

Phase transition in count approximation by Count-Min sketch with conservative updates

Éric Fusy¹ and Gregory Kucherov¹

¹LIGM, CNRS, Univ. Gustave Eiffel, Marne-la-Vallée, France
`{Eric.Fusy|Gregory.Kucherov}@univ-eiffel.fr`

Abstract

Count-Min sketch is a hash-based data structure to represent a dynamically changing associative array of counters. Here we analyse the counting version of Count-Min under a stronger update rule known as *conservative update*, assuming the uniform distribution of input keys. We show that the accuracy of conservative update strategy undergoes a phase transition, depending on the number of distinct keys in the input as a fraction of the size of the Count-Min array. We prove that below the threshold, the relative error is asymptotically $o(1)$ (as opposed to the regular Count-Min strategy), whereas above the threshold, the relative error is $\Theta(1)$. The threshold corresponds to the peelability threshold of random k -uniform hypergraphs. We demonstrate that even for small number of keys, peelability of the underlying hypergraph is a crucial property to ensure the $o(1)$ error. Finally, we provide an experimental evidence that the phase transition does not extend to non-uniform distributions, in particular to the popular Zipf's distribution.

1 Introduction

Count-Min sketch is a hash-based data structure to represent a dynamically changing associative array $\mathbf{a}$ of counters in an approximate way. The array $\mathbf{a}$ can be seen as a mapping from some set K of keys to $\mathbb{N}$, where K is drawn from a (large) universe U . The goal is to support *point queries* about the (approximate) current value of $\mathbf{a}(p)$ for a key p . Count-Min is especially suitable for the streaming framework, when counters associated to keys are updated dynamically. That is, *updates* are (key,value) pairs (p, ℓ) with the meaning that $\mathbf{a}(p)$ is updated to $\mathbf{a}(p) + \ell$.

Count-Min sketch was proposed in [11], see e.g. [9] for a survey. A similar data structure was introduced earlier in [8] named *Spectral Bloom filter*, itself closely related to *Counting Bloom filters* [18]. The difference between Count-Min sketch and Spectral Bloom filter is marginal: while a Count-Min sketch requires hash functions to have disjoint codomains (rows of Count-Min matrix), a Spectral Bloom filter has all hash functions mapping to the same array, as does the regular Bloom filter. In this paper, we will deal with the Spectral Bloom filter version but will keep the term Count-Min sketch as more common in the literature.

Count-Min sketch supports negative update values ℓ provided that at each moment, each counter $\mathbf{a}(p)$ remains non-negative (so-called *strict turnstile model* [22]). When updates are positive, the Count-Min update algorithm can be modified to a stronger version leading to smaller errors in

queries. This modification, introduced in [16] as *conservative update*, is mentioned [9], without any formal analysis given in those papers. This variant is also discussed in [8] under the name *minimal increase*, where it is claimed that it decreases the probability of a positive error by a factor of the number of hash functions, but no proof is given. We discuss this claim in the concluding part of this paper.

The case of positive updates is widespread in practice. In particular, a very common instance is *counting* where all update values are 1. This task occurs in different scenarios in network traffic monitoring, as well as other applications related to data stream mining [16]. In bioinformatics, we may want to maintain, on the fly, multiplicities of k -mers (words of length k) occurring in a big dataset [24, 1, 28]. We refer to [14] for more examples of applications.

While it is easily seen that the error in conservative update can only be smaller than in Count-Min, obtaining more precise bounds is a challenging problem. Count-Min guarantees, with high probability, that the additive error can be bounded by $\varepsilon \|\mathbf{a}\|_1$ for any ε , where $\|\mathbf{a}\|_1$ is the $L1$ -norm of $\mathbf{a}$ [11]. In the counting setting, $\|\mathbf{a}\|_1$ is the length of the input stream which can be very large, and therefore this bound provides a weak guarantee in practice, unless the distribution of keys is very skewed and queries are made on frequent keys (*heavy hitters*) [22, 6, 10]. It is therefore an important practical question to analyse the improvement provided by the conservative update strategy compared to the original Count-Min sketch.

Probably the first attempt towards this goal was made in [5], under assumption that all $\binom{n}{k}$ counter combinations are equally likely at each step (n size of the Count-Min array, k number of hash functions) which amounts to assuming uniform distribution on $\binom{n}{k}$ input keys. For the regime when the number of distinct keys in the input considerably exceeds the sketch size, it was experimentally observed in [5] that frequent keys have essentially no error while the (over)estimate for less frequent keys tends to be the same, for given input stream length and sketch size. On the other hand, it was experimentally shown that in that regime, the uniform distribution of keys presents the worst-case scenario producing the largest such estimate and a method was proposed to compute this estimate. Another method for bounding the error proposed in [14] is based on a simulation of spectral Bloom filters by a hierarchy of ordinary Bloom filters. However, the bounds provided are not explicit but are expressed via a recursive relation based on false positive rates of involved Bloom filters. Recent works [4, 3] propose formulas for computing error bounds depending on key probabilities assumed independent but not necessarily uniform, in particular leading to an improved precision bounds for detecting heavy hitters.

In this paper, we provide a probabilistic analysis of the conservative update scheme for counting under the assumption of *uniform distribution of keys* in the input. Our main result is a demonstration that the error in count estimates undergoes a phase transition when the number of distinct keys grows relative to the size of the Count-Min array. We show that the phase transition threshold corresponds to the *peelability threshold* for random k -uniform hypergraphs. For the *subcritical regime*, when the number of distinct keys is below the threshold, we show that the relative error for a randomly chosen key tends to 0 asymptotically, with high probability. This contrasts with the regular Count-Min algorithm producing a relative error shown to be at least 1 with constant probability.

For the *supercritical regime*, we show that the average relative error is lower-bounded by a constant (depending on the number of distinct keys), with high probability. We prove this result for $k = 2$ and conjecture that it holds for arbitrary k as well. We provide computer simulations showing that the expected relative error grows fast after the threshold, with a distribution showing

a peculiar multi-modal shape. In particular, keys with small (or zero) error still occur after the threshold, but their fraction quickly decreases when the number of distinct keys grows.

After defining Count-Min sketch and conservative update strategy in Section 2 and introducing hash hypergraphs in Section 3, we formulate the conservative update algorithm (or regular Count-Min, for that matter) in terms of a hypergraph augmented with counters associated to vertices. In Section 4, we state our main results and illustrate them with a series of computer simulations. All technical proofs are provided in a separate Section 5.

In addition, in Section 6, we study a specific family of 2-regular k -hypergraphs that are sparse but not peelable. For such graphs we show that while the relative error of every key is 1 with the regular Count-Min strategy, it is $1/k + o(1)$ for conservative update. While this result is mainly of theoretical interest, it illustrates that the peelability property is crucial for the error to be asymptotically vanishing. Finally, in Section 7, we turn to non-uniform distributions and provide experimental evidence that for Zipf's distribution, the phase transition in average error does not occur.

2 Count-Min and Conservative Update

We consider a (counting version of) Count-Min sketch to be an array A of size n of counters initially set to 0, together with hash functions $h_1, \dots, h_k$ mapping keys from a given universe to $[1..n]$. To count key occurrences in a stream of keys, regular Count-Min proceeds as follows. To process a key p , each of the counters $A[h_i(p)]$, $1 \leq i \leq k$, is incremented by 1. Querying the occurrence number $\mathbf{a}(p)$ of a key p returns the estimate $\hat{\mathbf{a}}_{CM}(p) = \min_{1 \leq i \leq k} \{A[h_i(p)]\}$. It is easily seen that $\hat{\mathbf{a}}_{CM}(p) \geq \mathbf{a}(p)$. A bound on the overestimate of $\mathbf{a}(p)$ is given by the following result adapted from [11].

Theorem 1 ([11]) *For $\varepsilon > 0$, $\delta > 0$, consider a Count-Min sketch with $k = \lceil \ln(\frac{1}{\delta}) \rceil$ and size $n = k \frac{\varepsilon}{\delta}$. Then $\hat{\mathbf{a}}_{CM}(p) - \mathbf{a}(p) \leq \varepsilon N$ with probability at least $1 - \delta$, where N is the size of the input stream.*

While Theorem 1 is useful in some situations, it has a limited utility as it bounds the error with respect to the stream size which can be very large.

Conservative update strengthens Count-Min by increasing only the smallest counters among $A[h_i(p)]$. Formally, for $1 \leq i \leq k$, $A[h_i(p)]$ is incremented by 1 only if $A[h_i(p)] = \min_{1 \leq j \leq k} \{A[h_j(p)]\}$ and is left unchanged otherwise. The estimate of $\mathbf{a}(p)$, denoted $\hat{\mathbf{a}}_{CU}(p)$, is computed as before: $\hat{\mathbf{a}}_{CU}(p) = \min_{1 \leq i \leq k} \{A[h_i(p)]\}$. It can be seen that $\hat{\mathbf{a}}_{CU}(p) \geq \mathbf{a}(p)$ still holds, and that $\hat{\mathbf{a}}_{CU}(p) \leq \hat{\mathbf{a}}_{CM}(p)$. The latter follows from the observation that on the same input, an entry of counter array A under conservative update can never get larger than the same entry under Count-Min.

3 Hash hypergraphs and CU process

With a counter array $A[1..n]$ and hash functions $h_1, \dots, h_k$ we associate a k -uniform *hash hypergraph* $H = (V, E)$ with vertex-set $V = \{1..n\}$ and edge-set $E = \{\{h_1(p), \dots, h_k(p)\}\}$ for all distinct keys p . Let $\mathcal{H}_{n,m}^k$ be the set of k -uniform hypergraphs with n vertices and m edges. We assume that the hash hypergraph is a uniformly random Erdős-Rényi hypergraph in $\mathcal{H}_{n,m}^k$, which we denote

by $H_{n,m}^k$, where m is the number of distinct keys in the input (for $k = 2$, we use the notation $G_{n,m} = H_{n,m}^2$). Even if this property is not granted by hash functions used in practice, it is a reasonable and commonly used hypothesis to conduct the analysis of sketch algorithms.

Below we show that the behavior of a sketching scheme depends on the properties of the associated hash hypergraph. It is well-known that depending on the m/n ratio, many properties of Erdős-Rényi (hyper)graphs follow a phase transition phenomenon [19]. For example, the emergence of a giant component, of size $O(n)$, occurs with high probability (hereafter, *w.h.p.*) at the threshold $\frac{m}{n} = \frac{1}{k(k-1)}$ [21].

Particularly relevant to us is the *peelability* property. Let $H = (V, E)$ be a hypergraph. The peeling process on H is as follows. We define $H_0 = H$, and iteratively for $i \geq 0$, we define V_i to be the set of leaves (vertices of degree 1) or isolated vertices in H_i , E_i to be the set of edges of H_i incident to vertices in V_i , and H_{i+1} to be the hypergraph obtained from H_i by deleting the vertices of V_i and the edges of E_i . A vertex in V_i is said to have *peeling level* i . The process stabilizes from some step I , and the hypergraph H_I is called the *core* of H , which is the largest induced sub-hypergraph whose vertices all have degree at least 2. If H_I is empty, then H is called *peelable*.

It is known [25] that peelability undergoes a phase transition. For $k \geq 3$, there exists a positive constant λ_k such that, for $\lambda < \lambda_k$, the random hypergraph $H_{n,\lambda n}^k$ is *w.h.p.* peelable as $n \rightarrow \infty$, while for $\lambda > \lambda_k$, the core of $H_{n,\lambda n}^k$ has *w.h.p.* a size concentrated around αn for some $\alpha > 0$ that depends on λ . The first peelability thresholds are $\lambda_3 \approx 0.818$, $\lambda_4 \approx 0.772$, etc., λ_3 being the largest.

For $k = 2$, for $\lambda < 1/2$, *w.h.p.* a proportion $1 - o(1)$ of vertices are in trees of size $O(1)$, (and a proportion $o(1)$ of the vertices are in the core), while for $\lambda \geq 1/2$, the core size is *w.h.p.* concentrated around αn for $\alpha > 0$ that depends on λ [27].

We note that properties of hash hypergraphs determine the behavior of some other hash-based data structures, such as Cuckoo hash tables [26] and Cuckoo filters [17], Minimal Perfect Hash Functions and Static Functions [23], Invertible Bloom filters [20], and others. We refer to [29] for an extended study of relationships between properties of hash hypergraphs and some of those data structures. In particular, peelability is directly relevant to certain constructions of Minimal Perfect Hash Functions as well as to good functioning of Invertible Bloom filters.

The connection to hash hypergraphs allows us to reformulate the Count-Min algorithm with conservative updates as a process, which we call *CU-process*, on a random hypergraph $H_{n,m}^k$, where n, m, k correspond to counter array length, number of distinct keys, and number of hash functions, respectively. Let $H = (V, E)$ be a hypergraph. To each vertex v we associate a counter c_v initially set to 0. At each step $t \geq 1$, a *CU-process* on H chooses an edge $e = \{v_1, \dots, v_k\} \in E$ in H , and increments by 1 those c_{v_i} which verify $c_{v_i} = \min_{1 \leq j \leq k} c_{v_j}$. For $t \geq 0$ and $v \in V$, $c_v(t)$ will denote the value of the counter c_v after t steps, and $o_e(t)$ the number of times edge $e \in E$ has been drawn in the first t steps. The counter $c_e(t)$ of an edge $e = \{v_1, \dots, v_k\}$ is defined as $c_e(t) = \min_{1 \leq i \leq k} c_{v_i}(t)$. Clearly, for each t and each e , $o_e(t) \leq c_e(t)$. The *relative error* of e at time t is defined as $R_e(t) = \frac{c_e(t) - o_e(t)}{o_e(t)}$. The following Lemma can be easily proved by induction on t .

Lemma 1 *Let $H = (V, E)$ be a hypergraph on which a CU-process is run. At every step t , for each vertex v , there is at least one edge e incident to v such that $c_e(t) = c_v(t)$.*

Observe that, when H is a graph ($k = 2$), Lemma 1 is equivalent to the property that vertex counters cannot have a strict local maximum, i.e., at every step t , each vertex v has at least one neighbour u such that $c_u(t) \geq c_v(t)$.

4 Phase transition of the relative error

4.1 Main results

Let $H = (V, E)$ be a hypergraph, $|V| = n, |E| = m$. Let $N \geq 1$. We consider two closely related models of input to perform the CU-process. In the *N-uniform model*, the CU process is performed on a random sequence of keys (edges in E) of length $N \cdot m$, each key being drawn independently and uniformly in E . In the *N-balanced model*, the CU-process is performed on a random sequence of length $N \cdot m$, such that each $e \in E$ occurs exactly N times, and the order of keys is random. In other words, the sequence of keys on which the CU-process is performed is a random permutation of the multiset made of N copies of each key of E . Clearly, both models are very close, since the number of occurrences of any key in the *N-uniform model* is concentrated around N (with Gaussian fluctuations of order $\sqrt{N}$) as N gets large. For both models, we use the notation $c_v^{(N)} = c_v(Nm)$ for the resulting counter of $v \in V$, $o_e^{(N)} = o_e(Nm)$ for the resulting number of occurrences of $e \in E$, $c_e^{(N)} = c_e(Nm)$ for the resulting counter of $e \in E$, and $R_e^{(N)} = R_e(Nm) = (c_e^{(N)} - o_e^{(N)})/o_e^{(N)}$ for the resulting relative error of e . In the *N-balanced model*, since each key $e \in E$ occurs N times, we have $R_e^{(N)} = (c_e^{(N)} - N)/N$.

Our main result is the following.

Theorem 2 (subcritical regime) *Let $k \geq 2$, and let $\lambda < \lambda_k$, where $\lambda_2 = 1/2$, and for $k \geq 3$, λ_k is the peelability threshold as defined in Section 3. Consider a CU-process on a random hypergraph $H_{n,\lambda n}^k$ under either *N-uniform* or *N-balanced model*, and consider the relative error $R_e^{(N)}$ of a random edge in $H_{n,\lambda n}^k$. Then $R_e^{(N)} = o(1)$ w.h.p., as both n and N grow¹.*

Note that with the regular Count-Min algorithm (see Section 2), in the *N-balanced model*, the counter value of a node v is $\tilde{c}_v^{(N)} = N \cdot \deg(v)$, and the relative error $\tilde{R}_e^{(N)}$ of an edge $e = (v_1, \dots, v_k)$ is always (whatever $N \geq 1$) equal to $\min(\deg(v_1), \dots, \deg(v_k)) - 1$, and is thus always a non-negative integer. For fixed $k \geq 2$ and $\lambda > 0$, and for a random edge e in $H_{n,\lambda n}^k$, the probability that all k vertices belonging to e have at least one incident edge apart from e converges to a positive constant $c(\lambda, k) = (1 - e^{-k\lambda})^k$. Therefore, $\tilde{R}_e$ is a nonnegative integer whose probability to be non-zero converges to $c(\lambda, k)$. Thus, Theorem 2 ensures that, for $\lambda < \lambda_k$, conservative updates lead to a drastic decrease of the error, from $\Theta(1)$ to $o(1)$.

For a given hypergraph $H = (V, E)$ with m edges, we define $\text{err}_N(H) = \frac{1}{m} \sum_{e \in E} R_e^{(N)}$ the *average* error over the edges of H . Theorem 2 states that a randomly chosen edge has a small error, but this does not formally exclude that a small fraction of edges may have large errors, possibly yielding $\text{err}_N(H)$ larger than $o(1)$. However, we believe that this is not the case. From the previous remark, it follows that the error of an edge $e = (v_1, \dots, v_k)$ is upper-bounded by $\min(\deg(v_1), \dots, \deg(v_k)) - 1$. Since the expected maximal degree in $H_{n,\lambda n}^k$ grows very slowly with n , one can expect that any set of $o(n)$ edges should have a contribution $o(1)$ w.h.p.. This is also supported by experiments given in the next section.

Based on Theorem 2 and the above discussion, we propound that a phase transition occurs for the average error, in the sense that it is $o(1)$ in the subcritical regime $\lambda < \lambda_k$, and $\Theta(1)$ in the supercritical regime $\lambda > \lambda_k$, w.h.p.. Regarding the supercritical regime, we are able to show that this indeed holds for $k = 2$ in the *N-balanced model*.

¹Formally, for any $\epsilon > 0$, there exists M such that $\mathbb{P}(R_e^{(N)} \leq \epsilon) \geq 1 - \epsilon$ if $n \geq M$ and $N \geq M$.

Theorem 3 (supercritical regime, case $k = 2$) *Let $\lambda > 1/2$. Then there exists a positive constant $f(\lambda)$ such that, in the N -balanced model, $\text{err}_N(G_{n,\lambda n}) \geq f(\lambda)$ w.h.p., as n grows².*

Our proof of Theorem 3 can be extended to any $k \geq 2$ and $\lambda > \frac{1}{k(k-1)}$ such that the giant component $G' = (V', E')$ in $H_{n,\lambda n}^k$ satisfies w.h.p. $|E'| - |V'| \geq a n$ for some positive constant a . For $k = 2$, any value $\lambda > 1/2$ has this property [27], while for $k = 3$ (respectively, $k = 4$), the analysis given in [2] ensures that this property holds for values of λ strictly above the peelability threshold, namely $\lambda > \widetilde{\lambda}_3 \approx 0.94$ (respectively, $\lambda > \widetilde{\lambda}_4 \approx 0.98$). Nevertheless, based on simulations presented below, we expect that Theorem 3 holds for $k \geq 3$ for all $\lambda > \lambda_k$ as well, however proving this would then require a different kind of argument.

4.2 Simulations

Here we provide several experimental results illustrating the phase transition stated in Theorems 2 and 3. Figure 1 shows plots for the average relative error $\text{err}_N(H_{n,m}^k)$ as a function of $\lambda = m/n$, for $k \in \{2, 3, 4\}$ for regular Count-Min and the conservative update strategies. Experiments were run for $n = 1000$ with the N -independent model (each edge drawn independently with probability $1/m$) and $N = 50,000$ (number of steps $N \cdot |E|$). For each λ , an average is taken over 15 random graphs.

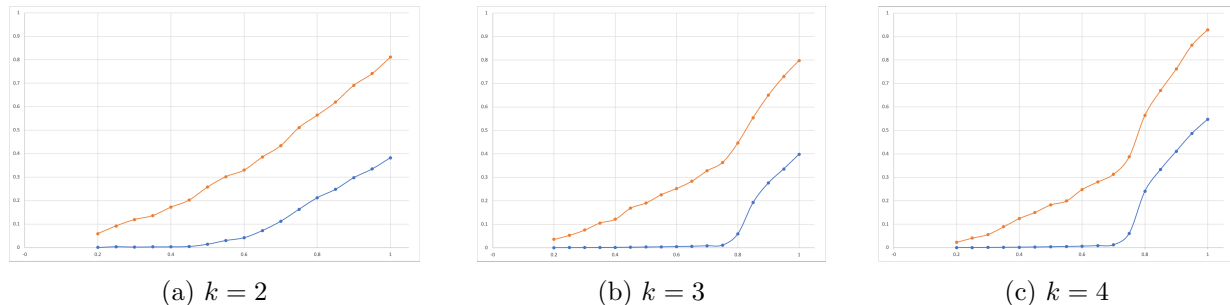


Figure 1: Average relative error as a function of $\lambda = m/n$ for regular Count-Min (red) and conservative update (blue), for $k \in \{2, 3, 4\}$.

The phase transitions are clearly seen to correspond to the critical threshold 0.5 for $k = 2$, and, for $k \in \{3, 4\}$, to the peelability thresholds $\lambda_3 \approx 0.818$, $\lambda_4 \approx 0.772$. Observe that the transition looks sharper for $k \geq 3$, which may be explained by the fact that the core size undergoes a discontinuous phase transition for $k \geq 3$, as shown in [25] (e.g. for $k = 3$, the fraction of vertices in the core jumps from 0 to about 0.13).

For the supercritical regime, we analysed the concentration around the average shown in Figure 1 by simulating CU-processes on 50 random graphs, for each λ . Figure 2 shows the results. We observe that the concentration tends to become tighter for growing λ . Furthermore, we experimentally studied the empirical distribution of individual relative errors, which turns out to have an interesting multi-modal shape for intermediate values of λ . Typical distributions for $k = 2, 3$ are illustrated in

²Formally, for any $\epsilon > 0$, there exists M such that $\mathbb{P}(\text{err}_N(G_{n,m}) \geq f(\lambda)) \geq 1 - \epsilon$ if $N \geq 1$ and $n \geq M$.

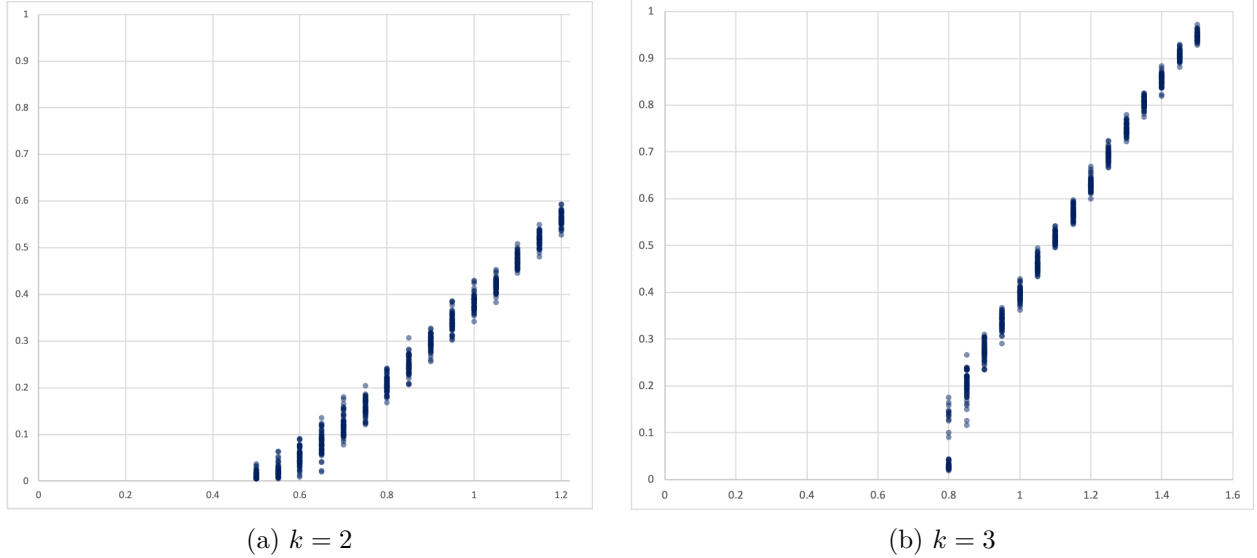
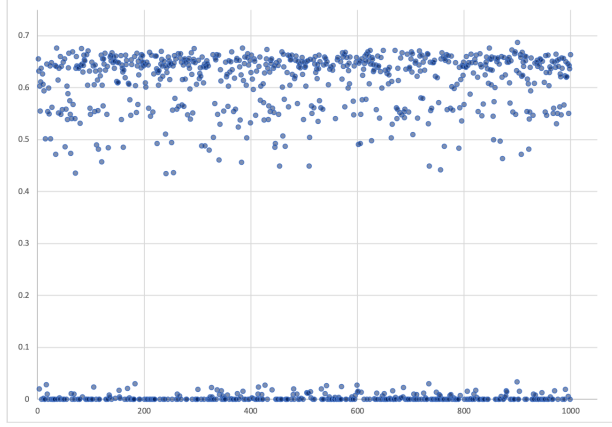


Figure 2: Concentration of the average error as a function of $\lambda = m/n$, for $k \in \{2, 3\}$.

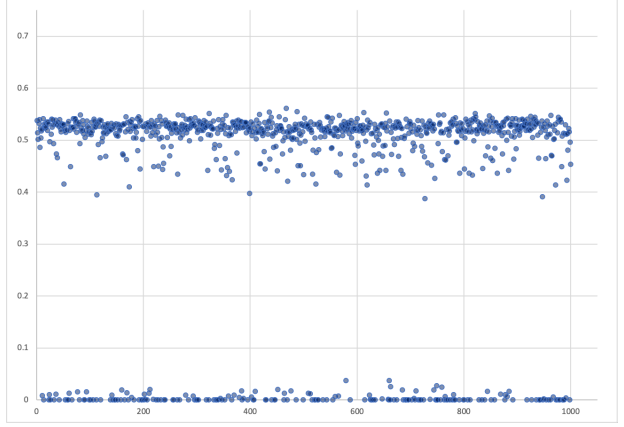
Figure 3 where each point corresponds to an edge, and the edges are randomly ordered along the x -axis. Each plot corresponds to an individual random graph.

When λ grows beyond the peelability threshold, a fraction of edges with small errors still remains but vanishes quickly: these include edges incident to at least one leaf (these have error 0) and peelable edges (these have error $o(1)$, by arguments to be given in Section 5.3). For intermediate values of λ , the distribution presents several modes: besides the main mode (largest concentration on plots of Figure 3), we observe a few other concentration values which are typically integers. While this phenomenon is still to be analysed, we explain it by the presence in the graphs of certain structural motifs that involve disparities in node degrees. Note that the fraction of values concentrated around the main mode is dominant: for example, for $k = 3, \lambda = 3$ (Figure 3d), about 90% of values correspond to the main mode (≈ 3.22). Finally, when λ becomes larger, these “secondary modes” disappear, and the distribution becomes concentrated around a single value. This is consistent with the tighter concentration observed earlier in Figure 2.

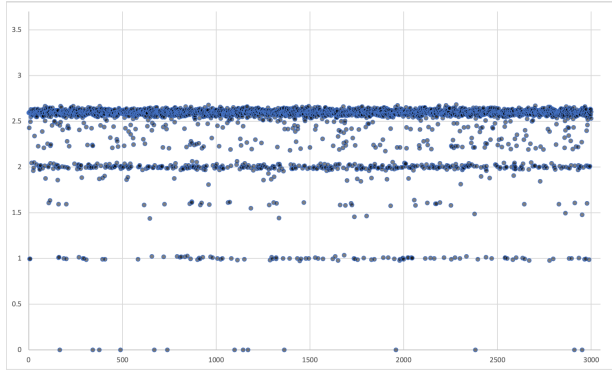
Finally, we report on another experiment supporting the conjecture of a positive average error in the supercritical regime. We simulated the CU-process on sparse random non-peelable 3-hypergraphs (i.e. $k = 3$), namely 2-regular 3-hypergraphs with $2n$ edges and $3n$ vertices (n parameter). These are sparsest possible non-peelable 3-hypergraphs, with degree 2 of each vertex. We observed that the average error for such graphs is concentrated around a constant value of ≈ 0.217 . Since the core size is linear in the supercritical regime, this experiment provides an evidence of a positive error in the general case. While this remains to be proved in general, in Section 6 we provide a proof for certain families of regular hypergraphs.



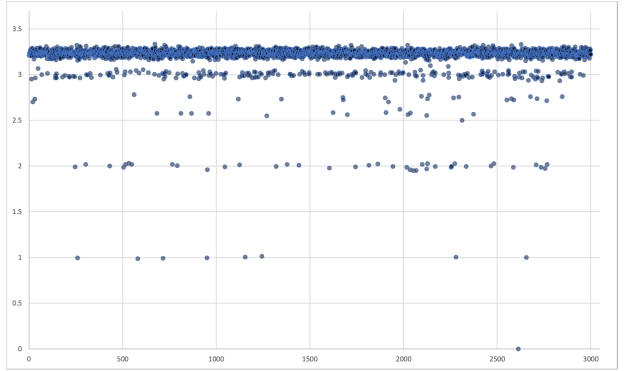
(a) $k = 2, \lambda = 1$



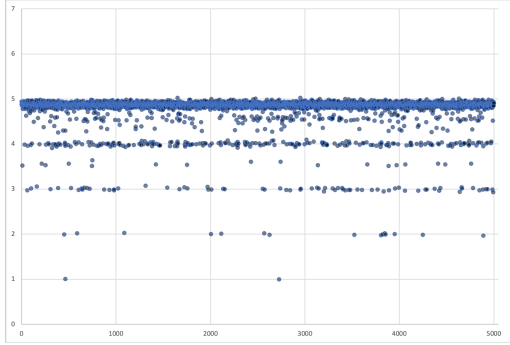
(b) $k = 3, \lambda = 1$



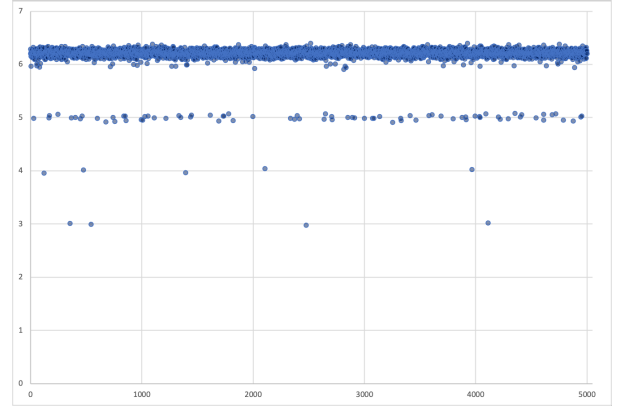
(c) $k = 2, \lambda = 3$



(d) $k = 3, \lambda = 3$



(e) $k = 2, \lambda = 5$



(f) $k = 3, \lambda = 5$

Figure 3: Distribution of relative errors of individual edges shown in a random order along x -axis.

5 Proofs of main results

Theorem 2 relies on properties of random hypergraphs. Case $k = 2$ corresponds to Erdős-Rényi random graphs $G_{n,\lambda n}$ [15] which have been extensively studied [19]. In particular, it is well known when $\lambda < 1/2$ and n gets large, $G_{n,\lambda n}$ is, w.h.p., a union of small connected components most

of which are constant-size trees. That is, a random edge in $G_{n,\lambda n}$ is, w.h.p., in a tree of size $O(1)$. Thus, the proof amounts to showing that, for a fixed tree T and a vertex $v \in T$, we have $c_v^{(N)}/N = 1 + o(1)$ w.h.p.. We prove this in Section 5.1 for both N -uniform and N -balanced models. The proof for $k \geq 3$, given in Section 5.3, requires more ingredients. An additional difficulty is that, for $\lambda < \lambda_k$, a random edge e in $H_{n,\lambda n}^k$ may be in the giant component (if $\lambda \in (\frac{1}{k(k-1)}, \lambda_k)$). However, we rely on the fact that the peeling level of e is $O(1)$ w.h.p., and prove that for a vertex v of bounded level, we have $c_v^{(N)}/N = 1 + o(1)$ w.h.p. as $N \rightarrow \infty$, where the $o(1)$ term does not depend on the size of the giant component.

5.1 CU-process on a fixed tree

5.1.1 Analysis in the N -uniform model

Consider a graph $G = (V, E)$, with m edges, on which the CU-process is run, in the N -uniform model. Recall that $c_v(t)$ (resp. $c_e(t)$) denotes the value of the counter for v (resp. for e) after t steps, and $o_e(t)$ is the number of occurrences e in the first t steps. The aim of this Section is to prove the following result.

Lemma 2 *Let $T = (V, E)$ be a tree, on which the CU-process is run, in the N -uniform model. Let $m = |E|$. Then, for every vertex v of T , there exist absolute positive constants a_v, b_v such that, for any $N \geq 1$ and $x > 0$, we have*

$$\mathbb{P}\left(\max_{t \in [0..Nm]} |c_v(t) - t/m| \geq x\sqrt{N}\right) \leq a_v \exp(-b_v x^2).$$

Lemma 2 implies that, in the N -uniform model, the final counter $c_v^{(N)}$ of every vertex v of T is concentrated around N , with (sub-)Gaussian fluctuations of order $\sqrt{N}$. The same holds for the final counter $c_e^{(N)} = \min(c_u^{(N)}, c_v^{(N)})$ of every edge $e = (u, v)$ of T . On the other hand, the number $o_e(Nm)$ of times e is chosen follows a binomial distribution $\text{Bin}(Nm, 1/m)$. Then, it is also concentrated around N , with Gaussian fluctuations of order $\sqrt{N}$. This implies that $R_e^{(N)} = O(1/\sqrt{N}) = o(1)$ w.h.p. as N gets large.

We say that a family of events $\mathcal{E}_M^N$ indexed by two parameters $N, M \geq 1$, is (N, M) -concentrated if there are absolute constants $a, b > 0$ such that, for every N, M , we have $\mathbb{P}(\mathcal{E}_M^N) \leq a \exp(-bx^2)$, where $x = M/\sqrt{N}$. For $f(t)$ a (possibly random) quantity depending on $t \in [0..Nm]$, we use the notation $\hat{f}(t) := f(t) - t/m$. Thus, to prove Lemma 2, we have to show that for a fixed tree T with m edges on which the N -independent CU-process is run, and for v a vertex of T , the event $\{\max_{t \in [0..Nm]} |\hat{c}_v(t)| \geq M\}$ is (N, M) -concentrated.

We proceed by induction on the peeling level i of vertices. A vertex v at level 0 is a leaf. Let e be its incident edge. It is easy to see that the counter of v increases exactly at the steps when e is drawn. Hence $c_v(t) = o_e(t)$ for $t \in [0..Nm]$. Doob's martingale maximal inequality combined with Hoeffding's inequality ensure that, for every edge e of T , we have

$$\mathbb{P}(\max_{t \in [0..Nm]} |\hat{o}_e(t)| \geq M) \leq 2 \exp(-2x^2/m), \quad \text{where } x = M/\sqrt{N}.$$

Hence, for any leaf v of T , the event $\{\max_{t \in [0..Nm]} |\hat{c}_v(t)| \geq M\}$ is (N, M) -concentrated. Moreover, for v a vertex and e an arbitrary edge incident to v , the fact that $c_v(t) \geq o_e(t)$ ensures that the event $\{\min_{t \in [0..Nm]} \hat{c}_v(t) \leq -M\}$ is (N, M) -concentrated. It thus remains to show that, for vertices of positive levels, the event $\{\max_{t \in [0..Nm]} \hat{c}_v(t) \geq M\}$ is (N, M) -concentrated.

The following statement will be useful to treat the inductive step.

Lemma 3 *Let G be a graph on which a CU-process is run. Let v be a vertex of G , with $v_1, \dots, v_{h+1}$ its neighbours one of which (say v_{h+1}) is distinguished, and with e denoting the edge $\{v, v_{h+1}\}$. For $t \in [0..Nm]$, let $d_v(t) = \max(c_{v_1}(t), \dots, c_{v_h}(t))$. Consider the event $\mathcal{E}_M^N$ that there exists $t \in [0..Nm]$ such that $c_v(t) \geq d_v(t) + M$, and the event $\mathcal{F}_M^N$ that there exists $t \in [0..Nm]$ such that $|\widehat{o}_e(t)| \geq M/4$ or $|\widehat{d}_v(t)| \geq M/4$. Then $\mathcal{E}_M^N$ implies $\mathcal{F}_M^N$.*

Proof: If $\mathcal{E}_M^N$ holds, let $t_0 \geq 0$ be such that $c_v(t_0) \geq d_v(t_0) + M$. Let $t' = \max\{t \leq t_0 \mid c_v(t) \leq d_v(t)\}$. The crucial point is that, in the interval $[t'..t_0]$, any step where c_v increases occurs when e is chosen (indeed, when $c_v(t) > \max(c_{v_1}(t), \dots, c_{v_h}(t))$, choosing an edge $\{v, v_i\}$ with $i \in [1..h]$ yields no increase of c_v). Hence, $c_v(t_0) - c_v(t') \leq o_e(t_0) - o_e(t')$. Since $c_v(t') = d_v(t')$ and $c_v(t_0) \geq d_v(t_0) + M$ we conclude that $o_e(t_0) - o_e(t') \geq d_v(t_0) - d_v(t') + M$. Thus we also have $\widehat{o}_e(t_0) - \widehat{o}_e(t') \geq \widehat{d}_v(t_0) - \widehat{d}_v(t') + M$. Hence $\max(|\widehat{o}_e(t_0)|, |\widehat{o}_e(t')|, |\widehat{d}_v(t_0)|, |\widehat{d}_v(t')|) \geq M/4$, so that $\mathcal{F}_M^N$ holds. $\square$

Let $i \geq 1$, and assume Lemma 2 holds for all vertices of level smaller than i . Let v be a vertex of level i , for which we want to prove that the event $\{\max_{t \in [0..Nm]} \widehat{c}_v(t) \geq M\}$ is (N, M) -concentrated. All neighbours $v_1, \dots, v_{h+1}$ of v have level at most $i - 1$, except for one, say v_{h+1} , with its level in $\{i - 1, i, i + 1\}$ (respectively corresponding to T_i having one vertex, two vertices, or at least three vertices).

Let e be the edge between v and v_{h+1} . Let $\mathcal{E}_M^N$ be the event that $\widehat{c}_v(t) \geq M$ for some $t \in [0..Nm]$. If this holds, then one of the two events $\{\widehat{c}_v(t) - \widehat{d}_v(t) \geq M/2\}$ or $\{\widehat{d}_v(t) \geq M/2\}$ holds. By Lemma 3, the first event implies that $|\widehat{d}_v(t)| \geq M/8$ or $|\widehat{o}_e(t)| \geq M/8$ for some $t \in [0..Nm]$. Hence, the event $\{\max_{t \in [0..Nm]} \widehat{c}_v(t) \geq M\}$ implies that either the event $\{\max_{t \in [0..Nm]} |\widehat{d}_v(t)| \geq M/8\}$ (which is also the union of the events $\{\max_{t \in [0..Nm]} |\widehat{c}_{v_j}(t)| \geq M/8\}$ for $j \in [1..h]$) or the event $\{\max_{t \in [0..Nm]} |\widehat{o}_e(t)| \geq M/8\}$ holds. Since these events are (N, M) -concentrated, we conclude (using the union bound) that the event $\{\max_{t \in [0..Nm]} \widehat{c}_v(t) \geq M\}$ is also (N, M) -concentrated, which concludes the proof of Lemma 2.

5.1.2 Analysis in the N -balanced model

Lemma 4 *Let T be a tree, on which the CU-process is performed, in the N -balanced model. Then, for every edge e of T , we have $R_e^{(N)} = o(1)$ w.h.p. as $N \rightarrow \infty$.*

Proof: Note that the N -balanced model is just the N -uniform model conditioned on all edges occurring exactly N times, which happens with probability $\Theta(n^{-m/2})$ if T has m edges. Let u be an extremity of e . By Lemma 2, there exists a positive constant b such that, in the N -uniform model,

$$\mathbb{P}(c_u(Nm) \geq N + N^{2/3}) = O(e^{-bn^{1/3}}).$$

Hence, in the N -balanced model, we have

$$\mathbb{P}(c_u^{(N)} \geq N + N^{2/3}) = O(n^{m/2} e^{-bn^{1/3}}) = o(1),$$

and thus

$$\mathbb{P}(R_e^{(N)} \geq N^{-1/3}) = \mathbb{P}(c_e^{(N)} \geq N + N^{2/3}) \leq \mathbb{P}(c_u^{(N)} \geq N + N^{2/3}) = o(1),$$

which ensures that $R_e^{(N)} = o(1)$ w.h.p. as $N \rightarrow \infty$. $\square$

5.2 Proof of Theorem 2 for $k = 2$

We use the well-known property [15] that, for fixed $\lambda \in (0, 1/2)$, a random edge e in $G_{n,\lambda n}$ is w.h.p. in a tree of size $O(1)$. Precisely, if we let $\mathcal{E}_{n,\lambda}^m$ be the event that e belongs to a tree of size at most m , then we have the property that, for every $\epsilon > 0$, there exist m and n_0 such that $\mathbb{P}(\mathcal{E}_{n,\lambda}^m) \geq 1 - \epsilon$ for $n \geq n_0$. By Lemma 4, there exists N_0 (depending on m) such that, in the N -balanced or N -uniform model, for every tree T with at most m edges, and every edge $e \in T$, we have $\mathbb{P}(R_e^{(N)} \geq \epsilon) \leq \epsilon$ for $N \geq N_0$. Hence, for $n \geq n_0$ and $N \geq N_0$, if we perform the CU-process on $G_{n,\lambda n}$ in the N -balanced model (note that the N -balanced model holds separately on every connected component), and draw a random edge e , we have

$$\mathbb{P}(R_e^{(N)} \geq \epsilon) \leq \mathbb{P}(R_e^{(N)} \geq \epsilon \mid \mathcal{E}_{n,\lambda}^m) + \mathbb{P}(\neg \mathcal{E}_{n,\lambda}^m) \leq 2\epsilon.$$

This means that $R_e^{(N)} = o(1)$ w.h.p., when n and N grow. In the N -uniform model, the same argument holds, using the fact that the total number of times edges in T are drawn is concentrated around Nm .

5.3 Proof of Theorem 2 for $k \geq 3$

The proof partly follows the same lines as for $k = 2$, but requires additional arguments, in particular a suitably extended notion of peelability. A *marked hypergraph* is a hypergraph $H = (V, E)$ where some of the vertices are marked. The subset of marked vertices is denoted V_∞ . When performing a CU-process on a marked hypergraph, the counters of unmarked vertices are (as usual) initially 0, while the counters at marked vertices are initially (and remain) $+\infty$. When peeling, the marked vertices are not allowed to be peeled (even when they are incident to a unique edge). We define $H_0 = H$, and iteratively for $i \geq 0$, we define V_i as the set of non-marked vertices that are leaves or isolated vertices in H_i , E_i as the set of edges of H_i incident to vertices in V_i , and H_{i+1} to be the hypergraph obtained from H_i by deleting the vertices in V_i and the edges in E_i . A vertex in V_i is said to have *level* i . Then H is called *peelable* if every unmarked vertex is peeled at some step.

Following the exact same lines as in the proof of Lemma 2 (induction on the vertex-levels, and a straightforward adaptation of Lemma 3 to hypergraphs) and Lemma 4, we obtain:

Lemma 5 *Let $H = (V, E)$ be a peelable connected marked hypergraph, on which the CU-process is run, in the N -uniform model. Let $m = |E|$, and for $t \in [0..Nm]$, let $c_v(t)$ be the value of the counter of v after t steps. Then, for every unmarked vertex v of H , there exist absolute positive constants a_v, b_v such that, for any $N \geq 1$ and $x > 0$, we have*

$$\mathbb{P}\left(\max_{t \in [0..Nm]} |c_v(t) - t/m| \geq x\sqrt{N}\right) \leq a_v \exp(-b_v x^2).$$

In the N -uniform or N -balanced model, for every unmarked vertex v of H , we have $c_v^{(N)}/N = 1 + o(1)$ w.h.p. as $N \rightarrow \infty$.

For H a hypergraph, and for v a vertex of finite level in H , a vertex $u \in H$ is called a *descendant* of v if there is a sequence $v_0 = u, v_1, \dots, v_r = v$ of vertices such that, for each $j \in [1..r-1]$, v_j and v_{j+1} are incident to a same edge, and the level of v_{j+1} is larger than the level of v_j .

Lemma 6 *For a hypergraph H , and for a vertex v of H of finite level, let D_v be the set formed by v and its descendants, and let E_v be the set of edges that are incident to at least one vertex from D_v . Let H_v be the marked hypergraph formed by the edges in E_v and their incident vertices, where the marked vertices are those not in D_v . Then H_v is peelable.*

Proof: Let u be a vertex in D_v , of level i . If $i = 0$, then u is a leaf in H . It is also a leaf in H_v , and is immediately peeled. Otherwise, except for possibly one incident edge, u has at least one neighbour v_e of level smaller than i in every incident edge e (note that v_e has to be in D_v). By induction, v_e and e are peeled in the peeling process of H_v . Hence, u is peeled as well during the peeling process on H_v (by induction as well, the level of a vertex in D_v is actually the same in H and in H_v). $\square$

Remark 1 *In a hypergraph $H = (V, E)$, the distance between two vertices u and v is the minimum number of edges to traverse in order to reach v from u . For $r \geq 1$ and $v \in V$, the ball $B_H(v, r)$ is the set of vertices at distance at most r from v . Clearly, if v is at level i , then every vertex in D_v is in $B_H(v, i)$, and every vertex of H_v is in $B(v, i + 1)$*

Lemma 7 (Lemma 3 in [25]) *Let $k \geq 3$, and let $\lambda \in (0, \lambda_k)$. For every $\epsilon > 0$, there exist $I(\epsilon)$ and $n_0(\epsilon)$ such that, for v a random vertex in $H_{n, \lambda n}^k$, the probability that v has level at most $I(\epsilon)$ is at least $1 - \epsilon$, for $n \geq n_0(\epsilon)$.*

Lemma 8 *Let $H = (V, E)$ be a hypergraph, and let E' be a subset of E . Let V' (resp. W) be the subset of vertices incident to at least one edge from E' (resp. incident only to edges in E'). Let $H' = (V', E')$ be the marked hypergraph where the unmarked vertices are those in W . Consider a CU-process on H , with S the sequence of items (each item an edge of H), and for $v \in V$ let c_v be the final value of the counter of v . Let S' be the subsequence of S composed by the items in E' . Consider the CU process on H' where the sequence of items is S' . For $v \in W$, let c'_v be the final value of the counter of v . Then $c'_v \geq c_v$.*

Proof: For an assignment $f : V \rightarrow \mathbb{N} \cup \{+\infty\}$ of initial values to the vertex counters, let $c_f : V \rightarrow \mathbb{N} \cup \{+\infty\}$ be the function giving the final vertex counters after performing the CU-process on an input sequence S . It is easy to check (by induction on the length of S) that the CU-process is monotonous: if $f(v) \leq g(v)$ for all $v \in V$, then $c_f(v) \leq c_g(v)$ for all $v \in V$. Now, if we define f as the function assigning initial value 0 to vertices in W , and initial value $+\infty$ to the other vertices, then $c'_v = c_f(v)$ for all $v \in W$. On the other hand, if f_0 denotes the function assigning initial value 0 to all vertices, then $c_v = c_{f_0}(v)$ for all $v \in V$. Hence, the monotonicity property ensures that $c'_v \geq c_v$ for all $v \in W$. $\square$

We can now conclude the proof of Theorem 2 for $k \geq 3$. Let $\epsilon > 0$. Let $I(\epsilon)$ and $n_0(\epsilon)$ be as given in Lemma 7. Let $s(\epsilon)$ be such that, for a random vertex $v \in H_{n, \lambda n}^k$, we have $\mathbb{P}(|B(v, I(\epsilon) + 1)| \leq s(\epsilon)) \geq 1 - \epsilon$ for all n . The existence of $s(\epsilon)$ easily follows from the known property that $H_{n, \lambda n}$ converges to a Galton-Watson branching process for the local topology, see e.g. [13, Prop 2.6]. Let v be a random vertex in $H_{n, \lambda n}$. Let $\mathcal{E}_{n, \epsilon}$ be the event that the level of v is at most $I(\epsilon)$, and the number of vertices in H_v is at most $s(\epsilon)$. Then, using Remark 1 and the union bound, we have $\mathbb{P}(\mathcal{E}_{n, \epsilon}) \geq 1 - 2\epsilon$ for $n \geq n_0(\epsilon)$. From Lemma 5, there exists N_0 such that, for every marked k -uniform hypergraph H with at most $s(\epsilon)$ vertices, and for every unmarked vertex v of H , we have $\mathbb{P}(c_v^{(N)}/N - 1 \geq \epsilon) \leq \epsilon$ for $N \geq N_0$. Let e be a random edge in $H_{n, \lambda n}^k$, and let v be a random vertex

of e . Note that v is distributed as a random vertex in $H_{n,\lambda n}^k$. With Lemma 6 and Lemma 8, this implies that, conditioned on $\mathcal{E}_{n,\epsilon}$, for $N \geq N_0$ we have $\mathbb{P}(c_v^{(N)}/N - 1 \geq \epsilon) \leq \epsilon$ in the N -balanced model, and also $\mathbb{P}(R_e^{(N)} \geq \epsilon) \leq \epsilon$ (since $c_e^{(N)} \leq c_v^{(N)}$). Since $\mathbb{P}(\neg \mathcal{E}_{n,\epsilon}) \leq 2\epsilon$, we conclude that, for $n \geq n_0(\epsilon)$ and $N \geq N_0$, we have $\mathbb{P}(R_e^{(N)} \geq \epsilon) \leq 3\epsilon$. Hence, in the N -balanced model, $R_e^{(N)} = o(1)$ w.h.p., when n and N grow. In the N -uniform model, the same argument holds again, using the fact that the number of times an edge in H_v is drawn is concentrated around Nm , with m the number of edges of H_v .

5.4 Proof of Theorem 3

The *excess* of a graph G is $\text{exc}(G) = |E| - |V|$.

Lemma 9 *Let $G = (V, E)$ be a graph. Then, for the N -balanced model, we have $\sum_{e \in E} R_e^{(N)} \geq \frac{1}{2} \text{exc}(G)$.*

Proof: During the CU process, each time an edge is drawn, the counter of at least one of its extremities is increased by 1. Hence $\sum_{v \in V} c_v^{(N)} \geq N|E|$. Hence, with the notation $R_v^{(N)} := c_v^{(N)}/N - 1$, we have $\sum_{v \in V} R_v^{(N)} \geq \text{exc}(G)$. Now, by Lemma 1, for each $v \in V$, there exists an edge e_v incident to v such that $c_{e_v}^{(N)} = c_v^{(N)}$ (if several incident edges have this property, an arbitrary one is chosen). Hence, $\sum_{v \in V} R_{e_v}^{(N)} \geq \text{exc}(G)$. Note that, in this sum, every edge occurs at most twice (since it has two extremities), thus $\sum_{e \in E} R_e^{(N)} \geq \frac{1}{2} \text{exc}(G)$. $\square$

For $\lambda > 1/2$, it is known [27, Theorem 6] that there is an explicit constant $\tilde{f}(\lambda) > 0$ such that the excess of the giant component $G' = (V', E')$ of $G_{n,\lambda n}$ is concentrated around $\tilde{f}(\lambda)n$, with fluctuations of order $\sqrt{n}$. Thus, $\text{exc}(G') \geq \frac{1}{2}\tilde{f}(\lambda)n$ w.h.p. as $n \rightarrow \infty$. Hence, by Lemma 9, w.h.p. as $n \rightarrow \infty$ (and for any $N \geq 1$), we have

$$\text{err}_N(G_{n,\lambda n}) \geq \frac{1}{\lambda n} \sum_{e \in E'} R_e^{(N)} \geq \frac{1}{2\lambda n} \text{exc}(G') \geq \frac{1}{4\lambda} \tilde{f}(\lambda) =: f(\lambda).$$

6 Analysis for some non-peelable hypergraphs

Analysing the asymptotic behaviour of the relative error of the CU-process on arbitrary hypergraphs seems to be a challenging task, even if we restrict ourselves to N -uniform and N -balanced models, as we do in this paper. Based on simulations, we expect that, for a fixed connected k -hypergraph $H = (V, E)$, and for $v \in V$, we have $c_v^{(N)}/N = C_v + o(1)$ w.h.p. as $N \rightarrow \infty$, for an explicit constant $C_v \in [1, \deg(v)]$. Since the number of increments at each step lies in $[1..k]$, constants C_v must verify $1 \leq \frac{1}{|E|} \sum_{v \in V} C_v \leq k$, where $\frac{1}{|E|} \sum_{v \in V} C_v$ can be seen as the average number of increments at a step. If H is peelable, then Lemma 5 implies that this concentration holds, with $C_v = 1$. We expect that, if no vertex of H is peelable, and if H is “sufficiently homogeneous”, then the constants C_v should be all equal to the same constant $C > 1$, and thus the relative error $R_e^{(N)}$ of every edge is concentrated around $C - 1 > 0$ w.h.p. as $N \rightarrow \infty$. This, in particular, is supported by an experiment reported at the end of Section 4.2.

In this Section, we show that this is the case for a family of regular hypergraphs which are very sparse ($O(\sqrt{|V|})$ edges) but have a high order (an edge contains $O(\sqrt{|V|})$ vertices). The *dual* of a hypergraph H is the hypergraph H' where the roles of vertices and edges are interchanged: the

vertices of H' are the edges of H , and the edges of H' are the vertices of H so that an edge of H' corresponding to a vertex v of H contains those vertices that correspond to edges incident to v in H . We consider here the hypergraph K'_n dual to the complete graph K_n .

Lemma 10 *Consider any CU-process on $K'_n = (V, E)$. At each step t , let $\text{minedges}(t)$ be the set of edges of E whose counter $c_e(t)$ is minimal over all edges. We have $|\text{minedges}(t)| \geq 2$. Let e be the edge selected at time t . Then, at time t , e is the only edge whose counter increases (by 1), except when $e \in \text{minedges}(t)$ and $|\text{minedges}(t)| = 2$, in which case the counter of the other edge in $\text{minedges}(t)$ also increases (by 1).*

Proof: Every vertex $v \in V$ has degree 2. Considering a vertex v of minimal counter, its two incident edges thus have to be in $\text{minedges}(t)$, therefore $|\text{minedges}(t)| \geq 2$.

Let e be the chosen edge at time t . Clearly, c_e increases (by 1). Moreover, for every edge e' such that $c_{e'}(t) \neq c_e(t)$, $c_{e'}$ does not increase. Let now e' be such that $c_{e'}(t) = c_e(t)$, but there exists an edge $e'' \notin \{e, e'\}$ with $c_{e''}(t) \leq c_{e'}(t)$. Let $v = e' \cap e''$. By Lemma 1, we must have $c_v(t) = c_{e'}(t)$. Since v is not incident to e , its counter does not increase, and neither does the counter of e' . If there does not exist $e'' \notin \{e, e'\}$ with $c_{e''}(t) \leq c_{e'}(t)$, then we must be in the situation where $e \in \text{minedges}(t)$ and $|\text{minedges}(t)| = 2$. Let $v = e \cap e'$. Then, by Lemma 1, we have $c_v(t) = c_e(t) = c_{e'}(t)$, and any vertex $v' \neq v$ must satisfy $c_{v'}(t) > c_v(t)$ (since v' has at least one incident edge different from e, e'). Thus, the counter of v increases, and the counter of e' also increases. $\square$

Theorem 4 *For any fixed $n \geq 2$, in the N -uniform model (resp. in the N -balanced model), the counter of each vertex $v \in K'_n$ satisfies $c_v^{(N)}/N = n/(n-1) + o(1)$ w.h.p. as $N \rightarrow \infty$. Hence, the relative error $R_e^{(N)}$ of each edge e in K'_n satisfies $R_e^{(N)} = 1/(n-1) + o(1)$ w.h.p. as $N \rightarrow \infty$.*

Proof: Consider a CU-process performed on $K'_n = (V, E)$. At each step t , let $p(t) = \min(c_v(t), v \in V) = \min(c_e(t), e \in E)$, and let $q(t) = \max(c_e(t), e \in E)$. It easily follows from Lemma 10 that $\sum_{e \in E} c_e(t) = t + p(t)$. Therefore, $(n-1)p(t) \leq t \leq (n-1)q(t)$, so that $\frac{t}{n-1} \in [p(t), q(t)]$. We now consider the CU process on K'_n in the N -uniform model. Let e_1, e_2 be two edges in E , and let $M \geq 1$. For $t \in [0..Nn]$, the event that $c_{e_1}(t) \geq c_{e_2}(t) + M$ implies the event that $o_{e_1}(t) - o_{e_1}(t') \geq o_{e_2}(t) - o_{e_1}(t') + M$ for some $t' \in [0..t]$ (indeed, letting t' be the last time before t where $c_{e_1} = c_{e_2}$, Lemma 10 ensures that all times in $[t'..t-1]$ where c_{e_1} increases are due to choosing e_1), which is the same as the event that $\widehat{o_{e_1}}(t) - \widehat{o_{e_1}}(t') \geq \widehat{o_{e_2}}(t) - \widehat{o_{e_1}}(t') + M$ for some $t' \in [0..t]$, with the notation of Section 5.1.1. Therefore, the event that $c_{e_2}(t) - c_{e_1}(t) \geq M$ for some $t \in [0..Nn]$ implies the event that $\max(|\widehat{o_{e_1}}(t)|, |\widehat{o_{e_2}}(t)|) \geq M/4$ for some $t \in [0..Nn]$. With the terminology of Section 5.1.1, for every $e \in E$, the event $\max_{t \in [0, Nn]} |\widehat{o_e}(t)| \geq M$ is (N, M) -concentrated. Hence, the event $\{\max_{t \in [0, Nn]} (c_{e_2}(t) - c_{e_1}(t)) \geq M\}$ is (N, M) -concentrated. This implies that $\mathbb{P}(c_{e_2}(Nn) - c_{e_1}(Nn) \geq N^{2/3}) = O(e^{-bN^{1/3}})$ for some $b > 0$, and (by the union bound) that $\mathbb{P}(q(Nn) - p(Nn) \geq N^{2/3}) = O(e^{-\tilde{b}N^{1/3}})$ for some $\tilde{b} > 0$. Note that $\frac{Nn}{n-1} \in [p(Nn), q(Nn)]$, and for every edge $e \in E$ we have $c_e(Nn) \in [p(Nn), q(Nn)]$. Therefore, $\mathbb{P}(|c_e(Nn) - \frac{Nn}{n-1}| \geq N^{2/3}) = O(e^{-\tilde{b}N^{1/3}})$. Hence, in the N -uniform model we have $R_e^{(N)} - \frac{1}{n-1} = O(N^{-1/3}) = o(1)$ w.h.p. as $N \rightarrow \infty$. In the N -balanced model, by the same argument as in Lemma 5.1.2, we have $\mathbb{P}(|c_e^{(N)} - \frac{Nn}{n-1}| \geq N^{2/3}) = O(N^{n/2}e^{-\tilde{b}N^{1/3}})$; hence, $R_e^{(N)} - \frac{1}{n-1} = O(N^{-1/3}) = o(1)$ w.h.p. as $N \rightarrow \infty$. $\square$

Note that with regular Count-Min, the relative error of every edge is $1 + o(1)$ (exactly 1 in the N -balanced model), since all vertices have degree 2. The statement and proof of Theorem 4 easily extend to dual complete hypergraphs as follows. Let $K_{n,r}$ be the complete r -uniform hypergraph on n vertices (for $n \geq r \geq 2$), and consider a CU-process on $K'_{n,r}$ in the N -uniform (resp. N -balanced) model. Then the counter of each vertex $v \in K'_{n,r}$ satisfies $c_v^{(N)}/N = n/(n-r+1) + o(1)$ w.h.p as $N \rightarrow \infty$, hence the relative error $R_e^{(N)}$ of any edge $e \in K'_{n,r}$ satisfies $R_e^{(N)} = \frac{r-1}{n-r+1} + o(1)$ w.h.p. as $N \rightarrow \infty$. For regular Count-Min, the relative error is $r-1 + o(1)$ (exactly $r-1$ in the N -balanced model), since every vertex in $K'_{n,r}$ has degree r .

7 Non-uniform distributions

An interesting and natural question is whether the phase transition phenomenon holds for non-uniform distributions as well. This question is of practical importance, as in many practical situations keys are not distributed uniformly. In particular, Zipfian distributions often occur in various applications and are a common test case for Count-Min sketches [12, 5, 14, 7, 4].

In Zipf's distributions, key probabilities in descending order are proportional to $1/i^\beta$, where i is the rank of the key and $\beta \geq 0$ is the *skewness* parameter. Note that for $\beta = 0$, Zipf's distribution reduces to the uniform one. It is therefore a natural question whether the phase transition occurs for Zipf's distributions with $\beta > 0$.

One may hypothesize that the answer to the question should be positive, as under Zipf's distribution, frequent keys tend to have no error, as it has been observed in earlier papers [5, 4, 3]. On the other hand, keys of the tail of the distribution have fairly similar frequencies, and therefore might show the same behavior as for the uniform case.

However, this hypothesis does not hold. Figure 4 shows the behavior of the average error for Zipf's distributions with $\beta \in \{0.2, 0.5, 0.7, 0.9\}$ vs. the uniform distribution ($\beta = 0$). The average error is defined here as the average error of all keys weighted by their frequencies³, i.e. $\text{err}_N(H) = \frac{1}{mN} \sum_{e \in E} o_e^{(N)} \frac{c_e^{(N)} - o_e^{(N)}}{o_e^{(N)}} = \frac{1}{mN} \sum_{e \in E} (c_e^{(N)} - o_e^{(N)})$. In other words, $\text{err}_N(H)$ is the expected error of a randomly drawn key from the entire input stream of length mN (taking into account multiplicities).

We observe that the phase transition behavior vanishes for $\beta > 0$. It turns out that even in the subcritical regime, frequent elements, while having no error themselves, heavily affect the error of certain rare elements, which raises the resulting average error. In the supercritical regime ($\lambda > 1$ in Figure 4) the opposite happens: the uniform distribution shows the largest average error. This is because an increasingly large fraction of the keys (those in the core of the associated hypergraph) contribute to the error, while for skewed distributions, frequent keys tend to have no error, and thus the larger β (with frequent keys becoming more predominant) the smaller the average error. Note that this is in accordance with the observation of [5] that the estimates for the uniform distribution majorate the estimates of infrequent keys for skewed distributions.

³This definition is natural for non-uniform distributions, as the error for a frequent key should have a larger contribution. Note that it is consistent with the definition of Section 4.1 in the N -balanced case, and in the N -uniform case it presents a negligible difference when N gets large.

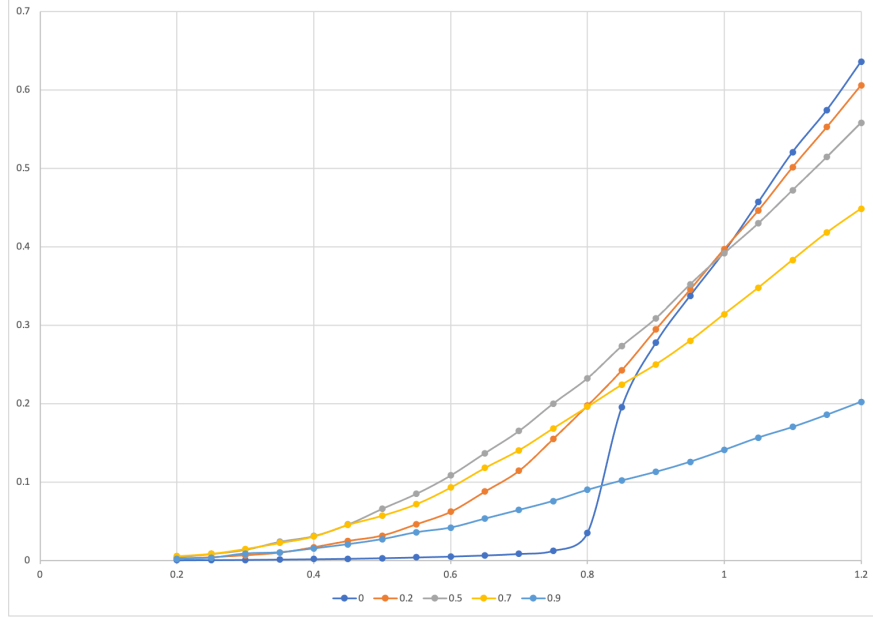


Figure 4: Average error as a function of $\lambda = m/n$, for Zipf's distributions with $\beta \in \{0.0, 0.2, 0.5, 0.7, 0.9\}$. Plots obtained for $n = 1000$, $k = 3$, $N = 50,000$.

8 Concluding remarks

We presented an analysis of conservative update strategy for Count-Min sketch under the assumption of uniform distribution of keys in the input stream. Our results show that the behaviour of the sketch heavily depends on the properties of the underlying hash hypergraph. Assuming that hash functions are k -wise independent, the error produced by the sketch follows two different regimes depending on the density of the underlying hypergraph, that is the number of distinct keys relative to the size of the sketch. When this ratio is below the threshold, the conservative update strategy produces a $o(1)$ relative error when the input stream and the number of distinct keys both grow, while the regular Count-Min produces a positive constant error. This gap formally demonstrates that conservative update achieves a substantial improvement over regular Count-Min.

We showed that the above-mentioned threshold corresponds to the peelability threshold for k -uniform random hypergraphs. One practical implication of this is that the best memory usage is obtained with three hash functions, due to the fact that λ_3 is maximum among all λ_k , and therefore $k = 3$ leads to the minimum number of counters needed to deal with a given number of distinct keys.

In [8] it is claimed, without proof, that the rate of positive errors of conservative update is k times smaller than that of regular Count-Min. This claim does not appear to be true. Note that Count-Min does not err on a key represented in the sketch if and only if the corresponding edge of the hypergraph includes a leaf (vertex of degree 1), while the conservative update can return an exact answer even for an edge without leaves. However, this latter event depends on the relative frequencies of keys and therefore on the specific distribution of keys and the input length. On the other hand, our experiments with uniformly distributed keys show that this event is relatively rare,

and the rate of positive error for Count-Min and conservative update are essentially the same.

One important assumption of our analysis is the uniform distribution of keys in the input. We presented an experimental evidence that for skewed distributions, in particular for Zipf’s distribution, the phase transition disappears when the skewness parameter grows. Therefore, the uniform distribution presents the smallest error in the subcritical regime. The situation is the opposite in the supercritical regime when the number of distinct keys is large compared to the number of counters: here the uniform distribution presents the largest average error. As mentioned earlier, for Zipf’s distribution, frequent keys have essentially no error, whereas in the supercritical regime, low frequency keys have all similar overestimates. This reveals another type of phase transition in error approximation for Zipf’s distribution, occurring between frequent and infrequent elements. Quantifying this transition is an interesting open question directly related to the accurate detection of *heavy hitters* in streams.

Acknowledgments We thank Djamel Bellazougui who first pointed out to us the conservative update strategy.

References

- [1] Sairam Behera, Sutanu Gayen, Jitender S Deogun, and NV Vinodchandran. Kmerestimate: A streaming algorithm for estimating k-mer counts with optimal space usage. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 438–447, 2018.
- [2] Michael Behrisch, Amin Coja-Oghlan, and Mihyun Kang. Local limit theorems for the giant component of random hypergraphs. *Comb. Probab. Comput.*, 23(3):331–366, 2014. doi:10.1017/S0963548314000017.
- [3] Younes Ben Mazziiane, Sara Alouf, and Giovanni Neglia. Analyzing Count Min Sketch with Conservative Updates. working paper or preprint, April 2022. URL: <https://hal.archives-ouvertes.fr/hal-03613957>.
- [4] Younes Ben Mazziiane, Sara Alouf, and Giovanni Neglia. A formal analysis of the count-min sketch with conservative updates. *CoRR*, abs/2203.14549, 2022. arXiv:2203.14549, doi:10.48550/arXiv.2203.14549.
- [5] Giuseppe Bianchi, Ken Duffy, Douglas J. Leith, and Vsevolod Shneer. Modeling conservative updates in multi-hash approximate count sketches. In *24th International Teletraffic Congress, ITC 2012, Kraków, Poland, September 4-7, 2012*, pages 1–8. IEEE, 2012. URL: <https://ieeexplore.ieee.org/document/6331813/>.
- [6] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.
- [7] Peiqing Chen, Yuhan Wu, Tong Yang, Junchen Jiang, and Zaoxing Liu. Precise error estimation for sketch-based flow measurement. In *Proceedings of the 21st ACM Internet Measurement Conference, IMC ’21*, page 113–121, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3487552.3487856.

- [8] Saar Cohen and Yossi Matias. Spectral bloom filters. In Alon Y. Halevy, Zachary G. Ives, and AnHai Doan, editors, *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003*, pages 241–252. ACM, 2003. doi:10.1145/872757.872787.
- [9] Graham Cormode. Count-min sketch. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, Second Edition*. Springer, 2018. doi:10.1007/978-1-4614-8265-9_87.
- [10] Graham Cormode and Marios Hadjieleftheriou. Finding frequent items in data streams. *Proceedings of the VLDB Endowment*, 1(2):1530–1541, 2008.
- [11] Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, April 2005. URL: <http://www.sciencedirect.com/science/article/pii/S0196677403001913>, doi:10.1016/j.jalgor.2003.12.001.
- [12] Graham Cormode and S. Muthukrishnan. Summarizing and mining skewed data streams. In Hillol Kargupta, Jaideep Srivastava, Chandrika Kamath, and Arnold Goodman, editors, *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005, Newport Beach, CA, USA, April 21-23, 2005*, pages 44–55. SIAM, 2005. doi:10.1137/1.9781611972757.5.
- [13] Amir Dembo and Andrea Montanari. Gibbs measures and phase transitions on sparse random graphs. *Brazilian Journal of Probability and Statistics*, 24(2):137–211, 2010.
- [14] Gil Einziger and Roy Friedman. A formal analysis of conservative update based approximate counting. In *International Conference on Computing, Networking and Communications, ICCNC 2015, Garden Grove, CA, USA, February 16-19, 2015*, pages 255–259. IEEE Computer Society, 2015. doi:10.1109/ICCNC.2015.7069350.
- [15] Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5(1):17–60, 1960.
- [16] Cristian Estan and George Varghese. New directions in traffic measurement and accounting. In Matthew Mathis, Peter Steenkiste, Hari Balakrishnan, and Vern Paxson, editors, *Proceedings of the ACM SIGCOMM 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, August 19-23, 2002, Pittsburgh, PA, USA*, pages 323–336. ACM, 2002. doi:10.1145/633025.633056.
- [17] Bin Fan, Dave G. Andersen, Michael Kaminsky, and Michael D. Mitzenmacher. Cuckoo filter: Practically better than bloom. In *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies, CoNEXT ’14*, page 75–88, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2674005.2674994.
- [18] Li Fan, Pei Cao, J. Almeida, and A.Z. Broder. Summary cache: a scalable wide-area web cache sharing protocol. *IEEE/ACM Transactions on Networking*, 8(3):281–293, 2000. doi:10.1109/90.851975.
- [19] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2015. doi:10.1017/CB09781316339831.

- [20] Michael T Goodrich and Michael Mitzenmacher. Invertible bloom lookup tables. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 792–799. IEEE, 2011.
- [21] Michał Karoński and Tomasz Łuczak. The phase transition in a random hypergraph. *Journal of Computational and Applied Mathematics*, 142(1):125–135, 2002. Probabilistic Methods in Combinatorics and Combinatorial Optimization. URL: <https://www.sciencedirect.com/science/article/pii/S0377042701004642>, doi:[https://doi.org/10.1016/S0377-0427\(01\)00464-2](https://doi.org/10.1016/S0377-0427(01)00464-2).
- [22] Hongyan Liu, Yuan Lin, and Jiawei Han. Methods for mining frequent items in data streams: an overview. *Knowledge and information systems*, 26(1):1–30, 2011.
- [23] Bohdan S Majewski, Nicholas C Wormald, George Havas, and Zbigniew J Czech. A family of perfect hashing methods. *The Computer Journal*, 39(6):547–554, 1996.
- [24] Hamid Mohamadi, Hamza Khan, and Inanc Birol. ntcad: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics*, 33(9):1324–1330, 2017.
- [25] Michael Molloy. Cores in random hypergraphs and boolean formulas. *Random Structures & Algorithms*, 27(1):124–135, 2005.
- [26] Rasmus Pagh and Flemming Friche Rodler. Cuckoo hashing. *Journal of Algorithms*, 51(2):122–144, 2004.
- [27] Boris Pittel and Nicholas C Wormald. Counting connected graphs inside-out. *Journal of Combinatorial Theory, Series B*, 93(2):127–172, 2005.
- [28] Yoshihiro Shibuya and Gregory Kucherov. Set-min Sketch: a Probabilistic Map for Power-Law Distributions with Application to k-mer Annotation. *bioRxiv*, page 2020.11.14.382713, November 2020. URL: <https://www.biorxiv.org/content/10.1101/2020.11.14.382713v1>, doi: 10.1101/2020.11.14.382713.
- [29] Stefan Walzer. *Random hypergraphs for hashing-based data structures*. PhD thesis, Technische Universität Ilmenau, Germany, 2020. URL: https://www.db-thueringen.de/receive/dbt_mods_00047127.