CONTINUOUS AND DISCONTINUOUS PHASE TRANSITIONS IN HYPERGRAPH PROCESSES

R.W.R. DARLING, DAVID A. LEVIN, AND JAMES R. NORRIS

National Security Agency, University of Utah, University of Cambridge

ABSTRACT. Let V denote a set of N vertices. To construct a *hypergraph process*, create a new hyperedge at each event time of a Poisson process; the cardinality K of this hyperedge is random, with generating function $\rho(x) \stackrel{\text{def}}{=} \sum \rho_k x^k$, where $P(K = k) = \rho_k$; given $K = k$, the k vertices appearing in the new hyperedge are selected uniformly at random from V. Assume $\rho_1 + \rho_2 > 0$. Hyperedges of cardinality 1 are called *patches*, and serve as a way of selecting root vertices. Identifiable vertices are those which are reachable from these root vertices, in a strong sense which generalizes the notion of graph component. Hyperedges are called identifiable if all of their vertices are identifiable. We use "fluid limit" scaling: hyperedges arrive at rate N, and we study structures of size $O(1)$ and $O(N)$. After division by N, numbers of identifiable vertices and hyperedges exhibit phase transitions, which may be continuous or discontinuous depending on the shape of the structure function $-\log(1 - x)/\rho'(x)$, $x \in (0, 1)$. Both the case $\rho_1 > 0$, and the case $\rho_1 = 0 < \rho_2$ are considered; for the latter, a single extraneous patch is added to mark the root vertex.

1

1. Introduction

The *k-core* of a graph is the largest subgraph with minimum degree at least k. Pittel et al. (1996) study the following algorithm for finding the 2-core of a graph:

1. If vertices of degree one exist, select one and remove the edge incident to it. This may cause the degree of other vertices to drop.
2. If there are no degree one vertices remaining, stop.
3. Repeat.

The graph obtained at the conclusion of this algorithm is the 2-core.

This algorithm is a special case of another, run on hypergraphs, called *hypergraph collapse* and first studied in Darling and Norris (2004). By a *hypergraph* we shall mean a map $\Lambda : 2^V \to \{0, 1, 2, \dots\}$, where V is a finite set of *vertices* and 2^V is the set of subsets of V. It will sometimes be helpful to think in terms of an *edge-labelling* of Λ, which is a choice of a set I and a map $e : I \to 2^V$ such that $\Lambda(A) = |\{i \in I : e(i) = A\}|$ for all A. Thus e describes a set of labelled subsets of V, which we call *hyperedges* and then Λ gives the number of hyperedges at each subset of V. Hyperedges of unit cardinality are called *patches*. Hypergraph collapse is the following algorithm:

1. If a patch exists, select one and remove it together with the unique vertex v it contains. This will cause any other hyperedge $e(i)$ containing v to be replaced by $e(i) \setminus \{v\}$.
2. If there are no patches remaining, stop.
3. Repeat.

Although we have described the algorithm in terms of an edge-labelled hypergraph, the possible moves for Λ do not depend on the edge-labelling chosen. The vertices which are removed by hypergraph collapse are called *identifiable*, and hyperedges which contain only identifiable vertices are also called *identifiable*. These definitions do not depend on the order in which patches are chosen during hypergraph collapse; see Darling and Norris (2004).

The core-finding algorithm of Pittel et al. (1996) is hypergraph collapse applied to the *dual* hypergraph. To obtain the dual, note that we can think of e as a subset of $V \times I$. The roles of V and I are now symmetric, so e also corresponds to an edge-labelling of a hypergraph Λ' in which the status of vertices and hyperedges is reversed. Vertices (resp. hyperedges) of Λ not in the core correspond to identifiable hyperedges (resp. identifiable vertices) of Λ'. More information about graph cores

can be found in Fountoulakis (2002), and hypergraph cores are considered by Cooper (2002).

The identifiable vertices obtained by hypergraph collapse also serve to generalize to hypergraphs the definition of graph component. A graph is a hypergraph having edges only of cardinality two, and consequently has no patches. However, if the single hyperedge $\{v\}$ is added to the graph [making it a hypergraph], then the identifiable vertices obtained by running hypergraph collapse on the augmented graph are exactly the vertices in the graph component containing v. The identifiable edges are all the edges of this graph component.

This motivates the following definition for patch-free hypergraphs: A vertex is *in the domain* of v if it is in the set of identifiable vertices when the hypergraph is augmented by the addition of the hyperedge $\{v\}$.

The purpose of this paper is to study the time-evolution of the set of identifiable vertices and the set of identifiable edges in a *Poisson hypergraph process*, which is a hypergraph-valued, continuous-time stochastic process. The vertex set is $V = \{1, 2, \ldots, N\}$, and the process depends on parameters $\{\rho_j\}_{j=1}^N$. Attached to each subset A of V is a Poisson clock run at rate $N\rho_{|A|}/\binom{N}{|A|}$, and these clocks are independent of one another. [Here $|A|$ denotes the cardinality of A.] When the clock associated to A "rings", a new hyperedge equal to A is added to the hypergraph. The overall rate at which hyperedges of cardinality k are added is then $N\rho_k$. We will call this process the Poisson(ρ) hypergraph process. This is a generalization of the ordinary random graph process, in which edges form between each pair of vertices independently at a fixed rate.

While for N fixed, this process depends only on the finite sequence $\{\rho_k\}_{k=1}^N$, we will be interested in the asymptotic behavior as $N \to \infty$, so we will assume that always the infinite sequence $\{\rho_k\}_{k=1}^\infty$ is given. Moreover, this sequence is required to be a probability distribution on $\{1, 2, \ldots\}$ with finite expectation and satisfying $\rho_1 + \rho_2 > 0$. The generating function $x \mapsto \sum_{k=1}^\infty \rho_k x^k$ will be denoted by ρ.

In Darling and Norris (2004), the *Poisson(β) random hypergraph* is defined, where $\{\beta_k\}$ is a sequence of positive real numbers. This is a random hypergraph with vertex set $\mathrm{V} = \{1, 2, \ldots, N\}$, so that for each $A \subset V$, the number of occurrences of the hyperedge A is a Poisson random variable with expectation $N\beta_{|A|}/\binom{N}{|A|}$, and these random variables are independent for different subsets of V. If $\{\Lambda_t\}_{t \geq 0}$ is a Poisson(ρ) hypergraph process, then for fixed $t \geq 0$, Λ_t is a Poisson($t\rho$) random hypergraph.

We separate out two distinct cases in the study of Poisson hypergraph processes, depending on whether $\rho_1 > 0$ or $\rho_1 = 0$. When $\rho_1 = 0$, the hypergraph never acquires patches, and provided the initial hypergraph is patch-free, the set of identifiable vertices is forever void. As the previous discussion of ordinary graphs suggests, it is natural to consider in such cases the set of vertices in the domain of a distinguished vertex.

We discuss now the case $\rho_1 > 0$. Our first result describes the evolution of the rescaled number of identifiable vertices and hyperedges in the Poisson(ρ) hypergraph process $\{\Lambda_t\}_{t \geq 0}$. Let

$$\tilde{T}_t^N = \frac{|\text{identifiable vertices in } \Lambda_t|}{N}$$

(1)

$$\tilde{Z}_t^N = \frac{|\text{identifiable hyperedges in } \Lambda_t|}{N}\,.$$

The *structure function* t, defined as

(2)
$$t(x) \stackrel{\text{def}}{=} \frac{-\log(1-x)}{\rho'(x)}\,, \quad x \in (0,1)\,,$$

plays a central role for hypergraph processes. [Recall that $\rho(x) = \sum_{k=1}^{\infty} \rho_k x^k$.] Typically t is not invertible, but there is a right-continuous monotonic function called the *lower envelope*:

(3)
$$g(s) \stackrel{\text{def}}{=} \inf\{x \in (0,1) \ : \ t(x) > s\}\,, \quad s \geq 0\,.$$

Also important for hypergraph processes is the *upper envelope*:

(4)
$$g^\star(s) \stackrel{\text{def}}{=} \sup\{x \in (0,1) \ : \ t(x) < s\} \vee 0\,, \quad s \geq 0\,.$$

We classify structure functions into three types: graph-like, bicritical, and exceptional. This taxonomy is given in Table 1. Figure 1 shows a bicritical structure function and the corresponding lower envelope.

Let $\Xi \subset \mathbb{R}_+$ denote the discontinuity set of g:

(5)
$$\Xi \stackrel{\text{def}}{=} \{s > 0 \ : \ g(s-) \neq g(s)\}\,,$$

where $g(s-) \stackrel{\text{def}}{=} \lim_{t \uparrow s} g(t)$.

For $s \in \Xi$, both $g(s-)$ and $g(s)$ are zeros of the function $x \mapsto \rho'(x) + \log(1-x)$. For the sake of simplicity of exposition, we shall assume below that there are never any zeros of this function strictly between $g(s-)$ and $g(s)$:

(6)
$$\{x \ : \ s\rho'(x) + \log(1-x) = 0\} \bigcap (g(s-), g(s)) = \varnothing\,, \quad \text{for all } s \in \Xi\,.$$

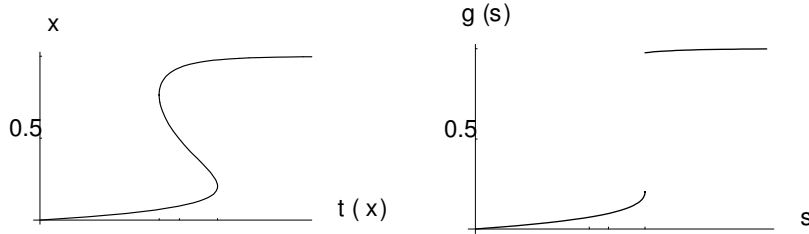| Type	Description	Example of $\rho(x)$
graph-like	t is strictly increasing, and g and $g^\star$ are continuous.	cubic with $3\rho_3 \leq \rho_2$
bicritical	g and $g^\star$ each have exactly one discontinuity.	cubic with $3\rho_3 > \rho_2$
exceptional	g or $g^\star$ has two or more discontinuities.	$\frac{x+5x^3+994x^{200}}{1000}$

TABLE 1. Classification of structure functions

FIGURE 1. *Left*: Bicritical structure function, with $t(x)$ on the horizontal axis, corresponding to a quartic polynomial $\rho(x)$ with $0 < \rho_1 < \rho_2 < \rho_3 < \rho_4$. *Right*: Lower envelope, showing the single discontinuity.

Also assume that Ξ has no accumulation points. This is true, for example, if $\sum_k k^2 \rho_k < \infty$.

Let $\{B_s, \ s \in \Xi\}$ denote a collection of independent Bernoulli(1/2) random variables, indexed by the discontinuity set (5). Define

(7)
$$\tilde{T}_t \overset{\text{def}}{=} g(t-) + B_t(g(t) - g(t-)), \quad t \in \Xi$$
$$\tilde{T}_t \overset{\text{def}}{=} g(t), \ t \notin \Xi.$$

In other words, at each point of discontinuity we choose the left limit or the right limit of g according to the flip of a fair coin. Finally, let

(8)
$$\tilde{Z}_t \overset{\text{def}}{=} t\rho(\tilde{T}_t) - (1 - \tilde{T}_t)\log(1 - \tilde{T}_t).$$

For a sequence of stochastic processes $\{X^N\}_{N=1}^{\infty}$, where $X^N = \{X_t^N\}_{t \geq 0}$, and a stochastic process $X = \{X_t\}_{t \geq 0}$, we write $X^N \xrightarrow{\text{f.d.d.}} X$ if the finite-dimensional distributions of X^N converge to those of X. For a sequence of random variables (or vectors) $\{X^N\}$, we write $X^N \xrightarrow{\text{d}} X$ to indicate that X^N converges in distribution to X.

Theorem 1. *Consider a Poisson hypergraph process such that $\rho_1 > 0$, and suppose* (6) *holds. As $N \to \infty$,*

$$(9) \qquad \{(\tilde{T}_t^N, \tilde{Z}_t^N)\}_{t \geq 0} \xrightarrow{\text{f.d.d.}} \{(\tilde{T}_t, \tilde{Z}_t)\}_{t \geq 0}.$$

Furthermore for any compact interval $I \subset [0, \infty) \setminus \Xi$,

$$(10) \qquad \sup_{t \in I} \left| (\tilde{T}_t^N, \tilde{Z}_t^N) - (g(t), t\rho(g(t)) - [1 - g(t)] \log(1 - g(t))) \right| \to 0$$

in probability as $N \to \infty$.

We now turn to the case of patch-free hypergraph processes, i.e. the regime where $\rho_1 = 0 < \rho_2$. In this case $g(s) = 0$ for all $s \in [0, (2\rho_2)^{-1})$. There are three possibilities for the behavior of g at $(2\rho_2)^{-1}$, enumerated in Table 2.

SUB-CASE OF $\rho_1 = 0 < \rho_2$	BEHAVIOR OF g
$3\rho_3 < \rho_2$	g is continuous at $(2\rho_2)^{-1}$, and right derivative is finite
$3\rho_3 = \rho_2$	$\rho_4, \rho_5, \ldots$ determine whether g is continuous at $(2\rho_2)^{-1}$
$3\rho_3 > \rho_2$	g is discontinuous at $(2\rho_2)^{-1}$

TABLE 2. The $0 = \rho_1 < \rho_2$ regime.

For simplicity, we focus on the case where g has a single discontinuity, located at $(2\rho_2)^{-1}$; i.e. $\Xi = \{(2\rho_2)^{-1}\}$. The general case follows the same pattern as Theorem 1, because after the number of identifiable vertices has reached $O(N)$, the subsequent evolution is much the same as the $\rho_1 > 0$ case.

In the $\rho_1 = 0$ and $\rho_2 > 0$ regime, another structure function besides (2) comes into play, namely the structure function t_2 of the graph which results from discarding all

hyperedges of cardinality more than two:

$$(11) \qquad t_2(x) \stackrel{\text{def}}{=} \frac{-\log(1-x)}{2\rho_2 x}, \quad x \in (0,1).$$

Since t_2 is monotonic, the corresponding lower envelope g_2 defined as

$$(12) \qquad g_2(s) \stackrel{\text{def}}{=} \inf\{x \in (0,1) : t_2(x) > s\}, \quad s \geq 0,$$

is continuous. As before, $g_2(s) = 0$ for $0 \leq s \leq (2\rho_2)^{-1}$, and $g_2(s) \to 1$ as $s \to \infty$; it describes the asymptotic proportion of vertices in the giant component of a random graph where the ratio of edges to vertices is $s\rho_2$.

We will construct in Section 7 an increasing process $\{M_t\}$ so that the distribution of M_t is

$$(13) \qquad P(M_t = n) = \begin{cases} e^{-2\beta_2 n} \, (2t\rho_2 n)^{n-1} / n! & \text{if } n \in \mathbb{N}, \\ \varphi_t & \text{if } n = \infty, \end{cases}$$

where φ_t is the largest solution x in $[0,1]$ of $2t\rho_2 x + \log(1-x) = 0$. [Notice that $\varphi_t = 0$ for $2t\rho_2 \leq 1$, and $0 < \varphi_t < 1$ otherwise.]

Write T_t^N for the number of vertices in the domain of v_0 in Λ_t, and write Z_t^N for the number of hyperedges identifiable from v_0 in Λ_t. Set $\bar{T}_t^N \stackrel{\text{def}}{=} N^{-1} T_t^N$ and $\bar{Z}_t^N \stackrel{\text{def}}{=} N^{-1} Z_t^N$. Also, define

$$\bar{T}_t \stackrel{\text{def}}{=} g(t)\mathbf{1}_{\{M_t=\infty\}};$$

$$\bar{Z}_t \stackrel{\text{def}}{=} \{t\rho(g(t)) - [1 - g(t)]\log(1 - g(t))\}\, \mathbf{1}_{\{M_t=\infty\}}.$$

Theorem 2. *Consider a Poisson hypergraph process such that $\rho_1 = 0 < \rho_2$, and suppose g has a single discontinuity located at $(2\rho_2)^{-1}$. Fix a distinguished vertex v_0. The number of vertices in the domain of v_0, and number of hyperedges identifiable from v_0, obey the following limits in distribution as $N \to \infty$:*

$$(14) \quad \{(T_t^N, Z_t^N)\}_{t \geq 0} \text{ converges weakly in } D\left([0,\infty), (\mathbb{N} \cup \{\infty\})^2\right) \text{ to } \{(M_t, M_t)\}_{t \geq 0},$$

where we adjoin ∞ to $\mathbb{N}$ as a compactifying point. Also

$$(15) \qquad \{(\bar{T}_t^N, \bar{Z}_t^N)\}_{t \geq 0} \stackrel{\text{f.d.d.}}{\longrightarrow} \{(\bar{T}_t, \bar{Z}_t)\}_{t \geq 0}.$$

Remark 1.1. Observe the difference between the limit law $\{(\bar{T}_t, \bar{Z}_t)\}_{t \geq 0}$ in (15) and the limit law $\{(\tilde{T}_t, \tilde{Z}_t)\}_{t \geq 0}$ in (9): $\tilde{T}_t$ conforms to the deterministic lower envelope $g(t)$, except at points in the finite discontinuity set, whereas $\bar{T}_t$ waits until the random time

$\chi \stackrel{\text{def}}{=} \inf\{t \geq 0 : M_t = \infty\}$, with distribution function $g_2(t)$, before jumping from 0 up to $g(t)$.

Remark 1.2. See Remark 5.1 as to whether the convergence (15) extends to weak convergence in the Skorohod space $D([0,\infty), \mathbb{R}^2_+)$.

The rest of this paper is organized as follows: Some definitions concerning hypergraphs are given in Section 2. We establish that certain key processes are Markov in Section 3. The case of hypergraphs and hypergraph processes with patches are treated in Section 4 and Section 5 respectively. Theorem 1 is proved in Section 5. Patch-free random hypergraphs and hypergraph processes are treated in Section 6 and Section 7, respectively. Theorem 2 is proved in Section 7. Finally, we mention future directions in Section 8.

2. Hypergraph definitions

Recall from the Introduction that the identifiable vertices are those vertices removed by the hypergraph collapse algorithm described there, and the identifiable hyperedges are those hyperedges consisting only of identifiable vertices.

Given a hypergraph Λ and a subset $S \subset V$, Λ^S denotes the hypergraph after all vertices in S are deleted. More precisely,

$$(16) \qquad \Lambda^S(A) \stackrel{\text{def}}{=} \sum_{B \supset A,\ B \setminus S = A} \Lambda(B), \quad A \subset V \setminus S.$$

We now more exactly specify the hypergraph collapse algorithm: select if possible a vertex v with $\Lambda(\{v\}) \geq 1$; replace V by $V \setminus \{v\}$ and Λ by $\Lambda^{\{v\}}$; then repeat. When the algorithm terminates, we obtain a set $V^\star$ consisting of the identifiable vertices, and a patch-free hypergraph $\Lambda^{V^\star}$ on $V \setminus V^\star$.

Suppose Λ is a patch-free hypergraph, and thus having no identifiable vertices. Given such a hypergraph Λ and a distinguished vertex v_0, we say that v is in the *domain* of v_0 in Λ if v is identifiable in the hypergraph $\Lambda + \mathbf{1}_{\{v_0\}}$ obtained by augmenting Λ by the hyperedge $\{v_0\}$. A hyperedge is said to be *identifiable* from v_0 if it is identifiable in $\Lambda + \mathbf{1}_{\{v_0\}}$.

Warning: For a general patch-free hypergraph, it is possible for vertex u to be in the domain of v, while v is not in the domain of u, although this cannot happen in graphs; see Figure 2.
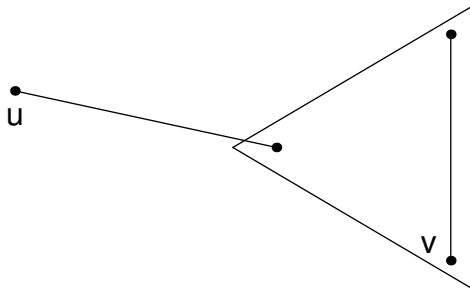
FIGURE 2. Adding a patch on v makes u identifiable, but not vice versa.

3. POISSON HYPERGRAPH PROCESSES: MARKOV PROPERTIES

For $\{\beta_k\}_{k=1}^{\infty}$ a sequence of non-negative numbers, a *Poisson(β) random hypergraph* is a random hypergraph Λ with vertex set $V = \{1, \ldots, N\}$ so that for $A \subset V$,

(i) the random variable $\Lambda(A)$ has a Poisson distribution with mean $N\beta_{|A|}/\binom{N}{|A|}$, and

(ii) $\{\Lambda(A) : A \subset V\}$ is a collection of independent random variables.

In what follows, $\{\rho_k\}_{k=1}^{\infty}$ will be a probability distribution on the positive integers which has finite mean and

$$(17) \qquad\qquad \rho_1 + \rho_2 > 0 \,.$$

We now give an explicit construction of the hypergraph-valued stochastic process described in the introduction. Let $K_1, K_2, \ldots$ be a sequence of independent random variables in $\{1, 2, 3, \ldots\}$ with common distribution $P(K_n = k) = \rho_k$, for all $n, k \in \mathbb{N}$. Denote by $A_1, A_2, \ldots$ a sequence of independent random subsets of V, such that A_n is chosen uniformly at random from the subsets of V of size K_n whenever $K_n \leq N$; the set A_n is not defined when $K_n > N$. Let $\{E_t\}_{t \geq 0}$ be a Poisson process, run at rate N, having arrival times $\tau_1, \tau_2, \ldots$. Define a stochastic process $\{\Lambda_t\}_{t \geq 0}$ with values in the set of hypergraphs with vertex set V by

$$\Lambda_t(A) \stackrel{\text{def}}{=} \sum_{n \,:\, \tau_n \leq t} \mathbf{1}_{\{A = A_n\}} \,.$$

Interpret $\Lambda_t(A)$ as the number of occurrences of hyperedge A by time t. In summary, for each $A \subset V$,

$$(18) \qquad\qquad \{\Lambda_t(A)\} \text{ is a Poisson process of rate } N\frac{\rho_{|A|}}{\binom{N}{|A|}} \,,$$

and all these Poisson processes are independent. We call $\{\Lambda_t\}_{t\geq 0}$ a *Poisson(ρ) hypergraph process*, where ρ denotes the generating function

$$(19) \qquad \rho(x) \stackrel{\text{def}}{=} \sum_{k\geq 1} \rho_k x^k \,.$$

The finite mean assumption is equivalent to $\rho'(1) < \infty$. For fixed $t \geq 0$, Λ_t is a Poisson($t\rho$) random hypergraph.

Whereas the hypergraph literature has tended to concentrate on the "k-uniform" case (i.e. $\rho_k = 1$ for some k), we find the superposition of k-uniform random hypergraphs for various different values of k can be handled without special effort, and leads to asymptotic properties absent from the k-uniform case. Moreover the Poisson structure simplifies our arguments, for example by allowing some summary statistics of $\{\Lambda_t\}_{t\geq 0}$ to be Markov processes in their own right: see Proposition 3.1. Poissonization is, of course, a well-established procedure – see Aldous (1989).

Previous literature has also concentrated on the case $\Lambda \leq 1$. We now sketch a way to deduce from our results for a Poisson(β) random hypergraph Λ some corresponding results for $\Lambda \wedge 1$. We note moreover that if $\rho_k = 1$ for some k then $\Lambda \wedge 1$ is exactly a k-uniform hypergraph. The set of identifiable vertices is the same for Λ and $\Lambda \wedge 1$ but Λ may have additional identifiable hyperedges. First consider patches. Throwing a Poisson($N\beta$) number of balls (i.e. patches) uniformly at random into N urns yields a Binomial($N, 1 - e^{-\beta_1}$) number of occupied urns (i.e. vertices covered by at least one patch). Hence the number of patches in Λ, less the number in $\Lambda \wedge 1$, divided by N, has limit in probability $\beta_1 + e^{-\beta_1} - 1$. On the other hand, the expected number of subsets of size at least 2 receiving at least 2 hyperedges is bounded uniformly in N. Hence, after rescaling by N^{-1}, only the extra patches in Λ can contribute in the limit and of course all of these do so.

Proposition 3.1. *Let T_t and Z_t denote the numbers of identifiable vertices and identifiable hyperedges for Λ_t. Both $\{T_t\}_{t\geq 0}$ and $\{(T_t, Z_t)\}_{t\geq 0}$ are Markov processes. The number of non-identifiable hyperedges in Λ_t, given that $T_t = m$, is conditionally Poisson, with mean*

$$(20) \qquad Nt \left[1 - \sum_{k\geq 1} \rho_k \frac{\binom{m}{k} + (N-m)\binom{m}{k-1}}{\binom{N}{k}} \right] \,.$$

When $m - N\gamma = o(N)$, for $\gamma \in [0,1]$, this reduces as $N \to \infty$ to

$$(21) \qquad Nt \left[1 - \rho(\gamma) - (1-\gamma)\rho'(\gamma) \right] + o(N) \,.$$

Remark 3.1. Because the total number of hyperedges in Λ_t is Poisson(Nt), Proposition 3.1 reduces the study of limits of identifiable hyperedges to study of limits of identifiable vertices. In particular, if $N^{-1}T_t$ converges in distribution as $N \to \infty$ to a random variable $\tilde{T}_t$, then necessarily

$$\text{(22)} \qquad N^{-1}Z_t \xrightarrow{\text{d}} t\left[\rho(\tilde{T}_t) + (1 - \tilde{T}_t)\rho'(\tilde{T}_t)\right].$$

Remark 3.2. It is easy to identify the generator of $\{T_t\}_{t \geq 0}$, rescale by division by N, and take a limit on any compact interval $I \subset \mathbb{R}_+ \setminus \Xi$ (see (5)); however this approach did not lead to a proof of Theorem 1, because of the difficulty of passing through discontinuous phase transitions.

To prepare for the proof, some measure-theoretic apparatus is needed. Let $(\Omega, \mathcal{F}, P)$ be the probability space on which the process $\{\Lambda_t\}_{t \geq 0}$ is defined. For any set $S \subset V$, and any $t \geq 0$, define the σ-field $\mathcal{F}_t^S \subset \mathcal{F}$ as

$$\mathcal{F}_t^S \overset{\text{def}}{=} \bigvee_{0 \leq s \leq t} \sigma\{\Lambda_s(A) \;:\; |A \setminus S| \leq 1\}.$$

Let $V_t^\star$ denote the set of vertices identifiable at time t. By construction, the event $\{V_t^\star = S\}$ occurs if and only if, among all sets containing all vertices covered by patches, S is the minimal subset of V for which $\Lambda_t(A) = 0$ whenever $|A \setminus S| = 1$. Thus $\{V_t^\star = S\} \in \mathcal{F}_t^S$.

When we consider $V_t^\star$ as a "stopping set" for a set-indexed process, it becomes natural to define another σ-field:

$$\mathcal{F}_{V_t^\star} \overset{\text{def}}{=} \left\{B \in \mathcal{F} \;:\; B \cap \{V_t^\star = S\} \in \mathcal{F}_t^S \text{ for all } S \subset V\right\}.$$

T_s and Z_s are $\mathcal{F}_{V_t^\star}$-measurable, for all $0 \leq s \leq t$. We may describe $\mathcal{F}_{V_t^\star}$ informally as the knowledge we have about $\{\Lambda_s\}_{0 \leq s \leq t}$ after performing hypergraph collapse at each time $s \in [0, t]$.

Lemma 3.2.

(i) *Fix any $t > 0$. Pick any collection of non-negative integers $\{k_A \;:\; A \subset V\}$, and set*

$$p(S) \overset{\text{def}}{=} P\left(\bigcap_{A \,:\, |A \setminus S| > 1} \{\Lambda_t(A) = k_A\}\right).$$

Then

$$P\left(\bigcap_{A\,:\,|A\setminus V_t^\star|>1}\{\Lambda_t(A)=k_a\}\;\middle|\;\mathcal{F}_{V_t^\star}\right)=p(V_t^\star)\,.$$

(ii) *Fix any $t>0$. The conditional distribution of the random hypergraph Λ_t^S (in the notation of (16)), given $\mathcal{F}_{V_t^\star}$, on the event $\{V_t^\star=S\}$, where $|S|=m$, is that of a Poisson(β) random hypergraph on $N-m$ vertices with parameters*

$$\beta_1\overset{\text{def}}{=}0$$

(23) $$\beta_j\overset{\text{def}}{=}\frac{t}{1-m/N}\binom{N-m}{j}\sum_{i\geq0}\rho_{i+j}\frac{\binom{m}{i}}{\binom{N}{j+i}}\,,\quad j\geq2\,.$$

For a random variable X, we write $X\sim\text{Poisson}(\mu)$ to indicate that the distribution of X is Poisson with expectation μ. Also we will write $X\sim\text{Binomial}(n,p)$ to indicate that X is a Binomial random variable with parameters n and p.

Proof of (i). Certainly $p(V_t^\star)$ is $\mathcal{F}_{V_t^\star}$-measurable. It remains to show that, for any $B\in\mathcal{F}_{V_t^\star}$,

$$\int_B p(V_t^\star)dP=P\left(B\cap\bigcap_{A\,:\,|A\setminus V_t^\star|>1}\{\Lambda_t(A)=k_a\}\right)\,.$$

Split the event on the right into disjoint events by intersecting with $\{V_t^\star=S\}$ for each $S\subset V$. For each S, $B\cap\{V_t^\star=S\}$ lies in $\mathcal{F}_t^S$, and therefore is independent of $\{\Lambda_t(A)=k_a\}$ for every A such that $|A\setminus S|>1$, by construction of a Poisson hypergraph process. The right side becomes

$$\sum_{S\subset V}p(S)P(B\cap\{V_t^\star=S\})$$

which is equal to the left side; (i) follows.   ■

Proof of (ii). Suppose $S\subset V$ and $A\subset V\setminus S$ with $|A|=j\geq2$. For any $C\subset S$ with $|C|=i$, (18) implies that

$$\Lambda_t(A\cup C)\sim\text{Poisson}\left(t\rho_{j+i}N/\binom{N}{j+i}\right)\,.$$

The result of part (i) implies that the random variables $\Lambda_t(A\cup C)$ are conditionally independent for different choices of C, given $\{V_t^\star=S\}\cap\mathcal{F}_{V_t^\star}$.

If $|S| = m$, there are $\binom{m}{i}$ choices of C, and following the notation of (16),

$$\Lambda_t^S(A) = \sum_{C \subset S} \Lambda_t(A \cup C) \sim \text{Poisson}\left(tN \sum_{i \geq 0} \rho_{j+i} \binom{m}{i} \Big/ \binom{N}{j+i} \right).$$

In a $\text{Poisson}(\beta)$ random hypergraph on $(N - m)$ vertices, the number of occurrences of A, where $|A| = j$, is Poisson with parameter

$$(N - m)\beta_j \Big/ \binom{N - m}{k}.$$

On comparison with the previous line, this verifies the formula (23) for β_j, when $j \geq 2$. Clearly there are no 1-hyperedges in Λ_t^S when $\{V_t^\star = S\}$, by definition of identifiability. Hence (ii) is established. ■

Proof of Proposition 3.1. Fix any $t > 0$. Suppose that $T_t = m$. The first jump in the process $\{(T_s, Z_s)\}_{s \geq t}$ can occur only when a new hyperedge arrives, and the arrival time is independent of the past. The law of the jump depends only on two things: the set A of vertices in the new hyperedge (which is independent of the past), and on the hypergraph Λ_t^S, where $S \stackrel{\text{def}}{=} V_t^\star$. Lemma 3.2(ii) establishes that the law of Λ_t^S, conditional on $\mathcal{F}_{V_t^\star}$ is fully determined by m, t, and the parameters $\{\rho_i\}_{i \geq 1}$; in particular it is conditionally independent of $\{(T_s, Z_s)\}_{0 \leq s \leq t}$ given that $\{T_t = m\}$. Hence the Markovian property of $\{T_t\}_{t \geq 0}$ and $\{(T_t, Z_t)\}_{t \geq 0}$ is established.

It follows from Lemma 3.2 that the total number of non-identifiable hyperedges in Λ_t, given that $\{T_t = m\}$, is conditionally Poisson, with mean $(N - m) \sum \beta_j$, for β_j as in (23). Write $k \stackrel{\text{def}}{=} i + j$, and switch the order of summation, to obtain

$$\left(1 - \frac{m}{N}\right) \sum \beta_j = t \sum_{k \geq 2} \rho_k \sum_{j=2}^{k} \binom{N - m}{j}\binom{m}{k - j} \Big/ \binom{N}{k}.$$

On considering the $\text{Hypergeometric}((N, N - m, k))$ distribution, we see that the inner sum is

$$1 - \left[\binom{m}{k} + (N - m)\binom{m}{k - 1} \right] \Big/ \binom{N}{k}.$$

The last expression is zero when $k = 1$, so $(N - m) \sum \beta_j$ takes the form (20). When $m - N\gamma = O(N)$, the last expression converges, as $N \to \infty$, to $1 - \gamma^k - k\gamma^{k-1}(1 - \gamma)$, and is bounded between 0 and 1. The Bounded Convergence Theorem yields (21). ■

4. Identifiability In Random Hypergraphs With Patches

In this section we review some material from Darling and Norris (2004).

Fix $t > 0$, and set $\Lambda \overset{\text{def}}{=} \Lambda_t$, $\beta_k \overset{\text{def}}{=} t\rho_k$. In this case, Λ is a Poisson(β) random hypergraph. Suppose we perform hypergraph collapse, described above, in the following special way: at each step the next vertex v to be deleted is selected with a probability proportional to the number of patches on v. This is called *randomized collapse*. The *debris* of a hypergraph is the number of hyperedges equal to the empty set. Set $\Lambda_0 \overset{\text{def}}{=} \Lambda$, and let $\{\Lambda_n\}_{n \in \mathbb{N}}$ denote the sequence of hypergraphs obtained. Set Y_n and Z_n to be the amount of patches and debris, respectively, in Λ_n; formally

$$Y_n \overset{\text{def}}{=} \sum_{v \in V} \Lambda_n(\{v\}), \quad \text{and} \quad Z_n \overset{\text{def}}{=} \Lambda_n(\varnothing).$$

The key observation in Darling and Norris (2004) is that $\{(Y_n, Z_n)\}_{n \in \mathbb{N}}$ is a Markov chain (but not the same one as in Proposition 3.1, for here t is fixed!), which stops at

$$(24) \qquad\qquad\qquad T \overset{\text{def}}{=} \inf\{n \,:\, Y_n = 0\}.$$

Moreover, conditional on $\{Y_n = m, Z_n = k\}$,

$$(25) \qquad\qquad \begin{aligned} Z_{n+1} &= k + 1 + W_{n+1}, \\ Y_{n+1} &= m - 1 - W_{n+1} + U_{n+1}. \end{aligned}$$

Here W_{n+1} and U_{n+1} are independent, with

$$(26) \qquad\qquad \begin{aligned} W_{n+1} &\sim \text{Binomial}\left(m - 1, \frac{1}{N - n}\right) \\ U_{n+1} &\sim \text{Poisson}\left((N - n - 1)t\lambda_2(N, n)\right) \end{aligned}$$

where

$$(27) \qquad\qquad \lambda_2(N, n) \overset{\text{def}}{=} N \sum_{i=0}^{n} \rho_{2+i} \binom{n}{i} \Big/ \binom{N}{i+2}.$$

By construction, $T = |V^\star|$, the number of identifiable vertices, and $Z \overset{\text{def}}{=} Z_T \overset{\text{def}}{=} \Lambda_T(\varnothing)$ is the number of identifiable hyperedges. For comparison, note that, by Proposition 3.1 the number of non-identifiable hyperedges in Λ, given that $T = N\gamma$, is conditionally Poisson, with mean

$$(28) \qquad\qquad N(t - \beta(\gamma) - (1 - \gamma)\beta'(\gamma)) + o(N).$$

We obtained a limit theorem for $\tilde{T}^N \stackrel{\text{def}}{=} N^{-1}T$ and $\tilde{Z}^N \stackrel{\text{def}}{=} N^{-1}Z$, where Z is the number of identifiable hyperedges. We state the result in a simple case. Set

$$\beta(x) \stackrel{\text{def}}{=} \sum_k \beta_k x^k, \quad x \in [0,1].$$

Assume that $\beta_1 > 0$ and that the derivative $\beta'(1) < \infty$. Then

(29) $$\{x \in [0,1) \,:\, \beta'(x) + \log(1-x) < 0\}$$

is non-empty, and its infimum is $g(t)$, as defined in (3). By our assumption (6), there is at most one $x \in [0, g(t))$ such that $\beta'(x) + \log(1-x) = 0$, namely $g(t-)$; this is different to $g(t)$ only if $t \in \Xi$, the set of discontinuity points of the lower envelope g.

Let $\tilde{T}$ be a random variable taking values $g(t)$ and $g(t-)$, each with probability $1/2$. As a special case of of Darling and Norris (2004, Theorem 2.2) we know:

Theorem 4.1. *The following limit in distribution holds as $N \to \infty$:*

(30) $$\left(\tilde{T}^N, \tilde{Z}^N \right) \stackrel{\text{d}}{\longrightarrow} \left(\tilde{T}, \beta(\tilde{T}) - (1 - \tilde{T}) \log(1 - \tilde{T}) \right).$$

Remark 4.1. Goldschmidt and Norris (2002) have shown that the limit for the rescaled number of identifiable hyperedges can be decomposed as follows: $(1 - \tilde{T}) \log(1 - \tilde{T})$ counts the *essential hyperedges*, i.e. those whose absence would have reduced the set of identifiable vertices, and $\beta(\tilde{T})$ counts the remainder.

Remark 4.2. Suppose in particular that $\Lambda \stackrel{\text{def}}{=} \Lambda_t$ and $\beta(x) \stackrel{\text{def}}{=} t\rho(x)$ for some $t \in \Xi$, the discontinuity set of g. Then (30) implies that the proportion of identifiable vertices has a limit in distribution which is random, taking the values $g(t)$ and $g(t-)$ each with probability $1/2$.

Remark 4.3. It suffices to derive the limit for $\tilde{T}^N$, since the limit for $\tilde{Z}^N$ follows from Proposition 3.1. To check this, recall that, by (22), if $\tilde{T}^N$ converges to $g(t)$, then the number of identifiable hyperedges, divided by N, converges to

(31) $$t\{\rho(g(t)) + [1 - g(t)]\rho'(g(t))\}.$$

However by definition of $g(t)$, $t\rho'(g(t)) = -\log(1 - g(t))$, so we have recovered the formula $\beta(\tilde{T}) - (1 - \tilde{T}) \log(1 - \tilde{T})$.

5. Identifiability In Hypergraph Processes With Patches

In this section we move from the static random hypergraph model of Theorem 4.1 to the Poisson(ρ) hypergraph process $\{\Lambda_t\}_{t\geq 0}$, providing here a proof of Theorem 1.

Extending the notation of the previous section, let $\tilde{T}_t^N$ and $\tilde{Z}_t^N$ denote the rescaled numbers of identifiable vertices and hyperedges for Λ_t, respectively, as defined in (1). Note that $t \mapsto \tilde{T}_t^N$ and $t \mapsto \tilde{Z}_t^N$ are increasing, right-continuous, stochastic processes. It follows from Proposition 3.1 that $\{(\tilde{T}_t^N, \tilde{Z}_t^N)\}_{t\geq 0}$ is a Markov process.

Proof of Theorem 1. Fix $0 \leq t_1 < \ldots < t_r$. We have to show the convergence in distribution

$$(32) \qquad\qquad \{(\tilde{T}_{t_i}^N, \tilde{Z}_{t_i}^N)\}_{i=1,\ldots,r} \xrightarrow{\text{d}} \{(\tilde{T}_{t_i}, \tilde{Z}_{t_i})\}_{i=1,\ldots,r}\,.$$

It suffices to do so when at least one of $\{t_i, t_{i+1}\}$ is not a discontinuity point, for every $i \in \{1, \ldots, r-1\}$. Proposition 3.1 showed that $\{(\tilde{T}_t^N, \tilde{Z}_t^N)\}_{t\geq 0}$ is Markov, and for any Markov process $\{Y_t\}_{t\geq 0}$ the conditional law of Y_{t_r} given $(Y_{t_1}, \ldots, Y_{t_{r-1}})$ is the same as the conditional law given $Y_{t_{r-1}}$. Hence it suffices to consider the case $r = 2$ such that $t_1 \notin \Xi$ or $t_2 \notin \Xi$, and these possibilities are both subsumed in the case $r = 3$ with $t_1, t_3 \notin \Xi$. Then only the marginal limit at time t_2, as given in Theorem 4.1 is random, so Theorem 4.1 implies the full convergence in distribution.

The second assertion follows from the first since all processes are increasing, and the limit is deterministic and continuous on I. ∎

Remark 5.1. The rescaled number of essential hyperedges, as studied by Goldschmidt and Norris (2002), has a limit $\{-(1 - \tilde{T}_t) \log(1 - \tilde{T}_t)\}_{t\geq 0}$ in the same sense as (9) and (10).

Remark 5.2. One may ask whether the convergence (9) extends to weak convergence in the Skorohod space $D([0,\infty), \mathbb{R}_+^2)$. Since $t \mapsto \tilde{T}_t^N$ and $t \mapsto \tilde{Z}_t^N$ are non-decreasing, the necessary and sufficient condition of Jacod and Shiryaev (1987, p. 306) may be applied, which would require that the sum of squared jumps of $\{\tilde{T}_t^N\}$ converges in law to the sum of squared jumps of $\{\tilde{T}_t\}$, and similarly for $\{\tilde{Z}_t^N\}$. Unfortunately the techniques presented in this paper do not seem to be able to confirm this; indeed, it seems plausible that, for arbitrarily large N, and for $t \in \Xi$, there is a probability bounded away from zero that $\tilde{T}_s^N$ makes more than one jump in going from $\approx g(t-)$ to $\approx g(t)$ at time $s \approx t$, and this would contradict the condition stated.

Remark 5.3. If (6) is false, one can reformulate the process (7), by consulting Darling and Norris (2004, Theorem 2.2) and prove a corresponding version of Theorem 1.

6. Domain Of A Vertex In A Hypergraph Without Patches

We revert to the fixed-time setting of Section 4. Suppose Λ is a Poisson(β) random hypergraph, such that

$$\beta_0 = \beta_1 = 0 < \beta_2, \quad \beta(x) \overset{\text{def}}{=} \sum_{k \geq 2} \beta_k x^k, \ x \in [0, 1].$$

Fix a vertex v_0. Write T^N for the number of vertices in the domain of v_0, and write Z^N for the number of hyperedges identifiable from v_0. Set $\bar{T}^N \overset{\text{def}}{=} N^{-1}T^N$ and $\bar{Z}^N \overset{\text{def}}{=} N^{-1}Z^N$. Both the *microscopic variables* (T^N, Z^N), and the *macroscopic variables* $(\bar{T}^N, \bar{Z}^N)$ have non-trivial limits as $N \to \infty$, which we now describe. The coefficient β_2 plays a distinguished role.

Lemma 6.1. *Let* $\{\xi_n\}_{n \in \mathbb{N}}$ *be a random walk on the integers, started at* $\xi_0 = 1$, *whose increments are of the form* $\xi_n - \xi_{n-1} = -1 + \text{Poisson}(2\beta_2)$. *Let* φ *be the largest root in* $[0, 1]$ *of* $2\beta_2 x + \log(1 - x) = 0$, *so* $\varphi = 0$ *for* $2\beta_2 \leq 1$, *and* $0 < \varphi < 1$ *otherwise. Then the first passage time to* 0,

(33) $$M \overset{\text{def}}{=} \inf\{n \geq 0 \ : \ \xi_n = 0\},$$

has the following distribution:

(34)
$$P(M = n) = e^{-2\beta_2 n} (2\beta_2 n)^{n-1} / n!, \quad n \in \mathbb{N};$$
$$P(M = \infty) = \varphi,$$

Remark. M is distributed as the total number of individuals in a branching process with one ancestor, and Poisson($2\beta_2$) offspring distribution. This distribution describes the sizes of small components in an Erdős-Rényi random graph; see Bollobás (2001).

Proof. The fact that $P(M = \infty) = \varphi$ is an elementary fact from the theory of branching processes. The formula for $P(M = n)$ is a special case of a formula of Dwass (1969), which is proved in detail on p. 300 of Devroye (1998). ∎

Assume that $\beta'(1) < \infty$. Then the set (29) is non-empty, and its infimum is $g \overset{\text{def}}{=} g(t)$, as defined in (3). Assume further that $\beta'(x) + \log(1 - x) > 0$ for all $x \in (0, g)$. If either of these assumptions fail, then the techniques of Darling and

Norris (2004), combined with some arguments given below, still establish the desired asymptotics. We omit the details.

Set

$$
(35) \qquad
\begin{aligned}
\bar{T} &\stackrel{\text{def}}{=} g\mathbf{1}_{\{M=\infty\}}\,; \\
\bar{Z} &\stackrel{\text{def}}{=} [\beta(g) - (1-g)\log(1-g)]\,\mathbf{1}_{\{M=\infty\}}\,.
\end{aligned}
$$

Theorem 6.2. *Consider a Poisson random hypergraph without patches, and fix a distinguished vertex v_0. The number of vertices in the domain of v_0, and number of hyperedges identifiable from v_0, obey the following limits in distribution as $N \to \infty$:*

$$
(36) \qquad (T^N, Z^N) \stackrel{\text{d}}{\longrightarrow} (M, M)\,; \quad (\bar{T}^N, \bar{Z}^N) \stackrel{\text{d}}{\longrightarrow} (\bar{T}, \bar{Z})\,.
$$

Here M is considered as a random variable taking values in the one-point compactification $\mathbb{N} \cup \{\infty\}$ of $\mathbb{N}$.

Proof. Step I. Set $\Lambda_0 \stackrel{\text{def}}{=} \Lambda + \mathbf{1}_{\{v_0\}}$, and let $\{\Lambda_n\}_{n \in \mathbb{N}}$ be a sequence of hypergraphs obtained by randomized collapse. Denote by Y_n^N and Z_n^N the numbers of patches and debris, respectively, in Λ_n. Then

$$
T^N \stackrel{\text{def}}{=} \inf\{n \geq 0 \,:\, Y_n^N = 0\}\,; \quad Z^N \stackrel{\text{def}}{=} Z_{T^N}^N\,.
$$

We know that $\{(Y_n^N, Z_n^N)\}_{n \geq 0}$ is a Markov chain, starting from $(1,0)$: the increments, conditional on $Y_n^N = m \geq 1$ and $Z_n^N = k$, are as given in (25) and (26).

For fixed $n \geq 0$ and $m \geq 1$, the random variable W_{n+1} defined in (26) converges to 0 in distribution as $N \to \infty$. Also

$$
(37) \qquad (N - n - 1)\lambda_2(N, n) \to 2\rho_2\,.
$$

so the random variable U_{n+1} defined in (26) converges to Poisson($2\beta_2$) in distribution as $N \to \infty$. Hence, for all $n \geq 0$,

$$
\{(Y_j^N, Z_j^N)\}_{0 \leq j \leq n} \stackrel{\text{d}}{\longrightarrow} \{(\xi_j, j)\}_{0 \leq j \leq n}
$$

which implies $(T^N, Z^N) \stackrel{\text{d}}{\longrightarrow} (M, M)$ as $N \to \infty$. If $2\beta_2 \leq 1$, then $P(M = \infty) = 0$, so the proof is complete. It only remains to prove the second convergence assertion in the case where $2\beta_2 > 1$, and $0 < \varphi < 1$.

Step II. Introduce an auxiliary time variable t, and let $\{\nu_t\}_{t\geq 0}$ be a Poisson process of rate N. Set

$$\bar{Y}_t^N \overset{\text{def}}{=} N^{-1} Y_{\nu_t}^N \,,$$

$$\bar{Z}_t^N \overset{\text{def}}{=} N^{-1} Z_{\nu_t}^N \,,$$

$$\bar{\nu}_t^N \overset{\text{def}}{=} N^{-1} \nu_t \,,$$

$$\tau^N \overset{\text{def}}{=} \inf\{t \geq 0 \,:\, \bar{Y}_t^N = 0\} \,.$$

With reference to Darling and Norris (2004), set

$$y(t) \overset{\text{def}}{=} (1 - t)(\beta'(t) + \log(1 - t)) \,;$$

$$z(t) \overset{\text{def}}{=} \beta(t) - (1 - t) \log(1 - t) \,.$$

By Theorem 6.1 and Remark 6.2 of Darling and Norris (2004), for all $\delta > 0$,

$$(38) \qquad \limsup_{N \to \infty} \frac{1}{N} \log \left(P\left(\sup_{t \leq \tau^N} \left\| (\bar{\nu}_t^N, \bar{Y}_t^N, \bar{Z}_t^N) - (t, y(t), z(t)) \right\| > \delta \right) \right) < 0 \,.$$

Observe that $\bar{\nu}_{\tau^N}^N = \bar{T}^N$, which will have the same limit in probability as does τ^N. We will show that, for all $\theta \in (\log(1 - \varphi)], 0)$, there exists $\delta > 0$ and N_0 such that

$$(39) \qquad\qquad P\big(\bar{T}^N \leq \delta\big) \leq e^\theta \,, \quad \text{for all } N \geq N_0 \,.$$

By (34) and the fact that $T^N \overset{\text{d}}{\longrightarrow} M$, we know that, for all $\delta > 0$ and all $\varphi' > \varphi$:

$$P\big(\bar{T}^N \leq \delta\big) \geq 1 - \varphi'$$

for all sufficiently large N. Also from (38) we obtain, for all $\delta > 0$,

$$P\big(\bar{T}^N \in (\delta, g - \delta) \cup (g + \delta, \infty)\big) \to 0$$

as $N \to \infty$. Hence the claim that $(\bar{T}^N, \bar{Z}^N) \overset{\text{d}}{\longrightarrow} (\bar{T}, \bar{Z})$ will follow as soon as we have proved (39); then (38) will strengthen this to show $(\bar{T}^N, \bar{Z}^N) \overset{\text{d}}{\longrightarrow} (\bar{T}, \bar{Z})$.

Step III. The remainder of the proof is to establish (39). Given $Y_n^N = m \geq 1$, set

$$\Phi^N(m, n) \overset{\text{def}}{=} E \exp\{\theta(-1 - W_{n+1} + U_{n+1})\}$$

$$= \exp\left\{ -\theta + F\left(m - 1, \frac{1}{N - n}, -\theta\right) + G((N - n - 1)\lambda_2(N, n), \theta) \right\} \,.$$

where

$$F(k, p, \theta) \overset{\text{def}}{=} k \log\left(1 - p + pe^\theta\right) \,; \quad G(\mu, \theta) \overset{\text{def}}{=} \mu(e^\theta - 1) \,.$$

Lemma 6.1 of Darling and Norris (2004) implies that

$$\sup_{n \leq N/2} \left| (N - n - 1)\lambda_2(N, n) - \left(1 - \frac{n}{N}\right)\beta''(n/N) \right| \to 0 \,,$$

as $N \to \infty$. Since $\theta > \log(1 - \varphi)$, there is $\bar{\varphi} < \varphi$ such that $\theta > \bar{\theta} \stackrel{\text{def}}{=} \log(1 - \bar{\varphi})$; by construction of φ, $2\beta_2\bar{\varphi} + \log(1 - \bar{\varphi}) > 0$, so $2\beta_2(1 - e^{\bar{\theta}}) + \bar{\theta} > 0$; in other words,

$$\exp\{-\bar{\theta} + G(2\beta_2, \bar{\theta})\} < 1 \,.$$

We can therefore find $\delta > 0$ and N_0 such that

(40) $\Phi^N(m, n) \leq 1 \,, \quad \text{for all } m, \ n \leq N\delta, \quad \text{for all } N \geq N_0 \,.$

Consider the martingale

$$M_n \stackrel{\text{def}}{=} e^{\bar{\theta} Y_n^N} \left(\prod_{k=0}^{n-1} \Phi^N(Y_k^N, k) \right)^{-1} ,$$

and set $R^N \stackrel{\text{def}}{=} \inf\{n \geq 0 \ : \ Y_n^N \geq N\delta\}$. It follows from (40) that, on the event $\{T^N \leq R^N \wedge N\delta\}$,

$$M_{T^N} \geq 1 \,, \quad \text{for all } N \geq N_0 \,.$$

Hence for $N \geq N_0$,

$$e^\theta > EM_0 = e^{\bar{\theta}} = EM_{T^N \wedge R^N \wedge N\delta} \geq P\big(T^N \leq R^N \wedge N\delta\big) \,.$$

However (38) implies that, for $\delta < g/2$, $P\big(R^N < T^N \leq N\delta\big) \to 0$, and (39) follows.
∎

7. Identifiability In Patch-Free Processes

We now focus on the case of patch-free hypergraph processes, proving in this section Theorem 2.

7.1. A Coupled Family of Random Walks.
Let $\{P_t(n)\}_{t \geq 0}$, $n \in \mathbb{N}$, be a family of independent Poisson processes, all of rate $2\rho_2 > 0$, and consider the coupled family of random walks $\{\xi_t(n)\}_{n \geq 0}$, for $t \in \mathbb{R}_+$, where $\xi_t(0) = 1$ for all n, and

(41) $\xi_t(n + 1) = \xi_t(n) + (P_t(n + 1) - 1)\mathbf{1}_{\{n < M_t\}} \,;$

(42) $M_t \stackrel{\text{def}}{=} \inf\{n \geq 0 \ : \ \xi_t(n) = 0\} \in \mathbb{N} \cup \{\infty\} \,.$

The marginal law of M_t is given by (34) with $\beta_2 \stackrel{\text{def}}{=} t\rho_2$. There is a relation between $\{\xi_t(n)\}_{n\geq 0}$ and the multigraph structure function: since $g_2(t)$ is the largest root in $[0,1]$ of $2t\rho_2 x + \log(1-x) = 0$, we have as a special case of (34):

Lemma 7.1. *The first time t at which $\{\xi_t(n)\}_{n\geq 0}$ escapes to infinity is related to the multigraph lower envelope* (12) *as follows:*

$$(43) \qquad\qquad\qquad P(M_t = \infty) = g_2(t)\,.$$

Moreover $t \mapsto M_t$ is an increasing process by the coupling, so $\chi \stackrel{\text{def}}{=} \inf\{t \geq 0 : M_t = \infty\}$ is a continuous random variable with distribution function $g_2(t)$.

7.2. **Notation.** We finally turn to the case of a Poisson(ρ) hypergraph process $\{\Lambda_t\}_{t\geq 0}$ without patches, i.e. such that

$$\rho_0 \stackrel{\text{def}}{=} \rho_1 \stackrel{\text{def}}{=} 0 < \rho_2\,, \quad \rho(x) \stackrel{\text{def}}{=} \sum_{k\geq 2} \rho_k x^k\,, x \in [0,1]\,.$$

Write T_t^N for the number of vertices in the domain of v_0 in Λ_t, and write Z_t^N for the number of hyperedges identifiable from v_0 in Λ_t. Set $\bar{T}_t^N \stackrel{\text{def}}{=} N^{-1}T_t^N$ and $\bar{Z}_t^N \stackrel{\text{def}}{=} N^{-1}Z_t^N$. Using (42), we define what will turn out to be the macroscopic limits for Theorem 2.

$$\bar{T}_t \stackrel{\text{def}}{=} g(t)\mathbf{1}_{\{M_t=\infty\}}\,;$$
$$\bar{Z}_t \stackrel{\text{def}}{=} \{t\rho(g(t)) - [1 - g(t)]\log(1 - g(t))\}\,\mathbf{1}_{\{M_t=\infty\}}\,.$$

Proof of Theorem 2.

Step I. Extending the notation of Theorem 6.2 let $\Lambda_t(n)$ denote the hypergraph that results from applying n steps of randomized collapse to $\Lambda_t + \mathbf{1}_{\{v_0\}}$; $Y_t^N(n)$ and $Z_t^N(n)$ count the number of patches, and the amount of debris, respectively in $\Lambda_t(n)$, and n is assumed to satisfy:

$$n \leq T_t^N \stackrel{\text{def}}{=} \inf\{n \geq 0 : Y_t^N(n) = 0\}\,.$$

Consider a finite set of time points $0 < t_1 < \ldots < t_r$. The hypergraph collapses of $\Lambda_{t_1} + \mathbf{1}_{\{v_0\}}, \ldots, \Lambda_{t_r} + \mathbf{1}_{v_0}$ are coupled together as follows: perform the $(n+1)$st step of randomized collapse by choosing a patch uniformly at random from the smallest unstable hypergraph. Poisson symmetries imply that this amounts to randomized

collapse for each of the unstable hypergraphs. Condition on the event:

$$(44) \qquad \bigcap_{i=1}^{r} \{ Y_{t_i}^N(n) = m_i , \ Z_{t_i}^N(n) = k_i \} .$$

For i such that $m_i = 0$, evidently $Y_{t_i}^N(n+1) = 0$ and $Z_{t_i}^N(n+1) = k_i$. For those i such that $m_i \geq 1$, we may write:

$$Y_{t_i}^N(n+1) = m_i - 1 - W_{t_i}^N(n+1) + U_{t_i}^N(n+1) ;$$
$$Z_{t_i}^N(n+1) = k_i + 1 + W_{t_i}^N(n+1) ,$$

where the random increments are distributed as follows. Take q to be the least $i \in \{1, 2, \ldots, r\}$ for which $m_i \geq 1$, and take $W_{t_q}^N(n+1)$ and $U_{t_q}^N(n+1)$ independent such that

$$(45) \qquad \begin{aligned} W_{t_q}^N(n+1) &\sim \mathrm{Binomial}\left(m_q - 1, \frac{1}{N-n} \right) ; \\ U_{t_q}^N(n+1) &\sim \mathrm{Poisson}\left((N-n-1)t_q \lambda_2(N-n) \right) , \end{aligned}$$

where $\lambda_2(N, n)$ is as in (27). Because of the coupling, we may take subsequent increments (for $i = q, \ldots, r-1$) to be independent and of the form:

$$W_{t_{i+1}}^N(n+1) - W_{t_i}^N(n+1) \sim \mathrm{Binomial}\left(m_{i+1} - m_i, \frac{1}{N-n} \right) ;$$
$$U_{t_{i+1}}^N(n+1) - U_{t_i}^N(n+1) \sim \mathrm{Poisson}\left((N-n-1)(t_{i+1} - t_i)\lambda_2(N, n) \right) .$$

Step II. Observe that the behavior of $\lambda_2(N, n)$ depends on whether $n \stackrel{\mathrm{def}}{=} O(1)$, or $n \stackrel{\mathrm{def}}{=} O(N)$. It follows from (37) and the calculations in Step I that, conditional on (44), the joint law of

$$\left((Y_{t_1}^N(n+1), Z_{t_1}^N(n+1)), \ldots, (Y_{t_r}^N(n+1), Z_{t_r}^N(n+1)) \right)$$

converges as $N \to \infty$ to the conditional law of

$$\left((\xi_{t_1}(n+1), k_1 + 1), \ldots, (\xi_{t_r}(n+1), k_r + 1) \right)$$

given that $\xi_{t_1}(n) = m_1, \ldots, \xi_{t_r}(n) = m_r$. Evidently $Z_{t_i}^N(0) = 0$ for all i. Since n was arbitrary, and since for each t both $\{\xi_t(n)\}_{n \geq 0}$ and $\{(Y_t^N(n), Z_t^N(n))\}_{n \geq 0}$ are Markov, we have now proved convergence in distribution as $N \to \infty$:

$$\{(Y_{t_1}^N(n), Z_{t_1}^N(n)), \ldots, (Y_{t_r}^N(n), Z_{t_r}^N(n))\}_{n \geq 0}$$

$$\stackrel{\mathrm{d}}{\longrightarrow} \{(\xi_{t_1}(n), n \wedge M_{t_1}), \ldots, (\xi_{t_r}(n), n \wedge M_{t_r})\}_{n \geq 0} .$$

In particular, in the notation of (42) and Section 7.2,

$$(46) \qquad ((T_{t_1}^N, Z_{t_1}^N), \ldots, (T_{t_r}^N, Z_{t_r}^N)) \xrightarrow{\text{d}} ((M_{t_1}, M_{t_1}), \ldots, (M_{t_r}, M_{t_r})) \, .$$

Step III. To prove (14) it suffices, in the light of (46), to prove tightness of $\{(T_t^N, Z_t^N)\}_{t \geq 0}$ with respect to the Skorohod topology of $D([0, \infty), (\mathbb{N} \cup \{\infty\})^2)$. On $(\mathbb{N} \cup \{\infty\})^2$, we shall use the metric

$$d\left((m, n), (p, q)\right) \overset{\text{def}}{=} \max\left\{ \left| \frac{1}{p} - \frac{1}{m} \right|, \left| \frac{1}{q} - \frac{1}{n} \right| \right\}$$

understanding that $1/\infty = 0$. We shall verify the condition of Aldous for tightness of $\{(T_t^N, Z_t^N)\}_{t \geq 0}$, as stated in Billingsley (1999), p. 176, or Kallenberg (2002), p. 314, with respect to this metric. Since $s \mapsto T_s^N$ and $s \mapsto Z_s^N$ are non-decreasing processes, the condition takes a slightly simpler form than usual: it suffices to show that, for each $\epsilon > 0$ and $\eta > 0$, there exist h and N_0 such that for every bounded sequence of optional times σ^N with respect to $\{(T_t^N, Z_t^N)\}_{t \geq 0}$, and for every $N \geq N_0$,

$$(47) \qquad P\left(\max\left\{ \left| \frac{1}{T_{\sigma+h}^N} - \frac{1}{T_\sigma^N} \right|, \left| \frac{1}{Z_{\sigma+h}^N} - \frac{1}{Z_\sigma^N} \right| \right\} \geq \epsilon \right) < \eta \, ,$$

where σ is short for σ^N in the subscripts.

Proposition 3.1 established that $\{(T_t^N, Z_t^N)\}_{t \geq 0}$ is a Markov process. By the strong Markov property, the conditional law of $T_{\sigma+h}^N - m^N$, given that $T_\sigma^N = m \overset{\text{def}}{=} m^N$, and $Z_\sigma^N = q^N$, is that same as that of the number of identifiable vertices in a Poisson($\hat{\beta}$) random hypergraph $\hat{\Lambda}^N$ on $\hat{N} \overset{\text{def}}{=} N - m$ vertices, where by the reasoning of Lemma 3.2 and the fact that $\rho_1 = 0$,

$$\hat{\beta}_1 \overset{\text{def}}{=} \frac{hN}{N-m} \sum_{k \geq 2} \rho_k \binom{m}{k-1} \binom{N-m}{1} \Big/ \binom{N}{k} \, .$$

Suppose $\epsilon > 0$ and $\eta > 0$ are given. In the case where $\min\{m^N, q^N\} > 1/\epsilon$, it follows that

$$(48) \qquad \max\left\{ \left| \frac{1}{T_{\sigma+h}^N} - \frac{1}{T_\sigma^N} \right|, \left| \frac{1}{Z_{\sigma+h}^N} - \frac{1}{Z_\sigma^N} \right| \right\} < \epsilon \, .$$

On the other hand, if $m^N \leq 1/\epsilon$, then

$$\hat{\beta}_1 \hat{N} \leq hN^2 \sum_{k \geq 2} \rho_k \binom{2/\epsilon}{k-1} \Big/ \binom{N}{k} = \frac{2h\rho_2}{\epsilon} + O(N^{-1}) \, .$$

Choose N_0 so large that, for $N \geq N_0$, the right side is not more than $3h\rho_2/\epsilon$; now it is true that, for any

$$h \leq \frac{-\epsilon \log(1 - \eta)}{3\rho_2} \, ,$$

and for any $N \geq N_0$, the probability that $\hat{\Lambda}^N$ has no patches, and hence no identifiable vertices nor identifiable hyperedges, is at least $1 - \eta$; in that case, $T_{\sigma+h}^N = T_\sigma^N$ and and $Z_{\sigma+h}^N = Z_\sigma^N$. In summary, for such N and h, (47) holds. Hence $\{(T_t^N, Z_t^N)\}_{t \geq 0}$ is tight, and (14) follows.

Step IV. As for (15) we need only check the convergence of finite-dimensional distributions, i.e. that

$$(49) \qquad \left((\bar{T}_{t_1}^N, \bar{Z}_{t_1}^N), \ldots, (\bar{T}_{t_r}^N, \bar{Z}_{t_r}^N)\right) \xrightarrow{\mathrm{d}} \left((\bar{T}_{t_1}, \bar{Z}_{t_1}), \ldots, (\bar{T}_{t_r}, \bar{Z}_{t_r})\right) .$$

for every finite set of time points $0 < t_1 < \ldots < t_r$. For the case $r = 1$, the validity of (49) follows from Theorem 6.2. For the sake of brevity, restrict our discussion of the case $r > 1$ to the $\bar{T}$ component; the argument for the $\bar{Z}$ component is similar. It suffices to show, for all $q = 2, \ldots, r$, and all $\epsilon > 0$, that

$$(50) \qquad P\left(\bigcap_{\substack{i,j \\ 1 \leq i < q \leq j \leq r}} \{\bar{T}_{t_i}^N < \epsilon\} \cap \{|\bar{T}_{t_j}^N - g(t_j)| < \epsilon\}\right) \to P\left(M_{t_{q-1}} < \infty = M_{t_q}\right) .$$

By our knowledge of the finite dimensional distributions from Theorem 6.2, the left side of is well approximated by

$$1 - P\left(\bar{T}_{t_{q-1}}^N \geq \epsilon\right) - P\left(\bar{T}_{t_q}^N \leq g(t_q) - \epsilon\right) ,$$

and for ϵ sufficiently small, this converges to the right side of (50).   ∎

8. FUTURE DIRECTIONS

We have not explained here the role of the upper envelope (4), even though it was included in the classification of structure functions. It is related to dual hypergraph collapse and the size of the core, as in Cooper (2002). We shall give the corresponding asymptotic results in a future paper.

ACKNOWLEDGMENTS

References

Aldous, D. 1989. *Probability Approximations via the Poisson Clumping Heuristic*, Springer.

Billingsley, P. 1999. *Convergence of Probability Measures*, Second edition, John Wiley.

Bollobás, B. 2001. *Random Graphs*, Second edition, Cambridge University Press.

Cooper, C. 2002. *The size of the cores of random graphs with a given degree sequence*, Preprint.

Darling, R.W.R. and J.R. Norris. 2004. *Structure of large random hypergraphs*, Ann. App. Probab., arXiv:math.PR/0109020, to appear.

Devroye, L. 1998. *Branching processes and their applications in the analysis of tree structures and tree algorithms*, Probabilistic Methods for Algorithmic Discrete Mathematics (M. Habib, C. Mc-Diarmid, J. Ramirez-Alfonsin, and B. Reed, eds.), Springer.

Duchet, P. 1995. *Hypergraphs*, Handbook of Combinatorics (R. Graham, M. Grötschel, and L. Lovász, eds.), Elsevier Science B.V.

Dwass, M. 1969. *The total progeny in a branching process*, Journal of Applied Probability **6**, 682–686.

Fountoulakis, N. 2002. *On the structure of the core of sparse random graphs*, Preprint.

Goldschmidt, C.A. and J.R. Norris. 2002. *Essential edges of Poisson random hypergraphs*, Random Structures & Algorithms, to appear.

Kallenberg, O. 2002. *Foundations of Modern Probability*, Second edition, Springer.

Jacod, J. and A.N. Shiryaev. 1987. *Limit Theorems for Stochastic Processes*, Springer.

Pittel, B., J. Spencer, and N. Wormald. 1996. *Sudden emergence of a giant k-core in a random graph*, Journal of Combinatorial Theory, Series B **67**, 111–151.

National Security Agency, P.O. Box 535, Annapolis Junction, MD 20701
  E-mail address: `rwrd@afterlife.ncsc.mil`

Department of Mathematics, The University of Utah, 155 S. 1400 E., Salt Lake City, UT 84112–0090
  E-mail address: `levin@math.utah.edu`

Statistical Laboratory, Centre For Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB
  E-mail address: `j.r.norris@statslab.cam.ac.uk`