Appendices

A DETAILED ANALYSES

A.1 Derivation of Lemma 1

We use Indicator random variables $X_{i,j}$ to determine whether the join of $i \in R, j \in S$ is made.

$$X_{i,j} = \begin{cases} 0 & i,j \text{ not joined in } \Psi. \\ 1 & i,j \text{ joined in } \Psi. \end{cases}$$

The tuples probes as long as they are stored.

$$E[X_{i,j}] = I[i.k = j.k] \cdot p(p_R q_s) = \frac{\varepsilon_r \varepsilon_s}{p} I[i.k = j.k].$$

By the linearity of expectation, the complete derivation is given by:

$$E[|\Psi|] = \sum E[X_{i,j}] = \sum_{i.k=j.k} E[X_{i,j}] = \frac{\varepsilon_R \varepsilon_S}{p} \gamma_{1,1} = \frac{\varepsilon_R \varepsilon_S}{p} J.$$

A.2 Derivation of Lemma 2

Obviously,

$$\mathrm{Var}[\hat{J}] = \frac{p^2}{(\varepsilon_R \varepsilon_S)^2} \mathrm{Var}[|\Psi|],$$

and we can obtain the result by calculating Var[|Ψ|].

$$\begin{aligned} \mathrm{Var}[|\Psi|] &= \sum \mathrm{Cov}(X_{r,s}, X_{r',s'}) \\ &= \sum E[X_{r,s}, X_{r',s'}] - E[X_{r,s}] E[X_{r',s'}]. \end{aligned}$$

There is a case where there are two tuples a, b with the same key in R, S respectively. a and b are joined if and only if a is sampled by σ_R and b is sampled by σ_S.

$$\begin{aligned} &E[X_{r,s}, X_{r',s'}](r \neq r', s \neq s', r.k \neq r'.k) \\ &= E[X_{r,s}] E[X_{r',s'}] \\ &E[X_{r,s}, X_{r',s'}](r \neq r', s \neq s', r.k = r'.k) \\ &= p q_R^2 q_S^2 I[r.k = s.k = r'.k = s'.k] \\ &E[X_{r,s}, X_{r,s'}](s \neq s') \\ &= p q_R q_S^2 I[r.k = s.k = s'.k] \\ &E[X_{r,s}, X_{r',s}](r \neq r') \\ &= p q_R^2 q_S I[r.k = s.k = r'.k] \\ &E[X_{r,s}, X_{r,s}] = E[X_{r,s}] \end{aligned}$$

Except for the components above, other items in the covariance form of $Var[|\Psi|]$ can be resolved to 0, so the final variance is given by:

$$\begin{aligned} &\mathrm{Var}[\hat{J}] \\ &= \frac{p^2}{(\varepsilon_R \varepsilon_S)^2} \mathrm{Var}[|\Psi|] \\ &= \frac{p^2}{(\varepsilon_R \varepsilon_S)^2} \Big((\gamma_{2,2} - \gamma_{2,1} - \gamma_{1,2} + \gamma_{1,1})(p q_R^2 q_S^2 - p^2 p_R^2 q_s^2) \\ &\quad + \gamma_{1,1}(p p_R q_s - p^2 p_R^2 q_s^2) \quad + (\gamma_{1,2} - \gamma_{1,1})(p q_R q_S^2 - p^2 p_R^2 q_s^2) \\ &\quad + (\gamma_{2,1} - \gamma_{1,1})(p q_R^2 q_S - p^2 p_R^2 q_s^2) \Big) \\ &= \frac{1-p}{p} \gamma_{2,2} + \frac{p - \epsilon_S}{p \epsilon_S} \gamma_{2,1} + \frac{p - \epsilon_R}{p \epsilon_R} \gamma_{1,2} + \frac{(p - \epsilon_S)(p - \epsilon_R)}{p \epsilon_S \epsilon_R} \gamma_{1,1}. \end{aligned}$$

A.3 Derivation of Lemma 3

In the random assumption, there is an equal chance for the two arrival orders of i, j.

$$E[X_{i,j}] = \frac{p(q_R + q_S)}{2} I[i.k = j.k] = \frac{\varepsilon_R + \varepsilon_S}{2} I[i.k = j.k].$$

In particular, it is important to note that according to Definition 2, stream joins need to take into account timestamps, we cannot assume that a join can be generated as long as one of i, j is sampled and it is wrong to use the principle of inclusion-exclusion getting the following result.

$$E[X_{i,j}] = p(q_R + q_S - q_R q_S)) I[i.k = j.k].$$

With the equal chance of r goes before s and s goes before r, the complete derivation is given by:

$$E[|\Psi|] = \sum E[X_{i,j}] = \sum_{i.k=j.k} E[X_{i,j}] = \frac{\varepsilon_R + \varepsilon_S}{2} \gamma_{1,1} = \frac{\varepsilon_R + \varepsilon_S}{2} J.$$

A.4 Derivation of Lemma 4

According to Appendix A.3,

$$\mathrm{Var}[\hat{J}] = 4\mathrm{Var}[|\Psi|]/(\varepsilon_R + \varepsilon_S)^2,$$

and we can obtain the result by calculating Var[|Ψ|].

$$\begin{aligned} \mathrm{Var}[|\Psi|] &= \sum \mathrm{Cov}(X_{r,s}, X_{r',s'}) \\ &= \sum E[X_{r,s}, X_{r',s'}] - E[X_{r,s}] E[X_{r',s'}]. \end{aligned}$$

There is a case where there are two tuples a, b with the same key in R, S respectively. For $\lambda = 1$, when a is sampled by σ_R and b is not sampled by σ_S, if a comes before b, then a join is generated. So, we should discuss $E[X_{r,s}, X_{r',s'}]$ with taking timestamps into account, which means the permutation of the arrival.

$$\begin{aligned} &E[X_{r,s}, X_{r',s'}](r \neq r', s \neq s', r.k \neq r'.k) \\ &\quad = E[X_{r,s}] E[X_{r',s'}] \\ &E[X_{r,s}, X_{r',s'}](r \neq r', s \neq s', r.k = r'.k) \\ &\quad = \frac{1}{4} p (q_R + q_S)^2 I[r.k = s.k = r'.k = s'.k] \\ &E[X_{r,s}, X_{r,s'}](s \neq s') \\ &\quad = \frac{1}{3} p (q_R + q_S q_R + q_S^2) I[r.k = s.k = s'.k] \\ &E[X_{r,s}, X_{r',s}](r \neq r') \\ &\quad = \frac{1}{3} p (q_R^2 + q_S q_R + q_S) I[r.k = s.k = r'.k] \\ &E[X_{r,s}, X_{r,s}] = E[X_{r,s}] \end{aligned}$$

So the final variance is given by:

$$\begin{aligned} &\mathrm{Var}[\hat{J}] \\ &= \quad 4\mathrm{Var}[|\Psi|]/(\varepsilon_R + \varepsilon_S)^2 \\ &= \quad \frac{4}{(\varepsilon_R + \varepsilon_S)^2} \\ &\quad \Big((\gamma_{2,2} - \gamma_{2,1} - \gamma_{1,2} + \gamma_{1,1}) \end{aligned}$$

$$\times\left(\frac{1}{4}p(q_R+q_S)^2-\frac{1}{4}p^2(q_R+q_S)^2\right)$$
$$+(\gamma_{1,2}-\gamma_{1,1})\left(\frac{1}{3}p(q_R+q_Sq_R+q_S^2)-\frac{1}{4}p^2(q_R+q_S)^2\right)$$
$$+(\gamma_{2,1}-\gamma_{1,1})\left(\frac{1}{3}p(q_R^2+q_Sq_R+q_S)-\frac{1}{4}p^2(q_R+q_S)^2\right)$$
$$+\gamma_{1,1}\left(\frac{1}{2}p(q_R+q_S)-\frac{1}{4}p^2(q_R+q_S)^2\right)\Bigg)$$
$$=\quad\left(\frac{1-p}{p}\right)\gamma_{2,2}+\frac{(\varepsilon_S-3\varepsilon_R)(\varepsilon_S+\varepsilon_R)+4p\varepsilon_R}{3p(\varepsilon_R+\varepsilon_S)^2}\gamma_{2,1}$$
$$+\frac{(\varepsilon_R-3\varepsilon_S)(\varepsilon_R+\varepsilon_S)+4p\varepsilon_S}{3p(\varepsilon_R+\varepsilon_S)^2}\gamma_{1,2}+\frac{-\varepsilon_R-\varepsilon_S-2p}{3p(\varepsilon_R+\varepsilon_S)}\gamma_{1,1}.$$

A.5 Detailed Form of $Var[\hat{J}]$ in General Cases

Besides the notations $\xi, \mu, f(x_1, x_2)$ used in Theorem 1 and Theorem 2, we need to introduce extra tools. Similar to $\xi_{r,s}$, me define $xi_{r,s,r'}$ to be the indicator variable $I[r$ arrives before s before r' or r' arrives before s before $r]$. In $\xi_{r,s,r'}$, we do not distinguish r and r', since their relative order does not affect their contribution, and we have $\xi_{s,r,r'}+\xi_{r,s,r'}+\xi_{r,r',s}=1$ With ξs using three subscripts, we define two more auxiliary functions:

$$f_1(x_1,x_2,x_3)=p\Bigg(\frac{\epsilon_R^2\left(\lambda_S\left(1-\frac{\epsilon_S}{p}\right)+\frac{\epsilon_S}{p}\right)x_1}{p^2}$$
$$+\frac{\epsilon_S\epsilon_R\left(\lambda_R\left(1-\frac{\epsilon_R}{p}\right)+\frac{\epsilon_R}{p}\right)x_2}{p^2}$$
$$+\frac{\epsilon_S\left(\lambda_R\left(1-\frac{\epsilon_R}{p}\right)+\frac{\epsilon_R}{p}\right)^2x_3}{p}\Bigg),$$
$$f_2(x_1,x_2,x_3)=p\Bigg(\frac{\epsilon_S^2\left(\lambda_R\left(1-\frac{\epsilon_R}{p}\right)+\frac{\epsilon_R}{p}\right)x_1}{p^2}$$
$$+\frac{\epsilon_R\epsilon_S\left(\lambda_S\left(1-\frac{\epsilon_S}{p}\right)+\frac{\epsilon_S}{p}\right)x_2}{p^2}$$
$$+\frac{\epsilon_R\left(\lambda_S\left(1-\frac{\epsilon_S}{p}\right)+\frac{\epsilon_S}{p}\right)^2x_3}{p}\Bigg).$$

With the same technique in Appendix A.2 and Appendix A.4, we calculate:

$$Var[\hat{J}]=\quad\frac{1-p}{p}\sum_{r,s,r',s'}I[r.k=s.k=r'.k=s'.k\wedge r\neq r'\wedge s\neq s']$$
$$\times\frac{f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})}\frac{f(\xi_{s',r'},\xi_{r',s'})}{f(\mu_{S,R},\mu_{R,S})}$$
$$+\sum_{r,r',s}I[r.k=r'.k=s.k\wedge r\neq r']$$
$$\times\frac{f_1(\xi_{r,r',s},\xi_{r,s,r'},\xi_{s,r,r'})-f(\xi_{s,r},\xi_{r,s})f(\xi_{s,r'},\xi_{r',s})}{f(\mu_{S,R},\mu_{R,S})^2}$$
$$+\sum_{r,s,s'}I[s.k=s'.k=r.k\wedge s\neq s']$$
$$\times\frac{f_2(\xi_{s,s',r},\xi_{s,r,s'},\xi_{r,s,s'})-f(\xi_{s,r},\xi_{r,s})f(\xi_{s',r},\xi_{r,s'})}{f(\mu_{S,R},\mu_{R,S})^2}$$
$$+\sum_{r,s}I[s.k=r.k]\times\frac{(1-f(\xi_{s,r},\xi_{r,s}))f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})^2}.$$

A.6 Parameter Setting Analysis

There are two parameters unknown (p and λ). As foregoing statements, p majorly affects variance, while λ majorly affects output size. The adjustment between these metrics formulates the adaption of Bi-Probe.

More precisely, the adaption is achieved by using p to maintain the variance level and modifying λ to adjust the output size. Therefore, λ is in major charge of the adaption, and we prefer λ to be an arbitrarily settable value. Our major concern with parameter setting is p. Thus, the focus is to study the impact of p on variance, and the means is to take the partial derivative of the variance for p. Then, we figure out the coefficient of p. Through a mindset of verifying the boundary conditions, we fill $\lambda=0$ and $\lambda=1$ to explore the basic behavior of p first. Notably, $\lambda=0$ and $\lambda=1$ both make $\frac{f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})}\frac{f(\xi_{s',r'},\xi_{r',s'})}{f(\mu_{S,R},\mu_{R,S})}$ become 1, and the rest parts are only $\frac{f_1(\xi_{r,r',s},\xi_{r,s,r'},\xi_{s,r,r'})}{f(\mu_{S,R},\mu_{R,S})}$, $\frac{f_2(\xi_{s,s',r},\xi_{s,r,s'},\xi_{r,s,s'})}{f(\mu_{S,R},\mu_{R,S})}$, and $\frac{f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})^2}$. Since f_1 and f_2 also have a strong symmetry, we focus on:

$$\frac{f_1(\xi_{r,r',s},\xi_{r,s,r'},\xi_{s,r,r'})}{f(\mu_{S,R},\mu_{R,S})}. \tag{7}$$

When setting $\lambda=0$, $\frac{f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})^2}=\frac{p}{\epsilon_R\epsilon_S}$ and (7)$=1/\epsilon_S$. When setting $\lambda=1$, $\frac{f(\xi_{s,r},\xi_{r,s})}{f(\mu_{S,R},\mu_{R,S})^2}=\frac{\epsilon_S\xi_{s,r}+\epsilon_R\xi_{r,s}}{(\epsilon_S\mu_{S,R}+\epsilon_R\mu_{R,S})^2}$ and

$$(7)=\frac{\frac{1}{p}(\epsilon_R^2\xi_{r,r',s}+\epsilon_S\epsilon_R\xi_{r,r',s})+\epsilon_S\xi_{s,r,r'}}{(\mu_{S,R}+\mu_{R,S})^2}. \tag{8}$$

Using the principle of inclusion-exclusion like Appendix A.2 and Appendix A.4, we can refill these constants back to the coefficient of p. It roughly conforms to the format $\frac{1}{p}(\gamma_{2,2}C_0-\gamma_{1,2}C_1-\gamma_{2,1}C_2+\gamma_{1,1}C_3)+p\gamma_{1,1}C_4$, where $C_0,...C_4$ are constants. Let $\lambda=0$, it makes no impact on variance, $C_0,..,C_3$ all become 1, and the solution of p is :

$$\sqrt{\frac{\epsilon_R\epsilon_S(\gamma_{2,2}-\gamma_{1,2}-\gamma_{2,1}+\gamma_{1,1})}{\gamma_{1,1}}}.$$

When λ takes effect, these constants fluctuate in a relatively small range, since even the $rho=1$ case does not change the level of p's coefficient. Not to mention the unpredictability of streaming data, the parameter setting has heuristic features at its root. Thus, by Occam's Razor, we cut the fluctuation and choose the direct way to set p by making $\lambda=0$.