Trustworthy Data Collection for Cyber Systems

Md Zakirul Alam Bhuiyan
Assistant Professor
Department of Computer and Information Sciences
Fordham University, New York, NY
https://sites.google.com/site/zakirulalam/
http://storm.cis.fordham.edu/~bhuiyan/
mbhuiyan3@fordham.edu, zakirulalam@gmail.com

Outline

- The Need for Trustworthy Data: Realization
- Challenges
- Potential Strategies
  - Trustworthy Data Collection
  - Protected Data Collection
  - Privacy-Preserving Data Mining
  - Guaranteeing Data Quality in Data Reduction

The Need for Trustworthy Data: Realization

○ We often talk about

- Security/reliability in communication, processing, storage
- Security and privacy for data and network outsourcing
- Security and privacy in crowdsourcing
- Security and privacy for mobile and wearable devices
- Security and privacy in cellular networks
- Security and privacy in cloud and edge computing
- Security and privacy in emerging wireless technologies
- Security and privacy in peer-to-peer and overlay networks
- Security and privacy in smart and connected health
- Security and privacy in smart cities, IoT, and RFID systems
- Security for critical infrastructures (smart grids, transportation, etc.)
- Security for software-defined and data center networks
- Security for routing and network management
- And so on..

The Need for Trustworthy Data: Realization

- We often talk about
  - How to achieve security and/or privacy in a cyber system?
    - There are huge works around everyday

- In most work, we consider data security and privacy
  - **During** the data processing, storing, and transmitting
  - **After** the data processing, storing, and transmitting

The Need for Trustworthy Data: Realization

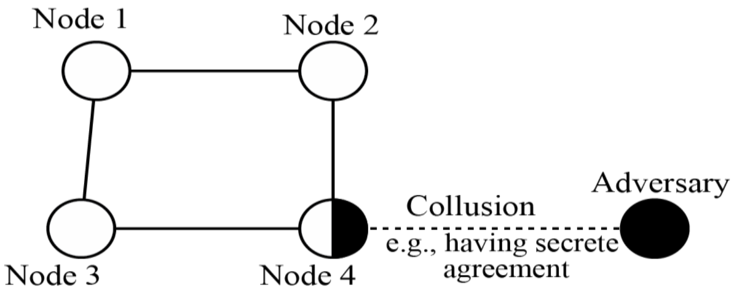
- **Questions:**
  - **How about if we don’t trust the data that we are about to process, store, and transmit?**
  - **How about if the data is already compromised or altered before being processed, stored, and transmitted?**
    - Decisions made in a cyber system based on the collected data may be meaningless, untrustworthy, i.e.,
      - We may process the compromised data
      - We may store the compromised data
      - We may encrypted the compromised data
      - We may transmit the compromised data

Outline

- The Need for Trustworthy Data: Realization
- Challenges
- Potential Strategies
  - Trustworthy Data Collection
  - Protected Data Collection
  - Privacy-Preserving Data Mining
  - Guaranteeing Data Quality in Data Reduction

Challenges (1)

- Integrity problem
  - **Security attacks**
    - Collusion attack, malicious attack
    - False data injection
    - Some sensors constantly provide
      - Truthful data while
      - Others may generate biased, compromised, or even fake data



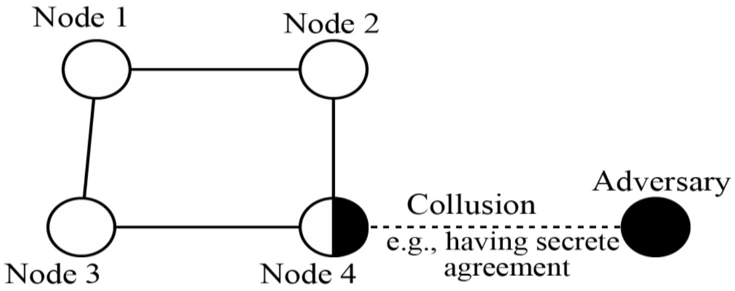
  - **Fault occurrences**
    - Data faults
    - System faults

Challenges (1)

- **Integrity problem**
  - **Untrustworthy data may have**
    - Illegal values
    - Violated attribute dependencies
    - Uniqueness violation
    - Referential integrity violation
    - Missing values
    - Misspellings
    - Cryptic values
    - Embedded values
    - Misfielded values
    - Word transpositions
    - Duplicate records
    - Contradicting records
    - Wrong references
    - Overlapping data/matching records
    - Name conflicts
    - Structural conflicts

Challenges (2)

- In which stage the data is altered and become untrustworthy?
  - At the acquisition
  - After the acquisition
  - At the transmission
  - During transmission
  - After transmission, and
  - Before aggregation



Challenges (3)

- How to identify the compromised data once the data reaches high-end storage, such as Cloud?

Challenges (4)

- How to ensure the trustworthiness of a cyber system in which data with integrity problem is already processed and ready for a decision-making

Low quality of data

Low quality of monitoring

Low quality of decision-making

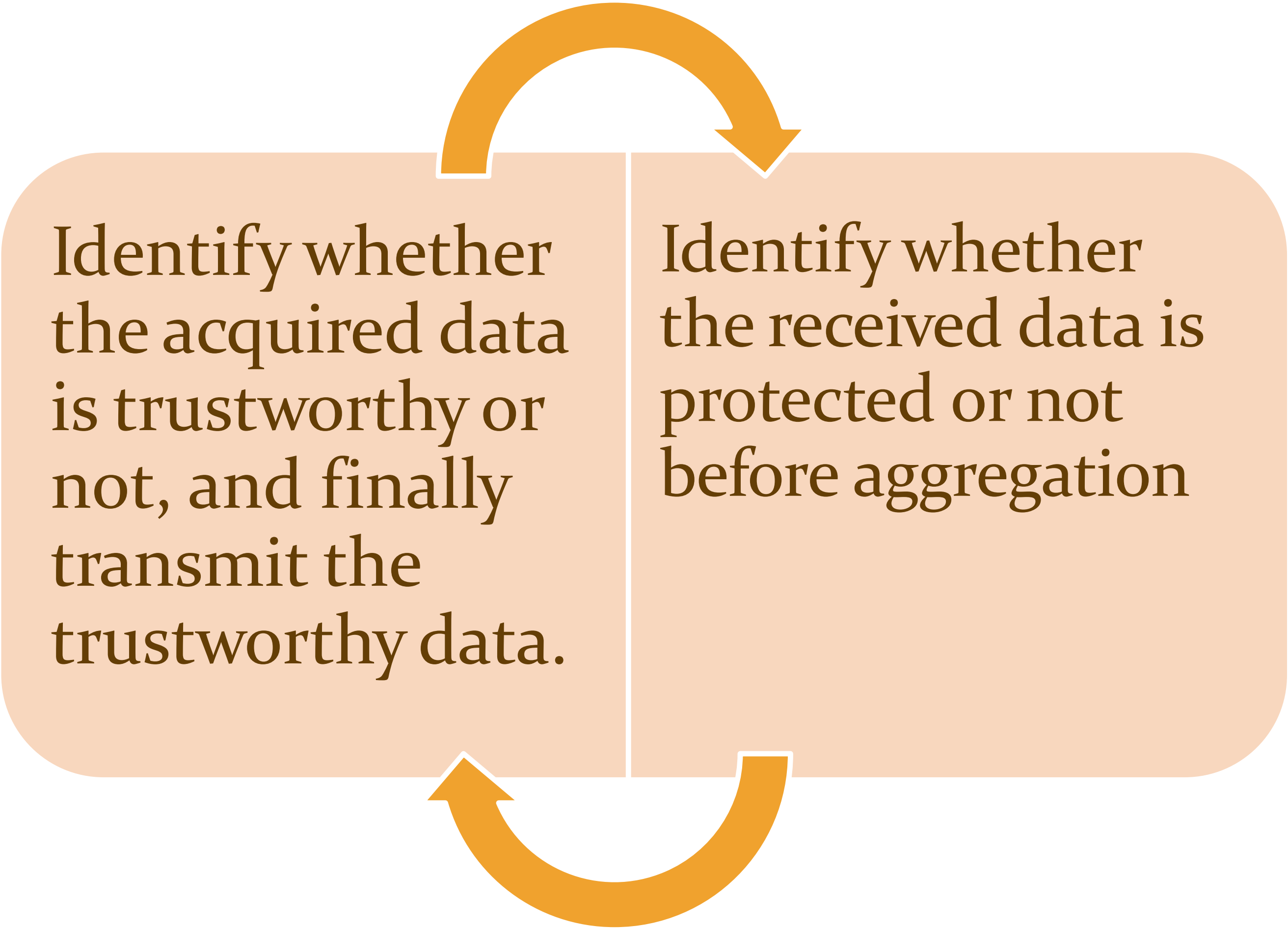
Real-time event undetecttion

Outline

- The Need for Trustworthy Data: Realization
- Challenges
- Potential Strategies
  - Trustworthy Data Collection
  - Protected Data Collection
  - Privacy-Preserving Data Mining
  - Guaranteeing Data Quality in Data Reduction

Strategies (1)

- Trustworthy data collection

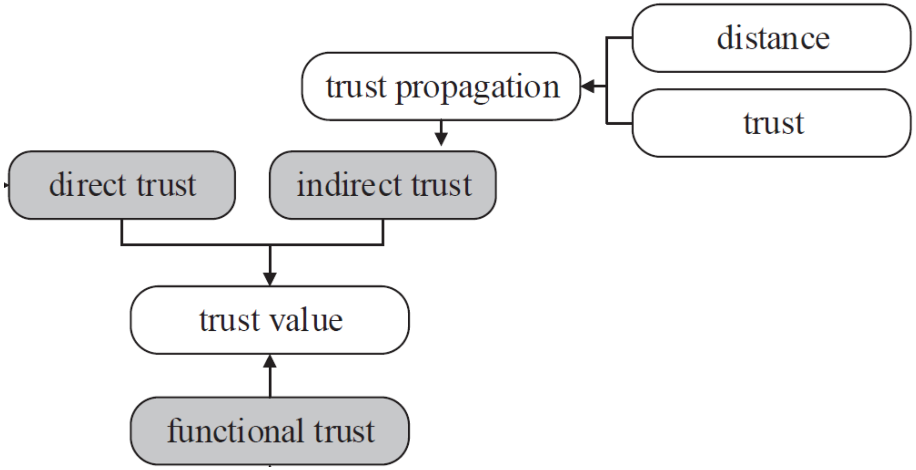


Identify whether the acquired data is trustworthy or not, and finally transmit the trustworthy data.

Identify whether the received data is protected or not before aggregation

Strategies (1)

- After the data acquisition or the transmission
  - TrustData evaluation
  - Truth discovery

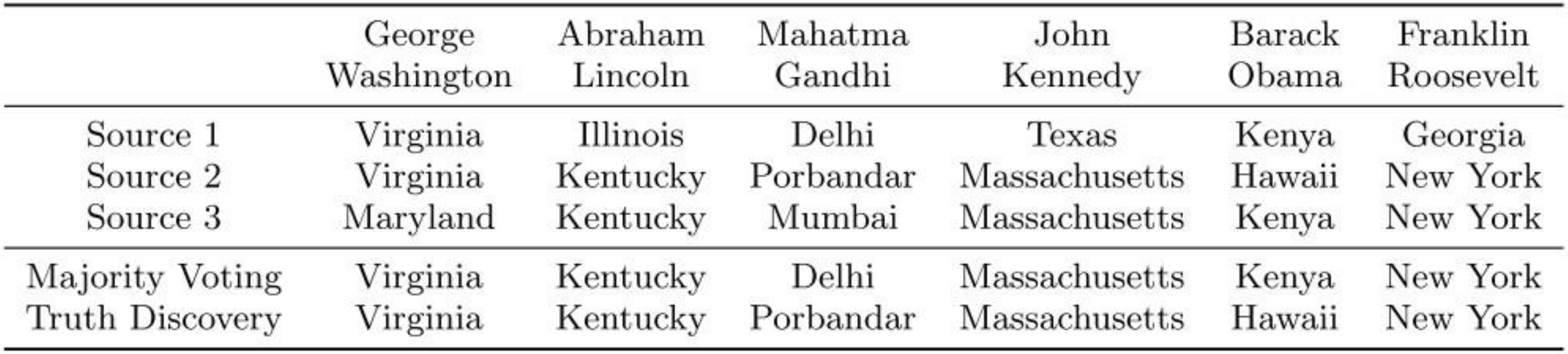


Strategies (1)

- Truth discovery
  - It is used in many domains in order to resolve conflicts with multiple noisy data or sources (sensors)
    - The medias provide billions of pieces of information, unfortunately, not all are reliable, relevant accurate, unbiased, or up-to-date
    - Before being used, the information are evaluated for truth.

Strategies (1)

- Truth discovery
  - Example: the birth place



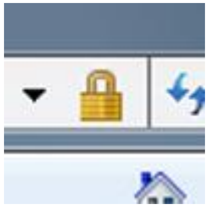
	George Washington	Abraham Lincoln	Mahatma Gandhi	John Kennedy	Barack Obama	Franklin Roosevelt
Source 1	Virginia	Illinois	Delhi	Texas	Kenya	Georgia
Source 2	Virginia	Kentucky	Porbandar	Massachusetts	Hawaii	New York
Source 3	Maryland	Kentucky	Mumbai	Massachusetts	Kenya	New York
Majority Voting	Virginia	Kentucky	Delhi	Massachusetts	Kenya	New York
Truth Discovery	Virginia	Kentucky	Porbandar	Massachusetts	Hawaii	New York

Strategies (2)

- Protected data collection
  - Data privacy @ the data acquisition

Privacy: what data goes where?

>> **What data collected** and goes where?



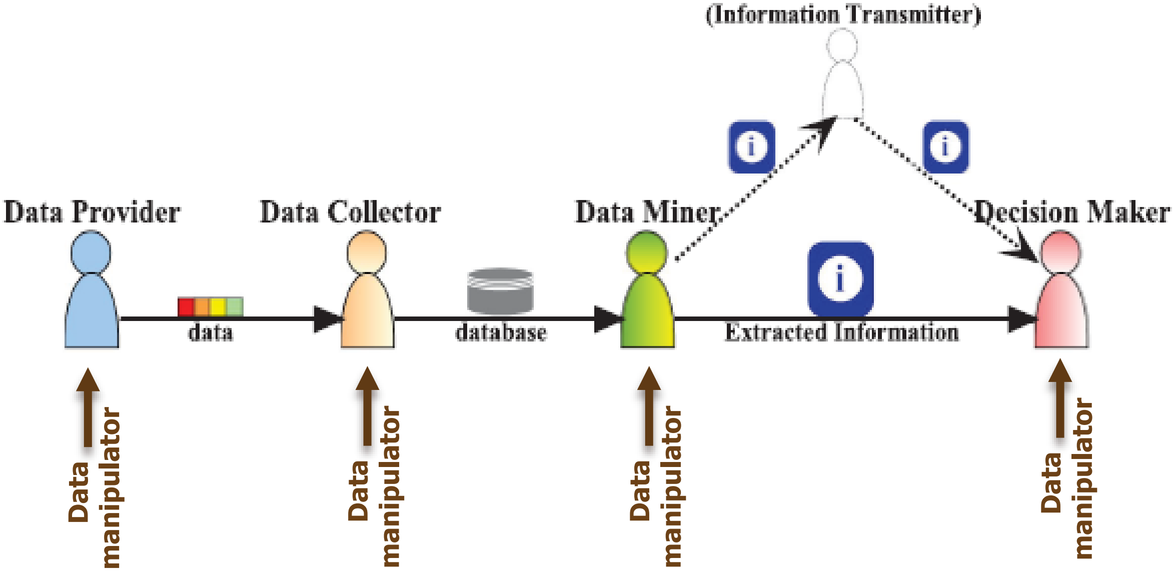
Security: protection against unauthorized access to data

>> protection against unauthorized access to already **privacy-breaching acquired** data??

How to provide protection to the privacy of data at the data acquisition?

Strategies (3)

- Privacy-Preserving Data Mining (PPDM)
  - The 4 types of users in data mining process



Data manipulator

Data manipulator

Data manipulator

Data manipulator

Strategies (4)

- Guaranteeing data quality in data reduction at the data acquisition
  - Energy consumption reduction
  - Wireless bandwidth reduction
  - Real-time decision making
  - Cost reduction

Strategies (4)

- Guaranteeing trustworthy decision-making from data reduced at the acquisition
  - At a low rate or high rate
    - 20Hz, 560Hz, 1024Hz
  - With narrow frequency
    - Single
  - Even-sensitive
    - Threshold (drop if low threshold)
  - Frequency content
    - High or low frequency content data

- Energy consumption reduction
- Wireless bandwidth reduction
- Real-time decision making
- Cost reduction

- Does the acquired data can lead to a trustworthy decision?

Conclusions

- May not be a good idea
  - To invest cost and time for processing, storing, and transmitting of unsecured and untrustworthy data
  - To encrypt untrustworthy data

We need trustworthy data for trustworthy cyber systems