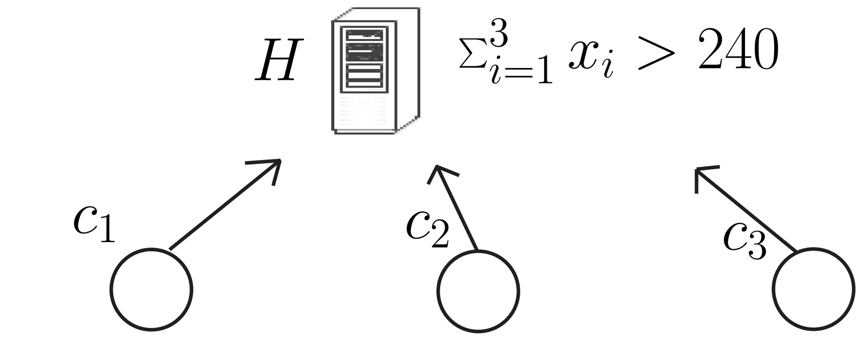
Efficient Threshold Monitoring for Distributed Probabilistic Data

Mingwang Tang, Feifei Li, Jeff M. Phillips, Jeffrey Jestes

THE UNIVERSITY OF UTAH®

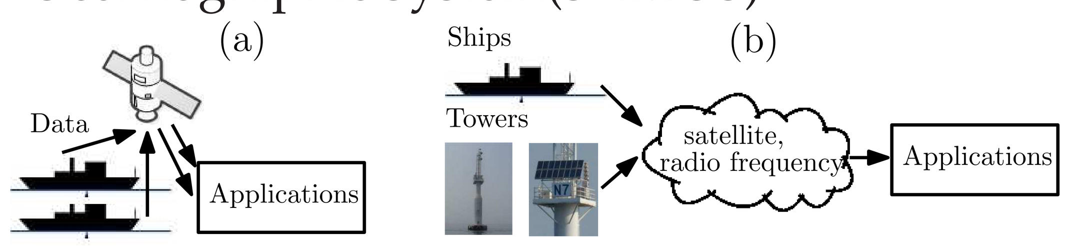
Introduction

- Distributed Threshold Monitoring (DTM):



	c_1	c_2	c_3
t_1	70	80	75
t_2	70	90	90
⋮	⋮	⋮	⋮
t_T	70	80	70

- The Shipboard Automated Meteorological and Oceanographic System(SAMOS)



- S. Jeyashanker et al., Efficient Constraint Monitoring Using Adaptive Thresholds, [ICDE2008]

- Attribute uncertain model and flat model

tuples	attribute score
d_1	$X_1 = \{(v_{1,1}, p_{1,1}), (v_{1,2}, p_{1,2}) \dots (v_{1,b_1}, p_{1,b_1})\}$
d_2	$X_2 = \{(v_{2,1}, p_{2,1}), (v_{2,2}, p_{2,2}) \dots (v_{2,b_2}, p_{2,b_2})\}$
.	...
d_t	$X_t = \{(v_{t,1}, p_{t,1}), (v_{t,2}, p_{t,2}) \dots (v_{t,b_t}, p_{t,b_t})\}$

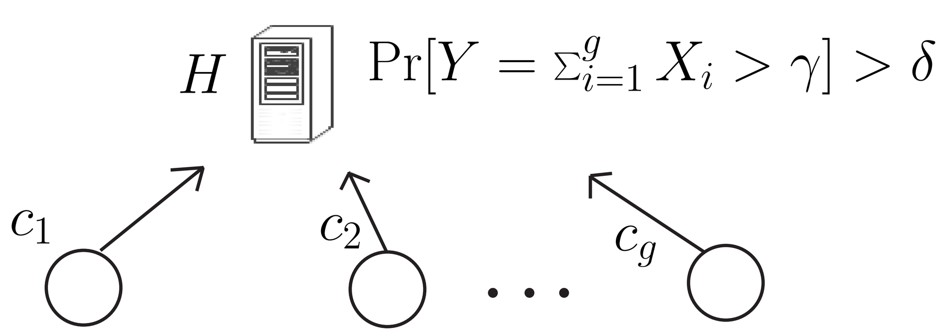
Exact Methods

- Compute $Pr[Y \geq \gamma]$ exactly: each client c_i simply sends X_i to H and H sum X_i and check against the (γ, δ) threshold.
- Both communication and computation expensive. Naively, it could be in $O(n^g)$ (n is the maximal $|X_i|$)
- When X_i's are represented by continuous pdfs, we leverage on the characteristic function of X_i to compute the pdf of Y.

Improved Adaptive Method (Iadaptive)

- $\sum_{i=1}^{g} \ln M_i(\beta 1) \leq \ln \delta + \beta_1 \gamma$
- $\sum_{i=1}^{g} \ln M_i(\beta 2) \leq \ln(1-\delta) + \beta_2 \gamma$
- A counter e of alarm instances is maintained in each period of k time instances.
- Periodically decide which monitoring instance to run and set the optimal value of β_1 and β_2

Distributed Probabilistic Threshold Monitoring (DPTM)

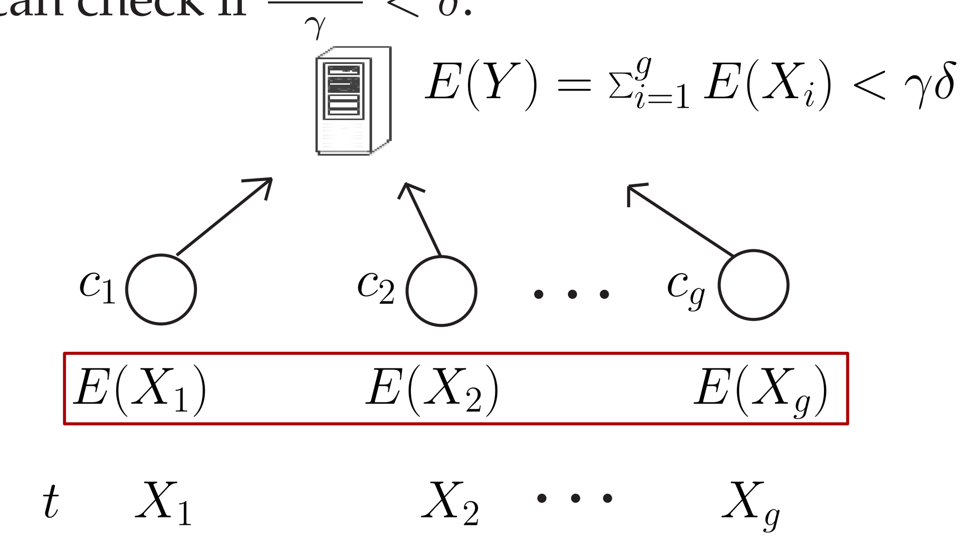


t_1	$X_{1,1}$	$X_{2,1}$	$X_{g,1}$
t_2	$X_{1,2}$	$X_{2,2}$	$X_{g,2}$
⋮	⋮	⋮	⋮
t_T	$X_{1,T}$	$X_{2,T}$	$X_{g,T}$

- $\Pr[Y > \gamma] <$ upperbound $< \delta$
  $\Pr[Y > \gamma] >$ lowerbound $> \delta$
  give two deterministic monitoring instances.
- The lowerbound (upperbound) is a function of some deterministic values derived based on X_i's → use DTM methods
- When derived deterministic monitoring instances fail to make a decision, still expensive to compute Y even with all X_i's → use sampling methods!

Baseline Method Based on Markov bound (Madaptive)

- Markov's inequality: $\Pr[Y > \gamma] \leq \frac{\mathbf{E}(Y)}{\gamma}$.
- H can check if $\frac{\mathbf{E}(Y)}{\gamma} < \delta$.



Improved Method

I One-sided Chebyshev's inequality:

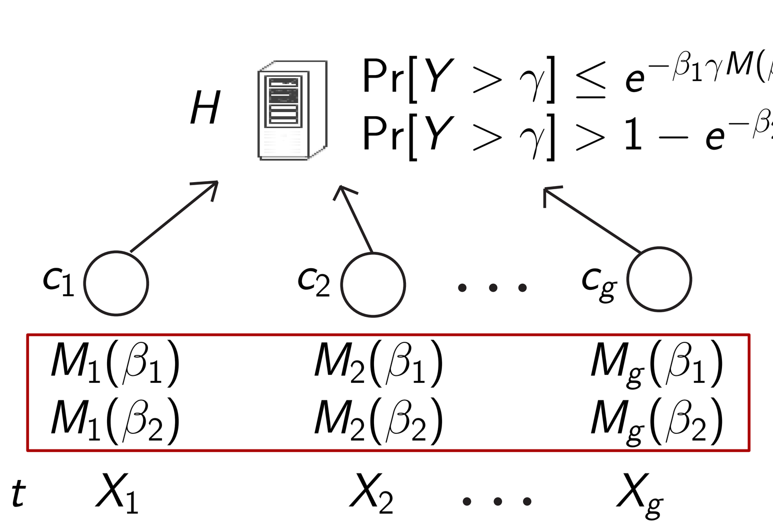
$\Pr[Y > \gamma] < \frac{\mathbf{Var}(Y)}{\mathbf{Var}(Y) + (\gamma - \mathbf{E}(Y))^2} . (\gamma > \mathbf{E}(Y))$

$\Pr[Y > \gamma] > 1 - \frac{\mathbf{Var}(Y)}{\mathbf{Var}(Y) + (\mathbf{E}(Y) - \gamma)^2} . (\mathbf{E}(Y) > \gamma)$

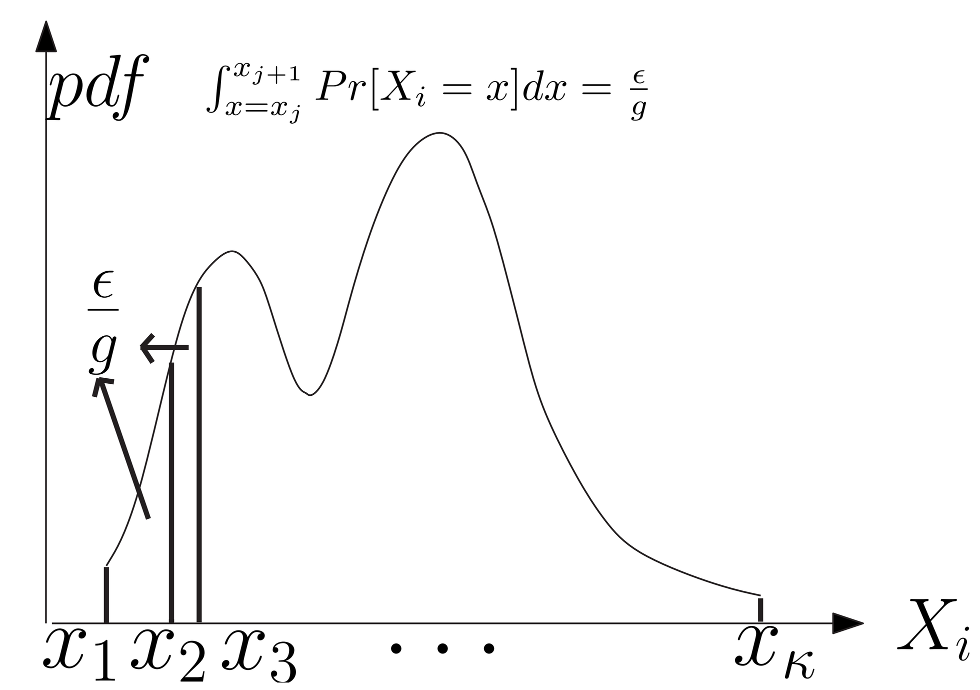
II The Chernoff bound using the moment-generating function.

$M(\beta) = \mathbf{E}(e^{\beta Y}) = \prod_{i=1}^{g} M_i(\beta)$ for any $\beta \in R$.

for any $\beta_1 > 0$ and $\beta_2 < 0$:



Deterministic Distributed ϵ-Sample (DDϵS)



- DDϵS gives $|\Pr[\tilde{Y} > \gamma] - \Pr[Y > \gamma]| \leq \epsilon$ with probability 1 in $O(g^2/\epsilon)$ bytes.

A Randomized Improvement of DDϵS (αDDϵS)

- $\int_{x=x_{i,j}}^{x_{i,j+1}} Pr[X_i = x]dx = \alpha$
- Choose the smallest sample point at random (within x_α).
- $\Pr[|\Pr[\tilde{Y} > \gamma] - \Pr[Y > \gamma]| \leq \epsilon] > 1 - \phi$ in $O(\frac{g}{\epsilon}\sqrt{2g \ln \frac{2}{\phi}})$ bytes.

Random Distributed ϵ-Sample (RDϵS)

- H asks for a random sample x_i from each client w.r.t. the distribution of X_i
- Repeating this sampling $\kappa = O(\frac{1}{\epsilon^2} \ln \frac{1}{\phi})$ times.
- $\Pr[|\Pr[\tilde{Y} > \gamma] - \Pr[Y > \gamma]| \leq \epsilon] \geq 1 - \phi$ using $O(\frac{g}{\epsilon^2} \ln \frac{1}{\phi})$ bytes.

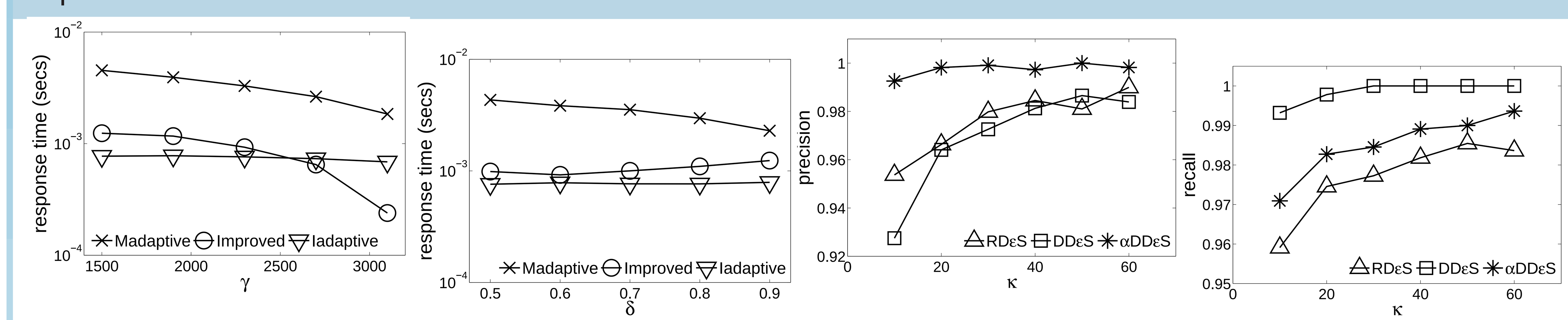
Default Experimental Parameters

Symbol	Definition	Default Value
τ	grouping interval	300
g	number of clients	10
δ	probability threshold	0.7
γ	score threshold	30% alarms (230g for WD)
κ	sample size per client	30

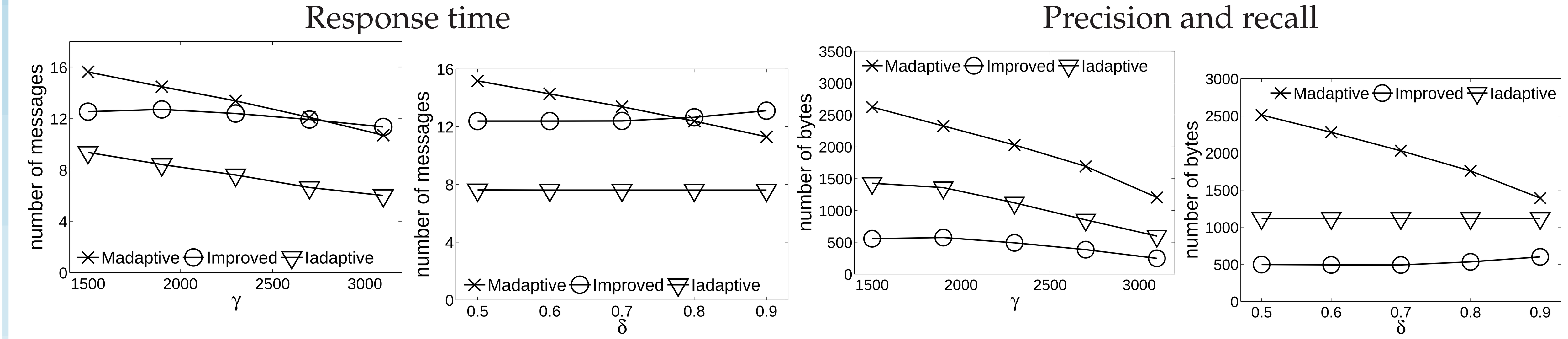
Datasets

- Real datasets (11.8 million records) from the SAMOS project.
- Each record contain four measurements: WD, WS, SS, TEM, which leads to four single probabilistic attribute datasets.

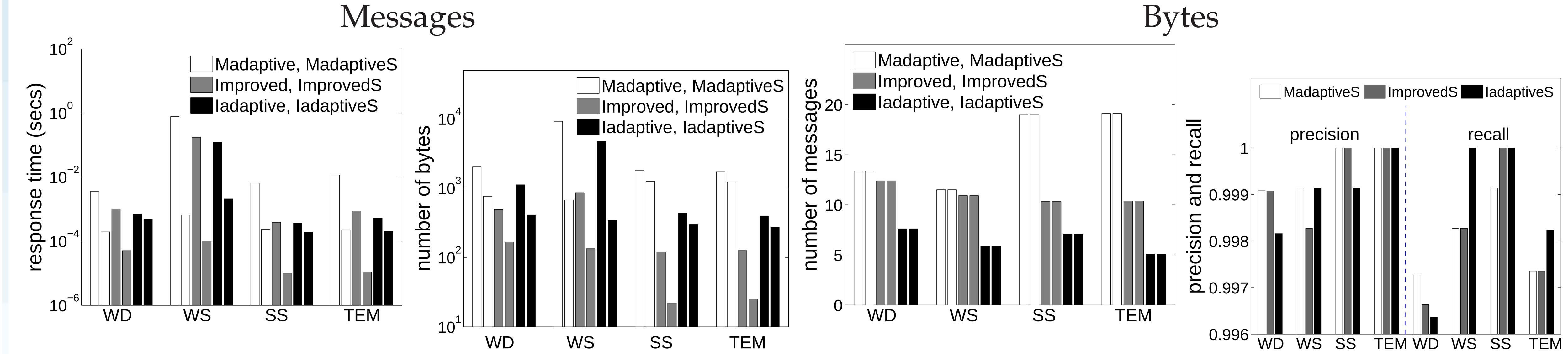
Experiments



Response time
Precision and recall



Messages
Bytes



Performance of all methods