Privacy Preserving Data Publishing (PPDP)

From k-Anonymity to t-closeness

1

Content

- **Motivation** of privacy preserving data publishing
  - Background
  - Privacy attack instances
- Existing **solutions**
  - *k*-Anonymity
  - *L*-Diversity
  - T-Closeness
- Challenges and Emerging Applications
- **Conclusion**

2

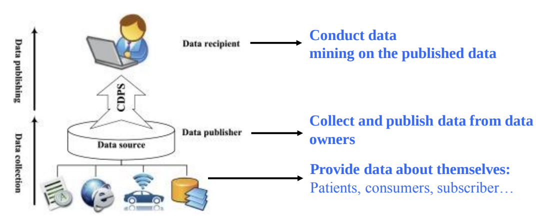
Content

- **Lots of data is being collected and warehoused**
  - Web data, e-commerce
  - purchases at department/grocery stores
  - Bank/Credit Card transactions
  - Social Network
  - Hospital

3

Motivation of PPDP

- Data collection and data publishing



4

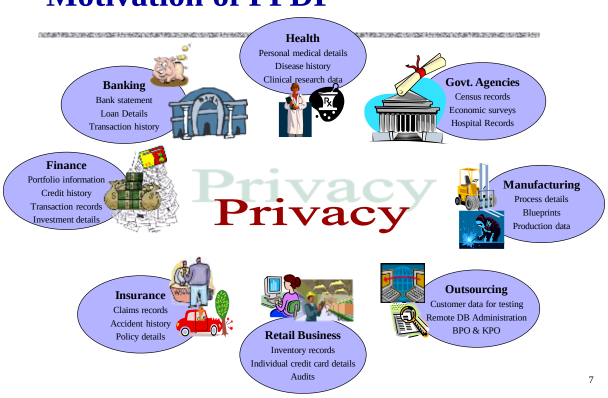
Motivation of PPDP

- There are two models of data holders:
  - In the untrusted model, the data holder is not trusted and may attempt to identify sensitive information from record owners.
  - In the trusted model, the data holder is trustworthy and record owners are willing to provide their personal information to the data holder;
    - however, the trust is not transitive to the data recipient.

Application Data Publish for Business

- Uncovering findings from data, help enable companies to **make smarter business decisions**:
  - Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.
  - Target identifies what are major customer segments within it's base and the unique shopping behaviors within those segments, which helps to guide messaging to different market audiences.
  - Proctor & Gamble utilizes time series models to more clearly understand future demand, which help plan for production levels more optimally.

Motivation of PPDP



Motivation of PPDP

- Objectives of PPDP
  - The privacy of the contributors are **protected**
  - The recipient gets **useful data**

Privacy VS Data Utility

Prohibit the disclosure or misuse of sensitive information about private individuals:
- SSN
- Disease
- …

Many types of research rely on the availability of private data:
- Demographic research
- Medical research
- Social network studies
- Web search studies

Government Regulations of Privacy

Country	Privacy Legislation
Australia	**Privacy Amendment Act of 2000**
European Union	**Personal Data Protection Directive 1998**
Hong Kong	**Personal Data (Privacy) Ordinance of 1995**
United Kingdom	**Data Protection Act of 1998**
United States	**Security Breach Information Act (S.B. 1386) of 2002** **Gramm-Leach-Bliley Act of 1999** **Health Insurance Portability and Accountability Act of 1996**

What About Privacy?

- **In PPDP,** the following three components need to be defined.
  - Sanitization mechanism: Given an original data set, a sanitization mechanism sanitizes the data set by making the data less precise. We call such a snapshot a release candidate.
  - Privacy criterion: Given a release candidate, the privacy criterion defines whether the release candidate is safe for release or not.
  - Utility metric: Given a release candidate, the utility metric quantifies the utility of the release candidate.

What About Privacy?

- **First thought:** anonymize the data
- **How?**
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address… what else?
  - Anything that identifies the person directly

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

Quasi-Identifiers

- Definition of Quasi-identifier:
  - A set of **non-sensitive** attributes $\{Q_1, Q_2, \ldots, Q_w\}$ of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population Ω.

Classification of Attributes

Sensitive attributes

- Medical records, salaries, etc.
- These attributes is what the researchers need, so they are always released directly

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

What About Privacy?

- **First thought:** anonymize the data
- **How?**
- Remove "personally identifying information" (PII)

- **Problem: Is this enough?**

Privacy Breach Example

Problem: Re-identification by Linking

Microdata: sensitive personal data held by an organization, e.g. medical records, transaction history. Often open to public access for reasons such as research.

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Privacy Breach Example

Latanya Sweeney's Attack (1997):

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	02142	9/15/61	female	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Lessons Learned

- **Any information released by the data curator can potentially be exploited by the adversary**
- Solution?
  - Publish a **modified** version of the data, such that:
    - the contributors' privacy is "adequately" protected
    - the published data is useful for its intended purpose(at least to some degree)
- Two issues:
  - **Privacy principle:** what do we mean by "adequately" protected privacy?
  - **Modification method:** how should we modify the data to ensure privacy while maximizing utility?

Lessons Learned

- **Any information released by the data curator can potentially be exploited by the adversary**
- Solution?
  - Publish a **modified** version of the data, such that:
    - the contributors' privacy is "adequately" protected
    - the published data is useful for its intended purpose(at least to some degree)
- Two issues:
  - **Privacy principle:** what do we mean by "adequately" protected privacy?
  - **Modification method:** how should we modify the data to ensure privacy while maximizing utility?

Existing solutions

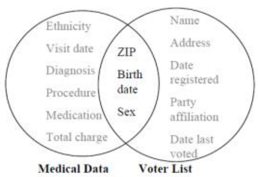
- Table 2.1 summarizes the attack models addressed by the privacy models.

Table 2.1: Privacy models

Privacy Model	Attack Model			
	Record linkage	Attribute linkage	Table linkage	Probabilistic attack
k-Anonymity [201, 217]	✓			
MultiR k-Anonymity [178]	✓			
ℓ-Diversity [162]	✓	✓		
Confidence Bounding [237]		✓		
(α, k)-Anonymity [246]	✓	✓		
(X, Y)-Privacy [236]	✓	✓		
(k, e)-Anonymity [269]		✓		
(ϵ, m)-Anonymity [152]		✓		
Personalized Privacy [250]		✓		
t-Closeness [153]		✓		✓
δ-Presence [176]			✓	
(c, t)-Isolation [46]	✓			✓
ϵ-Differential Privacy [74]			✓	✓
(d, γ)-Privacy [193]			✓	✓
Distributional Privacy [33]			✓	✓

Existing attacks

- Record linkage:
  - adversaries collect auxiliary information about a certain individual from multiple data sources and then combine that data to form a whole picture about their target, which is often an individual's personally identifiable information



Existing attacks

- Attribute linkage:
  - The adversary does not need to link an individual to a specific record, but can still determine the sensitive value associated with the individual.
    - For example: if Alice knows that Tom's record is: his zip code is 14852 and his age is 38, then without identifying which record is Tom's, Alice can still infer that Tom has Cancer.

TID	Zip code	Age	Condition
1	130***	< 30	Heart disease
2	130***	< 30	Viral infection
3	130***	< 30	Viral infection
4	148***	[30-40]	Cancer
5	148***	[30-40]	Cancer
6	148***	[30-40]	Cancer

Existing attacks

- Table linkage:
  - A table linkage occurs if an attacker can confidently infer the presence or the absence of the victim's record in the released table.

  - Suppose a hospital releases a data table with a particular type of disease.
  - Identifying the presence of the victim's record in the table is already damaging.

Existing attacks

- Example table linkage:
- Suppose the data publisher has released a 3-anonymous patient table (c)
- The attacker is presumed to also have access to an external public table (d)
- **Table (c) is in Table (d)**

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(d) 4-anonymous external table

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

- To launch a table linkage on a target victim, for Alice

The probability that Alice is present in (c) is 4/5=0.8
  - because there are 4 records in (c) and 5 records in (d) containing <Artist,Female,[30 − 35)>.

Existing attacks

- Example: Suppose the data publisher has released a 3-anonymous patient table (c)
- The attacker is presumed to also have access to an external public table (d)
- **Table (c) is in Table (d)**

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(d) 4-anonymous external table

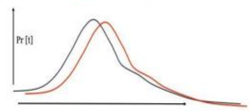
Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

Practice: what is the probability that Bob is present in (c)?

Existing attacks

- Probabilistic Attack:
  - Probabilistic attack is not like linkage attack which precisely knows individual information, then gain sensitive information combined with existed background knowledge,
  - but it focuses on changing adversary's probabilistic confidence of getting privacy information after acquiring published dataset.
  - Example: differential privacy

$$\frac{\Pr[\mathcal{K}(DB - Me) = t]}{\Pr[\mathcal{K}(DB + Me) = t]} \le e^{\varepsilon} \approx 1 \pm \varepsilon$$



Next

- Existing representative **solutions**
  - ***k*-Anonymity**
  - *L*-Diversity
  - T-Closeness

K-Anonymity: Intuition

- The information for each person contained in the released table **cannot be distinguished from at least k-1 individuals** whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

K-Anonymity Protection Model

- Private table: T
- Released table: RT
- Attributes: $A_1, A_2, \ldots, A_n$
- Quasi-identifier subset: $A_i, \ldots, A_j$

Let RT($A_1,\ldots,A_n$) be a table, $QI_{RT}=(A_i,\ldots, A_j)$ be the quasi-identifier associated with RT, $A_i,\ldots,A_j \subseteq A_1,\ldots,A_n$, and RT satisfy *k*-anonymity. Then, each sequence of values in RT[A_x] appears with at least *k* occurrences in RT[QI_{RT}] for $x=i,\ldots,j$.

Example of a k-Anonymous Table

	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

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Achieving k-Anonymity

- Suppression
- Generalization
- Swapping
- Randomization

Achieving k-Anonymity

- Suppression
  - In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'.

Example of Suppression (1)

Released table

	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

External data

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

Example of Suppression (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Suppressed table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

Achieving k-Anonymity

- Generalization
  - In this method, individual values of attributes are replaced by with a broader category.
  - For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' , etc.

Example of generalization

- Transform the QI values into less specific forms

Age	Zipcode	Disease
21	12000	dyspepsia
22	14000	bronchitis
24	18000	flu
23	25000	gastritis
41	20000	flu
36	27000	gastritis
37	33000	dyspepsia
40	35000	flu
43	26000	gastritis
52	33000	dyspepsia
56	34000	gastritis

Age	Zipcode	Disease
[21, 22]	[12k, 14k]	dyspepsia
[21, 22]	[12k, 14k]	bronchitis
[23, 24]	[18k, 25k]	flu
[23, 24]	[18k, 25k]	gastritis
[36, 41]	[20k, 27k]	flu
[36, 41]	[20k, 27k]	gastritis
[37, 43]	[26k, 35k]	dyspepsia
[37, 43]	[26k, 35k]	flu
[37, 43]	[26k, 35k]	gastritis
[52, 56]	[33k, 34k]	dyspepsia
[52, 56]	[33k, 34k]	gastritis

generalize

Achieving k-Anonymity

- Swapping
  - produces a release candidate by swapping some attribute values.
  - For example, the data publisher may swap the age values of Ann and Eshwar, swap the gender values of Bruce and Cary, and so on.
  - remove the relationship between quasi identifiers and sensitive data

Example of Swapping

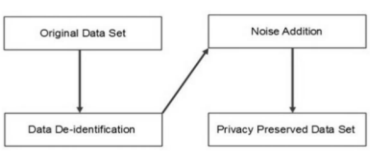
Age	Zipcode	Disease
[21, 22]	[12k, 14k]	dyspepsia
[21, 22]	[18k, 25k]	bronchitis
[23, 24]	[12k, 14k]	flu
[23, 24]	[18k, 25k]	gastritis
[36, 41]	[20k, 27k]	flu
[36, 41]	[20k, 27k]	gastritis
[37, 43]	[26k, 35k]	dyspepsia
[37, 43]	[26k, 35k]	flu
[37, 43]	[26k, 35k]	gastritis
[52, 56]	[33k, 34k]	dyspepsia
[52, 56]	[33k, 34k]	gastritis

Age	Zipcode	Disease
[21, 22]	[12k, 14k]	dyspepsia
[21, 22]	[12k, 14k]	bronchitis
[23, 24]	[18k, 25k]	flu
[23, 24]	[18k, 25k]	gastritis
[36, 41]	[20k, 27k]	flu
[36, 41]	[20k, 27k]	gastritis
[37, 43]	[26k, 35k]	dyspepsia
[37, 43]	[26k, 35k]	flu
[37, 43]	[26k, 35k]	gastritis
[52, 56]	[33k, 34k]	dyspepsia
[52, 56]	[33k, 34k]	gastritis

Swapping

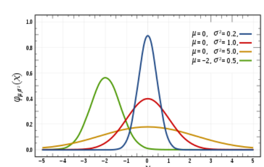
Achieving k-Anonymity

- Randomization
  - A release candidate of the randomization mechanism is generated by adding random noise to the data.
    - Gaussian noise



Example of Randomization

Age	Zipcode	Disease
21	12000	dyspepsia
22	14000	bronchitis
24	18000	flu
23	25000	gastritis
41	20000	flu
36	27000	gastritis
37	33000	dyspepsia
40	35000	flu
43	26000	gastritis
52	33000	dyspepsia
56	34000	gastritis



Example of Randomization

Age	Zipcode	Disease
21	12000	dyspepsia
22	14000	bronchitis
24	18000	flu
23	25000	gastritis
41	20000	flu
36	27000	gastritis
37	33000	dyspepsia
40	35000	flu
43	26000	gastritis
52	33000	dyspepsia
56	34000	gastritis

Age	Zipcode	Disease
22	13400	dyspepsia
23	13200	bronchitis
23	15650	flu
22	23200	gastritis
39	22300	flu
37	24400	gastritis
38	34400	dyspepsia
39	34500	flu
41	24500	gastritis
54	33500	dyspepsia
54	34600	gastritis

$\mu = 0, \sigma^2 = 2$ $\mu = 0, \sigma^2 = 2500$

Achieving k-Anonymity

- Randomization
  - An example

First Name	Last Name	SSN	Age	Major	GPA	Zip code	State	Scholarship Amount	Gender
John	Artist	Xxx-xx-xx9	23	Computer Science	4.00	21071	MA	30000.00	M
Peter	Chemist	Xxx-xx-xx10	33	Biology	3.35	31072	MD	50000.00	M
Evan	Biologist	Xxx-xx-xx11	32	Chemistry	2.19	21073	MA	25000.00	F
Joy	Music	Xxx-xx-xx12	45	History	2.99	21074	MD	67000.00	F
Eva	Pictures	Xxx-xx-xx13	23	Chemistry	3.67	21075	PA	78000.00	F
Sandra	Hollywood	Xxx-xx-xx14	21	Art	3.65	21076	MD	90888.00	F
Okello	Oscars	Xxx-xx-xx15	25	Music	4.00	21077	LA	90000.00	M
Mukisa	Grammys	Xxx-xx-xx16	30	Computer Science	2.79	21078	PA	10000.00	M
Jacinta	Historian	Xxx-xx-xx17	32	Biology	3.00	21079	CA	7000.00	F
Bosco	Activist	Xxx-xx-xx18	37	Art	2.98	21080	MD	11000.00	M

Table 1: *Original Data Set (All data for illustrative purposes).*

		Random Noise Between 1000 and 9000	
	Scholarship Amount	Random Noise	Perturbed Scholarship Amount
	30000.00	7296.88	37296.88
	50000.00	2087.4	52087.4
	25000.00	4403.32	29403.32
	67000.00	5389.89	72389.89
	78000.00	6639.16	84639.16
	90888.00	6228.52	97116.52
	90000.00	1554.68	91554.68
	10000.00	8977.3	18977.3
	7000.00	3455.81	10455.81
	11000.00	6932.86	17932.86
Mean	***20500***		***27614.87***
Standard Deviation	***13435.0288425444***		***13692.4298530319***
Variance	***180500000***		***187482635.2802***

Table 3: *Random noise between 1000 and 9000 added to Scholarship attribute.*

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Next

- Existing representative **solutions**
  - *k*-Anonymity
  - ***L*-Diversity**
  - T-Closeness

l-Diversity

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

L-Diversity

- T*: the Anonymized Table
- q*: the generalized value of q in the published table T*
- s: a possible value of the sensitive attribute
- n(q*,s'): number of tuples with sensitive attribute s' and non-sensitive attribute q*
- q*-block: the set of tuples in T* whose non-sensitive attribute values generalize to q*

L-Diversity

- Lack diversity: lack of diversity in the sensitive attribute manifests itself as follows:

$$\forall s' \neq s, \quad n_{(q^\star, s')} \ll n_{(q^\star, s)}$$

L-Diversity

- Then, **L-Diversity Principle** can be defined as:
  - A q*-block is L-diverse if contains at least L "well-represented" values for the sensitive attribute S.
  - A table is L-diverse if every q*-block is L-diverse.

An example

	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

4-anonymous table

	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

3 diverse table

- Using a 3-diverse table, we no longer are able to tell if Bob (a 31 year old American from zip code 13053) has cancer.
- We also cannot tell if Umeko(a 21 year old Japanese from zip code 13068) has a viral infection or cancer.

Probabilistic inference attacks over l-Diversity

- Each equivalence class has at least **l well-represented** sensitive values

...	Disease
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

10 records

8 records have HIV

2 records have other values

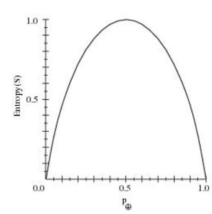
- Doesn't prevent probabilistic inference attacks
  - Infer: the patient has HIV with large possibility

Other Versions of L-Diversity

- Probabilistic L-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by 1/L

Other Versions of l-Diversity

- Suppose S is a collection of 14 examples of some Boolean concept, including 9 positive and 5 negative examples. Then the entropy of S relative to this Boolean classification is
  - Entropy([9+,5-])=-(9/14)log2(9/14)-(5/14)log2(5/14)=0.940



Other Versions of l-Diversity

- Entropy L-diversity
  - The entropy of the distribution of sensitive values in **each equivalence class** is at least log(L)

$$-\sum_{s \in S} p(q^*, s)\log(p(q^*, s')) \geq \log(l)$$

- $\frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}$ is the fraction of tuples in the q*-block with sensitive attribute value equal to s

- Problem of Entropy L-diversity
  - Here every q⋆-block has at least L distinct values for the sensitive attribute
  - This implies that for a table to be entropy L-Diverse, the entropy of the entire table must be at least log(L).
  - Therefore, entropy L-Diversity may be too restrictive to be practical.

Other Versions of L-Diversity

- Recursive (c,L)-diversity
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$ where r_i is the frequency of the i^{th} most frequent value
  - Intuition: the most frequent value does not appear too frequently

Limitations of L-Diversity

- L-diversity may be difficult and unnecessary to achieve.
  - A single sensitive attribute
    - Two values: HIV positive (1%) and HIV negative (99%)
    - Very different degrees of sensitivity
- L-diversity is unnecessary to achieve
  - 2-diversity is unnecessary for an equivalence class that contains only negative records
- L-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Sensitive Attribute Disclosure

L-diversity is insufficient to prevent attribute disclosure.

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

L-diversity does not consider semantics of sensitive values!

Next

- Existing representative **solutions**
  - *k*-Anonymity
  - *L*-Diversity
  - ***T*-Closeness**

Why t-Closeness?

- Pre-existing privacy measures k-anonymity and L-diversity have flaws.
  - k-anonymity-each equivalence class has at least k records to protect against identity disclosure.
    - k-anonymity is vulnerable to homogeneity attacks and background knowledge attacks.
  - L-diversity: distribution of a sensitive attribute in each equivalence class has at least L "well represented" values to protect against attribute disclosure.
    - L-diversity is vulnerable to skewness attacks and similarity attacks.

t-Closeness overview

- Privacy is measured by the information gain of an observer.
- We assume:
  - B0: Alice believes that Bob has the virus because he has been acting sick.
  - B1: Alice gets a summary report of the table and learns that only 1% of the population has the virus. This distribution is Q, the distribution of the sensitive attribute in the whole table. She believes that Bob is in that one percent.
  - B2: Alice takes a look at the table, and finds that Bob is in equivalence class 3 because he is 32 and lives in zip code 47623. She learns P, the distribution of the sensitive attribute values in this class. Based on P she decides that it is actually quite likely that Bob has the virus.

t-Closeness overview

- l-diversity limits the gain between B0 (belief before any knowledge of the table) and B2 (belief after examining the table and the relevant equivalence class) by requiring that P (distribution in the equivalence class) has diversity.
- Q (global distribution in the table) should be treated as public information.
- If the change from B0 to B1 is large, means that the Q contains lots of new information. But we can't control people's access to Q, so we shouldn't worry about it.
- Therefore should focusing on limiting the gain between B1 and B2. We can do so by limiting the difference between P and Q. The closer P and Q are, the closer B1 and B2 are.

t-Closeness definition

- An equivalence class is said to have **t-closeness**
  - if the distance between the distribution of a sensitive attribute (P) in this class and the distribution of the attribute in the whole table(Q) is no more than a threshold t.
  - A table is said to have t-closeness if all equivalence classes have t-closeness.

t-Closeness

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

Distance measurement

- Now that we’ve confirmed that limiting the difference between *P* and *Q* is the key to privacy, we need a way to measure the distance.
  - m: the number of sensitive values in an equivalence class
  - $P=(p_1,p_2,\ldots,p_m)$, $Q=(q_1,q_2,\ldots,q_m)$
- Here are some naive measurements:
  - Method 1: variational distance

$$D[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|.$$

Distance measurement

- Example

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

- Overall distribution of the Income attribute:
  Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

- The first equivalence class in Table 4 has distribution:
  P1 = {3k, 4k, 5k}

- The second equivalence class has distribution:
  P2 = {6k, 8k, 11k}

D(P1,Q)=0.5*(|1/3-1/9|+ |1/3-1/9| + |1/3-1/9| + |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|)=1/2
D(P2,Q)=0.5*(|1/3-1/9|+ |1/3-1/9| + |1/3-1/9|+|0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|)=1/2

We have D(P1,Q)= D(P2,Q)

Distance measurement

- Here are some naive measurements:
  - Method 2: Kullback-Leibler (KL) distance

$$D[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = H(\mathbf{P}) - H(\mathbf{P},\mathbf{Q})$$

	0	1
D1	0.1	0.9
D2	0.2	0.8
D3	0.9	0.1

- H(P) is the entropy of P

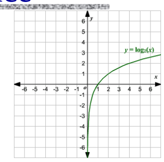
$$H(\mathbf{P}) = \sum_{i=1}^{m} p_i \log p_i$$

- H (P, Q) is the cross-entropy of P and Q

$$H(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{m} p_i \log q_i$$

Kullback-Leibler (KL) distance

$$H(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{m} p_i \log q_i$$



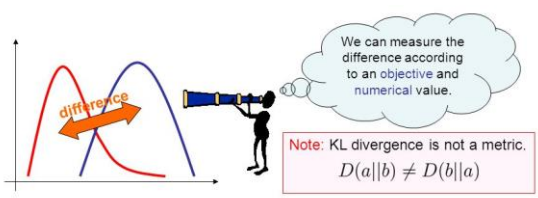
	0	1
D1	0.1	0.9
D2	0.2	0.8
D3	0.9	0.1

H(D1,D2)=0.1*ln(0.2)+0.9*ln(0.8)=-0.1609-0.2008=-0.3617

H(D1,D3)=0.1*ln(0.9)+0.9*ln(0.1)=-0.0105-2.0723=-2.0828

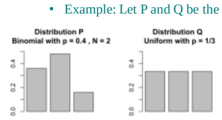
Kullback-Leibler (KL) distance

In the context of machine learning, D (P, Q) is often called the information gain achieved if Q is used instead of P.



Distance measurement

- Here are some naive measurements:
  - Method 2: Kullback-Leibler (KL) distance
    - Example: Let P and Q be the distributions shown in the table and figure



	0	1	2
Distribution P	0.36	0.48	0.16
Distribution Q	0.333	0.333	0.333

The KL divergence is calculated as follows. This example uses the natural log with base e, designated ln.

$$D_{\mathrm{KL}}(Q\|P) = \sum_i Q(i) \ln\left(\frac{Q(i)}{P(i)}\right)$$
$$= 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.333 \ln\left(\frac{0.333}{0.48}\right) + 0.333 \ln\left(\frac{0.333}{0.16}\right)$$
$$= -0.02596 + (-0.12176) + 0.24408$$
$$= 0.09637$$

Distance measurement

- However, these distance measures **do not reflect** the semantic distance among values.
  - Let's see an example

Distance measurement

See the example again

- Example

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

- Overall distribution of the Income attribute:
  Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

- The first equivalence class in Table 4 has distribution:
  P1 = {3k, 4k, 5k}

- The second equivalence class has distribution:
  P2 = {6k, 8k, 11k}

D(P1,Q)=0.5*(|1/3-1/9|+ |1/3-1/9| + |1/3-1/9| + |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|)=1/2
D(P2,Q)=0.5*(|1/3-1/9|+ |1/3-1/9| + |1/3-1/9|+|0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|+ |0-1/9|)=1/2

However, we we would like to have

D(P1,Q)>D(P2,Q)

Since: Our intuition is that P1 results in more information leakage than P2, because the values in P1 are all in the lower end.

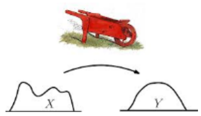
Distance measurement

- However, these distance measures **do not reflect** the semantic distance among values.
- The distance measures mentioned above would not be able to do so, because from their point of view values such as 3k and 6k are just different points and have no other semantic meaning.
- How to avoid it?
  - Earth Mover's distance is good!

Earth Mover's distance

- The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other.
  - Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space.
  - EMD measures the least amount of work needed to fill the holes with earth.
  - A unit of work corresponds to moving a unit of earth by a unit of ground distance.

Earth Mover's distance



- Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space.
- EMD measures the least amount of work needed to fill the holes with earth.
- A unit of work corresponds to moving a unit of earth by a unit of ground distance.

Earth Mover's distance

Definition of EMD:

- EMD can be formally defined using the well-studied transportation problem.
- $P=(p_1,p_2,\ldots,p_m)$, $Q=(q_1,q_2,\ldots,q_m)$

$$WORK(\mathbf{P},\mathbf{Q},F)=\sum_{i=1}^{m}\sum_{j=1}^{m}d_{ij}f_{ij}$$

subject to the following constraints:

$$f_{ij}\geq 0 \qquad 1\leq i\leq m, 1\leq j\leq m \qquad (c1)$$

$$p_i-\sum_{j=1}^{m}f_{ij}+\sum_{j=1}^{m}f_{ji}=q_i \qquad 1\leq i\leq m \qquad (c2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}f_{ij}=\sum_{i=1}^{m}p_i=\sum_{i=1}^{m}q_i=1 \qquad (c3)$$

Properties of EMD:

- d_{ij} is the ground distance between i in P and j in Q, which is defined as $|p_i-q_j|/(m-1)$
- f_{ij} is the flow of mass to transform i in P into j in Q using the minimal amount of work.
- *F* is the mass flow to transform *P* into *Q*.
- D[P,Q] = WORK(P,Q,F) is the work to transform P into Q
- D[P,Q] is between 0 and 1.
- For any P_1 and P_2, D[P,Q]<=max(D[P_1,Q],D[P_2, Q]).

Earth Mover's distance

- EMD gives us a method for determining the distance between two distributions but doesn't tell us how to determine the distance between two elements in the distributions.
- The way to do that will differ depending on the type of data we're using...

How to Calculate the EMD?

- To use t-closeness with EMD, we need to be able to calculate the EMD between two distributions.
  - EMD for Numerical Attributes
  - EMD for Categorical Attributes

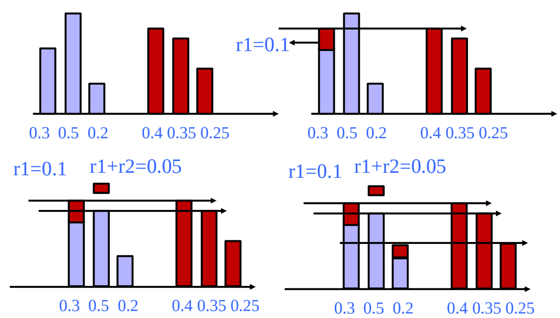
EMD for Numerical Attributes

- Numerical attribute values are ordered.
  - Let the attribute domain be $\{v_1, v_2 \ldots v_m\}$
  - Set $r_i = p_i - q_i$
  - The distance between P and Q can be calculated as:

$$D[\mathbf{P},\mathbf{Q}] = \frac{1}{m-1}(|r_1|+|r_1+r_2|+\ldots+|r_1+r_2+\ldots r_{m-1}|)$$

$$= \frac{1}{m-1}\sum_{i=1}^{i=m}\left|\sum_{j=1}^{j=i} r_j\right|$$

EMD for Numerical Attributes



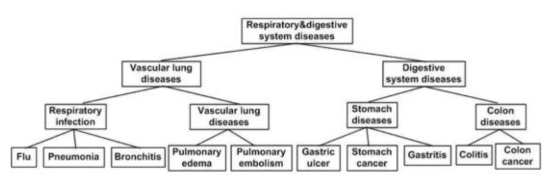
EMD for Categorical Attributes

- For categorical attributes(i.e., diseases), a total order often does not exist.
- We consider two distance measures.
  - Method 1: Equal Distance: The ground distance between any two value of a categorical attribute is defined to be 1.
    - As the distance between any two values is 1, for each point that $p_i - q_i > 0$, one just needs to move the extra to some other points.

$$D[\mathbf{P},\mathbf{Q}] = \frac{1}{2}\sum_{i=1}^{m}|p_i - q_i| = \sum_{p_i \geq q_i}(p_i - q_i) = -\sum_{p_i < q_i}(p_i - q_i)$$

EMD for Categorical Attributes

- Method 2: Hierarchical Distance: The distance between two values of a categorical attribute is based on the minimum level to which these two values are generalized to the same value according to the domain hierarchy.
  - Example: **Hierarchy for categorical attributes** ***Disease*****.**



EMD for Categorical Attributes

- Method 2: Hierarchical Distance
  - Several definitions:
    - we define the *extra* of a leaf node that corresponds to element i, to be $p_i - q_i$, and the *extra* of an internal node N to be the sum of *extras* of leaf nodes below N.
    - Child(N) is the set of all leaf nodes below node N

$$extra(N) = \begin{cases} p_i - q_i & \text{if N is a leaf} \\ \sum_{C \in Child(N)} extra(C) & \text{otherwise} \end{cases}$$

    - We further define two other functions for *internal nodes*:

$$pos\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) > 0} |extra(C)|$$

$$neg\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) < 0} |extra(C)|$$

EMD for Categorical Attributes

- Method 2: Hierarchical Distance
  - Several definitions:
    - The *extra* function has the property that the sum of *extra* values for nodes at the same level is 0.

$$cost(N) = \frac{height(N)}{H} \min(pos\_extra(N), neg\_extra(N))$$

  - Thus, the earth mover's distance can be written as:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_N cost(N)$$

Example of EMD

Remember this slide? Now let's calculate the EMD and create a *t*-close table.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

P1 = {3k, 4k, 5k}

P2 = {6k, 8k, 11k}

P_1 has more information leakage than P_2 because there are fewer people in that salary range and thus they are easier to identify, thus we should have $D[P_1,Q] > D[P_2,Q]$.

However, these algorithms just view 3k and 6k as different points and don't attach semantic meaning to them. They would calculate this wrong.

Example of EMD

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

P1 = {3k, 4k, 5k}

One optimal mass flow that transforms P_1 to Q is to move 1/9 probability mass across the following pairs:
3k->3k,3k->4k, 3k->5k
cost: 1/9*((3-3)+(4-3)+(5-3))/8
4k->6k,4k->7k 4k->8k
cost: 1/9*((6-4)+(7-4)+(8-4))/8
5k->9k,5k->10k 5k->11k
cost: 1/9*((9-5)+(10-5)+(11-5))/8

Total cost: 1/9*27/8=0.375

Remember: for numerical attributes, minimal work can be achieved by satisfying all elements of Q sequentially

Example of EMD

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

P2 = {6k, 8k, 11k}

One optimal mass flow that transforms P_2 to Q is to move 1/9 probability mass across the following pairs:
6k→3k, 6k→4K,6K→5k, cost=1/9*(3+2+1)/8
8k→6k, 8k→7k, 8k→8k cost=1/9*(2+1+0)/8
11k→9k, 11k→10k,11K→11K cost=1/9*(2+1)/8

The cost of this is 1/9× (12)/8 = 12/72 = 3/18 = 0.167.

Example of EMD

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}

P1 = {3k, 4k, 5k}

P2 = {6k, 8k, 11k}

In conclusion,
D[$P1$,Q] is 0.375 and D[$P2$,Q] has a distance of 0.167.
Therefore, P_2 **reveals less private data.**

Content

- **Motivation** of privacy preserving data publishing
  - Background
  - Privacy attack instances
- Existing **solutions**
  - *k*-Anonymity
  - *l*-Diversity
  - T-Closeness
- Challenges and Emerging Applications
- **Conclusion**

Challenges and Emerging Applications

- The problems of privacy preservation, re-identification, and inference control are not limited to non-aggregate microdata and contingency tables.
- In many of these new applications, the privacy goal is generally de-identification,
  - that is, the removal of personally identifiable information.

Challenges and Emerging Applications

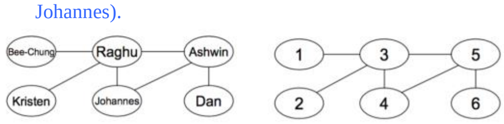
- Next, two representative emerging aplications
  - Social Network Privacy
  - Search Log Privacy

Social Network Privacy

- Social Network Privacy
  - Social networks describe entities (often people) and the relationships between them.
  - Social network analysis is often used to understand the nature of these relationships, such as patterns of influence in communities, or to detect collusion and fraud.
  - the release of data is often prevented by concerns about the privacy of individuals.

Social Network Privacy

- Naive De-Identification and Attacks
  - We will model a social network as a simple, undirected graph G = (V,E).
  - Nodes correspond to entities and edges represent connections between entities.
  - Each entity has an associated unique name (e.g., Raghu or Johannes).



(a) Original data (b) De-identified data

Social Network Privacy

- Naive De-Identification and Attacks
- Unfortunately, there are various ways in which this naive solution can be compromised.
  - Active attack:
    - an attacker actively manipulates the structure of the graph before the data are released
  - Passive attack
    - these attacks can be launched based on background knowledge related to the graph's structure

Search Log Privacy

- On July 29, 2006, AOL(America Online INC.) published three-month Web search queries of around 600 thousand users.
- For a given user, this data set contained the queries submitted by the user to the AOL search engine.

Table 7.1. Example search log.

	User ID	Query	Time	Rank	URL
1	User1	Tax ssn 111223333	2008-01-05 08:10		
2	User2	Restaurant arlington wi	2008-01-03 10:20	1	local.yahoo.com/...
3	User2	Restaurant arlington wi	2008-01-03 10:22	4	www.gorestaurants.net/...
4	User2	70 single men	2008-01-05 14:30		
5	User2	chen family tree	2008-01-06 20:01	1	chenfamilytree.com
6	User2	Nude pictures	2008-01-10 21:42		
7	User3	www.some-church.com	2008-01-08 10:35	1	www.some-church.com
8	User3	Tax for pastor	2008-01-13 22:50	8	answers.yahoo.com/...

[a]Rank and URL indicate the position and the URL on the search result page that the user clicked. Empty means no click.

Search Log Privacy

- To protect users' privacy, AOL replaced the AOL user names with randomly generated ID numbers.
- However, soon after the data set was released, many users together with their private queries were identified.

- As an example, the New York Times identified user No. 4417749 because this user searched for her family name, her hometown, and something about her age
  - By combining this information, it was not difficult to create a very short list of candidates that matched the information.

Search Log Privacy

- The AOL case signifies the need for appropriate search log anonymization. Existing privacy definitions do not apply directly to search logs.

- However, satisfactory solutions to search log publishing are still yet to be found.

Challenges in Emerging Applications

- The Curse of Dimensionality
  - With improving technology it is becoming easier to measure and record more information about each individual.
  - Thus, the number of attributes is growing, causing the size of the domain to increase exponentially.

Challenges in Emerging Applications

- Sequential Releases and Composability
  - The US Census Bureau publish data from the decennial census every 10 years
  - These sequential releases pose an additional privacy threat since user information can be linked across different releases.

Conclusion

- An overview of existing solutions for privacy preserving data publishing
  - *k*-Anonymity
  - *l*-Diversity
  - T-Closeness
- Challenges in Emerging Applications

Reference

- Li, Li, Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity" (ICDE 2007).
- Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity." (ICDE'06).
- L. Sweeney. "k-anonymity: a model for protecting privacy. "International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
- Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." Foundations and Trends® in Theoretical Computer Science 9, no. 3–4 (2014): 211-407.
- Dwork, Cynthia. "Differential privacy." In Encyclopedia of Cryptography and Security, pp. 338-340. Springer US, 2011.
- Li, Tiancheng, and Ninghui Li. "On the tradeoff between privacy and utility in data publishing." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 517-526. ACM, 2009.
- Diag, Pseudo Activity Age. "Privacy Preserving Data Publishing." (2008).
- Wang, Jian, Yongcheng Luo, Yan Zhao, and Jiajin Le. "A survey on privacy preserving data mining." In Database Technology and Applications, 2009 First International Workshop on, pp. 111-114. IEEE, 2009.

Q&A