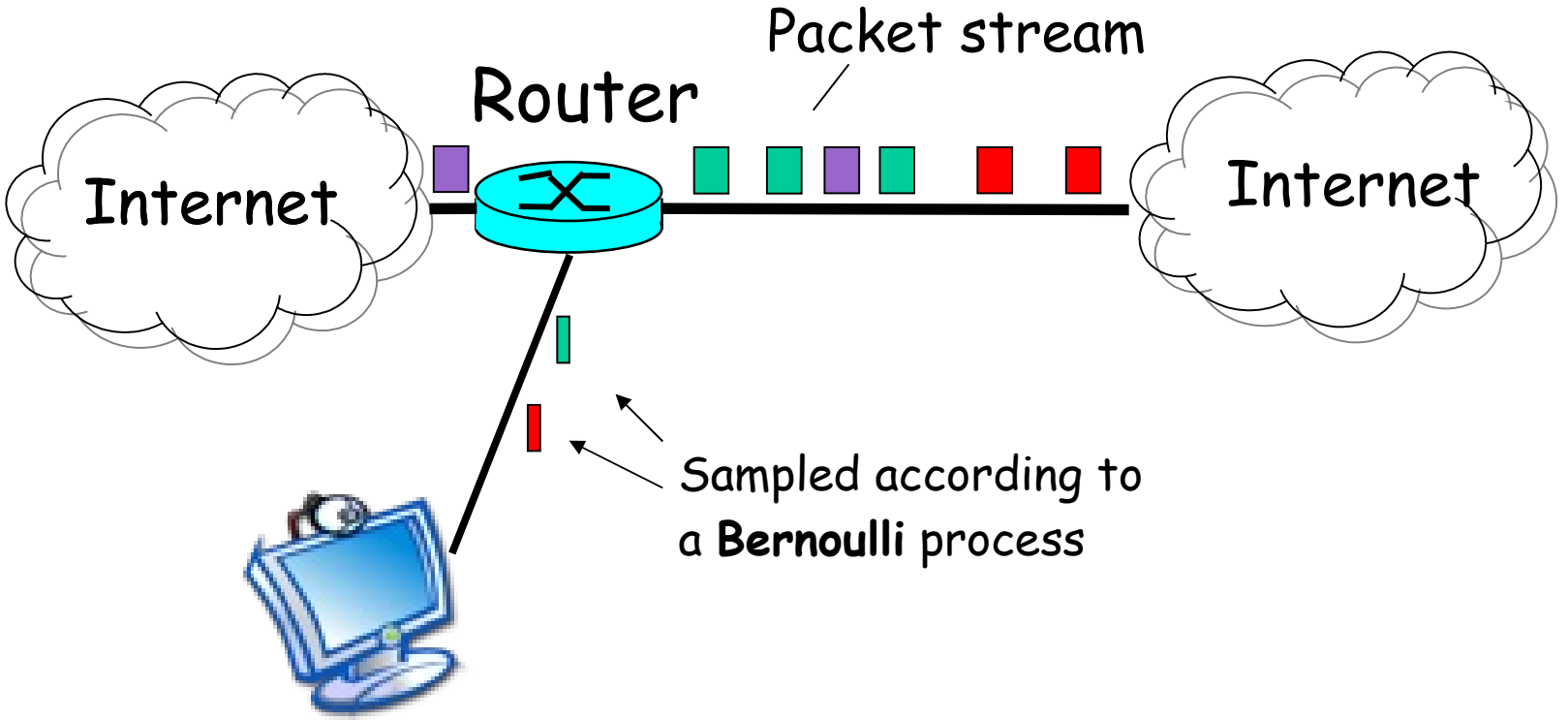
Fisher Information of Sampled Packets: an Application to Flow Size Estimation

Bruno Ribeiro Don Towsley
Umass Amherst

Tao Ye Jean Bolot
Sprint Labs

Sprint®

Motivation scenario



Traffic summary:

Classifies sampled packets into sampled flows → Estimate flow level statistics

Pkt sampling x flow size distribution

- Hard to estimate flow level information from sampled summaries.

Distribution of number of packets in a flow:

- "Inverting sampled traffic"
  - Nicolas Hohn and Darryl Veitch, IMC 2003
  - Inversion cannot find accurate estimates for flow size dist.
- "Estimating Flow Distributions From Sampled Flow Statistics"
  - Nick Duffield et al., SIGCOMM 2003, ToN 2005
    - Maximum likelihood estimator does a good job.
    - Inversion is not a good estimator.
      - Inversion estimates have high variance

Outline

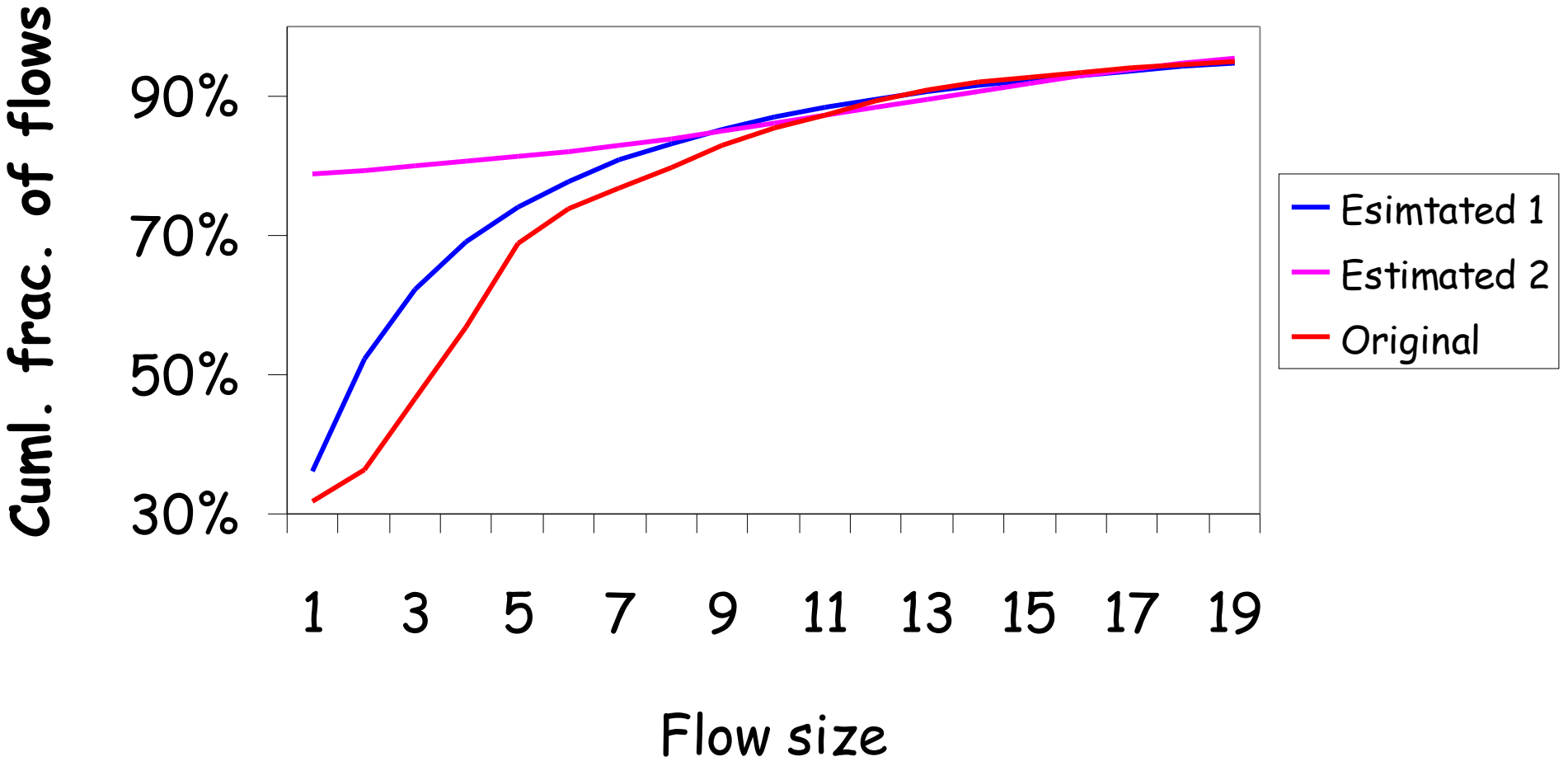
- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

Outline

- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

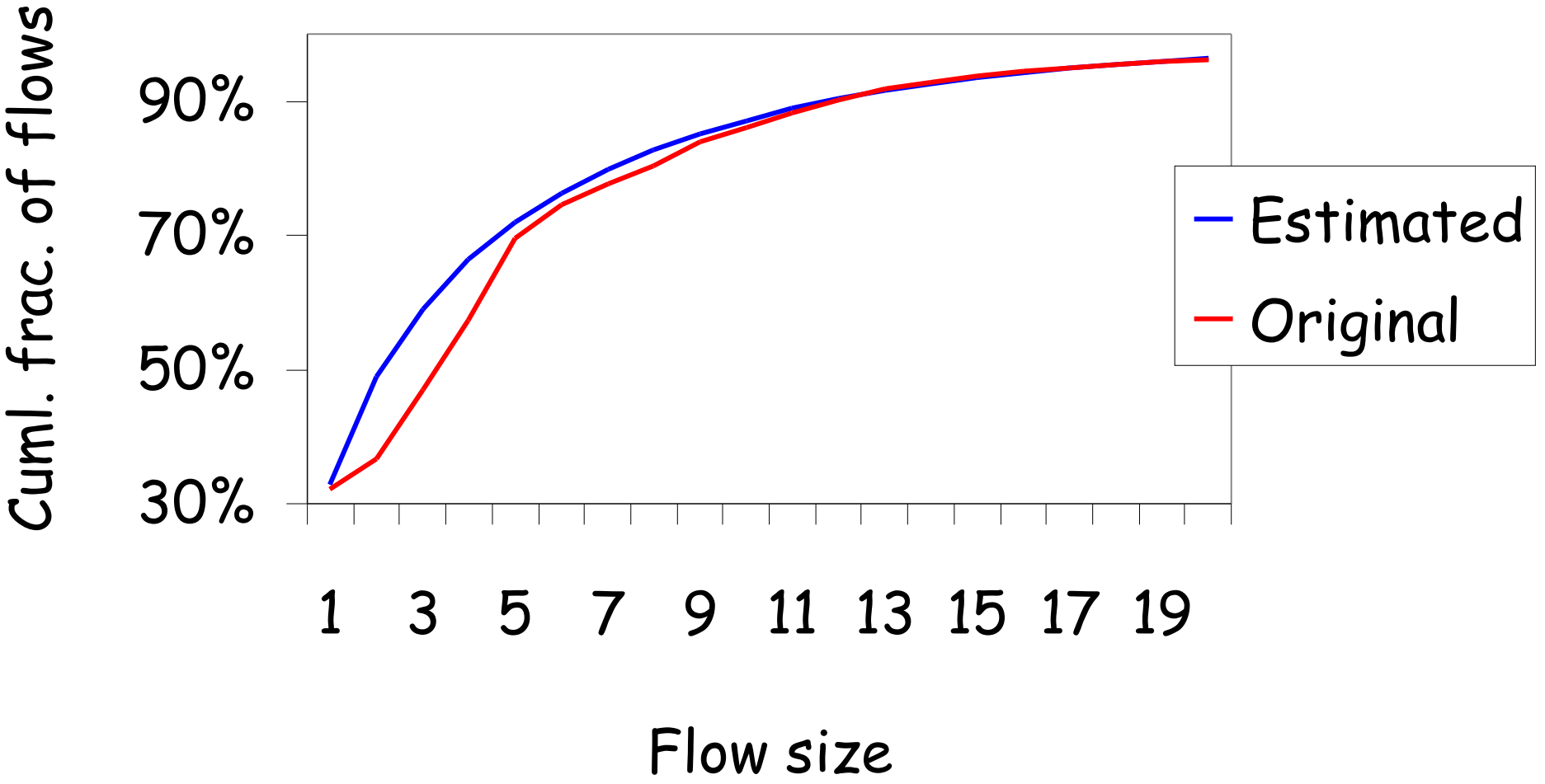
Best estimator to date

- Maximum likelihood estimator (Duffield et al., ToN 2005)
- Parameters: pkt sampling rate = 1/200; 128,000 sampled flows

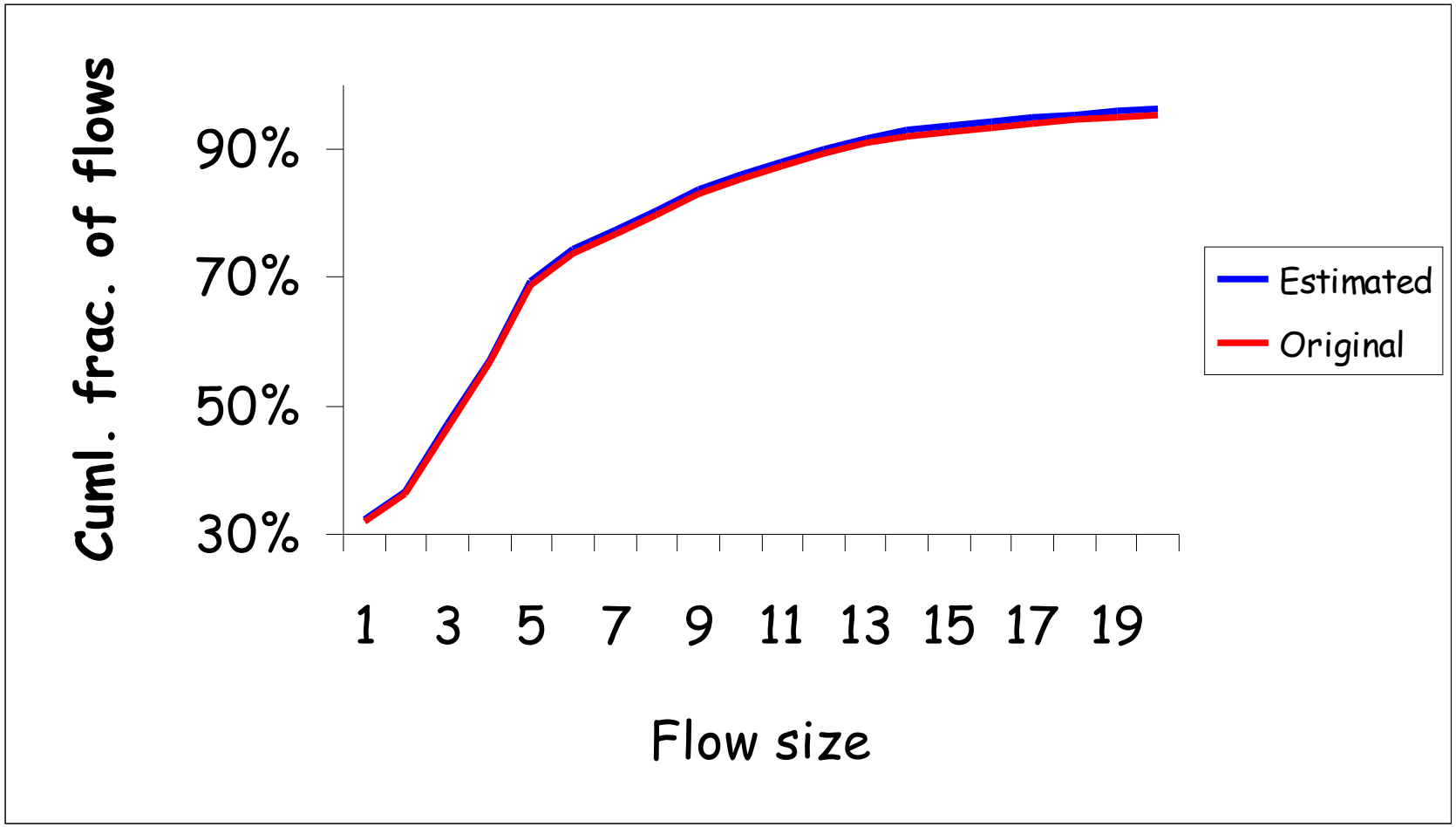


Many more samples

- Maximum likelihood estimator (Duffield et al., ToN 2005)
- Parameters: pkt sampling rate = 1/200, 1 **trillion** sampled flows (10^{12})



Exploiting TCP protocol info.

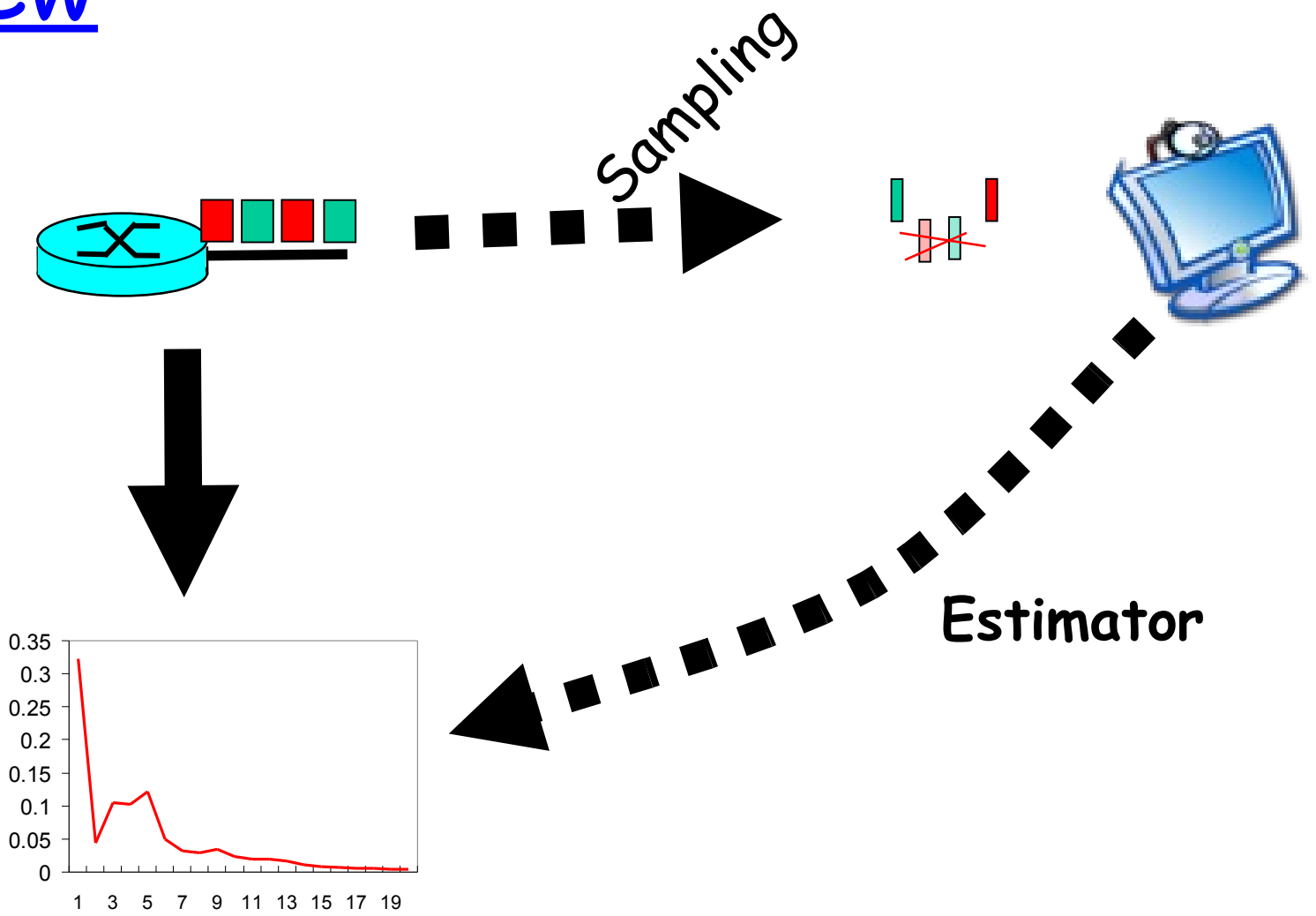


Parameters: pkt sampling rate = 1/200, 1 **million** sampled flows (10^6)

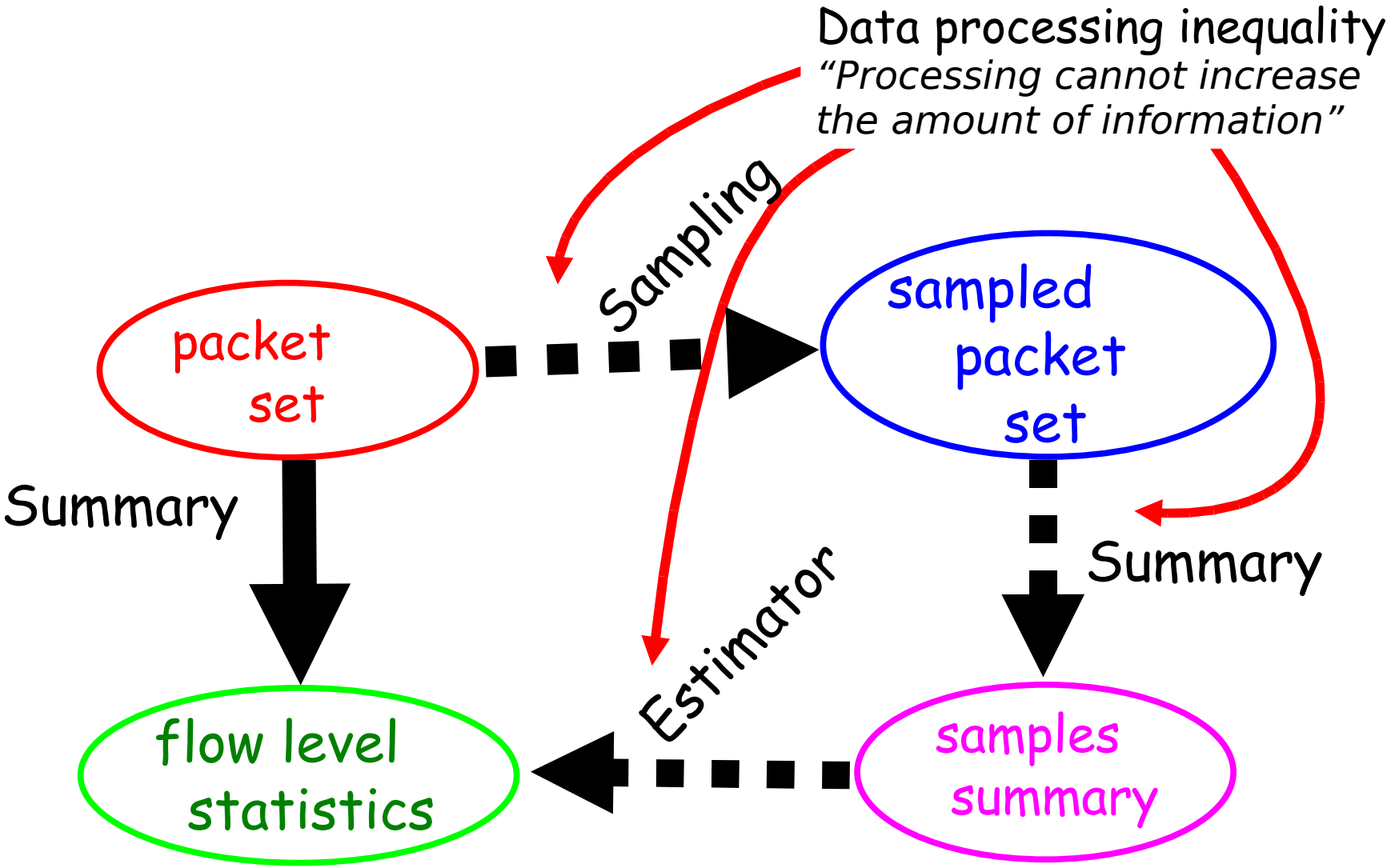
Outline

- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

Finding estimates – schematic view



Information model

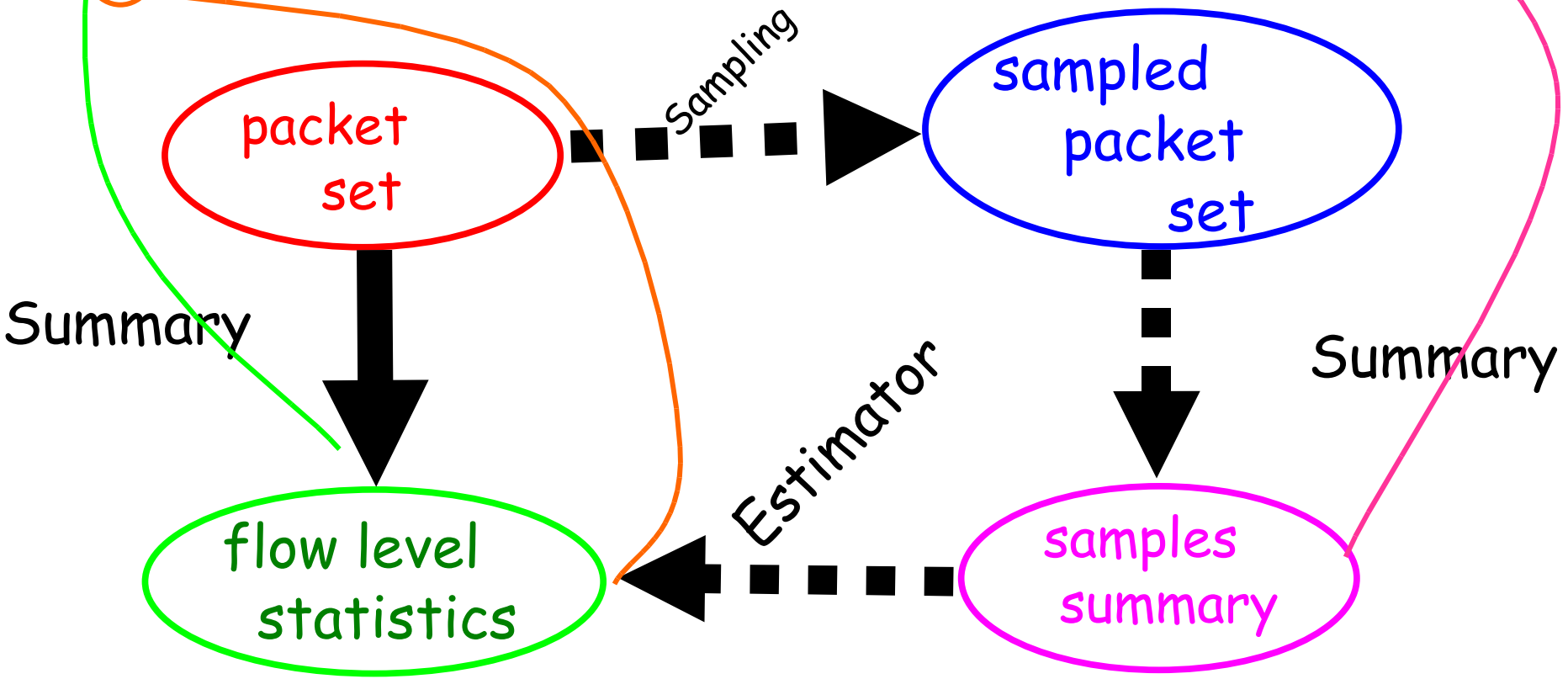


The math of the model

d_j - fraction of sampled flows with label j (sampled packets).

θ_i - fraction of flows with size i.

$\hat{\theta}_i$ - estimated fraction of flows with size i.



Model: e.g. packet count

- W - maximum flow size.
- $b_{i,j}$ - binomial probability of selecting *j* packets out of *i*.

$$d_j = \sum_{i=1}^{W} b_{i,j}\theta_i \text{ , or in vector form } \vec{d} = \mathbf{B}\vec{\theta}$$

- D - *number of sampled packets of a randomly chosen sampled flow.*
  - $P[D = j \mid \vec{\theta}] = d_j$ (likelihood function)
  - *Sampled flows are samples from D*

Other protocol information changes **B**

Accuracy of estimates (notation)

- Mean squared error (MSE): $E[(\theta_i - \hat{\theta}_i)^2]$
- Standard deviation error (STD): $\sqrt{\text{MSE}}$
- Confidence interval size: $\propto$STD (for $n \gg 1$)

Outline

- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

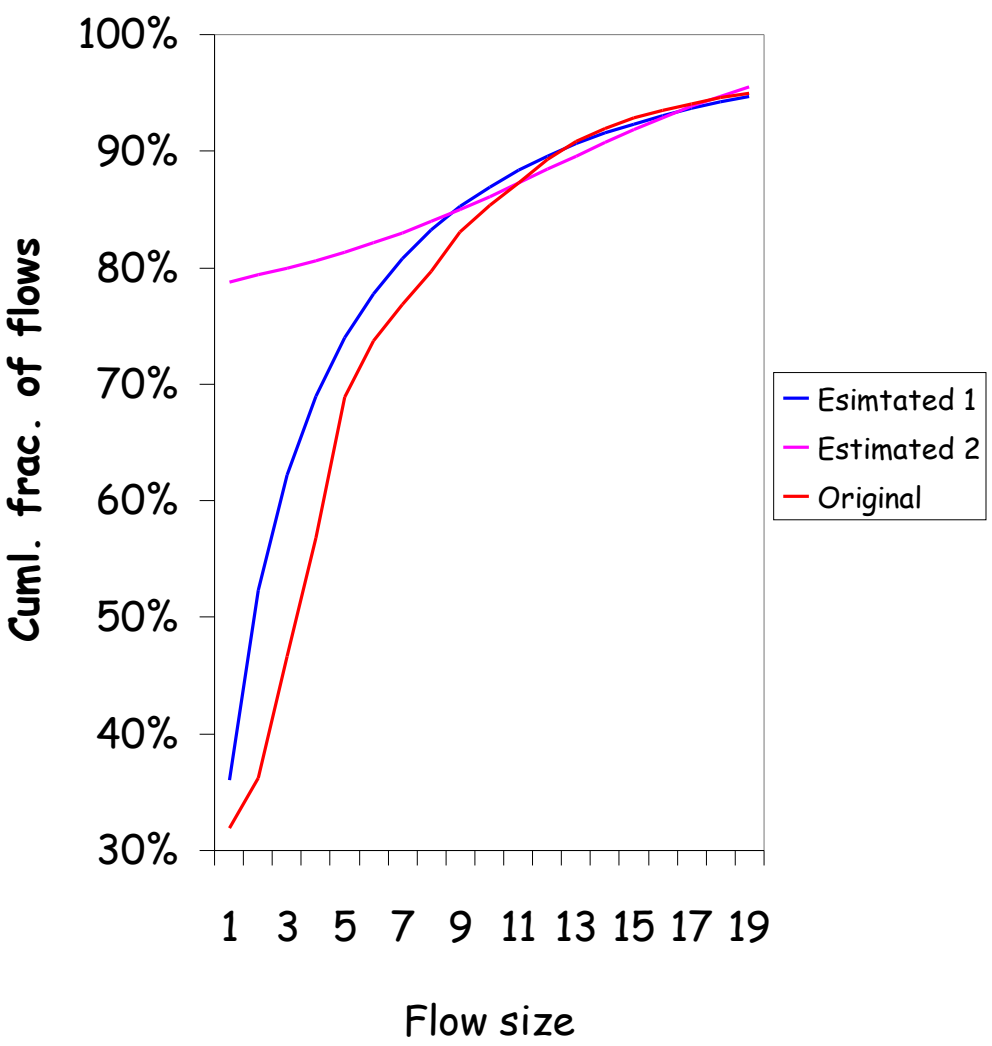
Fisher info. and Cramér-Rao bound

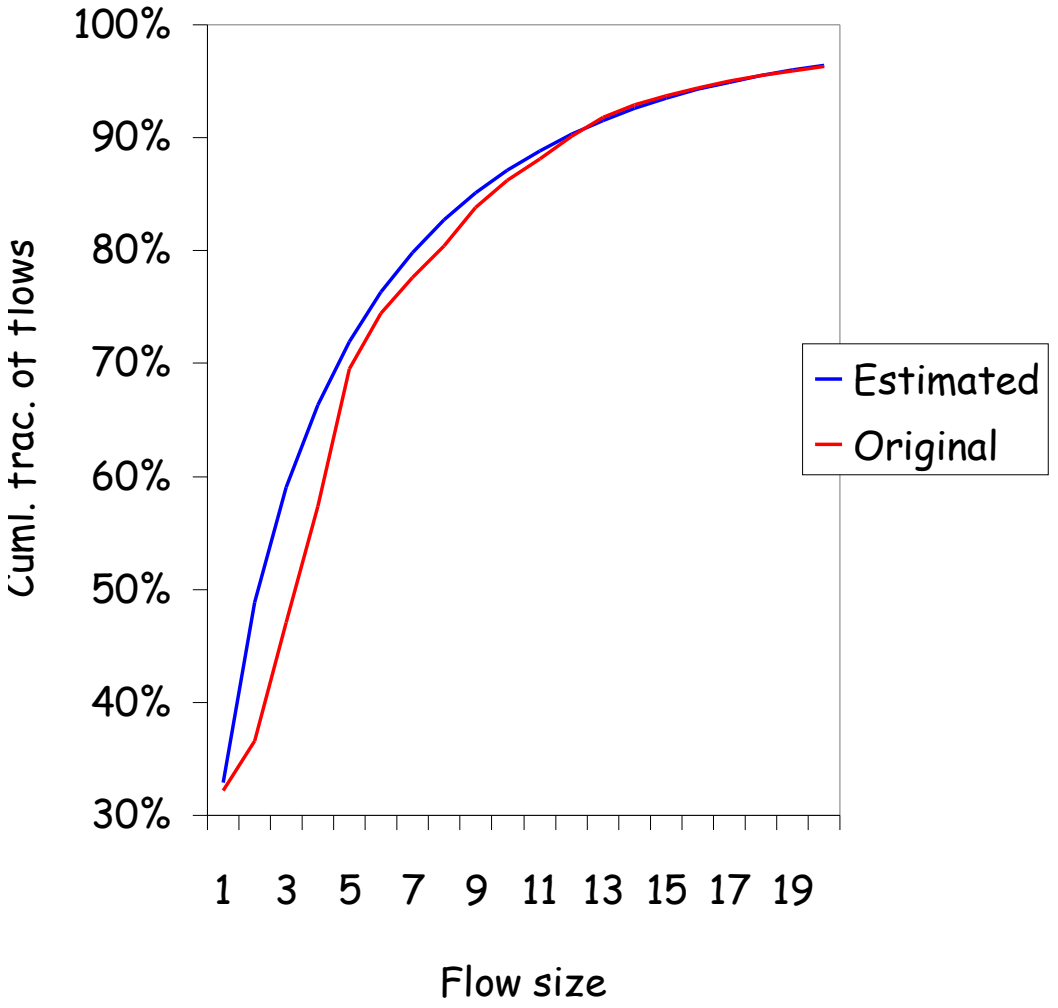
- *Fisher information (denoted* $\mathcal{I}(n, \vec{d}, \vec{\theta})$ *)*
  - *Amount of information that n samples from D carry about unobservable parameters* $\vec{\theta}$*.*
  - $\mathcal{I}(n, \vec{d}, \vec{\theta})$ is a WxW matrix
- Fisher information of n sampled flows:
  - $\mathcal{I}(n, \vec{d}, \vec{\theta}) = n\mathcal{I}(1, \vec{d}, \vec{\theta})$
- Cramér-Rao bound:
  - *The MSE of any unbiased estimator is lower bounded by the inverse of the Fisher information:*

$$MSE \geq -(\mathcal{I}(1, \vec{d}, \vec{\theta})^{-1})_{i,i}/n$$

STD x number of samples

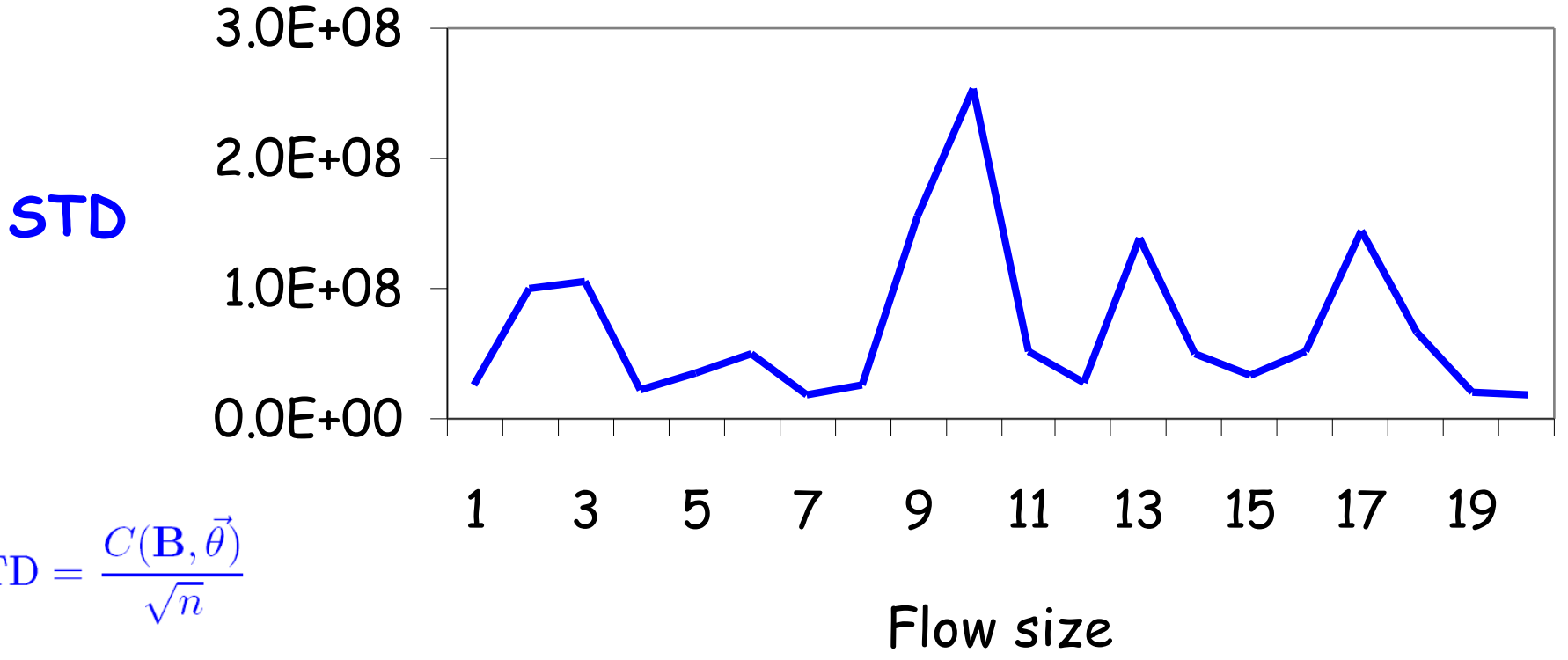
Cramér-Rao bound: $MSE \geq -(\mathcal{I}(1, \vec{d}, \vec{\theta})^{-1})_{i,i}/n$





Fisher info. of pkt counts example

Accuracy of best unbiased estimator over pkt counts summaries



$$\mathrm{STD} = \frac{C(\mathbf{B}, \vec{\theta})}{\sqrt{n}}$$

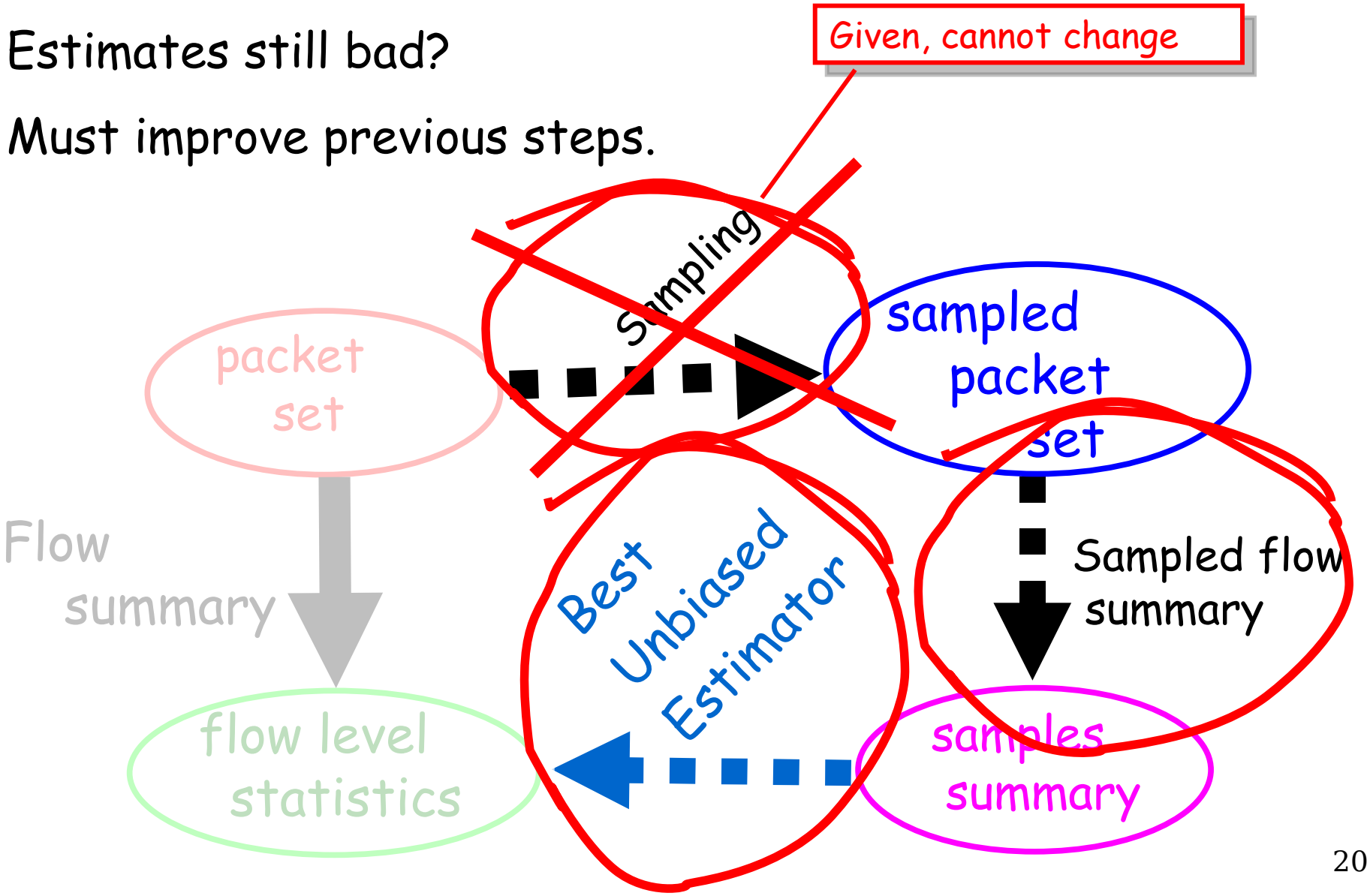
10^{20} sampled flows needed for small STD lower bound (STD ≥ 0.025).

Outline

- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

Assume best unbiased estimator

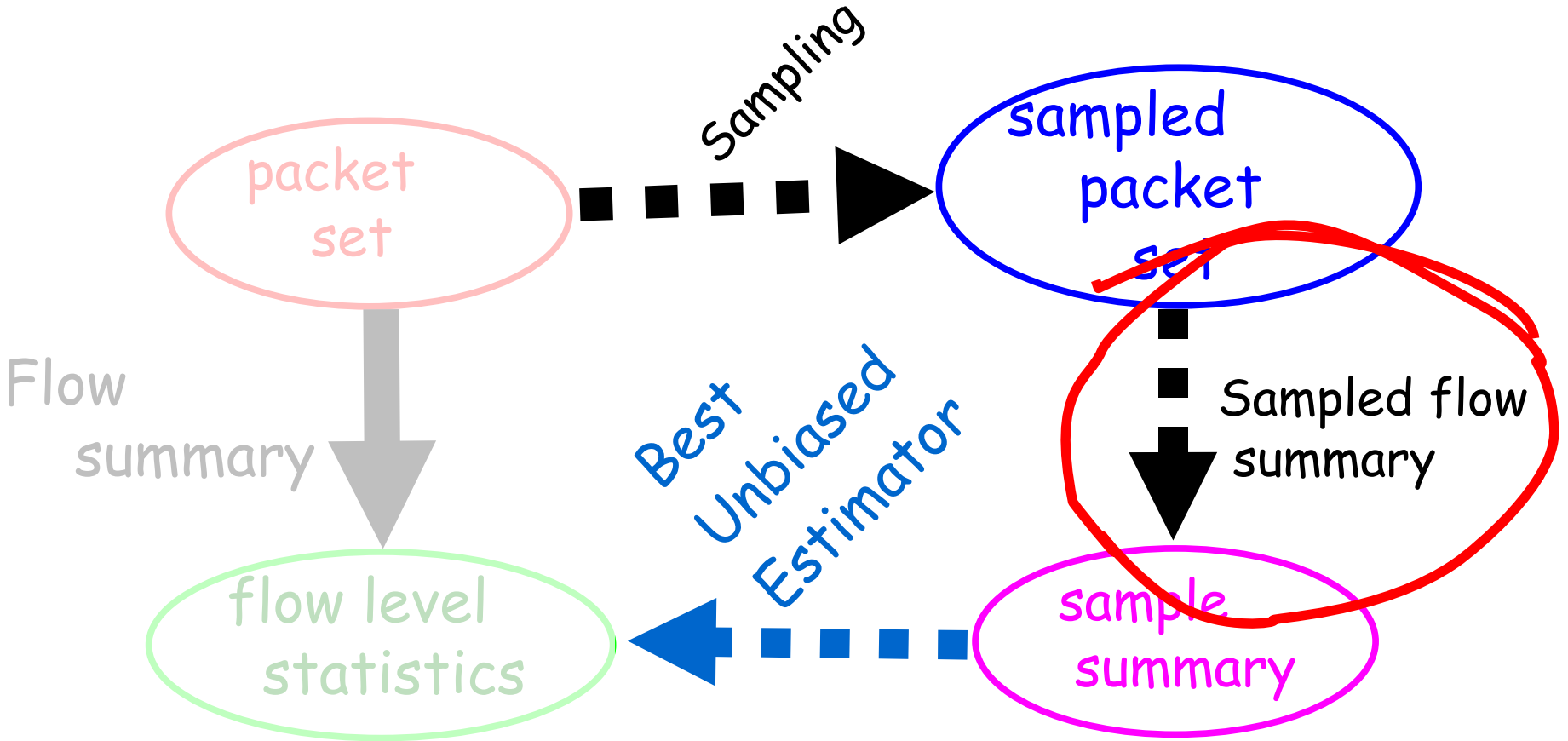
- Assume best unbiased estimator.
- Estimates still bad?
- Must improve previous steps.



How to improve accuracy

Change sampled flow summary.

- Make use of protocol information.



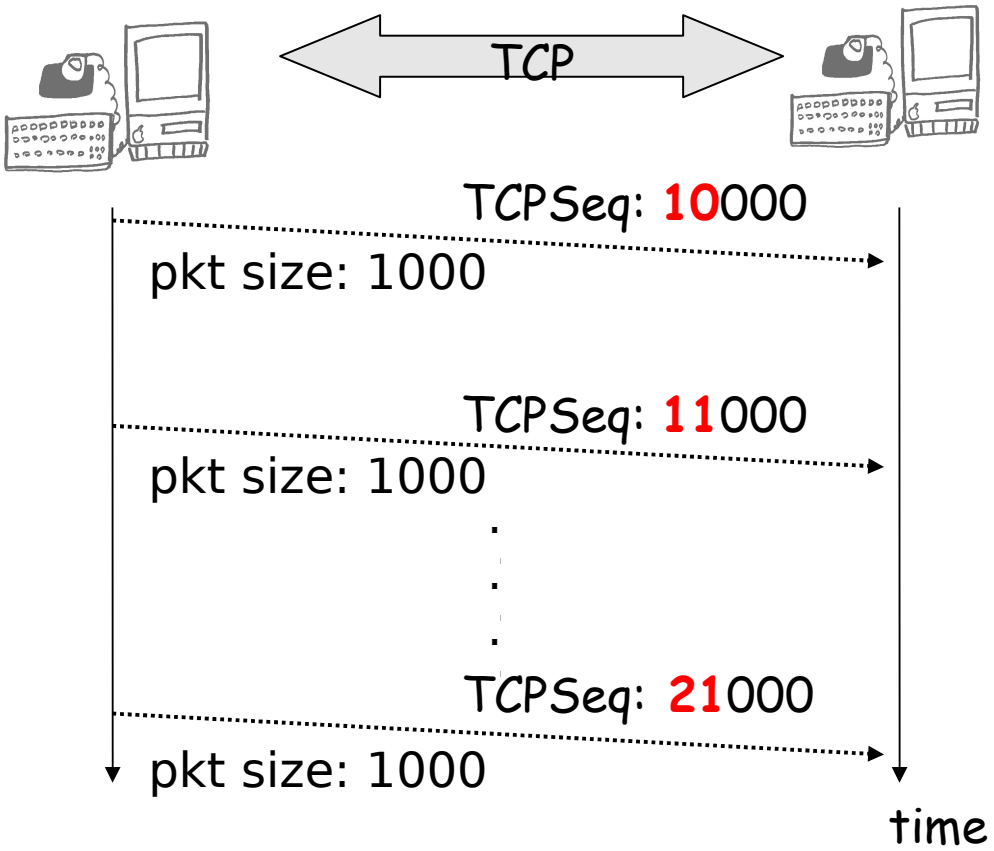
Add TCP protocol information

TCP Sequence Numbers?

Assumptions:

- All flow packets have same size*
- All flow packets transverse instrumented router
- TCP payload ≠ 0

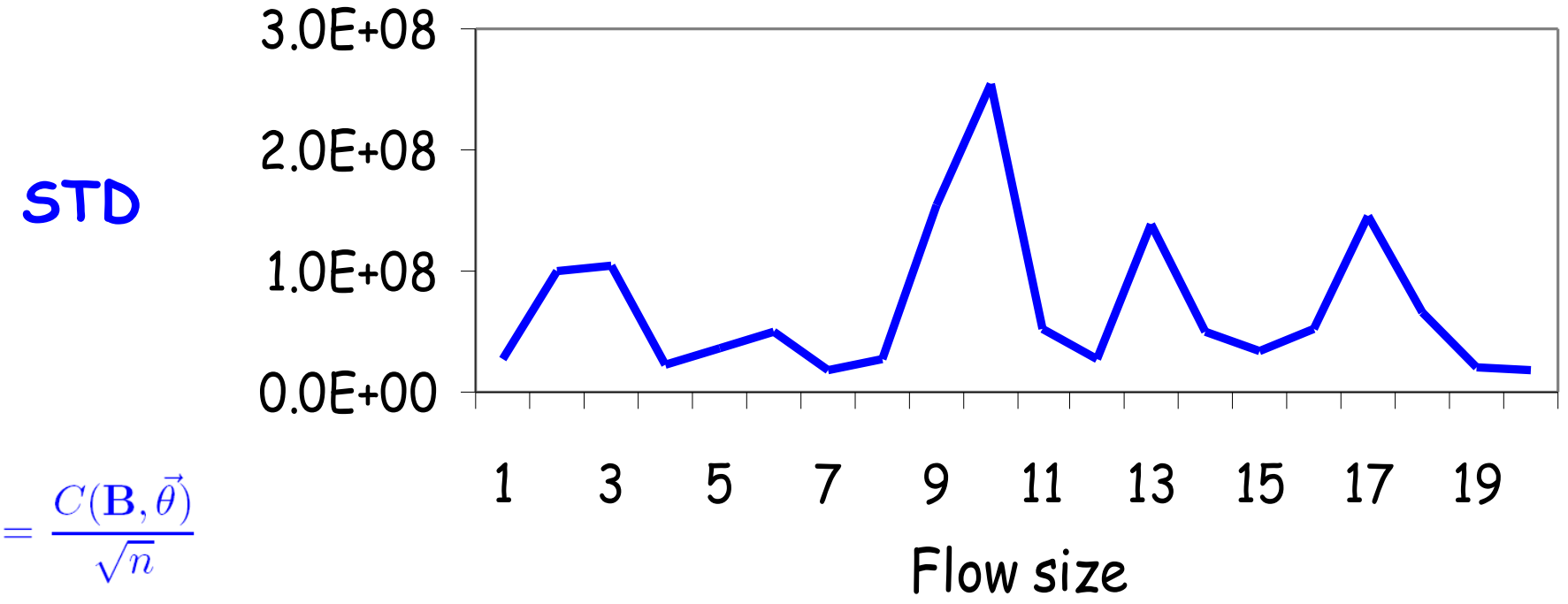
* SYN, FIN packets treated separately



- Can use other protocol information (SYN, FIN).

Adding protocol information

- Summary uses no protocol information (packet counts only).

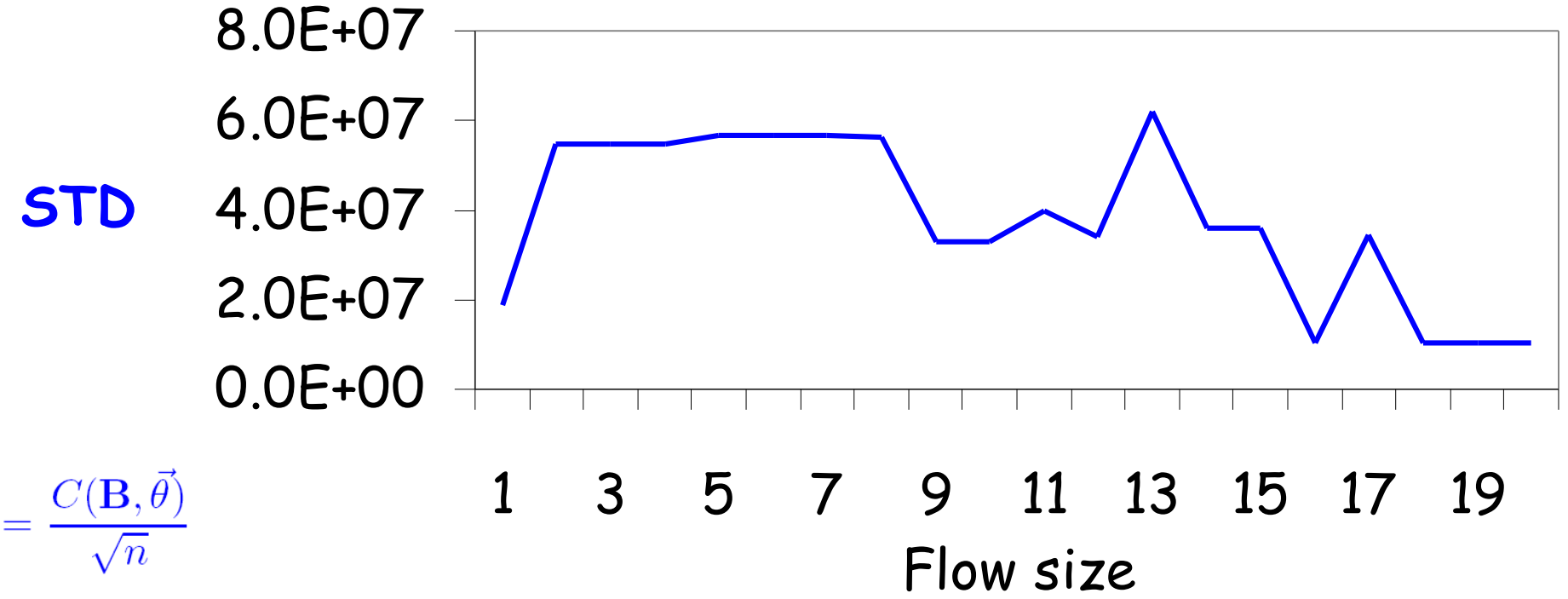


$$\mathrm{STD} = \frac{C(\mathbf{B}, \vec{\theta})}{\sqrt{n}}$$

10^{20} sampled flows needed for small STD lower bound (STD ≥ 0.025).

Adding protocol information

- Summary uses SYN flag information.
- Information a bit higher.

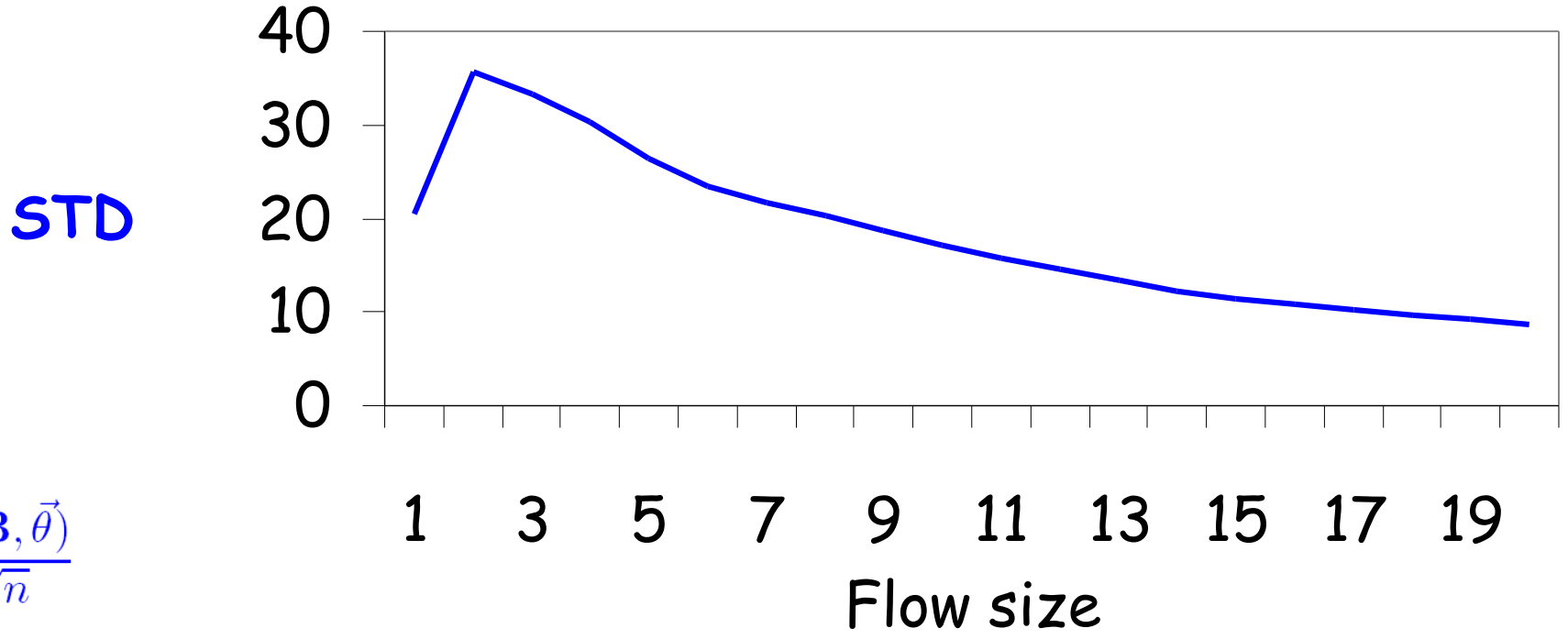


$$\mathrm{STD} = \frac{C(\mathbf{B}, \vec{\theta})}{\sqrt{n}}$$

4×10^{18} sampled flows needed for small STD lower bound (STD ≥ 0.025).

Adding protocol information

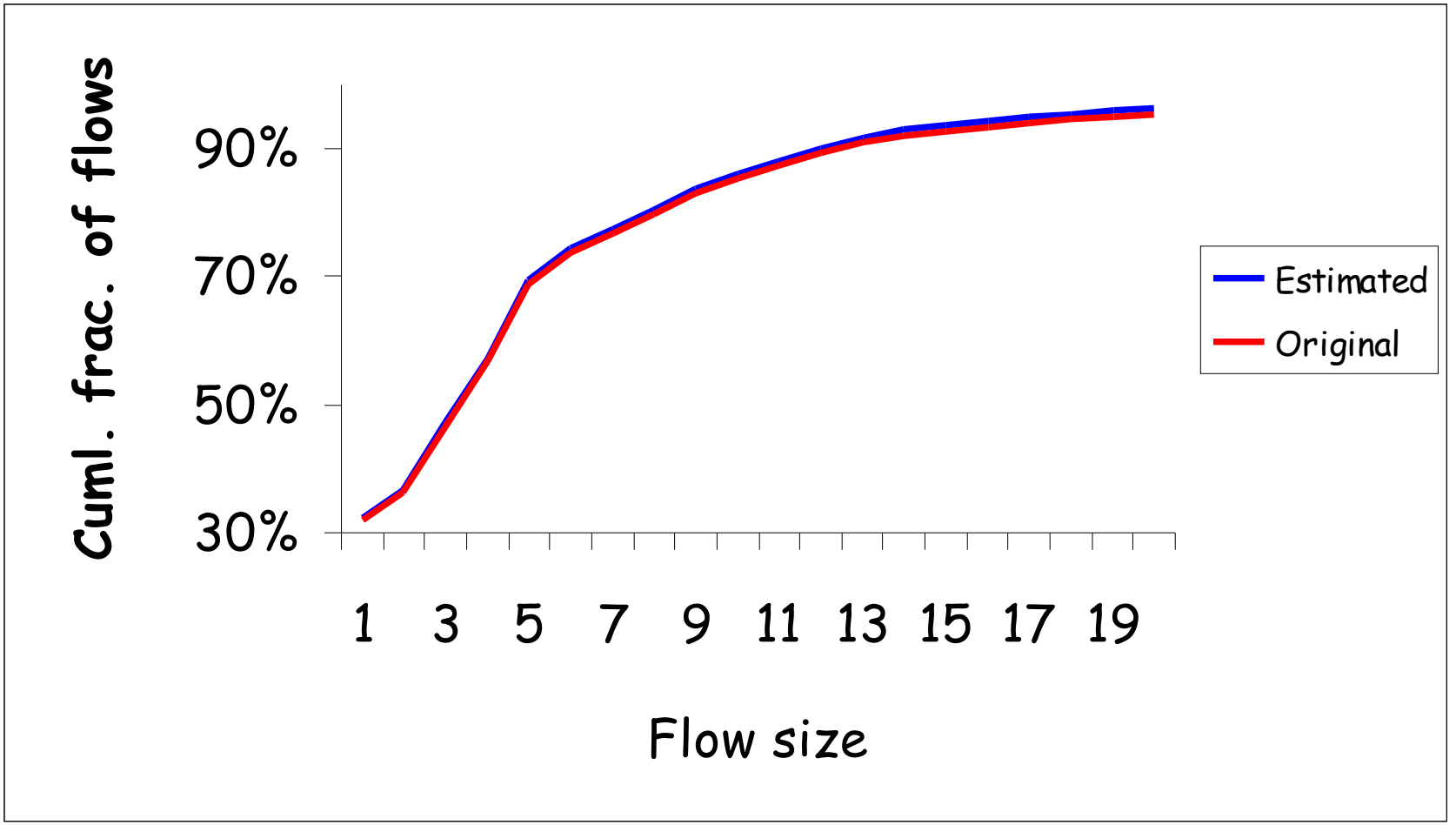
- Summary uses TCP Seq. Numb. + SYN flag information.
- Information much higher.



$$\mathrm{STD} = \frac{C(\mathbf{B}, \vec{\theta})}{\sqrt{n}}$$

2×10^6 sampled flows needed for small STD lower bound (STD ≥ 0.025).

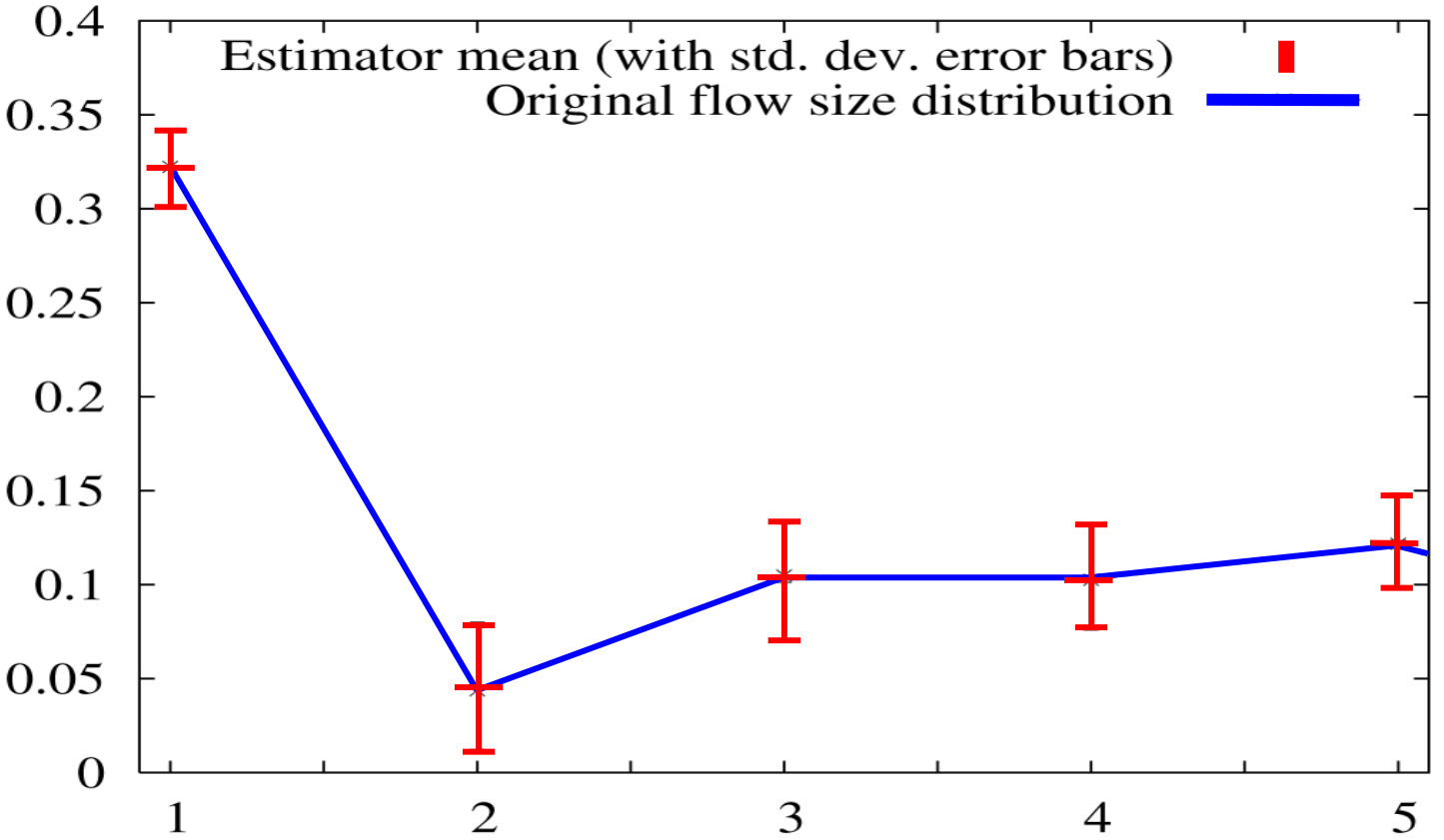
Estimates from TCP protocol info.



Parameters: pkt sampling rate = 1/200, 1 **million** sampled flows (10^6)

Practical implication

- Maximum Likelihood Estimator
  - 1 **million** sampled flows; pkt sampling rate = 1/200



Outline

- Accuracy of previous works
- Information model of sampling
- Fisher information and Cramer-Rao bound
- Solve the problem: Gathering more information from packet samples
- Conclusion

Conclusion:

- Fisher information:
  - Measure information gain.
  - Help design flow size distribution estimators.
- Packet sampling:
  - Shows packet counts summaries insufficient for accurate estimates.
    - Protocol information needed.
    - Some TCP fields carry valuable information.