# Hashing: Construction of A Hash Family With Smaller Description Length

Mingkun Ni, Yuanhao Zhang

{m8ni, y2384zha}@uwaterloo.ca

University of Waterloo

Waterloo, ON, Canada

August 9, 2020

## Abstract

**TO BE UPDATED: The following is a placeholder for abstract.** After reading through some related research papers, we decide to look into one of them, titled *Ball and bins: smaller hash families and faster evaluation*. This paper introduces two new constructions that we could guarantee $O(\log n/\log\log n)$ maximum load when throw $n$ balls into $n$ with either a smaller description length or a faster calculation time. It is well-known that, with high probability, a $O(\log n/\log\log n)$-wise independent hash family would guarantee max load of $O(\log n/\log\log n)$. Such $O(\log n/\log\log n)$-wise independent function can be described by $O(\log^2 n/\log\log n)$ bits, which already yields a dramatic improvement over a truly random function. This paper aims to find an even smaller description length or a faster calculation time of the function while maintaining the guarantee of $O(\log n/\log\log n)$ max load. The special part of this research, specifically about the first construction, is that it constructs an innovative structure of a multi-layer random graph. With such construction of multi-layer graph, each layer is considered as a different hash process with different input and output sizes, which is the number of bins in each layer. Thus, we want to look deep in the first construction of this research and elaborate more details.

# 1 - Introduction

A traditional analysis of randomized algorithm to map $m$ balls into $n$ bins independently and uniformly guarantees that each bin contains at most $O(\log n/\log\log n)$ balls with high probability, as known as the maximum load of the balls and bins problem. For a truly random hash function $h(x) : M \to N$, it would take $O(mlogn)$ space to store it. The traditional analysis with the use of truly random hash functions is impractical in various real-world applications because of the space to store the hash functions. Hence, a weaker notion of randomness, k-wise independence, is introduced to solve this issue. It is specifically well-studied in the case of mapping $n$ balls into $n$ bins that any $O(\log n/\log\log n)$-wise independent hash families can guarantee the maximum load of $O(\log n/\log\log n)$ with high probability.

This paper will continue to study the problems of mapping $n$ balls into $n$ bins with a construction of hash functions that require a smaller description length given the inspiration from the paper, titled *Ball and bins: smaller hash families and faster evaluation*. By using a $O(\log n/\log\log n)$-wise independent hash families, the hash functions can be described by $O(\log^2 n/\log\log n)$ bits, which itself yields a dramatic improvement over the description length of a truly random functions. We would like to provide an explicit family of hash functions to guarantee the same maximum load of $O(\log n/\log\log n)$ with high probability, and each hash function can be strictly described by $o(\log^2 n/\log\log n)$ bits. We provide an overview of the construction below.

## 1.1 - Construction: COPIED AND PASTED

Our construction is based on concatenating the outputs of $O(loglogn)$ functions which are gradually more independent: each function $f$ in our construction is described using d functions

$$f(x) = h_1(x) \circ ... \circ h_d(x),$$

where we view the output of each $h_i$ as a binary string, and a $\circ$ denotes the concatenation operator on binary strings. The first function $h_1$ is only $O(1)$-wise independent, and the level of independence gradually increases to $O(logn/loglogn)$-wise independence for the last function $h_d$. As we increase the level of independence, we decrease the output length of the functions from $\Omega(logn)$ bits for $h_1$ to $O(loglogn)$ bits for $h_d$. We instantiate these $O(loglogn)$ functions using $\epsilon$-biased distributions. The trade-off between the level of independence and the output length implies that each of these functions can be described using only $O(logn)$ bits and evaluated in time $O(logn)$.

### 1.2 - Contribution:

We note that the above construction is from the paper, titled *Ball and bins: smaller hash families and faster evaluation.* This paper was written to fully understand the construction presented in Section 1.1. This construction was the study results of L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. This paper will present a full and complete tour guide in understanding that this construction indeed guaranteed the same maximum load with a smaller description length. In the original paper, proofs of many lemmas and theorems are neglected. We have expanded many of lemmas and theorems given in the original paper with elaboration and proofs.

### 1.3 - Outline:

In Section 2, we will introduce a few pieces of terminology, definitions, lemmas, and theorems that we will be using in latter sections. Section 3 is the essential part of this research paper. It will contain the formal introduction of our construction and formal proof of why this construction works. It will first give a formal description of the construction, and we will be analyzing the construction in which we will be essentially explaining the interpretation of such construction. Finally, we will prove step-by-step that the construction guarantees the description length of hashing functions. Finally, Section 4 will introduce some extensional use of this construction. It will also analyze this construction with correspondence to the trade-off it makes to obtain the smaller description length of the function.

## 2 - Preliminary Tools

In this section, we present the relevant definitions, lemmas, theorems that we will use to prove the construction.

### 2.1 - Definitions:

GIVE EXAMPLE to differentiate from original paper like item 6
We use the unit RAM model throughout the paper. In the RAM model, we assume that we can access an arbitrary position of an array in $O(1)$ time. We also assume that word size is large enough such that it takes $O(1)$ word oepration. For the balls and bins problem, we are considering the case where there are exactly $n$ balls and $n$ bins. We want to achieve a maximum load of $O(logn/loglogn)$ with smaller than usual description length under this condition. The following

are some definitions and terms that we will be using frequently throughout the paper.

1. For a natural number u, we define the set of integers $\{1, 2, ..., n\}$ as [u].
2. We represent a uniform distribution over the set $\{0,1\}^n$ by $U_n$.
3. The term $x \in X$ represents a sample $x$ from a random variable $X$.
4. $SD(X, Y)$ represents the statistical distance between two random variables over finite domain

$$SD(X, Y) = \frac{1}{2} \sum_{w \in \Omega} |Pr(X \in \omega) - Pr(Y \in \omega)|$$

5. The term $x \in S$ means that we draw a random sample $x$ uniformly from a finite set $S$.
6. For two bit-string $x$ and $y$, $x \circ y$ represents the concatenation of x and y bit-string. For example, for $x = 1001$ and $y = 0111$,

$$x \circ y = 1001 \circ 0111 = 10010111$$

## 2.2 - More Definitions:

We present a few more definitions. These definitions are more complicated but essential to understanding the proof the construction.

**k-wise $\delta$-dependent:** Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

**$\epsilon$-biased distribution:** It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).

**min-entropy:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**$k$-source:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**$(J, k)$-block source:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**$(J, k, \epsilon)$-block source:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structure.

## 2.3 - Theorems:

In this section, we can present the key lemmas, corollaries, and theorems that we use to prove the construction. Refer to Section 2.1 and Section 2.2 if encountering unknown definitions.

**Corollary 2.1:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**Corollary 2.2:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**Lemma 2.3:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

**Corollary 2.4:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words,

combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

# 3 - Construction

In this section, we will first present the construction, mentioned in Section 1.1, based on the gradually increasing independence. This construction will guarantee a maximum load of $O(\log n/\log\log n)$ using $O(\log n/\log\log n)$ with a smaller description length. This constructions allows us to prove the following theorem:

**Theorem 3.1:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.
In what follows we provide a more formal description of our construction (see Section 3.1), and then analyze it for proving Theorem 3.1 (see Section 3.2).

## 3.1 - Formal Description of Construction

COPIED AND PASTED. We assume that n is a power of two, as otherwise we can choose the number of bins to be the largest power of two which is smaller than n, and this may affect the maximal load by at most a multiplicative factor of two. Let d = O(log log n).....

We will also visualize the construction as a tree of $d + 1$ layers. There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

## 3.2 - Analysis of Construction

There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text.

**Lemma 3.2:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text.

$$formula$$

**Proof.** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text.

**Step-by-Step Proof of Theorem 3.1:** There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc. There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.

# 4 - Extension and Appendix

This section presents some extensional use and appendix. There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text.

# References


[1] L. E. Celis, O. Reingold, G. Segev and U. Wieder, "Balls and Bins: Smaller Hash Families and Faster Evaluation," 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, Palm Springs, CA, 2011, pp. 599-608, doi: 10.1109/FOCS.2011.49.

[2] A. Pagh and R. Pagh. Uniform hashing in constant time and optimal space. SIAM Journal on Computing, 38(1):85–96, 2008.

[3] J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. SIAM Journal on Computing, 22(4):838–856, 1993.

[4] M. Dietzfelbinger and F. Meyer auf der Heide. A new universal class of hash functions and dynamic hashing in real time. In Proceedings of the 17th International Colloquium on Automata, Languages and Programming, pages 6–19, 1990.

[5] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos. Linear hash functions. Journal of the ACM, 46(5):667–683, 1999.

[6] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple construction of almost kwise independent random variables. Random Structures and Algorithms, 3(3):289–304, 1992.

[7] O. Goldreich and A. Wigderson. Tiny families of functions with random properties: A quality-size trade-off for hashing. Random Structures and Algorithms, 11(4):315–343, 1997.