Mergeable Summaries

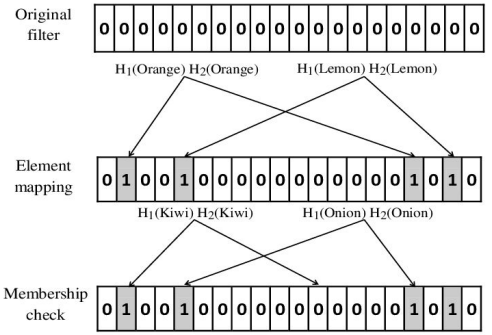
PODS 2022 Test-of-time presentation

Pankaj Agarwal (Duke), **Graham Cormode** (AT&T → Warwick/Meta),
Zengfeng Huang (HKUST → Fudan), Jeff M. Phillips (Utah),
Zheiwei Wei (HKUST → Renmin), Ke Yi (HKUST)

Overview

- What are mergeable summaries?
- What's in the PODS 2012 paper?
- What has happened since?

What are mergeable summaries?



A summary is a compact representation of a large volume of data so certain queries can be answered approximately

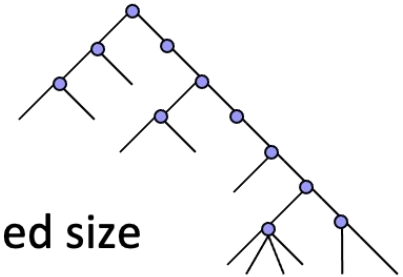
E.g., a collection of integers can be summarized by a sum and count to answer mean queries
A set of objects is summarized by a Bloom filter, which can answer set membership queries

From the abstract, “Informally speaking, **mergeability** requires that, given two summaries on two datasets, there is a way to merge the two summaries into a single summary on the two datasets combined together, while preserving the error and size guarantees.”

Generalizes streaming model, where a summary is merged with a single update at a time

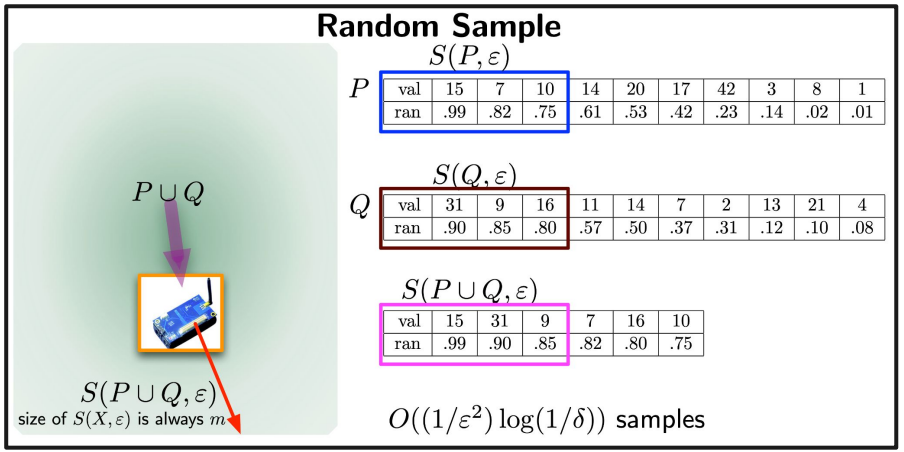
Mergeability

- Ideally, summaries are algebraic: associative, commutative
  - Allows arbitrary computation trees
    (see also synopsis diffusion [Nath+04], MUD model)
  - Distribution "just works", whatever the architecture



- Summaries should have bounded size
  - Ideally, independent of base data size
  - Or sublinear in base data (logarithmic, square root)
  - Should **not** depend linearly on number of merges
  - Rule out "trivial" solution of keeping union of input

Context



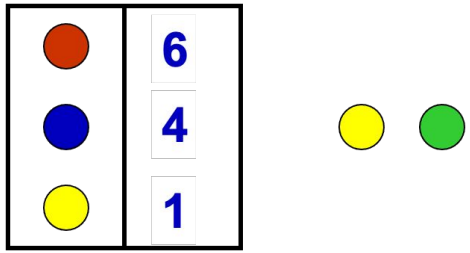
Did the paper “invent” mergeable summaries?

No, several examples were known already:

- Sketches based on random projections are mergeable by SUMming
- Bloom filters (based on hashing) and count distinct summaries are mergeable by OR-ing
- Sample are mergeable by careful subsampling of the union of samples

The paper highlighted the importance of mergability, showed mergeable results for new tasks

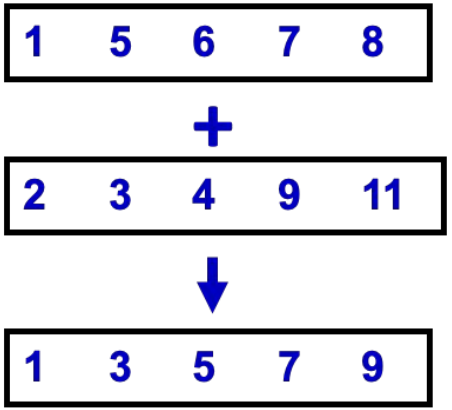
Results in the paper: heavy hitters



Two popular summaries for capturing item frequencies: Misra-Gries and SpaceSaving

- Both summaries keep a set of $1/\varepsilon$ items and counters and give an ε accuracy guarantee
- Both summaries are mergeable while retaining the exact same space-accuracy guarantees
  - Merge operation: pool the counters, combine overlaps, then prune back to fixed size
- In fact, they are isomorphic: they retain the same information (represented differently)
- This yields deterministic, fast, compact summaries for finding heavy hitters

Results in the paper: quantiles



Quantiles capture the order statistics of a one-dimensional distribution (i.e., median, quartiles etc.) with error ε

Prior work gave $O(1/\varepsilon \log^2(1/\varepsilon))$ sized (randomized) summaries

We showed a (randomized) $O(1/\varepsilon \log^{3/2}(1/\varepsilon))$-sized summary based on merging buffers

- Shaving off a $(\log^{1/2}(1/\varepsilon))$ factor
- More importantly, showed that this could be achieved with mergability

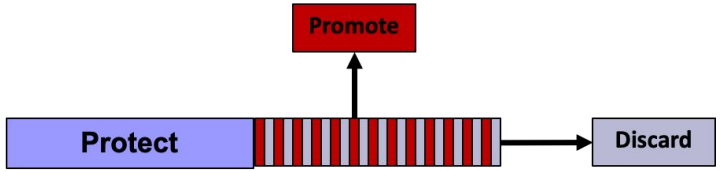
But $\log^{3/2}(1/\varepsilon)$ is not the last word on this problem!

Other results in the paper

- A deterministic bound for quantiles under some restrictions on the merging pattern
- Mergeable summaries for rectangle queries on point sets (ε-approximations)
- Mergeable summaries for range spaces of bounded VC dimension on point sets

Details in the full version with experimental evaluations (TODS 2013)

What happened next: quantiles



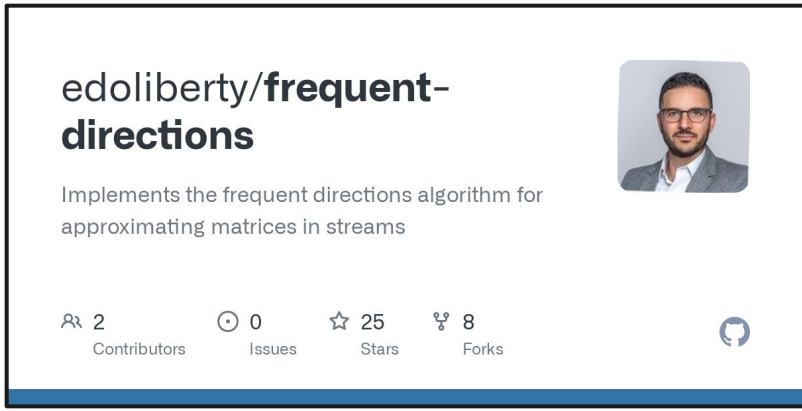
Several significant improvements for mergeable quantile summaries:

- Karnin, Lang, Liberty 16: $O(1/\varepsilon)$ randomized (KLL)
  Observed that buffers can vary in size, combined with sampling
- C. , Karnin, Liberty, Thaler, Vesely 21: $O(1/\varepsilon \log^{3/2} \varepsilon N)$ for relative error rank approximation
  Very close to the trivial lower bound of $\Omega(1/\varepsilon \log \varepsilon N)$!
  Best previous bounds were $O(1/\varepsilon \log (\varepsilon N) \cdot \min\{1/\varepsilon, \log^2 (\varepsilon N)\})$
  Still some small room for closing the gap!

Both build on the approach developed in the PODS'12 paper

What happened next: frequent directions

- A (mergeble) summary for a tall, skinny matrix that captures the “important directions”
- Due to Edo Liberty (KDD’13 best paper)
- “Inspired” by Misra-Gries summary and analysis
- Space-optimal, faster algorithm due to H.
  (an ICML’18 best paper awardee)

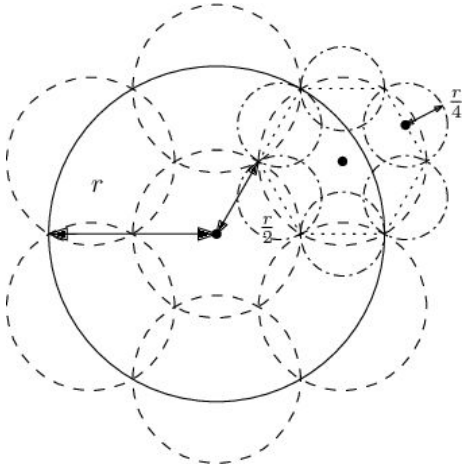


What happened next: composable coresets

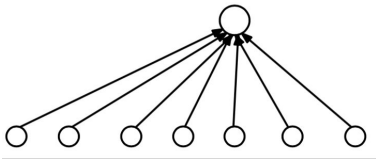
Composable coresets form a relaxation where error can accumulate (slowly) on merges

Composable coresets have been shown for:

- ε-kernels for convex hulls
- choosing diverse representatives
- vertex cover and matching
- submodular maximization



What happened next: k-way mergeability



Extension from pairwise merges to k-way mergeability: merging k summaries (H. and Y. 14/18)

- Each summary (on average) only needs to have size ~O(summary size/k^{α})
  Here α>0 depends on the discrepancy of the range space
  E. g. , for intervals (i.e., quantiles) and boxes α=1/2; for d-dim halfspaces α=1-1/(d+1)
- Almost matching lower bounds
  Builds on geometric discrepancy theory
  Shows a separation between deterministic and randomized summaries

What happened next: approximate counting

A basic problem: approximately counting a large quantity N with as few bits as possible

- Morris 1978: a compact randomized counter with O(log log N) bits
- Nelson and Yu 2022: a compact mergeable summary with improved error bounds

See the best paper talk later this session!

MORRIS'S COUNTER

```
1  Init():
2      c ← 0
3  Update(item):
4      increment c with probability 2^−c
5      ▷ and do nothing with probability 1 − 2^−c
6  Query():
7      return 2^c − 1
```

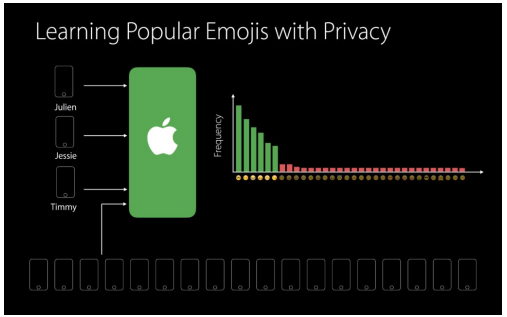
What happened next: privacy and security

Mergeable summaries have had many applications in privacy and security

- Privacy: smaller summary implies less noise needed to achieve (differential) privacy
- Security: linear summaries can be combined with (additive) homomorphic encryption

Some notable examples:

- Apple's implementation of private data collection via sketches
- Google's RAPPOR system based on Bloom filters
- Systems in federated learning and analytics using sketches



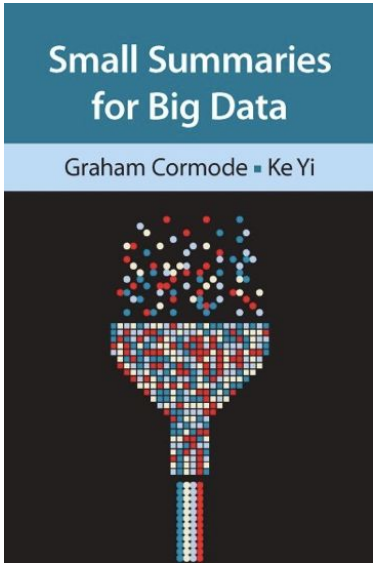
What happened next: a book of summaries!

With hindsight, many existing results are examples of mergeable summaries

This is the framing for C. and Y.'s book on "Small Summaries for Big Data"

An overview of summary techniques and discussions of their application

Many mergeable summaries are implemented in open-source libraries
E.g., Apache Data Sketches https://datasketches.apache.org/



A summary of mergeable summaries

The PODS 2012 paper laid out a manifesto for mergability, with novel results

Many paper and books subsequently adopted the notion of mergeability

Mergeable summaries are having increased influence in practice for a range of applications