Network Metrology

Privacy and Anonymization

This class

- Privacy and anonymization
  - Definitions and motivation
  - Anonymization methods
  - Case study: anonymization of packet traces

Definition: Personal data

- From CNIL website
  - “Any information relating to an identified or identifiable individual”;
  - “an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number (e.g. social security number) or one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (e.g. name and first name, date of birth, biometrics data, fingerprints, DNA…)”

Definition: Anonymization

- Process to sanitize data to ensure anonymity
  - Absence of identity
  - Prevent others from linking identity to actions of an individual

Privacy harms

- Identity theft
- Decisions based on personal data
  - Inaccuracies
  - Lack of transparency
- Blackmail
- Government surveillance
- Access to business intelligence, sensitive data
  - Embarrassing
  - Benefit competitors

Why anonymize data?

- Anonymization reduces risk in case of data leaks
- Benefits of sharing data
  - Good scientific practice
  - Get others to work on relevant problems
  - Get broader view

Challenges to sharing data

- Data protection laws
  - In Europe, data must be deleted after some time
- Hard to ensure data is fully anonymous
  - If released data can be combined with other datasets
- Often users/corporations own data
  - Hard to obtain permission to release data

ANONYMIZATION METHODS

What to anonymize?

- Identity-related information
  - E.g., IP address, user names, email
  - Require strict anonymization (best to remove)
- Personal-sensitive information
  - E.g., passwords, visited sites, downloaded content
  - Important to separate ID from sensitive information
- Organization-specific information
  - E.g., IP prefixes, location of peering points
- Business and security sensitive
  - E.g., routing config, capacity, policy decisions

Considerations to select anonymization method

- Trade-off: usefulness of data versus risks
  - Which info to protect
  - What can be shared but modified
  - What should not be lost
- Decisions depend on purpose of the analysis

Methods to anonymize identifiers

- Lossless transformation
  - Two-way hash function (using same seed)
  - Mapping is not lost, can do correlations
- Semi-lossy transformation
  - Removing portions of a string
  - E.g., removing URL suffix, remove last bits of IP
- Lossy transformation
  - E.g., Mapping string to numbers
  - One-way hashing with secret salt (random data).

Question: Is anonymization of identifiers sufficient?

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

Quasi-identifier

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

AOL example

- In 2006, AOL released 20 million search queries for 650.000 users
  - “Anonymized” by removing AOL id and IP address
  - Easily de-anonymized in a couple of days by looking at queries

The New York Times

Technology

A Face Is Exposed for AOL Searcher No. 4417749

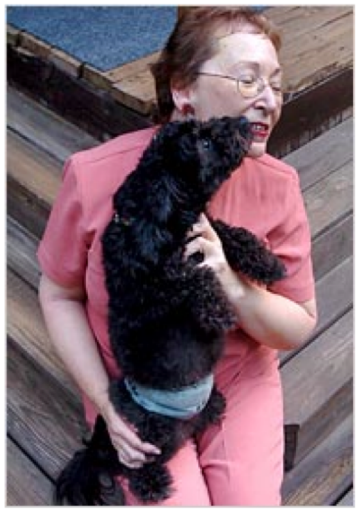
By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS

PRINT

REPRINTS

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.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

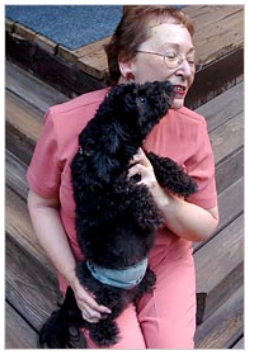
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

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 widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the

AOL User 4417749



- AOL query logs have the form

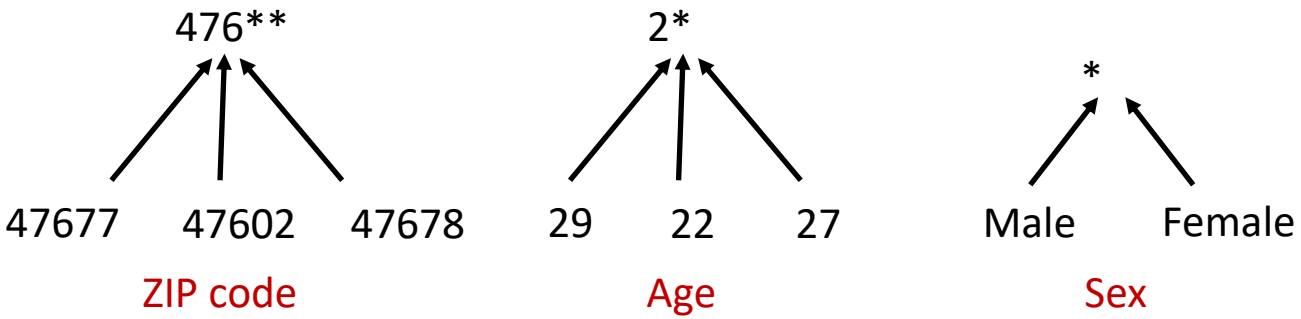
  <AnonID, Query, QueryTime, ItemRank, ClickURL>
  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold
    - Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User 4417749 is 62-year-old Thelma Arnold

K-Anonymity

- Intuition
  - The information for each person contained in the released data cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Any quasi-identifier present in the released data must appear in at least k records

Generalization

- Method
  - Replace quasi-identifiers with less specific, but semantically consistent values



Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with "outliers"
- Lots of algorithms in the literature
  - Aim to produce "useful" anonymizations
  - … usually without any clear notion of utility

Example of Generalization

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

Generalized table

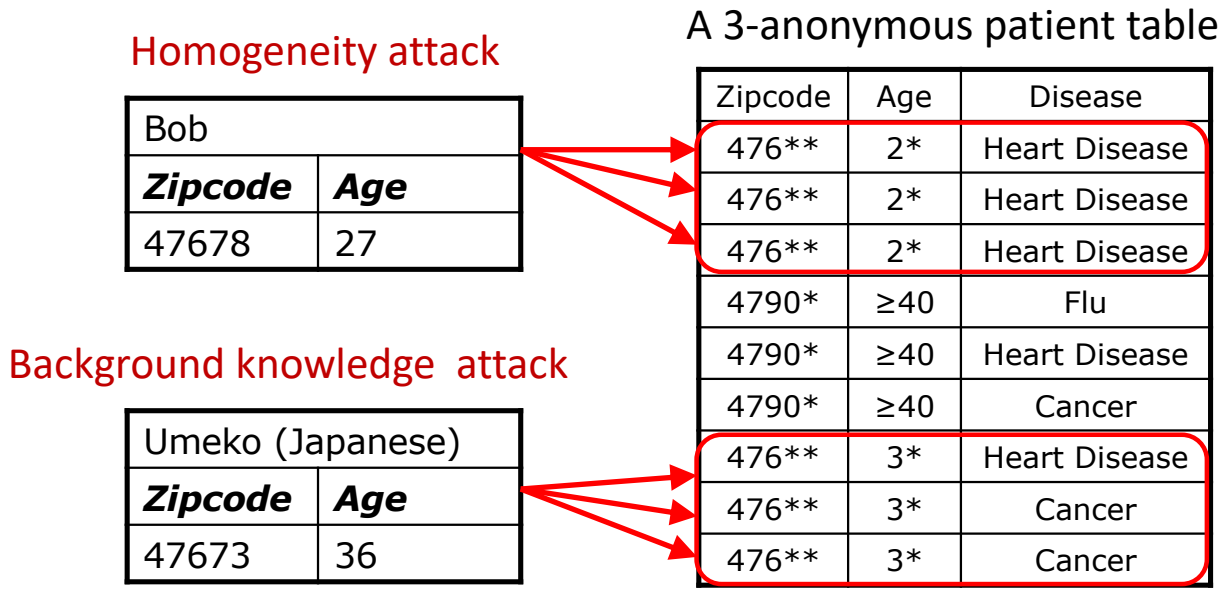
QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

!!

- 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

Question: Does k-anonymity guarantee privacy?

- Not if
  - Sensitive values in an equivalent class lack diversity
  - The attacker has background knowledge



Source: Vassiliadis, Panos. "l-diversity–privacy beyond k-anonymity."

l-Diversity

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

	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

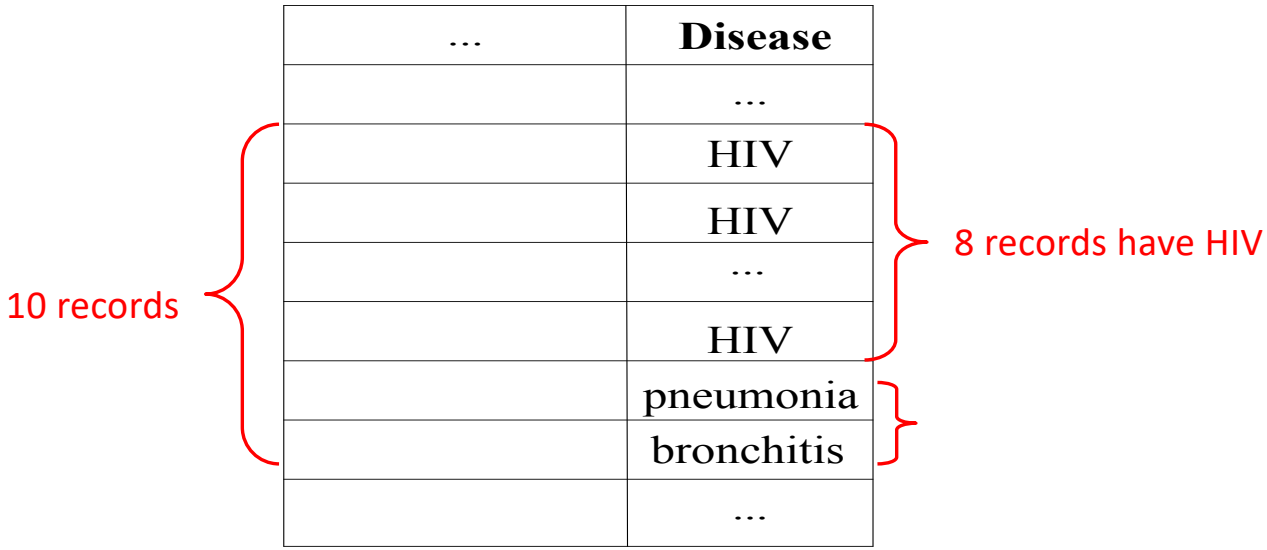
Figure 2. 4-anonymous Inpatient Microdata

	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

– Doesn't prevent probabilistic inference attacks



Limitations of *l*-Diversity

- 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

t-Closeness

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

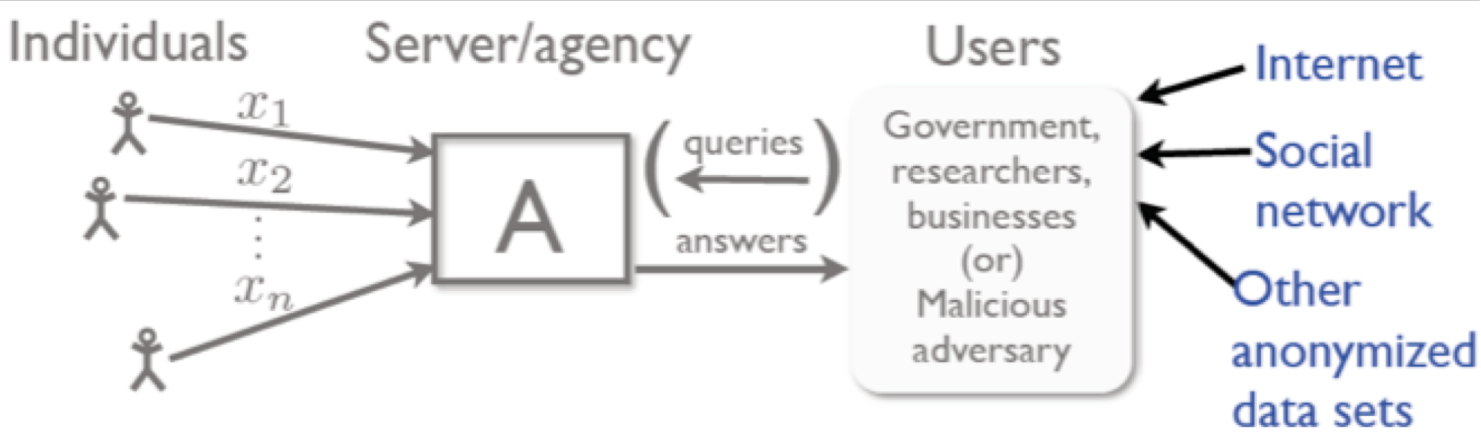
Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Source: Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.

Limitations of k-anonymity

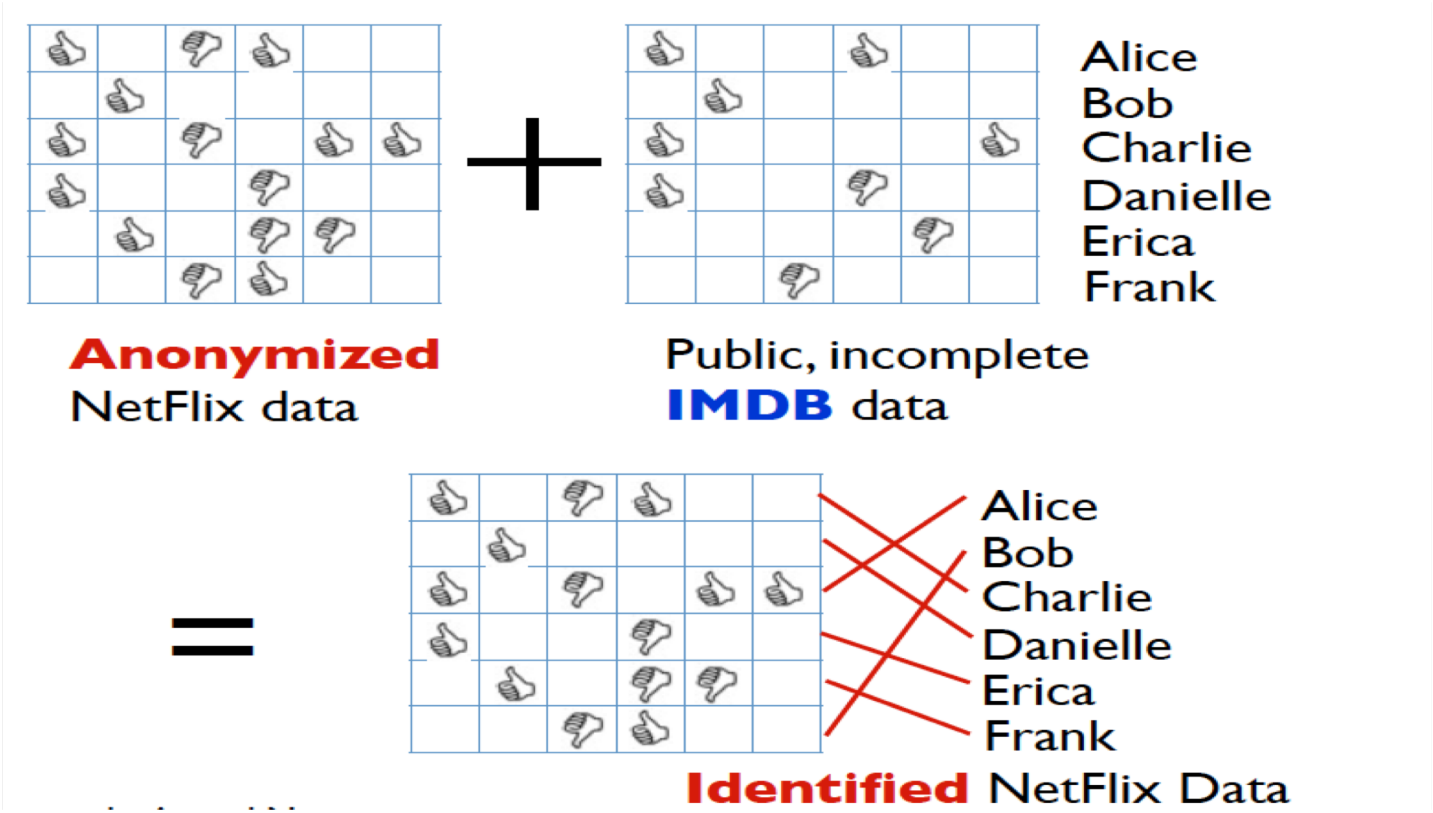
- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - "k-anonymous" dataset can leak sensitive information
- "Quasi-identifier" fallacy
  - Assumes a priori that attacker will not know certain information about his/her target
- Relies on locality
  - Destroys utility of many real-world datasets

Anonymization challenge: External information



- Users have external sources of data
  - Can't assume to know them all
  - Linkage attacks

Netflix example



Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008.

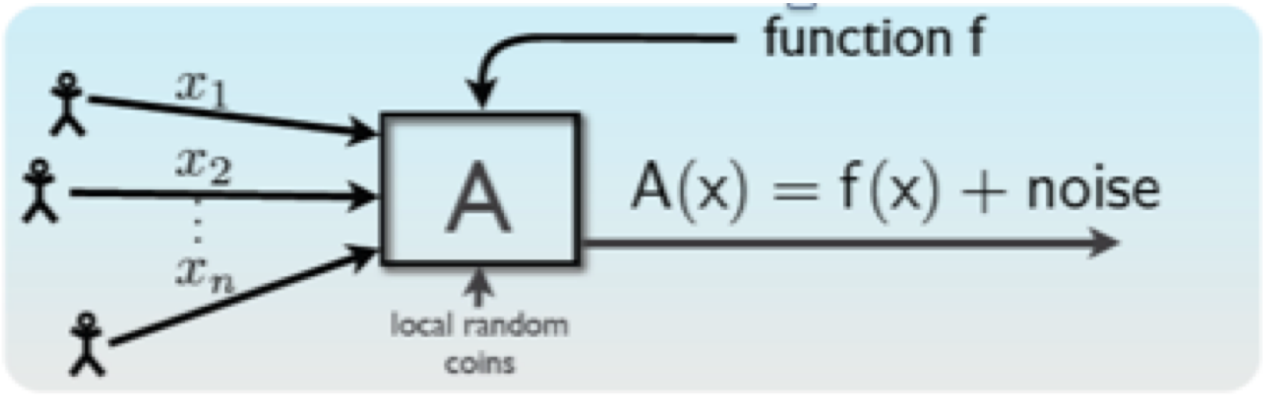
Differential privacy

- Approach to privacy-preserving data analysis.
- Enables release of statistical information about a sensitive dataset, while provably protecting individual-level information.
- Different privacy states that an adversary with access to the output of the data analysis will learn roughly the same information whether or not a single user's data was included or not.

Source: Dwork, Cynthia. "Differential privacy: A survey of results." *International Conference on Theory and Applications of Models of Computation*. Springer, Berlin, Heidelberg, 2008.

Differential privacy

- Intuition
  - Trusted mediator *A* receives queries from user and responds by adding some noise
  - Result is independent of any user's data



CASE STUDY: PACKET TRACES

Information in packet traces

- Full packets contain lots of sensitive info
  - Headers: connection endpoints, protocol
  - Payload: visited content, passwords, etc.

What to anonymize

- Payload most sensitive information
  - Better if removed completely
  - If not possible, get minimum necessary
    - E.g., HTTP host better than full URL
- Packet headers can be shared with care
  - MAC addresses
    - Potential to link records with the same MAC across datasets
  - IP addresses often need to be anonymized
  - IP addresses appear in other parts of the packet
    - IP options (e.g., record route)
    - ICMP/DNS packets

Anonymization of IP addresses

- Simply hashing IP addresses is too restrictive
  - IPs in the same prefix announced by the same AS
  - Many analysis require knowledge of IP prefix
- Prefix-preserving anonymization
  - All IP addresses that share a prefix in the raw data will also do so in the anonymized data
- Packet trace anonymization tools
  - tcpdpriv, ipsumdump, ip2anonip, Crypto-PAn, PktAnon

Summary

- Sharing data is important, but hard
- Anonymization of identifiers
  - Lossless: two-way hashing
  - Semi-lossy: e.g. prefix preserving
  - Lossy: mapping strings to numbers
- Privacy properties of anonymized datasets
  - k-anonymity, l-diversity, t-closeness
- Alternative model: send queries to data
  - ε-differential privacy

References

- CNIL data protection
  - http://www.cnil.fr/english/data-protection/
- “Internet measurement: Infrastructure, traffic & applications”, Chapters 8 and 9
- Class on k-anonymity and other cluster-based methods
  - https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0ahUKEwjEst2N8pnKAhUI_A4KHSzKAJQQFggrMAI&url=https%3A%2F%2Fwww.cs.utexas.edu%2F~shmat%2Fcourses%2Fcs380s_fall09%2F21kanon.ppt&usg=AFQjCNEwUcK9mXssQBgE2jS5gsvrsrXGmA&sig2=Fb4T1mhViTP5WJY9le4syg
- “Differentially-private network trace analysis”, SIGCOMM 2010
  - http://www.sigcomm.org/node/2880