Learning Graphical Models from a Distributed Stream

Yu Zhang[1], Srikanta Tirthapura[1], Graham Cormode[2]
[1] *Iowa State University*
[2] *University of Warwick*

Overview

This paper considers the problem of maintaining machine learning model from a distributed stream over a network with high latency. Under this scenario, the algorithm should be:

- Communication-efficient among distributed sites
- Able to accurately keep track of the continuously changing model

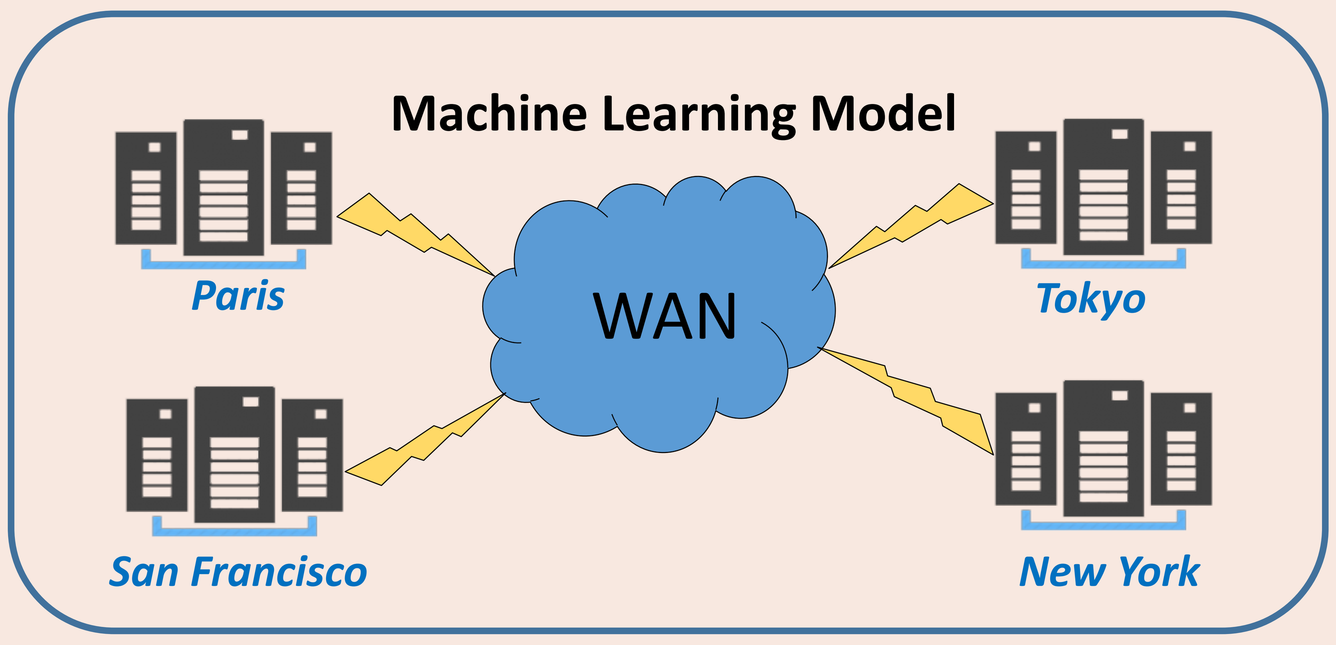

Figure 1: Machine learning from distributed stream over WAN

Problem Statement

Goals of Learning

- Learn the probabilistic graphical model [1]: Bayesian Networks [2]
- Discuss the problem of estimating parameters: conditional probability distribution (CPD) of each variable given its parents

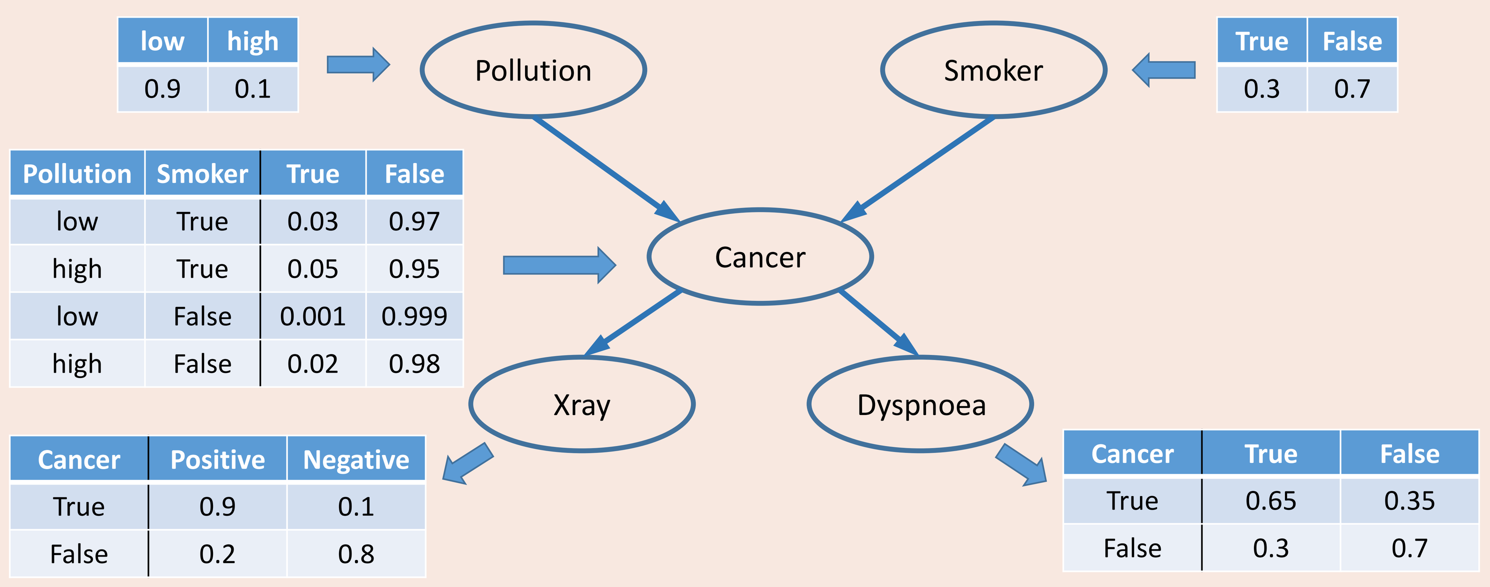

Figure 2: Cancer Bayesian Network with CPDs

Distributed Stream

- Each site receives an individual stream of observations
- A coordinator maintains a Bayesian Network and answers queries

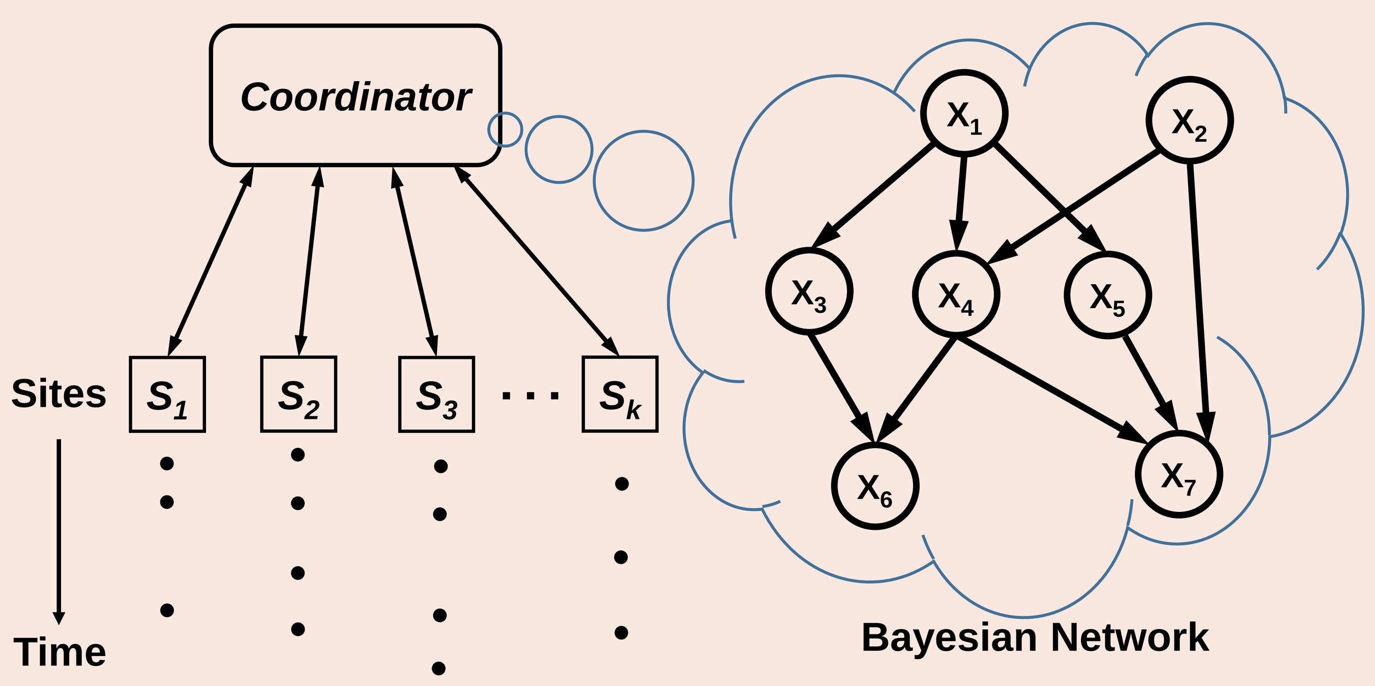

Figure 3: Learn Bayesian Network Parameters from Distributed Stream

Research Challenge and Solution Idea

Exact Counting: Report each event exactly to the coordinator
Drawback: High network communication cost becomes the bottleneck
Approximate Counter [3]: Substantially reduces the number of messages sent but maintains the value of counter with approximate error
Challenge: Design algorithms using approximate counters to reduce the communication while achieving the joint distribution as accurate as MLE
Solution Idea: Error and communication as an optimization problem

$$\text{Minimize communication} \quad \text{s.t.} \quad e^{-\epsilon} \le \frac{\tilde{P}(\boldsymbol{x})}{\hat{P}(\boldsymbol{x})} \le e^{\epsilon}$$

where ϵ is the error budget, $\boldsymbol{x}$ is the input vector, $\tilde{P}(\boldsymbol{x})$ is the probability using approximate counters and $\hat{P}(\boldsymbol{x})$ is the probability using MLE.

Communication Efficient Algorithms

We propose a set of algorithms that are different based on how they set the error parameter for each approximate counter.
Baseline: Divides the error budget uniformly across all variables
Uniform: Improved randomized analysis by utilizing the property of approximate counter: unbiased and variance bounded
Non-uniform: Uneven error parameter assignment, account for the cardinalities of different variables and the parents

Table 1: Theoretical Result Summary
(ϵ: error budget, n: number of variables, m: total number of events)

Algorithm	Approx. Factor of Counters	Communication (msgs.)
Exact	1	$O(mn)$
Baseline	$O(\frac{\epsilon}{n})$	$O(\frac{n^2}{\epsilon}\log m)$
Uniform	$O(\frac{\epsilon}{\sqrt{n}})$	$O(\frac{n^{3/2}}{\epsilon}\log m)$
Non-uniform	uneven	at least Uniform

Experiment

Experiment Setting

- Live implementation on an EC2 cluster on Amazon Web Services
- Training data is generated based on the ground truth
- Testing data contains 1000 events and estimate the probability of each event using parameters learnt by the distributed algorithms
- Error budget is set to 0.1

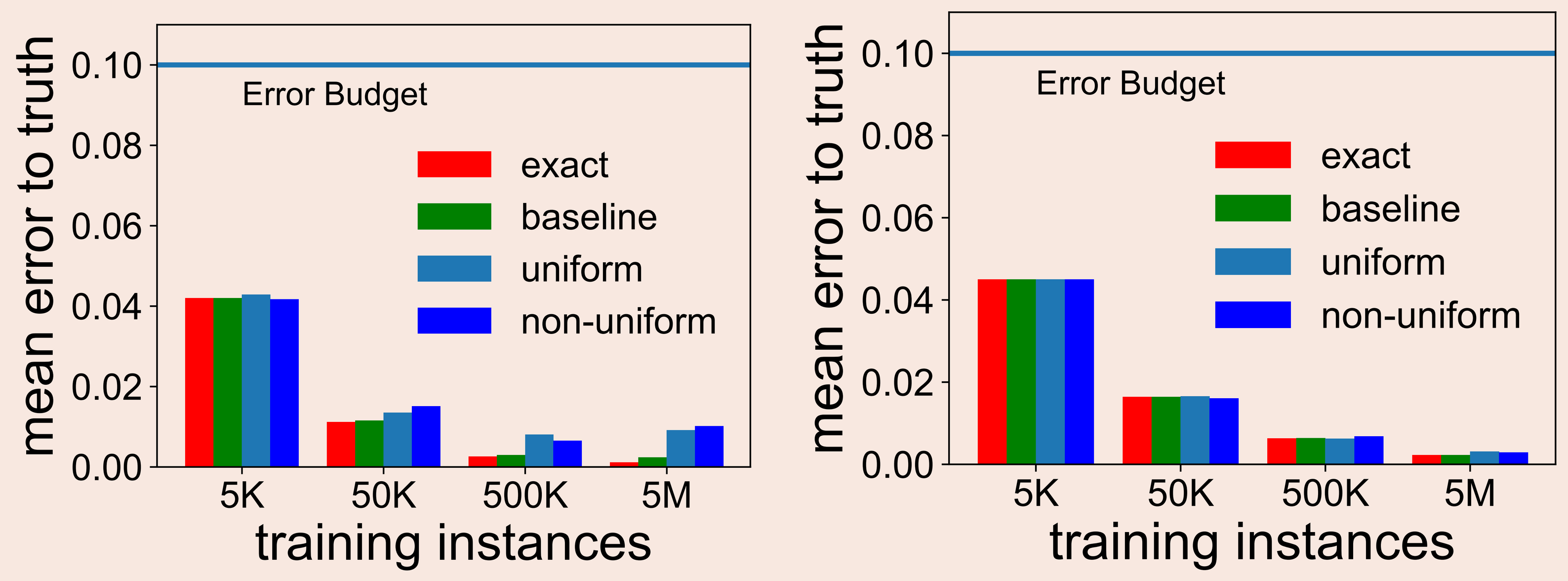

Figure 4: Error to the ground truth. (Left: Alarm, Right: Munin)

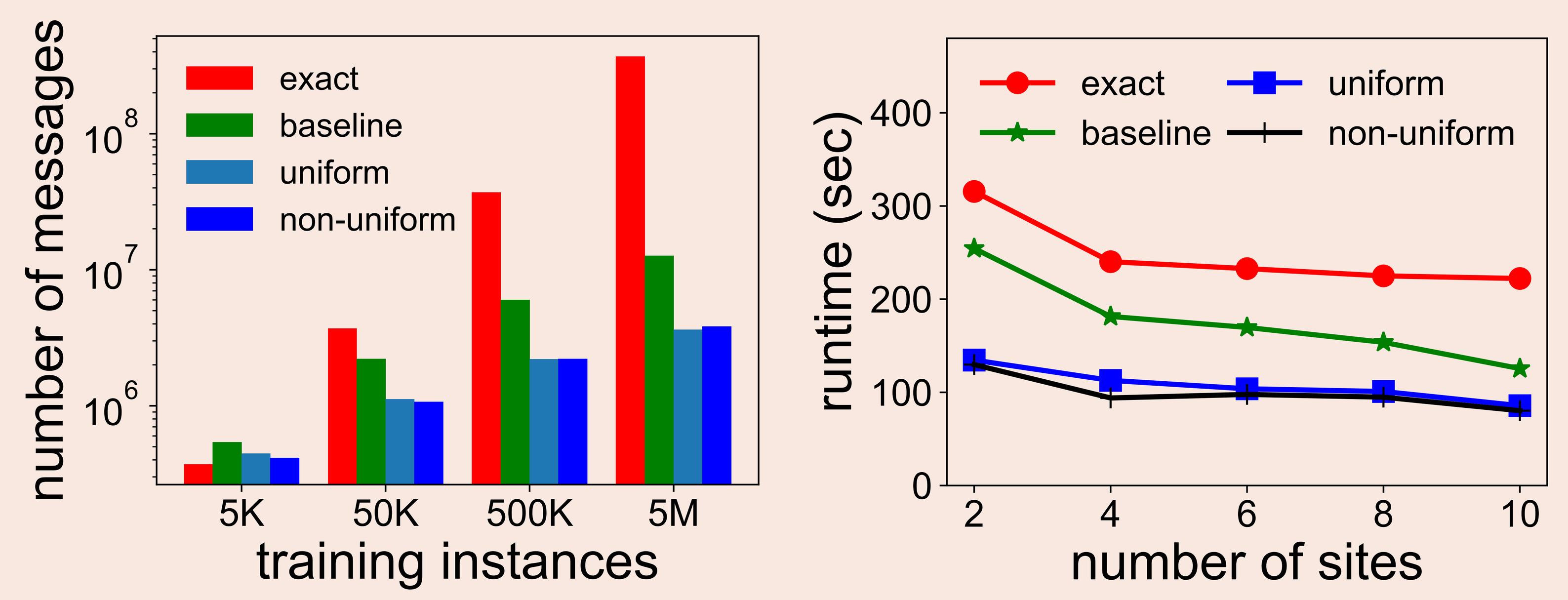

Figure 5: Communication cost and training time of different algorithms

Conclusion

- Compared to maintain the exact MLE, our algorithms reduce the communication that is only *logarithmic* in the number of events
- Offer provable guarantees on the estimation of joint probability
- Provide optimal strategy for maintaining parameters by utilizing the approximate counter properties and the information of Bayesian Network structure
- Empirically show the reduced communication and similar prediction errors as the MLE for estimation and classification tasks

References

[1] D Koller and N Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

[2] J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proc. of Cognitive Science Society (CSS-7)*, 1985.

[3] Zengfeng Huang, Ke Yi, and Qin Zhang. Randomized algorithms for tracking distributed count, frequencies, and ranks. In *PODS*, 2012.

Acknowledgements

The work of Dr. Cormode is supported in part by European Research Council grant ERC-2014-CoG 647557 and a Royal Society Wolfson Research Merit Award, and of Yu Zhang and Dr.Tirthapura are supported in part by the National Science Foundation through grants 1527541 and 1725702.