Dense Peelable Random Uniform Hypergraphs

Martin Dietzfelbinger
Technische Universität Ilmenau, Germany
martin.dietzfelbinger@tu-ilmenau.de

Stefan Walzer
Technische Universität Ilmenau, Germany
stefan.walzer@tu-ilmenau.de

Abstract

We describe a new family of k-uniform hypergraphs with independent random edges. The hypergraphs have a high probability of being *peelable*, i.e. to admit no sub-hypergraph of minimum degree 2, even when the edge density (number of edges over vertices) is close to 1.

In our construction, the vertex set is partitioned into linearly arranged *segments* and each edge is incident to random vertices of k consecutive segments. Quite surprisingly, the linear geometry allows our graphs to be peeled "from the outside in". The density thresholds f_k for peelability of our hypergraphs ($f_3 \approx 0.918$, $f_4 \approx 0.977$, $f_5 \approx 0.992$, ...) are well beyond the corresponding thresholds ($c_3 \approx 0.818$, $c_4 \approx 0.772$, $c_5 \approx 0.702$, ...) of standard k-uniform random hypergraphs.

To get a grip on f_k, we analyse an idealised peeling process on the random weak limit of our hypergraph family. The process can be described in terms of an operator on $[0,1]^{\mathbb{Z}}$ and f_k can be linked to thresholds relating to the operator. These thresholds are then tractable with numerical methods.

Random hypergraphs underlie the construction of various data structures based on hashing, for instance invertible Bloom filters, perfect hash functions, retrieval data structures, error correcting codes and cuckoo hash tables, where inputs are mapped to edges using hash functions. Frequently, the data structures rely on peelability of the hypergraph, or peelability allows for simple linear time algorithms. Memory efficiency is closely tied to edge density while worst and average case query times are tied to maximum and average edge size.

To demonstrate the usefulness of our construction, we used our 3-uniform hypergraphs as a drop-in replacement for the standard 3-uniform hypergraphs in a retrieval data structure by Botelho et al. [8]. This reduces memory usage from $1.23m$ bits to $1.12m$ bits (m being the input size) with almost no change in running time. Using $k > 3$ attains, at small sacrifices in running time, further improvements to memory usage.

2012 ACM Subject Classification Theory of computation → Data structures design and analysis

Keywords and phrases Random Hypergraphs, Peeling Threshold, 2-Core, Hashing, Retrieval, Succinct Data Structure

Digital Object Identifier 10.4230/LIPIcs.ESA.2019.38

1 Introduction

The *core* of a hypergraph $H = (V, E)$ is the largest sub-hypergraph of H with minimum degree at least 2. The core can be obtained by *peeling*, which means repeatedly choosing a vertex of degree 0 or 1 and removing it (and the incident edge if present) from the hypergraph, until no such vertex exists. If the core of H is empty, then H is called *peelable*.

The significance of peelability. Hypergraphs underlie many hashing based data structures and peelability is often necessary for proper operation or allows for simple linear time algorithms. We list a few examples.

- **Invertible Bloom Lookup Tables.** IBLTs [22] are based on Bloomier filters [10] which are based on Bloom filters [4]. Each element is inserted in several random positions in a hash table. Any cell stores the XOR of all elements that have been inserted into it. A LIST-ENTRIES query on an IBLT can recover all elements of the table precisely if the underlying hypergraph is peelable. Among other things, IBLTs have been used to construct error correcting codes [34] and to solve the set reconciliation and straggler identification problems [16].
- **Erasure Correcting Codes.** To construct capacity achieving erasure codes, the authors of [28] consider a hypergraph where V corresponds to parity check bits and E to message bits that were lost during transmission. A message bit is incident to precisely those check bits to which it contributed. Correct decoding hinges on peelability of the hypergraph.
- **Cuckoo Hashing and XORSAT.** In the context of cuckoo hash tables [14, 31, 36] and solving random XORSAT formulas [15, 19, 37], (partial) peelability of the underlying hypergraph makes placing all (some) keys or solving the linear system (eliminating some variables) particularly simple.
- **Retrieval and Perfect Hashing.** The retrieval problem (considered later in Section 7) occurs in the context of constructing perfect hash functions [3, 6, 7, 8, 30]. The known approaches involve finding a solution $z : V \to R$ for a system $(\sum_{v \in e} z(v) = f(e))_{e \in E}$ of equations where $H = (V, E)$ is a hypergraph, $f : E \to R$ a function and R a small set. If R is a field, then the incidence matrix of H needs to have full rank over R to guarantee the existence of a solution. If H is peelable however, then the existence of a solution is guaranteed even if R only has a group structure. Moreover, it can be computed in linear time.

In these contexts, the hypergraph typically has vertex set $[n] = \{1, \dots, n\}$ and for each element x of an input set S, an edge $e_x \subset [n]$ is created with incidences chosen via hash functions. For theoretical considerations, the edges $(e_x)_{x \in S}$ are often assumed to be independent random variables. This has proven to be a good model for practical settings, even though perfect independence is not achieved by most practical hash functions. An important choice left to the algorithm designer is the distribution of e_x.

Previous work. If the distribution is such that $\mathcal{O}(n)$ edges have size 2 or less (in particular if H is a graph with $\mathcal{O}(n)$ edges), then – due to the well-known "birthday paradox" – there is a constant probability that an edge is repeated. In that case, H is clearly not peelable. The simplest workable candidate for the distribution of e_x is therefore to pick a constant $k \geq 3$ and let e_x contain k vertices chosen independently and uniformly at random. We refer to these standard hypergraphs as *k-uniform Erdős-Renyi hypergraphs* $H^k_{n,cn}$ where c is the *edge density*, i.e. the number of edges over the number of vertices. Corresponding *peelability thresholds* c_k have been determined in [35] meaning if $c < c_k$ then $H^k_{n,cn}$ is peelable with high probability (whp), i.e. with probability approaching 1 as $n \to \infty$ and if $c > c_k$ then $H^k_{n,cn}$ is not peelable whp. The largest threshold is $c_3 \approx 0.818$. Since the edge density is often tightly linked to a performance metric (e.g. memory efficiency of a dictionary, rate of a code) a density closer to 1 would be desirable, but we know of only two alternative constructions.

To obtain erasure codes with high rates the authors of [28] construct for any $D \in \mathbb{N}$ hypergraphs with edge sizes in $\{5, \dots, D + 4\}$, average edge size $\approx \ln D + 3$ and edge density $1 - 1/D$ that are peelable whp. In particular, this yields peelable hypergraphs with edge densities arbitrarily close to 1. A downside is that the high maximum edge size can lead to worst case query times of $\Theta(D)$ in certain contexts. Motivated by this, the author of

Table 1 The erosion thresholds er_k and peelability thresholds f_k for k-ary fuse graphs satisfy $b_k \le \mathrm{er}_k \le f_k \le c_k^*$. The values B_k play a role in Section 5.

k	3	4	5	6	7
b_k	0.9179352469	0.9767692112	0.9924345766	0.9973757381	0.9990561294
c_k^*	0.9179352767	0.9767701649	0.9924383913	0.9973795528	0.9990637588
B_k	0.9179353065	0.9767711186	0.9924422067	0.9973833675	0.9990713882
$\Rightarrow f_k \approx$	0.917935	0.97677	0.99243	0.99738	0.99906

[39] looked into non-uniform hypergraphs with constant maximum edge size. Focusing on hypergraphs with two admissible edge sizes, he found for example that mixing edges of size 3 and size 21 yields a family of hypergraphs with peelability threshold ≈ 0.92.

Our construction. In this paper we introduce and analyse a new distribution on edges that yields k-uniform hypergraphs with high peelability thresholds that perform well in practical algorithms.

We call our hypergraphs *fuse graphs* (as in the cord attached to a firecracker). There is an underlying linear geometry and similar to how fire proceeds linearly through a lit fuse, the peeling process proceeds linearly through our hypergraphs, in the sense that vertices on the inside of the line tend to only become peelable after vertices closer to the end of the line have already been removed.

Formally, for $k \ge 3$, $\ell \in \mathbb{N}$ and $c \in \mathbb{R}^+$ we define the family $(F(n,k,c,\ell))_{n\in\mathbb{N}}$ of k-uniform fuse graphs as follows. The vertex set is $V = \{1,\dots,n(\ell+k-1)\}$ where for $i \in I := \{0,\dots,\ell+k-2\}$ the vertices $\{in+1,\dots,(i+1)n\}$ form the i-th *segment*[1]. The edge set E has size $cn\ell$. Each edge $e \in E$ is independently determined by one uniformly random variable $j \in J := \{0,\dots,\ell-1\}$ denoting the *type* of e and k independent random variables $o_0,\dots,o_{k-1}$ uniformly distributed in $[n]$, yielding $e = \{(j+t)n + o_t \mid t \in \{0,\dots,k-1\}\}$. In other words, e contains one uniformly random vertex from each segment $j, j+1,\dots,j+k-1$. There may be repeating edges but the probability that his happens is $\mathcal{O}(1/n)$. The edge density $c\frac{\ell}{\ell+k-1}$ approaches c for $\ell \gg k$.

Results. Let the *peelability threshold* for k-ary fuse graphs be defined as

$$f_k := \sup\{c \in \mathbb{R}^+ \mid \forall \ell \in \mathbb{N} : \Pr[F(n,k,c,\ell) \text{ is peelable}] \xrightarrow{n\to\infty} 1\}.$$

Our Main Theorem relates f_k to the *orientability threshold* c_k^* of k-ary Erdős-Renyi hypergraphs and the *erosion threshold* er_k defined in the technical part of our paper.

▶ **Theorem 1.** *For any $k \ge 3$ we have $\mathrm{er}_k \le f_k \le c_k^*$.*

The orientability thresholds c_k^* are known exactly [11, 19, 20] and we determine lower bounds on the erosion thresholds er_k. As shown in Table 1, this makes it possible to narrow down f_k to an interval of size 10^{-5} for all $k \in \{3,\dots,7\}$.

[1] Denoting the segment size by n instead of the number of vertices is more convenient. Note that $|V| = \Theta(n)$ still holds.

Outline. The paper is organised as follows. In Section 2 we idealise the peeling process by switching to the *random weak limit* of our hypergraphs, and capture the essential behaviour of the process in terms of an operator $\hat{\mathbf{P}}$ acting on functions $q : \mathbb{Z} \to [0, 1]$. For this operator, we identify the properties of being *eroding* and *consolidating* as well as corresponding thresholds er_k and co_k in Section 3. We then prove the "$\mathrm{er}_k \leq f_k$" part of our theorem in Section 4 and give numerical approximations of er_k and co_k in Section 5. The comparatively simple "$f_k \leq c_k^*$" part of our theorem is independent of these considerations and is proved in Section 6. Finally, in Section 7 we demonstrate how using our hypergraphs can improve the performance of practical retrieval data structures.

2 The Peeling Process and Idealised Peeling Operators

In this section we consider how the probabilities for vertices to "survive" $r \in \mathbb{N}$ rounds of peeling changes from one round to the next. In the classical setting this could be described by a function, mapping the old probability to the new one [35]. In our case, however, there are distinct probabilities for each segment of the graph. Thus we need a corresponding operator $\hat{\mathbf{P}}$ that acts on *sequences* of probabilities. Conveniently, it will be independent of n and ℓ.

We almost always suppress n, k, c, ℓ in notation outside of definitions, assuming n to be large. Big-$\mathcal{O}$ notation refers to $n \to \infty$ while k, c, ℓ are constant.

Consider the parallel peeling process $\mathsf{peel}(F)$ on $F = F(n, k, c, \ell)$. In each *round* of $\mathsf{peel}(F)$, all vertices of degree 0 or 1 are determined and then deleted simultaneously. Deleting a vertex implicitly deletes incident edges. We also define the *rooted peeling process* $\mathsf{peel}_v(F)$ for any vertex $v \in V$, which behaves exactly like $\mathsf{peel}(F)$ except that the special vertex v may only be deleted if it has degree 0, not if it has degree 1. For any $i \in I$ and $r \in \mathbb{N}_0$ we let $q^{(r)}(i) = q^{(r)}(i, n, k, c, \ell)$ be the probability that a vertex v of segment i survives r rounds of $\mathsf{peel}_v(F)$, i.e. is not deleted. Note that the probability is well-defined as vertices of the same segment are symmetric.

By definition, $q^{(0)}(i) = 1$ for all $i \in I$. Whether a vertex v of segment $i \in I$ survives $r > 0$ rounds is a function of its r-neighbourhood $N(n, v, r)$, i.e. the set of vertices and edges of F that can be reached from v by traversing at most r hyperedges.

It is standard to consider the *random weak limit* of F to get a grip on the distribution of $N(n, v, r)$ and thus on $q^{(r)}(i)$. Intuitively, we identify a (possibly infinite) random tree that captures the local characteristics of F for $n \to \infty$. See [1] for a good survey with examples and details on how to formally define the underlying topology and metric space. In the limit, the binomially distributed vertex degrees (e.g. $\mathrm{Bin}(cn\ell, \frac{1}{n\ell})$ for vertices of segment 0) become Poisson distributed ($\mathrm{Po}(c)$ for segment 0). Short cycles are not only rare but non-existent and certain weakly correlated random variables become perfectly independent.

▶ **Definition 2** (Limiting Tree). *Let $k, \ell \in \mathbb{N}$, $c \in \mathbb{R}^+$ and $i \in I$. The random (possibly infinite) hypertree $T_i = T_i(k, c, \ell)$ is distributed as follows.*

T_i has a root vertex $\mathrm{root}(T_i)$ *of segment*[2] *i which for each $j \in \{i - k + 1, \ldots, i\} \cap J$ has $d_j \sim \mathrm{Po}(c)$* child edges *of type j. Each child edge of type j is incident to $k - 1$ (fresh)* child vertices *of its own, one for each segment $i' \in \{j, \ldots, j + k - 1\} \setminus \{i\}$. The sub-hypertree at such a child vertex of segment i' is distributed recursively (and independently of its sibling-subtrees) according to $T_{i'}$.*

[2] In the current context, the segment of a vertex is an abstract label. There can be an unbounded number of vertices of each segment.

Since all arguments are standard in contexts where local weak convergence plays a role, we state the following lemma without proof. For instance, a full argument to show a similar convergence is given in [25]. See also [24] for the related technique of Poissonisation.

▶ **Lemma 3.** *Let $r \in \mathbb{N}$ be constant. Let further $N(n,v,r)$ be the r-neighbourhood of a vertex v of segment i in F and $T_i^{(r)}$ the r-neighbourhood of* $\mathrm{root}(T_i)$*, both viewed as undirected and unlabelled hypergraphs. Then $N(n,v,r)$ converges in distribution to $T_i^{(r)}$ as $n \to \infty$.*

We now direct our attention to survival probabilities in the idealised peeling processes $(\mathsf{peel}_{\mathrm{root}(T_i)}(T_i))_{i\in I}$, which are easier to analyse than those of $\mathsf{peel}_v(F)$.

▶ **Lemma 4.** *Let $r \in \mathbb{N}_0$ be constant and $q_T^{(r)}(i) = q_T^{(r)}(i,k,c,\ell)$ be the probability that* $\mathrm{root}(T_i)$ *survives r rounds of* $\mathsf{peel}_{\mathrm{root}(T_i)}(T_i)$ *for $i \in I$. Then*

$$q_T^{(r+1)}(i) = 1 - \exp\Bigg(- c \sum_{j\in\{i-k+1,\dots,i\}\cap J} \prod_{\substack{j\le i'<j+k\\ i'\neq i}} q_T^{(r)}(i')\Bigg) \qquad \text{for } i \in I.$$

Proof. Let $i \in I$ and $v = \mathrm{root}(T_i)$. Assume $j \in \{i-k+1,\dots,i\}\cap J$ is the type of some edge e incident to v. Edge e survives r rounds of $\mathsf{peel}_v(T_i)$ if and only if all of its incident vertices survive these r rounds. Since v itself may not be deleted by $\mathsf{peel}_v(T_i)$ as long as e exists, the relevant vertices are the $k-1$ child vertices, one for each segment $i' \in \{j,\dots,j+k-1\}-\{i\}$. Call these $w_1,\dots,w_{k-1}$ and denote the subtrees rooted at those vertices by $W_1,\dots,W_{k-1}$. Now consider the peeling processes $\mathsf{peel}_{w_1}(W_1),\dots,\mathsf{peel}_{w_{k-1}}(W_{k-1})$. Assume one of them, say $\mathsf{peel}_{w_s}(W_s)$, deletes w_s in round $r' \le r$, meaning w_s has degree 0 before round r'. It follows that w_s has degree at most 1 before round r' in $\mathsf{peel}_v(T_i)$, meaning $\mathsf{peel}_v(T_i)$ deletes e in round r' (or earlier). Conversely, if none of $\mathsf{peel}_{w_1}(W_1),\dots,\mathsf{peel}_{w_{k-1}}(W_{k-1})$ delete their root vertex within r rounds, then $w_1,\dots,w_{k-1}$ have degree at least 2 after round r of $\mathsf{peel}_v(T_i)$ and e survives round r of $\mathsf{peel}_v(T_i)$. This makes the probability for e to survive r rounds of $\mathsf{peel}_v(T_i)$ equal to $p_{ij} := \prod_{j\le i'<j+k, i'\neq i} q_T^{(r)}(i')$. Since the number m_{ij} of edges of type j incident to v has distribution $m_{ij} \sim \mathrm{Po}(c)$, the number m'_{ij} of edges of type j incident to v surviving r rounds of $\mathsf{peel}_v(T_i)$ is a correspondingly *thinned out* variable, namely $m'_{ij} \sim \mathrm{Bin}(m_{ij}, p_{ij})$, which means $m'_{ij} \sim \mathrm{Po}(cp_{ij})$.

The claim now follows by observing that v survives $r+1$ rounds of $\mathsf{peel}_v(T_i)$ if and only if at least one of its child edges survives r rounds of $\mathsf{peel}_v(T_i)$:

$$\begin{aligned} q_T^{(r+1)}(i) &= \Pr[v \text{ survives } r+1 \text{ rounds of } \mathsf{peel}_v(T_i)] = 1 - \Pr\Big[\bigcap_{j\in\{i-k+1,\dots,i\}\cap J} \{m'_{ij} = 0\}\Big] \\ &= 1 - \prod_{j\in\{i-k+1,\dots,i\}\cap J} \Pr[m'_{ij} = 0] = 1 - \prod_{j\in\{i-k+1,\dots,i\}\cap J} \exp(-cp_{ij}) = 1 - \exp(-c \sum_{j\in\{i-k+1,\dots,i\}\cap J} p_{ij}). \end{aligned}$$

Replacing p_{ij} with its definition completes the proof. ◀

For convenience we define, for $k \ge 3, \ell \in \mathbb{N}$ and $c \in \mathbb{R}^+$, the operator $\mathbf{P} = \mathbf{P}(k,c,\ell)$, which maps any $q : I \to [0,1]$ to $\mathbf{P}q : I \to [0,1]$ with

$$(\mathbf{P}q)(i) = 1 - \exp\Bigg(- c \sum_{j\in\{i-k+1,\dots,i\}\cap J} \prod_{\substack{j\le i'<j+k\\ i'\neq i}} q(i')\Bigg) \quad \text{for } i \in I.$$

Together Lemmas 3 and 4 imply that $\mathbf{P}$ can be used to approximate survival probabilities.

▶ **Corollary 5.** *Let $r \in \mathbb{N}_0$ be constant. Then for all $i \in I$*

$$\mathbf{P}^r q^{(0)}(i) \overset{\text{def}}{=} \mathbf{P}^r q_T^{(0)}(i) \overset{\text{Lem 4}}{=} q_T^{(r)}(i) \overset{\text{Lem 3}}{=} q^{(r)}(i) \pm o(1).$$

To obtain *upper* bounds on survival probabilities, we may remove the awkward restriction "$\cap J$" in the definition of $\mathbf{P}$. We define $\hat{\mathbf{P}} = \hat{\mathbf{P}}(k,c)$ as mapping $q : \mathbb{Z} \to [0,1]$ to $\hat{\mathbf{P}}q : \mathbb{Z} \to [0,1]$ with

$$(\hat{\mathbf{P}}q)(i) = 1 - \exp\Bigg(- c \sum_{j=i-k+1}^{i} \prod_{\substack{j \le i' < j+k \\ i' \neq i}} q(i') \Bigg) \quad \text{for } i \in \mathbb{Z}.$$

Note that $\hat{\mathbf{P}}$ does not depend on ℓ or n. To simplify notation, we assume that the old operator $\mathbf{P}$ also acts on functions $q : \mathbb{Z} \to [0,1]$, ignoring $q(i)$ for $i \notin I$, and producing $\mathbf{P}q : \mathbb{Z} \to [0,1]$ with $\mathbf{P}q(i) = 0$ for $i \notin I$. We also extend $q^{(0)}$ to be $\mathbb{1}_I : \mathbb{Z} \to [0,1]$, i.e. the characteristic function on I, essentially introducing vertices of segments $i \notin I$ which are, however, already deleted with probability 1 before the first round begins. Note that while $q^{(r)}(i)$ and $q_T^{(r)}(i)$ are by definition non-increasing in r, this is not the case for $(\hat{\mathbf{P}}^r q^{(0)})(i)$. For instance, $\hat{\mathbf{P}}^r q^{(0)}$ has support $\{-r, -r+1, \dots, \ell + k - 2 + r\}$, which grows with r.[3] The following lemma lists a few easily verified properties of $\hat{\mathbf{P}}$. All inequalities between functions should be interpreted point-wise.

▶ **Lemma 6.**

(i) $\forall q : \mathbb{Z} \to [0,1] \colon \mathbf{P}q \le \hat{\mathbf{P}}q$.

(ii) $\hat{\mathbf{P}}$ *commutes with the shift operators* $\ll$ *and* $\gg$ *defined via* $(\ll q)(i) = q(i+1)$ *and* $(\gg q)(i) = q(i-1)$. *In other words, we have* $\forall q : \mathbb{Z} \to [0,1] \colon \hat{\mathbf{P}}(\ll q) = \ll(\hat{\mathbf{P}}q) \wedge \hat{\mathbf{P}}(\gg q) = \gg(\hat{\mathbf{P}}q)$.

(iii) $\hat{\mathbf{P}}$ *is monotonic, i.e.* $\forall q, q' : \mathbb{Z} \to [0,1] \colon q \le q' \Rightarrow \hat{\mathbf{P}}q \le \hat{\mathbf{P}}q'$.

(iv) $\hat{\mathbf{P}}$ *respects monotonicity, i.e. if $q(i)$ is (strictly) increasing in i, then so is $(\hat{\mathbf{P}}q)(i)$.*

3 Two Fixed Points Battling for Territory

In this section we define the *erosion* and *consolidation thresholds* at which the behaviour of $\hat{\mathbf{P}}$ changes in crucial ways.

First, we require a few facts about the function $f : [0,1] \to [0,1]$ mapping $x \mapsto 1 - e^{-ckx^{k-1}}$. It appears in the analysis of cores in k-ary Erdős-Renyi hypergraphs $H^k_{n,cn}$, essentially mapping the probability ρ_r for a vertex to survive r rounds of peeling to the probability $\rho_{r+1} = f(\rho_r)$ to survive $r+1$ rounds of peeling, see [35, page 5][4].

The threshold c_k for the appearance of a core in $H^k_{n,cn}$ turns out to be the threshold for the appearance of a non-zero fixed point of f. The following is implicit in the analysis.

▶ **Fact 7** ([35, Proofs of Lemmas 3 and 4])**.**

(i) *For $c < c_k$, f has only the fixed point $f(0) = 0$, with $f'(0) < 1$.*

(ii) *For $c > c_k$, there are exactly three fixed points 0, $\xi_1 = \xi_1(k,c)$ and $\xi_2 = \xi_2(k,c)$ where $f'(\xi_1) > 1$ while $f'(0), f'(\xi_2) < 1$.*

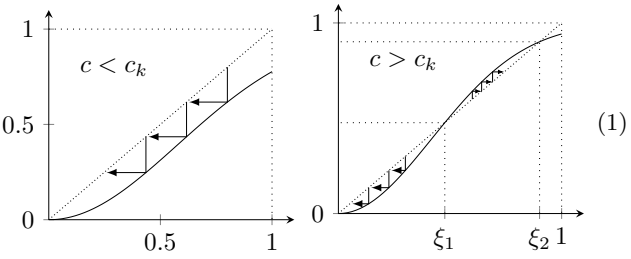
---

[3] It is still possible to interpret $\hat{\mathbf{P}}^r q^{(0)}(i)$ as survival probabilities in more symmetric extended versions $\hat{T}_i$ of the tree T_i, but we will not pursue this.

[4] Our setting corresponds to the choices $(r_{\text{Molloy}}, k_{\text{Molloy}}, c_{\text{Molloy}}) = (k, 2, c \cdot (k-1)!)$.

This implies the following behaviour of applying f repeatedly to a starting value x. This should be immediately clear from the sketches on the right.

$$f^r(x) \overset{r\to\infty}{\longrightarrow} \begin{cases} 0 & \text{if } c < c_k \wedge x \in [0,1], \\ 0 & \text{if } c > c_k \wedge x \in [0,\xi_1), \\ \xi_2 & \text{if } c > c_k \wedge x \in (\xi_1, 1]. \end{cases} \tag{1}$$



Note that f captures the behaviour of $\hat{\mathbf{P}}$ on constant functions $\text{const}_x(i) := x$, in the sense that $\hat{\mathbf{P}}\text{const}_x = \text{const}_{f(x)}$. For $c < c_k$ we therefore have for all $i \in I$

$$\mathbf{P}^r q^{(0)}(i) \overset{\text{Cor 5}}{=} q^{(r)}(i) \pm o(1) \text{ and } \mathbf{P}^r q^{(0)} \le \hat{\mathbf{P}}^r q^{(0)} \le \hat{\mathbf{P}}^r \text{const}_1 = \text{const}_{f^r(1)} \overset{r\to\infty}{\longrightarrow} \text{const}_0.$$

In conjunction with a later lemma, this is sufficient to show that F is peelable whp in this case. A similar argument for $c = c_k$ is possible as well. Our focus from now on is therefore on the interesting case $c > c_k$ where the three distinct fixed points 0, ξ_1, ξ_2 of f exist.

We give an intuitive account of the phenomenon underlying the following steps before continuing formally. Due to (1) we have

$$\hat{\mathbf{P}}^r \text{const}_x \overset{r\to\infty}{\longrightarrow} \begin{cases} \text{const}_0 & \text{for } x < \xi_1 \\ \text{const}_{\xi_2} & \text{for } x > \xi_1. \end{cases}$$

Now consider what happens if we iterate $\hat{\mathbf{P}}$ on a function that is "torn" between these two cases. Concretely, let us consider the function step_0^1 where we define $\text{step}_x^y : \mathbb{Z} \to [0,1]$ to have value y on $\mathbb{N}_0$ and value x on negative inputs. Should we expect $\hat{\mathbf{P}}^r \text{step}_0^1$ to converge to const_0 or const_{ξ_2} as r increases? It turns out both is possible, depending on c.

Speaking more generally, let $q : \mathbb{Z} \to [0,1]$ be any function. If $N(i) := \{i-k+1, \dots, i+k-1\} \setminus \{i\}$ then $\hat{\mathbf{P}}q(i)$ depends (monotonically) on $(q(i'))_{i' \in N(i)}$. It is clear that if $q(i') < \xi_1$ for all $i' \in N(i)$, then $\hat{\mathbf{P}}q(i) < \xi_1$ as well. Similarly, if $q(i') > \xi_1$ for all $i' \in N(i)$ then $\hat{\mathbf{P}}q(i) > \xi_1$. If, however, there are indices $i'_1, i'_2 \in N(i)$ with $q(i'_1) < \xi_1 < q(i'_2)$ then $\hat{\mathbf{P}}q(i)$ could be above or below ξ_1; in this case we call the index i *contested*.

The contested area of step_0^1 is $[-k+1, k-2]$. Iterating $\hat{\mathbf{P}}$ we obtain $\hat{\mathbf{P}}^r \text{step}_0^1$ for $r \in \mathbb{N}_0$. For all $r \in \mathbb{N}_0$ the contested area is an interval of size $2k-2$ with all values to the left of it (towards $-\infty$) less than ξ_1 and all values to the right of it (towards ∞) bigger than ξ_1. However, the contested area may *shift*. If the domain of values bigger than ξ_1 is shrinking (*"eroding"*), then we see convergence to const_0. If conversely it is growing ("consolidating"), then we see convergence to const_{ξ_2}. In Figure 1 we visualise these effects. There is only a small range of values c where both fixed points seem equally "strong" and the same area remains perpetually contested.

With this in mind, we make the following definitions. For a compact formulation in the coarse terms of shifts ("$\ll$", "$\gg$") and point-wise inequalities ("$<$", "$>$") we use slightly different step functions.

▶ **Definition 8.** *Let $k \ge 3$, $c \in \mathbb{R}^+$ and $\hat{\mathbf{P}} = \hat{\mathbf{P}}(k,c)$ as above. We say*

$$\hat{\mathbf{P}} \textit{ is eroding if } \exists R \in \mathbb{N}\colon \hat{\mathbf{P}}^R \text{step}^1_{\xi_1/2} < \gg \text{step}^1_{\xi_1/2}$$

and $\hat{\mathbf{P}}$ is consolidating if $\exists R \in \mathbb{N}\colon \hat{\mathbf{P}}^R \text{step}_0^{(\xi_1+\xi_2)/2} > \ll \text{step}_0^{(\xi_1+\xi_2)/2}$.

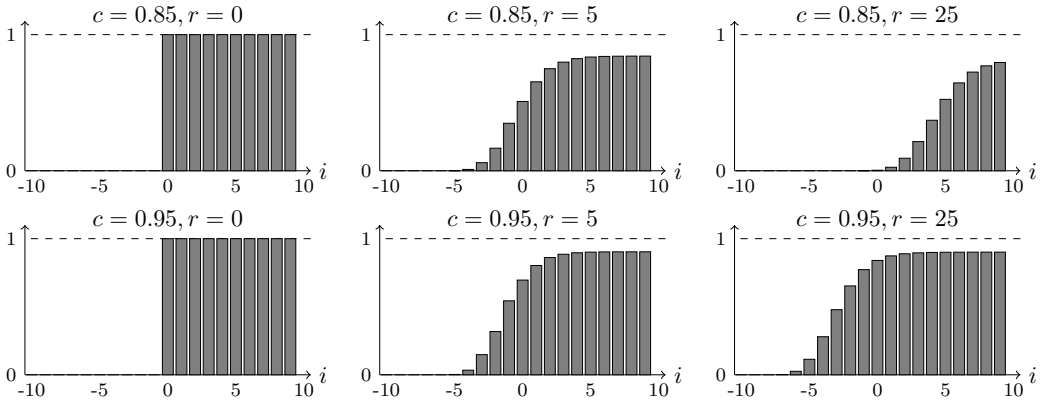


Figure 1 Depiction of $\hat{\mathbf{P}}^r \mathrm{step}_0^1$ for $c \in \{0.85, 0.95\}$ and $r \in \{0, 5, 25\}$ on the range $i \in \{-10, \dots, 10\}$. The phenomenon of *erosion* can be seen on the top with the plot seemingly moving towards the right between $r = 5$ and $r = 25$. Similarly, *consolidation* can be seen on the bottom.

We define the corresponding *erosion* and *consolidation thresholds* as

$$\mathrm{er}_k = \sup\{c \in \mathbb{R}^+ \mid \hat{\mathbf{P}}(k, c) \text{ is eroding}\}, \qquad \mathrm{co}_k = \inf\{c \in \mathbb{R}^+ \mid \hat{\mathbf{P}}(k, c) \text{ is consolidating}\}.$$

Note that $c < \mathrm{er}_k$ implies that $\hat{\mathbf{P}}(k, c)$ is eroding and $c > \mathrm{co}_k$ implies $\hat{\mathbf{P}}(k, c)$ is consolidating as would be expected. This uses that the definition of $\hat{\mathbf{P}}$ is monotonic in c.

The following lemma states that erosion (consolidation) are sufficient conditions for const_0 (const_{ξ_2}) to "win the battle" when iterating $\hat{\mathbf{P}}$ on step_0^1.

▶ **Lemma 9.** *Let $k \geq 3$.*

(i) *If $c < \mathrm{er}_k$ and $i \in \mathbb{Z}$, then $\hat{\mathbf{P}}^r \mathrm{step}_0^1(i) \overset{r\to\infty}{\longrightarrow} 0$.*

(ii) *If $c > \mathrm{co}_k$ and $i \in \mathbb{Z}$, then $\hat{\mathbf{P}}^r \mathrm{step}_0^1(i) \overset{r\to\infty}{\longrightarrow} \xi_2$.*

(iii) $\mathrm{er}_k \leq \mathrm{co}_k$.

Proof.

(i) Let $R \in \mathbb{N}$ be the witness to the fact that $\hat{\mathbf{P}}(k, c)$ is eroding and $i \in \mathbb{Z}$ arbitrary.

$$\begin{aligned}
\lim_{r\to\infty} (\hat{\mathbf{P}}^r \mathrm{step}_0^1)(i) &\leq \lim_{r\to\infty} (\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2})(i) = \lim_{r\to\infty} (\hat{\mathbf{P}}^r ((\hat{\mathbf{P}}^R)^{kr} \mathrm{step}^1_{\xi_1/2}))(i) \\
&\leq \lim_{r\to\infty} (\hat{\mathbf{P}}^r (\gg^{kr} \mathrm{step}^1_{\xi_1/2}))(i) = \lim_{r\to\infty} (\gg^{kr} (\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2}))(i) \\
&= \lim_{r\to\infty} (\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2})(i - kr) = \lim_{r\to\infty} (\hat{\mathbf{P}}^r \mathrm{const}_{\xi_1/2})(i - kr) \\
&= \lim_{r\to\infty} \mathrm{const}_{f^r(\xi_1/2)}(i - kr) = \lim_{r\to\infty} f^r(\xi_1/2) = 0.
\end{aligned}$$

When replacing $\mathrm{step}^1_{\xi_1/2}$ by $\mathrm{const}_{\xi_1/2}$ we exploited that $(\hat{\mathbf{P}}q)(i)$ depends only on the values $q(i')$ for $i' \in \{i - k + 1, \dots, i + k - 1\}$ and thus $(\hat{\mathbf{P}}^r q)(i)$ depends only on the values $q(i')$ for $i' \in [i - (k-1)r, i + (k-1)r]$.

(ii) The proof is analogous to the proof of (i).

(iii) This is clear, since the implications of (i) and (ii) are mutually exclusive. ◀

4 Erosion is Sufficient for Peeling

We now connect the phenomenon of erosion to the survival probabilities $q^{(R)}(i)$ we were originally interested in. For $c < \mathrm{er}_k$ and any $\ell \in \mathbb{N}$, they can be made smaller than any $\delta > 0$ in $R = R(\delta, \ell)$ rounds. For $c > \mathrm{co}_k$ and ℓ sufficiently large, no constant number of rounds suffices to reduce all survival probabilities below ξ_1.

▶ **Lemma 10.** *Let $k \geq 3$.*

(i) *If $c < \mathrm{er}_k$ then $\forall \ell \in \mathbb{N}, \delta > 0 \colon \exists R, N \in \mathbb{N} \colon \forall n \geq N, i \in I : q^{(R)}(i) < \delta$.*

(ii) *If $c > \mathrm{co}_k$ then $\exists L = L(k,c) : \forall \ell \geq L \colon \exists i \in I : \lim_{r\to\infty} \lim_{n\to\infty} q^{(r)}(i) > \xi_1$.*

Proof.

(i) Let $\ell \in \mathbb{N}$ and $\delta > 0$ be arbitrary constants. Using (i) from Lemma 9, there exists a constant R such that $\hat{\mathbf{P}}^R \mathrm{step}_0^1(i) \leq \delta/2$ for all $i \in I$. Therefore for $i \in I$:

$$q^{(R)}(i) \overset{\text{Cor 5}}{=} (\mathbf{P}^R q^{(0)})(i) + o(1) \leq (\hat{\mathbf{P}}^R q^{(0)})(i) + o(1) \leq (\hat{\mathbf{P}}^R \mathrm{step}_0^1)(i) \leq \delta/2 + o(1).$$

which implies the existence of an appropriate $N \in \mathbb{N}$.

(ii) Let $R \in \mathbb{N}$ be the witness to the fact that $\hat{\mathbf{P}}(k,c)$ is consolidating and let $\ell \geq L(k,c) := 4d$ for $d = (k-1)R$. Consider the function $q^* : \mathbb{Z} \to [0,1]$ defined as $q^* = \mathbb{1}_{\{d,\dots,\ell-d-1\}} \cdot (\xi_1 + \xi_2)/2$, i.e. the function with value $(\xi_1+\xi_2)/2$ on its support $\{d, \dots, \ell-d-1\}$ and 0 outside of it. For any $d \leq i < \ell - 2d$ we have

$$\begin{aligned}\mathbf{P}^R q^*(i) = \hat{\mathbf{P}}^R q^*(i) = \hat{\mathbf{P}}^R \mathrm{step}_0^{(\xi_1+\xi_2)/2}(i-d) &\geq \lll \mathrm{step}_0^{(\xi_1+\xi_2)/2}(i-d) \\ &= (\xi_1+\xi_2)/2 = q^*(i).\end{aligned}$$

For the first equality, we exploited that i is so far from the borders of $I = \{0, \dots, \ell-1\}$ that there is no difference between $\mathbf{P}$ and $\hat{\mathbf{P}}$. For the second equality we used that only the values of q^* on $\{i-d, \dots, i+d\}$ play a role and q^* is a (shifted) step function on that domain. By mirroring, the same argument can be made to get $\mathbf{P}^R q^*(i) \geq q^*(i)$ for $2d \leq i < \ell - d$ as well and thus the point-wise inequality $\mathbf{P}^R q^* \geq q^*$. Since $q^{(0)} \geq q^*$ we get

$$\lim_{r\to\infty} \lim_{n\to\infty} q^{(r)} \overset{\text{Cor 5}}{=} \lim_{r\to\infty} \mathbf{P}^r \mathbb{1}_I \geq \lim_{r\to\infty} \mathbf{P}^r q^* \geq q^*.$$

Since q^* exceeds ξ_1 on $\{d, \dots, \ell-d-1\}$, this implies the claim. ◀

While Lemma 10(i) is sufficient to show that all but a δ-fraction of the vertices is peeled whp if $c < \mathrm{er}_k$, we still need the following combinatorial argument that shows that whp no non-empty core is contained within the remaining vertices. Arguments such as these are standard, many similar ones can be found for instance in [18, 19, 23, 27, 29, 35, 32].

▶ **Lemma 11.** *For any $k \geq 3$, $\ell \in \mathbb{N}$ and $c \in (0,1)$ there exists $\delta = \delta(k,\ell) > 0$ such that the following holds whp. For any non-empty set $V' \subseteq V$ of at most $\delta|V|$ vertices of $F = (V,E)$, there exists $v \in V'$ of degree at most 1 in the sub-hypergraph of H induced by V'.*

Proof. In the course of the proof we will implicitly encounter positive upper bounds on δ in terms of k and ℓ. Any $\delta > 0$ small enough to respect these bounds is suitable. We consider the events $(W_{s,t})_{k \leq s \leq \delta|V|, \frac{2s}{k} \leq t \leq |E|}$ that some small set V' of size s induces t edges. If *none* of these events occurs, then all such V' induce less than $2|V'|/k$ edges and therefore induce hypergraphs with average degree less than 2, so a vertex of degree at most 1 exists in each of them.

It is thus sufficient to show that $\Pr[\bigcup_s \bigcup_t W_{s,t}] = \mathcal{O}(1/n)$. We shall use a first moment argument. First note that F has duplicate edges with probability $\binom{cn\ell}{2}(\ell n^k)^{-1} = \mathcal{O}(n^{-1})$, so we restrict our attention to F without duplicate edges. Given s and t there are $\binom{(\ell+k-1)n}{s}$ ways to choose V' and at most $\binom{s^k}{t}$ ways to choose which k-tuples of vertices in V' induce an edge. The probability that any given k-tuple actually does induce an edge is either zero if the k vertices are not of consecutive segments or $1 - (1-(\ell n^k)^{-1})^{cn\ell} \leq \frac{cn}{n^k} = \frac{1}{n^{k-1}}$. Thus, using constants $C, C', C'', C''' \in \mathbb{R}^+$ (that may depend on k and ℓ) where precise values do not matter, we get

$$\Pr[\bigcup_{s=k}^{\delta|V|}\bigcup_{t=\frac{2s}{k}}^{|E|} W_{s,t}] \le \sum_{s=k}^{\delta|V|}\sum_{t=\frac{2s}{k}}^{|E|} \Pr[W_{s,t}] \le \sum_{s=k}^{\delta|V|}\sum_{t=\frac{2s}{k}}^{|E|} \binom{(\ell+k-1)n}{s}\binom{s^k}{t}\left(\frac{1}{n^{k-1}}\right)^t$$
$$\le \sum_{s=k}^{\delta|V|}\sum_{t=\frac{2s}{k}}^{|E|} \left(\frac{e(\ell+k-1)n}{s}\right)^s \left(\frac{es^k}{tn^{k-1}}\right)^t \le \sum_{s=k}^{\delta|V|}\sum_{t=\frac{2s}{k}}^{|E|} \left(C\frac{n}{s}\right)^s \left(C'\frac{s^{k-1}}{n^{k-1}}\right)^t$$
$$\le 2\sum_{s=k}^{\delta|V|} \left(C\frac{n}{s}\right)^s \left(C'\frac{s^{k-1}}{n^{k-1}}\right)^{\frac{2s}{k}} = 2\sum_{s=k}^{\delta|V|} \left(C''\frac{n^k}{s^k}\frac{s^{2k-2}}{n^{2k-2}}\right)^{\frac{s}{k}} = 2\sum_{s=k}^{\delta|V|} \left(C'''\frac{s}{n}\right)^{\frac{s(k-2)}{k}}.$$

To get rid of the summation over t, we assumed $(s/n)^{k-1} \le \delta^{k-1} \le \frac{1}{2C'}$. Elementary arguments show that in the resulting bound, the contribution of summands for $s \in \{k, \dots, 2k\}$ is of order $\mathcal{O}(\frac{1}{n})$, the contribution of the summands with $s \in \{2k+1, \dots, O(\log n)\}$ are of order $\mathcal{O}(\frac{\log n}{n^2})$ (using $\frac{s}{n} \le \frac{\log n}{n}$) and the contribution of the remaining terms with $s \ge 3\log_2 n$ is of order $\mathcal{O}(2^{-\log_2 n}) = \mathcal{O}(\frac{1}{n})$ (using $C'''\frac{s}{n} \le C'''\delta(\ell+2) \le \frac{1}{2}$).

This gives $\Pr[\bigcup_{s,t} W_{s,t}] = \mathcal{O}(n^{-1})$, proving the claim. ◀

We are ready to prove the "$\mathrm{er}_k \le f_k$" of Theorem 1, stated here as a theorem of its own.

▶ **Theorem 12.** *For all $k \ge 3$ we have $\mathrm{er}_k \le f_k$.*

Proof. We need to prove that for any $c < \mathrm{er}_k$ and any $\ell \in \mathbb{N}$ the fuse graph $F = F(n,k,c,\ell)$ is peelable whp.

First, let $\delta = \delta(k,\ell)$ be the constant from Lemma 11 and $R = R(\delta/2,\ell)$ as well as $N = N(\delta/2,\ell)$ the corresponding constants from Lemma 10(i).

Assuming $n \ge N$ we have $q^{(R)}(i) \le \delta/2$ for all $i \in I$, meaning any vertex v from F is *not* deleted within R rounds of $\mathsf{peel}_v(F)$ with probability at most $\delta/2$. Since $\mathsf{peel}(F)$ deletes at least the vertices that any $\mathsf{peel}_v(F)$ for $v \in V$ deletes, the expected number of vertices not deleted by $\mathsf{peel}(F)$ within R rounds is at most $\delta|V|/2$.

Now standard arguments using Azuma's inequality (see e.g. [33, Theorem 13.7]) suffice to conclude that whp at most $\delta|V|$ vertices are not deleted by $\mathsf{peel}(F)$ within R rounds.

By Lemma 11 whp neither the remaining $\delta|V|$ vertices, nor any of its subsets induces a hypergraph of minimum degree 2. Therefore the core of F is empty. ◀

A natural follow-up question to Theorem 12 would be whether $\mathrm{er}_k = f_k$, which would also imply $f_k \le \mathrm{co}_k$. To establish this stronger claim, we would have to exclude the possibility that for certain densities c there is a function $r(n) = \omega(1)$ such that a constant fraction of vertices survive $r(n)$ rounds but are nevertheless deleted eventually. It seems plausible that arguments similar to [35, Lemma 4] can be used, but since our main goal is reached we do not pursue this now.

5 Approximating the Erosion and Consolidation Thresholds

We now approximate the thresholds er_k (and analogously co_k) with numerical methods. Note that if $c < \mathrm{er}_k$ (if $c > \mathrm{co}_k$), then this can be verified in a finite computation, because the correct value of R, together with a bound on the required precision of floating point operations (when rounding conservatively), constitutes a witness. Moreover, the function $\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2}$ can be represented by a finite number of reals, since it is constant on $(-\infty, -(k-1)r]$ and constant on $[(k-1)r, \infty)$.

To approximate er_k (and co_k) with high precision, more efficient approaches are required, however. We compute upper bounds on $\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2}$ by focusing on a finite domain $[-D, D]$ for some $D \in \mathbb{N}$ and rounding conservatively outside of it. Concretely we define $(a_r : \mathbb{Z} \to [0,1])_{r\in\mathbb{N}_0}$ (dependent on k, c and D) with $a_0 := \mathrm{step}^1_{\xi_1/2}$ (analogously $(b_r : \mathbb{Z} \to [0,1])_{r\in\mathbb{N}_0}$ with $b_0 := \mathrm{step}_0^{(\xi_1+\xi_2)/2}$). For $r \geq 0$ we let

$$a_{r+1}(i) := \begin{cases} a_{r+1}(-D) & \text{if } i < -D, \\ \hat{\mathbf{P}} a_r(i) & \text{if } -D \leq i \leq D, \\ 1 & \text{if } i > D. \end{cases} \qquad b_{r+1}(i) := \begin{cases} 0 & \text{if } i < -D, \\ \hat{\mathbf{P}} b_r(i) & \text{if } -D \leq i \leq D, \\ \hat{\mathbf{P}} b_r(D) & \text{if } i > D. \end{cases}$$

Due to the limited effective domain, each a_r is given by $2D+2$ values. It is easy to see that each a_r is monotonous and fulfils $a_{r+1} \leq \hat{\mathbf{P}} a_r$, which implies $\hat{\mathbf{P}}^r \mathrm{step}^1_{\xi_1/2} \leq a_r$. If we find $a_r(0) < \xi_1/2$, then by monotonicity we have $a_r \leq \gg \mathrm{step}^1_{\xi_1/2}$ and therefore:

$$\exists R \in \mathbb{N} : a_R(0) < \xi_1/2 \quad \Rightarrow \quad \exists R \in \mathbb{N} : \hat{\mathbf{P}}^R \mathrm{step}^1_{\xi_1/2} < \gg \mathrm{step}^1_{\xi_1/2} \quad \overset{\text{def}}{\Rightarrow} \quad c < \mathrm{er}_k.$$

(Analogously if $b_R(-1) > (\xi_1+\xi_2)/2$ then $c > \mathrm{co}_k$ follows.)

Experimental Results. For $D = 50$ and all $k \in \{3, \ldots, 7\}$ we computed, using double-precision floating point values, $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ for various c. For each pair (k, c), we either find that $\hat{\mathbf{P}}(k, c)$ is consolidating, it is eroding, or none of the two can be verified. The results suggest that $\mathrm{er}_k < c^*_k < \mathrm{co}_k$ where c^*_k is the orientability threshold for k-ary Erdős-Renyi hypergraphs.

Concretely, we considered for $j = 1, 2, 3, \ldots$ the values $c^*_k - 2^{-j}$ and tried to verify that they are less than er_k. The largest for which we succeeded is reported as b_k in Table 1 on page 3. The largest number of iterations required was $6 \cdot 10^7$. For the first value that could not be shown to be less than er_k, our approximations of the sequence of $(a_i)_{i\in\mathbb{N}}$ became stationary with $a[0] > \xi_1/2$, i.e. the double-precision floats did not change any more (the highest number of iterations to reach this point was $2 \cdot 10^8$). It is possible that the value is still less than er_k and our choice of D or the precision of our floats is simply insufficient. Further experiments with 128-bit floats and larger values of D suggest however, that there is a tiny but real gap between er_k, c^*_k and co_k and the natural conjecture of equality is misplaced.

In the same way we report the smallest value of the form $c^*_k + 2^{-j}$ for which we verified that it exceeds co_k as B_k in Table 1.

6 Peeling Necessitates Orientability of Erdős-Renyi Hypergraphs

We now prove the "$f_k \leq c^*_k$"-half of Theorem 1, stated as Theorem 14. Recall that an *orientation* of a hypergraph $H = (V, E)$ is an injective map $f : E \to V$ with $f(e) \in e$ for all $e \in E$ and that c^*_k is the threshold for orientability of k-uniform Erdős-Renyi hypergraphs.

After classical (2-ary) cuckoo hashing was discovered [36] (relying on $c^*_2 = \frac{1}{2}$), the thresholds for $k > 2$ were determined independently by [11, 19, 20], with generalisations to other graphs and hypergraphs studied in [9, 17, 25, 26, 40].

Note that if H is peelable then it is also orientable: Just orient each edge e to a vertex $v \in e$ such that v and e are deleted in the same round of $\mathsf{peel}(H)$.

Our proof of Theorem 14 relies strongly on a deep and remarkable theorem due to Lelarge [27]. To clarify its role in our proof, we restate it in weaker but sufficient form.

▶ **Theorem 13** (Lelarge [27, Theorem 4.1]). *Let $(G_n = (A_n, B_n, E_n))_{n\in\mathbb{N}}$ be a sequence of bipartite graphs with $|E_n| = O(|A_n|)$. Let further $M(G_n)$ be the size of a maximum matching in G_n. If the* random weak limit *ρ of $(G_n)_{n\in\mathbb{N}}$ is a* bipartite unimodular Galton-Watson tree*, then $\lim_{n\to\infty} \frac{M(G_n)}{|A_n|}$ exists almost surely and depends only on ρ.*

To see the connection, note that an orientation of a hypergraph is a left-perfect matching in its (bipartite) incidence graph.

▶ **Theorem 14.** *For all $k \geq 3$ we have $f_k \leq c_k^*$.*

Proof. Let $c = c_k^* + \varepsilon$. We need to show that there exists $\ell \in \mathbb{N}$ such that the fuse graph $F = F(n, k, c, \ell)$ is not peelable whp.

Let $H = H_{n,cn}^k$ be the k-ary Erdős-Renyi random hypergraph with density c. By choice of c, H is not orientable whp. More strongly even, there exists $\delta = \delta(\varepsilon) > 0$ such that the largest *partial orientation*, i.e. the largest subset of the edges that can be oriented, has size $(1-\delta)cn + o(n)$ whp, see for instance [27].

We set $\ell = \frac{k}{\delta c}$ and consider F as well as the hypergraph $\widetilde{F}$ where the vertices i and $i + n\ell$ for all $i \in \{1, \ldots, (k-1)n\}$ are merged. This "glues" the last $k-1$ segments of F on top of the first $k-1$ segments of F, making $\widetilde{F}$ a "seamless" version of our construction. Crucially, the *random weak limit* of $\widetilde{F}$ and H coincide, i.e. for any constant $R \in \mathbb{N}$ the distribution of the R-neighbourhood $N_{\widetilde{F}}(v, R)$ of a random vertex v of $\widetilde{F}$ has the same limit (as $n \to \infty$) as the distribution of the R-neighbourhood $N_H(v, R)$ of a random vertex v of H.[5] It now follows from [27, Theorem 4.1] that the size of the largest partial orientation of $\widetilde{F}$ is essentially also a $(1-\delta)$-fraction of the number of edges, namely $(1-\delta)c\ell n + o(n)$ whp. Switching from $\widetilde{F}$ back to F can increase the size of a largest partial orientation by at most $(k-1)n$ to $(1 - \delta + \frac{k-1}{c\ell})c\ell n + o(n) = (1 - \frac{\delta}{k})c\ell n + o(n)$ whp. Thus F is not orientable whp and therefore not peelable whp. ◀

7 Experiments

We used our hypergraphs to implement retrieval data structures and compare it to existing implementations.

A 1-*bit retrieval data structure* for a universe $\mathcal{U}$ is a pair of algorithms construct and query, where the input of construct is a set $S \subseteq \mathcal{U}$ of size $m = |S|$ and $f : S \to \{0, 1\}$. If construct succeeds, then the output is a data structure D_f such that query$(D_f, x) = f(x)$ for all $x \in S$. The output of query(D_f, y) for $y \in \mathcal{U} \setminus S$ may yield an arbitrary element of $\{0, 1\}$. The interesting setting is when the data structure may only occupy $\mathcal{O}(m)$ bits. See [8, 7, 12, 21, 38].

One approach is to map each element $x \in S$ to a set $e_x \subset [N]$ via a hash function, where $N = m/c$ for some desired edge density c. One then seeks a solution $z : [N] \to \{0, 1\}$ satisfying $\bigoplus_{v\in e_x} z(v) = f(x)$ for all $x \in S$. The bit-vector z and the hash function then form D_f. A query simply evaluates the left hand side of the equation for x to recover $f(x)$. To compute z, we consider the hypergraph $H = ([N], \{e_x, x \in S\})$. A peelable vertex $v \in [N]$ only contained in one edge e_x corresponds to a variable $z(v)$ only occuring in the equation associated with x. It is thus easy to see that if H is peelable, repeated elimination and back-substitution yields z in $\mathcal{O}(m)$ time.

[5] The common limit of the incidence graphs of $\widetilde{F}$ and H is the bipartite unimodular Galton-Watson tree described in [27, Section 4]. Standard arguments, e.g. from [24, 25] suffice to establish the identity.

Table 2 Overheads and average running times per key of various practical retrieval data structures.

	Configuration	Overhead	construct [µs/key]	query [ns]
Botelho et al. [8]	c = 0.81	23.5%	0.32	59
⟨Fuse Graphs⟩	$c = 0.910, k = 3, \ell = 100$	12.1%	0.29	55
⟨Fuse Graphs⟩	$c = 0.960, k = 4, \ell = 200$	5.7%	0.29	60
⟨Fuse Graphs⟩	$c = 0.985, k = 7, \ell = 500$	2.7%	0.38	74
Luby et al. [28]	$c = 0.9, D = 12$	11.1%	0.79	94
Luby et al. [28]	$c = 0.99, D = 150$	1.1%	0.87	109
Genuzio et al. [21]	$c = 0.91, k = 3, C = 10^4$	10.2%	1.30	58
Genuzio et al. [21]	$c = 0.97, k = 4, C = 10^4$	3.4%	2.20	64
the authors [13]	$c = 0.9995, \ell = 16, C = 10^4$	0.25%	2.47	56

We implemented the following peeling-based variations and report results in[6] Table 2. By the *overhead* of an implementation we mean $\frac{N'}{m} - 1$ where $N' \geq N$ is the total number of bits used, including auxiliary data structures.

Botelho et al. [8] H is a 3-ary Erdős-Renyi hypergraph with an edge density below the peelability threshold $c_3 \approx 0.818$. Construction via peeling and queries are very fast, but the overhead of 23% is sizeable (i.e. D_f occupies roughly $1.23m$ bits).

Fuse Graphs. The edges are distributed such that H is a fuse graph. Recall that the edge density is $c\frac{\ell}{\ell+k-1}$. Note that we let ℓ grow with k to keep the density close to c. We still keep ℓ in a moderate range, as our construction relies on $n \gg \ell$.

Luby et al. [28] The edges are distributed such that H is the peelable hypergraph from [28] already mentioned on page 2. To our knowledge these hypergraphs have not been considered in the context of retrieval. They seem to be particularly well suited to achieve very small overheads at the cost of larger construction and mean query times compared to our other approaches. Note that the largest edge size is $D + 4$ and the worst-case query time is therefore much larger than the reported average query time.

For reference, we also implemented two recent retrieval data structures that do not rely on peeling but solve linear systems [13, 21]. There, to counteract cubic solving time, the input is partitioned into chunks of size C. Especially [13] achieves much smaller overheads than what is feasible with peeling approaches, with the downside of being much slower and more complicated.

Overall, it seems using fuse graphs in retrieval data structures has a chance of outperforming existing approaches when moderate memory overheads of $\approx 5\%$ are acceptable.

However, more research is required to explore the complex space of possible input sizes, configurations of the data structures and trade-offs between overhead and runtime. Our implementations are configured reasonably, but arbitrary in some aspects. A full discussion is beyond the scope of this paper.

[6] Experiments were performed on an desktop computer with an Intel® Core i7-2600 Processor @ 3.40GHz. In all cases, the data set S contains the first $m = 10^7$ URLs from the `eu-2015-host` dataset gathered by [5] with ≈80 bytes per key, and $f : \mathcal{U} \to \{0, 1\}$ is taken to be the parity of the string length. As hash function we used MurmurHash3_x64_128 [2]. If more than 128 hash bits were needed, techniques resembling double-hashing were used to generate additional bits to avoid another execution of murmur. Reported query times are averages obtained by querying all elements of the data set once. They include the roughly 25 ns needed to evaluate murmur on average. The reported numbers are medians of 5 executions.

8 Conclusion

We introduced for all $k \in \mathbb{N}$ a new family of k-uniform hypergraphs where the vertex set is partitioned into a large but constant number of segments. Each edge chooses a random range of k consecutive segments and one random incidence in each of them.

While we have no asymptotic results on the resulting peelability thresholds f_k, at least for small k they are remarkably close to c_k^* with $0 \leq c_k^* - f_k \leq 10^{-5}$ for $k \in \{3, 4, 5, 6, 7\}$. In other words, f_k almost coincides with the *orientability* threshold c_k^* of Erdős-Renyi hypergraphs and significantly exceeds their peelability threshold c_k. Note that $c_k^* = 1 - (1 + o_k(1))e^{-k} \overset{k\to\infty}{\longrightarrow} 1$ (see [19, page 3]) while $c_k \overset{k\to\infty}{\longrightarrow} 0$ (see e.g. [35]). When plugging our hypergraphs into the retrieval framework by [8], we obtained corresponding improvements with respect to memory usage, with no discernible downsides.

Future Experiments. While our experiments on retrieval data structures are promising, it is unclear how robustly the advantages translate to other practical settings where peelable hypergraphs are used, say when implementing Invertible Bloom Lookup Tables [22]. There are hidden disadvantages of our hypergraphs not considered in this paper – for instance the number of rounds needed to peel our hypergraphs is higher, possibly hurting parallel peeling algorithms – as well as hidden advantages – peeling in external memory, a setting considered in [3], is easy due to the locality of the edges.

A Theoretical Question. Given our results, it is natural to suspect a fundamental connection between f_k and c_k^*. Quite possibly, the tiny gap that seems to remain between the values – clearly negligible from a practical perspective – is merely an artefact of the discreteness of segments in our construction.

This discreteness, while heavily used in our arguments, may in fact be dispensable. Indeed, we believe the key idea behind our hypergraphs is *limited bandwidth* where a hypergraph on vertex set $[n]$ has bandwidth at most d if each edge e satisfies $\max_{v\in e} v - \min_{v\in e} v < d$ (the incidence matrix can then be sorted to resemble a bandmatrix). Such a hypergraph can be generated by choosing for each edge a random range of d consecutive vertices and k incidences independently and uniformly at random from that range. In experiments with $k = 3$ and $d = \varepsilon n$, such hypergraphs performed similar to the hypergraphs we analysed (with $k = 3$ and $\ell \approx 1/\varepsilon$). Note that there are no discrete segments in the modified construction. It would be nice to see whether in such a variation peelability and orientability are more elegantly and more intimately linked.

References

1 David Aldous and J. Michael Steele. The objective method: Probabilistic combinatorial optimization and local weak convergence. In *Probability on Discrete Structures. Encyclopaedia of Mathematical Sciences (Probability Theory)*, volume 110, pages 1–72. Springer, Berlin, Heidelberg, 2004. doi:10.1007/978-3-662-09444-0_1.

2 Austin Appleby. MurmurHash3, 2012. URL: https://github.com/aappleby/smhasher/blob/master/src/MurmurHash3.cpp.

3 Djamal Belazzougui, Paolo Boldi, Giuseppe Ottaviano, Rossano Venturini, and Sebastiano Vigna. Cache-oblivious peeling of random hypergraphs. In *Data Compression Conference*, pages 352–361, 2014. doi:10.1109/DCC.2014.48.

4 Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 1970. doi:10.1145/362686.362692.

5 Paolo Boldi, Andrea Marino, Massimo Santini, and Sebastiano Vigna. BUbiNG: Massive crawling for the masses. In *Proc. 23rd WWW'14*, pages 227–228, 2014. doi:10.1145/2567948.2577304.

6 Fabiano Cupertino Botelho. *Near-Optimal Space Perfect Hashing Algorithms*. PhD thesis, Federal University of Minas Gerais, 2008. URL: http://cmph.sourceforge.net/papers/thesis.pdf.

7 Fabiano Cupertino Botelho, Rasmus Pagh, and Nivio Ziviani. Simple and space-efficient minimal perfect hash functions. In *Proc. 10th WADS*, pages 139–150, 2007. doi:10.1007/978-3-540-73951-7_13.

8 Fabiano Cupertino Botelho, Rasmus Pagh, and Nivio Ziviani. Practical perfect hashing in nearly optimal space. *Inf. Syst.*, pages 108–131, 2013. doi:10.1016/j.is.2012.06.002.

9 Julie Anne Cain, Peter Sanders, and Nicholas C. Wormald. The random graph threshold for k-orientiability and a fast algorithm for optimal multiple-choice allocation. In *Proc. 18th SODA*, pages 469–476, 2007. URL: http://dl.acm.org/citation.cfm?id=1283383.1283433.

10 Denis Xavier Charles and Kumar Chellapilla. Bloomier filters: A second look. In *Proc. 16th ESA*, 2008. doi:10.1007/978-3-540-87744-8_22.

11 Martin Dietzfelbinger, Andreas Goerdt, Michael Mitzenmacher, Andrea Montanari, Rasmus Pagh, and Michael Rink. Tight thresholds for cuckoo hashing via XORSAT. In *Proc. 37th ICALP (1)*, pages 213–225, 2010. doi:10.1007/978-3-642-14165-2_19.

12 Martin Dietzfelbinger and Rasmus Pagh. Succinct data structures for retrieval and approximate membership (extended abstract). In *Proc. 35th ICALP (1)*, pages 385–396, 2008. doi:10.1007/978-3-540-70575-8_32.

13 Martin Dietzfelbinger and Stefan Walzer. Constant-time retrieval with $O(\log m)$ extra bits. In *Proc. 36th STACS*, pages 24:1–24:16, 2019. doi:10.4230/LIPIcs.STACS.2019.24.

14 Martin Dietzfelbinger and Christoph Weidling. Balanced allocation and dictionaries with tightly packed constant size bins. *Theor. Comput. Sci.*, 380(1-2):47–68, 2007. doi:10.1016/j.tcs.2007.02.054.

15 Olivier Dubois and Jacques Mandler. The 3-XORSAT threshold. In *Proc. 43rd FOCS*, pages 769–778, 2002. doi:10.1109/SFCS.2002.1182002.

16 David Eppstein and Michael T. Goodrich. Straggler identification in round-trip data streams via Newton's identities and invertible Bloom filters. *IEEE Trans. on Knowl. and Data Eng.*, 23(2):297–306, 2011. doi:10.1109/TKDE.2010.132.

17 Daniel Fernholz and Vijaya Ramachandran. The k-orientability thresholds for $g_{n,p}$. In *Proc. 18th SODA*, pages 459–468, 2007. URL: http://dl.acm.org/citation.cfm?id=1283383.1283432.

18 Nikolaos Fountoulakis, Megha Khosla, and Konstantinos Panagiotou. The multiple-orientability thresholds for random hypergraphs. *Combinatorics, Probability & Computing*, 25(6):870–908, 2016. doi:10.1017/S0963548315000334.

19 Nikolaos Fountoulakis and Konstantinos Panagiotou. Sharp load thresholds for cuckoo hashing. *Random Struct. Algorithms*, 41(3):306–333, 2012. doi:10.1002/rsa.20426.

20 Alan M. Frieze and Páll Melsted. Maximum matchings in random bipartite graphs and the space utilization of cuckoo hash tables. *Random Struct. Algorithms*, 41(3):334–364, 2012. doi:10.1002/rsa.20427.

21 Marco Genuzio, Giuseppe Ottaviano, and Sebastiano Vigna. Fast scalable construction of (minimal perfect hash) functions. In *Proc. 15th SEA*, pages 339–352, 2016. doi:10.1007/978-3-319-38851-9_23.

22 Michael T. Goodrich and Michael Mitzenmacher. Invertible Bloom lookup tables. In *Proc. 49th Annual Allerton Conference on Communication, Control, and Computing*, pages 792–799, 2011. doi:10.1109/Allerton.2011.6120248.

23 Svante Janson and Malwina J. Luczak. A simple solution to the k-core problem. *Random Struct. Algorithms*, 30(1-2):50–62, 2007. doi:10.1002/rsa.20147.

24 Jeong Han Kim. Poisson cloning model for random graphs. In *Proc. ICM Madrid 2006 Vol. III*, pages 873–898, 2006. URL: https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM2006.3/ICM2006.3.ocr.pdf.

25 Mathieu Leconte. Double hashing thresholds via local weak convergence. In *51st Annual Allerton Conference on Communication, Control, and Computing*, pages 131–137, 2013. doi:10.1109/Allerton.2013.6736515.

26 Eric Lehman and Rina Panigrahy. 3.5-way cuckoo hashing for the price of 2-and-a-bit. In *Proc. 17th ESA*, pages 671–681, 2009. doi:10.1007/978-3-642-04128-0_60.

27 Marc Lelarge. A new approach to the orientation of random hypergraphs. In *Proc. 23rd SODA*, pages 251–264, 2012. doi:10.1137/1.9781611973099.23.

28 Michael Luby, Michael Mitzenmacher, Mohammad Amin Shokrollahi, and Daniel A. Spielman. Efficient erasure correcting codes. *IEEE Transactions on Information Theory*, 47(2):569–584, 2001. doi:10.1109/18.910575.

29 Tomasz Luczak. Size and connectivity of the k-core of a random graph. *Discrete Mathematics*, 91(1):61–68, 1991. doi:10.1016/0012-365X(91)90162-U.

30 Bohdan S. Majewski, Nicholas C. Wormald, George Havas, and Zbigniew J. Czech. A family of perfect hashing methods. *Comput. J.*, pages 547–554, 1996. doi:10.1093/comjnl/39.6.547.

31 Michael Mitzenmacher. Some open questions related to cuckoo hashing. In *Proc. 17th ESA*, 2009. doi:10.1007/978-3-642-04128-0_1.

32 Michael Mitzenmacher, Konstantinos Panagiotou, and Stefan Walzer. Load thresholds for cuckoo hashing with double hashing. In *16th SWAT*, pages 29:1–29:9, 2018. doi:10.4230/LIPIcs.SWAT.2018.29.

33 Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2017.

34 Michael Mitzenmacher and George Varghese. Biff (Bloom filter) codes: Fast error correction for large data sets. In *Proc. ISIT 2012*, pages 483–487, 2012. doi:10.1109/ISIT.2012.6284236.

35 Michael Molloy. Cores in random hypergraphs and boolean formulas. *Random Struct. Algorithms*, 27(1):124–135, 2005. doi:10.1002/rsa.20061.

36 Rasmus Pagh and Flemming Friche Rodler. Cuckoo hashing. *J. Algorithms*, 51(2):122–144, 2004. doi:10.1016/j.jalgor.2003.12.002.

37 Boris Pittel and Gregory B. Sorkin. The satisfiability threshold for k-XORSAT. *Combinatorics, Probability & Computing*, 25(2):236–268, 2016. doi:10.1017/S0963548315000097.

38 Ely Porat. An optimal Bloom filter replacement based on matrix solving. In *Proc. 4th CSR*, pages 263–273, 2009. doi:10.1007/978-3-642-03351-3_25.

39 Michael Rink. Mixed hypergraphs for linear-time construction of denser hashing-based data structures. In *Proc. 39th SOFSEM*, pages 356–368, 2013. doi:10.1007/978-3-642-35843-2_31.

40 Stefan Walzer. Load thresholds for cuckoo hashing with overlapping blocks. In *Proc. 45th ICALP*, pages 102:1–102:10, 2018. doi:10.4230/LIPIcs.ICALP.2018.102.