Efficient Algorithms for Querying Noisy Disturbed/Streaming Datasets
Classical computation model does not fit “big data”

The random-access memory (RAM) model

- A processor and an **unbounded size memory**
- **Centralized**, cost is in terms of # memory cells read/written
Classical computation model does not fit “big data”

The random-access memory (RAM) model
- A processor and an **unbounded size memory**
- **Centralized**, cost is in terms of \# memory cells read/written

Big data doesn’t fit!
- High-speed online data: incapable of storing everything
- Data is distributed in different machines
Streaming Model

E.g., what is the number of distinct elements?

- small RAM space
- small update time
- small query time
Big data computation model: streaming model

Streaming Model

- small RAM space
- small update time
- small query time

E.g., what is the number of distinct elements?

Captures many settings, e.g.,

Network routers

Ad auction

And flight logs, telephone switches, etc.
Big data computation model: \( k \)-site model

Communication \( \rightarrow \) time, energy, bandwidth, \ldots

Sensor networks  The MapReduce model.  The BSP model.  Cloud computation

Abstraction

\( C \)

\( k \)-site model

\( S_1 \)  \( S_2 \)  \( S_3 \)  \ldots  \( S_k \)

- small comm. cost
- small \# comm. rounds
- small local computation time
Real-world datasets are often noisy.
Imprecise references to the same real-world entities are ubiquitous in scientific and commercial databases.

music, images, ... after compressions, resize, reformat, etc.
Real-world datasets are often noisy.
Imprecise references to the same real-world entities are ubiquitous in scientific and commercial databases.

music, images, ...
after compressions, resize, reformat, etc.

“Indiana University Bloomington”
“Hoosier University”

Queries of the same meaning sent to Google
Cannot do a comprehensive data deduplication in the streaming/k-site model using small space/communication!

With my PhD students, we are working on the design of distributed and streaming algorithms that run directly on the noisy datasets, resolve the noise “on the fly”, and retain communication and space efficiency.

project funded by NSF
THANK YOU

CONTACT INFORMATION
Email: qzhangcs@Indiana.edu
http://homes.soic.indiana.edu/qzhangcs/