Temporal Update Dynamics under Blind Sampling

Xiaoyong Li, Daren B.H. Cline and Dmitri Loguinov

Internet Research Lab
Computer Science Department
Texas A&M University, College Station, TX 77843

Apr. 29, 2015

Agenda

- Introduction
- Overview
- Age Sampling
- Comparison Sampling: Constant Interval
- Comparison Sampling: Random Interval
- Conclusion

Introduction

- Source objects in many distributed systems experience periodic modification
  - In response to user actions, real-time events
  - Examples: web pages, DNS record
- The update process in the source can be viewed as a stochastic process N_U
  - We are interested in estimating the inter-update distribution $F_U(x)$ using a downloading process N_S with inter delay S_1, S_2, …
  - Previous work use Poisson N_U and constant S_i
- Challenges
  - Non-Poisson updates
  - Blind sampling: the inter-update delay is hidden from the observer

Motivation

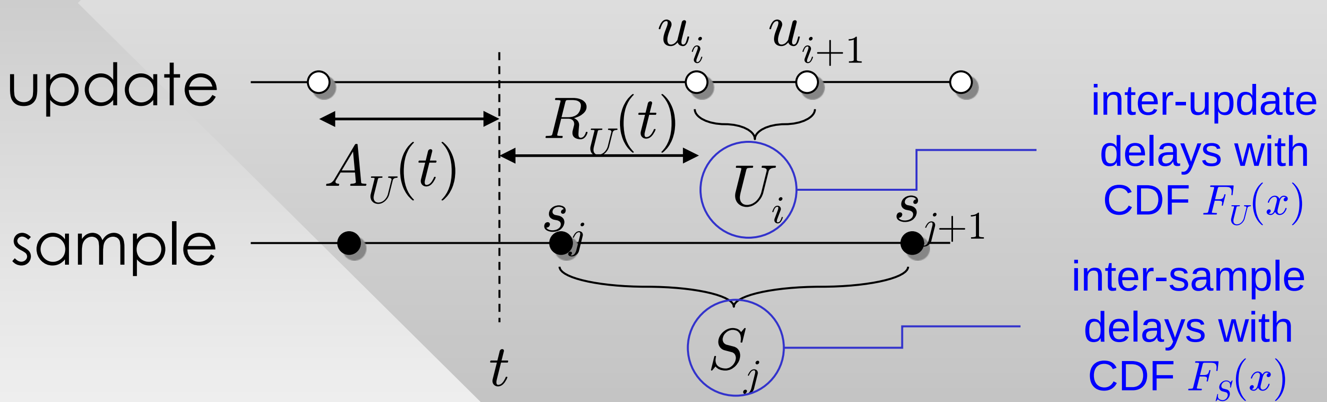
- Search engines
  - Periodically revisit web pages to reduce their staleness in the index
  - Need $F_U(x)$ to determine the download bandwidth to maintain staleness below a certain threshold
  - Exponential assumption leads to errors in the download bandwidth that are two orders of magnitude
- Data Centers
  - Replicate quickly changing databases among multiple nodes
  - Individual replica may not stay fresh for a long period because of the highly dynamic nature of the source
  - How many replicas should be queried by clients to obtain certain consistent level?

Agenda

- Introduction
- Overview
- Age Sampling
- Comparison Sampling: Constant Interval
- Comparison Sampling: Random Interval
- Conclusion

Notation

- Model
  - Source experiences random updates via process N_U
  - Observer samples the content via process N_S



- Age of U at t : $A_U(t)$ with distribution $G_U(x)$ as $t \to \infty$
- Obtain $G_U(x)$ and get $F_U(x)$ by inversing the following equation

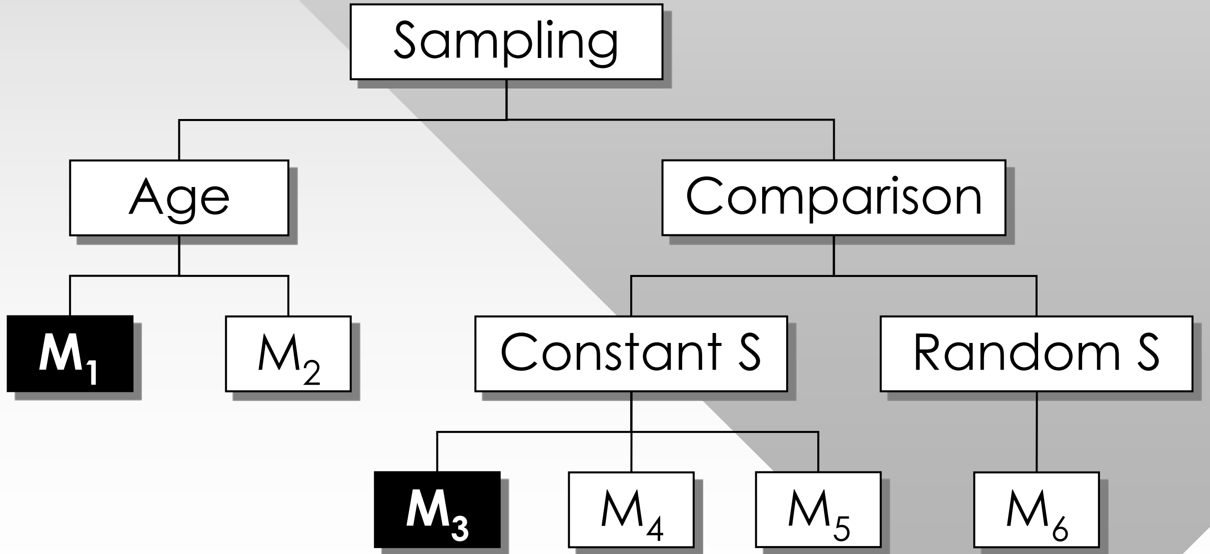
$$G_U(x) = \frac{1}{E[U]} \int_0^x (1 - F_U(y))dy$$

Assumptions

- We only have one update and download sequence (one sample-path), which leads to a possibility of phase-lock
  - $U_i = 1$ for $i > 0$ and $S_j = 2$ for $j > 0$
  - Update ages observed are all zero
- Definition 1: A random variable X is called *lattice* if there exists a constant c such that X/c is always an integer
- Assumption 1: At least one of U and S is non-lattice
  - The condition is satisfied with any continuous random variable, including exponential U in previous works

Roadmap

- Age sampling
  - Has access to the last-modification timestamp, which gives the update age at each sampling point $A_U(s_j)$
- Comparison sampling
  - Only use binary values between two successive samples



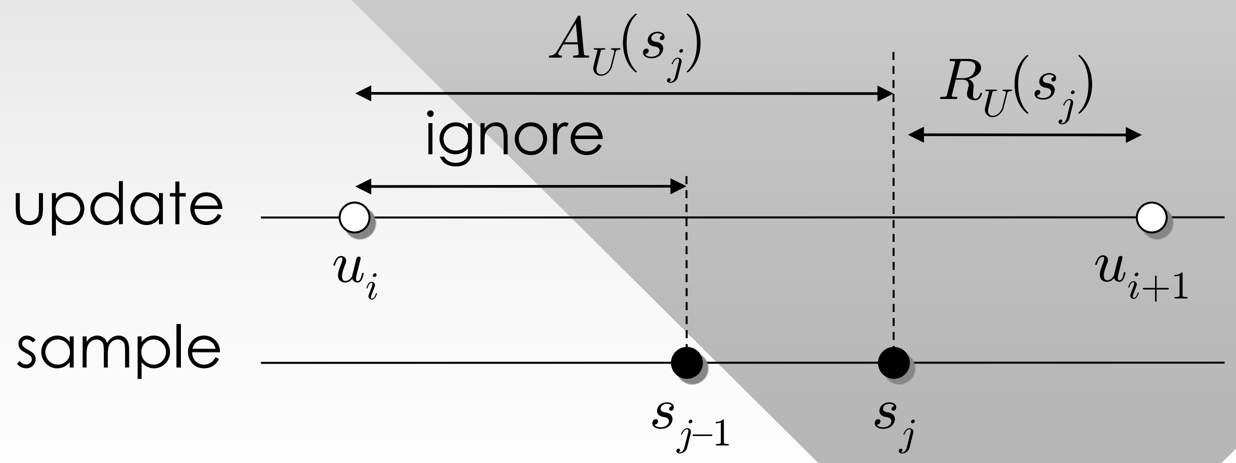
shaded boxes indicate Poisson-only techniques

Agenda

- Introduction
- Overview
- Age Sampling
- Comparison Sampling: Constant Interval
- Comparison Sampling: Random Interval
- Conclusion

M1

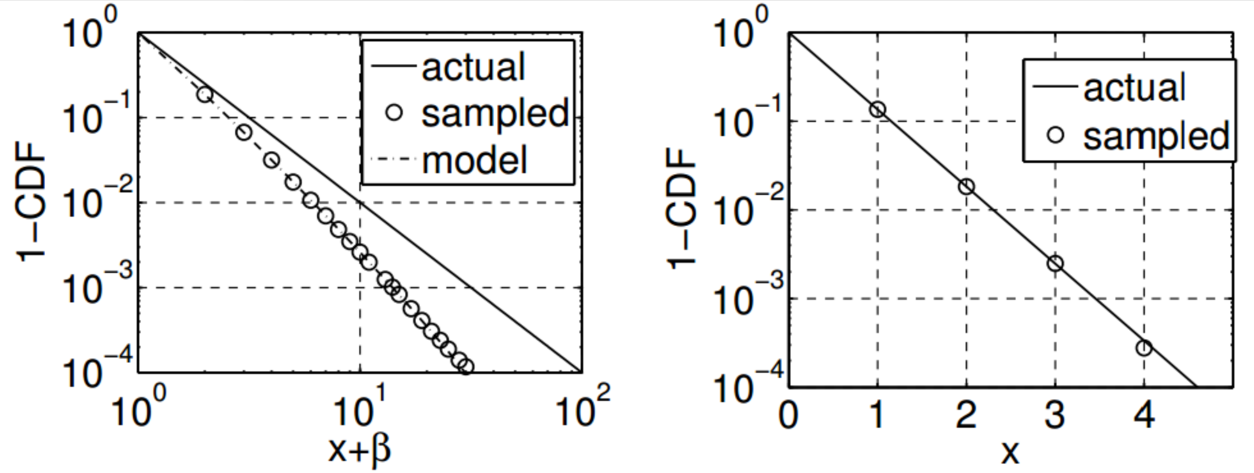
- When multiple sample points land in the same update interval, only retain the one with largest age
  - Keeps a subset of age samples
  - Proposed by previous studies to under Poisson updates
  - Used to estimate the mean of the update



M1

- Theorem 2: The tail distribution of the samples collected by M1 converges in probability to:

$$\bar{G}_1(x) = \frac{E[G_U(x+S)] - G_U(x)}{E[G_U(S)]}$$



(a) Pareto U (b) exponential U

Fig. 5. Simulation results of M_1 under exponential S ($\lambda = 1, \mu = 2$).

Bias in M1

- The tail of M1 is "sandwiched" between the update and age tails

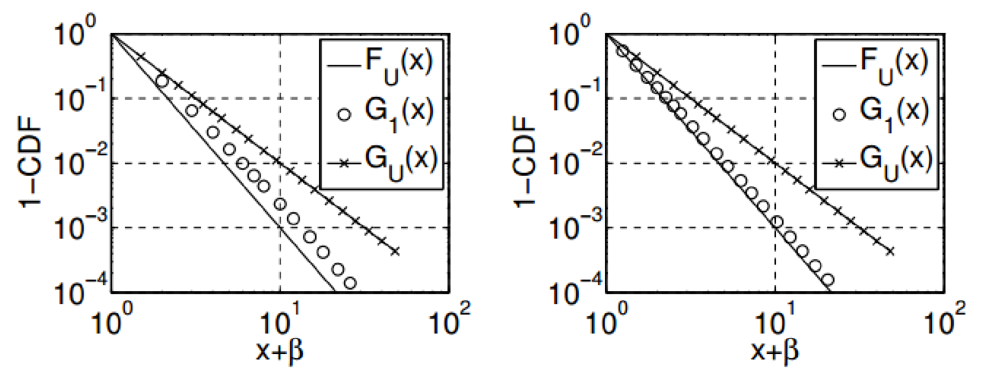


Fig. 6. Tail sandwich of M_1 under Pareto updates and constant S ($\mu = 2$).

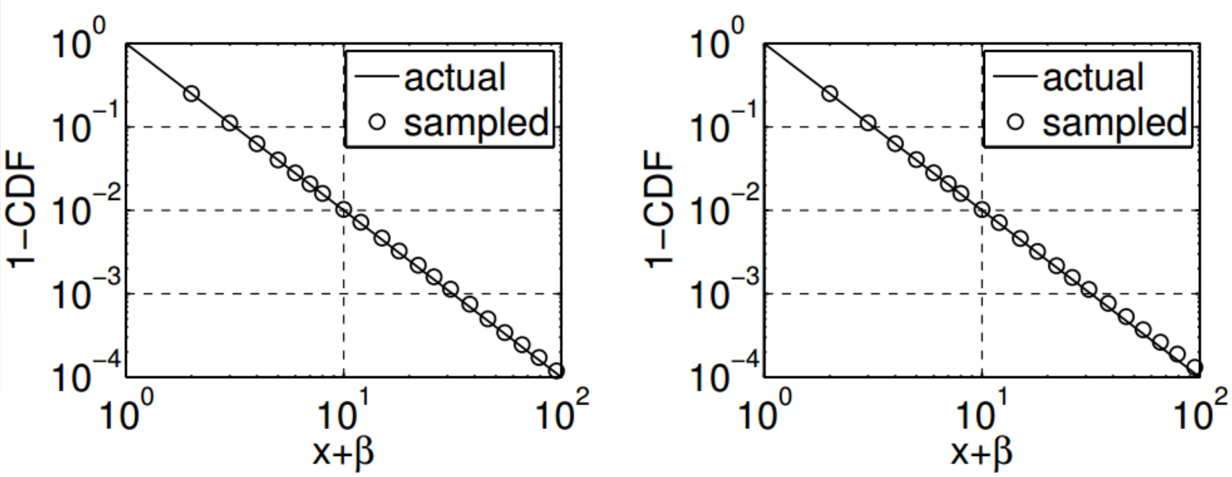
- The fraction of age samples retained by M1 :

$$p = P(R_U < S) = E[G_U(S)]$$

- For $p \to 1$, variable D_1 sampled by M1 converges in distribution to A_U. For $p \to 0$ and mild conditions on S, variable D_1 converges in distribution to U

M2

- Instead of using the largest age sample for each detected update, M2 use all available ages
- Theorem 5: Method M2 is consistent with respect to the update age distribution.



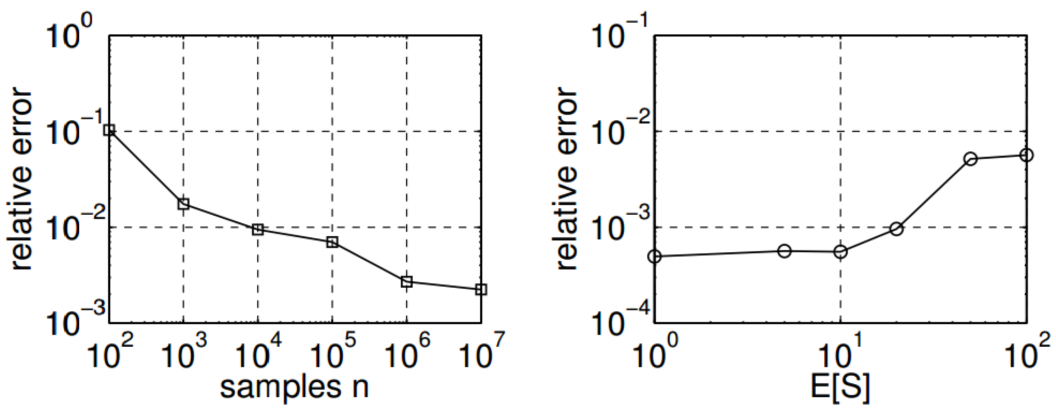
(a) exponential S (b) constant S

Fig. 7. Verification of (12) under Pareto updates and $\lambda = 1$.

Computer Science, Texas A&M University

M2

- M1 and M2 has the same network overhead because they both have to contact the source $N_S(t)$ times
- Effect of the observation window T and expected sampling interval S
  - relative error on the update age mean



(a) impact of T ($\lambda = 1$) (b) impact of S ($T = 10\text{K}$)

Fig. 9. Average relative error of $\zeta(T)$ of M_2 under Pareto U and exponential S ($\mu = 2, m = 1000$).

Agenda

- Introduction
- Overview
- Age Sampling
- Comparison Sampling: Constant Interval
- Comparison Sampling: Random Interval
- Conclusion

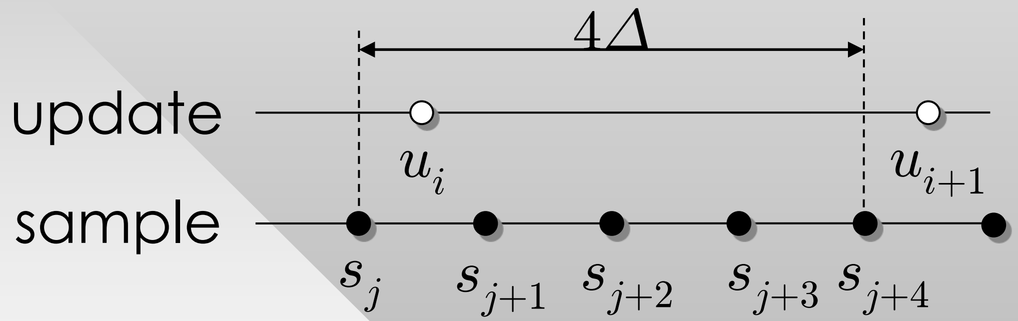
Basics

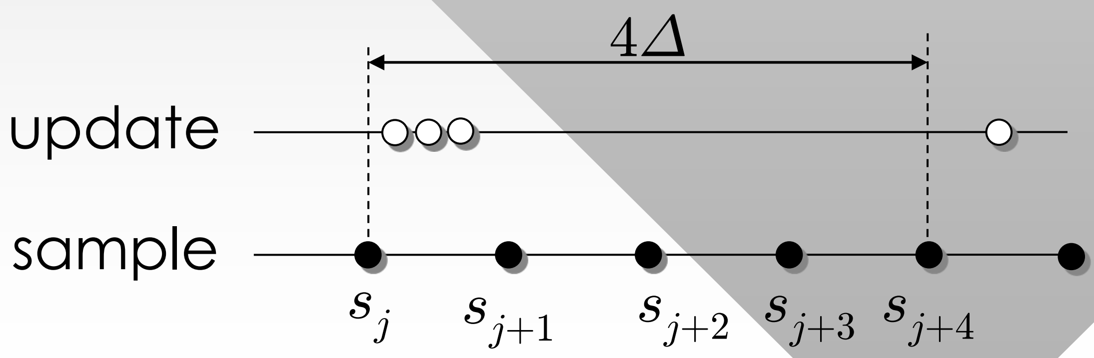
- Do not have access to age
- The inter-sample delay Δ is a constant
- Binary observations Q_{ij}
  - Indicates whether an update occurs between two sampling points s_i and s_j
- All observations related to update intervals are multiple of inter-sample delay $S = \Delta$
  - An estimator is Δ-consistent with respect to the target distribution if it can correctly reproduce it in all discrete points $x_n = n\Delta$ as $T \to \infty$

Computer Science, Texas A&M University

M3

- Round the distance between each adjacent pair of detected updates to the nearest multiple of Δ
  - Expected to produce the update distribution $F_U(x)$
  - Inaccurate when multiple updates occurs within one Δ





M3

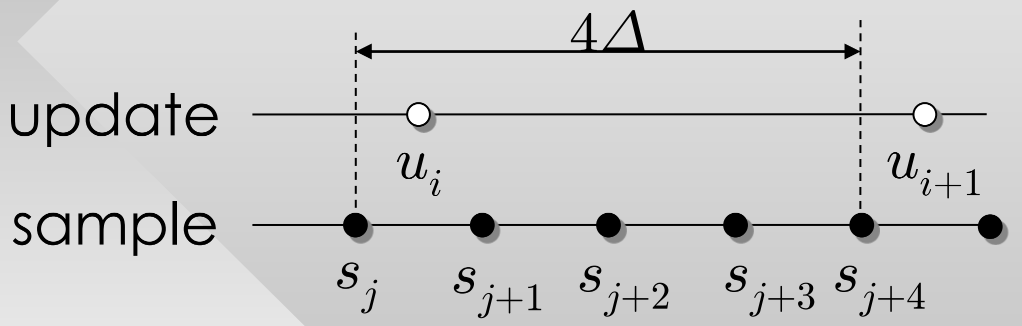
- Theorem 6: The tail distribution of M3 is a step-function

$$\bar{G}_3(x_n) = \frac{G_U(x_{n+1}) - G_U(x_n)}{G_U(\Delta)}$$

- Similar to M1, M3 is consistent when $F_U(x)$ is exponential
- When $\Delta \to \infty$, G_3 converges to $G_U(x)$
- When $\Delta \to 0$, G_3 converges to $F_U(x)$
- Neither scenario is usable in practice

M4

- Collect age samples at each sampling point
  - Four samples in the example: Δ, 2Δ, 3Δ, 4Δ



- Theorem 7: M4 is Δ-consistent with respect to the age distribution
  - The mean age of M3 is not necessarily larger than that of M4 e.g. Pareto update and $\Delta=1$, M3 and M4 produces mean age 1.33 and 1.63, respectively.

M5

- A closer look at M3 results $\bar{G}_3(x_n) = \frac{G_U(x_{n+1}) - G_U(x_n)}{G_U(\Delta)}$
- $G_U(x)$ can be recursively recovered using samples in M3
- Theorem 8: M5 is Δ-consistent with the age distribution.

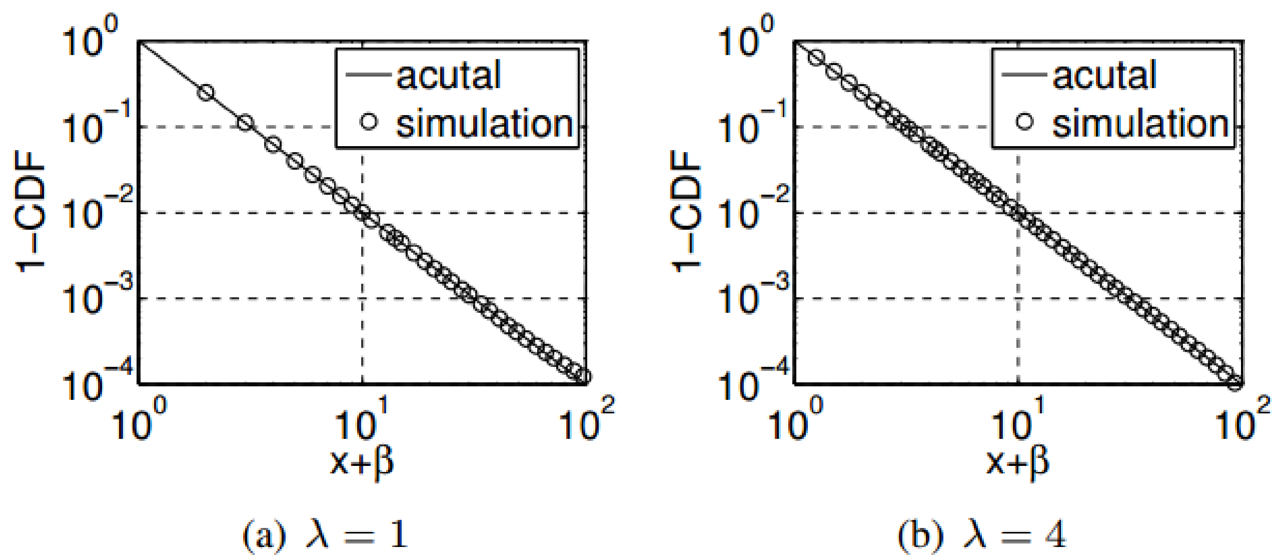


Fig. 12. Verification of (22) under Pareto U ($\mu = 2$).

Comparison between M4 and M5

- Weighted Mean Relative Difference between two distribution

$$W(T) = \frac{\sum_n |H(x_n, T) - G_U(x_n)|}{\sum_n (H(x_n, T) + G_U(x_n))/2}$$

- Kolmogorov-Smirnov statistic

$$\kappa(T) = \sup_x |H(x, T) - G_U(x)|$$

CONVERGENCE OF BOTH Δ-CONSISTENT METHODS UNDER PARETO U ($\mu = 2, \lambda = 1$)

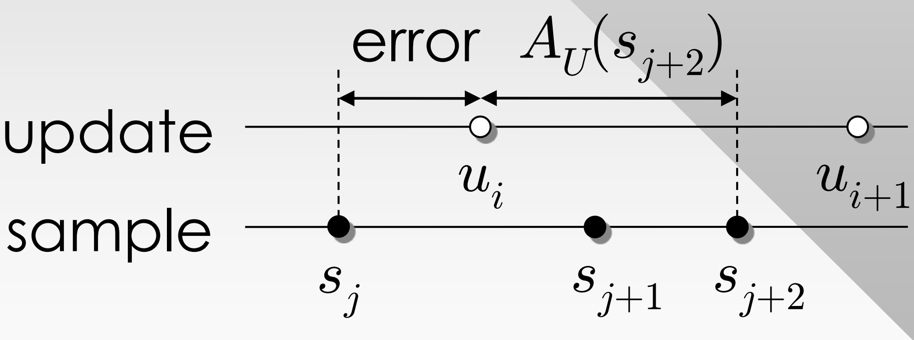
T	M_4		M_5	
	$w(T)$	$\kappa(T)$	$w(T)$	$\kappa(T)$
10^2	3.5×10^{-2}	6.4×10^{-2}	3.7×10^{-2}	6.7×10^{-2}
10^3	1.4×10^{-2}	2.2×10^{-2}	1.4×10^{-2}	2.2×10^{-2}
10^4	4.7×10^{-3}	7.2×10^{-3}	4.7×10^{-3}	7.3×10^{-3}
10^5	1.5×10^{-3}	2.4×10^{-3}	1.5×10^{-3}	2.4×10^{-3}
10^6	4.1×10^{-4}	5.8×10^{-4}	4.1×10^{-4}	5.8×10^{-4}
10^7	2.2×10^{-4}	2.6×10^{-4}	2.2×10^{-4}	2.6×10^{-4}

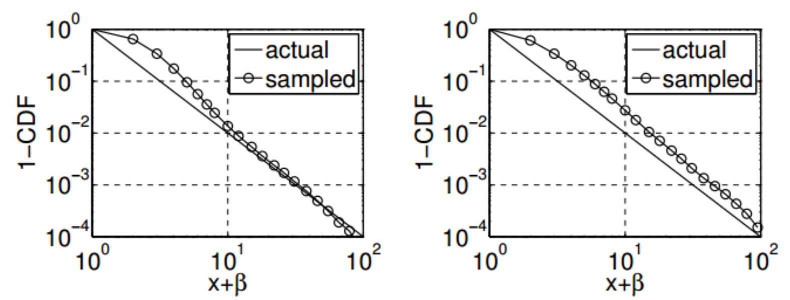
Agenda

- Introduction
- Overview
- Age Sampling
- Comparison Sampling: Constant Interval
- Comparison Sampling: Random Interval
- Conclusion

G-M4

- Straightforward Approach
  - Generalize M4 to random S
  - Approximate $A_U(s_j)$ by $s_j - s_j^*$; s_j^* is the most-recent sample point after which an update has been detected
  - Round-off error varies from interval to interval
  - Biased





(a) exponential S (b) Pareto S

Fig. 14. Bias of G-M_4 with Pareto updates ($\mu = 2, \lambda = 1$).

M6

- For a user defined constant h and fixed $y_n = nh$, count the number of inter-sample $W(y_n)$ with distances $s_j - s_i$ that round up to y_n and the number of them with an update $Z(y_n)$. Define $G_6(y_n) = Z(y_n)/W(y_n)$
  - Use n^2 samples, while all other methods have linear overhead
- Theorem 9: When $h \to 0$, and $F_S(x) > 0$, M6 is consistent with respect to the age distribution

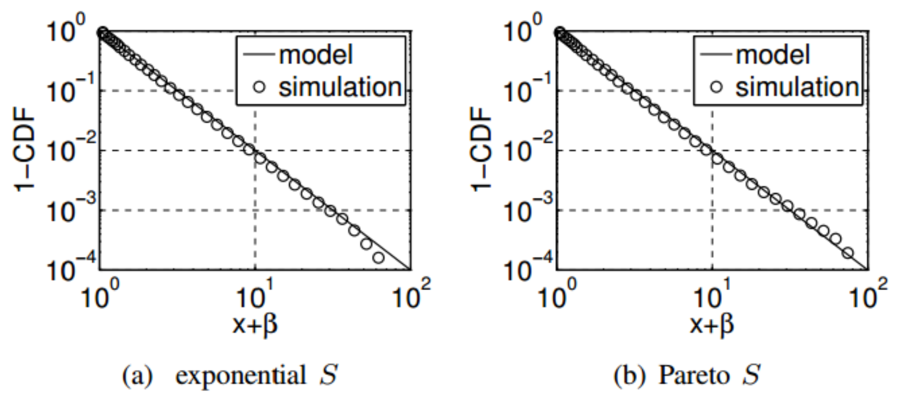


Fig. 15. Simulations of M_6 under Pareto updates ($h = 0.05, \mu = 2, \lambda = 1$).

Conclusion

- We studied the problem of estimating the update distribution at a remote source under blind sampling
- We analyzed prior approaches and showed them to be biased under general conditions
- We introduced novel modeling techniques and proposed several unbiased algorithms
- Future work includes analysis of convergence speed, investigation of non-parametric smoothing techniques for density estimation

Questions?