any small \( \varepsilon > 0 \),

\[
\begin{align*}
(3) & \quad P \left( \left( \sum_{i=1}^N 1_{\{Y_i = z\}} = N\alpha_z \forall z \in [n] \right) \right) \\
& \leq (1 + o(1)) \exp(-\varepsilon \log(1 - \varepsilon)N) P \left( \left( \sum_{i=1}^N 1_{\{Y_i = z\}} = N\alpha_z \forall z \in [n] \right) \right),
\end{align*}
\]

\[
\begin{align*}
(4) & \quad \left\| P \left( \left( \sum_{i=1}^N 1_{\{Y_i = z\}} = N\alpha_z \forall z \in [n] \right) - P \left( \left( \sum_{i=1}^N 1_{\{Y_i = z\}} = N\alpha_z \forall z \in [n] \right) \right) \right\|_{TV} \\
& \leq O(\varepsilon^2 n) + o(1).
\end{align*}
\]

Proof. For \( \vartheta(\cdot) \) as in (2.2) and \( \Gamma(\cdot, \cdot) \) as in (2.3) we take

\[
\mathcal{A}_n := E_n \cap \left\{ \sum_{i=1}^{\vartheta(|A|)} O_{n, n-i+1}/ \sum_{z \in [n]} O_z - \vartheta(|A|) \leq 4\Gamma(2\varepsilon, p^{out}(I))/r^{in} + 2\varepsilon \right\},
\]

where \( O_{n,1} \leq O_{n,2} \leq \cdots \leq O_{n,n} \) are the order statistics for \( O_1, \ldots, O_n \). In order to prove \( p^{in, out}_{\mathcal{A}_n}(\mathcal{A}|E_n) = 1 - o(1) \), first recall that \( p^{in} \) and \( p^{out} \) have the same mean \( r^{in} \). So using Chebyshev inequality, \( p^{in, out}_{\mathcal{A}_n}(\mathcal{A}|E_n) = O(1/n) \), and hence

\[
(2.4) \quad p^{in, out}_{\mathcal{A}_n}(\mathcal{A}|E_n) = O(1/n) = o(1),
\]

because \( p^{in, out}_{\mathcal{A}_n}(E_n) = O(1/\sqrt{n}) \) by the local central limit theorem.

Now we apply Theorem 1 of [28] for the function

\[
J(t) = \begin{cases} 
0 & \text{for } 0 \leq t \leq 1 - 2\varepsilon \\
1 & \text{for } 1 - \varepsilon \leq t \leq 1 \\
(t - 1)/\varepsilon + 2 & \text{for } 1 - 2\varepsilon \leq t \leq 1 - \varepsilon 
\end{cases}
\]

and the i.i.d. random variables \( O_1, \ldots, O_n \). Since \( p^{out} \) has finite second moment, it can be checked easily that the quantity given in (10) of [28], \( \sigma^2(J, p^{out}) \) is finite. This together with
Theorem 4 of [28] implies
\[ E_{\infty} (S_n - \mu)^2 = O(1/n), \]
where \( S_n := \frac{1}{n} \sum_{i=1}^{n} J(i/(n + 1))O_{n,i} \) and \( \mu := \int_{0}^{1} J(t)(p_{\text{out}})^*(t) \, dt \)
is as in (11) of [28]. Note that
\[ \int_{1-\varepsilon}^{1} (p_{\text{out}})^*(t) \, dt \leq \mu \leq \int_{1-2\varepsilon}^{1} (p_{\text{out}})^*(t) \, dt, \]
which means \( \Gamma(\varepsilon, p_{\text{out}}) \leq \mu \leq \Gamma(2\varepsilon, p_{\text{out}}). \)

Consequently, using Chebyshev inequality
\[
P_{\infty} (S_n > 2\mu) = O(1/n), \quad \text{which implies}
\]
\[
P_{\infty} \left( \frac{1}{n} \sum_{i=1}^{n} O_{n,n-i+1} > 2\Gamma(2\varepsilon, p_{\text{out}}) \right) = O(1/n), \quad \text{and consequently}
\]
\[
P_{\infty} \left( \frac{1}{n} \sum_{i=1}^{n} O_{n,n-i+1} > 2\Gamma(2\varepsilon, p_{\text{out}}) \right) e_n = o(1),
\]
as \( p_{\infty} (E_n) = O(1/\sqrt{n}). \) Combining the last display with (2.4) and noting that \( \vartheta(|A|) \leq \varepsilon n, \)
we see that \( p_{\infty} (\mathcal{A}) = 1 - o(1). \)

(1). It is easy to see from the construction of \( \mathcal{G}_n \) that if we write the labels of the sites in
\[ \bigcup_{t=1}^{K_{\mathcal{G}} A} \mathbb{Z}_t A \setminus R_t A ^{\text{in,out}} \]
in an increasing order, then a collision can occur at the \( k \)-th site (in this ordering)
with probability \( \leq \sum_{i=1}^{|A|+k-1} O_{n,n-i+1} / [\sum_{z} O_{z} - k]. \)
Therefore, for any \( 1 \leq t \leq K_{|A|} \) and \( x \in \mathbb{Z}_t A \setminus R_t A ^{\text{in,out}} \)
\[
P_{2,1,0} (x \in C_t A ^{\text{in,out}}) \leq \frac{\sum_{i=1}^{\vartheta(|A|)} O_{n,n-i+1} \sum_{z} O_{z} - \vartheta(|A|)}{
\sum_{z} O_{z} - \vartheta(|A|)}
\]
as \( |\bigcup_{t=1}^{K_{|A|}} \mathbb{Z}_t A \setminus R_t A ^{\text{in,out}}| \leq \vartheta(|A|) - |A|. \)
So the assertion follows from the definition of \( \mathcal{A} \).

(2). We can imitate the argument of Theorem 3.1.2 of [10] to see that under \( P_{2,1,0} \)
the number of self-loops and multiple edges are asymptotically independent, and both of them have asymptotic
Poisson distribution whose means are functions of the moments \( \sum_{k,l} k^l \mu_{k,l} p_{\text{out}}^{\text{in,out}}, i, j \in \{0, 1, 2\}. \)
So \( P_{2,1,0} (F_n) \) has a positive limit. This together with the fact that
\[
(Y_s^i, 1 \leq i \leq I_x, x \in [n]) \overset{d}{=} \left( \tilde{Y}_s^i, 1 \leq i \leq I_x, x \in [n] \right) \sum_{x=1}^{n} \sum_{i=1}^{I_x} 1_{\tilde{Y}_s^i = x} = O_x \forall z \in [n].
\]
gives the desired inequality.

(3). For a vector of positive integers \( y \), we write \( X_i (y) \) for the number of components of \( y \) which
are \( i \). We also write \( \bar{Y} = (\bar{Y}_1, \ldots, \bar{Y}_N) \) and \( \bar{Y}_{a:b} = (\bar{Y}_a, \bar{Y}_{a+1}, \ldots, \bar{Y}_b) \).
For any \( y \in [n]^{\mathbb{N}}, \)
\[
P \left( \bar{Y}_{1:x} = y \left| X_z (\bar{Y}) = N \alpha_z \forall z \in [n] \right. \right) = \frac{P \left( X_z (\bar{Y}_{x+1} N) = N \alpha_z - X_z (y) \forall z \in [n] \right)}{P \left( X_z (\bar{Y}) = N \alpha_z \forall z \in [n] \right)}.
\]
In order to bound the fraction in (2.5) recall that

\[
\text{Multinomial}(N; \alpha_1, \ldots, \alpha_n) \overset{d}{=} \left( Z_1, \ldots, Z_n \middle| \sum_{i=1}^{n} Z_i = N \right),
\]

where \( \{Z_i\}_{i=1}^{n} \) are independent and \( Z_i \sim \text{Poisson}(N\alpha_i) \). In that case, it is not hard to check that \( P(\sum_{i=1}^{n} Z_i = N) = (1 + o(1))/\sqrt{2\pi N} \) by Stirling’s formula. So, the ratio in (2.5) is

\[
(1 + o(1)) \sqrt{1 - \varepsilon} \prod_{z \in [n]} \frac{P(Y_z = N\alpha_z - X_z(y))}{P(Z_z = N\alpha_z)},
\]

where \( Y_i \sim \text{Poisson}(N(1 - \varepsilon)\alpha_i), Z_i \sim \text{Poisson}(N\alpha_i), i \in [n], \) and they are independent. The expression in the last display equals

\[
(1 + o(1)) \sqrt{1 - \varepsilon}[e^{\varepsilon}(1 - \varepsilon)]^N \exp(-\log(1 - \varepsilon)\varepsilon N) \prod_{z=1}^{n} \prod_{i=1}^{\alpha_z} \left(1 - \frac{i - 1}{N\alpha_z}\right).
\]

Using the inequality \( 1 - \varepsilon \leq e^{-\varepsilon} \) we get the desired bound.

(4). Using the bound of part (3), the total variation distance between these two measures is

\[
\frac{1}{2} \sum_{y \in [n]} \left| P\left(\tilde{Y}_{1 \leq N} = y \middle| X_z(\tilde{Y}_{1:N}) = N\alpha_z \forall z \in [n]\right) - 1 \right|
\]

\[
\leq 1 + \exp(-\varepsilon \log(1 - \varepsilon)N)\left| P\left(\tilde{Y}_{1 \leq N} \in A^c \right) \right|
\]

\[
+ \sup_{y \in A^c} \left| P\left(\tilde{Y}_{1 \leq N} = y \middle| X_z(\tilde{Y}_{1:N}) = N\alpha_z \forall z \in [n]\right) - 1 \right|
\]

for any set \( A \). Now recall from (2.8) that

\[
P\left(\tilde{Y}_{1 \leq N} = y \middle| X_z(\tilde{y}) = N\alpha_z \forall z \in [n]\right)
\]

\[
= (1 + o(1)) \sqrt{1 - \varepsilon}[e^{\varepsilon}(1 - \varepsilon)]^N \exp(-\log(1 - \varepsilon)\varepsilon N) \prod_{z=1}^{n} \prod_{i=1}^{\alpha_z} \left(1 - \frac{i - 1}{N\alpha_z}\right)
\]

Since \( e^{\varepsilon} \geq 1 - \varepsilon \), the first term in the right hand side of (2.10) lies between \( 1 - \varepsilon^2 N \) and 1, whereas the second term lies between \( \exp(\varepsilon^2 N) \) and \( \exp(2\varepsilon^2 N) \) when \( \varepsilon > 0 \) is small. Also the product term in (2.10) lies between 1 and

\[
1 - \sum_{z=1}^{n} \sum_{i=1}^{\alpha_z} \frac{i - 1}{N\alpha_z} = 1 - \sum_{z=1}^{n} \frac{X_z(y)(X_z(y) - 1)}{2N\alpha_z}.
\]
Consequently, if we take

\[ A_\eta := \left\{ y \in [n]^{\varepsilon N} : \frac{1}{2N} \sum_{z=1}^{n} X_z(y)(X_z(y) - 1) < \eta^{2}N \right\}, \]

then

\[
\left| \frac{P(\tilde{Y}_{1:}\varepsilon N = y|X_z(\tilde{Y}_{1:}\varepsilon N) = N \alpha_z \forall z \in [n]) - 1}{P(\tilde{Y}_{1:}\varepsilon N = y)} \right| = O(\varepsilon^{2}N)
\]

whenever \( y \in A_\eta \). So, in view of (2.9), it suffices to show that \( P(\tilde{Y}_{1:}\varepsilon N \in A_\eta^c) = O(\exp(-C\varepsilon N)) \) for some constant \( C > 0 \) and for some suitable choice of \( \eta \).

Note that the joint distribution of \( \{X_z(\tilde{Y}_{1:}\varepsilon N)\}_{z \in [n]} \) is \( \text{Multinomial}(\varepsilon N; \alpha_1, \ldots, \alpha_n) \), so using (2.6) and local central limit theorem

\[
P(\tilde{Y}_{1:}\varepsilon N \in A_\eta^c) = (1 + o(1))\sqrt{2\pi N}P \left( \sum_{i=1}^{n} \frac{\hat{Y}_i - 1}{\alpha_i} \geq \eta^{2}N \right),
\]

where \( \{\hat{Y}_i\}_{i=1}^{n} \) are independent and \( \hat{Y}_i \sim \text{Poisson}(\varepsilon N\alpha_i) \). Hence using standard large deviation argument, the above probability is at most \( \exp(-C(\eta)\varepsilon N) \) for some constant \( C(\eta) \) such that \( C(\eta) > 0 \) when \( \eta \) is large enough. This completes the argument.

\[ \blacksquare \]

**Remark 2.3.** The assertions (3) and (4) of Lemma 2.2 is true if we replace the index set \( \{1, 2, \ldots, \varepsilon N\} \) by (possibly random) \( \{i_1, i_2, \ldots, i_{\varepsilon N}\} \).

3. **Ingredients.** In this section, we will state and prove some of the basic lemmas which will be required in proving our main results.

**Lemma 3.1.** Let \( X \) be any nonnegative random variable such that \( 2(EX)^2 \leq EX^2 < \infty \). Then \( \log Ee^{-tX} \leq \text{var}(X)t^2/2 - E(X)t \) for any \( t > 0 \).

**Proof.** Let \( \mu = EX \) and \( \mu_2 = \sqrt{EX^2} \) so that \( \sigma^2 = \text{var}(X) = \mu_2^2 - \mu^2 \). We choose \( p = \mu^2/\mu_2^2 \) and \( \alpha = \mu_2^2/\mu \) so that \( Y := (1 - p)\delta_0 + p\delta_\alpha \) satisfies \( EY = \mu \) and \( EY^2 = \mu_2^2 \). By Benette’s inequality [5],

\[
\text{for any } t > 0, \log Ee^{-tX} \leq \log Ee^{-tY} = \log[(1 - p) + pe^{-at}] =: \varphi(t).
\]

Differentiating the function \( \varphi \) and noting that \( \rho \alpha = \mu \) and \( \mu \alpha = \mu_2^2 \) we get

\[
\varphi'(t) = \frac{-\rho e^{-at}}{(1 - p) + pe^{-at}} = \frac{\mu}{(1 - p)e^{at} + p}, \quad \varphi''(t) = \sigma^2 \frac{e^{at}}{(1 - p)e^{at} + p}.
\]

Also note that the quadratic function \( f(x) = [(1 - p)x + p]^2 - x \) has nonnegative slope at \( x = 1 \) if \( 2(1 - p) \geq 1 \), which is true by our hypothesis. So \( \varphi''(t) \leq \sigma^2 \) for any \( t \geq 0 \). Finally using Taylor series expansion for the function \( \varphi \) we see that for any \( t > 0 \),

\[
\varphi(t) = \varphi(0) + \varphi'(0)t + \varphi''(u)t^2/2 \text{ for some } u \in [0, t]
\]

\[ \leq -\mu t + \sigma^2 t^2/2. \]

This inequality together with (3.1) gives the desired result.

\[ \blacksquare \]
Lemma 3.2. For any $\kappa > 0$ and $\Delta \geq 1$ the function $\phi_{\kappa,\Delta}(\gamma) := \gamma[\log(\Delta/\gamma)]^\kappa$ is increasing for $\gamma \leq \Delta e^{-\kappa}$ and decreasing for $\gamma \geq \Delta e^{-\kappa}$. Hence $\phi_{\kappa,\Delta}(\gamma) \leq \Delta(\kappa/e)^\kappa$.

Proof. We get the conclusion using elementary method.

Lemma 3.3. For $\delta > 0$, $\vartheta(m) := \beta_1 m[\log(n/m)]^{\beta_2}, 0 < \gamma \leq 1$ and any integer $\Lambda \geq 1$ there is an $\epsilon_{3.3} > 0$ depending on $\Lambda, \beta_1, \beta_2, \gamma$ such that $m \leq \epsilon_{3.3} n$ and $M \in \mathbb{N}$ imply

$$P\left(\text{Binomial}(\Lambda M, (\vartheta(m)/n)^\gamma) \geq \frac{1}{\gamma}(1 + \delta)M\right) \leq \exp(-(1 + \delta/2)M \log(n/m)).$$

Proof. A standard large deviations result for the Binomial distribution, see e.g., Lemma 2.8.4 in [10] implies $P(\text{Binomial}(\Lambda M, q) \geq \Lambda Mr) \leq \exp(-\Lambda MH_q(r))$ for any $r > q$, where

$$H_q(r) := r \log \left(\frac{r}{q}\right) + (1 - r) \log \left(\frac{1 - r}{1 - q}\right).$$

When $r = (1 + \delta)/(\gamma \Lambda)$, the first term in the large deviation bound (3.2) is

$$\exp(-\Lambda M r \log(r/q)) \leq \exp\left(-\frac{1}{\gamma}(1 + \delta)M \left[\log\left(\frac{n}{m}\right)^\gamma - \log\frac{\Lambda \gamma \beta_1}{1 + \delta} - \beta_2 \gamma \log \log \frac{n}{m}\right]\right).$$

For the second term in the large deviation bound in (3.2) we note that $1/(1 - q) > 1$ and $(1 - r) \log(1 - r) \geq -1/e$ by Lemma 3.2 (with $\kappa = \Delta = 1$), and conclude

$$\exp\left(-\Lambda M(1 - r) \log\left(\frac{1 - r}{1 - q}\right)\right) \leq \exp(-\Lambda M(1 - r) \log(1 - r)) \leq \exp(\Lambda M/e).$$

Combining the last two estimates

$$P(\text{Binomial}(\Lambda M, (\vartheta(m)/n)^\gamma) > \frac{1}{\gamma}(1 + \delta)M)$$

$$\leq \exp\left(-(1 + \delta)M \log(n/m) + \beta_4 M + \beta_5 M \log \log \frac{n}{m}\right),$$

for constants $\beta_4$ and $\beta_5$.

Now we choose

$$\epsilon_{3.3} := \max\left\{\epsilon \in (0, e^{-2\beta_5/\delta}) : \epsilon \left[\log \frac{1}{\epsilon}\right]^{2\beta_5/\delta} \leq \exp(-2\beta_4/\delta)\right\}.$$ 

Clearly $\epsilon_{3.3} > 0$ and, in view of Lemma 3.2 with $\kappa = 2\beta_5/\delta$ and $\Delta = 1$, $m \leq \epsilon_{3.3} n$ implies

$$\left(\frac{n}{m}\right)^{2\beta_5/\delta} \leq \epsilon_{3.3} \left[\log \frac{1}{\epsilon_{3.3}}\right]^{2\beta_5/\delta} \leq \exp(-2\beta_4/\delta),$$

which in turn implies $\beta_4 + \beta_5 \log \log \left[\frac{n}{m}\right] \leq (\delta/2) \log(n/m)$. This completes the proof. 

$\blacksquare$
4. Choice of ‘good’ graphs $G_n$. For $B \subset A \subset [n]$, recall the definition of the forest $\{Z^A_t\}_{t=0}^{K[A]}$ (as described in Section 2.2) with associated subsets $\{O_t^{A,B}\}$ of ‘open’ sites. Let $\mathbf{p}$ be the limiting out-degree distribution for $\tilde{G}_n$, namely

$$
\mathbf{p} = \begin{cases} 
\mathbf{p}^\text{in} & \text{for RBN}^1 \\
\tilde{\mathbf{p}}^\text{in} = \left\{ (i^\text{in})^{-1} \sum_l l p_{k,l}^\text{in,\text{out}} \right\}_{k=2}^\infty & \text{for RBN}^2
\end{cases}
$$

with $p_0 = p_1 = 0$ and mean $r = \sum k k p_k > 2$.

**Proposition 4.1.** There are constants $c_1, c_2, \delta > 0$ such that for $q \in (1/r, 1/2)$, $K_m = c_1 \log_2(c_2 \log(n/m))$ and for $A \subset [n]$ if

$$
E_A := \bigcap_{B \subset A} \{B \cap [B] \ni (1-\delta)|A|\} \left\{ |O_{K[A]}^{A,B}| \geq (4/\delta) q^{-|A|} |B| \right\},
$$

then there is an $\epsilon_{4.1} > 0$ such that for any $a > 0$ the probability of

$$
G_n^1 := \bigcap_{A \subset [n]} \{ |O_n^{A}\} \leq \epsilon_{4.1} n \} E_A
$$

under $\mathbb{P}_{i,n}, i = 1, 2$, is $1 - o(1)$.

**Proof.** Let $\mathbf{p}_n = \{p_{n,k}\}_{k \geq 2}$ be the distribution

$$
\mathbf{p}_n := \begin{cases} 
\frac{1}{n^2} \sum_{z \in [n]} \delta_{I_z} & \text{for RBN}^1 \\
\sum_{z \in [n]} O_z \delta_{I_z} / \sum_{z \in [n]} O_z & \text{for RBN}^2
\end{cases}
$$

In view of Lemma 2.1 and 2.2, $\mathbf{p}_n$ approximates out-degree distribution for the graph $\tilde{G}_n$ with $p_{n,0} = p_{n,1} = 0$. Let $r_n := \sum k k p_{n,k} \in (2, \infty)$ be the mean of $\mathbf{p}_n$. It is easy to see that $r_n \to r$.

In this proof, we write $\mathbb{P}$ for the probability distribution on the forests $\{Z^A_t\}_{t \geq 0, A \subset [n]}$, when $\mathbf{p}_n$ is used as its offspring distribution. We also use $\tilde{\mathbb{P}}$ as a dummy replacement for $\mathbb{P}_{1.1}$ and $\mathbb{P}_{2.1.0}(|F_n|)$. Lemma 2.1 and 2.2 suggest that for any event $F$ involving the structure of the graph which depends on at most $\varepsilon n$ many vertices of the graph,

$$
\mathbf{p}_n^\text{in} \left\{ \left( I : \tilde{\mathbb{P}}(F) \leq C_{2.1} \mathbb{P}(F) \right) \right\} = 1 - o(1).
$$

Now fix $\eta \in (0, 1 - 1/qr)$, and

$$
\gamma := \begin{cases} 
1 & \text{for RBN}^1 \\
\frac{\alpha - 2}{\alpha - 1} & \text{for RBN}^2 \text{ when } p_{k}^\text{out} \sim c k^{-\alpha} \text{ and } \alpha > 3
\end{cases}
$$

When the tail of $\mathbf{p}^\text{out}$ is lighter than polynomial, then $\gamma$ is taken to be 1. Clearly $1/\gamma < 2 < r(1 - \eta)$, so we can choose $\delta \in (0, 1/10)$ such that $(1 + 5\delta)/(2\gamma) < 1$. We need to introduce some
more notations, let

\[ \tilde{r} := r_n(1 - \eta) \] so that \( q\tilde{r} > 1 \) for large enough \( n, \)

\( \rho > 0 \) be such that \( 2^{\rho - 1} \left( 1 - \frac{1 + 5\delta}{2\gamma} \right) \geq 1 \) and \( (q\tilde{r})^{\rho} \left( 1 - \frac{1 + 5\delta}{\gamma\tilde{r}} \right) > 1, \)

\( \sigma \geq 1 \) be such that \[ \left( (q\tilde{r})^{\rho} \left( 1 - \frac{1 + 5\delta}{\gamma\tilde{r}} \right) \right)^{\sigma} \geq r_n^{\rho}, \]

\( I_n(\eta) := \sup_{\theta} \left( \theta r (1 - \eta) - \log \left( \sum_k e^{\theta_k p_{n,k}} \right) \right) > 0 \) be the large deviation rate function for \( p_n \)

(4.4)

\[ k_m := \log_2 \left[ \frac{1 + 3\delta}{(\tilde{r} - \gamma^{-1}(1 + 5\delta))I_n(\eta)} \log \frac{n}{m} \right] \] and \( K_m = \rho \sigma k_m \) for \( m \leq n, \) so that

\[ \vartheta(m) := m + \sum_{l=1}^{K_m} (2r)^l \leq \beta_1 m[\log(n/m)]^{\beta_2} \]

for \( \beta_1 = \left( 1 + \frac{2r}{2r - 1} \right) \left( \frac{1 + 3\delta}{(\tilde{r} - \gamma^{-1}(1 + 5\delta))I_n(\eta)} \right)^{\rho \sigma \log_2(2r)} \) and \( \beta_2 := \rho \sigma \log_2(2r). \)

Suppose \( B \subset A \subset [n] \) are subsets such that \( |A| = m \) and \( |B| \geq (1 - \delta)m. \) For \( k \geq 1, \) define the events

\[ H_k^{A,B} := \left\{ \sum_{i=1}^{\rho} |C_{\rho(k-1)+i}| \leq \frac{1}{\gamma} (1 + 5\delta) |O_{\rho(k-1)}^{A,B}| \right\}, \]

\[ L_k^{A,B} := \left\{ |Z_{\rho(k-1)+j}| \geq \tilde{r} |O_{\rho(k-1)+j-1}^{A,B}| \right\} \] and \( L_k^{A,B} := \cap_{j=1}^{\rho} L_{k,j}^{A,B}. \)

Note that on the event \( H_k^{A,B}, \)

\[ |O_{\rho k}^{A,B}| \geq 2|O_{\rho(k-1)}^{A,B}| - |C_{\rho k}^{A,B}| \]

\[ \geq 2^2 |O_{\rho(k-2)}^{A,B}| - 2|C_{\rho k}^{A,B}| \]

\[ \geq \ldots \]

\[ \geq 2^\rho |O_{\rho(k-1)}^{A,B}| - \sum_{i=1}^{\rho} 2^{\rho-i} |C_{\rho(k-1)+i}^{A,B}| \]

\[ \geq 2^\rho |O_{\rho(k-1)}^{A,B}| - 2^{\rho-1} \sum_{i=1}^{\rho} |C_{\rho(k-1)+i}^{A,B}| \]

(4.5)

\[ \geq (2^\rho - 2^{\rho-1} \gamma^{-1}(1 + 5\delta)) |O_{\rho(k-1)}^{A,B}| \geq 2|O_{\rho(k-1)}^{A,B}|, \]

by the choice of \( \rho. \) Since \( |O_{t}^{A,B}| \leq (2r)|O_{t-1}^{A,B}| \) for any \( t \geq 1, \) a similar argument which leads to the
previous display suggests that the following inequalities are true on the event $H^A_B k \cap \bigcap_{j=1}^i L^A_B$.

\[
\begin{align*}
|O^A_B_{\rho(k-1)+i}| & \geq \tilde{r}^i |O^A_B_{\rho(k-1)+i-1} - |C^A_B\rho(k-1)+i| | \geq \tilde{r}^2 |O^A_B|_{\rho(k-1)+i-2} - \tilde{r} |C^A_B\rho(k-1)+i| | \geq \cdots \\
& \geq \tilde{r}^i |O^A_B_{\rho(k-1)}| | - \sum_{j=1}^i \tilde{r}^j |C^A_B\rho(k-1)+j| \\
& \geq \tilde{r}^i |O^A_B_{\rho(k-1)}| | - \tilde{r}^{i-1} \sum_{j=1}^i |C^A_B\rho(k-1)+j| \\
& \geq (\tilde{r}^i - \tilde{r}^{i-1} \gamma^{-1}(1 + 5\delta)) |O^A_B_{\rho(k-1)}| \\
& \geq \tilde{r} \left(1 - \frac{1 + 5\delta}{\gamma \tilde{r}}\right) |O^A_B_{\rho(k-1)}|,
\end{align*}
\]

(4.6) (4.7)

Taking $i = \rho$ in (4.6),

\[
|O^A_B_{\rho k}| \geq \tilde{r}^\rho \left(1 - \frac{1 + 5\delta}{\gamma \tilde{r}}\right) |O^A_B_{\rho(k-1)}| \text{ on the event } H^A_B k \cap L^A_B.
\]

Recalling $K_m = \rho \sigma k_m$ and using (4.5) and (4.8) repeatedly,

\[
|O^A_B_{K_m}| \geq |B| 2^{km} \left[\tilde{r}^\rho \left(1 - \frac{1 + 5\delta}{\gamma \tilde{r}}\right)\right]^{(\sigma-1)k_m} \text{ on the event } \cap_{k=1}^{k_m} H^A_B k \cap \bigcap_{k=k_m+1}^{\infty} L^A_B.
\]

Now note that

\[
q^{\rho \sigma} \left[\tilde{r}^{\rho \sigma} \left(1 - \frac{1 + 5\delta}{\gamma \tilde{r}}\right)\right]^{\sigma-1} \geq \left[(q\tilde{r})^{\rho \sigma} \left(1 - \frac{1 + 5\delta}{\gamma \tilde{r}}\right)\right]^{\sigma} r^{-\rho} \geq 1
\]

by the choices of $\rho$ and $\sigma$. So if

\[
(m/n) \leq \exp \left(-\frac{4(\tilde{r} - \gamma^{-1}(1 + 5\delta)) I_n(\eta)}{\delta(1 + 3\delta)}\right)
\]

so that $2^{km} \geq (4/\delta)$, then

\[
|O^A_B_{K_m}| \geq |B|(4/\delta) q^{-K_m} \text{ on the event } \cap_{k=1}^{k_m} H^A_B k \cap \bigcap_{k=k_m+1}^{\infty} L^A_B.
\]

(4.9) (4.10)

To estimate the probability of the event in (4.10) recall that $|C^A_{t+1}| / |O^A_{t+1}| \leq (2r)|O^A_t|$ for any $t \geq 0$ and by Lemma 2.1 and 2.2 each site is included in $C^A_{t+1} \subset C^A_t$ with probability at most $c(\vartheta(m)/n)^\gamma$. So for any $k \geq 1$, $\sum_{i=1}^r |C^A_{\rho(k-1)+i} |$ conditionally on $|O^A_{\rho(k-1)}|$ is stochastically dominated by the Binomial($\Lambda M$, $c(\vartheta(m)/n)^\gamma$) distribution, where $\Lambda = (2r) + (2r)^2 + \cdots + (2r)^\rho$ and $M = |O^A_{\rho(k-1)}|$. Hence, applying Lemma 3.3 with the above choices of $\Lambda$ and $M$ if

\[
m \leq \epsilon_{3.3}(\Lambda, 5\delta, \eta, \gamma)n,
\]

then

\[
\mathbb{P} \left\{ (H^A_B)^c \mid |O^A_B_{\rho(k-1)}| \right\} \leq \exp \left(-(1 + 5\delta/2)|O^A_B_{\rho(k-1)}| \log(n/m)\right).
\]
Since $|O^A_{\rho(k-1)}| \geq |B|$ on the event $\cap_{j=1}^{k-1} H^A_j$ by (4.5), the above inequality reduces to

$$\mathbb{P}\left((H^A_k)^c \cap \bigcap_{j=1}^{k-1} H^A_j\right) \leq \exp\left(-(1 + 5\delta/2)|B| \log(n/m)\right)$$

(4.12)

The last inequality follows from the fact that $|B| \geq (1 - \delta)m$ and $\delta \in (0, 1/10)$, which makes $(1 + 5\delta/2)(1 - \delta) \geq 1 + \delta$.

By the choice of $I_n(\eta)$, a standard large deviation argument for the sum of i.i.d. random variables yields

$$\mathbb{P}\left((I^A_{k,i})^c \cap |O^A_{\rho(k-1)+i-1}| \leq \exp\left(-|O^A_{\rho(k-1)+i-1}| I_n(\eta)\right)$$

for any $k \geq 1$ and $1 \leq i \leq \rho$. Now repeated applications of the inequality in (4.5) suggest that $|O^A_{\rho k_m}| \geq 2^{k_m}|B|$ on the event $\cap_{j=1}^{k_m} H^A_j$. In view of (4.5) and (4.7), for any $k > k_m$ and $1 \leq i \leq \rho$,

$$|O^A_{\rho(k-1)+i-1}| \geq (\tilde{r} - \gamma^{-1}(1 + 5\delta))|O^A_{\rho(k-1)}| \geq (\tilde{r} - \gamma^{-1}(1 + 5\delta))|O^A_{\rho k_m}|$$

on the event $\cap_{j=1}^{k_m+1} H^A_j \cap_{j=1}^{k-1} L^A_{k,j}$. So the inequality in the last display reduces to

$$\mathbb{P}\left((I^A_{k,i})^c \cap \bigcap_{j=1}^{k_m} L^A_{k,j} \cap \bigcap_{j=1}^{k-1} H^A_j\right) \leq \exp\left(-(1 + 3\delta)|B| \log(n/m)\right) \leq \exp(-(1 + \delta)m \log(n/m)).$$

(4.13)

The last two inequalities follow from the definition of $k_m$ and the facts that $|B| \geq (1 - \delta)m$, which implies $(1 + 3\delta)|B| \geq (1 + \delta)m$ for $\delta \in (0, 1/10)$. Applying Lemma 3.2 with $\kappa = \Delta = 1,$

$$m \log(n/m) = n\phi_{1,1}(m/n) \geq n\phi_{1,1}(1/n) = \log n \text{ for } m \leq n/e.$$  

(4.14)

Combining (4.10), (4.12) and (4.13) if $m/n$ is small satisfying (4.9), (4.11) and (4.14), then

$$\mathbb{P}\left(|O^A_{K_m}| < |B|(4/\delta)q^{-K_m}\right) \leq \mathbb{P}\left((\cap_{k=1}^{\sigma k_m} H^A_k \cap \bigcap_{k=k_m+1}^{\rho} I^A_k)^c\right)$$

$$\leq \sum_{k=1}^{\sigma k_m} \mathbb{P}\left((H^A_k)^c \cap \bigcap_{j=1}^{k-1} H^A_j\right)$$

$$+ \sum_{k=k_m+1}^{\rho} \sum_{i=1}^{\sigma k_m} \mathbb{P}\left((I^A_{k,i})^c \cap \bigcap_{j=1}^{k-1} L^A_{k,j} \cap \bigcap_{j=1}^{k} H^A_j\right)$$

$$\leq [\sigma + \rho(\sigma - 1)]k_m \exp(-(1 + \delta)m \log(n/m))$$

$$\leq [\sigma + \rho(\sigma - 1)] \log_2[C \log n]$$

$$\exp(-(1 + 3\delta/4)m \log(n/m))n^{-\delta/4}$$

$$\leq \exp(-(1 + 3\delta/4)m \log(n/m)).$$

for large enough $n$. Since the event considered in the last display involves at most $\vartheta(m)$ vertices of the graph, the above estimate together with (4.3) implies

$$\tilde{\mathbb{P}}\left(|O^A_{K_m}| < |B|(4/\delta)q^{-K_m}\right) \leq \exp(-(1 + 3\delta/8)m \log(n/m)),$$
with \( p_{in}^{in}/p_{out}^{in.out} \) probability \( 1 - o(1) \) provided \( m/n \leq \varepsilon \) is small.

Using this estimate and union bound we see that if \( m/n \) is small, then

\[
\tilde{\mathbb{P}} \left( \bigcup_{A \in \{\{A \cap [n]: |A| = m\} E_A^{m} \} \right) \\
\leq \tilde{\mathbb{P}} \left( \bigcup_{m' \in \{([1-\delta)m, m\} \cup \{\{A,B\}: B \subseteq [A \cap [n]: |A| = m, |B| = m\}} \left\{ |O_{K_m}^{A,B} \leq (4/\delta) q^{-K_m} |B| \right\} \right) \\
\leq \sum_{m' \in \{([1-\delta)m, m\} \} \binom{n}{m} \binom{m}{m'} \exp \left( (1 + 3\delta/8) m \log \frac{n}{m} \right).
\]

(4.15)

It is easy to check that \( \left( \frac{l}{m} \right) \leq \frac{L}{\delta \pi} \leq (L \varepsilon / \delta l)^l \) for any positive integers \( l \leq L \) and the function \( \phi_{1,\varepsilon} (\cdot) \) defined in Lemma 3.2 is increasing on \((0, 1)\). So for \( m' \geq (1 - \delta)m \),

\[
\binom{n}{m} \leq \left( \frac{ne}{m} \right)^m \text{ and } \binom{m}{m'} \leq \left( \frac{me}{m-m'} \right)^{m-m'} = \exp \left[ m \phi_{1,\varepsilon} \left( \frac{m-m'}{m} \right) \right] \\
\leq \exp (m \phi_{1,\varepsilon}(\delta)) \leq (e/\delta)^{\delta m}.
\]

Also there are at most \( m \leq e^m \) choices for \( m' \). Using these bounds the right hand side of (4.15)

\[
\leq \exp [m + m \log (ne/m) + m \delta \log (e/\delta) - (1 + 3\delta/8) m \log (n/m)] \\
\leq \exp [-3(\delta/8) m \log (n/m) + \Delta_1 m]
\]

for some constant \( \Delta_1 \). If \( m/n \leq \exp (-8\Delta_1/\delta) \), then the right hand side of the last display is

\[
\leq \exp [-(\delta/4) m \log (n/m)].
\]

Therefore, if \( \epsilon_{4.1} \) is chosen small enough, then for any \( m \leq \epsilon_{4.1} n \),

\[
\tilde{\mathbb{P}} \left( \bigcup_{A \in \{\{A \cap [n]: |A| = m\} E_A^{m} \} \right) \leq \exp [-(\delta/4) m \log (n/m)].
\]

Combining this with the fact that \( m \mapsto m \log (n/m) \) is increasing for \( m \leq n/e \) (by Lemma 3.2),

\[
\tilde{\mathbb{P}} ((\mathcal{G}_n^{1} \cap \mathcal{G}_n^{2} ) \leq \sum_{m \in [(\log n)^a, \epsilon_{4.1} \epsilon_{4.1} n]} \tilde{\mathbb{P}} \left( \bigcup_{A \in \{\{A \cap [n]: |A| = m\} E_A^{m} \} \right) \leq \sum_{m \in [(\log n)^a, \epsilon_{4.1} \epsilon_{4.1} n]} \exp [-(\delta/4) (\log n)^a \log (n/\log n)^a)] \\
\leq n \exp [-(\delta/4) (\log n)^{1+\alpha} (1 + o(1))] = o(1/\sqrt{n}).
\]

This together with (4.3) completes the proof. 

Recall the definition of \( \pi (\cdot, \cdot) \) from (1.11) and let \( CL_x^k := \bigcup_{t=0}^k \tilde{Z}_l^{(x)} \) be the oriented cluster of depth \( k \) starting from \( x \in [n] \) in the graph \( \tilde{\mathbf{G}}_n \). For any sequence \( \{t_n\} \) as in the statement of our theorems, define the events

\[
A_x := \left\{ |\tilde{Z}_l^{(x)}| 2a \log \log n/ \log (q\xi) \geq (\log n)^a \right\}, \quad \bar{A}_x := \left\{ |\tilde{Z}_l^{(x)}| t_n/2a \log \log n/ \log (q\xi) \neq \emptyset \right\},
\]

\[
A_{x,y} := \left\{ CL_x^k 2a \log \log n/ \log (q\xi) \cap CL_y^k 2a \log \log n/ \log (q\xi) = \emptyset \right\}.
\]

**Proposition 4.2.** For \( \pi (\cdot, \cdot) \) as in (1.11), \( p \) as in (4.1), \( q > 1/r \) and \( \varepsilon > 0 \) let \( \pi := \pi (p, q) \) and

\[
\mathcal{G}_n^2 := \left\{ \mathcal{G}_n : n(\pi - \varepsilon) \leq \sum_{x \in [n]} \mathbb{P}_{\mathcal{G}_n, q}(A_x), \sum_{x \in [n]} \mathbb{P}_{\mathcal{G}_n, q}(\bar{A}_x) \leq n(\pi + \varepsilon) \right\} \cap \left\{ \sum_{x,y \in [n], x \neq y} 1_{A_{x,y}^n} \leq n^{9/5} \right\}.
\]

Then \( \mathbb{P}_{1,n}(\mathcal{G}_n^2) = 1 - o(1) \) for \( i = 1, 2 \).
PROOF. In this proof also the notations $P$ and $\hat{P}$ serve the same purpose as they did in the proof of Proposition 4.1. $\mathbb{E}$ and $\hat{\mathbb{E}}$ denote the corresponding expectations. Also let $s_n = 2a \log \log n / \log (q \bar{r})$, $t_n$ and $\log \log n / \log (q \bar{r})$.

First we note that if

$$C_x := \left\{ \left| CL_{2a \log \log n / \log (q \bar{r})}^x \right| \leq n^{1/4} \right\},$$

then $P(C_x), \hat{P}(C_x) \geq 1 - n^{-1/8}$,

using Markov inequality. This bound together with (4) of Lemma 2.2 with $\varepsilon = n^{-3/4}$ implies (4.17)

$$\left| \mathbb{E}(P_{g_n,q}(A_x)) - \hat{\mathbb{E}}(P_{g_n,q}(A_x)) \right| = o(1) + \left| \mathbb{E}(P_{g_n,q}(A_x)) \mathbf{1}_{C_x} - \hat{\mathbb{E}}(P_{g_n,q}(A_x)) \mathbf{1}_{C_x} \right| = o(1).$$

Now if $B_x$ denotes the event that collision does not occur in the cluster $CL_{2a \log \log n / \log (q \bar{r})}$, then combining (4.16) with Lemma 2.1 and 2.2,

$$P(B_x^c) = o(1) + P(B_x^c \cap C_x) = o(1) + \hat{P}(B_x^c \cap C_x) = o(1).$$

On the event $B_x$, the law of $\xi_{X}^{(x)}$, $0 \leq t \leq 2a \log \log n / \log (q \bar{r})$ under the annealed measure $P \times P_{\xi_{t}}^{g_n,q}$ is same as that of a branching process with offspring distribution $(1 - q)\delta_0 + q\nu_n$, where $\nu_n$ is as in (4.2). So if $\{Z_t\}_{t \geq 0}$ is such a branching process with $Z_0 = 1$, then its survival probability is $\pi(p_n, q)$ and using (4.18),

$$\mathbb{E}[P_{g_n,q}(A_x)] = P \left( Z_n > (q \bar{r})^{s_n/2} \right) + o(1).$$

Imitating the proof of Proposition 2.1 in [20], the right hand side is $\pi(p_n, q) + o(1)$. This together with (4.17) and the fact that $\pi(p_n, q) \to \pi$ as $n \to \infty$ implies (4.19)

$$\hat{\mathbb{E}}[P_{g_n,q}(A_x)] = \pi + o(1),$$

using similar argument $\hat{\mathbb{E}}[P_{g_n,q}(\tilde{A}_x)] = \pi + o(1)$.

Also using (4.16) and Lemma 2.1 and 2.2

$$\hat{P}(A_{x,y}^c) \leq \hat{P}(C_x^c) + \hat{P}(C_y^c) + \hat{P}(A_{x,y}^c \cap C_x \cap C_y)$$

$$\leq 2n^{-1/8} + cn^{1/4} \left( \frac{n^{1/4}}{n} \right)^{\alpha - 2}/(\alpha - 1) \leq cn^{-1/8}$$

for some constant $c > 0$. Using the last inequality and following the argument which leads to (2.13) in [6], if $x_1, x_2 \in [n]$ are such that $x_1 \neq x_2$, then

$$\hat{\mathbb{E}}[P_{g_n,q}(A_{x_1})P_{g_n,q}(A_{x_2})] - \hat{\mathbb{E}}[P_{g_n,q}(A_{x_1})]\hat{\mathbb{E}}[P_{g_n,q}(A_{x_2})]$$

$$\leq \hat{P}(A_{x_1,x_2}^c)[1 + 1/\hat{P}(A_{x_1,x_2})] \leq cn^{-1/8}.$$
and hence combining with (4.20) and using the standard second moment argument,

\[
\tilde{P} \left( \sum_{x,y \in [n], x \neq y} 1_{A_{x,y}} > n^{-1/10} \frac{n}{2} \right) = o(1).
\]

This completes the proof. ■

5. Proofs of the Theorems. Let \( \mathcal{F} \) be a forest consisting of \( m \) rooted directed trees and let \( \mathcal{F}_{k,i} \) denote the set of vertices of the \( i \)-th tree which are at oriented distance \( k \) from the root level and \( \mathcal{F}_k = \bigcup_{i=1}^m \mathcal{F}_{k,i} \).

**Lemma 5.1.** If \( P_{\xi, q} \) denotes the law of \( \{ \xi_i^A, A \subset \mathcal{F}_0, t \geq 0 \} \) on the directed forest \( \mathcal{F} \) and if \( |\mathcal{F}_k| \geq 2q^{-k}|\mathcal{F}_0| \), then

\[
P_{\xi, q} \left( \left\{ \xi_k \right\} \leq |\mathcal{F}_0| \right) \leq \exp \left( -cq^k|\mathcal{F}_k|^2/\sum_{i=1}^m |\mathcal{F}_{k,i}|^2 \right).
\]

**Proof.** Let \( m = |\mathcal{F}_0| \). For \( x \in \mathcal{F}_k \) let \( Y_x := \{ x \in \xi_k \} \) and for \( 1 \leq i \leq m \) let \( N_i := \sum_x Y_x 1 \{ x \in \mathcal{F}_{k,i} \} \). It is easy to see that if \( l(x, y) \) equals half of the distance between \( x \) and \( y \) in the forest ignoring the orientation of the edges, then

\[
E_{\xi, q} Y_x = q^k, \quad E_{\xi, q} (Y_x Y_{y}) = \begin{cases} 
q^k & \text{if } x = y \\
q^{k+l(x,y)-1} & \text{if } 1 \leq l(x, y) \leq k, \text{ so that} \\
q^{2k} & \text{otherwise}
\end{cases}
\]

\[
E_{\xi, q} N_i = q^k |\mathcal{F}_{k,i}|, \quad E_{\xi, q} N_i^2 = \sum_{x, z \in \mathcal{F}_{k,i}} E_{\xi, q} (Y_x Y_z) \in \left[q^{2k-1}|\mathcal{F}_{k,i}|^2, q^k|\mathcal{F}_{k,i}|^2\right].
\]

By our hypothesis, \( \sum_{i=1}^m E_{\xi, q} N_i \geq 2m \). So applying Lemma 3.1

\[
P_{\xi, q} \left( \left\{ \xi_k \right\} \leq m \right) = P_{\xi, q} \left( \sum_{i=1}^m N_i \leq m \right) \leq \exp \left( tm + \sum_{i=1}^m \log E_{\xi, q} e^{-tN_i} \right)
\]

\[
\leq \exp \left( tm + \sum_{i=1}^m [-tE_{\xi, q} N_i + (E_{\xi, q} N_i^2 - [E_{\xi, q} N_i]^2) t^2/2] \right)
\]

\[
\leq \exp \left( -\sum_{i=1}^m \left(t/2\right) E_{\xi, q} N_i - \left(E_{\xi, q} N_i^2 - [E_{\xi, q} N_i]^2\right) t^2/2 \right)
\]

for any \( t \geq 0 \). Optimizing the last expression with respect to \( t \) and noting that \( at - bt^2/2 \leq a^2/2b \) for any \( a, b > 0 \) we have

\[
(5.1) \quad P_{\xi, q} \left( \left\{ \xi_k \right\} \leq m \right) \leq \exp \left( -\frac{\left(\sum_{i=1}^m E_{\xi, q} N_i\right)^2}{8 \sum_{i=1}^m (E_{\xi, q} N_i^2 - [E_{\xi, q} N_i]^2)} \right)
\]

\[
(5.2) \quad \leq \exp \left( -\frac{q^{2k}}{8} \frac{\left(\sum_{i=1}^m |\mathcal{F}_{k,i}|\right)^2}{\sum_{i=1}^m |\mathcal{F}_{k,i}|^2} \right).
\]
PROPOSITION 5.2. Let $\epsilon_{4.1}$ and $G^1_n$ be as in Proposition 4.1 and $K_m$ be as in (4.4). There are constants $C_{5.2}, b > 0$ such that if $\mathcal{G}_n \in G^1_n$ and $A \subset [n]$ has size $m \leq \epsilon_{4.1} n$, then
\[ P_{\mathcal{G}_n,q} \left( \left| \xi_{K_m} \right| \leq m \right) \leq \exp \left( -C_{5.2} m (\log(m/n))^{-b} \right) . \]

PROOF. For $\mathcal{G}_n \in G^1_n$, any $A \subset [n]$ with $|A| = m \leq \epsilon_{4.1} n$ and $\delta$ as in Proposition 4.1, define
\[ \tau_A := \left\{ x \in A : |O^A_{K_m}(x)| \geq (4/\delta) q^{-K_m} \right\} . \]
Clearly $|\tau_A| \geq \delta|A|$, because otherwise $B = A \setminus \tau_A$ will have $|B| \geq (1 - \delta)|A|$ and
\[ |O^A_{K_m}(x)| \leq \sum_{x \in B} |O^A_{K_m}(x)| < (4/\delta) q^{-K_m} |B| \]
by the definition of $\tau_A$ and this contradicts the fact that $\mathcal{G}_n \in G^1_n$.

Let $\mathcal{F}$ be the subgraph of $\{ Z^A_{t,\tau_A} \}_{t=0}^{K_m}$ induced by the vertex set
\[ \bigcup_{x \in \tau_A} \left( \bigcup_{i=0}^{K_m-1} O^A_t(x) \cup \Pi \left( (4/\delta) q^{-K_m} \cup O^A_{K_m}(x) \right) \right) . \]
So $\mathcal{F}$ is a labeled directed forest with depth $K_m$ such that $|\mathcal{F}_0| \geq \delta m$ and $|\mathcal{F}_{K_m,i}| = \left( (4/\delta) q^{-K_m} \right)$ for all $i$. Applying Lemma 5.1 with $k$ replaced by $K_m$ and $m$ replaced by $\delta m$, and noting that $\xi_{K_m} \xi_{K_m}$ stochastically dominates $\xi_{K_m}$,
\[ P_{\mathcal{G}_n,q} \left( \left| \xi_{K_m} \right| \leq m \right) \leq \exp \left( -\frac{1}{8} q^{K_m} \delta m \right) . \]
This proves the result. ■

PROOF OF THEOREM 1.4 AND 1.5. We take $\mathcal{G}_n := G^1_n \cap G^2_n$, where $G^1_n$ and $G^2_n$ are as in Proposition 4.1 and 4.2 respectively, and we will see that
\[ \Delta := \frac{1}{2} C_{5.2} \epsilon_{4.1} \log(1/\epsilon_{4.1})^{-b} \]
will suffice, where $C_{5.2}, b$ are as in Proposition 5.2. Clearly $P_{t,n}(\mathcal{G}_n) = 1 - o(1)$. Define
\[ T_x := \inf \left\{ t \geq 1 : \left| \xi_t \right| \geq \epsilon_{4.1} n \right\} . \]
We take $T_x = \infty$ if $\xi_t$ never reaches $\epsilon_{4.1} n$. Recalling the definition of the event $A_x$ from (4.19) and then applying Proposition 4.1, $\mathcal{G}_n \in \mathcal{G}_n$ implies
\[ P_{\mathcal{G}_n,q} \left( A_x \cap \left\{ T_x > 2a \log \log n / \log(q^r) + \sum_{m=\log n}^{\epsilon_{4.1} n - 1} K_m \right\} \leq \sum_{m=\log n}^{\epsilon_{4.1} n} \exp \left( -C_{5.2} m (\log(n/m))^{-b} \right) \leq n \exp \left( -C_{5.2} (\log n)^a (\log(n/\log n)^a)^{-b} \right) = o(1/n) \]
(5.3)
if \( a \) is large enough. For \( i \geq 1 \) if \( \left| \xi^{\{x\}}_{T_x+(i-1)K_m} \right| \geq \epsilon_{4.1} n \), then we can again apply Proposition 4.1 with \( A \) replaced by any subset of \( \xi^{\{x\}}_{T_x+(i-1)K_m} \) consisting of \( \epsilon_{4.1} n \) many vertices to have

\[
P_{\mathcal{G}_n,q} \left( \left\{ \left| \xi^{\{x\}}_{T_x+iK_m} \right| < \epsilon_{4.1} n \right\} \cap \left\{ \left| \xi^{\{x\}}_{T_x+(i-1)K_m} \right| \geq \epsilon_{4.1} n \right\} \right) \leq \exp \left( -C_{5.2} \epsilon_{4.1} \left[ \log(1/\epsilon_{4.1}) \right]^{-b} n \right),
\]

which in turn implies

\[
P_{\mathcal{G}_n,q} \left( \left\{ \left| \xi^{\{x\}}_{T_x+i\Delta_n K_m} \right| < \epsilon_{4.1} n \right\} \cap \{ T_x < \infty \} \right) = e^{\Delta_n} e^{-2\Delta_n} = o(1/n).
\]

Combining (5.3) and (5.4) and using union bound,

\[
P_{\mathcal{G}_n,q} \left( \bigcup_{x \in [n]} A_x \cap \left\{ \xi_{\exp(\Delta_n)}^{\{x\}} = \emptyset \right\} \right) \leq n o(1/n) = o(1).
\]

This together with the duality relationship between \( \xi_t \) and \( \xi_t^{\rightarrow} \) suggests

\[
P_{\mathcal{G}_n,q} \left( \xi_{\exp(\Delta_n)}^{[n]} \supset \{ x \in [n] : A_x \text{ occurs} \} \right) = P_{\mathcal{G}_n,q} \left( \xi_{\exp(\Delta_n)}^{\rightarrow} \neq \emptyset \text{ if } A_x \text{ occurs} \right) = 1 - o(1).
\]

Now in order to estimate the size of \( \{ x \in [n] : A_x \text{ occurs} \} \), we will use a second moment argument for \( \sum_{x \in [n]} 1_{A_x} \). Note that

\[
E_{\mathcal{G}_n,q} \left[ \sum_{x \in [n]} 1_{A_x} - \sum_{x \in [n]} P_{\mathcal{G}_n,q}(A_x) \right]^2 = \sum_{x,y \in [n]} [P_{\mathcal{G}_n,q}(A_x \cap A_y) - P_{\mathcal{G}_n,q}(A_x) P_{\mathcal{G}_n,q}(A_y)].
\]

Recalling the definition of the event \( A_{x,y} \) from (4.20) if \( A_{x,y} \) occurs, then the corresponding summand in the above sum is 0, otherwise the summands are at most 1. Keeping this observation in mind and using the fact that \( \mathcal{G}_n \in \mathcal{G}_n^2 \),

\[
E_{\mathcal{G}_n,q} \left[ \sum_{x \in [n]} 1_{A_x} - \sum_{x \in [n]} P_{\mathcal{G}_n,q}(A_x) \right]^2 \leq n + \sum_{x,y \in [n], x \neq y} 1_{A_{x,y}} \leq n + \binom{n}{2} o(1).
\]

The same estimate is true if we replace \( A_x \) by \( \tilde{A}_x \) in the above display. Also from Proposition 4.2 \( n(\pi - \varepsilon) \leq \sum_{x \in [n]} P_{\mathcal{G}_n,q}(A_x) \) and \( \sum_{x \in [n]} P_{\mathcal{G}_n,q}(\tilde{A}_x) \leq n(\pi + \varepsilon) \) for \( \mathcal{G}_n \in \mathcal{G}_n^2 \). Therefore, by Chebyshev inequality

\[
P_{\mathcal{G}_n,q} \left( (\pi - 2\varepsilon)n \leq \sum_{x \in [n]} 1_{A_x}, \sum_{x \in [n]} 1_{\tilde{A}_x} \leq (\pi + 2\varepsilon)n \right) = 1 - o(1).
\]

Combining this with (5.5)

\[
P_{\mathcal{G}_n,q} \left( \left| \xi_{\exp(\Delta_n)}^{[n]} \right| \geq n(\pi - 2\varepsilon) \right) = 1 - o(1).
\]

Also, using duality between \( \xi_t \) and \( \xi^{\rightarrow}_t \) again, for any \( t \geq t_n \)

\[
P_{\mathcal{G}_n,q} \left( \left| \xi_t^{[n]} \right| > n(\pi + 2\varepsilon) \right) = P_{\mathcal{G}_n,q} \left( \sum_{x \in [n]} 1_{\tilde{A}_x} > (\pi + 2\varepsilon)n \right) = o(1).
\]
So the required result follows from attractiveness of the threshold contact process.

To prove the last assertion suppose \( q < \frac{1}{r} \). Then we take

\[
\mathcal{G}_n := \left\{ \mathcal{G}_n : E_{\mathcal{G}_n,q} \left| \sum_{x \in \{x\}} \xi C \log n \right| \leq n^{C \log(qr)/2} \text{ for all } x \in [n] \right\}
\]

for some constant \( C > 0 \). In order to see that \( P_{i,n}(\mathcal{G}_n) = 1 - \Theta(1) \), recall the definition of \( P \) and \( \tilde{P} \) from the proof of Proposition 4.1. Using union bound and Markov inequality,

\[
P( (\mathcal{G}_n)^c ) \leq \sum_{x \in [n]} n^{-C \log(qr)/2} E_{\mathcal{G}_n,q} \left| \sum_{x \in \{x\}} \xi C \log n \right| \leq n \cdot n^{-C \log(qr)/2} \cdot n^{-C \log(qr)}.
\]

Since a branching process starting from \( x \) with offspring distribution \( p_n \) stochastically dominates \( \sum_{i \in \{x\}} \xi C \log n \), \( \tilde{P}( (\mathcal{G}_n)^c ) \) has the same upper bound. This together with the fact that \( r_n \to r \) implies \( P_{i,n}(\mathcal{G}_n) = 1 - \Theta(1) \). Finally for \( \mathcal{G}_n \in \mathcal{G}_n \), again using union bound and Markov inequality

\[
P_{\mathcal{G}_n,q}(\xi C \log n \neq \emptyset) = P_{\mathcal{G}_n,q} \left( \sum_{x \in \{x\}} \xi C \log n \neq \emptyset \right) \leq \sum_{x \in [n]} E_{\mathcal{G}_n,q} \left| \sum_{x \in \{x\}} \xi C \log n \right| = o(1)
\]

if \( C \) is chosen large enough.

\[\blacksquare\]

**Acknowledgments.** The author thanks Shankar Bhamidi for various helpful comments and discussions while writing this article.

**References.**


