Mergeable Summaries

Pankaj K. Agarwal
Duke University
pankaj@cs.duke.edu

Graham Cormode
AT&T Labs–Research
pankaj@cs.duke.edu

Zengfeng Haung
HKUST
huangzf@cse.ust.hk

Jeff M. Phillips
University of Utah
jeffp@cs.utah.edu

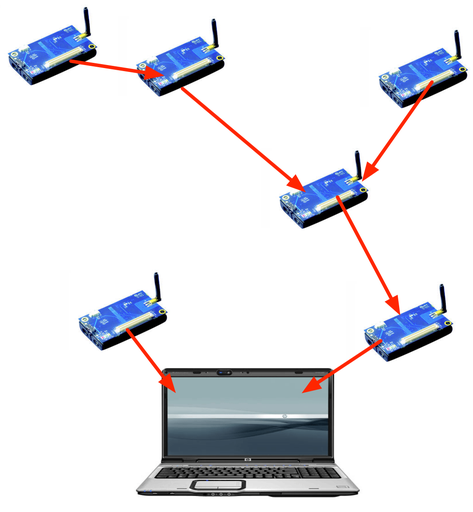
Zhewei Wei
HKUST
wzxac@cse.ust.hk

Ke Yi
HKUST
yike@cse.ust.hk

Motivation

Data summarization Compute a compact summary S of the data D that preserves its important properties, and to use the summary for answering queries.

Streaming Model S can be updated to reflect the new arrival without recourse to the underlying D.



Sensor networks

When a node receives the summaries from its children it merges these with its own summary, and forwards the result to its parent.

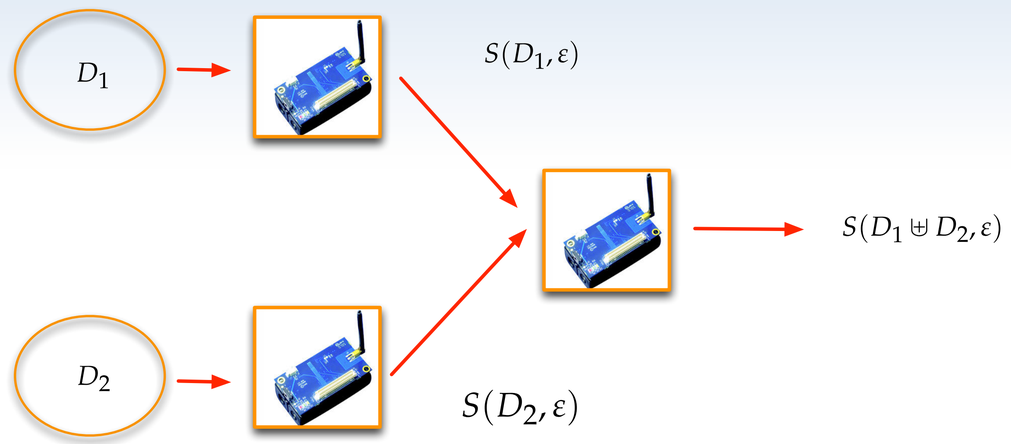
Distributed Computation

Each machine receives a share of the input and builds a summary of its share. We need to merge the summaries to form a summary on the whole input.

Question: Can we *merge* the ε-summaries of two (separate) data sets to obtain an ε-summary of the union of the two data sets, without increasing the size of the summary or its approximation error?

Problem Statement

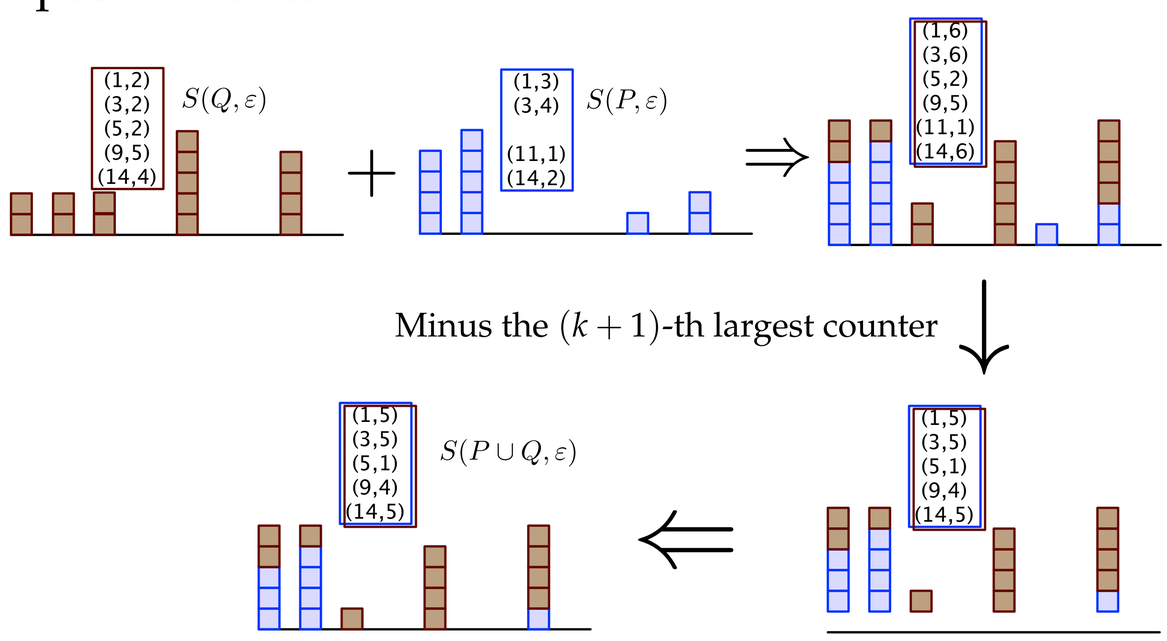
Let$S()$ denote a summarization method. Given D and an error parameter ε, We use $S(D,\varepsilon)$ to denote any valid summary for data set D with error ε produced by this method, and use $k(n,\varepsilon)$ to denote the maximum size of any $S(D,\varepsilon)$ for any D of n items.



We say that $S()$ is *mergeable* if there exists an algorithm $\mathcal{A}$ that produces a summary $S(D_1 \uplus D_2,\varepsilon)$ from any two input summaries $S(D_1,\varepsilon)$ and $S(D_2,\varepsilon)$. Note that, by definition, the size of the merged summary produced by $\mathcal{A}$ is at most $k(|D_1|+|D_2|,\varepsilon)$.

Heavy Hitters

Merging Algorithm for MG Sketch We first combine the two summaries by adding up the corresponding counters. This could result in up to $2k$ counters. We then perform a prune operation: Take the $(k+1)$-th largest counter, say C_{k+1}, and subtract it from all counters, and then remove all non-positive ones.



Theorem 1 *The MG summaries are mergeable with the above merging algorithm.*

Table 1: Best constructive summary size upper bounds under different models. The bounds in red are from this paper.

problem	offline	streaming	mergeable
heavy hitters	$1/\varepsilon$	$1/\varepsilon$ (MG82, SpacSaving06)	$1/\varepsilon$
quantiles (deterministic)	$1/\varepsilon$	$(1/\varepsilon)\log(\varepsilon n)$ (GK01)	$(1/\varepsilon)\log u$ (Q-digest04) $(1/\varepsilon)\log(\varepsilon n)$ (restricted merging)
quantiles (randomized)	$1/\varepsilon$	$1/\varepsilon \cdot \log^{3/2}(1/\varepsilon)$	
ε-approximations (rectangles)	$(1/\varepsilon)\log^{2d}(1/\varepsilon)$	$(1/\varepsilon)\log^{2d+1}(1/\varepsilon)$ (Suri et. al. 04)	$(1/\varepsilon)\log^{2d+3/2}(1/\varepsilon)$
ε-approximations (VC-dim ν)	$1/\varepsilon^{\frac{2\nu}{\nu+1}}$	$1/\varepsilon^{\frac{2\nu}{\nu+1}}\log^{\nu+1}(1/\varepsilon)$ (Suri et. al. 04)	$1/\varepsilon^{\frac{2\nu}{\nu+1}}\log^{3/2}(1/\varepsilon)$
ε-kernels	$1/\varepsilon^{\frac{d-1}{2}}$	$1/\varepsilon^{\frac{d-1}{2}}\log(1/\varepsilon)$ (Zarrabi-Zadeh08)	$1/\varepsilon^{\frac{d-1}{2}}$ (w/assumptions on data)