中国科学技术大学
University of Science and Technology of China

伪随机理论及其运用
以最小哈希族的构造为例

部分结果已被 SODA 2026 接收 (Chen, Huang, and Li, 2026)

报告人：黄盛唐 (PB22000196)
指导老师：陈雪 特任教授

中国科学技术大学

2026 年 6 月 3 日

- **随机化**是计算机科学的核心方法之一：算法、密码学、概率方法 ...
- 真随机比特**昂贵且稀缺** $\Longrightarrow$ **去随机化**：用尽量少的随机性模拟真随机的行为。

- **随机化**是计算机科学的核心方法之一：算法、密码学、概率方法 …
- 真随机比特**昂贵且稀缺** $\Longrightarrow$ **去随机化**：用尽量少的随机性模拟真随机的行为。

定义 – 伪随机生成器（PRG）

$G:\{0,1\}^{\ell}\to D$ 让任何 $f\in\mathcal{F}$ 都无法区分 $G(U_\ell)$ 与真随机的 $x\sim D$（误差 ε）。ℓ 称为 **种子长度**。

$$\forall f\in\mathcal{F},\quad \mathbb{E}[f(G(U_\ell))]=\mathop{\mathbb{E}}_{x\sim D}[f(x)]\pm\varepsilon.$$

- **哈希函数族** $\Longleftrightarrow$ **PRG**：从 $\mathcal{H}=\{h:[N]\to[M]\}$ 中均匀采样 h，等价于以 $\log|\mathcal{H}|$ 比特种子在 $[M]^N$ 中生成一个向量。
- 因此哈希族**规模的对数** $=$ 种子长度 $=$ 随机性使用，是衡量构造质量的核心量。

定义 – [Broder, Charikar, Frieze, and Mitzenmacher, 1998]

$\mathcal{H} = \{h : [N] \to [M]\}$ 是误差为 δ 的**最小哈希族**，若对任意 $X \subseteq [N]$ 与 $y \notin X$，

$$\Pr_{h \sim \mathcal{H}} \left[h(y) < \min_{x \in X} h(x) \right] = \frac{1 \pm \delta}{|X| + 1}.$$

将 y 推广为大小至多 k 的子集 Y $(X \cap Y = \varnothing)$，即得 k-**最小哈希族。**

定义 – [Broder, Charikar, Frieze, and Mitzenmacher, 1998]

$\mathcal{H} = \{h : [N] \to [M]\}$ 是误差为 δ 的**最小哈希族**，若对任意 $X \subseteq [N]$ 与 $y \notin X$，

$$\Pr_{h\sim\mathcal{H}}\left[h(y) < \min_{x\in X} h(x)\right] = \frac{1 \pm \delta}{|X| + 1}.$$

将 y 推广为大小至多 k 的子集 Y $(X \cap Y = \varnothing)$，即得 k-**最小哈希族。**

广泛应用：

- 集合相似度估计、网页去重
- 流式 ℓ_0 采样、草图算法
- 图聚类、稀有度近似、数据挖掘

最经典应用：Jaccard 相似度

对集合 $A, B \subseteq [N]$，

$$\Pr_{h}\left[\min_{a\in A} h(a) = \min_{b\in B} h(b)\right] = (1 \pm \delta) \cdot \frac{|A \cap B|}{|A \cup B|}.$$

$\Longrightarrow$ 天然要求**乘法误差。**

路线一：基于有限次独立性（bounded independence）

- Indyk (2001)：$t = O(\log 1/\delta)$ 次独立 $\Rightarrow$ 误差 δ 的最小哈希族；
- Pătraşcu and Thorup (2016)：$t = \Omega(\log 1/\delta)$ 必要 $\Rightarrow$ 紧致；
- Feigenblat, Porat, and Shiftan (2011)：对 k-最小哈希给出 $O(\log 1/\delta + k \log\log 1/\delta)$ 次独立上界。

局限：（i）k-最小哈希所需独立次数上下界仍有 **gap**；（ii）t 次独立需 $\Theta(t \log NM)$ 比特，当 $\delta = o(1)$ 时**无法达到最优种子长度** $O(\log(NM/\delta))$。

已有显式构造的两条路线及其局限

路线一：基于有限次独立性（bounded independence）

- Indyk (2001)：$t = O(\log 1/\delta)$ 次独立 $\Rightarrow$ 误差 δ 的最小哈希族；
- Pătraşcu and Thorup (2016)：$t = \Omega(\log 1/\delta)$ 必要 $\Rightarrow$ 紧致；
- Feigenblat, Porat, and Shiftan (2011)：对 k-最小哈希给出 $O(\log 1/\delta + k \log\log 1/\delta)$ 次独立上界。

局限：（i）k-最小哈希所需独立次数上下界仍有 **gap**；（ii）t 次独立需 $\Theta(t \log NM)$ 比特，当 $\delta = o(1)$ 时**无法达到最优种子长度** $O(\log(NM/\delta))$。

路线二：基于组合矩形（combinatorial rectangles）的 PRG

- Saks, Srinivasan, Zhou, and Zuckerman (1999)：归约到组合矩形 PRG，得到 $O(\log^{3/2} NM)$ 种子长度，多项式小的误差（$1/\mathrm{poly}(N)$）；
- Gopalan and Yehudayoff (2020)（SOTA）：$O(\log NM \cdot \log\log NM)$ 种子长度，多项式小的误差。

局限：最小哈希要求 $\delta/|X|$ 量级的乘法误差，对应极小加法误差；该路线**无法消除** $\log\log NM$ **因子。**

定理 – 非形式化

基于有限次独立性的构造，要实现误差为 δ 的 k-最小哈希族，

$\boxed{\Theta\left(k+\log 1/\delta\right)}$ 次独立性是充分且必要的。

定理 – 非形式化

基于有限次独立性的构造，要实现误差为 δ 的 k-最小哈希族，

$\boxed{\Theta\left(k+\log 1/\delta\right)}$ 次独立性是充分且必要的。

哈希族	所需独立次数	来源
最小哈希（$k=1$）	$\Theta(\log 1/\delta)$（紧）	Ind01; PT16
k-最小哈希	$O(\log 1/\delta + k\log\log 1/\delta)$（上界，不紧）	FPS11
k-最小哈希	$\boldsymbol{\Theta(k+\log 1/\delta)}$（紧）	**本文**

定理 – 非形式化，设 $M = \mathrm{poly}(N)$

存在最小哈希族的显式构造，种子长度为 $\boxed{O(\log N)}$（最优），乘法误差为 $2^{-O(\log N/\log\log N)}$（亚常数，几乎多项式小）。

定理 – 非形式化，设 $M = \mathrm{poly}(N)$

存在最小哈希族的显式构造，种子长度为 $\boxed{O(\log N)}$（最优），乘法误差为 $2^{-O(\log N/\log\log N)}$（亚常数，几乎多项式小）。

方法	种子长度	误差	来源
有限次独立性	$\Theta(\log 1/\delta \cdot \log N)$	δ	Ind01; FPS11
组合矩形的 PRG	$O(\log N \cdot \log\log N)$	$1/\mathrm{poly}(N)$	SSZZ99; GY20
非显式构造	$\Omega(\log(N/\delta))$	δ	—
本文	$\boldsymbol{O(\log N)}$	$\boldsymbol{2^{-O(\log N/\log\log N)}}$	—

- 首次在最优种子长度下，将乘法误差压低至亚常数级别；
- 推广到 k-最小哈希：$k = \log^{O(1)} N$ 时，种子长度 $O(k \log N)$，误差量级不变。

感谢各位老师的聆听！

敬请批评指正。

[BCFM98] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. In *Proceedings of the thirtieth annual ACM symposium on Theory of Computing*, pages 327–336, 1998.

[CHL26] Xue Chen, Shengtang Huang, and Xin Li. Explicit Min-wise Hash Families with Optimal Size. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2512–2539. SIAM, 2026.

[FPS11] Guy Feigenblat, Ely Porat, and Ariel Shiftan. Exponential time improvement for min-wise based algorithms. *Information and Computation*, 209(4):737–747, 2011.

[GY20] Parikshit Gopalan and Amir Yehudayoff. Concentration for Limited Independence via Inequalities for the Elementary Symmetric Polynomials. *Theory of Computing*, 16:1–29, 2020.

[Ind01] Piotr Indyk. A Small Approximately Min-Wise Independent Family of Hash Functions. *Journal of Algorithms*, 38(1):84–90, 2001.

[PT16] Mihai Pătraşcu and Mikkel Thorup. On the k-Independence Required by Linear Probing and Minwise Independence. *ACM Transactions on Algorithms*, 12(1):8:1–8:27, 2016.

[SSZZ99] Michael E. Saks, Aravind Srinivasan, Shiyu Zhou, and David Zuckerman. Low Discrepancy Sets Yield Approximate Min-Wise Independent Permutation Families. In *RANDOM/APPROX*, pages 11–15, 1999.