[MSc Thesis]
Developing a Probabilistic Solver for Approximate Traffic Monitoring

francesco.dadalt@inf.ethz.ch

June 2024

1 Context

Consider a data stream of key-value pairs (e.g. IP-addresses and IP-packet-payloads). In various settings the need arises to compute the sum over all values associated to some key (i.e. the total data-volume sent by some IP-address), for example when one needs to detect malicious users in the network that do not behave fairly or when detecting DDoS attacks. We refer to the sum of values associated to a keys as its volume. Keeping track of the exact volume for each key can become very memory-intensive when one has for example millions of unique Internet flows going over a network link every second. To that end, a zoo of probabilistic compressive algorithms (referred to as sketches), that solve the problem of estimating the volumes associated to keys in the stream, have been developed. The goal of this thesis is to develop a probabilistic solver that analyzes a specific class of volume-sketches.

2 Problem

The exact problem is the following: We consider a class of probabilistic compressive data structures as used by the CountMin-Sketch [1], Count-Sketch [2], and CountBayes-Sketch [3] amongst others. These data structures consist of a two-dimensional array of counters. If we flatten these counters into one big array c then the act of compressing the volumes associated to keys a in a data stream can be described as

$$c = H\ a$$

where H is a random projection matrix, usually implicitly defined by a pseudorandom hash function. On a high level, sketching algorithms compute $\hat{H}^{-1}c = \hat{H}^{-1}Ha \approx a$ and therefore attempt to recover a from c. This recovery is however in general lossy because $dim(c) < dim(a)$. We are interested in computing information- and probability-theoretic metrics regarding the operations $c = H\ a$ and $\hat{H}^{-1}c$ such as the information gain from observing c, optimal Bayesian predictors for a, and information loss arising due to the use of sub-optimal pseudorandom hash functions for H.

3 Your Task

Your task is to develop a solver that computes exact Bayesian posterior probability distributions for the volumes of keys a. The focus lies of precision; computational efficiency is a secondary interest. To that end you will perform a theoretic analysis of the problem at hand and construct a Markov-Chain Monte-Carlo (MCMC) based algorithm that generates samples of the posterior Bayesian probability distributions. In particular, based on internal research you will likely have to implement Reversible-Jump MCMC which allows generating samples from variable-dimension spaces. The desired end-goal is a tool that can be used to provide insights into the theoretic nature of volume-sketching algorithms. Such a tool could be of use to the research community in order to better understand bounds on optimal performance and what kind of issues lead to performance degradation.

4 What You Gain

You gain hands on experience with MCMC methods and in particular RJ-MCMC which are cornerstone techniques used in stochastic simulation, Bayesian statistics, financial statistics, and genomics to name a few. Furthermore, excellent completion of the project should give you a good chance at publishing your work at a conference.

5 Requirements

- Strong probability-theory background
- Comfortable with MCMC and Bayesian inference
- Comfortable with linear algebra
- Willing to learn about RJ-MCMC

6 Reach Out

Send motivational letter, CV, and academic transcripts to francesco.dadalt@inf.ethz.com.

[1] Graham Cormode and S. Muthukrishnan. "An improved data stream summary: the count-min sketch and its applications". In: *Journal of Algorithms* 55.1 (2005), pp. 58–75. ISSN: 0196-6774. DOI: https://doi.org/10.1016/j.jalgor.2003.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0196677403001913.

[2] Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding Frequent Items in Data Streams". In: *Automata, Languages and Programming*. Ed. by Peter Widmayer et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 693–703. ISBN: 978-3-540-45465-6.

[3] Francesco Da Dalt, Simon Scherrer, and Adrian Perrig. "Bayesian Sketches for Volume Estimation in Data Streams". In: *Proc. VLDB Endow.* 16.4 (Dec. 2022), pp. 657–669. ISSN: 2150-8097. DOI: 10.14778/3574245.3574252. URL: https://doi.org/10.14778/3574245.3574252.