COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 19: Public Data Release and Individual Anonymity

Plan today

- Public release of wrangled data – anonymity issues and pitfalls
  - How can it be maintained?
  - Case study of location and trajectory datasets

The problem

- ***The public is concerned that computer scientists can purportedly identify individuals hidden in anonymized data with "astonishing ease."***

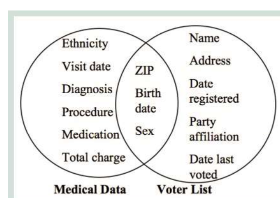
https://fpf.org/wp-content/uploads/The-Re-identification-of-Governor-Welds-Medical-Information-Daniel-Barth-Jones.pdf

Example 1: Massachusetts mid 1990s

- Mid 1990s: Massachusetts Group Insurance Commission releases records about history of hospital visits of State employees
  – Governor of Massachusetts assured public that personal identifiers had been deleted
    - name, address, social security number deleted
    - *Zip code (post code), birth date, sex retained*

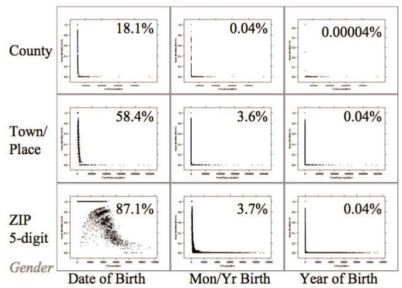
https://upload.wikimedia.org/wikipedia/commons/3/39/WilliamWeld.jpg

- 1997: Latanya Sweeney, a PhD student, went searching for the Governor's records in this dataset
  - Purchased voter rolls of city where he lived
    - Name, address, postcode, birth date, sex in rolls
    - Only 6 people had same birth date as Governor
    - Only 3 were men
    - Of these, only one lived in his zipcode …….



Example 2: Census Data

- Sweeney continued her research in privacy:
  - https://www.youtube.com/watch?v=tivCK_fBBfo



- She did a study of records from the1990 USA census, concluding that
  - 87% of Americans uniquely identified by zip code, birth date and sex
  - 53% of Americans uniquely identified by city, birth date and sex
  - Led to changes in privacy legislation in the USA
  - http://latanyasweeney.org/work/identifiability.html

- Australia
  - Privacy Act 1988, census data
    - *http://www.abs.gov.au/websitedbs/censushome.nsf/home/privacy*

THE UNIVERSITY OF MELBOURNE

Example 3: Netflix Dataset

- 2006: Netflix publicly releases 6 years of data about its customers viewing habits
  - Cinematch is the bit of software embedded in the Netflix Web site that analyzes each customer's movie-viewing habits and recommends other movies that the customer might enjoy.
  - https://www.nytimes.com/2008/11/21/technology/21iht-23netflixt.18049332.html
  - An anonymous id is created for each user
  - Sampled 10% of their data
  - Slight data perturbation
- Aim: Help to build better collaborative filtering algorithms (10% improve to cinematch.
  - 1 million dollar prize for a model

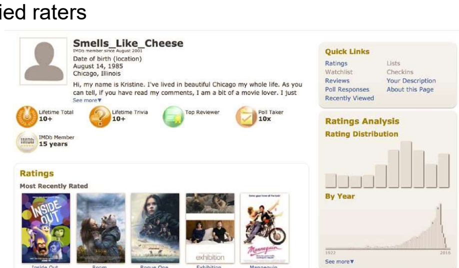
Anonymous ID	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma	….
A1	3	2	-	-	-	1	-	
A2	-	-	1	2	-	-	-	
A3	1	-	-	3	2	1	-	

THE UNIVERSITY OF MELBOURNE

Linking Netflix data with IMDb public data

- Two researchers, Narayanan and Shmatikov:
  - https://arxiv.org/pdf/cs/0610105v2.pdf

- Given knowledge about a person's "public" movie habits on IMDb, showed it was possible uncover their "private" movie habits in the Netflix dataset
  - 8 movie ratings (≤ 2 wrong ratings, dates ± 2 weeks):
    - 99% re-identified raters



THE UNIVERSITY OF MELBOURNE

Measures of anonymity for individuals

- Removing explicit identifiers from a dataset is not enough
- Solutions
  - *k-anonymity*
  - *l-diversity*

- Terminology
  - Explicit identifier: Unique for an individual
    - name, national ID, TFN, account numbers
  - Quasi –identifier: A combination of non sensitive attributes that can be linked with external data to identify an individual
    - E.g {Gender, Age, Zip code} combination from earlier
  - Sensitive attribute(s)
    - Information that people don't wish to reveal (e.g. medical condition)

THE UNIVERSITY OF MELBOURNE

Problem: If the data gets into the wrong hands

- If I know target is a 35 year old American living in zip 13068
  - Can infer they have cancer
- If I know target is a 28 year old Russian living in zip 13053
  - Can infer they have heart disease

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

THE UNIVERSITY OF MELBOURNE

k-anonymity

- *``Produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful.''*
- A table satisfies *k*-anonymity if every record in the table is indistinguishable from at least *k* − 1 other records with respect to every set of quasi-identifier attributes; such a table is called a *k-anonymous* table.
- Hence, for every combination of values of the quasi-identifiers in the *k*-anonymous table, there are at least *k* records that share those values.

THE UNIVERSITY OF MELBOURNE

k-anonymity example

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

THE UNIVERSITY OF MELBOURNE

What level of *k*-anonymity is satisfied here?

- Sensitive attribute: COMP20008 Grade
- Quasi identifier: {Gender, Age, Hair Colour}

Student Name	Gender	Age	Hair Colour	COMP20008 Grade
7930c	Male	20	Brown	78
1a985	Male	20	Brown	88
04ed9	Female	19	Red	75
82260	Female	19	Red	85
e461e	Female	19	Red	80
1e609	Female	21	Brown	80

k=1, 2, 3 or 4?

Another example: What level of anonymity?

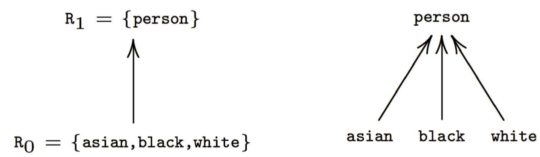
- Sensitive attribute: Problem
- Quasi identifier: {Race, Birth, Gender, ZIP}

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

https://epic.org/privacy/reidentification/Sweeney_Article.pdf

How to Achieve *k*-anonymity

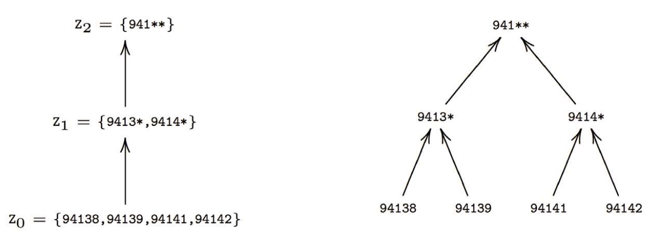
- Generalization
  - Make the quasi identifiers less specific
  - Column level
  - Example: race



http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf

How to Achieve *k*-anonymity- continued

- Generalization
  - Example: Zip code



  - When generalizing 94138 which one is a better strategy?
    - 9413*
    - *4138

http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf

How to Achieve *k*-anonymity- continued

- Suppression
  - Remove (suppress) the quasi identifiers completely
  - Moderate the generalization process
  - Limited number of outliers
  - Row, column and cell level
  - Example:
    - Removing the last two lines
    - Generalizing zip code to *941***
    - Generalizing race to *person*

Race:R_0	ZIP:Z_1
asian	9414*
asian	9414*
asian	9413*
asian	9413*
asian	9413*
black	9413*
black	9413*
white	*9413**
white	*9414**

http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf

k-anonymity: recap

- In the worst case, if data gets into the wrong hands, can only narrow down a quasi identifier to a group of *k* individuals
- Data publisher needs to
  - Determine quasi identifier(s)
  - Choose parameter *k*

Attack on *k*-anonymity I: Homogeneity attack

- k-anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.
  - Alice knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 9,10,11, or 12
    - Alice can conclude that Bob has cancer if she sees the data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	3*	$*$	Cancer
10	130**	3*	$*$	Cancer
11	130**	3*	$*$	Cancer
12	130**	3*	$*$	Cancer

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

Attack on *k*-anonymity II: Background attack

- *k-anonymity does not protect against attacks based on background knowledge.*
  - Alice knows that Umeko is a 21 year- old Japanese female who currently lives in zip code 13068.
  - She knows that that Umeko's information is contained in record number 1,2,3, or 4.
  - She concludes that Umeko has a viral infection, since Japanese have very low incidence of heart disease

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	3*	$*$	Cancer
10	130**	3*	$*$	Cancer
11	130**	3*	$*$	Cancer
12	130**	3*	$*$	Cancer

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

Solution: *l*-diversity

- Make the sensitive attribute diverse within each group

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	$*$	Heart Disease
4	1305*	≤ 40	$*$	Viral Infection
9	1305*	≤ 40	$*$	Cancer
10	1305*	≤ 40	$*$	Cancer
5	1485*	> 40	$*$	Cancer
6	1485*	> 40	$*$	Heart Disease
7	1485*	> 40	$*$	Viral Infection
8	1485*	> 40	$*$	Viral Infection
2	1306*	≤ 40	$*$	Heart Disease
3	1306*	≤ 40	$*$	Viral Infection
11	1306*	≤ 40	$*$	Cancer
12	1306*	≤ 40	$*$	Cancer

Figure 3. 3-Diverse Inpatient Microdata

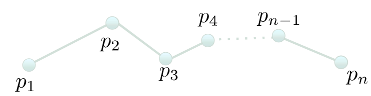
l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

Location & Trajectory Privacy

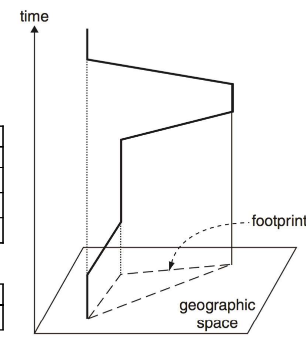
- What about datasets that record information about an individual in time and space?
- Location data being collected and stored throughout the day
  - GPS-enabled smart phones, cars, and wearable devices
  - Wi-Fi access points
  - Cell towers
  - Geo-tagged tweets, Facebook status, location check-ins …

Trajectory

- A function from time to geographical space



ID	GPS-Latitude	GPS-Longitude	Time
111478	33.692771	-111.993959	11:52
111478	33.692752	-111.993895	11:54
111478	33.692723	-111.993581	11:56
111478	33.692804	-111.993464	11:58
		⋮	
111478	33.69314	-111.993223	12:28
111478	33.69317	-111.993192	12:30

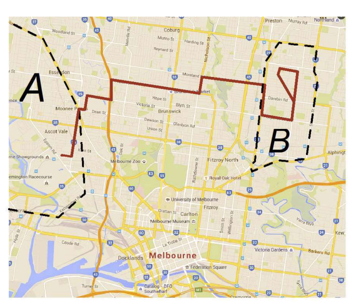


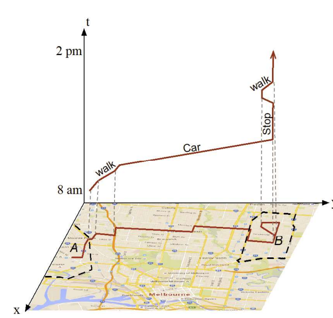
Location & Trajectory Data: Privacy Concerns

- Status quo of current mobile systems
  - Able to continuously monitor, communicate, and process information about a person's location
  - Have a high degree of spatial and temporal precision and accuracy
  - Might be linked with other data
- Analyzing and sharing location datasets has significant privacy implications
  - Personal safety, e.g., stalking, assault
  - Location-based profiling, e.g., Facebook
  - Intrusive inferences, e.g. individual's political views, personal preferences, health conditions

Inference Attacks - Example

- An user's Monday to Thursday trips
  - Home/work location pair may lead to a small set of potential individuals -> only {Bob, Alice} travel from A to B

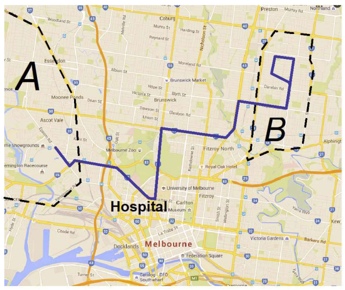


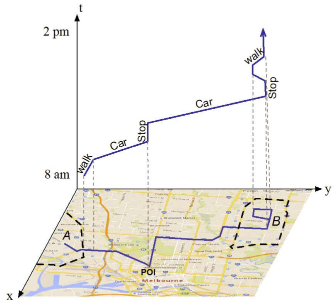


THE UNIVERSITY OF MELBOURNE

Inference Attacks - Example

- The same user's Friday trips
  - Regular visit to a heart hospital -> Alice is Japanese, so most probably the user is Bob

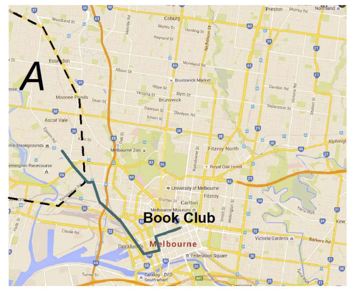


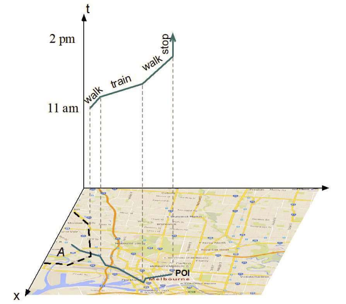


THE UNIVERSITY OF MELBOURNE

Inference Attacks - Example

- *Bob*'s Saturday trips
  - We can learn about his habits, preferences, etc.

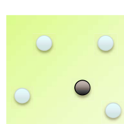




THE UNIVERSITY OF MELBOURNE

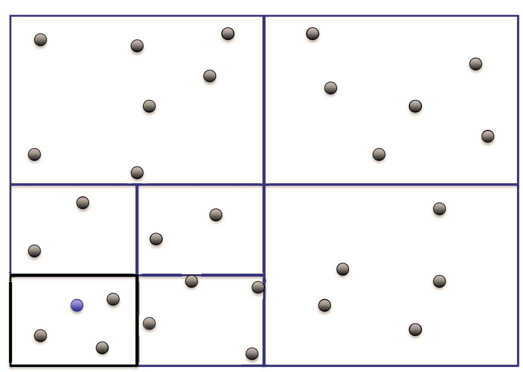
Anonymity: Cloaking

- *k*-anonymity
  - Individuals are *k*-anonymous if their location information cannot be distinguished from $k-1$ other individuals
- Spatial cloaking
  - Gruteser & Grunwald use quadtrees
  - Adapt the spatial precision of location information about a person according to the number of other people in the same quadrant
- Temporal cloaking
  - Reduce the frequency of temporal information



THE UNIVERSITY OF MELBOURNE

Spatial Cloaking (k_{min} = 4)



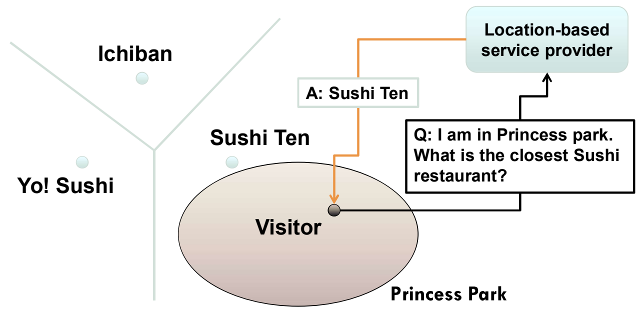
Obfuscation

- Idea
  - Mask an individual's precise location
  - Deliberately degrade the quality of information about an individual's location (imperfect information)
  - Identity can be revealed
- Assumption
  - Spatial imperfection ≈ privacy
  - The greater the imperfect knowledge about a user's location, the greater the user's privacy

Actual Location: (x,y)
Reported Location: Region

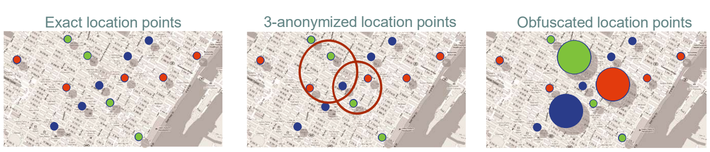
Motivation for Obfuscation

- Finding the closest Sushi restaurant

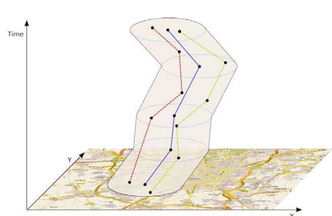


Overview of Privacy Models

- Location privacy vs. trajectory privacy



- Clustering k similar trajectories:
  - At each timestamp a point with the least distance to all trajectories is reported
- **Question:**
  - shortcomings of spatio-temporal cloaking obfuscation?



Summary

- To reduce risk of re-identification of individuals in released datasets
  - Choose value of *k*
  - Manipulate data to make it *k*-anonymous, either
    - Replace categories by broader categories
    - Suppress attributes with a * (limited utility)
  - Further manipulate data to make it *l*-diverse
    - Ensure there are at least *l* different values of the sensitive attribute in each group
- Privacy is difficult to maintain in high-dimensional datasets like trajectory datasets
  - Cloaking provides spatial *k*-anonymity
  - Obfuscation ensures location imprecision

THE UNIVERSITY OF MELBOURNE

Acknowledgements

This lecture was prepared using some material adapted from:

- Masachusette story
  - https://epic.org/privacy/reidentification/ohm_article.pdf
- From a social science perspective
  - http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
- *l*-diversity
  - https://www.cs.cornell.edu/~vmuthu/research/ldiversity.pdf