Open Questions & Research Areas in Privacy-Preserving Data Mining

Yehuda Lindell

IBM T.J.Watson

Extended Models

- Adversarial power:
  - Until now, almost only semi-honest, static
  - Extensions:
    - Malicious
    - Honest looking
    - Others…
- Trust:
  - Better to assume nothing.
  - Practically, optimistic models or others are worth consideration

Extended Models

- Composition:
  - Until now, only stand-alone
  - Extensions: really need security under concurrent general composition
    - Implies a necessity for a common reference string, or to look for weaker definitions
  - Also relates to “non-cryptographic” methods (perturbation methods may be carried out independently in similar databases…)

Application versus Theory

Finally, someone actually wants to **use** secure multiparty computation

- I.e., someone wants to use it, rather than us wanting them to use it

Can we provide real solutions to real users?

- Danger of “expert systems”

Applying Secure Computation

- Necessary conditions:
  - We need to find out what **models** are truly realistic/acceptable, & in what settings
    - Does the semi-honest model really suffice for government agencies and privacy law
  - We need to understand what **problems** are really of interest (ID3 versus C4.5)
    - Involves also learning more about data mining
  - We need to understand how we can **fit** into the data mining process (e.g., tweaking, cross validation).

Applying Secure Computation

- Can we build a prototype for a realistic scenario, and see how it works?
  - In data mining, implementation is essential for determining usability
  - Here too, many real problems may only be revealed upon implementation
    - Can imagine an intermediary step whereby the data mining computation will actually not take place securely (actually pool data). Users will not see this.
    - Drawback: can only test where it is not really needed.