

Privacy and Confidentiality in a Homeland Security Context

Stephen E. Fienberg

Department of Statistics, CALD, Cylab

Carnegie Mellon University

Pittsburgh, PA USA

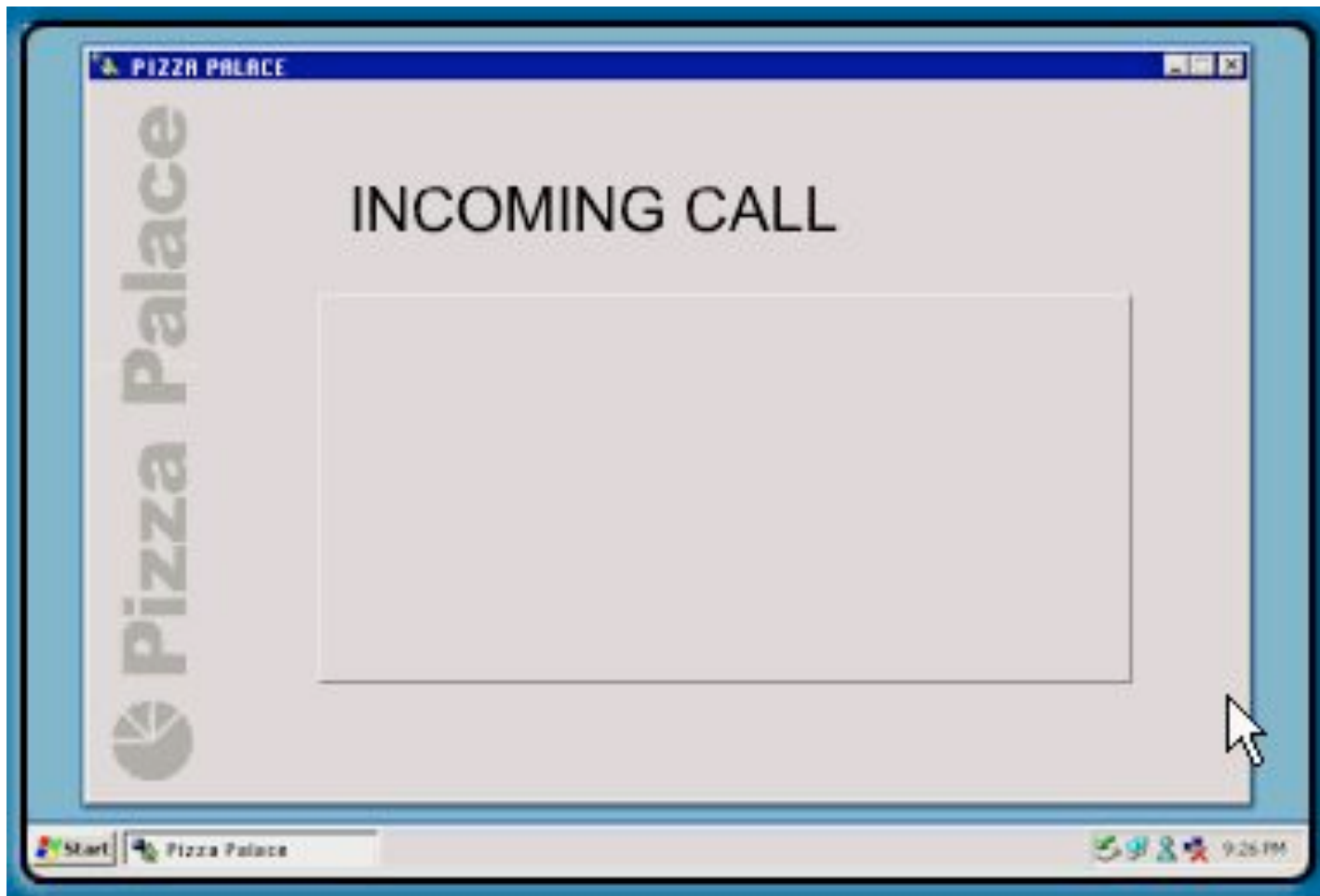
and

IMA, University of Minnesota

Overview

- **Some background homeland security, datamining, and privacy concerns.**
- **Splitting and linking data bases:**
 - Matching and confidentiality.
 - Matching and privacy-preserving datamining.
 - Selective revelation?
- **Statistical approach to the problem:**
 - Ideas and results for categorical data.
 - Link to datamining for association rules.
 - Network data for detecting terrorists.
- **Lessons?**

Myth or Reality?



Courtesy of the American Civil Liberties Union.

Personal Data for the Taking

By **TOM ZELLER Jr.**

May 18, 2005

- A class at Johns Hopkins was able to build detailed dossiers on Baltimore citizens using only public databases.



```

LAT N, CITY
HERE MONTH
ORDER BY TEMP
<string.h> EXEC SQL
long station_id; long mon;
TABLE STATION ID INTEGER
STATE CHAR(2), LAT N REAL
TABLE STATS (ID INTEGER
MONTH INTEGER CHECK
TEMP F SPOUSE CHECK (TEMP
RAIN I REAL CHECK (RAIN I
PRIMARY KEY (ID, MONTH)); INSERT
ES (13, 1, 57.4, 0.31); SELECT LAT
FROM STATS, STATION WHERE MONTH
ATS.ID = STATION.ID ORDER BY TEMP
<stdio.h> #include<string.h> EXEC SQL
SECTION; long station_id; long mon; float
; char city_name[21]; long SQLCODE; EXEC
CLARE SOCIAL SECURITY NUMBER CONNECT
needed, goes here * strcpy(city_name,
QL SELECT ID INTO :station_id FROM STAT
CITY = :city_name; if (SQLCODE = -100)
s no station for city %s\n", city_name); exit
For the city %s, Station ID is %ld\n", city
_id); printf("And here is the weather data:
L DECLARE HOME ADDRESS CURSOR FOR SELECT
, RAIN I FROM STATS WHERE ID = :station
ER BY MONTH; EXEC SQL OPEN XYZ; while
DE != 100) EXEC SQL FETCH XYZ INTO :mo
ain; if (SQLCODE == 100) printf("end of
ntf("month = %ld, temperature = %f, rain
non,temp,rain); EXEC SQL CLOSE XYZ; ex
TABLE STATION ID INTEGER PRIMARY KEY,
) STATE CHAR(2), LAT N REAL, LONG W
TE TABLE STATS (ID INTEGER REFERENCE
ID), MONTH INTEGER CHECK (MONTH BE
), TEMP F REAL CHECK (TEMP F BETWE
), RAIN I REAL CHECK (RAIN I BETWEEN 0
PRIMARY KEY (ID, MONTH)); INSERT INTO
ES (13, 1, 57.4, 0.31); SELECT LAT N,
FROM STATS, STATION WHERE MONTH
ATS.ID PERSONAL POLITICS ORDER BY TEMP F,
<stdio.h> #include<string.h> EXEC SQL
SECTION; long station_id; long mon;
; char city_name[21]; long SQLCODE;
CLARE SECTION; main / * the CONNECT
needed, goes here * strcpy(city_name,
QL SELECT ID INTO :station_id FROM STAT
CITY = :city_name; if (SQLCODE == 100)
s no station for city %s\n", city_name)
For the city %s, S ACCOUNTS %ld\n", city
_id); printf("And here is the weather
L DECLARE XYZ CURSOR FOR SELECT
, RAIN I FROM STATS WHERE ID
ER BY MONTH; EXEC SQL OPEN XYZ;
DE != 100) EXEC SQL FETCH XYZ INTO
ain; if (SQLCODE == 100) printf("end of
ntf("month = %ld, temperature = %f,
TATION ID INTEGER PRIMARY KEY,
TATE CHAR(2), LAT N REAL, LONG
TE TABLE STATS (ID INTEGER
MONTH CHILDREN CHECK
TEMP F REAL CHECK (TEMP F
RAIN I REAL CHECK (RAIN I BET
PRIMARY KEY (ID, MONTH)); INSERT
ES (13, 1, 57.4, 0.31); SELECT
FROM STATS, STATION WHERE MOI
ATS.ID = STATION.ID ORDER BY
TABLE STATION ID INTEGER
Y CHAR(20), STATE CHAR(2), LAT N
CREATE TABLE
STATION(ID), MONTH
1 AND 12

```

In Plain Sight

Personal information is held in thousands of public databases — not just those compiled by big data brokers like ChoicePoint — and new limits are being proposed on Internet access to such data.

Land Deeds

Many county governments have made these records available online, often as scanned images of the original document.

Occupational Licenses

The A.C.L.U. has filed a lawsuit against the State of Alaska, which, like many states, posts the addresses of licensed professionals.

Voter Registrations

Party affiliations and campaign contributions are a part of the public record, but privacy advocates say some citizens do not vote because of this.

Court Records

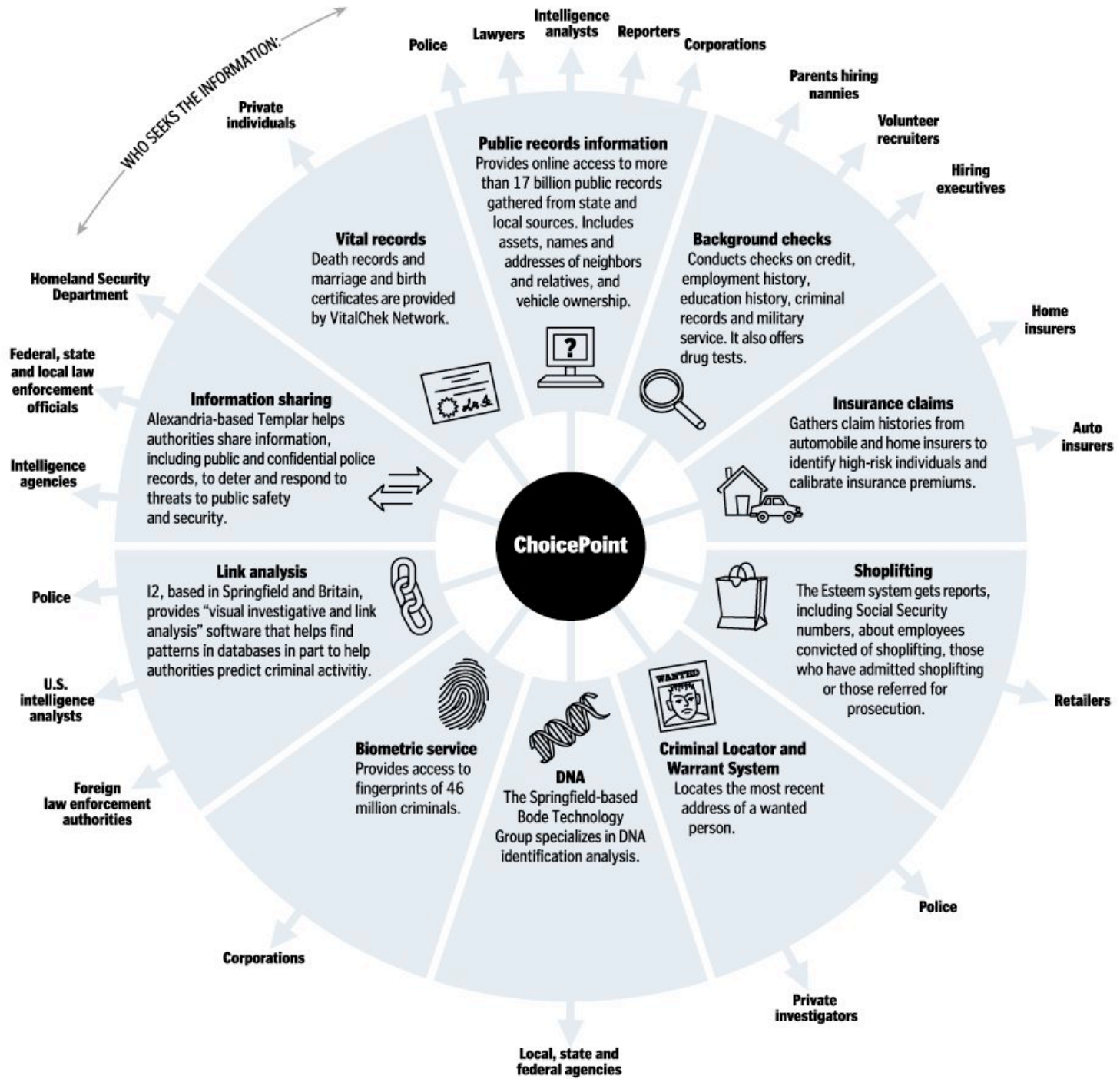
These often contain not only Social Security numbers, but also information about dependent children, birth dates, bank accounts, and even unique identifiers like a mother's maiden name.

Sources: Privacy Rights Clearinghouse; Virginia Watchdog

Transactions Data

- **Some examples:**
 - **Online banking.**
 - **Scanner data.**
 - **EZ-Pass for highway fee**
 - **Airline trips.**
 - **RFIDs in passports?**
- **Who controls what?**
- **Who has access to what?**

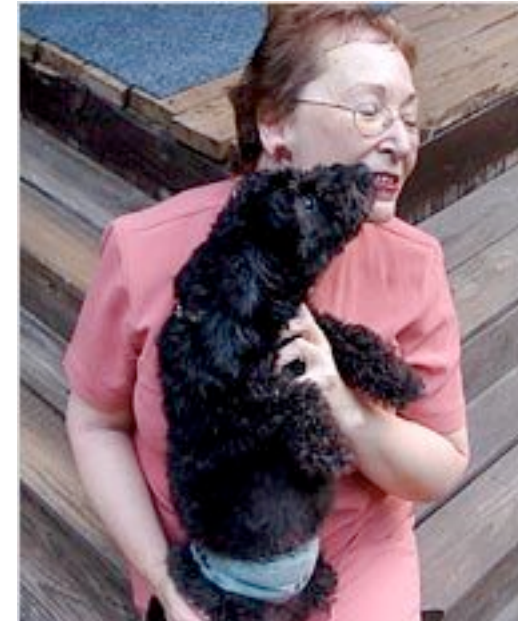




2006 AOL Data Release Fiasco

A Face Is Exposed for AOL Searcher No. 4417749 - New York Times

The New York Times
nytimes.com



August 9, 2006

A Face Is Exposed for AOL Searcher No. 4417749

By [MICHAEL BARBARO](#) and [TOM ZELLER Jr.](#)

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga.," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her grandchildren. "I was thinking about my grandchildren," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the weekend and apologized for its unauthorized move by a team that had hoped it would benefit academic researchers.

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans which continue to circulate online, underscore how much people unintentionally reveal when they use search engines — and how risky it can be for companies like AOL, Google, and others to compile such data.

Those risks have long nipped privacy advocates against online marketers and other Internet

What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizarre.com
4417749	jcpenny	2006-04-28	14:16:05	1	http://www.jcpenny.com
4417749	tortus and turtles	2006-04-29	13:12:47	1	http://www.manchesterterrier.com
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26	1	http://www.manchesterterrier.com
4417749	fingers going numb	2006-05-02	17:35:47	1	http://www.manchesterterrier.com
4417749	dances by laura	2006-05-02	17:59:32	1	http://www.manchesterterrier.com
4417749	dances by lori	2006-05-02	17:59:57	1	http://www.manchesterterrier.com
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:50	1	http://solosingles.com
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going through a divorce. I thought there was a place called 'Dances by Lori' for sinolees."

Homeland Security and the Search for Terrorists

- **Total Information Awareness (TIA)**
 - Aborted DARPA program to link databases for detection, classification and identification of terrorists.
 - Designed to use transaction data.
- **Multistate Anti-Terrorism Information Exchange (MATRIX) Project**
 - System to link public/private records, drivers license information, motor vehicle information, criminal offender information, business/corporate data.
 - Datamining application, called FCIC Plus.
 - Developed with the help of the FBI, INS, DEA, and the U.S. Secret Service.
- **Airline Watch Lists**



Record Linkage

- ***Record linkage***: attempt to determine whether pairs of *data records* describe the same entity, i.e., find record pairs that are *co-referent*:
 - **Entities**: usually people (or organizations or...).
 - **Data records**: names, addresses, job titles, birthdays, ...
- **Main applications**:
 - *Joining two **or more** heterogeneous relations*
 - *Removing duplicates from a single relation*
- **AKA**: data matching, merge/purge, duplicate detection, data cleansing, ETL (extraction, transfer, and loading), de-duping, ...

Methods for Linking Databases

- **Using unique identifiers? ID numbers, names, etc.**
- **Probabilistic matching or record linkage:**
 - What methods? With what levels of accuracy and what assumptions.
 - Fellegi-Sunter methods have implicit assumptions of overlap; rely on accuracy of variables for blocking to enable comparison of records.
 - Logistic regression-like models.
 - “String-distance” metrics.
- **What do ChoicePoint and Seisint really do??**
 - Contract between MATRIX and Seisint (the private company supplying MATRIX with data now owned by LexusNexis) states that it cannot guarantee the “correctness or completeness” of data in system.



Fellegi-Sunter Key Ideas I

- **Represent** every pair of records using vector of features (variables) that describe similarity between individual record fields, e.g., Boolean (e.g., last-name matches), discrete (e.g., first- n -characters-of-name-agree), or continuous (e.g., string-edit-distance-between-first-names).
- **Place** feature vectors for record pairs into three classes: matches, non-matches, and possible matches.

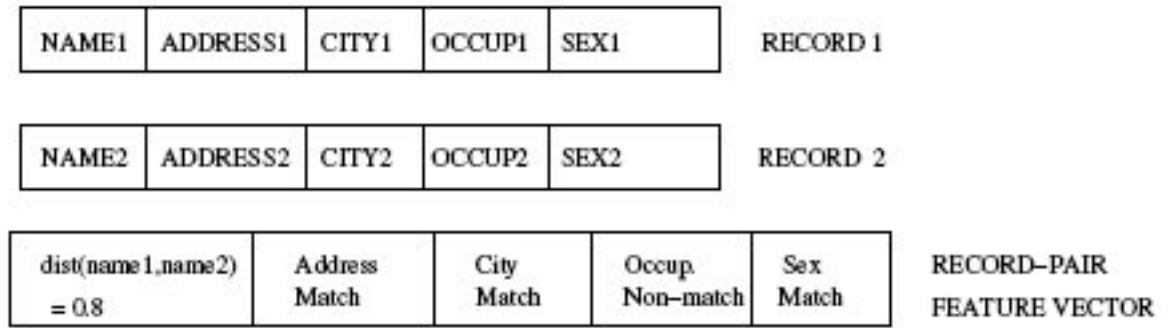
Fellegi-Sunter II

- **Perform** record-pair classification by calculating the ratio $P(\gamma | M)/P(\gamma | U)$ for each candidate record pair, where:
 - γ is feature vector for pair.
 - $P(\gamma | M)$ and $P(\gamma | U)$ are probabilities of observing that feature vector for matched and non-matched pair, respectively.
- **Choose** two thresholds based on desired error levels— T_μ and T_λ —optimally separate ratio values for equivalent, possibly equivalent, and non-equivalent record pairs.

Fellegi-Sunter III

- **When no training data in form of duplicate and non-duplicate record pairs are available, matching can be unsupervised:**
 - **Estimate** conditional probabilities for feature values using observed frequencies.
- **Because most record pairs are clearly non-matches, we need not consider them for matching.**
 - **Way to manage this is to “block” data bases (e.g., using geography or some other variable in both data bases) so that only records in comparable blocks are compared.**

Hierarchical Graphical Model for Record Linkage



- **New modeling approaches to matching.**
- **Link to analytical tools to be applied to matched database.**

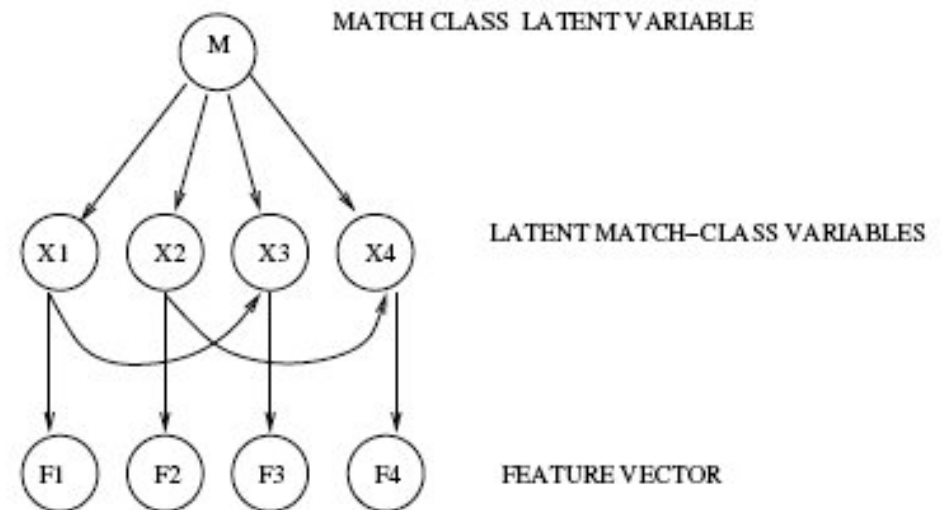


Figure 3: Hierarchical generative model for Record Linkage

Protection from Datamining: “Need to Know” Principle

- Keeping different kinds of data together allows potential intruder (including government) to make inferences nobody *needs* to make.

Name/ID	Location	Job Skills	Medical	Languages
John Brown #9321	Ft. Bragg	Radio operator	Allergic penicillin	None
Anton van Graal #6531	Ft. Bragg	Tank repair	Broken wrist, April 1999	Dutch
...

Why are explosives experts are assigned to location X?

Why are Indonesian speakers now being trained as radio operators?

may limit

Splitting DB ~~limits~~ Potential Damage From Intruders....

Name/ID	Location	Job Skills	Medical	Languages
John Brown #9321	Ft. Bragg	Radio operator	Allergic penicillin	None
Anton van Graal #6531	Ft. Bragg	Tank repair	Broken wrist, April 1999	Dutch
...

Name/ID	Location
John Brown #9321	Ft. Bragg
Anton van Graal #6531	Ft. Bragg
...	...

Name/ID	Medical
John Brown #9321	Allergic penicillin
Anton van Graal #6531	Broken wrist, April 1999
...	

Name/ID	Languages
John Brown #9321	None
Anton van Graal #6531	Dutch
...	... 16

Encryption

- Useful for “secure” transmission of data, and as a tool in “privacy-preserving” methods.
- But every encryption method has its limits!!
- And encryption can’t protect privacy if people get access to un-encrypted files.

COMPUTER SECURITY

Flaw Found in Data-Protection Method

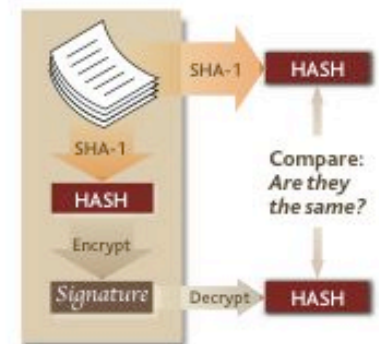
Cryptographers are making a hash of things again. Last month three code breakers demonstrated a way to break the Secure Hash Algorithm (SHA-1), a government-approved standard cryptographic function crucial to many electronic transactions, including digital signature schemes and password verification. Although the finding doesn't mean that SHA-1 is unusable, it has prompted the cryptographic community to suggest finding more secure versions of SHA. “The research community is going to have to think very hard about this,” says Massachusetts Institute of Technology cryptographer Ron Rivest. “We clearly have to replace SHA-1.”

A hash function is a mathematical device that takes a chunk of text (which can be huge) and, through a series of arithmetic manipulations, turns that text into a number (which is small). Hash functions allow computer programs to verify that large blocks of text or data are unaltered without needing to store the large files themselves. For example, you might know that a particular hash function spits out the number 634,331,206 when given the authentic text of *War and Peace*. If someone gives you a file of text, you just run the hash function on it. If the number 634,331,206 doesn't come out, the file can't be an unaltered copy of *War and Peace*.

That is what many operating systems do with passwords: Rather than storing passwords in an easy-to-steal file, they store the passwords' hash values instead. Even if hackers get hold of the list of hash values, they don't know how to turn those values into valid passwords and get into the system. Of course, this is true only if the program can't be run backward. To guarantee a one-way hash function, the National Institute of Standards and Technology (NIST) in the early 1990s introduced SHA-1.

Now, more than a decade later, three researchers from China and the United States have devised the first successful attack on SHA-1. Although they don't force the algorithm to run backward, in an unpublished paper circulating among computer-security experts they show how to do a

related trick. “What we have done is shown something called a collision,” says Yiqun Lisa Yin, an independent consultant in Greenwich, Connecticut. “Two different messages map to the same outcome.” In other words, Yin and her two colleagues, Xiaoyun Wang and Hongbo Yu of Shandong University in East China, came up with a way to find different blocks of text that have identical hash values. In theory, hackers could use the trick to forge stamps of authenticity for electronic documents.



Hashed Hancock. A digital signature scheme using hash functions (such as SHA-1) and ciphers may be vulnerable to forgery.

An attacker, by pure brute force, would expect to find one such collision in 2^{80} attempts. The team shows how to reduce that value to 2^{69} tries—still out of the range of supercomputers, but close enough to worry experts. Rivest thinks NIST should hold a competition to design a next-generation hash algorithm. NIST has no plans for such a competition, says Edward Roback, chief of NIST's Computer Security Resource Center, but is encouraging users to switch to beefed-up versions of SHA: “It's not like SHA is completely broken, but any time the security of an algorithm is less than expected, it's a concern.”

—CHARLES SEIFE

Privacy-Preserving Datamining

- **Special multi-party computation protocols using “secure transmission,” e.g., encryption.**
 - Presumes semi-honest participants.
- **Designed to calculate quantities on merged data bases without sharing data per se:**
 - Horizontally and Vertically partitioned cases.
 - **Real world applications are far more complex.**

Secure Multi-Party Computation I

- Technique consists of protocol for exchanging messages among parties A , B , C , etc.
- Assume the parties are semi-honest, i.e., they correctly follow the protocol specification, yet attempt to learn additional information by analyzing the messages that are passed.
- Parties A and B have encryption functions E_A (known only to A) and E_B (known only to B) such that for all \mathbf{x} , $E_A(E_B(\mathbf{x})) = E_B(E_A(\mathbf{x}))$.

Secure Multi-Party Computation II

- A 's database consists of list \mathbf{A} and B 's consists of list \mathbf{B} .
- A sends B message $E_A(\mathbf{A})$.
- B computes $E_B(E_A(\mathbf{A}))$ and then sends to A two messages: $E_B(E_A(\mathbf{A}))$ and $E_B(\mathbf{B})$.
- A then applies E_A to $E_B(\mathbf{B})$, yielding $E_B(E_A(\mathbf{A}))$ and $E_A(E_B(\mathbf{B}))$.
- A computes $E_B(E_A(\mathbf{A})) \cap E_A(E_B(\mathbf{B}))$.
- Since A knows the order of items in \mathbf{A} , A also knows the order of items in $E_B(E_A(\mathbf{A}))$ and can quickly determine $\mathbf{A} \cap \mathbf{B}$.

Secure Multi-Party Computation III

- **The main problems with this approach are**
 - It is asymmetric, i.e., B must trust A to send $A \cap B$ back.
 - It presumes semi-honest behavior.
- **For examples of related ideas in secure multi-party computation of regressions, see the work at NISS.**
 - **Leakage with each round of transmission.**

Other Problems with PPDM

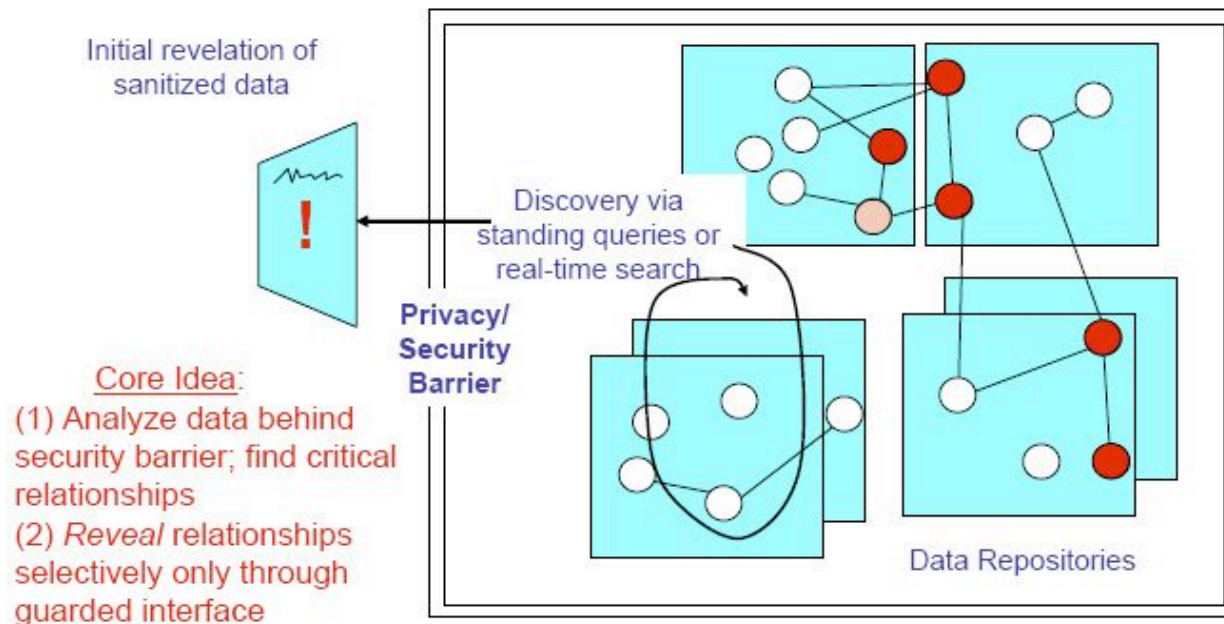
- Presumes data are error-free and thus can be matched perfectly in absence of privacy concerns.
 - Can extend ideas to include probabilistic matching/linkage.
- Focuses on sharing “results” of some computation, e.g.,
 - Dot products and regression computations.
 - Description of an association rule.
- So-called security proofs are designed to protect database owners, *but not individuals*.

TIA Security Program

- **Three-fold plan:**
 - *Inference control* to prevent unauthorized individuals from completing queries that would allow identification of ordinary citizens.
 - *Access control* to return sensitive identifying data only to authorized users.
 - *Immutable audit trail* for accountability.
- **Methods:**
 - PPDM.
 - “Privacy appliances.”
 - *Selective revelation.*

Can We Have Selective Revelation?

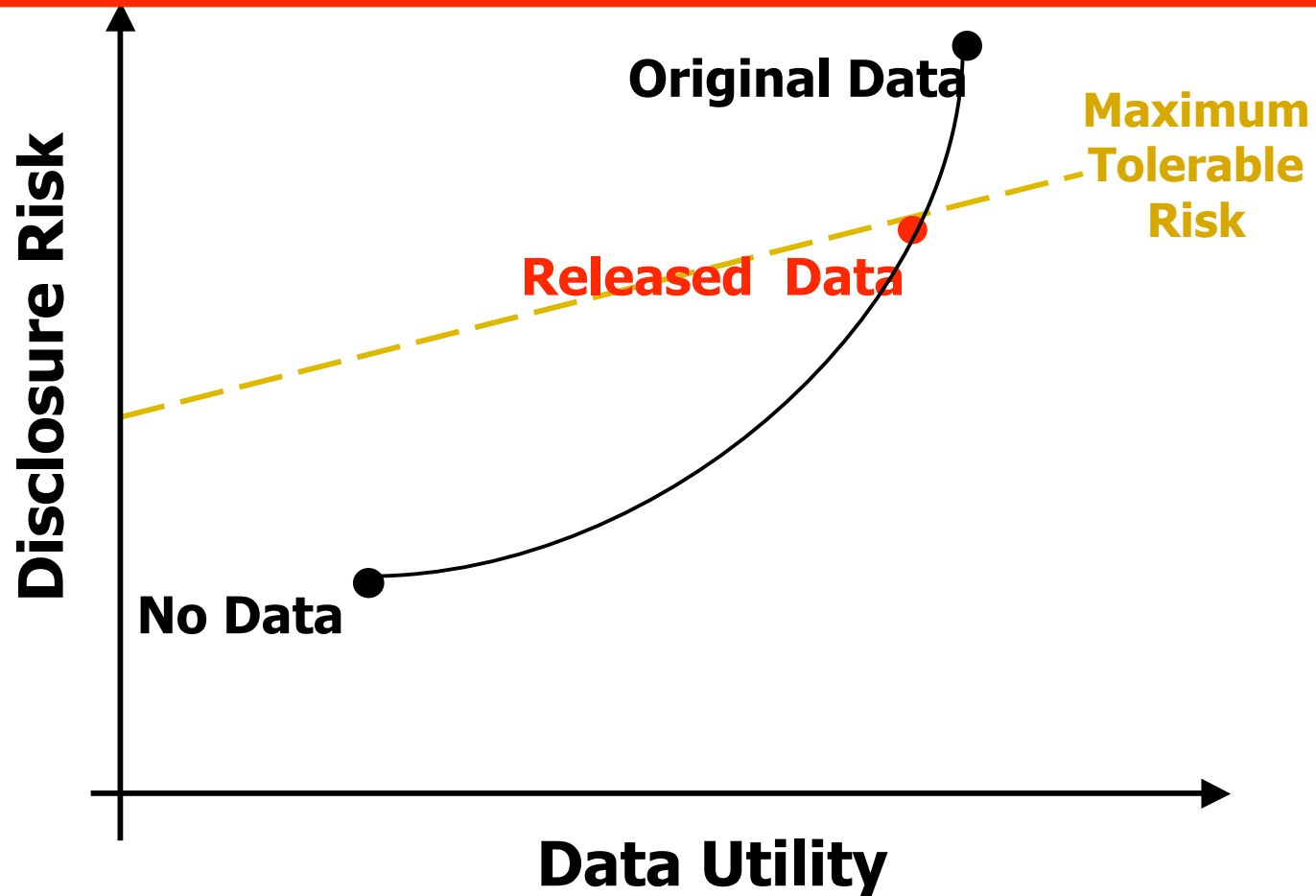
- Simplistic strategy advocated as part of TIA initiative in part based on PPDM.
- Mining for **groups** and for **individuals**.



Datamining Questions

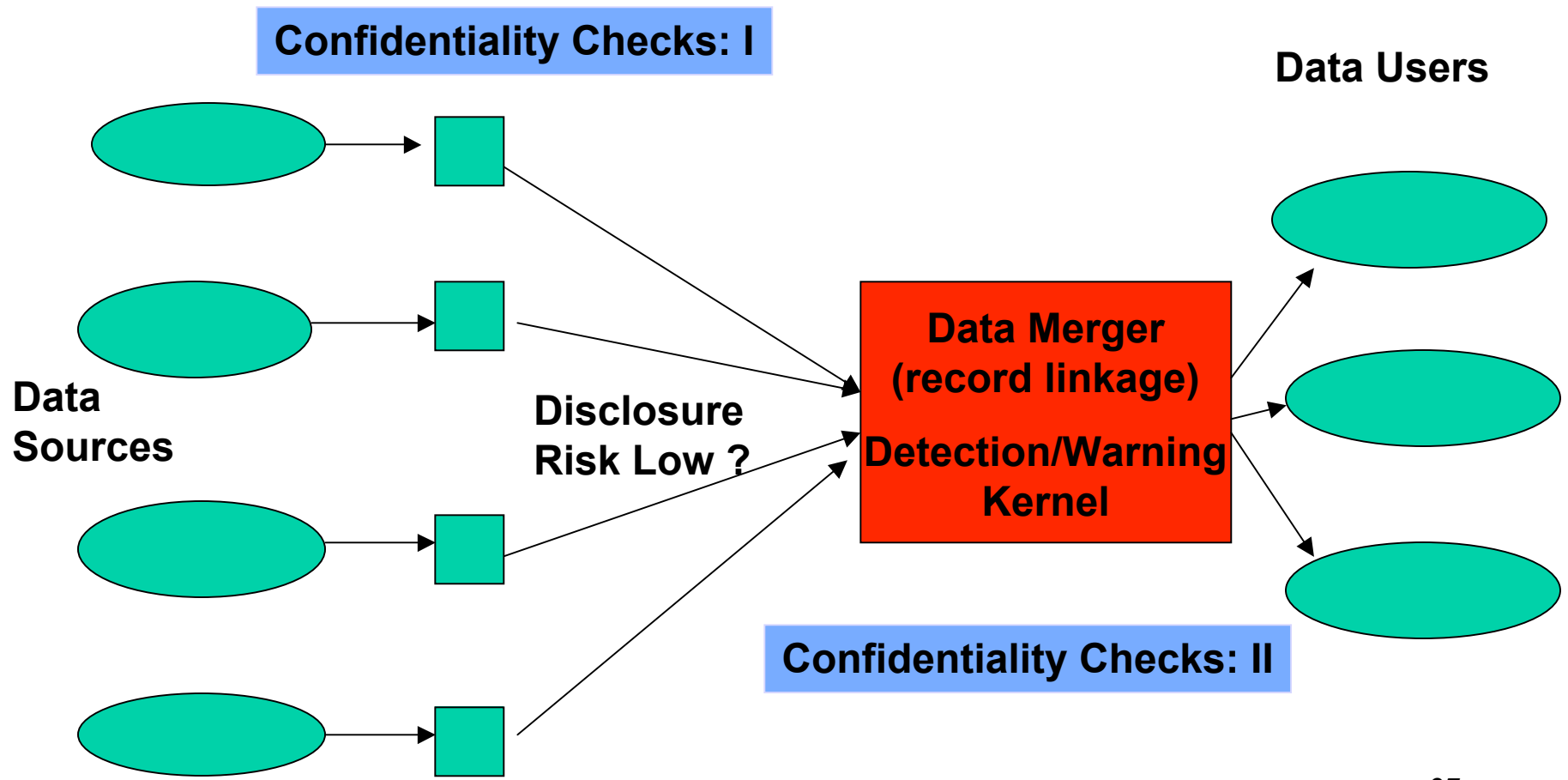
- **Individuals:**
 - What data are available on **David Madigan** and **Fred Roberts**?
 - Need names and/or related identifiers.
- **Groups:**
 - Can we build a good **logistic regression** model for predicting who is a *potential* terrorist?

R-U Confidentiality Map



(Duncan, et al. 2001, 2004)

Conceptual Statistical Kernel



What to Report?

- **All released information is potentially useful to intruder, but some is essential to data analysts.**
 - ***Kerckhoffs' Principle:*** Encryption systems should not depend on secrecy; they should not be compromised even if intruder knows exactly how they work (except for the key).
 - ***Statistical Principle:*** Report at least as much information as analyst needs to do proper inferences—including information on transformations and methods for data disclosure limitation.

Margins and Multiway Contingency Tables

- **Releasing marginal tables allows fitting of log-linear models:**
 - Finding “safe” sets of margins to release in sense that we can use them to “tightly” bound small cell values, or even infer these values with high probability. **Fienberg, Dobra, Slavkovic**
 - **Offers counterexamples to security associated with selective revelation as strategy.**
 - Use R-U utility ideas to assess trade-offs.
- **This is in effect what we have described as splitting databases!**

Two-Way Fréchet Bounds

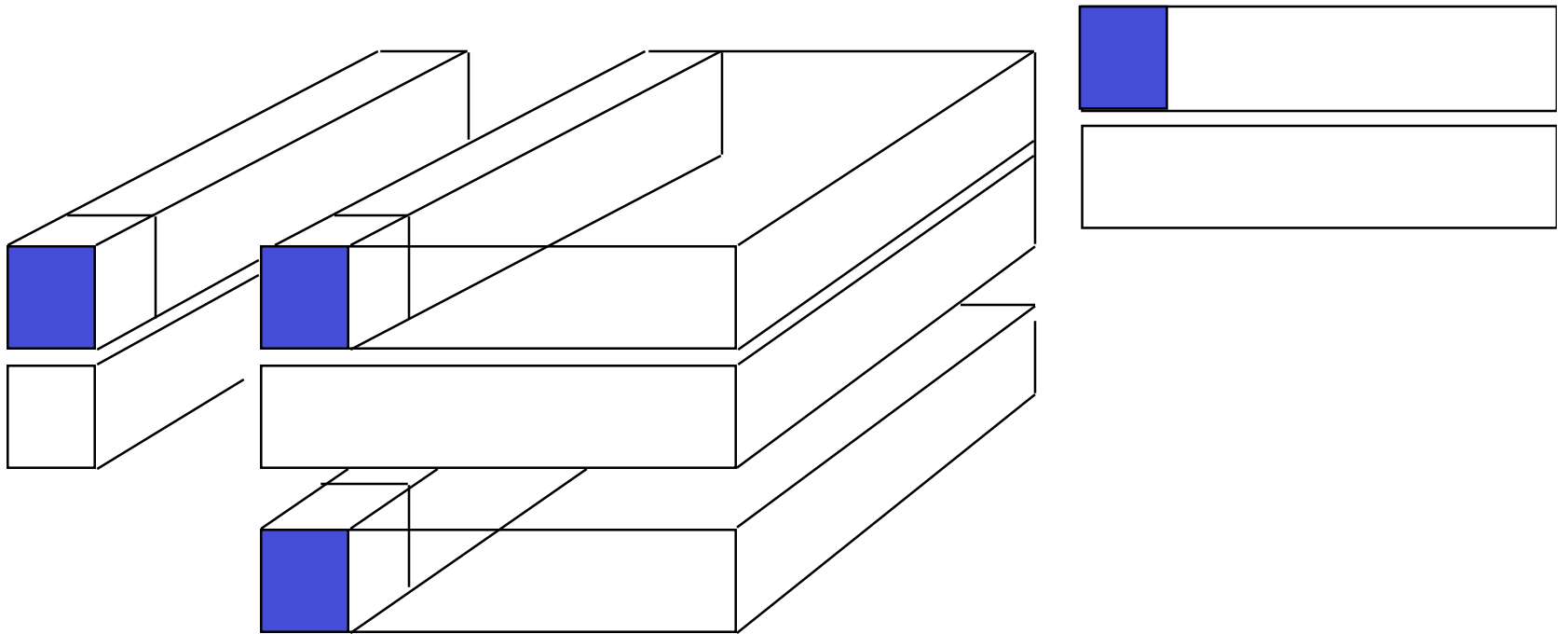
- For 2×2 tables of counts $\{n_{ij}\}$ given the marginal totals $\{n_{1+}, n_{2+}\}$ and $\{n_{+1}, n_{+2}\}$:

$$\begin{array}{cc|c} n_{11} & n_{12} & n_{1+} \\ n_{21} & n_{22} & n_{2+} \\ \hline n_{+1} & n_{+2} & n \end{array}$$

$$\min(n_{i+}, n_{+j}) \geq n_{ij} \geq \max(n_{i+} + n_{+j} - n, 0)$$

- Interested in multi-way generalizations involving higher-order, overlapping margins.

Bounds for Tables With Non-Negative Entries: Three-Dimensional Illustration



Methods for scaling up approach for large $k > 3$.

Example: Clinical Trial for Analgesic Drug Effectiveness

Center	Status	Treatment	Response		
			Poor	Moderate	Excellent
1	1	Active	3	20	5
		Placebo	11	14	8
	2	Active	3	14	12
		Placebo	6	13	5
2	1	Active	12	12	0
		Placebo	11	10	0
	2	Active	3	9	4
		Placebo	6	9	3

Sources: Koch et al. (1983); Fienberg and Slavkovic (Chance, 2004)

Statistical Analysis

- Interested in effect of T on R!!

- Possible margins release:

[CST][CSR][CTR] [CST][CSR][TR] [CST][CSR]

Releasing more margins produces much tighter bounds.

- Log-linear model analysis reveals:

- “Good Model”: [CST][CSR]

- Target model of Inference: [CST][CSR][TR]

- $\Delta G^2 = 5.4$ with 2 d.f.



Conditional Inference and Counting Tables

- We can use new tools from algebraic geometry to count how many tables there are having the observed margins.

LatTE

Center	Status	Treatment	Response		
			Poor	Moderate	Excellent
1	1	Active	3	20	5
		Placebo	11	14	8
	2	Active	3	14	12
		Placebo	6	13	5
2	1	Active	12	12	0
		Placebo	11	10	0
	2	Active	3	9	4
		Placebo	6	9	3

- **[CST][CSR]**
65,419,200 tables
- **[CST][CSR][TR]**
108,490 tables
- **[CST][CSR][CTR]**
980 tables

Bounds For Released Margins: [CST][CSR][CTR]

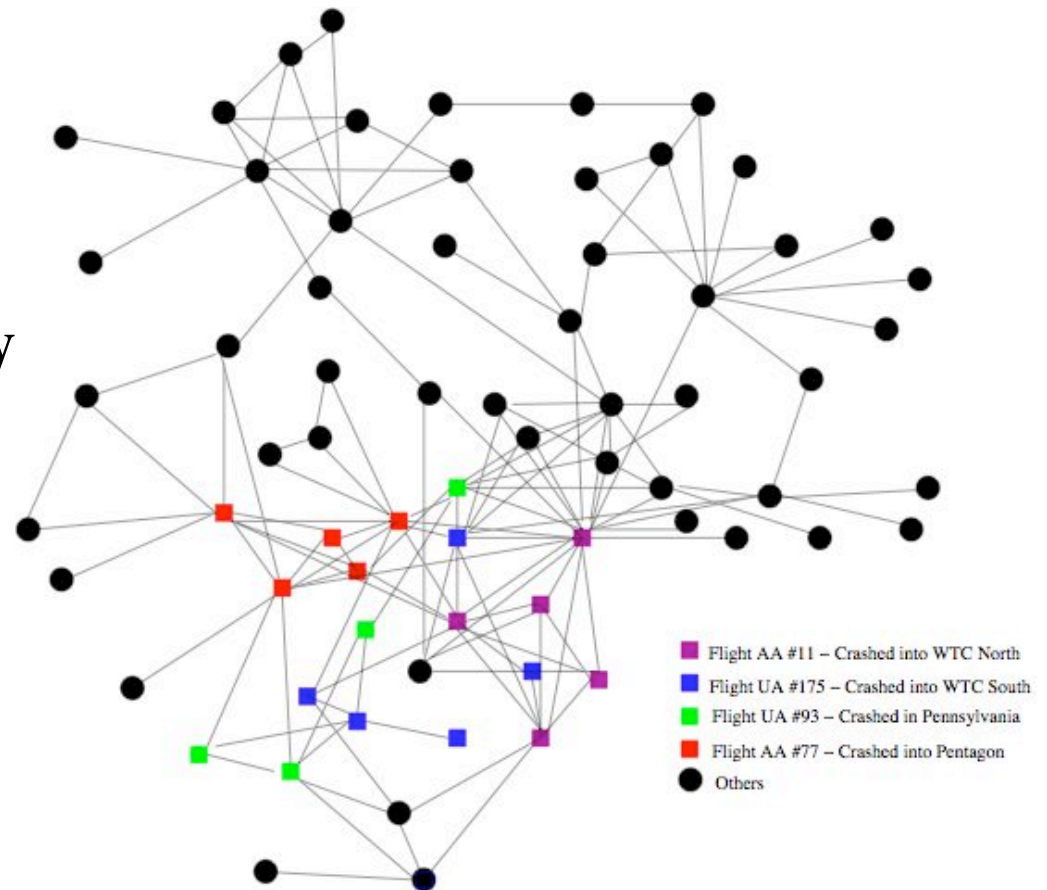
Center	Status	Treatment	Response		
			Poor	Moderate	Excellent
1	1	Active	[0,6]	[9,28]	[0,13]
		Placebo	[8,14]	[6,25]	[0,13]
	2	Active	[0,6]	[6,25]	[4,17]
		Placebo	[3,9]	[2,21]	[0,13]
2	1	Active	[6,15]	[9,18]	[0,0]
		Placebo	[8,17]	[4,13]	[0,0]
	2	Active	[0,9]	[3,12]	[4,4]
		Placebo	[0,9]	[6,15]	[3,3]

Association Rules and Multiway Contingency Tables

- **Association rules are conditional tables; we need margins for their assessment.**
 - **Fienberg and Slavkovic (2005).** *Datamining and Knowledge Discovery.*
 - **Extension to privacy-preserving association rule mining, and for logistic regression.**
 - **Fienberg, Fulp, Slavkovic, Wroebel (2006)**
- **We've come full circle to PPDM issues.**

Privacy and Network Data

- **Can we use network data to catch terrorists?**
 - With what data?
 - How many layers?
- **Searching for new terrorist cells?**
 - What signature?
- **Protecting non-terrorists?**



Privacy Protection for Network Data?

- **Anonymization doesn't work!**
 - Backstrom, Dwork, Kleinberg (forthcoming)

Current Status of Disclosure Protection Methodology

- **Major advances in last decade, many of which are rooted in real statistical theory.**
 - **Serious lag in their implementation.**
- **Matching and record linkage remain serious threat to confidentiality and privacy.**
- **Best statistically-grounded methods, even if they scaled to large datasets, still cannot offer protection demanded by some laws.**

Combining Approaches?

- **Can we combine PPDM, encryption, etc., with statistical approaches to disclosure limitation to address the protection problem?**
- **Need to think of new approaches to confidentiality and privacy, in layered fashion to provide maximal accuracy with reasonable protection:**
 - **Access to sensitive information varies by level and “type of access permitted.”**
 - **Restrictions on use vary by level.**
 - **Audit trails.**

Conclusions I

- **Security breaches in systems operated by Acxiom, ChoicePoint, and LexusNexus have filled newspapers.**
- **Lessons from such privacy breaches extend easily to virtually all electronically accessible databases, especially those that are part of proposed homeland security systems.**
 - **Violations may be just tip of “privacy-violation” iceberg.**
 - **Calls for government intervention and legal restrictions with respect to data warehousing and datamining.**

Conclusions II

- **Data warehouses (e.g., for homeland security) have been assembled through aggregation of information from separate data bases and transactional data systems.**
 - **They depend heavily on matching and record linkage methods that intrinsically statistical in nature, and whose accuracy deteriorates rapidly in presence of serious measurement error.**
 - **Datamining tools can't make up for bad data and poor matches.**

Conclusions III

- **We need new computational and statistical technologies to protect linked multiple data bases from privacy protection in face of commercial and government queries.**
 - **Slogans like “selective revelation” are not enough without technical backup.**
 - **Real promise in integration of research ideas emanating from statistical disclosure and cryptography communities.**
 - **Technologies that result from such collaborative research must be part of public domain, because only then can we evaluate their adequacy.**

The End

- **Many related papers are available online at:**
<http://www.stat.cmu.edu/~fienberg/DLindex.html>
<http://www.niss.org/dgii/index.html>
- **Record Linkage programs from William Winkler (US Census Bureau)**
- **SecondString: Open-source Java toolkit of approximate string-matching methods for matching names and records:**
<http://secondstring.sourceforge.net/>

Current Situation

- **Data on individuals and enterprises are widely available from electronic databases.**
- **In U.S. at least, public cannot distinguish among data from:**
 - **ChoicePoint, etc.**
 - **U.S. Homeland Security efforts.**
 - **Confidential statistical databases (e.g., from Census Bureau, NCHS, etc.).**
- **Governments can rescind confidentiality provisions of agencies, e.g., U.S. Patriot Act.**
- **Privacy and hunt for terrorists are in conflict!** 45