A BF Sketches for Single Sets

We provide extended results for BF for single sets.

A.1 Enhancing the Estimator by Swamidass [113]

The estimator by Swamidass et al. [113], is *divergent*[6] in its original form. To alleviate this, we replace $B_{X,1}$ with $\widetilde{B}_{X,1} \equiv B_{X,1} - \mathbb{I}[B_{X,1} = B_X]$, where, for a given proposition P, $\mathbb{I}[P]$ is 1 if P holds, and 0 otherwise. $\widetilde{B}_{X,1}$ only differs from $B_{X,1}$ in the unlikely case of $B_{X,1} = B_X$. Thanks to this modification, our estimator $\widehat{|X|}$ has, unlike Swamidass et al.'s, a finite expectation (as it is bounded).

The final form of the estimator is

$$\widetilde{\widehat{|X|}} = -\frac{B_X}{b} \log\left(1 - \frac{\widetilde{B_{X,1}}}{B_X}\right)$$

A.2 Proof of consistency and asymptotic unbiasedness

We need to show that $\widehat{|X|}_S = -\frac{B_X}{b} \log\left(1 - \frac{B_{X,1}}{B_X}\right)$ is consistent and asymptotically unbiased as $B_X \to \infty$. We provide here an intuitive formulation based on the false positive probability which can be easily made more rigorous by direct application of the definition of consistency which we omit for the sake of simplicity. First of all, as shown in eq.(4), we can notice that $\widehat{|X|}_S \sim \widehat{|X|}_L$ as the Bloom Filter size diverges. This means that the proof is valid for both estimators because they are asymptotically equivalent. Now we can look at the probability of false positives as $B_X \to \infty$ for fixed and finite b and $|X|$:

$$\lim_{B_X \to \infty} \left[1 - \left(1 - \frac{1}{B_X}\right)^{b|X|}\right]^b \to 0$$

The result above tells us that false positive matches cannot happen anymore in the limit. Each element of $|X|$ will then be hashed in a *personal* bit and counting the number of ones in B_X (and dividing by b in case of multiple hash functions) will always deliver $|X|$ at a given precision as $|X|$ is fixed and $B_X \to \infty$. Thus we can conclude that $\frac{B_{X,1}}{b} \xrightarrow{p} |X|$ which proves consistency. Asymptotic unbiasedness follows from consistency in our case as the variance of both estimators is bounded (see the proof of Proposition 1). The same reasoning can be easily extended to show consistency and asymptotic unbiasedness also for $\widehat{|X \cap Y|}_{AND}$ and $\widehat{|X \cap Y|}_{OR}$ presented in section 5.1.

A.3 Proposition 1

Proof. We now prove Proposition 1 from Section 4. Before bounding the mean squared error, we need to prove several simple bounds. Let $\mu = E[B_{0,X}] = B_X \left(1 - \frac{1}{B_X}\right)^{b|X|}$. It holds:

$$\begin{aligned}\mu \geq B_X \left(1 - \frac{1}{B_X}\right)^{0.499 B_X \log B_X} &\geq B_X \exp(-\frac{0.499 \log B_X}{1 - 1/B_X}) \\ &= B_X^{0.501 - o(1)}\end{aligned}$$

Let us fix some $\varepsilon > 0$. Let $\mathcal{E}$ be the event that $B_{0,X} \geq \mu/(1+\varepsilon)$. Kamath et al. [76, Theorem 2] prove that:
$P(\mathcal{E}) \geq 1 - \exp(-\Omega(\mu^2/B_X)) \geq 1 - \exp(-B_X^{\Omega(1)})$.
We have $\widehat{|X|} = -\frac{B_X}{b} \log(B_{X,0}/B_X + \mathbb{I}[B_{X,0} = 0]) \leq B_X \log B_X$ and by our assumption, $|X| \leq b|X| \leq 0.499 B_X \log B_X$. It thus holds $(\widehat{|X|} - |X|)^2 \leq O(B_X^2 \log^2 B_X)$. Let $\kappa = -\frac{B_X}{b} \log\left(1 - \frac{1}{B_X}\right)^{b|X|} = -B_X |X| \log\left(1 - \frac{1}{B_X}\right)$. Moreover for $B_X \to \infty$, we have $\log(1 - 1/B_X) = -1/B_X + O(1/B_X^2)$. Therefore, it holds $\kappa = |X| + o(1)$. Now we can bound the mean squared error as follows:

$$\begin{aligned}
&E[(\widehat{|X|} - |X|)^2] && (15)\\
&= E[(\widehat{|X|} - |X|)^2 | \mathcal{E}] P(\mathcal{E}) + E[(\widehat{|X|} - |X|)^2 | \neg\mathcal{E}] P(\neg\mathcal{E}) && (16)\\
&\leq (1+\varepsilon) E[(\widehat{|X|} - \kappa)^2 | \mathcal{E}] + \frac{1+\varepsilon}{\varepsilon} E[(\kappa - |X|)^2 | \mathcal{E}] && (17)\\
&\quad + O(B_X^2 \log^2 B_X) \cdot \exp(-B_X^{\Omega(1)}) && (18)\\
&\leq \frac{(1+\varepsilon) B_X^2}{b^2} E[(\log(B_{X,0}/B_X) - \log(1 - 1/B_X)^{b|X|})^2 | \mathcal{E}] && (19)\\
&\quad + O((\kappa - |X|)^2) + \exp(-B_X^{\Omega(1)}) && (20)\\
&\leq \frac{(1+\varepsilon) B_X^2}{b^2} E[(\log(B_{X,0}/B_X) - \log(1 - 1/B_X)^{b|X|})^2 | \mathcal{E}] && (21)\\
&\quad + O(|X|/B_X) && (22)\\
&\leq \frac{(1+\varepsilon)^2 B_X^2}{b^2} e^{2b|X|/B_X} E[(B_{X,0}/B_X - (1 - 1/B_X)^{b|X|})^2 | \mathcal{E}] && (23)\\
&\quad + O(|X|/B_X) && (24)\\
&\leq \frac{(1+\varepsilon)^2 B_X^2}{b^2} e^{2b|X|/(B_X - 1)} && (25)\\
&\quad \cdot E[(B_{X,0}/B_X - (1 - 1/B_X)^{b|X|})^2] / P[\mathcal{E}] + O(|X|/B_X) && (26)\\
&= \big((1+\varepsilon)^2 + o(1)\big) \frac{B_X^2}{b^2} e^{2b|X|/(B_X - 1)} && (27)\\
&\quad \cdot E[(B_{X,0}/B_X - (1 - 1/B_X)^{b|X|})^2] + O(|X|/B_X) && (28)\\
&= \big((1+\varepsilon)^2 + o(1)\big) \frac{e^{2b|X|/(B_X - 1)}}{b^2} Var[B_{X,0}] + O(|X|/B_X) && (29)\\
&\leq \big((1+\varepsilon)^2 + o(1)\big) e^{2b|X|/(B_X - 1)} && (30)\\
&\quad \cdot \left(e^{-\frac{b|X|}{B_X}} \frac{B_X}{b^2} - B_X/b^2 - |X|/b\right) + O(|X|/B_X) && (31)\\
&\leq \big((1+\varepsilon)^2 + o(1)\big) \left(e^{|X|b/(B_X - 1)} \frac{B_X}{b^2} - B_X/b^2 - |X|/b\right) && (32)\\
&\quad + O(|X|/B_X) && (33)\\
&\leq \big((1+\varepsilon)^2 + o(1)\big) \left(e^{|X|b/(B_X - 1)} \frac{B_X}{b^2} - B_X/b^2 - |X|/b\right) && (34)
\end{aligned}$$

[6] An estimator whose moments are not finite. In the case of the estimator Swamidass et al. [113], the expectation of $\widehat{|X|}$, and thus also the higher moments, diverge, which happens for $B_{X,1} = B_X$

where eq. (18) holds because for any $a, b, c \in \mathbb{R}$ and $\varepsilon > 0$, it holds[7] $(a-b)^2 \leq (1+\varepsilon)(a-c)^2 + \frac{1+\varepsilon}{\varepsilon}(c-b)^2$. Eq. (23) holds because on $\mathcal{E}$, given $B_{X,0} \geq \mu/(1+\varepsilon)$, $\log(B_{X,0}/B_X)$ is c-lipschitz for $c = (1+\varepsilon)B_X/\mu \leq (1+\varepsilon)e^{\frac{2b|X|}{B_X(1-1/B_X)}} = (1+\varepsilon)e^{\frac{2b|X|}{B_X-1}}$. Eq. (29) holds because $E[B_{X,0}/B_X] = (1-1/B_X)^{b|X|}$ and eq. (31) holds because $Var(B_{X,0}) \sim B_X e^{-\frac{b|X|}{B_X}} - B_X\left(\frac{b|X|}{B_X}+1\right)e^{-\frac{2b|X|}{B_X}}$ [71]. By sending $\varepsilon \to 0$, we get that[8]:

$$E[(\widehat{|X|} - |X|)^2] \leq (1+o(1))\left(e^{|X|b/(B_X-1)}\frac{B_X}{b^2} - B_X/b^2 - |X|/b\right)$$

□

A.4 Proposition 2

Proof. We now prove Proposition 2 from Section 4. We start by the well known MSE decomposition:

$$E\left[\left(\widehat{|X|}_\bullet - |X|\right)^2\right] = E\left[\left(\widehat{|X|}_\bullet - |X|\right)\right]^2 + Var(\widehat{|X|}_\bullet) \tag{35}$$

Now notice that $E[B_{0,X}] = B_X\left(1-\frac{1}{B_X}\right)^{b|X|}$. Then, since $\widehat{|X|}_\bullet = \delta_{B_X,b}\, B_{X,1}$, we can easily derive:

$$\begin{aligned} E\left[\delta_{B_X,b}\, B_{X,1}\right] &= E\left[\delta_{B_X,b}\, (B_X - B_{X,0})\right] \\ &= \delta_{B_X,b}\, B_X\left[1-\left(1-\frac{1}{B_X}\right)^{b|X|}\right] \end{aligned}$$

On the other hand, to bound the variance of the simplified estimator, we follow the same reasoning outlined in Proposition 1. Indeed it holds that $Var(B_{X,0}) \sim B_X\left[e^{-\frac{|X|b}{B_X}} - \left(1+\frac{|X|b}{B_X}\right)e^{-\frac{2|X|b}{B_X}}\right]$ as shown in [71]. Now notice that $Var(B_{X,1}) = Var(B_X - B_{X,0}) = Var(B_{X,0})$. At this point we can substitute in eq. (35) the squared bias and variance of $\widehat{|X|}_\bullet$ to conclude that:

$$\begin{aligned} E\left[\left(\widehat{|X|}_\bullet - |X|\right)^2\right] \leq &\left\{|X| - \delta_{B_X,b}\, B_X\left[1-\left(1-\frac{1}{B_X}\right)^{b|X|}\right]\right\}^2 \\ &+\delta^2_{B_X,b}\, B_X\left[e^{-\frac{|X|b}{B_X}} - \left(1+\frac{|X|b}{B_X}\right)e^{-\frac{2|X|b}{B_X}}\right] \end{aligned}$$

which ends the proof. To enhance the interpretation of the bound, we use that fact that $\left(1-\frac{1}{B_X}\right)^{b|X|} \sim e^{-\frac{|X|b}{B_X}}$ in the statement of Proposition 2.

□

B MinHash Sketches for Set Intersection

B.1 Expectation formula

Since in the case of k-hash, $|M_X \cap M_Y| \sim Bin(\ k\ ,\ J_{X,Y}\ )$, and for 1-hash, $|M^1_X \cap M^1_Y| \sim Hypergeometric(|X \cup Y|, |X \cap Y|, k)$, we have:

$$\mathbb{E}[\widehat{|X \cap Y|}_{kH}] = (|X|+|Y|)\sum_{s=0}^{k}\binom{k}{s}(J_{X,Y})^s(1-J_{X,Y})^{k-s}\frac{s}{k+s} \tag{36}$$

$$\mathbb{E}[\widehat{|X \cap Y|}_{1H}] = (|X|+|Y|)\sum_{s=0}^{k}\frac{\binom{|X\cap Y|}{s}\binom{|X\cup Y|-|X\cap Y|}{k-s}}{\binom{|X\cup Y|}{k}}\frac{s}{k+s} \tag{37}$$

There exists an involved closed form expression for equation (36) which is beyond the scope of this paper. We refer the interested reader to [41] for a clear derivation of a similar problem.

B.2 Proof of consistency and asymptotic unbiasedness

We start to show that $\widehat{|X \cap Y|}_{kH}$ is consistent. This follows respectively from Proposition 4 statement. Indeed by taking the limit for $k \to \infty$ with fixed and finite $|X|$ and $|Y|$ we obtain:

$$\lim_{k\to\infty} P\left(\left|\widehat{|X \cap Y|}_{kH} - |X \cap Y|\right| \geq t\right) \leq \lim_{k\to\infty} 2e^{-\frac{k\,t^2}{2(|X|+|Y|)^2}} \to 0$$

The above implies that $\widehat{|X \cap Y|}_{kH} \xrightarrow{p} |X \cap Y|$. On the other hand, for $\widehat{|X \cap Y|}_{1H}$ we are in the *sampling without replacement* scheme. This means that the population size (i.e. $|X \cup Y|$) is finite and by taking the limit for $k \to |X \cup Y|$ in Proposition 5, with fixed and finite $|X|$ and $|Y|$, we have already sampled the entire population contrarily to the k-Hash case. Thus $\widehat{|X \cap Y|}_{1H}$ is also a consistent estimator of $|X \cap Y|$. Then, for both estimators, the asymptotic unbiasedness follows from consistency and by noticing that both $\widehat{|X \cap Y|}_{kH}$ and $\widehat{|X \cap Y|}_{1H}$ have a bounded variance.

B.3 Sub-Gaussian preliminaries

We recall some key notions of sub-gaussian random variables as they are necessary for the following proofs. First of all, we define $\psi_X(\lambda) = \log(\mathbb{E}[e^{\lambda X}])$ as the logarithmic moment generating function (i.e. cumulant) of a generic random variable X. For example, if Z is a centered normal random variable with variance σ^2, we have that $\psi_Z(\lambda) = \frac{\lambda^2\sigma^2}{2}$. It can be shown, we refer the interested reader to chapter 2 of [35] for a detailed explanation, that Chernoff's inequality in this case implies, for all $t > 0$, that:

$$P(Z \geq t) \leq e^{-\frac{t^2}{2\sigma^2}} \tag{38}$$

The bound above, characterize the decay of the tail probabilities of a centered normal random variable. If the tail probabilities of a generic centered random variable X, decrease at least as rapidly as the ones in (38) then X is *sub-gaussian*. More formally, a centered random variable X is said to be *sub-gaussian* with variance factor σ^2 if:

$$\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2} \qquad \forall\lambda \in \mathbb{R} \tag{39}$$

We underline that (39) only requires $Var(X) \leq \sigma^2$. Moreover, if we call $\mathcal{G}(\sigma^2)$ the collection of random variables for which (39) holds (e.g. all bounded random variables belongs to $\mathcal{G}(\sigma^2)$), we can state that:

[7] This inequality is equivalent to $(1+\varepsilon)(a-c)^2 + \frac{1+\varepsilon}{\varepsilon}(c-b)^2 - (a-b)^2 \geq 0$. The left-hand side can be simplified to $\frac{(\varepsilon a+b-c(1+\varepsilon))^2}{\varepsilon}$ and the inequality thus holds.

[8] It is well known that if $f(x) \leq (1+\varepsilon)g(x)$ for any $\varepsilon > 0$, then $f(x) \leq (1+o(1))g(x)$.

Lemma B.1. *Let $X_1, \ldots, X_n$ be independent random variables, such that for every i we have $X_i \in \mathcal{G}(\sigma_i^2)$. Then $\sum_{i=1}^n X_i \in \mathcal{G}(\sum_{i=1}^n \sigma_i^2)$.*

This is due to the fact that (39) implies a bound on the moment generating function whose properties help to verify the above statement. For a more detailed discussion on B.1 and for alternative characterizations of sub-gaussianity in terms of growth of moments, we refer the interested reader to chapter 2 of [35].

B.4 Concentration bounds for k-Hash and 1-Hash

We present below the proof of Propositions 4 and 5. First, we show to following lemma which we will also use later.

Lemma B.2.

$$P(|\hat{J}_1 - J| \geq t), P(|\hat{J}_k - J| \geq t) \leq 2e^{-2t^2 k} \tag{40}$$

Proof. The random variables $k\hat{J}_1$ and $k\hat{J}_k$ follow the hypergeometric and binomial distributions, respectively. Applying the Hoeffding's inequalities in the binomial case, we get the desired inequality. The Serfling's bound can be applied in the case of the hypergeometric distribution. The Serfling's bound always gives better bounds than the Hoeffding's, proving the inequality for $\hat{J}_1$. □

We can now show concentration of the sum of the estimators and, therefore also of the individual estimators (by fixing $n = 1$).

Theorem B.3. *Let $Y_1 = \sum_i^n C_i \frac{\hat{J}_1}{1+\hat{J}_1}$, $Y_k = \sum_i^n C_i \frac{\hat{J}_k}{1+\hat{J}_k}$. Then for any non-negative constants C_i and $S = \sum_{i=1}^n C_i \frac{J}{1+J}$*

$$P(|Y_1 - S| > t), P(|Y_k - S| > t) \leq 2\exp\left(-\frac{kt^2}{2(\sum_i^n C_i)^2}\right) \tag{41}$$

Proof. The function $\frac{X}{1+X}$ is 1-Lipschitz and it, therefore, holds that

$$\left|\frac{X}{1+X} - \frac{X'}{1+X'}\right| \leq |X - X'|$$

The concentration results from Lemma B.2 therefore also hold for $\frac{\hat{J}_1}{1+\hat{J}_1}$ and $\frac{\hat{J}_k}{1+\hat{J}_k}$. The random variables

$$\frac{\hat{J}_1}{1+\hat{J}_1} - \frac{J}{1+J}$$
$$\frac{\hat{J}_k}{1+\hat{J}_k} - \frac{J}{1+J}$$

are therefore subgaussian with coefficient $1/\sqrt{k}$ (see eq. (38), (39) and in general section B.3). Multiplying by C_i's and taking the sum, we get that $Y_1 - S$ and $Y_k - s$ are subgaussian with coefficients

$$\frac{\sum_{i=1}^n C_i}{\sqrt{k}}$$

The theorem follows by Lemma B.1. □

C KMV Sketches for Single Sets

We also state an existing result on the KMV sketching; we use it later to provide a KMV sketch for $|X \cap Y|$ [46]. The hash function used with a KMV maps elements from X to real numbers in $(0, 1]$ u.a.r.[9]. Thus, the hashes should be evenly spaced and one can estimate $|X|$ by dividing the size $k - 1$ of K_X by the largest hash in K_X.

$$\widehat{|X|}_K = \frac{k-1}{\max\{x | x \in K_X\}} \tag{42}$$

As noted in [46, §2.1], the k-th smallest value follows the beta distribution $Beta(\alpha, \beta)$ with shape parameters $\alpha = k$ and $\beta = |X| - k + 1$. Now we can get concentration bounds for the estimator: indeed, similarly to [28], we can show that:

Proposition 6. *Consider $\widehat{|X|}_K$ in Eq. (42), then the probability of deviation from the true set size, at a given distance $t \geq 0$, is*

$$P\left(\left|\widehat{|X|} - |X|\right| \leq t\right) = I_{u(|X|,k,t/|X|)}(k, |X| - k + 1) - I_{l(|X|,k,t/|X|)}(k, |X| - k + 1)$$

where $u(|X|, k, t/|X|) = \frac{k-1}{|X|-t}$ and $l(|X|, k, t/|X|) = \frac{k-1}{|X|+t}$ and $I_x(a, b)$ is the regularized incomplete beta function.

In the case of a KMV estimator bound, we can evaluate:

$$I_x(k, |X| - k + 1) = \sum_{i=k}^{|X|} \binom{|X|}{i} x^i (1-x)^{|X|-i}$$

D KMV Sketches for $|X \cap Y|$

Given $\mathcal{K}_X$ and $\mathcal{K}_Y$ of size k_X and k_Y, one can construct a KMV $\mathcal{K}_{X\cup Y}$ by taking the $k = \min\{k_X, k_Y\}$ smallest elements from $K_X \cup K_Y$. $\widehat{|X \cup Y|}_K$, $\widehat{|X|}_K$ and $\widehat{|Y|}_K$ can be computed using the following equations (note that the second one uses the exact sizes of X, Y instead of their estimators).

$$\widehat{|X \cap Y|}_K = \widehat{|X|}_K + \widehat{|Y|}_K - \widehat{|X \cup Y|}_K \tag{43}$$

$$\widehat{|X \cap Y|}_K = |X| + |Y| - \widehat{|X \cup Y|}_K \tag{44}$$

We present now a simple upper bound (using union bound) on the probability that $\widehat{|X \cap Y|}_K$ deviates by more than t from the true value. Yet, if we know the exact size of X and Y (a reasonable assumption for graph algorithms as the degrees can be easily precomputed), we can get a considerably better bound. The following is a simple application of the identity $|X \cap Y| = |X| + |Y| - |X \cup Y|$ and Proposition 6 on the estimator of $|X \cup Y|$:

Proposition 7. *Let $\widehat{|X \cap Y|}_K$ be the estimator from (44), then*

$$P\left(|\widehat{|X \cap Y|}_K - |X \cap Y|| \geq t\right) = I_{u(|X\cup Y|,k,t/|X\cup Y|)}(k, |X \cup Y| - k + 1) - I_{l(|X\cup Y|,k,t/|X\cup Y|)}(k, |X \cup Y| - k + 1)$$

[9] uniformly at random

E Results & Derivations for Triangle Counts

E.1 Proof of consistency and asymptotic unbiasedness

Any estimator for triangle count analyzed in PG, is simply a sum of cardinalities $\widehat{|X \cap Y|}$ for different neighborhoods X and Y (cf. Section 3):

$$\widehat{TC}_{\star} = \frac{1}{3} \sum_{(u,v)\in E} \widehat{|N_u \cap N_v|}_{\star}$$

where $\star$ indicates a specific $\widehat{|X \cap Y|}_{\star}$ estimator (cf. Table 3). Since we have already proven consistency and asymptotic unbiasedness for each of the $\widehat{|X \cap Y|}_{\star}$ estimators presented in PG, we now can address jointly the consistency of the triangle count estimators. To do so we just need to acknowledge the fact that a sum of consistent estimators is itself a consistent estimator. Indeed this is a direct consequence of the more general *Slutsky theorem* which enable us to state that $\widehat{TC}_{\star} \xrightarrow{p} TC$. The asymptotic unbiasedness then follows from consistency and by noticing that all $\widehat{TC}_{\star}$ estimators have a bounded variance (see all the proofs presented below).

E.2 MinHash

We can show the concentration of the sum of the set intersection estimators using theorem B.3 presented in Appendix B. Then for the edge $e_i = uv$, we define $C_i = \deg(u) + \deg(v)$ thus giving us $S = \frac{1}{3} \sum_{i=1}^{n} C_i \frac{J}{1+J} = TC$. We will not consider the scaling factor $\frac{1}{3}$ till the final expressions of the bounds to ease the notation. Thus we can write:

$$\sum_{i=1}^{m} C_i = \sum_{i=1, e_i=uv}^{m} \deg(u) + \deg(v) = \\ = \sum_{v\in V} \deg(v)^2$$

Combining the above result with theorem B.3, we obtain the triangle count bound for MinHash presented in Theorem 6.1.

However this bound can be improved if we assume more independence, which will be satisfied in the case of triangle counting when the maximum degree is not too large. We now prove a tighter bound under these conditions.

Theorem E.1. *Let $Y_1 = \sum_i^n C_i \frac{\hat{J}_1}{1+\hat{J}_1}$, $Y_k = \sum_i^n C_i \frac{\hat{J}_k}{1+\hat{J}_k}$, and assume we partition the set of estimators into groups $\mathcal{X}_1, \cdots, \mathcal{X}_\chi$ such that estimators from each set are mutually independent. Then for any non-negative constants C_i and $S = \sum_{i=1}^n C_i \frac{J}{1+J}$*

$$P(|Y_1 - S| > t), P(|Y_k - S| > t) \leq 2\exp\left(-\frac{k(\max(0, t - 2S/k))^2}{2(\sum_i^\chi \sqrt{\sum_{d\in\mathcal{X}_i} C_d^2})^2}\right) \\ \leq 2\exp\left(-\frac{k(\max(0, t - 2S/k))^2}{2\chi \sum_i^n C_i^2}\right)$$

Proof. We modify the proof of Theorem B.3 by instead considering the random variables

$$\frac{\hat{J}_1}{1+\hat{J}_1} - \frac{J}{1+J} - \mu_1$$
$$\frac{\hat{J}_k}{1+\hat{J}_k} - \frac{J}{1+J} - \mu_k$$

where μ_1 and μ_k are chosen so as to make this random variable have mean zero.

We then first sum estimators from each group separately using Lemma B.1, which gives us subgaussian coefficient of $\sqrt{\sum_{d\in\mathcal{X}_i} C_d^2}$. Adding the groups together, we get using again Lemma B.1, that the subgaussian coefficient is $\sigma_\mathcal{X} = \sum_i^\chi \sqrt{\sum_{d\in\mathcal{X}_i} C_d^2}$. To finish the proof of the first inequality, we have to show a bound on $\sum C_i\mu_1$ and $\sum C_i\mu_k$. We show the argument for the case of 1-hash, the argument for k-hash is analogous. Note that μ_1 is the jensen gap of $\frac{\hat{J}_1}{1+\hat{J}_1}$. Since $\mathrm{Var}(\hat{J}_1) \leq J/k$, by Theorem 1 from [82], we have $-J/k \leq E[\frac{\hat{J}_k}{1+\hat{J}_k}] - \frac{J}{1+J} \leq 0$. We can bound $-J/k \geq 2/k\frac{J}{1+J}$. Therefore, we can bound

$$-2S/k \leq \sum C_i\mu_1 \leq 0$$

To prove the second inequality, we define the following optimization problem

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} \sqrt{x_n} \\ \text{subject to} \quad & \sum x_i = c \end{aligned}$$

Set $x_i = \sum_{d\in\mathcal{X}_i} C_d^2$ and $c = \sum C_i^2$. We see that for every possible assignment of the estimators to the sets $\{\mathcal{X}_i\}_{i=1}^{\chi}$, we have a feasible solution with objective value equal to the subgaussian coefficient $\sigma_\mathcal{X}$. Therefore, the subgaussian coefficient for any assignment to the groups is dominated by the optimum of this optimization problem.

Optimum of this optimization problem is when all x_i's have the same value – otherwise one can pick i, j such that $x_i < x_j$ and $0 < \varepsilon \leq (x_j - x_i)/2$ and then replace x_i by $x_i + \varepsilon$ and similarly x_j by $x_j - \varepsilon$, increasing the objective while retaining feasibility. This gives us objective value of $\chi\sqrt{\sum_{i=1}^n C_i^2/\chi} = \sqrt{\chi \sum_i^n C_i^2}$ □

To show the final expression of the bound, we use Theorem E.1. Then by Vizing's theorem, $\chi \leq \Delta + 1$ and by the same substitution done for the first bound, we have:

$$\sum_{i=1}^{m} C_i^2 = \sum_{i=1, e_i=uv}^{m} (\deg(u) + \deg(v))^2 \leq \\ \leq \sum_{i=1, e_i=uv}^{m} 2(\deg(u)^2 + \deg(v)^2) = 2\sum_{v\in V} \deg(v)^3$$

Indeed combining the above result with Theorem E.1, we obtain the triangle count bound for MinHash presented in Theorem 6.1 if the maximum degree is Δ.

E.3 Bloom Filters

To show the concentration for the Bloom Filters triangle count estimators (i.e. $\widehat{TC}_\bullet$), we need to bound their mean squared error:

$$E[(TC - \widehat{TC}_\bullet)^2] = Var(\widehat{TC}_\bullet) + (E[\widehat{TC}_\bullet] - TC)^2$$

Now we can use directly the expectation $E\left[|\widehat{N_u \cap N_v}|_\bullet\right]$ and the standard deviation $\sigma_{(u,v)_\bullet}$ of $|\widehat{N_u \cap N_v}|_\bullet$ as described in Section 5.1 to obtain the final expression for each specific estimator. As done for the MinHash case, we will not consider the scaling factor $\frac{1}{3}$ till the final expressions of the bounds to ease the notation. In the particular case of $|\widehat{N_u \cap N_v}|_{AND}$ we have:

$$E[|\widehat{N_u \cap N_v}|_{AND} = \delta_{B_{N_u \cap N_v}, b}\ B_{N_u \cap N_v}\left(1 - e^{-\frac{|N_u \cap N_v| b}{B_{N_u \cap N_v}}}\right) \tag{45}$$

and

$$\begin{aligned} Var(\sum_{(u,v)\in E} |\widehat{N_u \cap N_v}|_{AND}) &\leq \sum_{(u,v)\in E}\sum_{(q,r)\in E} \sqrt{Var(|\widehat{N_u \cap N_v}|_{AND})} \\ &\quad \cdot \sqrt{Var(|\widehat{N_q \cap N_r}|_{AND})} \\ &= \sum_{(u,v)\in E}\sum_{(q,r)\in E} \sigma_{(u,v)}\sigma_{(q,r)} \end{aligned} \tag{46}$$

where the above holds for the covariance inequality (also known as Cauchy-Schwarz inequality). Moreover, thanks to the result in [71], we can write the actual expression of the variance as:

$$\begin{aligned} Var(|\widehat{N_u \cap N_v}|_{AND}) \sim\ & \delta^2_{B_{N_u \cap N_v}, b}\ B_{N_u \cap N_v} \\ & \cdot \left[e^{-\frac{|N_u \cap N_v| b}{B_{N_u \cap N_v}}} - \left(1 + \frac{|N_u \cap N_v| b}{B_{N_u \cap N_v}}\right) e^{-\frac{2|N_u \cap N_v| b}{B_{N_u \cap N_v}}}\right] \end{aligned}$$

Thus we can now combine eq. (45) and (46) to obtain a bound for the mean squared error of the estimator. Finally thanks to Chebychev inequality, we have the final expression of the bound presented in Theorem 6.1 for the AND estimator. The bound also holds for $\widehat{TC}_L$ considering $E\left[|\widehat{N_u \cap N_v}|\right]$ and $\sigma_{(u,v)}$ as the expectation and the standard deviation of $|\widehat{N_u \cap N_v}|_L$. Similarly, to obtain the bounds for the BF based on OR, one can use the above equations and substitute $|N_u \cap N_v|$ with $|N_u \cup N_v|$ as described in Section 5.1.