

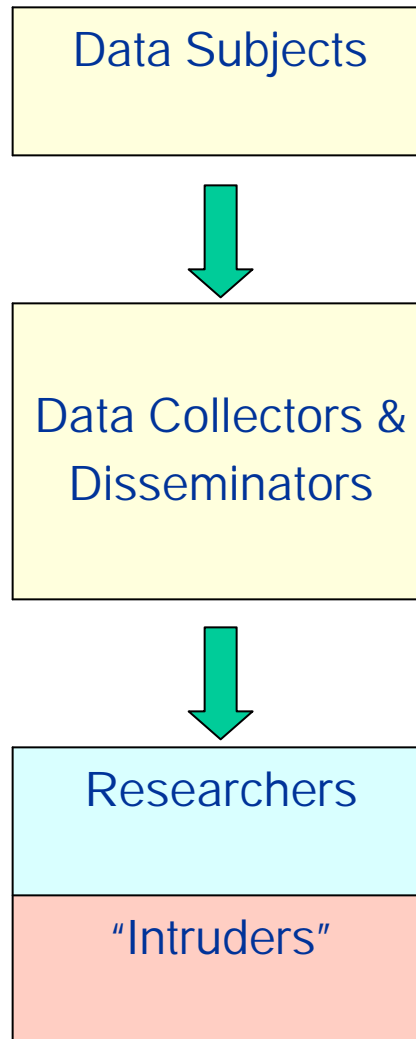
Pointers from Research on Data Confidentiality and Data Quality

Ashish Sanil

National Institute of Statistical Sciences

[based on work done with Adrian Dobra, Steve Fienberg, Shanti Gomatam, Alan Karr and Jaeyong Lee]

Data Confidentiality Problem (Dissemination)



Data Confidentiality Problem (Dissemination)

Problem

Solution approach

Data Subjects



Data Collectors & Disseminators



Researchers
"Intruders"

Confidentiality of subjects
Analytical usefulness of the data



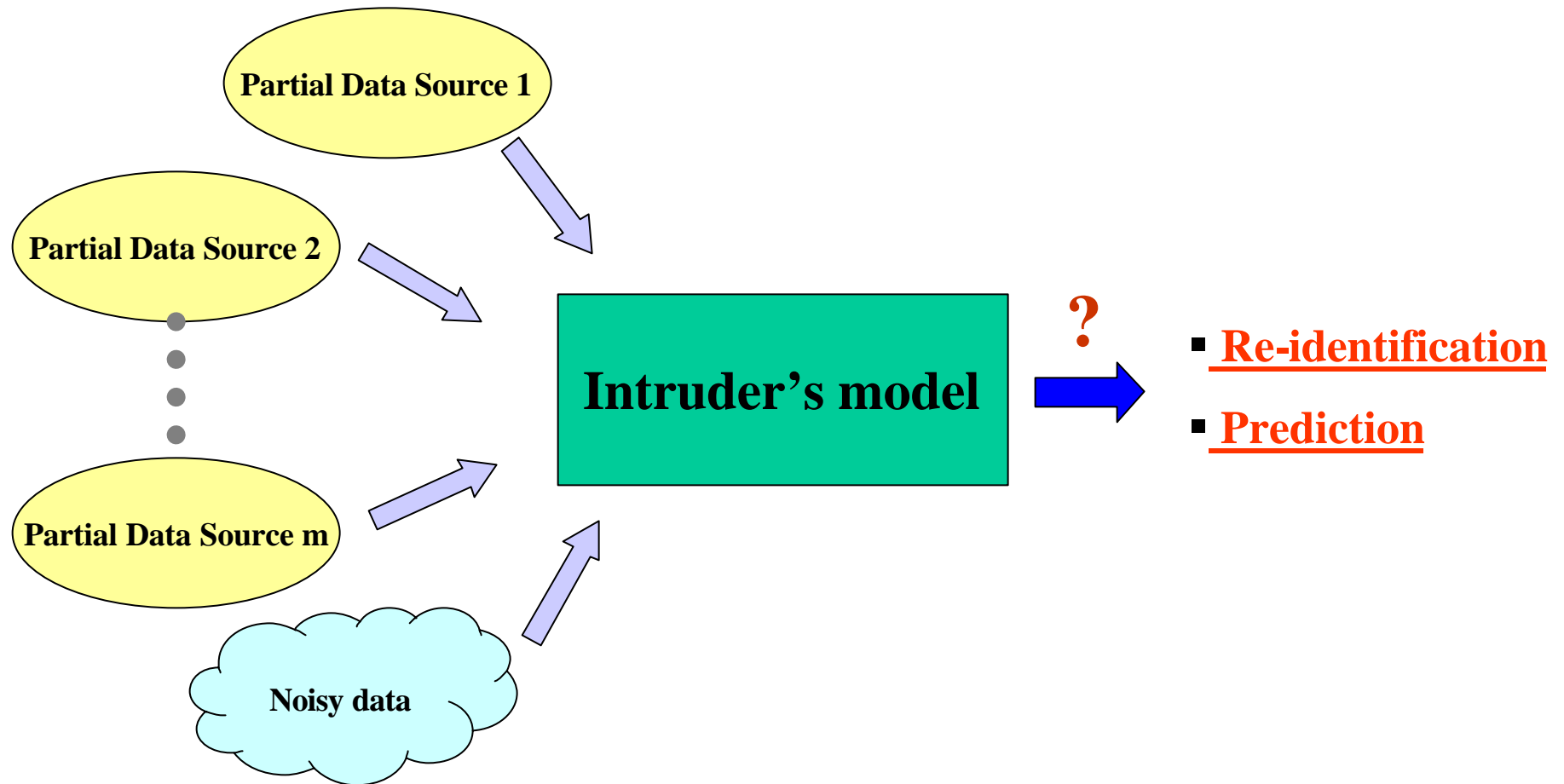
Consider intruder Strategies (risk)
Consider researchers' analysis methods (utility)

Learn population characteristics
Uncover information on individuals



Statistical analysis
Intruder models

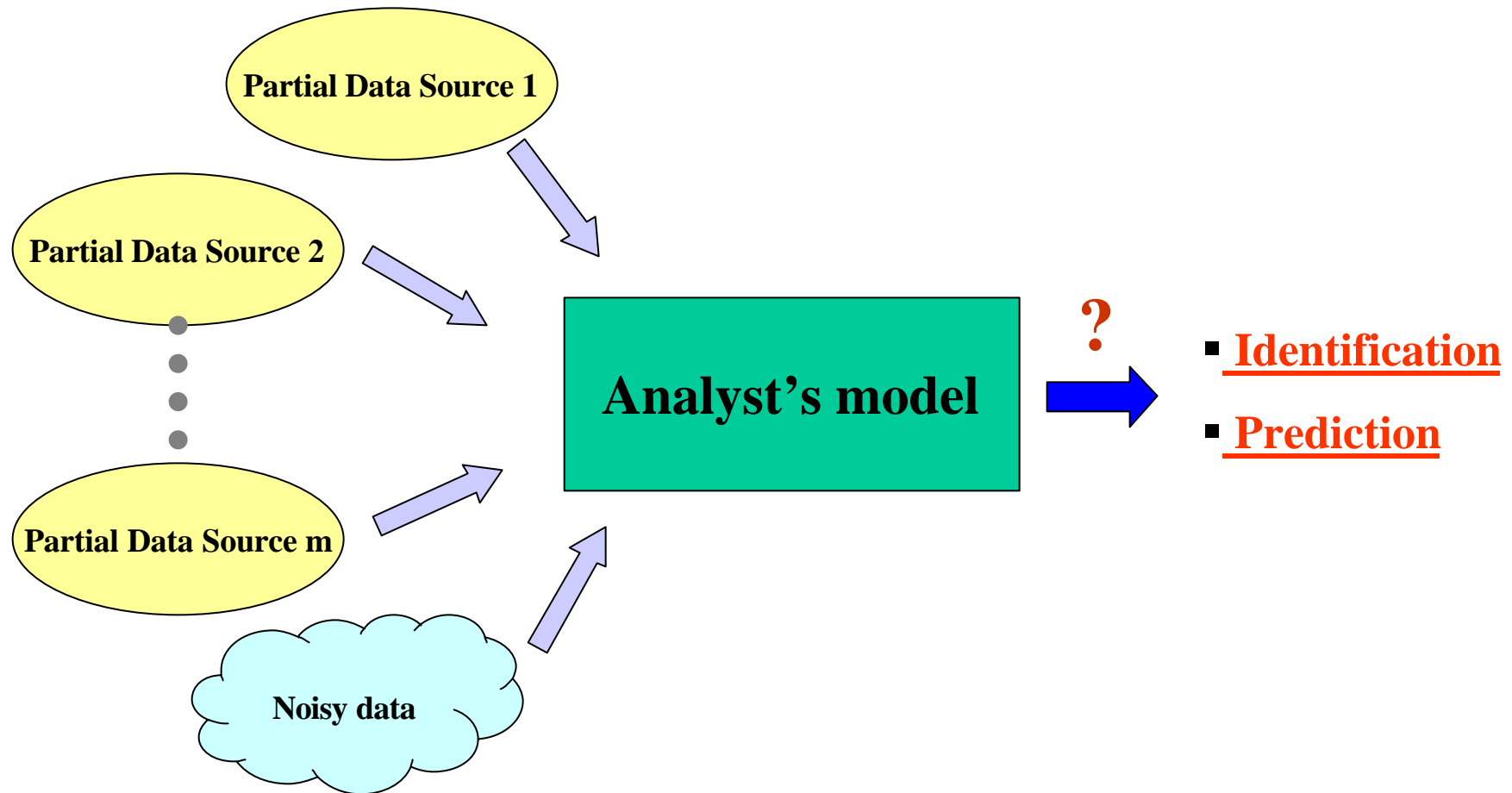
Modeling “Intruder” behavior



Security and Privacy/Confidentiality

- Use of databases of confidential, high-quality, high-resolution data on individuals
 - Legal and ethical issues
 - Privacy-preserving access and data-mining
- Extracting useful information from readily available, possibly low-quality and incomplete data

Data Integration Problem



Data Quality Issues

- **DQ Problem:** Evaluate data quality (consistency, accuracy, etc.) and try to improve it
- **DC \leftrightarrow DQ Link:** Like the intruder, we need to have a model/procedure to ascertain how well we can do with the imperfect data

Modeling Tools

	Noisy Data	Partial Data
Deterministic	<ul style="list-style-type: none">• Rules-based validity checks	<ul style="list-style-type: none">• Techniques for finding upper and lower bounds
Inference	<ul style="list-style-type: none">• Record linkage• Robust methods• Outlier detection	<ul style="list-style-type: none">• Dis-aggregation methods• Reconstruction techniques

Modeling Tools

	Noisy Data	Partial Data
Determin.	<ul style="list-style-type: none"> • Rules-based validity checks 	<ul style="list-style-type: none"> • Finding upper and lower bounds
Inference	<ul style="list-style-type: none"> • Record linkage • Robust methods • Outlier detection 	<ul style="list-style-type: none"> • Dis-aggregation methods • Reconstruction techniques

- Verifying data types and ranges as defined in metadata
 - Check consistency (e.g., temporal constraints)
 - Parse and standardize (e.g., addresses)
- Can detect anomalies/data distortion measures
- Part of Extract-Transform-Load (ETL) process in data warehousing systems
- Often first step in record linkage

Modeling Tools

	Noisy Data	Partial Data
Determin.	<ul style="list-style-type: none">• Rules-based validity checks	<ul style="list-style-type: none">• Techniques for finding upper and lower bounds
Inference	<ul style="list-style-type: none">• Record linkage• Robust methods• Outlier detection	<ul style="list-style-type: none">• Dis-aggregation methods• Reconstruction techniques

- Most probabilistic record linkage based on Fellegi-Sunter model
 - Match records in data files A and B
 - Consider all pairs in $A \times B$
 - Estimate probability of observing certain patterns (say, partial substring match) given true match, and given non-match
 - Decision rules for using the probabilities to declare record pairs as match, non-match or undecided
- Implementation and scalability challenges
- Large CS/Statistics literature

Modeling Tools

	Noisy Data	Partial Data
Determin.	<ul style="list-style-type: none">• Rules-based validity checks	<ul style="list-style-type: none">• Techniques for finding upper and lower bounds
Inference	<ul style="list-style-type: none">• Record linkage• Outlier detection• Robust methods	<ul style="list-style-type: none">• Dis-aggre• Reconstruction techniques

- Outlier detection methods: Statistical analogs of validity checks
- Need to determine if sensitive relationships in the data can be learned by using robust statistical techniques

Modeling Tools

	Noisy Data	Partial Data
Determin.	<ul style="list-style-type: none">• Rules-based validity checks	<ul style="list-style-type: none">• Techniques for finding upper and lower bounds
Inference	<ul style="list-style-type: none">• Record linkage• Robust methods• Outlier detection	<ul style="list-style-type: none">• Dis-aggregation methods• Reconstruction techniques

- Reconstruction techniques such as Iterative Proportional Fitting (prediction from a log-linear model)
- Missing value imputation methods
- Dis-aggregation strategies, e.g., simulating possible populations that satisfy the aggregation constraints

Example Scenario

Tracking shipments

- Vessels originating from two ports: O1, O2
- Two ports of destination: D1, D2
- Carrying three kinds of cargo: X, Y, Z
- Partial information available in the form of cross-tabulated numbers

Example: Three Data Sources

	X	Y	Z
D1	4	10	3
D2	2	14	29

	O1	O2
D1	6	11
D2	28	17

	X	Y	Z
O1	3	19	12
O2	3	5	20

Example: Three Data Sources

	X	Y	Z
D1	4	10	3
D2	2	14	29

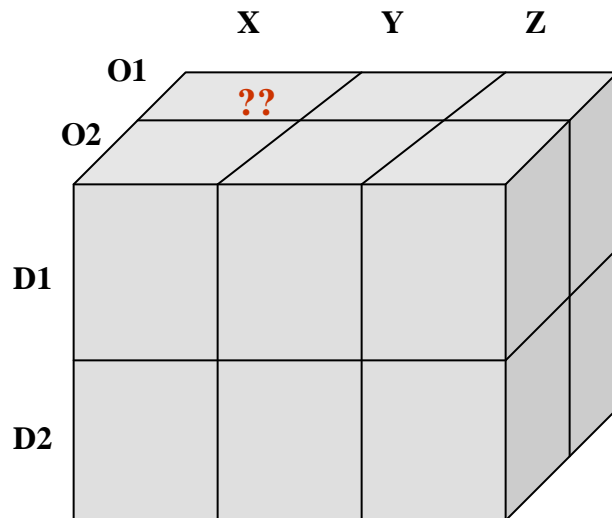
How many of X from O1 → D1?

	X	Y	Z
O1	3	19	12
O2	3	5	20

	O1	O2
D1	6	11
D2	28	17

Example: Three Data Sources

	X	Y	Z
D1	4	10	3
D2	2	14	29



	O1	O2
D1	6	11
D2	28	17

	X	Y	Z
O1	3	19	12
O2	3	5	20

Problem Formulation and Solution

- Denote the cell counts in the 3-way table by $n_{i,j,k}$ $i=\{D1,D2\}, j=\{O1,O2\}, k=\{X,Y,Z\}$
- Objective: $\max/\min n_{D1,O1,X}$
- Subject to linear constraints on $n_{i,j,k}$ that preserve the marginal totals; $n_{i,j,k}$ non-negative integers [E.g., $n_{D1,O1,X} + n_{D1,O2,X} = 4$]
- Use an Integer Programming solver to solve
- Result: $1 \leq n_{D1,O1,X} \leq 1$

Problem Solution (contd.)

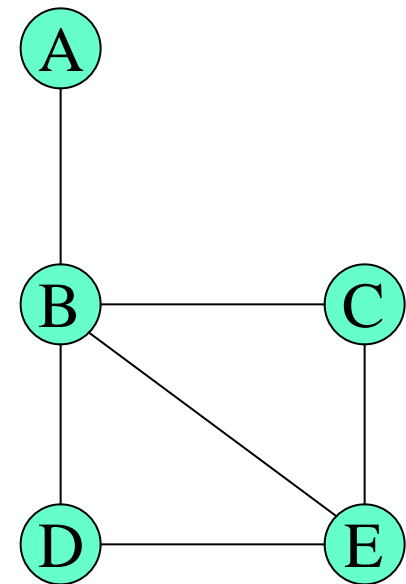
- Example constructed to demonstrate the extreme case: All elements of the 3-way table are exactly determined from the 2-way marginals!!
- More typically, we obtain sharp bounds on the cell counts
- Tightness of the bounds depends on:
 - Dimension of marginals available
 - Number of marginals available
 - Sparseness of the full table

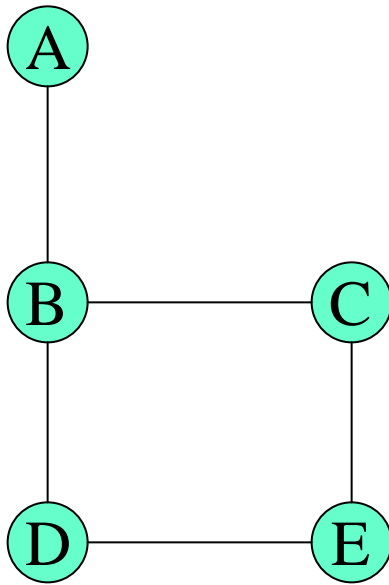
Related Techniques

- **Simulation:** Can run a Markov Chain Monte Carlo simulation to explore the space of all tables that satisfy the marginal constraints (via Gröbner basis technology)
- **Iterative Proportional Fitting** for table reconstruction
- **Scalability:**
 - Heuristic Algorithms for bounds: “Shuttle Algorithm” - seems to work reasonably well when all $(k-1)$ dimensional marginals are known for a k -dimensional table
 - Network Flow formulations for special cases
 - Linear Programming (ignore integrality constraints)
 - Decomposable Graphical Models

Special Case: Decomposable Graphical Models

- A large class of sets of available marginals can be represented as undirected graphs
- If the graph is decomposable (triangulated) then explicit formulas are available for the bounds (*Fienberg-Dobra* work)
- [Graph on the right corresponds to the availability of the (A,B) , (B,D,E) , (B,C,E) marginal tables]





Not decomposable!

“Magnitude” tables

Cells contain real-valued, additive quantities

- Linear Programming can be used for bounds
- Cells with small count and/or dominant contributors are at higher risk of exposure (Statistical Disclosure Control has the (n,p)-rule which says, e.g., “Cells with $n < 3$ and where one of the three accounts for more than $p=0.7$ of the content should be considered risky”)

Concluding Remarks

(DC,DQ) methods can be useful

- Need to explicitly modify them
 - Problem-specific knowledge
 - Discard DC-specific characteristics
- Hopefully, added resources can also be used for tackling problems of scalability, etc.

References

- <http://www.niss.org/dg> : papers, references on cell-bounds and many other things
- <http://www.cs.cmu.edu/~wcohen/matching> : Annotated bibliography on record linkage
- **Leon Willenborg and Ton de Waal:**
 - "Statistical Disclosure Control in Practice" (1996), Springer
 - "Elements of Statistical Disclosure Control" (2000), Springer