COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Data Matching

ISTA 488: Information Privacy with Applications

David Sidi (dsidi@email.arizona.edu)

Small mention of interesting things

- Carpenter v. US
- Send me proof that you've posted your key
- Final remarks
- Best module / topic?
- Worst module / topic?
- TCEs (at the end)

Warm-up

- None; TCE’s instead

Personally-identifiable information

- Information that can be used to pick out a specific person is personally-identifiable
- Two ways it can “pick out”
  - Direct identifiers
  - A combination of pseudoidentifiers

Direct Identifiers

- Single piece information that appears only in conjunction with other information if that information applies to a specific person
  - Social security number / Driver's license numbers / National ID numbers
- 'Piece' is vague in this usage

COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Hilda Schrader Whitcher is an employee at Woolworth.

075-05-1120 is an employee at Woolworth.

(we are all Hilda now)

A word of caution

- What counts as a direct identifier is not apparent from looking at it in isolation; sometimes intuitions are wrong
- Birthday paradox: when combinatorics attacks
- This American Life “Things I mean to know” @12:15 - 17:45

Pseudoidentifiers

- Sets of attribute that, in combination, function like a direct identifier
- example?

- Occupation: Professional Basketball player (NBA)
- Height:
  - Muggsy Bogues: 5’3”
  - Manute Bol: 7’7”
- Occupation, year active, height is an indirect identifier for these two

Newcombe's version of Muggsy

- In English speaking countries, which of these has more distinguishing power?
  - Zbigniew Zabrinsky
  - John Smith
- Newcombe used odds ratios, with cutoffs used to indicate links
- Fellegi and Sunter: showed optimality under fixed upper bounds on the false link (match) rates and the false non-link (non-match) rates

See Peter Christen, **Data Matching**

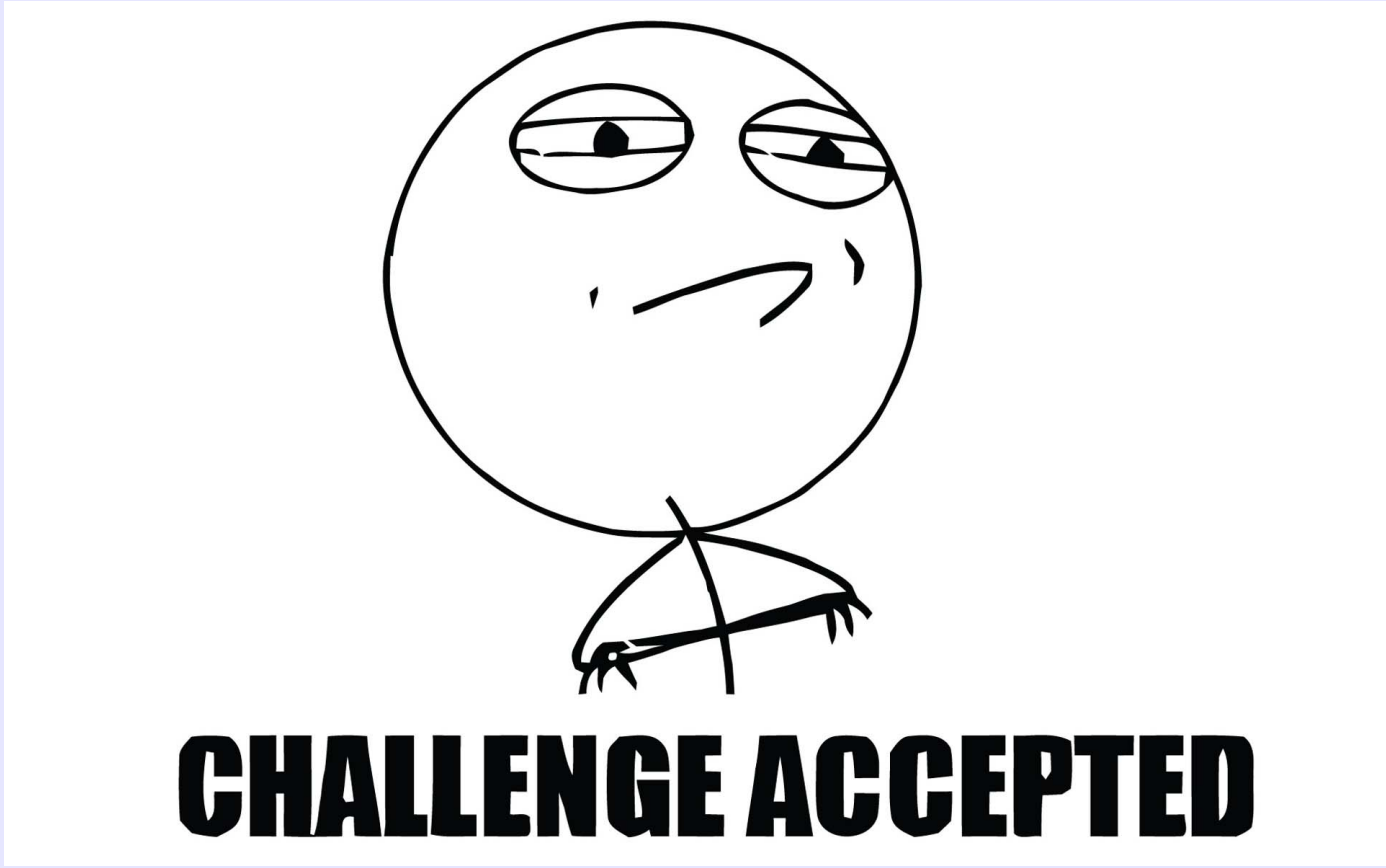
Pseudoidentifiers

- Sets of attribute that, in combination, function like a direct identifier
- Varies depending on attribute values!
- More nontrivial attributes means more likely that they comprise a pseudoidentifier
  - why is that significant?

Another word of caution

- Just as what is a direct identifier can be unintuitive, what is not an indirect identifier can be unintuitive
- Example?

- Pre-HIPAA
- “De-identified” hospital records
- Attributes included ZIP, DOB, gender



Latanya Sweeney, seeing the word “de-identified” (dramatic reenactment)

https://www.youtube.com/watch?v=Pn4p4VgSyCs
@ 3:00 - 5:00

Governor William Weld, after his data was identified (dramatic reenactment)

Another word of caution

- Just as what is a direct identifier can be unintuitive, what is not an indirect identifier can be unintuitive
- Example?
- Moral: “de-identification” by removing direct identifiers is not enough; Pseudoidentifiers can still connect data to an individual
  - HIPAA is representative: not data that can be used alone or in combination with other information to identify an individual subject.

See Bambauer, ‘Tragedy of the Data Commons’, 23.

Pseudoidentifiers

- Sets of attribute that, in combination, function like a direct identifier
- Varies depending on attribute values!
- More nontrivial attributes means more likely that they comprise a pseudoidentifier
  - why is that significant?
- How do we build pseudoidentifiers? Linkage

- Pseudoidentifiers are build by combining attributes
  - example for web browsing?

List of fingerprinting sources

1. UserAgent
2. Language
3. Color Depth
4. Screen Resolution
5. Timezone
6. Has session storage or not
7. Has local storage or not
8. Has indexed DB
9. Has IE specific 'AddBehavior'
10. Has open DB
11. CPU class
12. Platform
13. DoNotTrack or not
14. Full list of installed fonts (maintaining their order, which increases the entropy), implemented with Flash.
15. A list of installed fonts, detected with JS/CSS (side-channel technique) - can detect up to 500 installed fonts without flash
16. Canvas fingerprinting
17. WebGL fingerprinting
18. Plugins (IE included)
19. Is AdBlock installed or not
20. Has the user tampered with its languages [1]
21. Has the user tampered with its screen resolution [1]
22. Has the user tampered with its OS [1]
23. Has the user tampered with its browser [1]
24. Touch screen detection and capabilities
25. Pixel Ratio
26. System's total number of logical processors available to the user agent.
27. Device memory

- Pseudoidentifiers are build by combining attributes
- As a matter of US law, only records that are **public** are part of the body of information that can be used to assess “personally-identifiable information”
- But all bets are off if you’re trying to deanonymize using data matching: information gathering can be cumulative, building on itself
- Let’s look at how data integration works in more detail

Data Matching in context

- (Information extraction)
- Schema matching
- Data matching
- Data fusion

Data Integration

Data Matching is for structured data

- Data is in database tables, with records over several attributes
- This may require information extraction from unstructured data (raw text, images of text, ...)

attributes

records

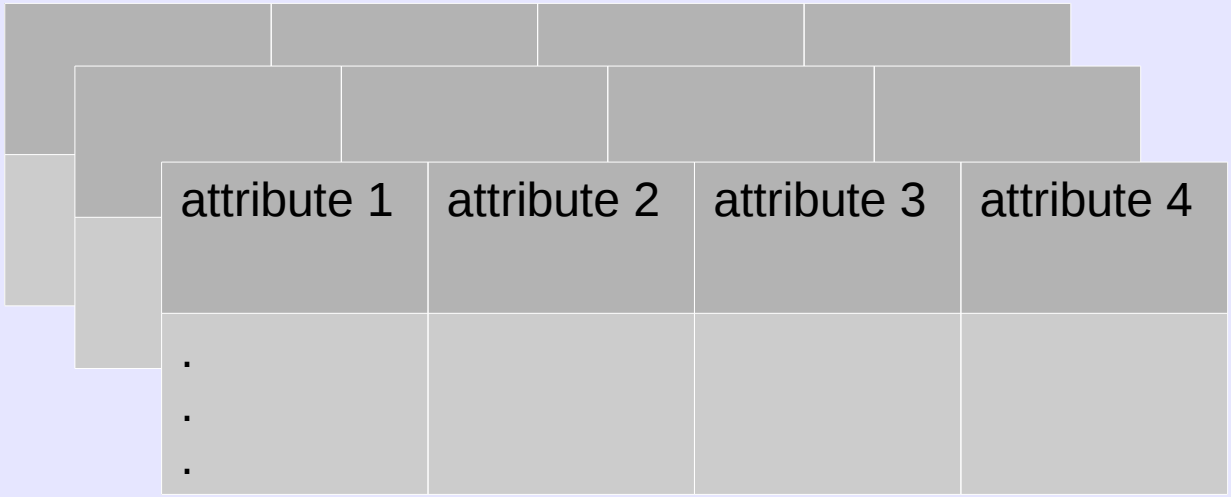
	Name	Address
1	John Doe	5 Main St., USA
2	Joe Bloggs	5 Brick Ln., UK

Records depend on what's useful

- Records can be
  - people (customers, taxpayers, criminals, travelers, ...)
  - corporations
  - web pages
  - web searches
  - bibliographic records
  - publications
  - ...

Setting

- N databases d_1, d_2, ... , d_N
- Each database d_i has r_i records



Aim of data matching

- Identify and match individual records that describe the same entities

PatTbl

PatientID	Name	DOB	Age	Gender	StreetAddress	Suburb	Postcode
P1273489	John Smith	8/10/1960	51	M	8/42 Miller Street	Melbourne	3011
Q6549234-2	Mick Meyer	30/01/1948	63	M	10 Port Road	Ferny Grove	7004
P7693427-8	Joanna Smith	12/11/1984	27	F	76 George Crest	Sydeny	2020

AdmittedPatients

PID	Surname	GivenName	BirthDate	Sex	AID
25198	Smith	Jo Anna	19841112	1	A347
55642	Smith	John W.	19601008	0	A135
15907	Meier	Michael	19480101	0	A810
99801	Meyer	Mike	19790320	0	A135

Addresses

AID	Street	Location
A135	42 Miller St	3000 Melbourne
A347	16 George Crs	2000 Sydney
A810	PO Box 553	7000 Brisbane

Aim of data matching

- Identify and match individual records that describe the same entities

PatTbl

PatientID	Name	DOB	Age	Gender	StreetAddress	Suburb	Postcode
P1273489	John Smith	8/10/1960	51	M	8/42 Miller Street	Melbourne	3011
Q6549234-2	Mick Meyer	30/01/1948	63	M	10 Port Road	Ferny Grove	7004
P7693427-8	Joanna Smith	12/11/1984	27	F	76 George Crest	Sydeny	2020

AdmittedPatients

PID	Surname	GivenName	BirthDate	Sex	AID
25198	Smith	Jo Anna	19841112	1	A347
55642	Smith	John W.	19601008	0	A135
15907	Meier	Michael	19480101	0	A810
99801	Meyer	Mike	19790320	0	A135

Addresses

AID	Street	Location
A135	42 Miller St	3000 Melbourne
A347	16 George Crs	2000 Sydney
A810	PO Box 553	7000 Brisbane

Problem: How do we know if we’re right?

Simplest nontrivial matching task

- Two databases are accurate, complete, unchanging, robust, and consistent over time
- The same unique entity identifiers are used in each of them
- Matching reduces to a database join

Simplest nontrivial matching task

- Two databases are accurate, complete, unchanging, robust, and consistent over time
- ~~The same unique entity identifiers are used in each of them~~
- Must use shared attributes

Simplest nontrivial matching task

- ~~Two databases are accurate, complete, unchanging, robust, and consistent over time~~
- ~~The same unique entity identifiers are used in each of them~~
- Must use shared attributes
- Must handle “dirty data”

Definitions (from Torra)

- Identity: for a given record R, a set of attribute values that reduce the anonymity set for R to contain only a single element.

Definitions (from Torra)

- Identity: for a given record R, a set of attribute values that reduce the anonymity set for R to contain only a single element.
- Height: 7’7”

  Weight: 200 lbs

  Nationality: Sudanese

- Every entity (record in a DB) can have several identities in this sense
- Of course, any superset of an identity is an identity

COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

TCE’s / Thank you all