Robust Set Reconciliation

SIGMOD 2014

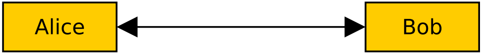
Christian Konrad

Reykjavik University

Joint work with: Di Chen, Ke Yi (both Hong Kong), Wei Yu (Aarhus), Qin Zhang (Indiana University)

Set Reconciliation

Data Synchronization Problem:



$S_A \subseteq \mathcal{U}$ $S_B \subseteq \mathcal{U}$

Goal: Alice and Bob learn $S_A \oplus S_B = (S_A \setminus S_B) \cup (S_B \setminus S_A)$

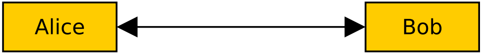
- Well-studied problem: $\mathrm{O}(|S_A \oplus S_B|)$ communication cost
- Many applications e.g. data consistency in distributed databases

Techniques:

- Ordered Data: Error Correcting Codes
- Unordered Data: Invertible Bloom Lookup Table

Set Reconciliation (2)

Example:



$S_A = \{2, 43, 119, 321, 599\}$ $\qquad$ $S_B = \{2, 44, 119, 222, 319\}$

Sets can be reconciled with communication cost $\mathrm{O}(|S_A \oplus S_B|)$

Sets are very similar:

- Two exact matches: $2, 119$
- Two almost matches: $43 \approx 44, 321 \approx 319$
- One true difference: $599 \neq 222$

Our Goal: Reconciliation that only considers the *true* differences with small communication cost

Synchronization of Image Databases

Alice

Bob

Difficulties:

- Same image, different encodings (bmp, jpeg, . . .)
- In general: rounding errors, introduction of noise

Communication Cost Constraint:

Given a communication budget, reconciliate as many true differences as possible

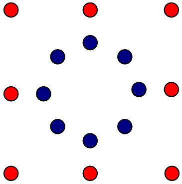
Robust Set Reconciliation

Input:

- Alice and Bob hold $S_A, S_B \subseteq [\Delta]^d$ on d-dim. grid of length Δ
- Communication budget k

Similarity measure: Earth-Mover-Distance

$\mathrm{EMD}(S_A, S_B) :=$ weight of minimum weight matching between S_A and S_B



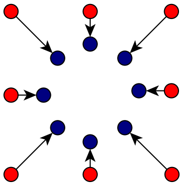
Robust Set Reconciliation

Input:

- Alice and Bob hold $S_A, S_B \subseteq [\Delta]^d$ on d-dim. grid of length Δ
- Communication budget k

Similarity measure: Earth-Mover-Distance

$\mathrm{EMD}(S_A, S_B) :=$ weight of minimum weight matching between S_A and S_B



$\mathrm{EMD}(S_A, S_B) =$ Sum of the lengths of the arrows

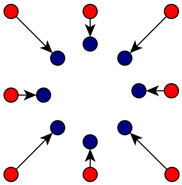
Robust Set Reconciliation

Input:

- Alice and Bob hold $S_A, S_B \subseteq [\Delta]^d$ on d-dim. grid of length Δ
- Communication budget k

Similarity measure: Earth-Mover-Distance

$\mathrm{EMD}(S_A, S_B) :=$ weight of minimum weight matching between S_A and S_B



$\mathrm{EMD}(S_A, S_B) =$ Sum of the lengths of the arrows

Robust Set Reconciliation: Alice sends message M to Bob with $|M| = \tilde{\mathrm{O}}(k)$. Then Bob finds a set S'_B so that $\mathrm{EMD}(S_A, S'_B)$ is minimized

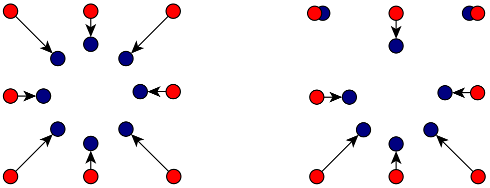
Optimal Solution

Communication budget limited by $\tilde{O}(k)$:
We cannot expect to reconciliate more than k point-pairs

k-**residual EMD:**

$$\mathrm{EMD}_k(S_A, S_B) := \min_{S_B^k} \mathrm{EMD}(S_A, S_B^k),$$

where S_B^k is obtained from S_B by relocating at most k points:



$\mathrm{EMD}(S_A, S_B)$ $\qquad$ $\mathrm{EMD}_2(S_A, S_B)$

"Remove the k heaviest edges"

Our Goal: Approximation Scheme. Bob finds S_B' so that

$$\mathrm{EMD}(S_A, S_B') \leq C \cdot \mathrm{EMD}_k(S_A, S_B)$$

Results

Upper Bound: We have designed a one-way protocol with

- Communication Cost $\mathrm{O}(kd \log(n\Delta^d) \log \Delta)$ so that
- Bob computes S'_B and

$$\mathrm{EMD}(S_A, S'_B) \leq \mathrm{O}(d) \cdot \mathrm{EMD}_k(S_A, S_B).$$

- The runtimes of both Alice and Bob is $\mathrm{O}(dn \log \Delta)$.

Lower Bound: Any possibly randomized one-way communication protocol that computes an $\mathrm{O}(1)$ approximation has communication cost

$$\mathrm{O}(k \log(\Delta^d / k) \log \Delta).$$

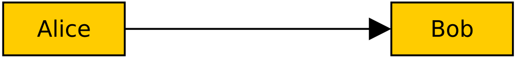
$\rightarrow$ For typical settings $d = \mathrm{O}(1), n = \Delta^{\mathrm{O}(1)}, k = \mathrm{O}(\Delta^{d-\epsilon})$ UB is tight

Experiments:

- Comparision to a baseline method that uses lossy compression
- Image reconciliation

Key Technique 1: Classical (One-way) Reconciliation

Ordered Data:



$u \in \mathcal{U}^n$ $\quad$ $v \in \mathcal{U}^n$

There is a one-way protocol so that:

- Communication Cost is $\tilde{\mathrm{O}}(k)$,
- If $d_H(u, v) \leq k$ then Bob can learn Alice's input,
- If $d_H(u, v) > k$ then Bob can report that $d_H(u, v) > k$.

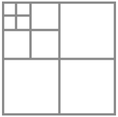
(d_H : Hamming distance)

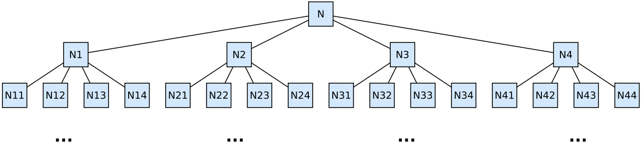
Technique:

- Forward Error Correction such as a Reed-Solomon code
- Invertible Bloom Lookup Table (near linear time for decoding/decoding)

Key Technique 2: Quad-trees

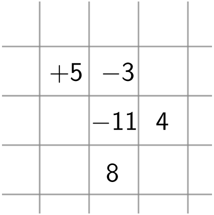
Quad-trees:





- A layer corresponds to a resolution of the point set
- Alice and Bob construct quad-trees T_A, T_B for their inputs S_A, S_B
- A layer of the difference tree $(T_A - T_B)$ indicates "surplus" and "deficit cells"

Correction given layer L of Alice's tree:
Subtract this layer from own layer L and do corrections as follows: Move points from surplus cells to center of deficit cells
Note: Additional error introduced since exact position is unknown

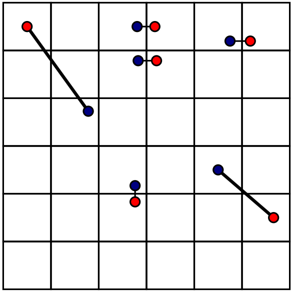


Key Technique 3: Random Shift

Let $M = (m_i)_i$ be a min-cost perfect matching between S_A and S_B

Interesting Layer: Consider layer in difference tree $(T_A - T_B)$ that reflects the k heaviest edges of M (Hamming distance $= \Theta(k)$)

Technical Difficulty: False Positives



$\rightarrow$ Perform a random shift of the grid

Summary: Algorithm

Alice:

1. *Random Shift:* Alice shifts all points by u.a.r. chosen γ
2. *Build Quad-tree*
3. *Invertible Bloom Lookup Table:* For every layer L of the quad-tree, build an IBLT that allows Bob to recover Alice's layer L if Bob's layers L differs by at most ck (for a constant c)
4. *Send Message:* Alice sends γ and the IBLT's to Bob

Bob:

1. *Random Shift*
2. *Build Quad-tree*
3. *Decode IBLTs:* Bob decodes the IBLTs and determines the highest layer L' so that Hamming distance is at most ck
4. *Move points:* Move points from surplus cells to deficit cells (center)
5. *Reverse Random Shift*

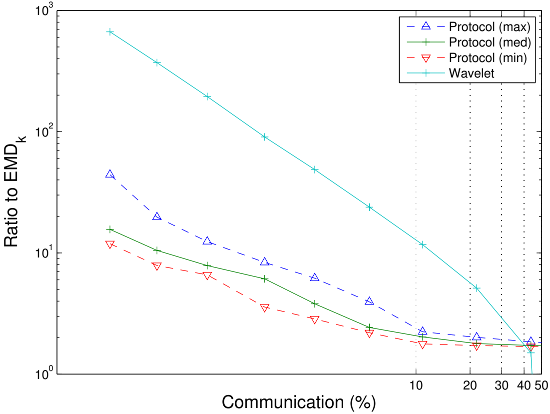
Redundancy factor c: Account for moving points to center of cells

Summary: Algorithm

- One-way two-party communication protocol for $\mathrm{O}(d)$-approximation
- Algorithm cannot compute EMD nor residual EMD
- Computing EMD in one-way two-party communication model is a hard problem: constant approximation has communication cost polynomial in Δ

One dimensional Experiment

- Alice's point set: 1D data set with $n = 10^6$ points
- Inject $k = 100$ true differences by randomly picking k points and moving them to an arbitrary location
- For all other nodes inject noise in $[-1, 1]$
- Baseline Method based on lossy Haar Wavelet Compression



Reconciliation of Image Database

Data Set:

- Alice has 10.000 high quality JPEG images
- Bob has a copy of this set which is modified as follows:
  - All images are recompressed with 95%-quality JPEG compression
  - k images are replaced by different ones

Adaption of the Algorithm:

- Images are mapped to 6-dimensional feature space
- Algorithm adapted to two-way communication

		Budget				
		2%	4%	6%	8%	10%
	5	0%	56%	**92%**	**100%**	**100%**
	10	2%	34%	84%	**100%**	**100%**
k	15	0%	28%	80%	**100%**	**100%**
	20	0%	19%	67%	**98%**	**99%**
	25	0%	5%	66%	87%	**99%**

Table: Recovery rate for image reconciliation

Conclusion

Summary:

- Robust set reconciliation method that works well in practice
- Lower Bound illustrating that communication budget is almost tight

Open Questions:

- Can $O(d)$-approximation be improved? (e.g. $(1+\epsilon)$-approx.)
- Improvement via multiple communication rounds?

Conclusion

Summary:

- Robust set reconciliation method that works well in practice
- Lower Bound illustrating that communication budget is almost tight

Open Questions:

- Can $O(d)$-approximation be improved? (e.g. $(1+\epsilon)$-approx.)
- Improvement via multiple communication rounds?

Thank you for your attention.