CERIAS

the center for education and research in information assurance and security

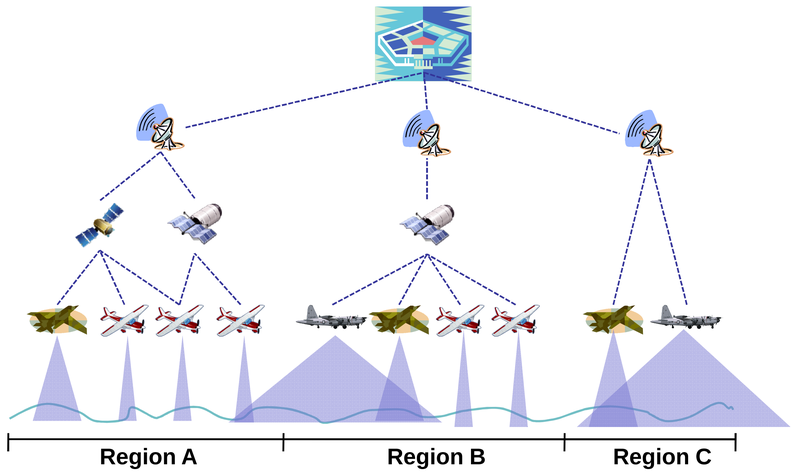
Provenance-Based Confidence Policy Management in Data Streams

Hyo-Sang Lim, Yang-Sae Moon, and Elisa Bertino
CERIAS, Purdue University
hslim@cs.purdue.edu, ysmoon@cs.purdue.edu, bertino@cs.purdue.edu

Problem Descriptions

- Confidence of a data item is highly depend on not only its values but also its provenance (e.g., trustworthiness of source/intermediate nodes).
- Especially in data stream environments,
  1) data elements arrive incrementally and
  2) trustworthiness of nodes can be dynamically changed as time goes on.
- To provide accurate confidence information for continuous query processing in DSMS, we need to supports incremental assignment confidence scores for nodes and data items

An Example : Battlefield Monitoring Sensor Network



Provenance-Based Confidence Policy

Confidence Policy: a novel notion that supports confidence scores in data management and query processing.

A confidence policy restricts access to the query results by specifying the minimum confidence level of a certain task.

a continuous query Q_i + **a confidence policy** $< r, pu, \beta >$ (β: desired confidence level; pu: purpose; r: role)

Data Stream Provenance

- **Network node provenance** : where the data item generated and passed
- **Operation provenance** : which operation are conducted for the data item

Challenges

1) How to calculate confidence scores for network nodes and data items?
2) How to use the scores in continuous query processing?

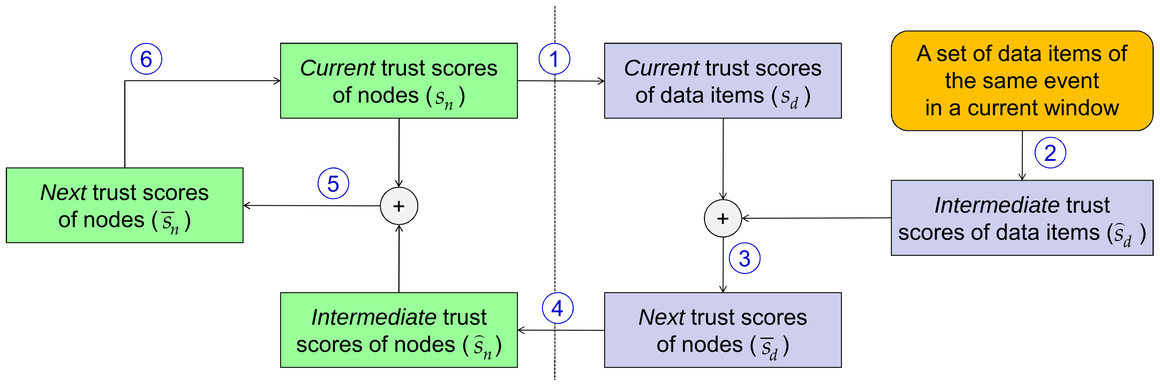
Our Approaches

Identify data items belonging to each single event

- Using an observation that the network configuration for sending and receiving data inherently implies semantics of an application
- Grouping network nodes so as to represent individual events and assigning each data item to a specific event by examining its provenance

Evolving scores : trust scores are gradually evolved in a cyclic framework

- Trust scores of data items from those of network nodes
- Trust scores of network nodes from those of data items



Principles to Calculate Scores

We use two similarity properties: ***data similarity*** inferred from data values and ***path similarity*** inferred from data provenance

- Data similarity comes from a simple intuition that the more data items having similar values, the higher trust scores.
- Path similarity comes from an observation that different paths, but similar data values may increase trustworthy of data items

	Similar Value	Different Value
Similar Path	score ↑	score ↓↓↓ (conflict)
Different Path (but same event)	score ↑↑↑ (cross check)	score ↓

Future Work

- Developing sophisticate methods for dynamic score calculation
- Developing query and policy evaluation
- Systematic quality adjustment
- Combining event granularity issue with the efficient delivery issue

PURDUE UNIVERSITY

CERIAS

Discovery Park e-Enterprise Center