A. Proofs

Theorem 5 *Give any two finite sets $S_1, S_2 \in \Omega$, with $A = |S_1 \cup S_2| > a = |S_1 \cap S_2| > 0$ and $|\Omega| = D \to \infty$. The limiting variance of the estimators from densification and improved densification when $k = D \to \infty$ is given by:*

$$\lim_{k\to\infty} Var(h) = \frac{a}{A}\left[\frac{A-a}{A(A+1)}\right] > 0 \quad (15)$$

$$\lim_{k\to\infty} Var(h^+) = \frac{a}{A}\left[\frac{3(A-1)+(2A-1)(a-1)}{2(A+1)(A-1)} - \frac{a}{A}\right] > 0 \quad (16)$$

Proof: When $k = D$, then $N_{emp} = D - A$. Substituting this value in the variance formulas from (Shrivastava & Li, 2014c) and taking the limit as $D = k \to \infty$, we get the above expression after manipulation. When $0 < R = \frac{a}{A} < 1$, they both are strictly positive. □

Theorem 6

$$Pr\big(h^*(S_1) = h^*(S_2)\big) = \frac{|S_1 \cap S_2|}{|S_1 \cap S_2|} = R \quad (17)$$

$$Var(h^*) = \frac{R}{k} + A\frac{R}{k^2} + B\frac{R\bar{R}}{k^2} - R^2 \quad (18)$$

$$\lim_{k\to\infty} Var(h^*) = 0 \quad (19)$$

where N_{emp} is the number of simultaneous empty bins between S_1 and S_2 and the quantities A and B are given by

$$A = \mathbb{E}\left[2N_{emp} + \frac{N_{emp}(N_{emp}-1)}{k-N_{emp}}\right]$$

$$B = \mathbb{E}\Bigg[(k-N_{emp})(k-N_{emp}-1) + 2N_{emp}(k-N_{emp}-1) + \frac{N_{emp}(N_{emp}-1)(k-N_{emp}-1)}{k-N_{emp}}\Bigg]$$

Proof:

The collision probability is easy using a simple observation that values coming from different bin numbers can never match across S_1 and S_2, i.e. $h^*_i(S_i) \neq h^*_j(S_2)$ if $i \neq j$, as they have disjoint different range. So whenever, for a simultaneous empty bin i, i.e. $E_i = 1$, we get $h^*_i(S_1) = h^*_i(S_2)$ after reassignment, the value must be coming from same non-empty bin, say numbers k which is not not empty. Thus,

$$Pr(h^*_i(S_1) = h^*_i(S_2)) = Pr(h^*_k(S_1) = h^*_k(S_2)|E_k = 0) = R$$

The variance is little involved. From the collision probability, we have the following is unbiased estimator.

$$\hat{R} = \frac{1}{k}\sum_{j=0}^{k-1} \mathbb{1}\{h^*_j(S_1) = h^*_j(S_2)\}. \quad (20)$$

For variance, define the number of simultaneously empty bins by

$$N_{emp} = \sum_{j=0}^{k-1} \mathbb{1}\{E_j = 1\}, \quad (21)$$

where $\mathbb{1}$ is the indicator function. We partition the event $\big(h^*_j(S_1) = h^*_j(S_2)\big)$ into two cases depending on E_j. Let M^N_j (**Non-empty Match at** j) and M^E_j (**Empty Match at** j) be the events defined as:

$$M^N_j = \mathbb{1}\{E_j = 0 \text{ and } h^*_j(S_1) = h^*_j(S_2)\} \quad (22)$$

$$M^E_j = \mathbb{1}\{E_j = 1 \text{ and } h^*_j(S_1) = h^*_j(S_2)\} \quad (23)$$

Note that, $M^N_j = 1 \implies M^E_j = 0$ and $M^E_j = 1 \implies M^N_j = 0$. From the LSH property of estimator we have

$$\mathbb{E}(M^N_j|E_j = 0) = \mathbb{E}(M^E_j|E_j = 1) = \mathbb{E}(M^E_j + M^N_j) = R \ \ \forall j \quad (24)$$

It is not difficult to show that,

$$\mathbb{E}\left(M^N_j M^N_i \middle| i \neq j, E_j = 0 \text{ and } E_i = 0\right) = R\tilde{R},$$

where $\tilde{R} = \frac{a-1}{f1+f2-a-1}$. Using these new events, we have

$$\hat{R} = \frac{1}{k}\sum_{j=0}^{k-1}\left[M^E_j + M^N_j\right] \quad (25)$$

We are interested in computing

$$Var(\hat{R}) = \mathbb{E}\left(\left(\frac{1}{k}\sum_{j=0}^{k-1}\left[M^E_j + M^N_j\right]\right)^2\right) - R^2 \quad (26)$$

For notational convenience we will use m to denote the event $k - N_{emp} = m$, i.e., the expression $\mathbb{E}(.|m)$ means $\mathbb{E}(.|k - N_{emp} = m)$. To simplify the analysis, we will first compute the conditional expectation

$$f(m) = \mathbb{E}\left(\left(\frac{1}{k}\sum_{j=0}^{k-1}\left[M^E_j + M^N_j\right]\right)^2 \middle| m\right) \quad (27)$$

By expansion and linearity of expectation, we obtain

$$k^2 f(m) = \mathbb{E}\left[\sum_{i\neq j} M^N_i M^N_j \middle| m\right] + \mathbb{E}\left[\sum_{i\neq j} M^N_i M^E_j \middle| m\right] + \mathbb{E}\left[\sum_{i\neq j} M^E_i M^E_j \middle| m\right] + \mathbb{E}\left[\sum_{i=1}^{k}\left[(M^N_j)^2 + (M^E_j)^2\right] \middle| m\right]$$

$M_j^N = (M_j^N)^2$ and $M_j^E = (M_j^E)^2$ as they are indicator functions and can only take values 0 and 1. Hence,

$$\mathbb{E}\left[\sum_{j=0}^{k-1}\left[(M_j^N)^2 + (M_j^E)^2\right]\middle|m\right] = kR \tag{28}$$

The values of the first three terms are given by the following 3 expression using simple binomial enpension and using the fact that we are dealing with indicator random variable which can only take values 0 or 1.

$$\mathbb{E}\left[\sum_{i\neq j} M_i^N M_j^N\middle|m\right] = m(m-1)R\tilde{R} \tag{29}$$

$$\mathbb{E}\left[\sum_{i\neq j} M_i^N M_j^E\middle|m\right] = 2m(k-m)\left[\frac{R}{m} + \frac{(m-1)R\tilde{R}}{m}\right] \tag{30}$$

Let p be the probability that two simultaneously empty bins i and j finally picks the same non-empty bin for reassignment. Then we have

$$\mathbb{E}\left[\sum_{i\neq j} M_i^E M_j^E\middle|m\right] = (k-m)(k-m-1)\left[pR + (1-p)R\tilde{R}\right] \tag{31}$$

because with probability $(1-p)$, it uses estimators from different simultaneous non-empty bin and in that case the $M_i^E M_j^E = 1$ with probability $R\tilde{R}$. We know that Algorithm 1 which uses 2-universal hashing the value of $p = \frac{1}{m}$. This is because any pairwise assignment is perfectly random with 2-universal hashing.

Substituting for all terms with value of p and rearranging terms gives the required expression.

When $k = D$, then $N_{emp} = D - A$. Substituting this value in the variance formulas and taking the limit as $D = k \to \infty$, we get 0 for all R.

Theorem 7

$$Var(h^*) \leq Var(h^+) \leq Var(h) \tag{32}$$

Proof: We have $p* = \frac{1}{m} \leq p^+ = \frac{1.5}{m+1} \leq p = \frac{2}{m+1}$. The value of p^+ and p comes from analysis in (Shrivastava & Li, 2014c)

Theorem 8 *Among all densification schemes, where the reassignment process for bin i is independent of the reassignment process of any other bin j, Algorithm 1 achieves the best possible variance.*

Under any independent re-assignment, the probability that two empty bins chooses the same non-empty bin out of m non-empty bins is lower bounded by $\frac{1}{m}$ which is achieved by optimal densification.