Understanding Circulant Permutation to Extend the CMinHash Estimator

Keegan Kang
Department of Mathematics and Statistics
Bucknell University
Lewisburg, PA 17837
kk054@bucknell.edu

Avery Hood
Department of Mathematics and Statistics
Bucknell University
Lewisburg, PA 17837
AveryHood1013@gmail.com

Benedict H. W. Wong
A*STAR Research Entities
Singapore, 138632
Benedict_Wong@a-star.edu.sg

Abstract

CMinHash promises a new direction for designing MinHash algorithms with reduced variance and storage space via circulant permutations. However, the original variance analysis is difficult to extend to other MinHash estimators. We present a new approach that gives explicit and easily comparable approximate variance expressions, and use it to analyze the Minner estimator. Numerical experiments show that our variance expression matches the empirical mean square error (MSE) of estimates. Our work thereby demonstrates how to apply circulant permutation to other MinHash estimators and compute their approximate variance.

1 Introduction

Consider two unweighted sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \Omega$, where $\Omega = \{1, 2, \ldots, D\}$. Let $|\mathcal{S}_1| = f_1$, $|\mathcal{S}_2| = f_2$, and $a = |\mathcal{S}_1 \cap \mathcal{S}_2|$, and $f = |\mathcal{S}_1 \cup \mathcal{S}_2|$. A common measure of similarity between these two sets is the Jaccard similarity, given by $J = \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} = \frac{a}{f_1 + f_2 - a}$. As D is usually large, directly computing the pairwise similarity between these sets is not computationally feasible, and the task is to instead estimate the Jaccard similarity of these sets via hashing algorithms. Measuring this similarity is useful in many applications, even in non statistical (or computer science) disciplines [16, 17, 21]. The traditional MinHash algorithm [3] uses random permutations to estimate J between any two sets, and since then there has been a large body of work on estimating J, which we divide into three categories.

New MinHash Approaches: These examples of MinHash algorithms describe a novel (at the time) approach to MinHash. For example, *k-min sketches / Conditional Random Sampling* records more information with one random permutation [4, 11], *b-bit minwise hashing* which stores the last b bits of information [8, 9], *MinMaxHash* uses the "other end" of the random permutation [5], *One Permutation Hashing* uses just one permutation [12, 18, 19], the *Minner estimator* and records more data for asymmetric hashing [1]. These contributions usually involve some kind of trade-off (e.g. data is required to be sparse for a better estimate, or greater computational time for reduced storage space), i.e. these methods do not work uniformly well for *all* datasets.

"Framework" that complements existing MinHash methods: These techniques can be applied to an existing algorithm to always improve its performance. For example, the technique of *control variates and maximum likelihood estimation* can be applied to MinHash (and other similar algorithms) to always reduce their variance [6, 10], the *combined estimator setup* in [20] uses more information to

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

get a better estimator when set sizes are sufficiently unequal, and *CMinHash* shows that "circulantly" moving rightwards across a permutation always reduces variance of estimates [14, 15].

Theoretical work which gives insight on the structure of MinHash: For example, [2] gives a concise theorem on query time, space usage, and update time of similarity search algorithms, [13] explains how to construct a MinHash type estimator across continuous probability distributions.

Our work belongs to the last two categories, where we re-examine the approach taken in the CMinHash paper [15], and propose a way to extend circulant permutations to other MinHash estimators.

Permutation	6	1	7	4	5	8	3	2
Set 1	•	○	○	•	○	•	○	**•**
Set 2	○	**•**	•	•	○	○	•	•

Draw Order	1	2	3	4	5	6	7	8
Set 1	○	**•**	○	•	○	•	○	•
Set 2	**•**	•	•	•	○	○	•	○

Figure 1: We have $|\mathcal{S}_1| = 4$, $|\mathcal{S}_2| = 5$. By treating a *random permutation* (left) as the indices where we *sample without replacement* (right), $\min\{\pi(\mathcal{S}_1)\} = 2$, $\min\{\pi(\mathcal{S}_2)\} = 1$, which is just the index where the first • appears.

1.1 Related Work

Without loss of generality, let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \Omega$, and we want to estimate the *Jaccard similarity* J between these sets. These sets can be represented as D sequential objects consisting of one of four types: $\left\{ \begin{matrix} \bullet \\ \bullet \end{matrix}, \begin{matrix} \bullet \\ \circ \end{matrix}, \begin{matrix} \circ \\ \bullet \end{matrix}, \begin{matrix} \circ \\ \circ \end{matrix} \right\}$, where the • in the corresponding row denotes an element of that set.

Theorem 1. *(Broder's MinHash) Suppose we have K independent random permutations $\pi_1, \dots, \pi_K$ on $[D]$. Let $\theta^{(\mathrm{MH})} = \sum_{t=1}^{K} \mathbb{1}_{\{\min\{\pi_t(\mathcal{S}_1)\}=\min\{\pi_t(\mathcal{S}_2)\}\}}/K$. Then $\mathbb{E}\left(\theta^{(\mathrm{MH})}\right) = J$, with $\mathrm{Var}\left(\theta^{(\mathrm{MH})}\right) = J(1-J)/K$.*

A *random permutation* π on $[D]$ can be thought of as a sequence of draws under *sampling without replacement*, where $\min\{\pi(\mathcal{S}_1)\}$ is the draw where the first • appears in the first row, and $\min\{\pi(\mathcal{S}_2)\}$ as the draw where the first • appears in the second row. Figure 1 shows an example where $D = 8$. Let the set $\mathcal{F}_+ = \left\{ \begin{matrix} \bullet \\ \bullet \end{matrix} \right\}$, and $\mathcal{F}_- = \left\{ \begin{matrix} \circ \\ \bullet \end{matrix}, \begin{matrix} \bullet \\ \circ \end{matrix} \right\}$. Each permutation in Broder's MinHash is equivalent to a trial where a "success" is drawing an element in $\mathcal{F}_+$ before an element in $\mathcal{F}_-$, and an estimator for J is found by finding the probability that an element from $\mathcal{F}_+$ **is seen before an element in** $\mathcal{F}_-$. Hence by Theorem 1, we get an estimate for J by recording the proportion of "successes" for K independent random permutations (trials), $1 \le t \le K$.

More generally, other MinHash algorithms can be represented by different D sequential objects, $\mathcal{F}_+$ and $\mathcal{F}_-$, and an estimator for J can be found by finding the probability that n_1 elements from $\mathcal{F}_+$ are seen before n_2 elements in $\mathcal{F}_-$ (we give an example in Section 3).

Permutation	3	2	6	1	7	4	5	8
Set 1	•	○	○	**•**	○	•	○	•
Set 2	○	•	•	**•**	○	○	•	•

Draw Order	1	2	3	4	5	6	7	8
Set 1	**•**	○	•	•	○	○	○	•
Set 2	**•**	•	○	○	•	•	○	•

Figure 2: The permutation π in Figure 2 is shifted rightward from Figure 1 by 2 units. Then $\min\{\pi_{\to 2}(\mathcal{S}_1)\} = \min\{\pi_{\to 2}(\mathcal{S}_2)\} = 1$.

CMinHash extends Broder's MinHash by reducing the K independent permutations needed to two independent permutations, given by σ, π, and we write this as CMinHash(σ, π) following the notation in [15]. First, the indices in the sets $\mathcal{S}_i$ are permuted by the random permutation σ. Second, the random permutation π is generated, and is re-used K times by shifting π circulantly rightwards by 1 unit K times. More formally, let $\pi_{\to t}$ denote shifting the permutation t units to the right. Figure 2 gives an example of shifting the permutation in Figure 1 rightwards by 2 units. By linearity of expectations, $\mathbb{E}_{\sigma,\pi}\left(\sum_{t=1}^{K} \mathbb{1}_{\{\min\{\pi_{\to t}(\mathcal{S}_1)\}=\min\{\pi_{\to t}(\mathcal{S}_2)\}\}}/K\right) = J$, hence recording the proportion

of "successes" under K shifts gives an estimate of J. A less compact version is on Page 13 in Appendix A.1. However, the variance of this estimate is slightly complicated, although lower than Broder's MinHash (see Page 14 in Appendix A.2 for expression).

1.2 Our Contributions And Paper Organization

CMinHash(σ, π) promises a new direction to design hashing algorithms using the 2 permutations (σ, π) and K circulant shifts of π (saving on storage space), but having a smaller variance than K independent permutations under traditional MinHash (and similar algorithms).

While the algorithm behind CMinHash(σ, π) is straightforward to implement (see Algorithm 2 in Appendix A.1), the theoretical analysis of the variance forms a substantial bulk of the original paper.

Moreover, the approach the original paper adopts in terms of random permutations makes the proofs of their Lemma 2.4 (used to compute expectations under σ) and Theorem 3.1 (used to compute the variance of the estimates under σ and π) difficult to replicate in other settings due to the random permutation paradigm and combinatorial calculations. Moreover, the result of Theorem 3.1 (variance of the CMinHash estimator) in [15] gives the variance as a combinatorial sum under constraints. Appendix A.2 notes that one equality used in their proof does not hold in general. This makes it difficult to theoretically compare the CMinHash variance against variance of other algorithms, since Theorem 3.4 and Proposition 3.6 show that the variance (given by the combinatorial sum) is only lower than Broder's MinHash estimate, and says nothing about other MinHash algorithms. The full statements of relevant lemmas, theorems, and propositions in [15] can be found in Appendix A.2.

Hence, researchers using a circulant permutation to improve other MinHash estimators would need to independently show that using a circulant permutation theoretically gives a lower variance compared to different MinHash algorithms, *which is non-trivial if following similar approaches as in the original paper*. Secondly, researchers may also want to compare the performance of two algorithms both using a circulant permutation framework. This is difficult to do theoretically, as the circulant permutation variance is given as a combinatorial sum under constraints if using the approach in [15].

While simulations can be run to compare the empirical performances of a proposed circulant permutation estimator against baselines and sidestep theory, there is no guarantee that a proposed estimator always has reduced variance in all simulated cases, since estimators can have different performance when datasets vary (e.g. sparsity of dataset). This leads us to give the following contributions.

Contribution One: We give a new proof of Lemma 2.4 in [15], using a different paradigm in Section 2. This allows **simpler computation of expectations under the permutation** σ. This approach can be applied to other MinHash estimators, and we give an example in Section 3.

Contribution Two: We derive a simpler expression for the variance in Theorem 3.1 in [15] under the assumption that $D \to \infty$, which is common (e.g. see [9, 12, 19]). Our result does not require the equality discussed in Appendix A.2, and is slightly stronger in this respect. This allows **easier analysis and comparison of the variance under the permutations** σ **and** π. More specifically, this gives a **simpler variance expression which is easier to analyze (and compare)** with a remainder of order f/D^2 when $f/D \to 0$, and of order D^{-2} when f/D is bounded away from zero, which we explain in Section 2. We remark that this gives approximately accurate variances even for small $D = 128$. This approach can also be applied to other MinHash estimators to get variance expressions for the corresponding circulant permutation estimator, with an example in Section 3.

Contribution Three: We show how our approach can be used to **analyze MinHash estimators based on circulant permutation**, which we demonstrate on an asymmetric hashing setup with the Minner estimator. We pick the Minner estimator for two reasons: a) to show how to do variance analysis when the sets $\mathcal{F}_+$ and $\mathcal{F}_-$ are different, and b) how to extend the circulant permutation setup with counts, i.e. finding the expected number of elements seen in $\mathcal{F}_-$ before an element in $\mathcal{F}_+$.

Overall, our paper **provides a direct way to compute expectations and variance under circulant permutation, making it easier to analyze and extend MinHash estimators.**

Section 2 (and A.1-A.2) give a broader overview of CMinHash. Section 3 applies our approach to the Minner estimator involving counts of events. Section 4 discusses the importance of our results.

2 CMinHash in Detail

The sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \{1, 2, \ldots, D\}$ described on Page 1 can be visualized as D objects in sequential order (see Figure 1). Let $a = \#$ of $\begin{smallmatrix}\bullet\\\bullet\end{smallmatrix}$, $f - a$ be the $\#$ of $\begin{smallmatrix}\circ\\\bullet\end{smallmatrix}$ and $\begin{smallmatrix}\bullet\\\circ\end{smallmatrix}$, and $D - f$ be the $\#$ of $\begin{smallmatrix}\circ\\\circ\end{smallmatrix}$.

Definition 1. *Definition 2.1 in [15] (modified). Define the location vector $\vec{x}$, where $x_i \in \{\widetilde{O}, \widetilde{\times}, \widetilde{-}\}$. x_i corresponds to the type of the ith object, which is a $\widetilde{O}$ if $\begin{smallmatrix}\bullet\\\bullet\end{smallmatrix}$, $\widetilde{\times}$ if $\begin{smallmatrix}\circ\\\bullet\end{smallmatrix}$ or $\begin{smallmatrix}\bullet\\\circ\end{smallmatrix}$, and $\widetilde{-}$ if $\begin{smallmatrix}\circ\\\circ\end{smallmatrix}$.*

Let $\theta^{(\mathrm{CMH})} := \sum_{t=1}^{K} \mathbb{1}_{\{\min\{\pi_{\to t}(\mathcal{S}_1)\}=\min\{\pi_{\to t}(\mathcal{S}_2)\}\}}/K$ which is the CMinHash estimator. We write $\mathbb{1}_t \equiv \mathbb{1}_{\{\min\{\pi_{\to t}(\mathcal{S}_1)\}=\min\{\pi_{\to t}(\mathcal{S}_2)\}\}}$ for compactness. $\mathbb{E}\left(\theta^{(\mathrm{CMH})}\right)$ is J due to linearity of expectations (equal probability of J regardless of starting point after applying π), but computing $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right)$ requires knowing $\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_t\right)$, $1 \le s, t \le K$, as $\mathbb{1}_s, \mathbb{1}_t$ are correlated.

Definition 2. *Definition 2.2 in [15]. For $A, B \in \{\widetilde{O}, \widetilde{\times}, \widetilde{-}\}$, let $\{(A, B) \mid \Delta\}$ be the set $\{(i, j) : (x_i, x_j) = (A, B), j - i = \Delta\}$. For $1 \le \Delta \le K - 1$, define*

$$\begin{aligned}
\mathcal{L}_0(\Delta) &= \{(\widetilde{O}, \widetilde{O})|\Delta\}, \quad \mathcal{L}_1(\Delta) = \{(\widetilde{O}, \widetilde{\times})|\Delta\}, \quad \mathcal{L}_2(\Delta) = \{(\widetilde{O}, \widetilde{-})|\Delta\},\\
\mathcal{G}_0(\Delta) &= \{(\widetilde{-}, \widetilde{O})|\Delta\}, \quad \mathcal{G}_1(\Delta) = \{(\widetilde{-}, \widetilde{\times})|\Delta\}, \quad \mathcal{G}_2(\Delta) = \{(\widetilde{-}, \widetilde{-})|\Delta\},\\
\mathcal{H}_0(\Delta) &= \{(\widetilde{\times}, \widetilde{O})|\Delta\}, \quad \mathcal{H}_1(\Delta) = \{(\widetilde{\times}, \widetilde{\times})|\Delta\}, \quad \mathcal{H}_2(\Delta) = \{(\widetilde{\times}, \widetilde{-})|\Delta\},
\end{aligned}$$

with the following constraints for all $1 \le \Delta \le K - 1$,

$$\begin{aligned}
|\mathcal{L}_0| + |\mathcal{L}_1| + |\mathcal{L}_2| &= |\mathcal{L}_0| + |\mathcal{G}_0| + |\mathcal{H}_0| = a,\\
|\mathcal{G}_0| + |\mathcal{G}_1| + |\mathcal{G}_2| &= |\mathcal{L}_2| + |\mathcal{G}_2| + |\mathcal{H}_2| = D - f,\\
|\mathcal{H}_0| + |\mathcal{H}_1| + |\mathcal{H}_2| &= |\mathcal{L}_1| + |\mathcal{G}_1| + |\mathcal{H}_1| = f - a.
\end{aligned}$$

The nine tuples (A, B) in Definition 2 correspond to symbols at $x_i, x_{i+\Delta}$ in the location vector $\vec{x}$.

Lemma 1. *Lemma 2.4 in [15]. For any $1 \le s < t \le K$, with $t - s = \Delta$, then $\mathbb{E}_\pi\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = \frac{|\mathcal{L}_0(\Delta)| + (|\mathcal{G}_0(\Delta)| + |\mathcal{L}_2(\Delta)|)J}{f + |\mathcal{G}_0(\Delta)| + |\mathcal{G}_1(\Delta)|}$.*

We give a simpler proof in Appendix A.3 with a pictorial intuition in Table 4, and explain where our approach differs to [15]. Following the sampling without replacement paradigm where a random permutation can be thought of as the *draw order*, then for a fixed Δ, then applying π means we are drawing the indices $(x_s, x_{s+\Delta})$ from the location vector $\vec{x}$ in some random order, and we know how many of each $(A, B) = (x_s, x_{s+\Delta})$ indices there are using Definition 2. The first entry in the tuple (A, B) then corresponds to what is seen when π is shifted s times to the right, and the second entry corresponds to what is seen when π is shifted $s + \Delta$ times to the right. Since a $\widetilde{O}$ in the location vector means that the hashed values collide, then $\mathbb{E}_\pi\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) > 0$ can be thought of as the **probability of a sequence of draws** of the nine different types of D tuples where the first $\widetilde{O}$ appears before the first $\widetilde{\times}$ for both entries of the tuple, as the random permutation π is a sequence of draws without replacement. We call such a sequence of draws **a success**, with two examples below.

- $(\widetilde{-}, \widetilde{-}), (\widetilde{-}, \widetilde{\mathbf{O}}), (\widetilde{\mathbf{O}}, \widetilde{-})$
- $(\widetilde{-}, \widetilde{-}), (\widetilde{-}, \widetilde{\mathbf{O}}), (\widetilde{-}, \widetilde{\times}), (\widetilde{\mathbf{O}}, \widetilde{O})$. The first $\widetilde{O}$ for the second entry of the tuples appeared before the first $\widetilde{\times}$ for the second entry of the tuples.

More examples are given by Table 2 and Table 3 in Appendix A.3. Hence, we just need to compute the probabilities of seeing the tuple we want before other invalid tuples, which is much simpler. The intuition behind this simplification can be seen behind this following analogy: Consider a bag with one blue marble, two red marbles, and N green marbles. The probability of drawing a blue marble before a red marble is always $\frac{1}{3}$, regardless of N. We can do the same with $\widetilde{O}$ and $\widetilde{\times}$, regardless of $\widetilde{-}$.

Next, we compute the variance of the CMinHash(σ, π) estimator under K shifts adopting a different approach, which we outline here. Suppose D is large. Under the random permutation σ, both elements in the tuple $(x_s, x_{s+\Delta})$ can be thought of drawing entries $x_s, x_{s+\Delta}$ without replacement

from a population of size D with three different objects $\widetilde{O}, \widetilde{\times}, \widetilde{-}$. More specifically, each set $\{(A,B) \mid \Delta\}, 1 \le \Delta \le K-1$ can be expressed as

$$||\{(A,B) \mid \Delta\}| = \sum_{s=1}^{D} \mathbb{1}_{\{(x_s, x_{s+\Delta})=(A,B)\}}. \tag{1}$$

Therefore, we can take expectations across (functions of) (1). For example,

$$\mathbb{E}_{\sigma,\pi}\left(|\mathcal{L}_0(\Delta)|\right) = \sum_{s=1}^{D} \mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_{\left\{(x_s, x_{s+\Delta})=(\widetilde{O},\widetilde{O})\right\}}\right) = \frac{a(a-1)}{D-1}.$$

For the sparse case, we first condition on the positions of the f entries which are not $\widetilde{-}$. This gives an exact expression whose first term can be computed by considering when two such entries are next to each other. For the non sparse case, we use the delta method on the same exact expression. The details are in Appendix A.4.

Lemma 2. *Suppose we pre-process the location vector with the random permutation σ. Let $B_D = 2D - f - 1$, and suppose $0 < J < 1$. For all $1 \le \Delta \le K-1$, if $f/D \to 0$, then*

$$\mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = \begin{cases} J^2 - \dfrac{J(1-J)f}{2(f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & D \text{ even and } \Delta = D/2, \\ J^2 - \dfrac{J(1-J)f}{(2f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & \text{otherwise}, \end{cases}$$

$$\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_{s+\Delta}\right) = \begin{cases} -\dfrac{J(1-J)f}{2(f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & D \text{ even and } \Delta = D/2, \\ -\dfrac{J(1-J)f}{(2f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & \text{otherwise}. \end{cases}$$

If f/D is bounded away from zero, then

$$\mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = J^2 - \frac{J(1-J)}{B_D} + O\left(D^{-2}\right),$$
$$\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_{s+\Delta}\right) = -\frac{J(1-J)}{B_D} + O\left(D^{-2}\right).$$

We prove this in Appendix A.4.

Definition 3. *For $0 < K \le D$, let $\lambda_{D,K} = \mathbb{1}_{\{D \text{ even}\}} \frac{(2K-D)_+}{K}$, where $(x)_+ := \max\{x, 0\}$.*

Definition 3 accounts for the proportion of terms in the covariance where $\Delta = D/2$.

Theorem 2. *Let $0 < a < f \le D$. Suppose we pre-process the location vector with the random permutation σ, and generate an independent permutation π. Let $B_D = 2D - f - 1$. If $f/D \to 0$,*

$$\mathrm{Var}_{\sigma,\pi}\left(\theta^{(\mathrm{CMH})}\right) = \frac{J(1-J)}{K}\left[1 - \frac{f}{D-1}\left\{\frac{K-1-\lambda_{D,K}}{2f-1} + \frac{\lambda_{D,K}}{2(f-1)}\right\}\right] + O\left(\frac{f}{D^2}\right).$$

If f/D is bounded away from zero, then

$$\mathrm{Var}_{\sigma,\pi}\left(\theta^{(\mathrm{CMH})}\right) = \frac{J(1-J)}{K}\left(1 - \frac{K-1}{B_D}\right) + O\left(D^{-2}\right).$$

Proof. We give the proof in Appendix A.4. □

The proof of Lemma 2 also gives an exact negative expression for every covariance. Hence, for $K > 1$, the variance under CMinHash(σ, π) is smaller than the variance $J(1-J)/K$ under K independent permutations. If f/D is bounded away from zero, the first-order variance reduction can be thought of as approximately scaled by $(1 - (K-1)/(2D-f-1))$. The leading variance ratio relative to MinHash is independent of a when D, f, K are fixed, which also verifies the Uniform Superiority (Theorem 3.4) and Consistent Improvement (Proposition 3.6) statements in [15].

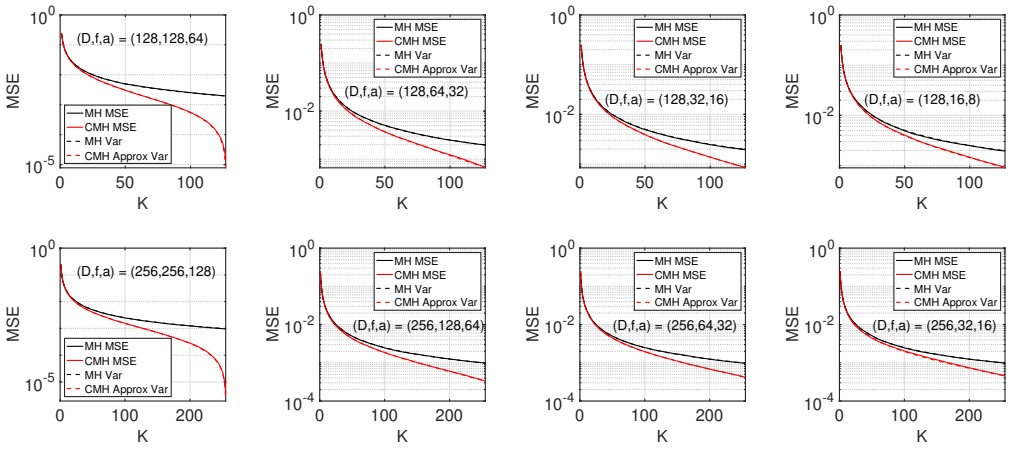

Figure 3: Empirical MSE vs theoretical approximate variance in Theorem 2 of MinHash and CMinHash (σ, π). The dotted lines overlap almost perfectly (up to Monte Carlo error), which shows our theoretical approximate variance matches the empirical variance.

We first construct a similar experiment to [15], generating synthetic vectors with varying D, f, a in Figure 3, averaging over 10000 iterations, and overlaying the theoretical approximation with the empirical mean square error (MSE) of the estimates. The dotted lines overlap almost perfectly with the solid lines at $D = 128$ and $D = 256$.

We do more experiments in Appendix A.5 (Figure 6) to investigate if the approximation is accurate in edge cases when $f/D \to 0$. At $D = 128$, for $f = 2, 4$, the variance approximation is slightly higher than the empirical MSE. But at $D = 1024$ for $f = 2, 4$, both lines overlap each other almost perfectly. We do more experiments on the Words dataset [7] (Figure 7 in Appendix A.5), where the theoretical variance approximation and empirical MSE lines also overlap each other almost perfectly.

Hence, we believe our approximation of the theoretical variance is practical enough to do comparisons.

We argue that our approach in Lemma 2 should be how MinHash algorithms are analyzed, especially with large D, *as large D necessitated the need of MinHash in the first place*.

3 Applying the Circulant Permutation Framework To Asymmetric Hashing

A natural question to ask is if our work is generalizable to other MinHash algorithms. We show our approach can be applied to the Minner estimators [1] (asymmetric hashing), and theoretically demonstrate a circulant permutation framework with reduced variance.

The asymmetric hashing setup considers a baseline set $\mathcal{S}_1$ relative to $\mathcal{S}_2, \ldots, \mathcal{S}_n$, and estimates the set overlap a_{1i} (which can then be used to estimate the Jaccard similarity J_{1i}, since $a_{1i} = \frac{(f_1+f_i)J_{1i}}{1+J_{1i}}$), under the assumption that the set cardinalities $f_1, \ldots, f_n$ are all known. We give more details of the Minner estimators in Appendix B.1.

Draw Order	1	2	3	4	5	6	7	8
Set 1	○	●	○	●	●	●	○	●
Set 2	○	○	○	○	○	○	○	●

Figure 4: We have $\min\{\pi_t(\mathcal{S}_2)\} = 8$, with $y_t = 1$, which can be thought of as the presence (or absence) of a • at draw 8 for $\mathcal{S}_1$, and $m_t = 4$, which can be thought of as counting how many • were seen in previous draws 1-7 for $\mathcal{S}_1$.

Theorem 3. *Minner estimators (heavily modified from [1]) Suppose we have K independent random permutations $\pi_1, \ldots, \pi_K$ on $[D]$, and we want to estimate the set overlap a between $\mathcal{S}_1$ and $\mathcal{S}_2$, where*

$\mathcal{S}_1$ is the baseline set. Let $y_t = \mathbb{1}_{\{\min(\pi_t(\mathcal{S}_2))\in\pi_t(\mathcal{S}_1)\}}$, and let m_t be the count of elements in $\pi_t(\mathcal{S}_1)$ strictly less than $\min(\pi_t(\mathcal{S}_2))$. Then the two estimators $f_2\sum_{t=1}^{K}\frac{y_t}{K}$ and $f_1-(f_2+1)\sum_{t=1}^{K}\frac{m_t}{K}$ both have expectation a. Their variances are, respectively, $\frac{a(f_2-a)}{K}$ and $\frac{f_2(f_1-a)(f_1+f_2-a+1)}{K(f_2+2)}$.

We remark that thinking of the random permutation as the draw order in sampling without replacement (see Figure 4) makes understanding the Minner estimators straightforward.

3.1 Applying the Circulant Permutation framework to a different setup of the location vector

Without loss of generality, consider estimating the set overlap between $\mathcal{S}_1$ and $\mathcal{S}_2$, where we look at the random variable y_t. The question we aim to resolve is: *How can the circulant permutation framework be applied when we are not recording both* $\min\{\pi_{\to t}(\mathcal{S}_i)\}$?

In Section 2, we introduced $\mathcal{F}_+$ and $\mathcal{F}_-$, where we find the probability that an element from $\mathcal{F}_+$ is seen before an element from $\mathcal{F}_-$. In the context of y_t, we set $\mathcal{F}_+ = \left\{ \begin{smallmatrix} \bullet \\ \bullet \end{smallmatrix} \right\}$, and $\mathcal{F}_- = \left\{ \begin{smallmatrix} \circ \\ \bullet \end{smallmatrix} \right\}$, as we continue drawing if we see a $\begin{smallmatrix} \bullet \\ \circ \end{smallmatrix}$ or $\begin{smallmatrix} \circ \\ \circ \end{smallmatrix}$. Let $C_s = \#$ of elements of $\mathcal{F}_+$ in D (success), $C_f = \#$ of elements of $\mathcal{F}_-$ in D (failure), and $C_n = D - C_s - C_f$ (neutral). There are $C_s = a$ $\widetilde{O}$s, $C_f = f_2 - a$ $\widetilde{\times}$s, and $C_n = D - f_2$ $\widetilde{-}$s. To generalize this, the constraints of the sets are now

$$\begin{aligned} |\mathcal{L}_0| + |\mathcal{L}_1| + |\mathcal{L}_2| &= |\mathcal{L}_0| + |\mathcal{G}_0| + |\mathcal{H}_0| = C_s, \\ |\mathcal{G}_0| + |\mathcal{G}_1| + |\mathcal{G}_2| &= |\mathcal{L}_2| + |\mathcal{G}_2| + |\mathcal{H}_2| = C_n, \\ |\mathcal{H}_0| + |\mathcal{H}_1| + |\mathcal{H}_2| &= |\mathcal{L}_1| + |\mathcal{G}_1| + |\mathcal{H}_1| = C_f. \end{aligned}$$

It is clear that under one permutation, the variance of drawing an element from $\mathcal{F}_+$ before $\mathcal{F}_-$ is $\frac{C_sC_f}{(C_s+C_f)^2}$ regardless of the size of C_n. Let V_{ind} be the variance of 1 trial, and V_K the variance of K independent trials (i.e. under K independent permutations). This leads to the following corollary.

Corollary 1. *Suppose we pre-process the location vector with a random permutation σ, and let $\mathbb{1}_s \equiv \mathbb{1}_{\{\text{Element in } \mathcal{F}_+ \text{ is seen before element in } \mathcal{F}_- \text{ at } \pi_{\to s}\}}$. Then for any $1 \le s < t \le K$, with $t - s = \Delta$,*

$$\mathbb{E}_\pi\left(\mathbb{1}_s\mathbb{1}_{s+\Delta}\right) = \frac{|\mathcal{L}_0(\Delta)| + (|\mathcal{G}_0(\Delta)| + |\mathcal{L}_2(\Delta)|)\frac{C_s}{C_s+C_f}}{C_s + C_f + |\mathcal{G}_0(\Delta)| + |\mathcal{G}_1(\Delta)|}.$$

Let $\lambda_{D,K}$ be defined as in Definition 3, and let $B_D = D + C_n - 1$. If $(C_s + C_f)/D \to 0$, then

$$\text{Var}_{\sigma,\pi}\left(\sum_{t=1}^{K}\frac{\mathbb{1}_t}{K}\right) = V_K\left[1 - \frac{C_s+C_f}{D-1}\left\{\frac{K-1-\lambda_{D,K}}{2(C_s+C_f)-1} + \frac{\lambda_{D,K}}{2(C_s+C_f-1)}\right\}\right] + O\left(\frac{C_s+C_f}{D^2}\right).$$

If $(C_s + C_f)/D$ is bounded away from zero, then

$$\text{Var}_{\sigma,\pi}\left(\sum_{t=1}^{K}\frac{\mathbb{1}_t}{K}\right) = V_K\left(1 - \frac{K-1}{B_D}\right) + O(D^{-2}).$$

Proof. We substitute $a = C_s$ and $f = C_s + C_f$ in Lemma 1, Lemma 2, and Theorem 2. Since $D - f = C_n$, the quantity $2D - f - 1$ becomes $D + C_n - 1 = B_D$. □

By the exact covariance identity (16), the variance under $K > 1$ circulant shifts is always reduced compared to K independent permutations. The size of the reduction depends on $(C_s + C_f)$. In particular, MinHash algorithms where $\mathcal{F}_+$ and $\mathcal{F}_-$ are small relative to D have a small variance reduction when CMinHash(σ, π) is applied.

With respect to the Minner estimator y_t, we can apply Corollary 1 and immediately get the theoretical approximation of the variance of y_t by letting $C_s = a$, $C_f = f_2 - a$.

3.2 Applying the Circulant Permutation framework to counts

Now consider estimating the set overlap between $\mathcal{S}_1$ and $\mathcal{S}_2$, where we look at the random variable m_t. The question we aim to resolve is: *How can the circulant permutation framework be applied when we are recording* **counts** *of elements in $\mathcal{F}_-$ before the first element in $\mathcal{F}_+$ is seen?*

We still construct the two sets $\mathcal{F}_+$, $\mathcal{F}_-$, but give a different meaning for elements in $\mathcal{F}_-$. Let $\mathcal{F}_+ = \left\{ \begin{smallmatrix} \bullet \\ \bullet \end{smallmatrix}, \begin{smallmatrix} \circ \\ \bullet \end{smallmatrix} \right\}$, and $\mathcal{F}_- = \left\{ \begin{smallmatrix} \bullet \\ \circ \end{smallmatrix} \right\}$ with $C_s = f_2$ and $C_f = f_1 - a$. Then the setup becomes: "count how many elements in $\mathcal{F}_-$ are seen before the first element in $\mathcal{F}_+$". Equivalently this would be "count how many $\widetilde{\times}$ are seen before the first $\widetilde{O}$ is drawn" for the location vector $\vec{x}$.

From Lemma 5 in Appendix B.2, the variance of observing θ elements from $\mathcal{F}_-$ before the first $\mathcal{F}_+$ is $\frac{C_s C_f (C_s + C_f + 1)}{(C_s+1)^2(C_s+2)}$. Let V_{ind} be the variance of 1 trial, and V_K the variance of K independent trials.

Lemma 3. *Suppose we pre-process the location vector with a random permutation σ. Let the random variable $\theta_s \equiv$ number of $\widetilde{\times}$ seen before the first $\widetilde{O}$ under the random permutation $\pi_{\to s}$. Let $B_\Delta = 2C_s - |\mathcal{L}_0| + 1$. Then for any $1 \le s < t \le K$, with $t - s = \Delta$,*

$$\mathbb{E}_\pi\left(\theta_s \theta_{s+\Delta}\right) = \frac{|\mathcal{H}_1|}{B_\Delta} + \frac{2C_f^2 - C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2|\mathcal{H}_1|}{(C_s+1)(B_\Delta+1)} + \frac{C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2|\mathcal{L}_1||\mathcal{H}_0|}{(C_s+1)B_\Delta(B_\Delta+1)}.$$

If C_s and C_f are fixed as $D \to \infty$, then, when D is even and $\Delta = D/2$,

$$\text{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = -V_{\text{ind}} \frac{(C_s+1)(C_s+2)}{2C_s(2C_s+1)} \frac{2C_s + C_f}{(C_s + C_f + 1)(D-1)} + O(D^{-2}),$$

and otherwise,

$$\text{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = -V_{\text{ind}} \frac{(C_s+2)(C_s+C_f)}{(2C_s+1)(C_s+C_f+1)(D-1)} + O(D^{-2}).$$

If both C_s/D and C_f/D are bounded away from zero, then

$$\text{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = \begin{cases} -V_{\text{ind}} \left(\frac{C_s(2D-C_s)+C_f(D-C_s)}{(C_s+C_f)(2D-C_s)^2}\right) + O\left(\frac{1}{D^2}\right), & D \text{ even and } \Delta = D/2, \\ -V_{\text{ind}} \left(\frac{1}{2D-C_s}\right) + O\left(\frac{1}{D^2}\right), & \text{otherwise.} \end{cases}$$

Theorem 4. *Suppose we pre-process the location vector with a random permutation σ. Let the random variable $\theta_t \equiv$ number of $\widetilde{\times}$ seen before the first $\widetilde{O}$ under the random permutation $\pi_{\to t}$. Let $K > 1$, and let $\lambda_{D,K}$ be defined as in Definition 3. If C_s and C_f are fixed as $D \to \infty$, then*

$$\text{Var}_{\sigma,\pi}\left(\sum_{t=1}^{K} \frac{\theta_t}{K}\right) = V_K \left[1 - \frac{1}{D-1}\left\{(K - 1 - \lambda_{D,K}) \frac{(C_s+2)(C_s+C_f)}{(2C_s+1)(C_s+C_f+1)} + \lambda_{D,K} \frac{(C_s+1)(C_s+2)}{2C_s(2C_s+1)} \frac{2C_s+C_f}{C_s+C_f+1}\right\}\right] + O(D^{-2}).$$

If both C_s/D and C_f/D are bounded away from zero, then

$$\text{Var}_{\sigma,\pi}\left(\sum_{t=1}^{K} \frac{\theta_t}{K}\right) = V_K \left[1 - \frac{K-1}{2D-C_s}\left\{1 - \frac{\lambda_{D,K} C_f D}{(K-1)(C_s+C_f)(2D-C_s)}\right\}\right] + O(D^{-2}).$$

Proof. The proof is similar to Theorem 2, where we add up the covariance terms. □

Each term within the square brackets in Theorem 4 is < 1. Hence the circulant permutation estimate gives a lower variance asymptotically. Substituting $C_s = f_2$ and $C_f = f_1 - a$ gives the variance approximation of m_t.

To verify the approximations, we run experiments on synthetic data at $D = 128$ with varying f_1, f_2, a values over 10000 iterations, where we similarly overlay the approximate theoretical variance when

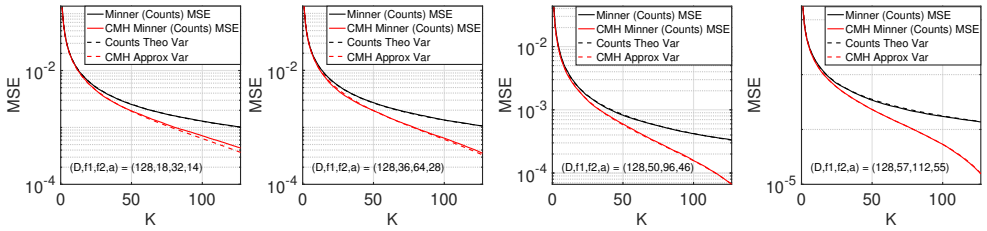

Figure 5: Empirical MSE vs theoretical approximate variance in Theorem 4 for count data. The dotted lines overlap almost perfectly for the last three plots.

$\frac{C_s+C_f}{D} = \frac{f_1+f_2-a}{D}$ do not go to zero with the empirical MSEs in Figure 5. Unlike Figure 3, where the approximate theoretical variance overlaps with the empirical MSE, we see that at $f_1 = 18, f_2 = 32$, we see the theoretical variance curves do not overlap the empirical MSEs. However, how good is our variance approximation for $\frac{C_s+C_f}{D} \to 0$ if D is large, and f_1, f_2 are small?

In Appendix B.3, we run another experiment averaged over 10000 iterations where we fix $(f_1, f_2, a) = (6, 24, 6)$, and increase this tuple up to $(f_1, f_2, a) = (96, 128, 132)$, with $D \in \{512, 1024, 2048, 4096\}$. In general, as D increases, both our approximation variance ratios match the empirical MSEs. Finally, we do experiments with the Words dataset as in Section 2 (Figure 14 in Appendix B). For the word pairs we choose, both theoretical approximate variance curves overlap the empirical MSE curves perfectly, even for the `HONG-KONG` word pair, where $\frac{f_1-a}{D} = \frac{41}{65536}$ is sparse.

4 Discussion and Conclusion

Each observation θ_s recorded by the circulant permutation estimator is correlated with another observation θ_t, due to the circulant shift. Thus in order to theoretically compute (or compare) the variance of such an estimate, the corresponding covariance terms $\mathrm{Cov}_\pi(\theta_s, \theta_t) = \mathbb{E}_\pi(\theta_s\theta_t) - \mathbb{E}_\pi(\theta_s)\,\mathbb{E}_\pi(\theta_t)$ need to be computed. Computing $\mathbb{E}_\pi(\theta_s\theta_t)$ is the first hurdle in circulant permutation algorithms, yet is needed to compute the variances for such estimators.

But by viewing MinHash under the paradigm of the draw order under sampling without replacement, we simplify the proof to compute $\mathbb{E}_\pi(\theta_s\theta_t)$ in [15] via our approach in Lemma 1. Moreover, by using indicator functions and considering exchangeability, we can further compute $\mathbb{E}_\pi(\theta_s\theta_t)$ for counts as well (our Lemma 3), showing our approach applies beyond the original CMinHash setting.

The second hurdle with circulant permutations is finding the variance of the corresponding estimate. While Theorem 3.1 in [15] states a combinatorial expression for the variance, one equality in its proof does not hold in general (see Appendix A.2). Even apart from this point, the expression is not easily tractable, nor easy to analyze or compare against other MinHash algorithms, e.g. [15] had Theorem 3.4, Proposition 3.6, and many, many experiments in their paper just to compare CMinHash with MinHash (without even comparing against other MinHash algorithms). Researchers building on the circulant permutation framework would face similar technical difficulties when comparing variances.

Our approach suggests that exact accuracy is not needed, particularly because MinHash algorithms are used when D is large in the first place. Lemma 2 shows that by letting D be large, we can do a straightforward computation to compute approximations of the variance. We demonstrate how to compute an approximation when set sizes are sparse relative to D, and when they are non sparse. Both approximations are accurate for practical purposes, allowing theoretical comparison of variances with other MinHash algorithms in the large D scenario.

Furthermore, our approach is not limited to CMinHash and applies to other MinHash estimators based on circulant permutation. The location vector setup of $\widetilde{O}, \widetilde{\times}, \widetilde{-}$ can be characterized as: "**stop** at the sth $\widetilde{O}$, *do something* at a $\widetilde{\times}$, and continue at a $\widetilde{-}$". Our proof approach in Lemma 1 can be modified to a "**stop** at the sth $\widetilde{O}$" scenario, and our approach with indicator random variables in Lemma 3 can account for a wide range of *do somethings*, by *conditioning on the element* $\widetilde{\times}$ *itself*, providing a direct way to compute expectations and variance for such estimators.

Thus, we believe our work makes circulant permutation estimators easier to analyze and extend. While state of the art algorithms are important, it is equally important to be able to understand and extend these algorithms. We provide a clear method to compute expectations and variance, and our proof techniques can be used directly to extend circulant permutations for other MinHash algorithms.

Limitations: Generally, D needs to be large for accurate approximations. Our experiments show that even $D = 65536$ (words data) gives good approximations. In practice, D is much larger (e.g. 2^{32}).

Acknowledgments and Disclosure of Funding

We would like to thank Bucknell's Presidential Fellow program, which paid a stipend to Avery Hood working on this paper.

References

[1] Thomas Ahle. Minner: Improved similarity estimation and recall on MinHashed databases, 2021. URL `https://thomasahle.com/papers/minner.pdf`. Manuscript.

[2] Thomas Dybdahl Ahle and Jakob Bæk Tejs Knudsen. Subsets and supermajorities: Optimal hashing-based set similarity search. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 728–739. IEEE, 2020. doi: 10.1109/FOCS46700.2020.00073.

[3] Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Conference on Compression and Complexity of Sequences*, pages 21–29. IEEE, 1997. doi: 10.1109/SEQUEN.1997.666900.

[4] Edith Cohen and Haim Kaplan. Summarizing data using bottom-k sketches. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing*, pages 225–234. Association for Computing Machinery, 2007. doi: 10.1145/1281100.1281133.

[5] Jianqiu Ji, Jianmin Li, Shuicheng Yan, Qi Tian, and Bo Zhang. Min-max hash for Jaccard similarity. In *2013 IEEE 13th International Conference on Data Mining*, pages 301–309. IEEE, 2013. doi: 10.1109/ICDM.2013.119.

[6] Keegan Kang, Sergey Kushnarev, Wei Pin Wong, Rameshwar Pratap, Haikal Yeo, and Chen Yijia. Improving hashing algorithms for similarity search via MLE and the control variates trick. In *Proceedings of the 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 814–829. PMLR, 2021. URL `https://proceedings.mlr.press/v157/kang21a.html`.

[7] Ping Li and Kenneth W. Church. Using sketches to estimate associations. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 708–715, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://aclanthology.org/H05-1089/`.

[8] Ping Li and Arnd Christian König. b-bit minwise hashing. In *Proceedings of the 19th International Conference on World Wide Web*, pages 671–680. Association for Computing Machinery, 2010. doi: 10.1145/1772690.1772759.

[9] Ping Li and Arnd Christian König. Theory and applications of b-bit minwise hashing. *Communications of the ACM*, 54(8):101–109, 2011. doi: 10.1145/1978542.1978566.

[10] Ping Li and Christian König. Accurate estimators for improving minwise hashing and b-bit minwise hashing. arXiv:1108.0895, Aug. 2011. Submitted 3 Aug. 2011. DOI: 10.48550/arXiv.1108.0895.

[11] Ping Li, Kenneth W. Church, and Trevor J. Hastie. Conditional random sampling: A sketch-based sampling technique for sparse data. In *Advances in Neural Information Processing Systems 19*, pages 873–880, 2006. URL `https://proceedings.neurips.cc/paper/2006/hash/aa6b7ad9d68bf3443c35d23de844463b-Abstract.html`.

[12] Ping Li, Art B. Owen, and Cun-Hui Zhang. One permutation hashing. In *Advances in Neural Information Processing Systems 25*, pages 3122–3130, 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/hash/eaa32c96f620053cf442ad32258076b9-Abstract.html`.

[13] Ping Li, Xiaoyun Li, and Gennady Samorodnitsky. $\mathcal{P}$-MinHash algorithm for continuous probability measures: Theory and application to machine learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1177–1187. Association for Computing Machinery, 2022. doi: 10.1145/3511808.3557413.

[14] Xiaoyun Li and Ping Li. C-OPH: Improving the accuracy of one permutation hashing (OPH) with circulant permutations. *arXiv preprint arXiv:2111.09544*, 2021. URL `https://arxiv.org/abs/2111.09544`.

[15] Xiaoyun Li and Ping Li. C-MinHash: Improving minwise hashing with circulant permutation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12857–12887. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22m.html.

[16] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, 2016. doi: 10.1186/s13059-016-0997-x.

[17] Mahmudur Rahman Hera and David Koslicki. Estimating similarity and distance using FracMinHash. *Algorithms for Molecular Biology*, 20(1):8, 2025. doi: 10.1186/s13015-025-00276-8.

[18] Anshumali Shrivastava. Optimal densification for fast and accurate minwise hashing. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3154–3163. PMLR, 2017. URL https://proceedings.mlr.press/v70/shrivastava17a.html.

[19] Anshumali Shrivastava and Ping Li. Densifying one permutation hashing via rotation for fast near neighbor search. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 557–565. PMLR, 2014. URL https://proceedings.mlr.press/v32/shrivastava14.html.

[20] Noah Thomas, Eric Yang, Benedict H. W. Wong, and Keegan Kang. A second look at Broder's MinHash without alignment. *Journal of Computer Science and Technology*, forthcoming. doi: 10.1007/s11390-026-6603-0. URL https://jcst.ict.ac.cn/article/doi/10.1007/s11390-026-6603-0. Accepted manuscript.

[21] Tom Tseng, Laxman Dhulipala, and Julian Shun. Parallel index-based structural graph clustering and its approximation. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1851–1864. Association for Computing Machinery, 2021. doi: 10.1145/3448016.3457278.

A Appendix: CMinHash and Alternative Proofs

A.1 The CMinHash(σ, π) algorithm

CMinHash(σ, π) is simple to implement, as it just requires a small tweak from Broder's original MinHash. We place Algorithm 1 and Algorithm 2 side by side for comparison. The pre-processing stage for CMinHash (and by extension, algorithms using circulant permutation) consists of lines 1-4 in Algorithm 2, which replaces the independent permutation within the loop (e.g. Line 6 in Algorithm 1).

Algorithm 1: Broder's MinHash

Input: $\mathcal{S}_1, \mathcal{S}_2 \subseteq \{1, 2, \ldots, D\}$
$\theta^{(\mathrm{MH})} = 0$
Output: $\theta^{(\mathrm{MH})}$

1
2
3
4
5 **for** $t = 1$ **to** K **do**
6 Generate permutation $\pi_t : [D] \to [D]$
7 $\theta_t \leftarrow \mathbb{1}_{\{\min\{\pi_t(\mathcal{S}_1)\} = \min\{\pi_t(\mathcal{S}_2)\}\}}$
8 $\theta^{(\mathrm{MH})} \mathrel{+}= \theta_t$
9 $\theta^{(\mathrm{MH})} \leftarrow \theta^{(\mathrm{MH})} / K$

Algorithm 2: CMinHash(σ, π)

Input: $\mathcal{S}_1, \mathcal{S}_2 \subseteq \{1, 2, \ldots, D\}$
$\theta^{(\mathrm{CMH})} = 0$
Output: $\theta^{(\mathrm{CMH})}$

1 Generate permutation $\sigma : [D] \to [D]$
2 $\mathcal{S}_1 \leftarrow \sigma(\mathcal{S}_1)$
3 $\mathcal{S}_2 \leftarrow \sigma(\mathcal{S}_2)$
4 Generate permutation $\pi : [D] \to [D]$
5 **for** $t = 1$ **to** K **do**
6
7 $\theta_t \leftarrow \mathbb{1}_{\{\min\{\pi_{\to t}(\mathcal{S}_1)\} = \min\{\pi_{\to t}(\mathcal{S}_2)\}\}}$
8 $\theta^{(\mathrm{CMH})} \mathrel{+}= \theta_t$
9 $\theta^{(\mathrm{CMH})} \leftarrow \theta^{(\mathrm{CMH})} / K$

As both $\mathbb{E}\left(\theta^{(\mathrm{MH})}\right)$ and $\mathbb{E}\left(\theta^{(\mathrm{CMH})}\right)$ are unbiased estimators of J due to linearity of expectation, the main result of [15] is to show that $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right) < \mathrm{Var}\left(\theta^{(\mathrm{MH})}\right)$ for $K < D$ (always satisfied, since if $K > D$, computing J directly is more efficient), provided $0 < J < 1$.

This is a surprising and impactful result, since K in Broder's MinHash is the number of independent permutations, while K in CMinHash(σ, π) is the number of circular shifts of **one** permutation. This means that the two independent permutations σ, π can lead to a lower variance compared to K independent permutations, and storing these two independent permutations in memory is cheaper than storing K independent permutations. Conventional thought assumes that because the circulant shift is used for one permutation π, the variance necessarily needs to be higher.

However, we have not seen many papers building on the work of [15] (e.g. using circulant permutations for *other* MinHash type algorithms), and we hypothesize this is due to the technical difficulty of the approaches in the paper, e.g. deriving a variance theorem equivalent to Theorem 3.1 in [15], and then showing that the theoretical variance is indeed lower than other benchmark estimates.

A.2 Lemmas, Theorems and Propositions in the CMinHash(σ, π) paper

We give the original lemmas, theorems and propositions used in the seminal CMinHash paper, since we refer to them in the main text, and only modify them slightly to fit our notation.

We start with the following definition.

Definition 4. *Definition 2.2 in [15]*
For $A, B \in \{\widetilde{O}, \widetilde{\times}, \widetilde{-}\}$*, let* $\{(A, B) \mid \Delta\}$ *be the set* $\{(i, j) : (x_i, x_j) = (A, B), j - i = \Delta\}$*. For* $1 \le \Delta \le K - 1$*, define*

$$\begin{aligned}
\mathcal{L}_0(\Delta) &= \{(\widetilde{O}, \widetilde{O}) | \Delta\}, \quad \mathcal{L}_1(\Delta) = \{(\widetilde{O}, \widetilde{\times}) | \Delta\}, \quad \mathcal{L}_2(\Delta) = \{(\widetilde{O}, \widetilde{-}) | \Delta\}, \\
\mathcal{G}_0(\Delta) &= \{(\widetilde{-}, \widetilde{O}) | \Delta\}, \quad \mathcal{G}_1(\Delta) = \{(\widetilde{-}, \widetilde{\times}) | \Delta\}, \quad \mathcal{G}_2(\Delta) = \{(\widetilde{-}, \widetilde{-}) | \Delta\}, \\
\mathcal{H}_0(\Delta) &= \{(\widetilde{\times}, \widetilde{O}) | \Delta\}, \quad \mathcal{H}_1(\Delta) = \{(\widetilde{\times}, \widetilde{\times}) | \Delta\}, \quad \mathcal{H}_2(\Delta) = \{(\widetilde{\times}, \widetilde{-}) | \Delta\},
\end{aligned}$$

with the following constraints for all $1 \le \Delta \le K - 1$*,*

$$\begin{aligned}
|\mathcal{L}_0| + |\mathcal{L}_1| + |\mathcal{L}_2| = |\mathcal{L}_0| + |\mathcal{G}_0| + |\mathcal{H}_0| = a, \\
|\mathcal{G}_0| + |\mathcal{G}_1| + |\mathcal{G}_2| = |\mathcal{L}_2| + |\mathcal{G}_2| + |\mathcal{H}_2| = D - f, \\
|\mathcal{H}_0| + |\mathcal{H}_1| + |\mathcal{H}_2| = |\mathcal{L}_1| + |\mathcal{G}_1| + |\mathcal{H}_1| = f - a.
\end{aligned}$$

With this definition, [15] gave an expression for $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right)$.

Theorem 5. *Theorem 3.1 in [15] (as stated)*
Suppose $0 < a < f \leq D$ *and* $J \notin \{0, 1\}$. *Then*

$$\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right) = \frac{J}{K} + \frac{(K-1)\widetilde{\mathcal{E}}}{K} - J^2,$$

where $l = \max(0, D - 2f + a)$ *and*

$$\widetilde{\mathcal{E}} = \sum_{\Xi} \left(\frac{l_0}{f + g_0 + g_1} + \frac{a(g_0 + l_2)}{(f + g_0 + g_1)f} \right) \left(\sum_{s=l}^{D-f-1} \frac{\begin{pmatrix} D-f \\ s \end{pmatrix} \begin{pmatrix} f-a-1 \\ D-f-s-1 \end{pmatrix}}{\begin{pmatrix} D-a-1 \\ D-f-1 \end{pmatrix}} \right.$$
$$\left. \times \frac{\begin{pmatrix} s \\ n_1 \end{pmatrix} \begin{pmatrix} D-f-s \\ n_2 \end{pmatrix} \begin{pmatrix} D-f-s \\ n_3 \end{pmatrix} \begin{pmatrix} f-a-(D-f-s) \\ n_4 \end{pmatrix} \begin{pmatrix} a-1 \\ a-l_1-l_2 \end{pmatrix}}{\begin{pmatrix} D-1 \\ a \end{pmatrix}} \right),$$

with the feasible set $\Xi = \{l_0, l_2, g_0, g_1\}$ *satisfies the constraints in Definition 4 and*

$$\begin{aligned} n_1 &= g_0 - (D - f - s - g_1), \\ n_2 &= D - f - s - g_1, \\ n_3 &= l_2 - g_0 + (D - f - s - g_1), \\ n_4 &= l_1 - (D - f - s - g_1). \end{aligned}$$

The proof of this theorem can be found in [15].

Remark: During the rebuttal period of NeurIPS 2026, we wanted to further convince reviewers that our approach for the Minner estimator (Section 3) was still simpler to understand, compared to the combinatorial expression we would get, following Li's approach. We never were able to replicate Li's approach for the Minner estimator, and we thought this was because we did not understand enough combinatorics.

Since ChatGPT 5.6 SOL was released, we asked it to apply Li's approach to the Minner estimator, in order to craft a better rebuttal for reviewers. Unfortunately, SOL found an error in Theorem 3.1, invalidating their approach.

The following italics is output (slightly modified) from SOL, and the authors take full responsibility for it.

In the proof of Theorem 3.1, it is stated that $\mathbb{E}_{\sigma,\pi}(\mathbb{1}_s \mathbb{1}_{s+\Delta})$ *is the same for all* $1 \leq \Delta \leq K-1$ *because* σ *is random. This does not hold in general. For example, take* $D = 4, K = 3, \mathcal{S}_1 = \{1, 2\}, \mathcal{S}_2 = \{1\}$. *Then* $a = 1, f = 2$ *and* $J = 1/2$. *After fixing the position of* $\widetilde{O}$, *the three possible location vectors are* $(\widetilde{O}, \widetilde{\times}, \widetilde{-}, \widetilde{-})$, $(\widetilde{O}, \widetilde{-}, \widetilde{\times}, \widetilde{-})$ *and* $(\widetilde{O}, \widetilde{-}, \widetilde{-}, \widetilde{\times})$. *We compute the expectation and variance by first principles. By Lemma 1,*

$$\begin{aligned} \mathbb{E}_{\sigma,\pi}(\mathbb{1}_1 \mathbb{1}_2) &= \frac{1}{3}\left(\frac{1}{6} + \frac{1}{4} + \frac{1}{6}\right) = \frac{7}{36}, \\ \mathbb{E}_{\sigma,\pi}(\mathbb{1}_1 \mathbb{1}_3) &= \frac{1}{3}\left(\frac{1}{4} + 0 + \frac{1}{4}\right) = \frac{1}{6}, \end{aligned}$$

The variance is hence

$$\begin{aligned} \mathrm{Var}_{\sigma,\pi}(\theta^{(\mathrm{CMH})}) &= \frac{J}{3} + \frac{2}{9}\left(2\mathbb{E}_{\sigma,\pi}(\mathbb{1}_1 \mathbb{1}_2) + \mathbb{E}_{\sigma,\pi}(\mathbb{1}_1 \mathbb{1}_3)\right) - J^2 \\ &= \frac{13}{324}. \end{aligned}$$

On the other hand, applying Theorem 3.1 gives $\mathrm{Var}_{\sigma,\pi}(\theta^{(\mathrm{CMH})}) = 5/108$.

In contrast, the approach in our paper considers every value of Δ. To be precise, the derivation in the proof of Theorem 3.1 uses $\Delta = 1$, but the variance also involves the corresponding expectations for $2 \leq \Delta \leq K-1$. Even apart from this point, the variance is expressed as a combinatorial sum under several constraints, which makes it difficult to compare theoretically against the variance given by other MinHash algorithms. Moreover, the current approach taken to derive the result in Theorem 3.1 would need to be replicated for other MinHash type algorithms using circulant permutation, which is not straightforward.

In fact, in order for [15] to show that $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right) < \mathrm{Var}\left(\theta^{(\mathrm{MH})}\right)$, they state and prove the following lemma and theorem.

Lemma 4. *Lemma 3.3 (Strict Increment) in [15]*
Let $f > a > 0$ and K be arbitrary and fixed. Denote $\widetilde{\mathcal{E}}_D$ as in Theorem 3.1, with D as a parameter. Then $\widetilde{\mathcal{E}}_{D+1} > \widetilde{\mathcal{E}}_D$ for all $D \geq f$.

Theorem 6. *Theorem 3.4 (Uniform Superiority) in [15] (rephrased slightly)*
For any two sets $\mathcal{S}_1, \mathcal{S}_2$, with $J \neq 0$ or $J \neq 1$, and $K > 1$, it holds that $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right) < \mathrm{Var}\left(\theta^{(\mathrm{MH})}\right)$.

Finally, the CMinHash(σ, π) estimator has some nice properties, which [15] gives in the following two propositions.

Proposition 1. *Proposition 3.2 (Symmetry) in [15]*
$\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right)$ *is the same for the (D, f, a) data pair and the $(D, f, f-a)$ data pair, for $0 \leq a \leq f \leq D$.*

Proposition 2. *Proposition 3.6 (Consistent Improvement) in [15]*
Suppose f is fixed. In terms of a, the variance ratio $\frac{\mathrm{Var}\left(\theta^{(\mathrm{MH})}\right)}{\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right)}$ is a constant for any $0 < a < f$.

Contribution Two of our paper presents an alternative approach to Theorem 3.1, giving a very accurate approximation to $\mathrm{Var}\left(\theta^{(\mathrm{CMH})}\right)$, which can be replicated across different MinHash type algorithms which we discuss in Section 3 and Section 4 (our **Contribution 3**).

Moreover, the exact covariance identity in (16) makes Theorem 3.4, Proposition 3.2, and Proposition 3.6 immediate, while the approximations avoid the constrained combinatorial sum in Theorem 3.1. We discuss this in Section 2.

A.3 A Simpler Approach to Lemma 2.4

We give a simpler approach for the following lemma.

Lemma 1. *Lemma 2.4 in [15]. For any $1 \leq s < t \leq K$, with $t - s = \Delta$, then $\mathbb{E}_\pi\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = \frac{|\mathcal{L}_0(\Delta)| + (|\mathcal{G}_0(\Delta)| + |\mathcal{L}_2(\Delta)|)J}{f + |\mathcal{G}_0(\Delta)| + |\mathcal{G}_1(\Delta)|}$.*

Before we embark on a proof, we give intuition on how we prove this lemma, which helps if we want to generalize this approach beyond MinHash.

Following the sampling without replacement paradigm where a random permutation can be thought of as the "draw order", then for a fixed Δ, we can imagine drawing the indices $(x_s, x_{s+\Delta})$ in some random order. The first entry in the tuple $(A, B) = (x_s, x_{s+\Delta})$ then corresponds to what is seen sequentially when π is shifted s times to the right, and the second entry corresponds to what is seen when π is shifted $s + \Delta$ times to the right.

For example, take $D = K = 5$, $s = 1$ and $\Delta = 4$. After applying σ, suppose that the location vector is $\vec{x} = (\widetilde{-}, \widetilde{-}, \widetilde{O}, \widetilde{\times}, \widetilde{O})$. Suppose one random permutation π draws the indices in the order $1, 2, \ldots, 5$ under $\pi_{\to s}$. Shifting this same permutation rightwards by another 4 units gives $\pi_{\to s+\Delta}$, under which the indices are drawn in the order $5, 1, 2, 3, 4$. We take the first entry at the index drawn under $\pi_{\to s}$ and the second entry at the index 4 units to its right, returning to index 1 after index 5. Thus Draw 1 gives (x_1, x_5), Draw 2 gives (x_2, x_1), and the remaining draws follow the same order. Table 1 gives all five tuples. We bold the first $\widetilde{O}$ in each entry.

The first $\widetilde{O}$ in the first entry appears at Draw 3, before the first $\widetilde{\times}$ at Draw 4. The first $\widetilde{O}$ in the second entry appears at Draw 1, before the first $\widetilde{\times}$ at Draw 5. Hence $\mathbb{1}_s = \mathbb{1}_{s+\Delta} = 1$, so $\mathbb{1}_s \mathbb{1}_{s+\Delta} = 1$ for this π. The first $\widetilde{O}$ occurs at different draws for the two entries.

Draw	First entry under $\pi_{\to s}$	Second entry under $\pi_{\to s+\Delta}$	Tuple
Draw 1	$\widetilde{-}$	$\widetilde{\mathbf{O}}$	$(\widetilde{-}, \widetilde{\mathbf{O}})$
Draw 2	$\widetilde{-}$	$\widetilde{-}$	$(\widetilde{-}, \widetilde{-})$
Draw 3	$\widetilde{\mathbf{O}}$	$\widetilde{-}$	$(\widetilde{\mathbf{O}}, \widetilde{-})$
Draw 4	$\widetilde{\times}$	$\widetilde{O}$	$(\widetilde{\times}, \widetilde{O})$
Draw 5	$\widetilde{O}$	$\widetilde{\times}$	$(\widetilde{O}, \widetilde{\times})$

Table 1: Five successive tuples from one random permutation, for $D = 5$, $s = 1$ and $\Delta = 4$. The first entry corresponds to $\pi_{\to s}$ and the second entry to $\pi_{\to s+\Delta}$.

Each draw gives the next tuple from this same random permutation. The subscripts s and $s + \Delta$ in the two indicators denote the two circulant shifts. Each Example column in Tables 2 and 3 shows a sequence of tuple draws from one random permutation.

Since a $\widetilde{O}$ in the location vector means that the hashed values collide, then $\mathbb{E}_\pi\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) > 0$ means that both $\min\{\pi_{\to s}(\mathcal{S}_1)\} = \min\{\pi_{\to s}(\mathcal{S}_2)\}$ and $\min\{\pi_{\to s+\Delta}(\mathcal{S}_1)\} = \min\{\pi_{\to s+\Delta}(\mathcal{S}_2)\}$ must hold.

Hence, $\mathbb{E}_\pi\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) > 0$ can be thought of as the **probability of a sequence of draws** of the nine different types of D tuples where the first $\widetilde{O}$ appears before the first $\widetilde{\times}$ in both entries of the tuple, as the random permutation π is a sequence of draws without replacement. We call such a sequence of draws **a success**, otherwise a failure. Five examples of successes and failures are in Table 2. We bold the first $\widetilde{O}$ which occurs.

	Example 1	Example 2	Example 3	Example 4	Example 5
Draw 1	$(\widetilde{-}, \widetilde{-})$	$(\widetilde{-}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$
Draw 2	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\widetilde{\times}, \widetilde{\times})$	$(\widetilde{O}, \widetilde{-})$	$(\widetilde{-}, \widetilde{-})$
Draw 3	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{-}, \widetilde{\times})$		$(\widetilde{O}, \widetilde{-})$	$(\widetilde{-}, \widetilde{-})$
Draw 4		$(\widetilde{\mathbf{O}}, \widetilde{-})$		$(\widetilde{\times}, \widetilde{\mathbf{O}})$	$(\widetilde{-}, \widetilde{\mathbf{O}})$

Table 2: Five examples of sequences of draws. Each example is a vertical column.

Example 1 and **Example 2** are functionally the same (and are both successes). While **Example 2** has a $\widetilde{\times}$ in the second entry at Draw 3, it does not matter, because the first $\widetilde{O}$ in the second entry was already drawn in Draw 2.

Example 3 is an example of a failure, as a $\widetilde{\times}$ appears before an $\widetilde{O}$ in the second entry at Draw 2.

Example 4 and **Example 5** are also functionally the same. We also do not worry if an $\widetilde{\times}$ appears next to an $\widetilde{O}$ (as in Draw 4) for **Example 4**.

In fact, since it does not matter if a $\widetilde{\times}$ appears after an $\widetilde{O}$, we do not need to worry about elements appearing after the first $\widetilde{O}$, and hence can view our examples as in Table 3.

By treating the tuples (A, B) as sampling without replacement, and computing the appropriate probability of success, we can now prove Lemma 1 easily.

Proof. Similar to Li's approach, we consider three cases, and condition on $\mathbb{1}_s$. More specifically, we distinguish cases where the success for the **first entry of the tuple** happens when we draw

	Example 1	Example 2	Example 3	Example 4	Example 5
Draw 1	$(\widetilde{-}, \widetilde{-})$	$(\widetilde{-}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$
Draw 2	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\ \ , \widetilde{\times})$	$(\ \ , \widetilde{-})$	$(\ \ , \widetilde{-})$
Draw 3	$(\widetilde{\mathbf{O}}, \ \ )$	$(\widetilde{-}, \ \ )$		$(\ \ , \widetilde{-})$	$(\ \ , \widetilde{-})$
Draw 4		$(\widetilde{\mathbf{O}}, \ \ )$		$(\ \ , \widetilde{\mathbf{O}})$	$(\ \ , \widetilde{\mathbf{O}})$

Table 3: Five examples of sequences of draws, where we do not record the type of entry after the first $\widetilde{O}$ is drawn. Each example is a vertical column.

$\{(\widetilde{O}, \widetilde{\times}), (\widetilde{O}, \widetilde{O}), (\widetilde{O}, \widetilde{-})\}$ (before any $\widetilde{\times}$ for the **first entry**), and consider the probability for a success where where the first $\widetilde{O}$ is drawn before a $\widetilde{\times}$ for the **second entry of the tuple**. Table 4 gives a pictorial intuition of the five cases.

	Case 1	Case 2a	Case 2b	Case 3a	Case 3b
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\widetilde{-}, \widetilde{-})$	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$(\widetilde{-}, \widetilde{\mathbf{O}})$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$(\widetilde{-}, \ \ )$	$\vdots$	$(\widetilde{-}, \ \ )$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Draw t	$(\widetilde{\mathbf{O}}, \widetilde{\times})$	$(\widetilde{\mathbf{O}}, \widetilde{\mathbf{O}})$	$(\widetilde{\mathbf{O}}, \widetilde{O})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$	$(\widetilde{\mathbf{O}}, \widetilde{-})$
					$\vdots$
					$(\ \ , \widetilde{-})$

Table 4: Pictorial intuition of the five (sub)cases of **successes** we consider, as we condition on the first $\widetilde{O}$ seen in the first entry of the tuple. A bold $\widetilde{\mathbf{O}}$ denotes the first $\widetilde{O}$ seen for that entry. A blank entry in a tuple means it does not matter what that entry is since an $\widetilde{O}$ was drawn before.

Case One and Case Two follow along similar lines in their proof, but Case Three is vastly simplified. For each case, we show the product of probabilities "sequentially" in our reasoning.

We remind the reader that tuples which do not affect the hash colliding are not involved in the below probability calculations. As an analogy, consider a bag with one blue marble, two red marbles, and N green marbles. The probability of drawing a blue marble before a red marble is always $\frac{1}{3}$, regardless of N.

Case One: The first $\widetilde{O}$ for the **first entry of the tuple** is seen at $(\widetilde{O}, \widetilde{\times})$. A success for the **second entry of the tuple** can only happen before $(\widetilde{O}, \widetilde{\times})$, hence we need the probability of seeing the first $\widetilde{O}$ appearing before any $\widetilde{\times}$ for the **second entry of the tuple**. This is the product of the probabilities of i) drawing a $(\widetilde{-}, \widetilde{O})$ before any invalid tuples in $\{(\widetilde{-}, \widetilde{\times}), (\widetilde{\times}, \widetilde{O}), (\widetilde{\times}, \widetilde{-}), (\widetilde{\times}, \widetilde{\times}), (\widetilde{O}, \widetilde{O}), (\widetilde{O}, \widetilde{-}), (\widetilde{O}, \widetilde{\times})\}$, and ii) drawing a $(\widetilde{O}, \widetilde{\times})$ before any in-

valid tuples in $\{(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{-})\}$, which is given by

$$p_1 = \frac{|\mathcal{G}_0|}{f+|\mathcal{G}_0|+|\mathcal{G}_1|}\frac{|\mathcal{L}_1|}{f} = \frac{|\mathcal{G}_0||\mathcal{L}_1|}{(f+|\mathcal{G}_0|+|\mathcal{G}_1|)f}. \tag{2}$$

Case Two: The first $\widetilde{O}$ for the **first entry of the tuple** is seen at $(\widetilde{O},\widetilde{O})$. A success is always guaranteed once $(\widetilde{O},\widetilde{O})$ is drawn, provided there are no invalid $\widetilde{\times}$ before an $\widetilde{O}$ for the **second entry of the tuple**. Hence there are two possible disjoint sequences ending with $(\widetilde{O},\widetilde{O})$ that are successes.

a) A sequence where no $(\widetilde{-},\widetilde{\times})$ is seen before $(\widetilde{O},\widetilde{O})$, where we draw $(\widetilde{O},\widetilde{O})$ before invalid tuples $\{(\widetilde{-},\widetilde{\times}),(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{-}),(\widetilde{O},\widetilde{\times})\}$. The probability is given by

$$p_2 = \frac{|\mathcal{L}_0|}{f+|\mathcal{G}_1|}. \tag{3}$$

b) A sequence where at least one $(\widetilde{-},\widetilde{\times})$ is seen before $(\widetilde{O},\widetilde{O})$. This means a $(\widetilde{-},\widetilde{O})$ must be seen before it. This is the product of the probabilities of i) first drawing $(\widetilde{-},\widetilde{O})$ before invalid tuples in $\{(\widetilde{-},\widetilde{\times}),(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{-}),(\widetilde{O},\widetilde{-})\}$, ii) next drawing $(\widetilde{-},\widetilde{\times})$ before invalid tuples in $\{(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{-}),(\widetilde{O},\widetilde{\times})\}$, and iii) finally drawing $(\widetilde{O},\widetilde{O})$ before invalid tuples in $\{(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{-}),(\widetilde{O},\widetilde{\times})\}$, which is given by

$$p_3 = \frac{|\mathcal{G}_0|}{f+|\mathcal{G}_0|+|\mathcal{G}_1|}\frac{|\mathcal{G}_1|}{f+|\mathcal{G}_1|}\frac{|\mathcal{L}_0|}{f} = \frac{|\mathcal{G}_0||\mathcal{G}_1||\mathcal{L}_0|}{(f+|\mathcal{G}_0|+|\mathcal{G}_1|)(f+|\mathcal{G}_1|)f}. \tag{4}$$

Case Three: The first $\widetilde{O}$ for the **first entry of the tuple** is seen at $(\widetilde{O},\widetilde{-})$. Then a success occurs when either the first $\widetilde{O}$ for the **second entry of the tuple** appears before the first $(\widetilde{O},\widetilde{-})$, or it appears after, together with no $\widetilde{\times}$ appearing before the first $\widetilde{O}$ for the **first entry of the tuple**. Hence there are two possible disjoint sequences which are successes.

a) A sequence which ends at $(\widetilde{O},\widetilde{-})$, which means an $(\widetilde{-},\widetilde{O})$ must be seen before it. This is the product of probabilities of i) first drawing $(\widetilde{-},\widetilde{O})$ before any invalid tuples $\{(\widetilde{-},\widetilde{\times}),(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{-}),(\widetilde{O},\widetilde{\times})\}$, and ii) drawing $(\widetilde{O},\widetilde{-})$ before invalid tuples in $\{(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{\times})\}$, given by

$$p_4 = \frac{|\mathcal{G}_0|}{f+|\mathcal{G}_0|+|\mathcal{G}_1|}\frac{|\mathcal{L}_2|}{f} = \frac{|\mathcal{G}_0||\mathcal{L}_2|}{(f+|\mathcal{G}_0|+|\mathcal{G}_1|)f} \tag{5}$$

b) A sequence which ends after $(\widetilde{O},\widetilde{-})$, which means the first $\widetilde{O}$ for the **second entry of the tuple** appears after it. This is the product of probabilities of i) first drawing $(\widetilde{O},\widetilde{-})$ before invalid tuples $\{(\widetilde{-},\widetilde{O}),(\widetilde{-},\widetilde{\times}),(\widetilde{\times},\widetilde{O}),(\widetilde{\times},\widetilde{-}),(\widetilde{\times},\widetilde{\times}),(\widetilde{O},\widetilde{O}),(\widetilde{O},\widetilde{\times})\}$, and ii) drawing an element in $\{(\widetilde{O},\widetilde{O}),(\widetilde{-},\widetilde{O}),(\widetilde{\times},\widetilde{O})\}$ before invalid tuples in $\{(\widetilde{O},\widetilde{\times}),(\widetilde{-},\widetilde{\times}),(\widetilde{\times},\widetilde{\times})\}$, given by

$$p_5 = \frac{|\mathcal{L}_2|}{f+|\mathcal{G}_0|+|\mathcal{G}_1|}\frac{a}{f} = \frac{|\mathcal{L}_2|a}{(f+|\mathcal{G}_0|+|\mathcal{G}_1|)f}. \tag{6}$$

Adding (2), (3), (4), (5), (6) gives $\frac{|\mathcal{L}_0|+(|\mathcal{G}_0|+|\mathcal{L}_2|)J}{f+|\mathcal{G}_0|+|\mathcal{G}_1|}$.

□

A.4 An Alternative Approach to Theorem 3.1

Lemma 2. *Suppose we pre-process the location vector with the random permutation σ. Let $B_D = 2D - f - 1$, and suppose $0 < J < 1$. For all $1 \le \Delta \le K - 1$, if $f/D \to 0$, then*

$$\mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = \begin{cases} J^2 - \dfrac{J(1-J)f}{2(f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & D \text{ even and } \Delta = D/2, \\ J^2 - \dfrac{J(1-J)f}{(2f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & \text{otherwise}, \end{cases}$$

$$\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_{s+\Delta}\right) = \begin{cases} -\dfrac{J(1-J)f}{2(f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & D \text{ even and } \Delta = D/2, \\ -\dfrac{J(1-J)f}{(2f-1)(D-1)} + O\left(\dfrac{f}{D^2}\right), & \text{otherwise}. \end{cases}$$

If f/D is bounded away from zero, then

$$\mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = J^2 - \frac{J(1-J)}{B_D} + O(D^{-2}),$$
$$\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_{s+\Delta}\right) = -\frac{J(1-J)}{B_D} + O(D^{-2}).$$

Proof. Lemma 1 gives $\mathbb{E}_{\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right)$. Define

$$N_\Delta = |\mathcal{L}_0(\Delta)| + J\left(|\mathcal{G}_0(\Delta)| + |\mathcal{L}_2(\Delta)|\right), \tag{7}$$
$$M_\Delta = f + |\mathcal{G}_0(\Delta)| + |\mathcal{G}_1(\Delta)|. \tag{8}$$

Thus

$$\mathbb{E}_{\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = \frac{N_\Delta}{M_\Delta}.$$

We first condition on the f positions of the location vector whose entries are $\widetilde{O}$ or $\widetilde{\times}$. Let

$$X_\Delta = \sum_{i=1}^{D} \mathbb{1}_{\{x_i \neq \widetilde{-},\ x_{i+\Delta} \neq \widetilde{-}\}}. \tag{9}$$

The quantity X_Δ counts the tuples $(x_i, x_{i+\Delta})$ for which both entries are $\widetilde{O}$ or $\widetilde{\times}$. There are $f - X_\Delta$ tuples whose first entry is $\widetilde{-}$ and whose second entry is $\widetilde{O}$ or $\widetilde{\times}$. To see this, each of the f entries equal to $\widetilde{O}$ or $\widetilde{\times}$ has one entry at distance Δ before it, and X_Δ of those preceding entries are also $\widetilde{O}$ or $\widetilde{\times}$. Therefore

$$|\mathcal{G}_0(\Delta)| + |\mathcal{G}_1(\Delta)| = f - X_\Delta. \tag{10}$$

Substitution of (10) into (8) gives

$$M_\Delta = 2f - X_\Delta. \tag{11}$$

Conditional on the positions counted by (9), the a entries equal to $\widetilde{O}$ are placed among the f positions equal to $\widetilde{O}$ or $\widetilde{\times}$ by sampling without replacement. For a tuple counted by X_Δ, the probability that both entries are $\widetilde{O}$ is $a(a-1)/(f(f-1))$. Also, $f - X_\Delta$ tuples go from $\widetilde{-}$ to either $\widetilde{O}$ or $\widetilde{\times}$, and another $f - X_\Delta$ tuples go from either $\widetilde{O}$ or $\widetilde{\times}$ to $\widetilde{-}$. In each case, the entry which is not $\widetilde{-}$ equals $\widetilde{O}$ with probability a/f. Hence

$$\mathbb{E}_{\sigma}\left(|\mathcal{L}_0(\Delta)| \mid X_\Delta\right) = X_\Delta \frac{a(a-1)}{f(f-1)}, \tag{12}$$

$$\mathbb{E}_{\sigma}\left(|\mathcal{G}_0(\Delta)| + |\mathcal{L}_2(\Delta)| \mid X_\Delta\right) = 2(f - X_\Delta)\frac{a}{f}. \tag{13}$$

Using $J = a/f$ in (7), (12) and (13) give

$$\mathbb{E}_\sigma\left(N_\Delta \mid X_\Delta\right) = J^2(2f - X_\Delta) - \frac{J(1-J)}{f-1}X_\Delta. \tag{14}$$

The denominator in (11) is fixed once X_Δ is fixed. Dividing (14) by (11) and then averaging over the positions gives

$$\mathbb{E}_{\sigma,\pi}\left(\mathbb{1}_s \mathbb{1}_{s+\Delta}\right) = J^2 - \frac{J(1-J)}{f-1}\mathbb{E}_\sigma\left(\frac{X_\Delta}{2f - X_\Delta}\right), \tag{15}$$

$$\mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_{s+\Delta}\right) = -\frac{J(1-J)}{f-1}\mathbb{E}_\sigma\left(\frac{X_\Delta}{2f - X_\Delta}\right). \tag{16}$$

We next approximate the expectation in (16). Write $Y_i = \mathbb{1}_{\left\{x_i \neq \widetilde{-},\ x_{i+\Delta} \neq \widetilde{-}\right\}}$, so that $X_\Delta = \sum_i Y_i$ by (9). For each tuple $(x_i, x_{i+\Delta})$, the probability that both entries are $\widetilde{O}$ or $\widetilde{\times}$ is $f(f-1)/(D(D-1))$. Summing this probability over the D possible values of i gives

$$\mathbb{E}_\sigma\left(X_\Delta\right) = D\frac{f(f-1)}{D(D-1)} = \frac{f(f-1)}{D-1}. \tag{17}$$

Suppose first that D is odd or $\Delta \neq D/2$. In the sum $X_\Delta(X_\Delta - 1) = \sum_{i \neq j} Y_i Y_j$, there are $2D$ choices of (i,j) for which the tuples $(x_i, x_{i+\Delta})$ and $(x_j, x_{j+\Delta})$ have one common position. The remaining $D(D-3)$ choices use four distinct positions. If $u^{\underline{r}} = u(u-1)\cdots(u-r+1)$, sampling without replacement therefore gives

$$\mathbb{E}_\sigma\left(X_\Delta(X_\Delta - 1)\right) = 2D\frac{f^{\underline{3}}}{D^{\underline{3}}} + D(D-3)\frac{f^{\underline{4}}}{D^{\underline{4}}} = \frac{f(f-1)^2(f-2)}{(D-1)(D-2)}. \tag{18}$$

For every integer $0 \leq x \leq f$,

$$\frac{x}{2f-x} = \frac{x}{2f-1} + \frac{x(x-1)}{(2f-1)(2f-x)}. \tag{19}$$

Apply (19) to X_Δ, use (17) for the first term, and use $2f - X_\Delta \geq f$ together with (18) for the second term. This gives

$$\frac{1}{f-1}\mathbb{E}_\sigma\left(\frac{X_\Delta}{2f - X_\Delta}\right) = \frac{f}{(2f-1)(D-1)} + R_\Delta, \tag{20}$$

$$0 \leq R_\Delta \leq \frac{(f-1)(f-2)}{(2f-1)(D-1)(D-2)} = O\left(\frac{f}{D^2}\right). \tag{21}$$

Now suppose that D is even and $\Delta = D/2$. The tuple $(x_i, x_{i+D/2})$ is accompanied by the tuple $(x_{i+D/2}, x_i)$, so X_Δ is even. Let $Q_\Delta = X_\Delta/2$. Two different pairs of these tuples have four different positions. There are $(D/2)(D/2-1)$ choices of the first and second pair, and hence

$$\begin{aligned}\mathbb{E}_\sigma\left(X_\Delta(X_\Delta - 2)\right) &= 4\mathbb{E}_\sigma\left(Q_\Delta(Q_\Delta - 1)\right) \\ &= 4\frac{D}{2}\left(\frac{D}{2} - 1\right)\frac{f^{\underline{4}}}{D^{\underline{4}}} = \frac{f(f-1)(f-2)(f-3)}{(D-1)(D-3)}.\end{aligned} \tag{22}$$

For every even integer $0 \leq x \leq f$,

$$\frac{x}{2f-x} = \frac{x}{2(f-1)} + \frac{x(x-2)}{2(f-1)(2f-x)}. \tag{23}$$

Equations (17), (22), and (23) give

$$\frac{1}{f-1}\mathbb{E}_\sigma\left(\frac{X_\Delta}{2f - X_\Delta}\right) = \frac{f}{2(f-1)(D-1)} + R_{D/2}, \tag{24}$$

$$0 \leq R_{D/2} \leq \frac{(f-2)(f-3)}{2(f-1)(D-1)(D-3)} = O\left(\frac{f}{D^2}\right). \tag{25}$$

Substituting (20) and (24), with their respective bounds (21) and (25), into (15) and (16) proves the two expressions when $f/D \to 0$. The two bounds hold for every $1 \le \Delta \le K-1$ because the pair count only distinguishes the case in (18) from the case in (22).

Now suppose that f/D is bounded away from zero. Define

$$h(x) = \frac{x}{2f-x}, \quad \mu_\Delta = \mathbb{E}_\sigma\left(X_\Delta\right) = \frac{f(f-1)}{D-1},$$

where the value of μ_Δ is from (17). Direct algebra gives

$$\frac{h(\mu_\Delta)}{f-1} = \frac{1}{2D-f-1} = \frac{1}{B_D}. \tag{26}$$

The same pair count used in (18) and (22) gives

$$\mathrm{Var}_\sigma\left(X_\Delta\right) = \begin{cases} \dfrac{2f(D-f)(f-1)(D-f-1)}{(D-1)^2(D-3)}, & D \text{ even and } \Delta = D/2, \\ \dfrac{f(D-f)(f-1)(D-f-1)}{(D-1)^2(D-2)}, & \text{otherwise.} \end{cases} \tag{27}$$

For $0 \le x \le f$,

$$h''(x) = \frac{4f}{(2f-x)^3}, \quad 0 < h''(x) \le \frac{4}{f^2}. \tag{28}$$

We Taylor expansion about μ_Δ, take expectations, and using the fact $\mathbb{E}_\sigma\left(X_\Delta - \mu_\Delta\right) = 0$ and (28) gives

$$0 \le \mathbb{E}_\sigma\left(h(X_\Delta)\right) - h(\mu_\Delta) \le \frac{2}{f^2}\mathrm{Var}_\sigma\left(X_\Delta\right). \tag{29}$$

If f/D is bounded below by a positive constant, the two expressions for the variance in (27) are $O(D)$. Dividing (29) by $f-1$ therefore gives

$$0 \le \frac{\mathbb{E}_\sigma\left(h(X_\Delta)\right) - h(\mu_\Delta)}{f-1} \le \frac{2\mathrm{Var}_\sigma\left(X_\Delta\right)}{f^2(f-1)} = O(D^{-2}). \tag{30}$$

Combining (26) and (30) gives

$$\frac{1}{f-1}\mathbb{E}_\sigma\left(\frac{X_\Delta}{2f-X_\Delta}\right) = \frac{1}{B_D} + O(D^{-2}). \tag{31}$$

Finally, substituting (31) into (15) and (16) gives the two expressions when f/D is bounded away from zero. □

Theorem 2. *Let $0 < a \le f \le D$. Suppose we preprocess the location vector with the random permutation σ, and generate an independent permutation π. Let $B_D = 2D - f - 1$. If $f/D \to 0$,*

$$\mathrm{Var}_{\sigma,\pi}\left(\theta^{(\mathrm{CMH})}\right) = \frac{J(1-J)}{K}\left[1 - \frac{f}{D-1}\left\{\frac{K-1-\lambda_{D,K}}{2f-1} + \frac{\lambda_{D,K}}{2(f-1)}\right\}\right] + O\left(\frac{f}{D^2}\right).$$

If f/D is bounded away from zero, then

$$\mathrm{Var}_{\sigma,\pi}\left(\theta^{(\mathrm{CMH})}\right) = \frac{J(1-J)}{K}\left(1 - \frac{K-1}{B_D}\right) + O(D^{-2}).$$

Proof. This follows from expanding $\mathrm{Var}_{\sigma,\pi}\left(\sum_{t=1}^K \mathbb{1}_t/K\right) = \frac{1}{K^2}\sum_{s=1}^K\sum_{t=1}^K \mathrm{Cov}_{\sigma,\pi}\left(\mathbb{1}_s, \mathbb{1}_t\right)$, using covariance terms in Lemma 2. The term $\lambda_{D,K}$ accounts for the proportion of covariance terms where D is even and $\Delta = D/2$. □

A.5 Additional Experiments to Verify Variance Approximation

While Figure 3 shows that our variance approximation for $\frac{f}{D} \to 0$ does match the empirical MSE even at small $D = 128$, we also need to understand how it behaves at edge cases (i.e. when f_1, f_2 approaches zero). While this usually isn't the case in practical settings, we want to investigate how well we can rely on the given approximation.

Hence, we plot (f, a) pairs of $(2, 1), (4, 2), (6, 3), (8, 4)$ at $D = 128$ and $D = 1024$ respectively. We see that at $f = 2, 4$, the approximate theoretical variance lines are slightly off, but still extremely close to the empirical MSE, but at $f = 6, 8$, both lines overlap each other almost perfectly. Hence, we believe our approximation of the theoretical variance is practical enough to do comparisons.

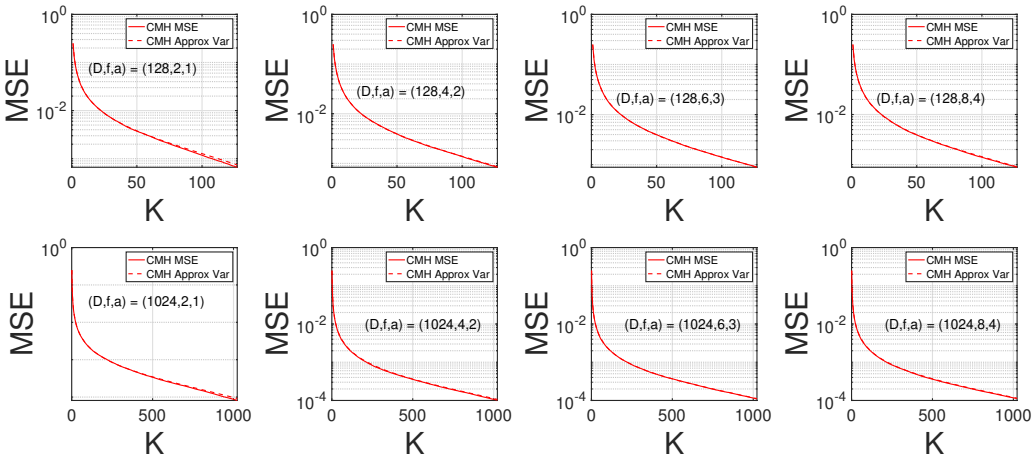

Figure 6: Empirical MSE vs theoretical approximate variance in Theorem 2 of MinHash and CMinHash (σ, π), at edge cases when f is small. At $f = 2, 4$, the dotted lines do not overlap perfectly, but they do at $f = 6, 8$.

Finally, we plot both theoretical approximations for the variance in Thereom 2 and the empirical MSE for select word pairs from the words dataset [7] in Figure 7. Here, $D = 65536$, while the word pairs have $\frac{f_i}{D} \in [0.0143, 0.3226]$, i.e. where $\frac{f}{D} \to 0$ and where $\frac{f}{D}$ does not tend to zero.

We average across 10,000 iterations as well. We can see that both theoretical approximations as well as the empirical MSE lines overlap perfectly.

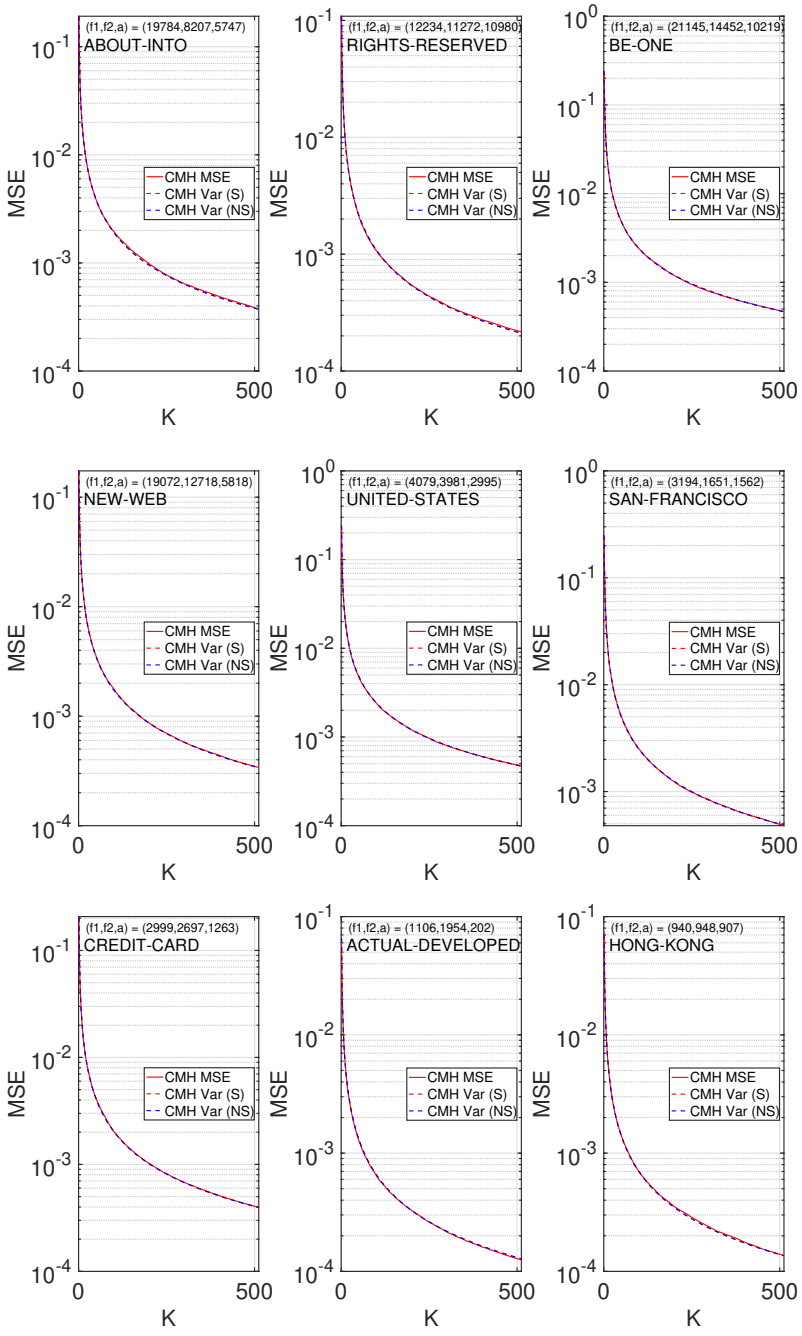

Figure 7: Empirical MSE versus both theoretical approximate variances in Theorem 2 of MinHash and CMinHash (σ, π) for selected word pairs. CMH Var (S) is the the *sparse* case where $\frac{f_1}{D} \to 0$. CMH Var (NS) is the the *non-sparse* case where $\frac{f_1}{D}$ does not tend to zero. Both theoretical approximations and empirical MSE lines overlap perfectly.

B Appendix: Circulant Permutation for The Minner Estimators

This appendix covers Section 3 in more detail.

B.1 Understanding the Minner Estimators

To better understand the Minner estimators, we give the following lemma before explaining in detail.

Lemma 5. *Suppose we have a random permutation π on $[D]$. Let $y = \mathbb{1}_{\{\min(\pi(\mathcal{S}_2))\in\pi(\mathcal{S}_1)\}}$, and m to be the count of elements in $\pi(\mathcal{S}_1)$ strictly lesser than $\min(\pi(\mathcal{S}_2))$. Then $\mathbb{E}(y) = \frac{a}{f_2}$, $\mathbb{E}(m) = \frac{f_1-a}{f_2+1}$, $\mathrm{Var}(m) = \frac{(f_1-a)(f_1+f_2-a+1)f_2}{(f_2+1)^2(f_2+2)}$.*

Draw Order	1	2	3	4	5	6	7	8
Set 1	○	●	○	●	●	●	○	●
Set 2	○	○	○	○	○	○	○	●

Figure 4: We have $\min\{\pi_t(\mathcal{S}_2)\} = 8$, with $y_t = 1$, which can be thought of as the presence (or absence) of a • at draw 8 for $\mathcal{S}_1$, and $m_t = 4$, which can be thought of as counting how many • were seen in previous draws 1-7 for $\mathcal{S}_1$.

We also replicate Figure 4 here, as viewing the random variables under the "drawing objects" from the sampling without replacement paradigm makes the proof straightforward.

Proof. Set $\mathcal{F}_+ = \left\{ \begin{smallmatrix} \bullet \\ \bullet \end{smallmatrix} \right\}$, and $\mathcal{F}_- = \left\{ \begin{smallmatrix} \circ \\ \bullet \end{smallmatrix} \right\}$, with $|\mathcal{F}_+| = a$ and $|\mathcal{F}_-| = f_2 - a$. Then y is simply the probability of seeing an element from $\mathcal{F}_+$ drawn before an element in $\mathcal{F}_-$, which is $\frac{a}{f_2-a+a} = \frac{a}{f_2}$.

Now, set $\mathcal{F}_+ = \left\{ \begin{smallmatrix} \bullet \\ \bullet \end{smallmatrix} , \begin{smallmatrix} \circ \\ \bullet \end{smallmatrix} \right\}$, and $\mathcal{F}_- = \left\{ \begin{smallmatrix} \bullet \\ \circ \end{smallmatrix} \right\}$, with $|\mathcal{F}_+| = f_2$ and $|\mathcal{F}_-| = f_1 - a$. Then m is the count of the number of elements in $\mathcal{F}_-$ seen before the first element in $\mathcal{F}_+$.

Suppose we index each element in $\mathcal{F}_-$ by i, $1 \le i \le f_1 - a$. We can write the random variable m as

$$m = \sum_{i=1}^{f_1-a} \mathbb{1}_{\{\text{element } i \text{ is drawn before any element in } \mathcal{F}_+\}}.$$

There are f_2 elements in $\mathcal{F}_+$, and we require the single element i in $\mathcal{F}_-$ to be drawn first, hence

$$\mathbb{E}\left(\mathbb{1}_{\{\text{element } i \text{ is drawn before any element in } \mathcal{F}_+\}}\right) = \frac{1}{|\mathcal{F}_+|+1} = \frac{1}{f_2+1},$$

by exchangeability. Now by linearity of expectations, $\mathbb{E}(m) = \sum_{i=1}^{f_1-a} \frac{1}{f_2+1} = \frac{f_1-a}{f_2+1}$.

To find $\mathrm{Var}(m)$, we need to find $\mathbb{E}(m^2)$. Certainly,

$$\mathbb{E}(m^2) = \sum_{i=1}^{f_1-a} \mathbb{E}\left(\mathbb{1}_{\{\text{element } i \text{ is drawn before any element in } \mathcal{F}_+\}}\right) + \sum_{i\neq j}^{f_1-a} \mathbb{E}\left(\mathbb{1}_{\{\text{element } i \text{ is drawn before any element in } \mathcal{F}_+\}} \mathbb{1}_{\{\text{element } j \text{ is drawn before any element in } \mathcal{F}_+\}}\right).$$

Now $\mathbb{E}\left(\mathbb{1}_{\{\text{element } i \text{ is drawn before any element in } \mathcal{F}_+\}} \mathbb{1}_{\{\text{element } j \text{ is drawn before any element in } \mathcal{F}_+\}}\right) = \frac{2}{(f_2+1)(f_2+2)}$, as this can be thought of as out of the f_2 elements in $\mathcal{F}_+$, and the two elements i, j the i, j elements must be drawn first. Therefore,

$$\mathbb{E}(m^2) = \frac{f_1-a}{f_2+1} + \frac{2(f_1-a)(f_1-a-1)}{(f_2+1)(f_2+2)},$$

and

$$\mathrm{Var}\,(m) = \mathbb{E}\left(m^2\right) - \mathbb{E}\,(m)^2 = \frac{(f_1 - a)f_2(f_1 + f_2 - a + 1)}{(f_2+1)^2(f_2+2)}.$$

□

We remark that both these estimators have different performance based on the sizes of the sets $\mathcal{S}_1, \mathcal{S}_2$. However, it is not in the scope of this paper to analyze which estimator is better given a, f_1, f_2, D, or propose better Minner estimators. The paper focuses on **extending the circulant permutation framework when counts are involved**.

B.2 Proof of Lemma 3

We recall the definition of the nine sets again, but this time using our notation C_s, C_f and C_n (success, failure, neutral) for generalizability.

For $A, B \in \{\widetilde{O}, \widetilde{\times}, \widetilde{-}\}$, let $\{(A,B) \mid \Delta\}$ be the set $\{(i,j) : (x_i, x_j) = (A,B), j - i = \Delta\}$. For $1 \le \Delta \le K-1$, define

$$\begin{aligned}
\mathcal{L}_0(\Delta) &= \{(\widetilde{O},\widetilde{O})|\Delta\}, \quad \mathcal{L}_1(\Delta) = \{(\widetilde{O},\widetilde{\times})|\Delta\}, \quad \mathcal{L}_2(\Delta) = \{(\widetilde{O},\widetilde{-})|\Delta\},\\
\mathcal{G}_0(\Delta) &= \{(\widetilde{-},\widetilde{O})|\Delta\}, \quad \mathcal{G}_1(\Delta) = \{(\widetilde{-},\widetilde{\times})|\Delta\}, \quad \mathcal{G}_2(\Delta) = \{(\widetilde{-},\widetilde{-})|\Delta\},\\
\mathcal{H}_0(\Delta) &= \{(\widetilde{\times},\widetilde{O})|\Delta\}, \quad \mathcal{H}_1(\Delta) = \{(\widetilde{\times},\widetilde{\times})|\Delta\}, \quad \mathcal{H}_2(\Delta) = \{(\widetilde{\times},\widetilde{-})|\Delta\},
\end{aligned}$$

with the following constraints for all $1 \le \Delta \le K-1$

$$\begin{aligned}
|\mathcal{L}_0| + |\mathcal{L}_1| + |\mathcal{L}_2| &= |\mathcal{L}_0| + |\mathcal{G}_0| + |\mathcal{H}_0| = C_s,\\
|\mathcal{G}_0| + |\mathcal{G}_1| + |\mathcal{G}_2| &= |\mathcal{L}_2| + |\mathcal{G}_2| + |\mathcal{H}_2| = C_n,\\
|\mathcal{H}_0| + |\mathcal{H}_1| + |\mathcal{H}_2| &= |\mathcal{L}_1| + |\mathcal{G}_1| + |\mathcal{H}_1| = C_f.
\end{aligned}$$

Lemma 3. *Suppose we pre-process the location vector with a random permutation σ. Let the random variable $\theta_s \equiv$ number of $\widetilde{\times}$ seen before the first $\widetilde{O}$ under the random permutation $\pi_{\to s}$. Let $B_\Delta = 2C_s - |\mathcal{L}_0| + 1$. Then for any $1 \le s < t \le K$, with $t - s = \Delta$,*

$$\mathbb{E}_\pi\left(\theta_s\theta_{s+\Delta}\right) = \frac{|\mathcal{H}_1|}{B_\Delta} + \frac{2C_f^2 - C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2\,|\mathcal{H}_1|}{(C_s+1)(B_\Delta+1)} + \frac{C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2\,|\mathcal{L}_1|\,|\mathcal{H}_0|}{(C_s+1)B_\Delta(B_\Delta+1)}.$$

If C_s and C_f are fixed as $D \to \infty$, then, when D is even and $\Delta = D/2$,

$$\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = -V_{\mathrm{ind}}\frac{(C_s+1)(C_s+2)}{2C_s(2C_s+1)}\frac{2C_s + C_f}{(C_s + C_f + 1)(D-1)} + O(D^{-2}),$$

and otherwise,

$$\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = -V_{\mathrm{ind}}\frac{(C_s+2)(C_s+C_f)}{(2C_s+1)(C_s+C_f+1)(D-1)} + O(D^{-2}).$$

If both C_s/D and C_f/D are bounded away from zero, then

$$\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = \begin{cases} -V_{\mathrm{ind}}\left(\frac{C_s(2D-C_s)+C_f(D-C_s)}{(C_s+C_f)(2D-C_s)^2}\right) + O\left(\frac{1}{D^2}\right), & D \text{ even and } \Delta = D/2,\\ -V_{\mathrm{ind}}\left(\frac{1}{2D-C_s}\right) + O\left(\frac{1}{D^2}\right), & \text{otherwise.}\end{cases}$$

Proof. Our proof will follow a similar approach to the proof of Lemma 1 with respect to sequence of draws.

For ease of notation, we define the sets

$$\begin{aligned}
\mathcal{O}_1 &= \mathcal{L}_0 \cup \mathcal{L}_1 \cup \mathcal{L}_2 \text{ (tuples where first entry is } \widetilde{O}),\\
\mathcal{O}_2 &= \mathcal{L}_0 \cup \mathcal{G}_0 \cup \mathcal{H}_0 \text{ (tuples where second entry is } \widetilde{O}),\\
\mathcal{X}_1 &= \mathcal{H}_0 \cup \mathcal{H}_1 \cup \mathcal{H}_2 \text{ (tuples where first entry is } \widetilde{\times}),\\
\mathcal{X}_2 &= \mathcal{L}_1 \cup \mathcal{G}_1 \cup \mathcal{H}_1 \text{ (tuples where second entry is } \widetilde{\times}),
\end{aligned}$$

and define the variable

$$U = |\mathcal{O}_1 \cup \mathcal{O}_2| = 2C_s - |\mathcal{L}_0|\,.$$

For an element i in $\mathcal{X}_1$, we write the random variable $I_i^{(1)} \equiv \mathbb{1}_{\{i \text{ appears before elements in } \mathcal{O}_1\}}$. Similarly, for an element j in $\mathcal{X}_2$, we write $I_j^{(2)} \equiv \mathbb{1}_{\{j \text{ appears before elements in } \mathcal{O}_2\}}$. The two random variables can therefore be written as

$$\theta_s = \sum_{i \in \mathcal{X}_1} I_i^{(1)}, \quad \theta_{s+\Delta} = \sum_{j \in \mathcal{X}_2} I_j^{(2)}. \tag{32}$$

Multiplying the two sums in (32) gives

$$\mathbb{E}_\pi\left(\theta_s \theta_{s+\Delta}\right) = \sum_{i \in \mathcal{X}_1 \cap \mathcal{X}_2} \mathbb{E}_\pi\left(I_i^{(1)} I_i^{(2)}\right) + \sum_{\substack{i \in \mathcal{X}_1,\, j \in \mathcal{X}_2 \\ i \neq j}} \mathbb{E}_\pi\left(I_i^{(1)} I_j^{(2)}\right). \tag{33}$$

For the first sum in (33), the equality $i = j$ requires the tuple $(\widetilde{\times}, \widetilde{\times})$, so there are $|\mathcal{H}_1|$ choices. For any one of them, $I_i^{(1)} I_i^{(2)} = 1$ when i is the first tuple among $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \{i\}$. Exchangeability of these $U + 1$ tuples gives probability $1/(U+1)$. Hence

$$\sum_{i \in \mathcal{X}_1 \cap \mathcal{X}_2} \mathbb{E}_\pi\left(I_i^{(1)} I_i^{(2)}\right) = \frac{|\mathcal{H}_1|}{U+1}. \tag{34}$$

We now evaluate the second sum in (33). As in the proof of Lemma 1, we consider four cases depending on whether i is in $\mathcal{O}_2$ and whether j is in $\mathcal{O}_1$.

Case One: Suppose $i \notin \mathcal{O}_2$ and $j \notin \mathcal{O}_1$. Let E be the earliest tuple seen out of $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \{i\} \cup \{j\}$. Since both i and j are outside $\mathcal{O}_1 \cup \mathcal{O}_2$, we have

$$\mathbb{P}\left(E = i\right) = \mathbb{P}\left(E = j\right) = \frac{1}{U+2}.$$

A valid sequence occurs in two disjoint ways. First, i is drawn before any tuple in $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \{j\}$, and then j is drawn before any tuple in $\mathcal{O}_2$. Second, j is drawn before any tuple in $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \{i\}$, and then i is drawn before any tuple in $\mathcal{O}_1$. Thus the probability for each ordered pair in this case is

$$\frac{1}{U+2}\frac{1}{C_s+1} + \frac{1}{U+2}\frac{1}{C_s+1} = \frac{2}{(C_s+1)(U+2)}.$$

There are

$$(|\mathcal{H}_1| + |\mathcal{H}_2|)(|\mathcal{G}_1| + |\mathcal{H}_1|) - |\mathcal{H}_1| = (C_f - |\mathcal{H}_0|)(C_f - |\mathcal{L}_1|) - |\mathcal{H}_1|$$

ordered pairs with $i \notin \mathcal{O}_2$, $j \notin \mathcal{O}_1$, and $i \neq j$. Hence the contribution from Case One is

$$\frac{2\left((C_f - |\mathcal{H}_0|)(C_f - |\mathcal{L}_1|) - |\mathcal{H}_1|\right)}{(C_s+1)(U+2)}. \tag{35}$$

Case Two: Suppose $i \in \mathcal{O}_2$ and $j \notin \mathcal{O}_1$. Since $i \in \mathcal{X}_1 \cap \mathcal{O}_2$, this means $i \in \mathcal{H}_0$. A valid sequence must first draw j before any tuple in $\mathcal{O}_2$, and then draw i before any tuple in $\mathcal{O}_1$. Since only j is added to $\mathcal{O}_1 \cup \mathcal{O}_2$, the probability for each ordered pair in this case is

$$\frac{1}{U+1}\frac{1}{C_s+1} = \frac{1}{(C_s+1)(U+1)}.$$

There are $|\mathcal{H}_0|\,(C_f - |\mathcal{L}_1|)$ such ordered pairs. Hence the contribution from Case Two is

$$\frac{|\mathcal{H}_0|\,(C_f - |\mathcal{L}_1|)}{(C_s+1)(U+1)}. \tag{36}$$

Case Three: Suppose $i \notin \mathcal{O}_2$ and $j \in \mathcal{O}_1$. By the same reasoning as Case Two, $j \in \mathcal{L}_1$, and the contribution is

$$\frac{(C_f - |\mathcal{H}_0|)\,|\mathcal{L}_1|}{(C_s+1)(U+1)}. \tag{37}$$

Case Four: Suppose $i \in \mathcal{O}_2$ and $j \in \mathcal{O}_1$. This case contributes zero. Since $i \in \mathcal{O}_2$, the event $\mathbb{1}_{\{j \text{ before } \mathcal{O}_2\}} = 1$ requires j to appear before i. However, since $j \in \mathcal{O}_1$, the event $\mathbb{1}_{\{i \text{ before } \mathcal{O}_1\}} = 1$ requires i to appear before j.

Adding (34), (35), (36), and (37), and substituting $U = 2C_s - |\mathcal{L}_0|$, gives

$$\begin{aligned}\mathbb{E}_\pi\left(\theta_s\theta_{s+\Delta}\right) = \frac{|\mathcal{H}_1|}{2C_s - |\mathcal{L}_0| + 1} + \frac{2C_f^2 - C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2\,|\mathcal{H}_1|}{(C_s+1)(2C_s - |\mathcal{L}_0| + 2)} \\ + \frac{C_f(|\mathcal{L}_1| + |\mathcal{H}_0|) - 2\,|\mathcal{L}_1|\,|\mathcal{H}_0|}{(C_s+1)(2C_s - |\mathcal{L}_0| + 1)(2C_s - |\mathcal{L}_0| + 2)}.\end{aligned} \tag{38}$$

This proves the first expression in Lemma 3.

Now suppose we pre-process the location vector with the random permutation σ. Then we want to find the expectation of (38) with respect to σ. For ease of notation, let

$$Z = (Z_0, Z_1, Z_2, Z_3) = \left(|\mathcal{L}_0|, |\mathcal{L}_1|, |\mathcal{H}_0|, |\mathcal{H}_1|\right), \tag{39}$$

and let $F(Z)$ denote the right hand side of (38). When $Z = (0,0,0,0)$,

$$F(0,0,0,0) = \frac{2C_f^2}{(C_s+1)(2C_s+2)} = \left(\frac{C_f}{C_s+1}\right)^2. \tag{40}$$

First suppose that C_s and C_f are fixed while $D \to \infty$, and that D is odd or $\Delta \neq D/2$. Let

$$R_\Delta = Z_0 + Z_1 + Z_2 + Z_3.$$

Thus R_Δ is the number of tuples $(x_i, x_{i+\Delta})$ whose two entries are $\widetilde{O}$ or $\widetilde{\times}$. When $R_\Delta = 1$, the four possible values of Z and their changes from (40) are

pair	Z	$F(Z) - F(0,0,0,0)$
$(\widetilde{O}, \widetilde{O})$	$(1,0,0,0)$	$\dfrac{C_f^2}{(C_s+1)^2(2C_s+1)}$
$(\widetilde{O}, \widetilde{\times})$	$(0,1,0,0)$	$-\dfrac{C_sC_f}{(C_s+1)^2(2C_s+1)}$
$(\widetilde{\times}, \widetilde{O})$	$(0,0,1,0)$	$-\dfrac{C_sC_f}{(C_s+1)^2(2C_s+1)}$
$(\widetilde{\times}, \widetilde{\times})$	$(0,0,0,1)$	$\dfrac{C_s^2}{(C_s+1)^2(2C_s+1)}$.

(41)

The four rows in (41) are obtained by substituting $Z = (1,0,0,0)$, $Z = (0,1,0,0)$, $Z = (0,0,1,0)$, and $Z = (0,0,0,1)$ into the right hand side of (38), and then subtracting (40). For example, the $(\widetilde{O}, \widetilde{\times})$ row is

$$\begin{aligned}F(0,1,0,0) - F(0,0,0,0) &= \frac{2C_f^2 - C_f}{2(C_s+1)^2} + \frac{C_f}{2(C_s+1)^2(2C_s+1)} - \frac{C_f^2}{(C_s+1)^2} \\ &= -\frac{C_sC_f}{(C_s+1)^2(2C_s+1)}.\end{aligned}$$

For each Δ, sampling without replacement gives

$$\begin{aligned}\mathbb{E}_\sigma\left(|\mathcal{L}_0|\right) &= \frac{C_s(C_s-1)}{D-1}, & \mathbb{E}_\sigma\left(|\mathcal{L}_1|\right) &= \frac{C_sC_f}{D-1}, \\ \mathbb{E}_\sigma\left(|\mathcal{H}_0|\right) &= \frac{C_sC_f}{D-1}, & \mathbb{E}_\sigma\left(|\mathcal{H}_1|\right) &= \frac{C_f(C_f-1)}{D-1}.\end{aligned} \tag{42}$$

The same calculation as in (18), with f replaced by $C_s + C_f$ gives

$$\mathbb{E}_\sigma\left(R_\Delta(R_\Delta - 1)\right) = O(D^{-2}).$$

Therefore $\mathbb{P}(R_\Delta \geq 2) = O(D^{-2})$. When $R_\Delta = 0$, the change in F is zero by (40). When $R_\Delta = 1$, it is the corresponding entry of (41). When $R_\Delta \geq 2$, F is bounded because C_s and C_f are constant. Multiplying the four entries in (41) by the four expectations in (42) and adding them gives

$$\begin{aligned}\operatorname{Cov}_{\sigma,\pi}(\theta_s, \theta_{s+\Delta}) &= \frac{1}{(C_s+1)^2(2C_s+1)(D-1)}\left\{C_f^2 C_s(C_s-1)\right. \\ &\qquad \left. -2C_s^2C_f^2 + C_s^2C_f(C_f-1)\right\} + O(D^{-2}) \\ &= -\frac{C_sC_f(C_s+C_f)}{(C_s+1)^2(2C_s+1)(D-1)} + O(D^{-2}). \end{aligned} \tag{43}$$

Now suppose that D is even and $\Delta = D/2$. The tuple $(x_i, x_{i+D/2})$ is accompanied by $(x_{i+D/2}, x_i)$. Let Q_Δ be the number of pairs of positions $\{i, i+D/2\}$ for which both entries are $\widetilde{O}$ or $\widetilde{\times}$. When $Q_\Delta = 1$, the possible values of Z and their changes are

entries in the pair	Z	$F(Z) - F(0,0,0,0)$
$(\widetilde{O}, \widetilde{O})$	$(2,0,0,0)$	$\dfrac{C_f^2}{C_s(C_s+1)^2}$
$(\widetilde{O}, \widetilde{\times})$	$(0,1,1,0)$	$-\dfrac{2C_sC_f+1}{(C_s+1)^2(2C_s+1)}$
$(\widetilde{\times}, \widetilde{\times})$	$(0,0,0,2)$	$\dfrac{2C_s^2}{(C_s+1)^2(2C_s+1)}$.

(44)

For example, a pair containing $\widetilde{O}$ and $\widetilde{\times}$ gives the two tuples $(\widetilde{O}, \widetilde{\times})$ and $(\widetilde{\times}, \widetilde{O})$, so $Z = (0,1,1,0)$. Substitution into (38) gives

$$\begin{aligned} F(0,1,1,0) - F(0,0,0,0) &= \frac{2C_f^2 - 2C_f}{2(C_s+1)^2} + \frac{2C_f - 2}{2(C_s+1)^2(2C_s+1)} - \frac{C_f^2}{(C_s+1)^2} \\ &= -\frac{2C_sC_f+1}{(C_s+1)^2(2C_s+1)}. \end{aligned}$$

There are $D/2$ pairs of positions of the form $\{i, i+D/2\}$. Sampling the two entries without replacement gives respective expectations

$$\frac{C_s(C_s-1)}{2(D-1)}, \quad \frac{C_sC_f}{D-1}, \quad \frac{C_f(C_f-1)}{2(D-1)} \tag{45}$$

for $(\widetilde{O}, \widetilde{O})$, $(\widetilde{O}, \widetilde{\times})$, and $(\widetilde{\times}, \widetilde{\times})$, respectively. The calculation in (22) with f replaced by $C_s + C_f$ gives

$$\mathbb{E}_\sigma(Q_\Delta(Q_\Delta - 1)) = O(D^{-2}).$$

Thus the cases with $Q_\Delta \geq 2$ affect the expectation only by $O(D^{-2})$. Multiplying the entries in (44) by the three expectations in (45) and adding them gives

$$\operatorname{Cov}_{\sigma,\pi}(\theta_s, \theta_{s+\Delta}) = -\frac{C_f(2C_s+C_f)}{2(C_s+1)(2C_s+1)(D-1)} + O(D^{-2}). \tag{46}$$

Lemma 5 gives the one shift variance

$$V_{\mathrm{ind}} = \frac{C_sC_f(C_s+C_f+1)}{(C_s+1)^2(C_s+2)}. \tag{47}$$

Dividing the leading term of (43) by (47) gives

$$\operatorname{Cov}_{\sigma,\pi}(\theta_s, \theta_{s+\Delta}) = -V_{\mathrm{ind}}\frac{(C_s+2)(C_s+C_f)}{(2C_s+1)(C_s+C_f+1)(D-1)} + O(D^{-2}), \tag{48}$$

while dividing (46) by (47) gives

$$\operatorname{Cov}_{\sigma,\pi}(\theta_s, \theta_{s+\Delta}) = -V_{\mathrm{ind}}\frac{(C_s+1)(C_s+2)}{2C_s(2C_s+1)}\frac{2C_s+C_f}{(C_s+C_f+1)(D-1)} + O(D^{-2}). \tag{49}$$

Equations (48) and (49) give the next two expressions in Lemma 3 when C_s and C_f are fixed.

Finally, we now suppose that both C_s/D and C_f/D are bounded away from zero. We continue to use Z from (39) and F from (38). The expectation of Z is

$$\mu = \mathbb{E}_\sigma(Z) = \frac{1}{D-1}\left(C_s(C_s-1),\ C_sC_f,\ C_sC_f,\ C_f(C_f-1)\right). \tag{50}$$

We use $P(A_1,\ldots,A_r)$ to denote the probability that a sequence of r draws without replacement from the location vector gives $A_1,\ldots,A_r$ in this order, where every A_i is $\widetilde{O}$, $\widetilde{\times}$, or $\widetilde{-}$. For example,

$$P(\widetilde{O},\widetilde{\times}) = \frac{C_sC_f}{D(D-1)}$$

For any two entries $A, B \in \{\widetilde{O},\widetilde{\times},\widetilde{-}\}$, let

$$Z_{AB} = \sum_{i=1}^{D} \mathbb{1}_{\{(x_i,x_{i+\Delta})=(A,B)\}}. \tag{51}$$

Thus Z_{AB} is the number of times (A,B) occurs among the tuples $(x_i, x_{i+\Delta})$.

Suppose first that D is odd or $\Delta \neq D/2$. There are D^2 terms after expanding $Z_{AB}Z_{CE}$ in (51). If the two starting positions are equal, there are D terms, and they are nonzero only if $A=C$ and $B=E$. If the second starting position is Δ after the first, there are D terms, and the common entry requires $B=C$. If the first starting position is Δ after the second, there are D terms, and the common entry requires $E=A$. The remaining $D(D-3)$ terms use four distinct positions. Adding these four cases gives

$$\begin{aligned}\mathbb{E}_\sigma(Z_{AB}Z_{CE}) &= D\mathbb{1}_{\{A=C,B=E\}}P(A,B) + D\mathbb{1}_{\{B=C\}}P(A,B,E)\\ &\quad + D\mathbb{1}_{\{E=A\}}P(C,A,B) + D(D-3)P(A,B,C,E).\end{aligned} \tag{52}$$

Now suppose that D is even and $\Delta = D/2$. There are D terms with the same starting position. There are another D terms in which $(x_j, x_{j+\Delta}) = (x_{i+\Delta}, x_i)$, which requires $A=E$ and $B=C$. The remaining $D(D-2)$ terms have four distinct positions. Therefore

$$\begin{aligned}\mathbb{E}_\sigma(Z_{AB}Z_{CE}) &= D\mathbb{1}_{\{A=C,B=E\}}P(A,B) + D\mathbb{1}_{\{A=E,B=C\}}P(A,B)\\ &\quad + D(D-2)P(A,B,C,E).\end{aligned} \tag{53}$$

Equations (52) and (53), together with their corresponding draw probabilities, give the sixteen second moments of the four entries in (39). For example, when D is odd or $\Delta \neq D/2$,

$$\mathbb{E}_\sigma(|\mathcal{L}_0|\,|\mathcal{L}_1|) = DP(\widetilde{O},\widetilde{O},\widetilde{\times}) + D(D-3)P(\widetilde{O},\widetilde{O},\widetilde{O},\widetilde{\times}),$$

because the second $\widetilde{O}$ in $(\widetilde{O},\widetilde{O})$ can be the first $\widetilde{O}$ in $(\widetilde{O},\widetilde{\times})$. When D is even and $\Delta = D/2$,

$$\mathbb{E}_\sigma(|\mathcal{L}_1|\,|\mathcal{H}_0|) = DP(\widetilde{O},\widetilde{\times}) + D(D-2)P(\widetilde{O},\widetilde{\times},\widetilde{\times},\widetilde{O}),$$

because $(\widetilde{O},\widetilde{\times})$ and $(\widetilde{\times},\widetilde{O})$ can be the two tuples formed from the same two positions.

Let $W = Z - \mu$. Taylor expansion of $F(\mu+W)$ and taking expectations give

$$\begin{aligned}\mathbb{E}_\sigma(F(Z)) &= F(\mu) + \frac{1}{2}\sum_{r,q=0}^{3} F_{rq}(\mu)\mathrm{Cov}_\sigma(Z_r,Z_q)\\ &\quad + \frac{1}{6}\sum_{r,q,v=0}^{3} F_{rqv}(\mu)\mathbb{E}_\sigma(W_rW_qW_v) + R_D.\end{aligned} \tag{54}$$

Here F_{rq} and F_{rqv} denote the respective partial derivatives. Because $|\mathcal{L}_0| \le C_s$, every denominator in (38) is at least of order C_s. Since C_s/D is bounded away from zero, the derivatives of orders two, three, and four are $O(D^{-2})$, $O(D^{-3})$, and $O(D^{-4})$, respectively. Expanding products of three and four sums of the form (51), and separating the common index cases in the same way as in (52) and (53), gives

$$\mathbb{E}_\sigma(W_rW_qW_v) = O(D), \quad \mathbb{E}_\sigma\left(\|W\|^4\right) = O(D^2). \tag{55}$$

It follows from (55) that the third order term in (54) is $O(D^{-2})$. The fourth order Taylor remainder satisfies $R_D = O(D^{-2})$ for the same reason. Thus, up to $O(D^{-2})$, only $F(\mu)$ and the covariance term remain in (54).

Now, we can use a computer to find (54). First, compute every covariance term as

$$\mathrm{Cov}_\sigma\left(Z_{AB}, Z_{CE}\right) = \mathbb{E}_\sigma\left(Z_{AB} Z_{CE}\right) - \mathbb{E}_\sigma\left(Z_{AB}\right)\mathbb{E}_\sigma\left(Z_{CE}\right),$$

using (50) with (52) or (53). Substitute these covariances into the second term of (54), use $C_n = D - C_s - C_f$, and collect the terms of order D^{-1}. This gives

$$\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = \begin{cases} -\dfrac{C_f\{C_s(2D - C_s) + C_f(D - C_s)\}}{C_s^2(2D - C_s)^2} + O(D^{-2}), & D \text{ even and} \\ & \Delta = D/2, \\ -\dfrac{C_f(C_s + C_f)}{C_s^2(2D - C_s)} + O(D^{-2}), & \text{otherwise.} \end{cases} \tag{56}$$

Equation (56) follows by combining (50), the respective draw probabilities, the second moments (52) and (53), and the expansion (54), and noting that $\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = \mathbb{E}_\sigma\left(F(Z)\right) - \left(\frac{C_f}{C_s+1}\right)^2$.

Next, the variance in (47) satisfies

$$V_{\mathrm{ind}} = \frac{C_f(C_s + C_f)}{C_s^2} + O(D^{-1}), \tag{57}$$

$$\frac{V_{\mathrm{ind}}}{C_s + C_f} = \frac{C_f}{C_s^2} + O(D^{-2}). \tag{58}$$

Substituting (57) and (58) into (56) gives

$$\mathrm{Cov}_{\sigma,\pi}\left(\theta_s, \theta_{s+\Delta}\right) = \begin{cases} -V_{\mathrm{ind}}\dfrac{C_s(2D - C_s) + C_f(D - C_s)}{(C_s + C_f)(2D - C_s)^2} + O(D^{-2}), & D \text{ even and} \\ & \Delta = D/2, \\ -V_{\mathrm{ind}}\dfrac{1}{2D - C_s} + O(D^{-2}), & \text{otherwise.} \end{cases}$$

□

B.3 Additional Experiments to Verify Variance Approximation

Since the Minner estimator depends on $C_s = f_2$ and $C_f = f_1 - a$, we investigate edge cases when $f_1 - a$, f_2 are small relative to D, as D increases. All experiments are done over 1,000 iterations (an order of magnitude less than the plots in Appendix A.5, since dealing with large K means more expensive permutations). We can see that the theoretical variance approximations match the empirical MSEs well, as in Figures 9-13.

In the legend of these plots, `Counts MSE` denotes the empirical MSE of the Minner count estimator, `Counts Var` denotes the theoretical variance of the Minner count estimator, `CP Counts MSE` denotes the empirical MSE of the count estimator with circulant permutations, `CP Counts Var (S)` denotes the theoretical variance approximation when $\frac{C_s+C_f}{D} \equiv \frac{f_1+f_2-a}{D} \to 0$, and `CP Counts Var (NS)` denotes the theoretical variance approximation when $\frac{f_1+f_2-a}{D}$ does not tend to zero.

Since we are looking at counts when $\frac{f_1+f_2-a}{D}$ is small, `CP Counts Var (S)` should be a better approximation. However, we find that both `CP Counts Var (S)` and `CP Counts Var (NS)` are equally good estimators as they are indistinguishable from each other.

Unlike the experiments in Appendix A.5, where the theoretical variance curves are almost indistinguishable from the empirical MSEs, we do not expect the count approximations here to overlap perfectly for moderate D, especially when $\frac{f_1+f_2-a}{D}$ is small, given that we use a log scale for the y axes in these plots. However, the approximation is still accurate for practical comparisons. In particular, `Counts Var` gives a useful reference for the amount of Monte Carlo discrepancy expected from 1,000 iterations.

We also plot the theoretical approximation and the empirical MSE for select word pairs from the words dataset [8] in Figure 14.

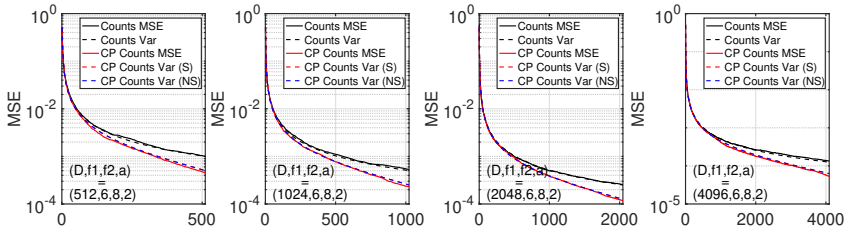

Figure 9: Empirical MSE vs theoretical approximate variance in Theorem 4 at edge cases when f_i are small respective to D. We use $(f_1, f_2, a) = (6, 8, 2)$.

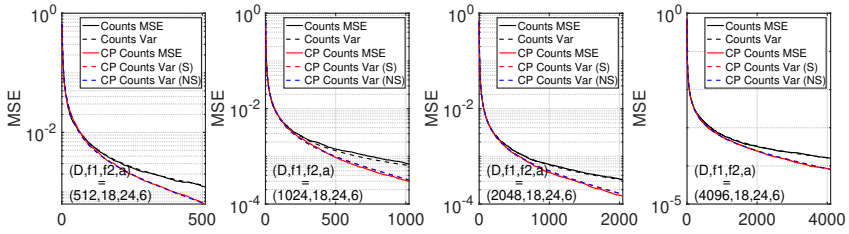

Figure 10: Empirical MSE vs theoretical approximate variance in Theorem 4 at edge cases when f_i are small respective to D. We use $(f_1, f_2, a) = (18, 24, 6)$.

Both the theoretical approximate variance curves overlaps with the empirical MSEs for all word pairs, even the HONG-KONG word pair, where $\frac{f_1 - a}{D} = \frac{41}{65536} \approx 6.3 \times 10^{-4}$ is small.

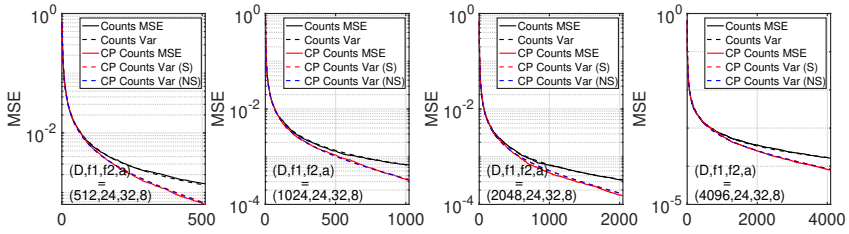

Figure 11: Empirical MSE vs theoretical approximate variance in Theorem 4 at edge cases when f_i are small respective to D. We use $(f_1, f_2, a) = (24, 32, 8)$.

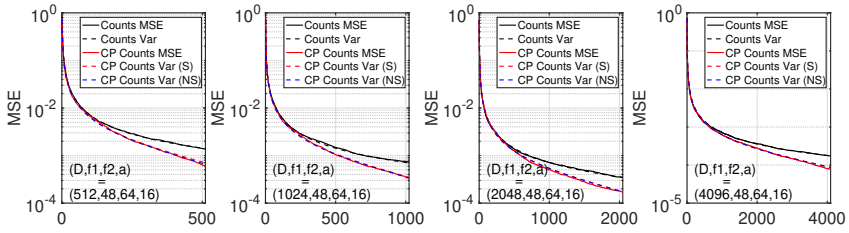

Figure 12: Empirical MSE vs theoretical approximate variance in Theorem 4 at edge cases when f_i are small respective to D. We use $(f_1, f_2, a) = (48, 64, 16)$.

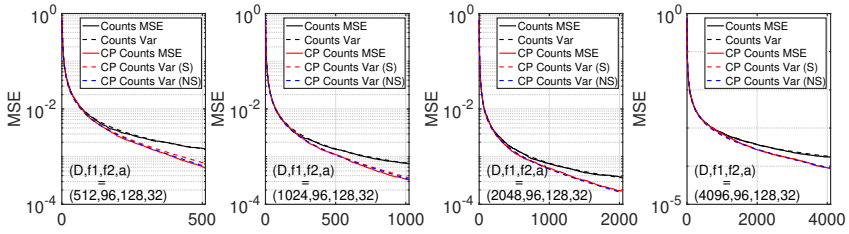

Figure 13: Empirical MSE vs theoretical approximate variance in Theorem 4 at edge cases when f_i are small respective to D. We use $(f_1, f_2, a) = (96, 128, 32)$.

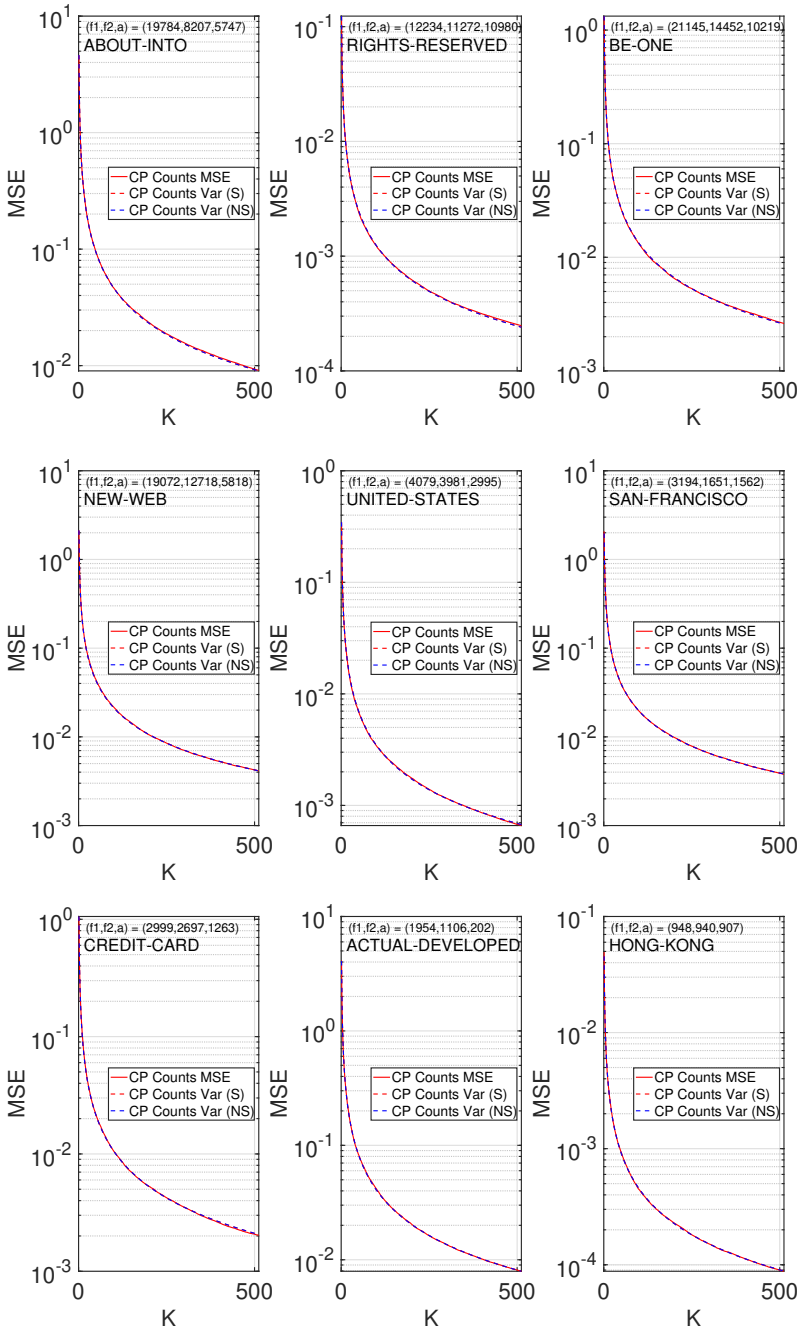

Figure 14: Empirical MSE vs theoretical approximate variance in Theorem 4 for selected word pairs. Both theoretical approximations and empirical MSE lines overlap perfectly.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract claims a new approach to approximate the variance of MinHash under circulant permutations. We demonstrate this in Section 2. We also claim that we describe how to apply our approach to different MinHash estimators. We describe how to do so in Section 4 for asymmetric hashing (Minner estimator).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are in the last section of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: A full and complete proof for all lemmas, theorems, and corollaries are in the appendix. The assumption that D is large is clearly stated.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our experiments verify if our approximate variances are close to the empirical MSEs of existing estimators (CMinHash, and circulant permutation applied to the Minner estimator). We review how circulant permutations work in Appendix A1, and give the approximate variances in Theorems 2 and 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our data is mostly synthetic, and the words dataset is an existing dataset used by [13]. We describe our experiments clearly in Sections 2 and 3. Our experiments are mostly similar to [13], hence code is not provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Our data is mostly synthetic, and the words dataset is an existing dataset used by [13]. We describe our experiments clearly in Sections 2 and 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Because we are showing an approximation (which is a fixed value given f_1, f_2, a, D) of the actual variance of estimators, no error bars are needed. We compute the respective MSEs for MinHash and Minner estimates averaged over 10,000 iterations. If our approximation was correct, the respective average of the empirical MSEs should overlap with our theoretical variance. Any confidence intervals (or error bars) around the average of the 10,000 empirical MSEs would be of the form $\overline{\text{MSE}} \pm \epsilon$. But if our approximate variance overlap the empirical MSEs, then the value of ϵ doesn't matter in this case (in fact, with 10,000 iterations, any ϵ would be sufficiently small to be mathematically negligible for visual comparison).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments mainly verify that our approximate variance matches empirical results, and no large scale compute resources are needed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work serves to make the analysis of the variance of randomized algorithms involving circulant permutations easier.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This is a theoretical paper.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the algorithm given and use the Words dataset in [13]. To the best of our knowledge, the words dataset first appears in [7].

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our approach is theoretical.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not use human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not use human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Originally, we did not use an LLM for our research (apart from formatting, phrasing, etc). However, during our rebuttal period (also mentioned on Page 13), we asked 5.6 SOL for a combinatorial implementation of Minner, in order to better craft a rebuttal for our reviewers. It turned out that Li and Li's variance proof for CMinHash had a slight error. We included the counter example in the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.