k-ANONYMITY: A model for protecting privacy (By L.Sweeney)

Presented by : Navreet Kaur

ROADMAP

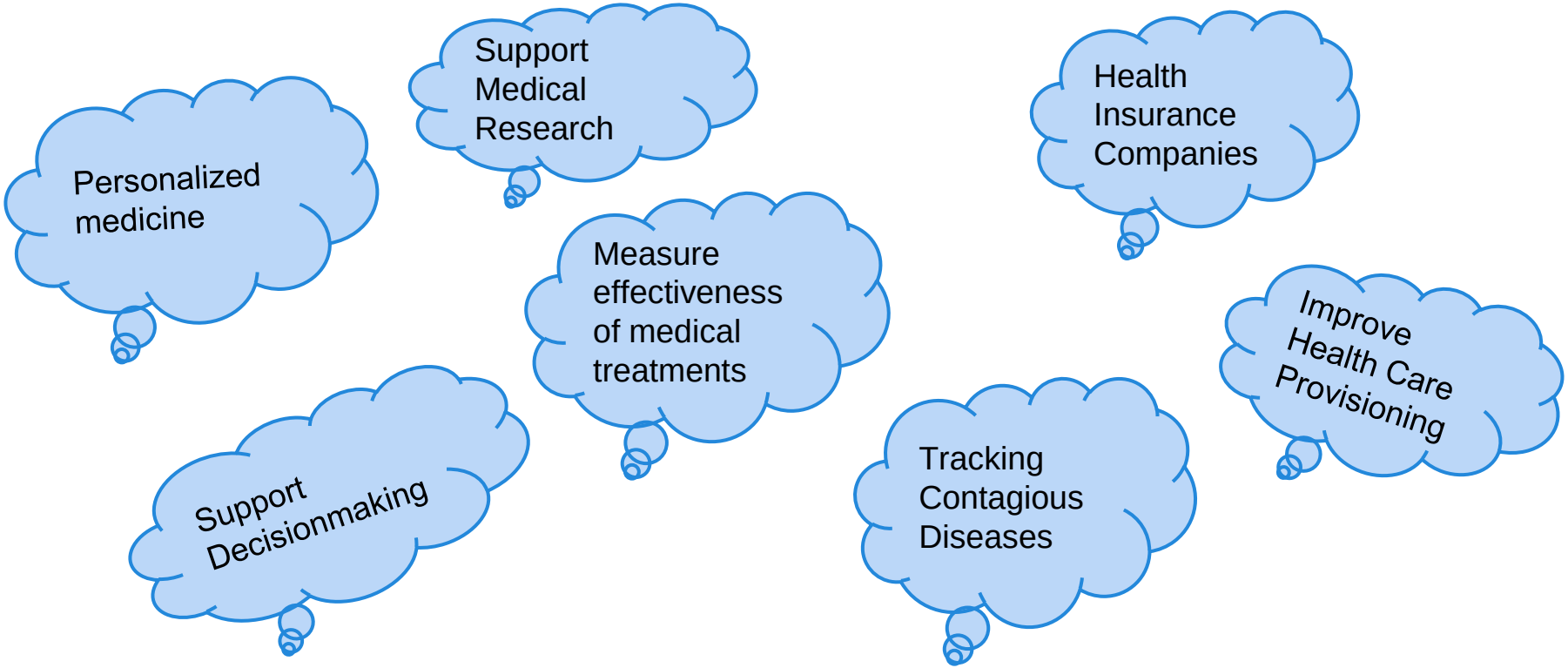
- Data sharing and data privacy
- Related background work
- k-anonymity model
- Possible attacks against k-Anonymity
- Weaknesses of k-Anonymity
- Extensions
- Conclusion

Data Sharing :

- Data sharing is making data used for scholarly research available to other investigators.*
- An exponential growth in number and variety of data collection containing person specific information.
- Collection of data is beneficial both in research and business.

* http://en.wikipedia.org/wiki/Data_sharing

Eg : Why Medical Data Sharing ?



Mutual supply chain data sharing — 1988

Increase in joint sales, reduced inventories and costs — 1990

Strategic sales information was disseminated in industry-wide reports even to Wal-mart's competitors* — 2001

Reconsidered data sharing strategy

* Grean et al. Supply chain partnership between P&G and Wal-Mart. Chapter 3, Integrated Series in Information Systems. 2002

Objective :

- Maximizing data utility while limiting disclosure risk to an acceptable level.

- How can a data holder release a version of its private data with guarantees that subjects of data cannot be re-identified and data is practically useful ?

Existing Works :

❑ Statistical Databases :

This technique involves various ways of adding noise while still maintaining some statistical invariance.

Limitations :

- Destroys integrity of data.

Existing works (contd) :

❑ Multi-level databases :

➔ Data is stored at different security classifications and users have different security clearances (Denning & Lunt).

➔ Suppression :Sensitive information and all information that allows inference of sensitive information is not released(Su and Ozsoyoglu).

Limitations :

- Protection only against known attacks.
- Suppression reduces quality of data.

Existing Works (contd):

❑ Computer Security :

Computer security is not privacy protection.

- It ensures that the recipient of information has the authority to receive information.
- Only prevents direct disclosures.

Privacy Protection : Release all the information such that identities of people who are subjects of data are protected.

k- Anonymity :

- It is a framework for constructing and evaluating algorithms & systems that release information such that released information limits what can be revealed about the properties of entities that are to be protected.

- Eg: If you want to identify a person and the only information you have is gender and zip code - there should be at least k number of people meeting the requirement.

Quasi Identifier :

- Attributes which appear in private data and also appear in public data are candidates for linking, these attributes constitute the Quasi Identifier and disclosure of these attributes should be controlled.

- Eg : {YOB, Gender, 3-digit Zip code} unique for 0.04% of US citizens

vs

{DOB, Gender, 5-digit Zip code} unique for 87% of US citizens*

*Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. IJUFKS. 2002

Hospital Patient Data

DOB	SEX	ZIP	DISEASE
10/21/74	M	528705	DIABETES
1/22/86	F	528718	BROKEN ARM
8/12/74	M	528745	HEPATITIS
5/7/74	M	528760	FLU
4/13/86	F	528652	FLU
9/5/74	F	528258	BRONCHITIS

Voter Registration Data

NAME	DOB	SEX	ZIP
BETH	10/21/74	M	528705
BOB	4/5/85	M	528975
KEELE	8/7/74	F	528741
MIKE	6/6/65	M	528985
LOLA	9/6/76	F	528356
BILL	8/7/69	M	528459

- Beth has diabetes

Hospital Patient Data

YOB	SEX	ZIP	DISEASE
1974	M	5287**	DIABETES
1986	F	5287**	BROKEN ARM
1974	M	5287**	HEPATITIS
1974	M	5287**	FLU
1986	F	5286**	FLU
1974	F	5282**	BRONCHITIS

Voter Registration Data

NAME	DOB	SEX	ZIP
BETH	10/21/74	M	528705
BOB	4/5/85	M	528975
KEELE	8/7/74	F	528741
MIKE	6/6/65	M	528985
LOLA	9/6/76	F	528356
BILL	8/7/69	M	528459

Release of Data Preventing linking of data.

k-Anonymity Protection Model :

Let RT (A1…….An) be a table, QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT [QI_{RT}] appears with at least k occurrences in RT[QI_{RT}],where :

- PT is private table.
- RT,GT1,GT2 are released tables.
- QI : Quasi Identifier
- (A1,A2,.....An) : Attributes

Assumption : Data holder has already identified the Quasi Identifier.

	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*}

For every combination of values of quasi identifiers in the 2-anonymous table,there are at least 2 records that share those values.

Fig from- *Sweeney: k-Anonymity: a Model for Protecting Privacy*

Attacks against k-anonymity:

- Unsorted matching attack : This attack is based on the order in which the tuples appear in the released table.

  Solution : Randomly sort the tuples of the solution table.

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of *k*-anonymity tables based on PT

Fig from- [Sweeney: k-Anonymity: a Model for Protecting Privacy]

Attacks against k-anonymity (contd):

- Complementary Release Attack : Subsequent releases of private data might compromise k-anonymity protection.

Solution :

- Consider attributes of previously released tables before releasing the new table.
- Base the subsequent releases on the initially released table.

Contemporary Attack (contd.) :

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Fig from- *Sweeney: k-Anonymity: a Model for Protecting Privacy*

Contemporary Attack (Contd.) :

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Fig from- *Sweeney: k-Anonymity: a Model for Protecting Privacy*

Attacks against k-anonymity(contd):

- Temporal attack : Data collections are dynamic. Adding,changing or removing tuples may compromise k-anonymity.

Solution :

- All the attributes released in an initial table should be considered as quasi identifiers for subsequent releases.
- Subsequent releases should be based on initial releases.

Conclude : K-Anonymity ensures that individuals cannot be identified by linking attacks

A little more…..

Limitations of k-anonymity:

- Homogeneity Attack :

	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

A. Machanavajjhala et al. *l*-Diversity: Privacy Beyond *k*-Anonymity. *ICDE* 2006

Limitations of k-Anonymity (contd.)

- Background Knowledge :

	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

A. Machanavajjhala et al. *l*-Diversity: Privacy Beyond *k*-Anonymity. *ICDE* 2006

Weaknesses of the paper :

- How to identify a set of “Quasi Identifier”?
- Dealing with large number of Quasi Identifiers could be problematic.

It generalizes or suppresses quasi identifiers to protect data which reduces quality of data.

Major Contribution

- This paper was one of the most initial attempts in privacy protection.
- It is used as a base for most of the privacy protection models.

Extensions to k-Anonymity model:

- l-Diversity
- t-Closeness
- a-k Anonymity
- e-m Anonymity, range diversity
- Personalized privacy

Conclusion

- Data sharing is important.
- Data utility needs to be maximised while private data should be protected.
- For every combination of values of quasi identifiers in the k-anonymous table , there are at least k records that share those values.
- k-anonymity protects data against linking attacks.
- But it was extended further as :
  - > k-anonymity can leak information due to lack of diversity.
  - > k-anonymity does not protect against attacks based on background knowledge.

Questions ?