

NISS

Data Confidentiality, Data Integration,
Data Mining, Data Quality:
Implications for Counterterrorism

Alan F. Karr

National Institute of Statistical Sciences

karr@niss.org

NISS

Data:

Implications for Counterterrorism

Alan F. Karr

National Institute of Statistical Sciences

karr@niss.org

Outline

- Thesis of the talk
- The problem areas
 - DC = data confidentiality
 - DI = data integration
 - DM = data mining
 - DQ = data quality
- The pairwise intersections: Illustrative examples of what we can and CANNOT do

The Thesis of The Talk

- All of DC, DI, DM, DQ are relevant to counterterrorism
- DC, DI, DQ are known to interact
 - Some aspects identified but none well understood
- All of {DC, DI, DQ} interact with DM
 - Not only understanding but also even identification of central issues lacking
- The interactions are central challenges to statistical scientists and their collaborators involved with counterterrorism

DC, DI, DM and DQ

Testbed Database 1

- **CPS-8** (Current Population Survey): excerpt from 1993 CPS
 - 48,842 data records (not realistic!)
 - 8 categorical attributes (not realistic!)
 - 2880 cells in full table (not realistic!)
 - 1695 cells with non-zero counts (not realistic!)
 - 735 at risk data elements
 - 361 in cells with count 1
 - 374 in cells with count 2

Testbed Database 2

- **FARS** (Fatality Analysis and Reporting System): 1999 data
 - FARS-A: Accident Table
 - FARS-D: Driver Table
 - FARS-P: Person Table
 - FARS-V: Vehicle Table

DC: Data Confidentiality

- Fundamental issue: Tradeoffs between
 - Disclosure risk: data subject IDs, attribute values
 - Data utility: to researchers, public, counterterrorists, ...
- Approaches to statistical disclosure limitation (SDL)
 - Restricted access: special sites, licensing, ...
 - Restricted data
 - Tables: release only selected marginals, cell suppression, ...
 - Aggregation: top-coding, geographical, ...
 - Altered data, typically by introduction of randomness
 - PRAM: data swapping, synthetic data, jittering, ...
 - Disseminate “safe analyses” rather than data
 - Regression servers

Example: Tabular Data

- Database: large (40 dimensions x 4 categories each) contingency table (2^{80} cells!)
- Static optimal tabular releases: set of marginals that
 - Maximizes utility
 - Satisfies upper bound constraint on risk
- Utility
 - Simple: number of marginals released
 - Useful: “Correctness” of inference using log-linear models

Tabular Data - 2

- Risk: two kinds
 - Disclosure risk: capability to bound small count (1 or 2) cells in full table
 - Transparency risk: Are statements of the form “Inference is correct” risky?
- Reference
 - Dobra, Fienberg, Karr, Sanil (2002). Software systems for tabular data releases. *IJUFKS* **10(5)** 529-544

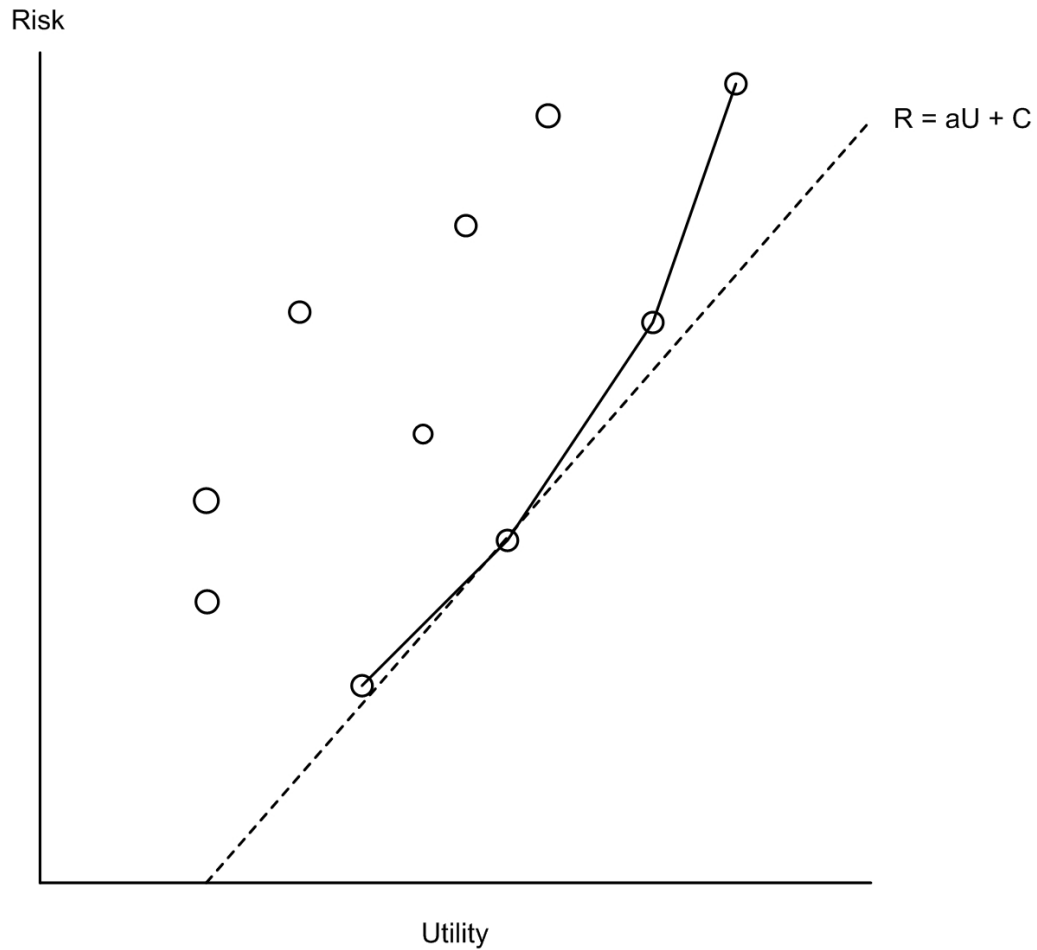
Example: Data Swapping

- Basic idea: switch subset of attributes between randomly selected pairs of records at microdata level
- Rationale: reduces disclosure risk—intruder cannot be certain that any record is real
- Side effect: distorts data, reducing utility
 - Changes (only) joint distributions that involve both swapped and unswapped attributes

Risk-Utility Framework

- Choice problem: must select
 - Swap rate
 - Swapped attributes
 - Optionally, constraints on unswapped attributes
- Characterize each candidate release by
 - Disclosure risk. Example: unswapped records in small count cells in post-swap table.
 - Data utility. Example: dis-utility = data distortion, as measured by Hellinger distance, or ...
- No unique solution, but restrict attention to frontier of undominated releases

Example Frontier



References

- Gomatam, Karr (2003). Distortion measures for categorical data swapping. Submitted to *JOS*.
- Gomatam, Karr, Sanil (2003). A risk-utility framework for categorical data swapping. Submitted to *JOS*.
- Sanil, Gomatam, Karr, Liu (2003). **NISS WebSwap: A Web Service for data swapping.** *J. Statist. Software* **8(7)**.

DC and Counterterrorism

- Very complex problem
 - Highly charged politically
 - Multiple stakeholders: individuals, data holders, government agencies, ...
- Fundamental inconsistencies between DC and some counterterrorism efforts, especially those meant to prevent terrorism

DI: Data Integration

- Fundamental issue: Management of and inference from data created by combining multiple, “related” databases, often assembled by different organizations for different purposes
- Approaches from different disciplines
 - Database: joins
 - Data warehousing: parsing and standardization
 - Statistics: probabilistic record linkage
 - AI/NL: machine translation

Example: DI via Machine Translation

- View the data mapping problem as a variant of the cross-language mapping problem of machine translation
- Use EM-based statistical algorithms
- Reference: E. Hovy (2003). Semi-automatic data integration. *Proc. dg.o 2003*.

DI and Counterterrorism

- To many, DI is central to counterterrorism
- Paucity of automatic methods
- Scale is a significant issue
 - Example: aberrant purchasing patterns among connected set of people requires joining, across multiple sellers,
 - Transaction table (what was bought)
 - Customer table (who bought it)
 - Store table (where it was bought)

DM: Data Mining

- Fundamental issue: discovery of information and knowledge in large, complex (and often unstructured) data sets
- Purposes
 - Pattern discovery
 - Identification of anomalous data (“looking for needles in haystacks”)
 - Use simple analyses, in order to overcome scalability problems with complex statistical procedures

Example: Association Rules

- D = database
- A, B subsets of D
- Association rule: $A \Rightarrow B$
 - Confidence: $C = |A \cap B| / |A| = P(B | A)$
 - Support: $S = |A \cap B| / |D| = P(A \cap B)$

Example: CPS-8

Age

Salary

	<25	25-55	>55
<\$50K	8339	23912	4904
>=\$50K	93	9629	1965

Support

	<25	25-55	>55
<\$50K	.170	.490	.100
>=\$50K	.002	.197	.040

Confidence given Age

	<25	25-55	>55
<\$50K	0.99	.71	.71
>=\$50K	.01	.29	.29

DM and Counterterrorism

- DM tools are too blunt for *very rare* patterns
 - Susceptible to false positives
- DM may not be necessary or effective
 - We have good tools to identify *known* patterns
 - Disease surveillance
 - Computer intrusion detection
 - T. Tether (5/6/03): DARPA is pursuing an “approach of searching for evidence of specified patterns” and “detecting in stages”

DQ: Data Quality

- Fundamental issue: Characterize and improve capability of data to be used effectively, economically and rapidly to inform and evaluate decisions
 - Multi-dimensional: accuracy, accessibility, relevance, timeliness, metadata, documentation, user capabilities, user expectations, cost, domain knowledge
 - Multi-disciplinary: statistics, computer science, total quality management (TQM)

Scale and Ubiquity of DQ Problems

- FARS-A, 1999 (one version)
 - 18,433 records, 49 attributes
- Intactness: Ignoring (lat,long), which is present in only 71 records, only 30% of records
 - Have no missing values
 - Pass 3 simple consistency checks (Example: temporal precedence)
- 4 attributes have the same value for all records
 - Produces high confidence, high support, but meaningless association rules

FARS-A 1999 Excerpt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CSTATE	CNUM	SEQNUM	VEHNUM	LNUM	PNUM	CITY	COUNTY	ACCDAT	ACCTIME	VEHFORM	PFORMS	NMOTFO	NHS
5369	53	183	0	0	1	0	690	61	6041999	327	1	1	0	1
5370	53	215	0	0	1	0	740	73	5261999	1205	1	2	0	1
5371	53	242	0	0	1	0	0	11	7101999	1125	2	5	0	1
5372	53	359	0	0	1	0	2230	53	9021999	2002	3	8	0	1
5373	53	383	0	0	1	0	1960	33	8261999	45	1	2	1	1
5374	53	412	0	0	1	0	0	61	8211999	30	1	3	0	1
5375	53	429	0	0	1	0	0	53	10171999	141	2	6	0	1
5376	53	431	0	0	1	0	0	67	10991999	9999	1	1	0	1
5377	53	446	0	0	1	0	2310	33	10231999	1815	4	6	0	1
5378	53	486	0	0	1	0	0	67	10111999	1910	3	5	0	1
5379	53	510	0	0	1	0	0	11	11261999	1359	2	6	0	1
5380	53	518	0	0	1	0	0	67	12101999	1347	1	1	0	1
5381	53	527	0	0	1	0	1000	15	12171999	1305	3	5	0	1
5382	6	1327	0	0	1	0	0	113	8011999	1	1	5	1	1
5383	32	34	0	0	1	0	0	3	1111999	945	1	1	0	1
5384	5	382	0	0	1	0	2320	119	9121999	1421	1	2	0	1
5385	17	4	0	0	1	0	0	63	1011999	2010	1	2	0	1
5386	17	16	0	0	1	0	3105	31	1051999	1415	1	2	1	1
5387	17	36	0	0	1	0	0	117	1131999	820	1	1	0	1
5388	17	45	0	0	1	0	0	167	1181999	355	1	2	0	1
5389	17	128	0	0	1	0	0	119	2211999	555	1	1	0	1
5390	17	238	0	0	1	0	0	43	4121999	2332	4	6	1	1
5391	17	380	0	0	1	0	0	135	5171999	45	1	1	0	1
5392	17	410	0	0	1	0	0	197	5291999	310	1	2	0	1
5393	17	459	0	0	1	0	0	197	6171999	102	4	4	0	1
5394	17	482	0	0	1	0	9340	167	6241999	1629	2	12	0	1
5395	17	659	0	0	1	0	8730	119	8071999	1352	1	2	0	1
5396	17	696	0	0	1	0	2610	163	8161999	5	2	7	0	1
5397	17	779	0	0	1	0	0	105	9011999	1435	1	1	0	1
5398	17	867	0	0	1	0	0	197	9301999	2332	2	2	0	1
5399	17	879	0	0	1	0	8410	31	10111999	2305	1	2	1	1
5400	17	1159	0	0	1	0	8410	31	11211999	544	2	3	0	1

Other FARS DQ Problems

- Numerical codes for categorical variables
 - Number of lanes: 7 means “7 or more”
- Multiple representations for missing attributes, some of which are valid data values
 - Age = 99
- Partially missing attributes
 - ACCDATE = 10991999
- “Unjoinable” tables that should be in 1-1 correspondence
 - FARS-D and FARS-V: in some versions, join is empty

DQ and Counterterrorism

- Poor DQ hampers counterterrorism
- Fundamental DQ gaps
 - Quantification: almost no meaningful DQ metrics
 - Models of the causes or effects of poor DQ
 - Do know how to use EDA to understand DQ problems
 - How to deal with costs
 - Multiple stakeholders, who shift costs to one another
 - Elusive costs
 - Uncertain costs

Overarching “Don’t Knows”

- DC, DI, DM and DQ for data that are not categorical or numerical
 - Examples: text, images, video, audio
- How to deal with costs
 - False positives
 - False negatives
- How to create benchmarks to evaluate different strategies
- DC, DI, DM and DQ for large quantities of machine-generated data (“data streams”)

DC-DI Interactions

DC \cap DI: What We Do Know

- Tools for DI also can be used to break DC
 - Example: Record linkage
 - Medical database from Cambridge, MA contained
 - Date of birth
 - Zip code
 - Linkage (by hand) to public voter list identified 90+% of records
- Databases to integrate with abound

DC \cap DI: What We Don't Know

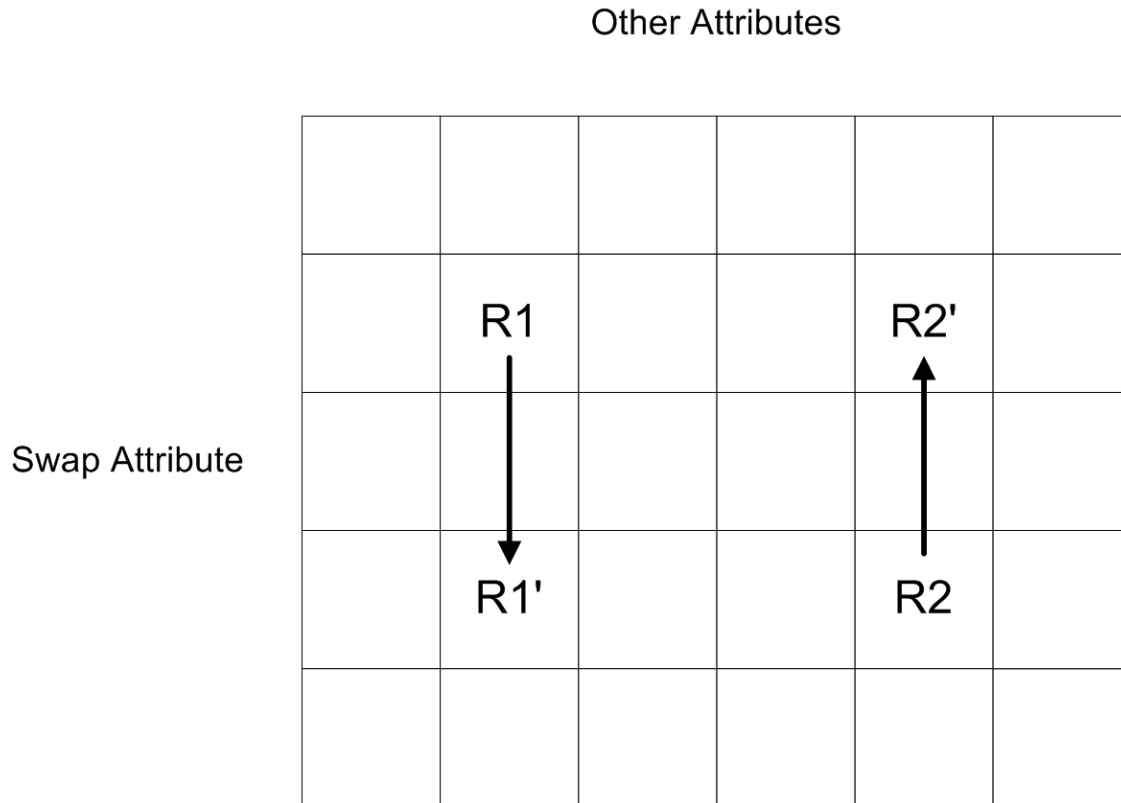
- Right abstractions for the problem
 - “Correct” universe of DI candidates to break DC
 - Risk: (Probability of) re-identification too simplistic
 - Alternatives
 - Cost
 - Effort
 - What about false positives?
- How to differentiate among methods for DI

DC-DM Interactions

DC \cap DM: What We Do Know

- SDL creates anomalies
- Initial step: Privacy-preserving data mining
 - Local computation (Clifton, et al.)
 - “PRAM”-based approaches (Agrawal & Srikant, Privacy Preserving Data Mining, *Proc. ACM SIGMOD Conf. on Management of Data*, 2000)

SDL-Created Anomalies



Privacy-Preserving Association Rules

- Problem
 - Multiple but identical (!?) databases
 - Find item pairs (A_i, A_j) with *global* (across all the databases) support $\geq s\%$
 - Only local computations, performed by database owners
 - Protect
 - Data items
 - Value of support at each site
 - Database sizes

Local Computation: The Algorithm

- Compute: $1(\sum_k C_k(i, j) \geq s \sum_k N_k)$
- Procedure
 - Site 1
 - Generate (large) random number R
 - Calculate and transmit to site 2: $X_1 = R + C_1(i, j) - sN_1$
 - Site m
 - Calculate and transmit to site $m+1$: $X_m = X_{m-1} + C_m(i, j) - sN_m$
 - Site 1
 - Check whether $X_k \geq R$
 - Distribute result to other sites

DC \cap DM: What We Don't Know

- Abstractions for disclosure risk associated with DM
 - Relationships in the data
- Whether, compared to other statistical analyses, DM is
 - Inherently more threatening to confidentiality
 - A different kind of threat
- Effect of most SDL strategies on DM
- Whether DM can be used to defeat SDL
 - Example: Can DM detect swapped records?

Example: CPS After Swapping

Age

	<25	25-55	>55
Salary <\$50K	7848	24292	5015
Salary >=\$50K	584	9249	1854

Support

	<25	25-55	>55
<\$50K	0.161	.497	.103
>=\$50K	.012	.189	.038

Confidence given Age

	<25	25-55	>55
<\$50K	.93	.72	.73
>=\$50K	.07	.28	.27

DC-DQ Interactions

DC \cap DQ: What We Do Know

- Poor DQ is said to protect DC
 - “Poor data quality is the best form of SDL”
- In the survey world, some DC \cap DQ issues are well understood
 - Non-response
 - Effects of instrument, collection technique, ...

DC \cap DQ: What We Don't Know

- The extent to which perceived protection of confidentiality affects DQ
 - Anecdotal evidence is that respondents lie if confidentiality is problematic
- Whether poor DQ really does protect DC
 - Can “false” records be identified (e.g., by DM)?
- Right level to think about DC
 - Individual
 - Population

DI-DM Interactions

DI \cap DM: What We Do Know

- Many disconnects between DI and DM
 - Database joins: scalability problem
 - Relational data mining (Siebes): limited DM capability
 - Record linkage: presumes clean correspondence between attributes [, good model for $P\{\text{Match}|S_1, S_2\}$]
- Difficulties with DI inhibit DM
 - Example: Ford Explorer/Firestone tire problem
 - FARS-V (NHTSA)
 - Warranty/service data (Manufacturers, dealers, JD Power,...)
 - Manufacturing data (Manufacturers)
 - Common thread = VIN (Example: 1FTDF15Y0KNB)

DI \cap DM: What We Don't Know

- How to track the effects of particular DI methods
- How to do DM when DI is impossible
 - Example: patterns in fatal accidents involving transit vehicles
 - FARS-A: CITY, COUNTY, TRID
 - FARS-V: MAKE, MODEL, BDTYP, MODYR
 - NTD: TRS_ID, cPER_VEH, dPER_VEH (patron and employee in-employee in-vehicle fatalities, per year)
 - Neither joins nor record linkage is possible!
 - What to do???

DI-DQ Interactions

DI \cap DQ: What We Do Know

- Inability to do DI is a DQ problem
- Metadata quality may be the real issue
 - Absence of foreign keys
 - Inability to match “the same attributes”
 - Example: gasoline and petrol
 - Example: different units
 - Different definitions of “the same” attribute
 - Different attribute categories
- Changes over time compound the problem

DI \cap DQ: What We Don't Know

- Whether DI improves DQ
 - More precisely, what effects different methods for DI have on DQ
- How to do DI in ways that don't impair DQ

DM-DQ Interactions

DM \cap DQ: What We Do Know

- DQ problems are not susceptible to automation
- Many data sets are not (completely) real
- In real data sets, many points are anomalous
- Anomalous data are likely to be wrong

“Unreal” Data

- SDL applied
 - Data swapping
 - Synthetic data
- Statistical adjustments
 - Imputation of missing responses
- Data clean-up
 - DQ edits. Examples:
 - Address corrections
 - (male,hysterectomy) → (female,hysterectomy)

Prevalence of Anomalous Points

- CPS-8: 1.5% of data
 - 361 data elements in cells with count 1
 - 374 data elements in cells with count 2
- Worse in higher dimensions
 - Example: 14-dimensional CPS data
 - 435,000,000 cells
 - 299,285 data elements
 - 72,739 data elements (24.3%) in cells with count 1 or 2

Anomalous Data Are Often Wrong

- Example
 - FARS LTIME = [FARS-P:(DEATHDATE,DEATHTIME - FARS-A:(ACCDATE,ACCTIME)], in FARS-P
 - Sort FARS-P by LTIME
 - Largest value = 72000: (CSTATE=54, CNUM=321)
 - FARS-A: ACCDATE = *11*/7/99, ACCTIME = 19:27, NFAT = 2
 - FARS-P: For both fatalities, DEATHDATE = *12*/7/99, DEATHTIME = 19:27, LTIME = 72000
- Implication: Large numbers of false positives in low quality data—lots of things look like needles, but aren't needles!

DM \cap DQ: What We Don't Know

- Whether DM can be used to
 - Characterize DQ
 - Improve DQ
- How to build robust DM tools

The Grand Challenge

$DM \cap DC \cap DI \cap DQ$
(\cap Counterterrorism)