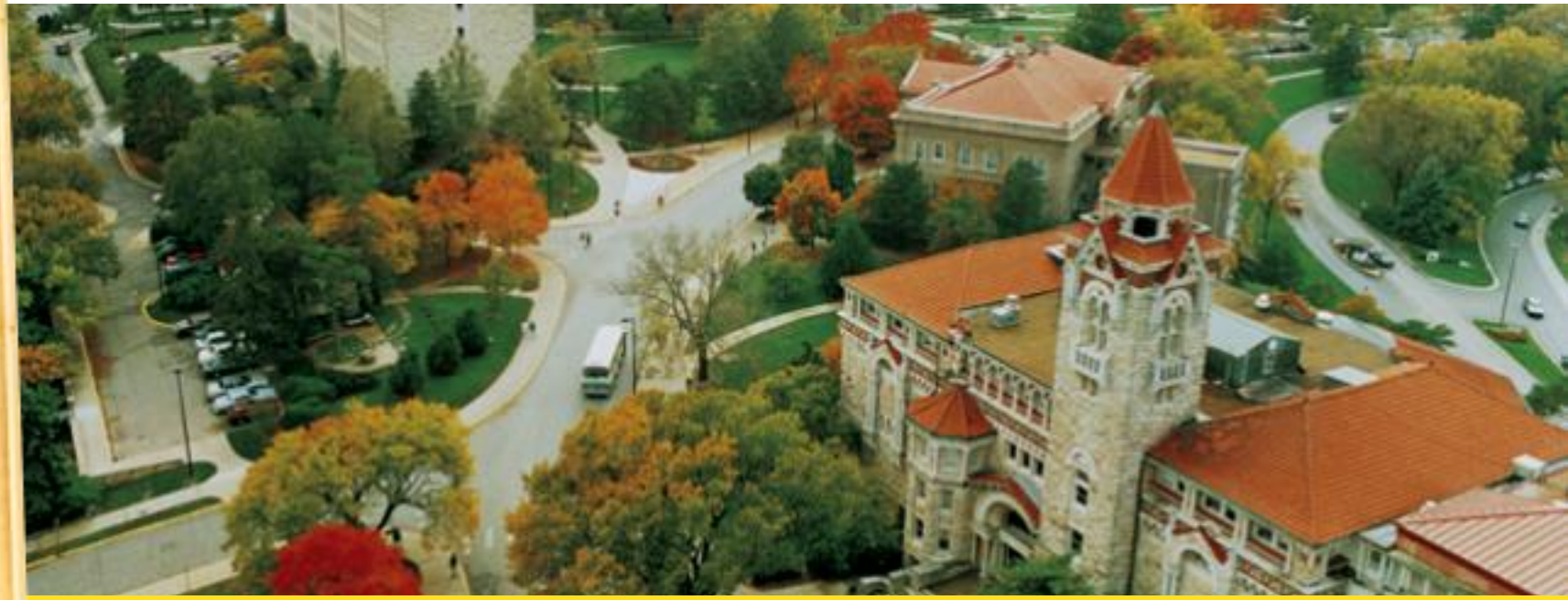
Privacy I: Data Privacy

Bo Luo
Associate Professor, EECS
Director, Information Assurance Lab, ITTC
The University of Kansas, Lawrence, KS, USA
bluo@ku.edu; http://www.ittc.ku.edu/~bluo



Privacy

- What is privacy?
  - Sensitive personal information?
  - Identifiable information?
  - Information access and information flow?
  - Usage of information
- Anonymity: data anonymity, network anonymity
- Private information publishing/sharing
- Privacy preserving data mining

Regulatory requirements

- Healthcare Information Portability and Accountability Act (HIPAA)
- Governs use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual; For treatment (generally requires consent); To public health / legal authorities
  - Use permitted where “there is no reasonable basis to believe that the information can be used to identify an individual”
- Safe Harbor Rules
  - Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
    - Name, location smaller than 3 digit postal code, dates finer than year, identifying numbers
    - Shown not to be sufficient (Sweeney) Also not necessary

Data anonymity

- Data Collection: a large amount of person-specific data has been collected (over a period of time).
- Data Mining: data and knowledge extracted by data mining techniques represent a key asset to the society.
  - Analyzing trends/patterns.
  - Formulating public policies.
- Regulatory Laws: some collected data must be made public.
  - Census data

Data anonymity

- Privacy
  - The data usually contains sensitive information about respondents.
  - Respondents' privacy may be at risk.
- Two opposing goals
  - To allow researchers to extract knowledge about the data
  - To protect the privacy of every individual

Individual Privacy: protect the "record"

- Individual item in database must not be disclosed
- Not necessarily a person
  - Information about a corporation
  - Transaction record
- Disclosure of parts of record may be allowed
  - Individually identifiable information?

Individually Identifiable Information

- Microdata table
  - Identifier (ID), Quasi-Identifier (QID), Sensitive Attribute (SA)

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

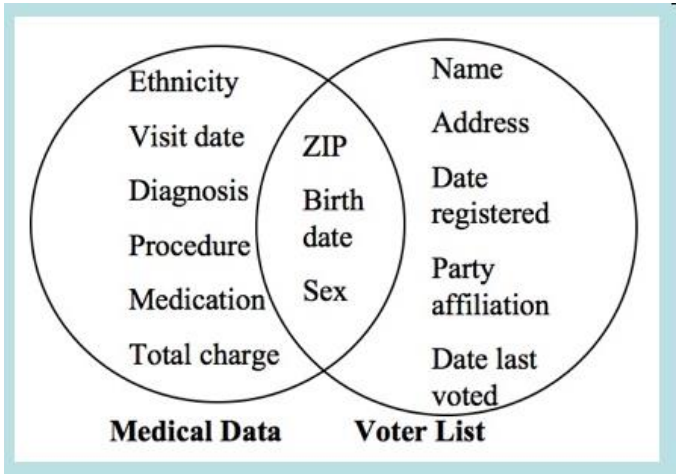
Individually Identifiable Information

- First try: remove the identifiers

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

Individually Identifiable Information

- Latanya Sweeney @ CMU
  - Collection of personal information is mandated in 37 states.
  - Data is de-identified, and publicly available.
  - She purchased voter's registration data from Mass. and compared with medical record.
  - 87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}.



Individually Identifiable Information

- Removing identifiers is not enough!

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

Classes of Solutions

- Data Obfuscation
  - Nobody sees the real data
- Summarization
  - Only the needed facts are exposed
- Data Separation
  - Data remains with trusted parties

Data Obfuscation

- Goal: Hide the protected information
- Approaches
  - Randomly modify data
  - Swap values between records
  - Controlled modification of data to hide secrets
  - Constrains: should not change statistical distribution, should not interfere legitimate use of data
- Problems
  - Does it really protect the data?
  - Can we learn from the results?

Data Obfuscation

- Example: US Census Bureau Public Use Microdata
- US Census Bureau summarizes by census block
  - Minimum 300 people; ranges rather than values
- For research, “complete” data provided for sample populations
  - Identifying information removed: limitation of detail: geographic distinction, continuous interval; Top/bottom coding (eliminate sparse/sensitive values)
  - Swap data values among similar individuals: if individual determined, sensitive values likely incorrect

Data Summarization

- Goal: Make only innocuous summaries of data available
- Approaches
  - Overall collection statistics
  - Limited query functionality
- Problems
  - Can we deduce data from statistics?
  - Is the information sufficient?

Data Summarization

- Example: Statistical Queries
- User is allowed to query protected data
  - Queries must use statistical operators that summarize results
    - Example: Summation of total income for a group doesn't disclose individual income
  - Multiple queries can be a problem
    - Request total salary for all employees of a company
    - Request the total salary for all employees but the president
    - Now we know the president's salary

Data Summarization

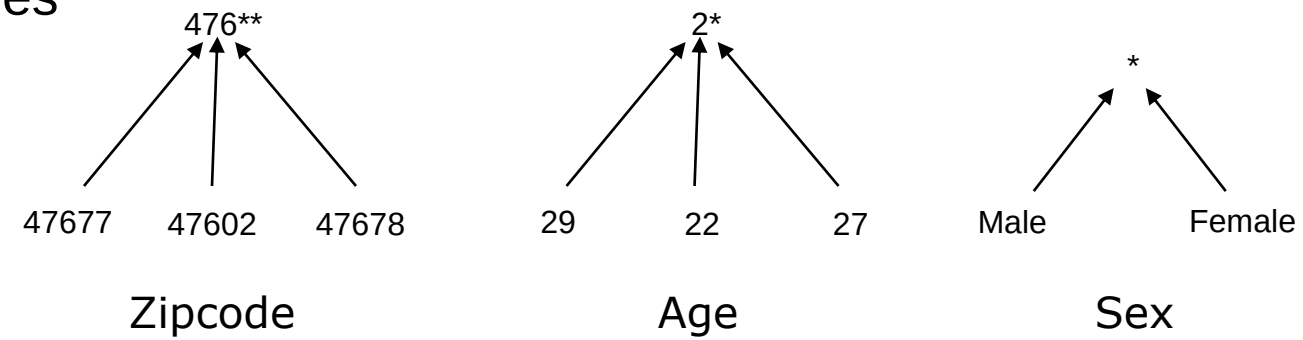
- Controls
  - Query restriction – Identify when a set of queries is safe
    - Result generated from at least k items
    - Items used to generate result have at most r items in common with those used for previous queries
  - Data perturbation: introducing noise into the original data
  - Output perturbation: leaving the original data intact, but introducing noise into the results

Data Separation

- Goal: Only trusted parties see the data
- Approaches
  - Data held by owner/creator
  - Limited release to trusted third party
  - Operations/analysis performed by trusted party
- Problems
  - Will the trusted party be willing to do the analysis?
  - Do the analysis results disclose private information?

k-Anonymity & Generalization

- k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalent class
  - k-Anonymity ensures that linking cannot be performed with confidence > $1/k$.
- Generalization 替换为不太具体但语义一致的值
  - Replace with less-specific but semantically-consistent values



k-Anonymity & Generalization

- 3-Anonymous table
  - Suppose that the adversary knows Alice's QI values (47677, 29, F).
  - The adversary does not know which one of the first 3 records corresponds to Alice's record.

The 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

The 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

k-Anonymity & Generalization

- This is wrong
  - 3-anonymity on each quasi-identifier
  - Uniquely identifiable on the combination

The Microdata

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	F	Ovarian Cancer
476**	2*	M	Ovarian Cancer
476**	3*	F	Prostate Cancer
479**	3*	M	Flu
479**	3*	F	Heart Disease
479**	2*	M	Heart Disease

k-Anonymity & Generalization

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

均衡性攻击

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

l-Diversity

- Principle
  - Each equivalence class has at least l well-represented sensitive values
- Distinct l-diversity
  - Each equivalence class has at least l distinct sensitive values
  - Probabilistic inference

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

10 records

8 records have HIV

2 records have other values

l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by $1/l$.
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least log(l)
- Recursive (c,l)-diversity
  - The most frequent value does not appear too frequently
  - $r_1 < c(r_l + r_{l+1} + ... + r_m)$ where r_i is the frequency of the i-th most frequent value.

l-Diversity

- Limitations:
- A single sensitive attribute
  - Two values: HIV positive (1%) and HIV negative (99%)
  - Very different degrees of sensitivity
- l-diversity may be unnecessary to achieve
  - 2-diversity is unnecessary for an equivalence class that contains only negative records
- l-diversity may be difficult to achieve
- Skewness attack
- Similarity attack

Privacy

- k-Anonymity and l-Diversity are powerful and popular solutions (from technical point of view).
- For more reading: t-closeness.

Case Study: The Netflix Prize

- The Netflix Prize: who has the best prediction algorithm?
  - 100M ratings from 480K users on 17K movies
  - Data was (not so) carefully sanitized: anonymized, modified dates, partial data.
  - Movie information (title and year) was provided
- Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset), Oakland 2008.
  - Netflix was sued and Netflix Prize II was canceled.
- Anonymization is NOT enough!

NETFLIX

Perfect Privacy?

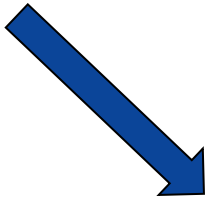
- Still remember Shannon Secrecy?
  - Probability of guessing the plaintext knowing the ciphertext = probability of guessing plaintext without knowing ciphertext.
  - Probability of any message giving a ciphertext is the same
- Consider an object O_k, and a characteristic D, which is a survey characteristic. For the object O_k this characteristic assumes the value D_k. If the release of the statistics S makes it possible to determine the value of D_k more accurately than it is possible without access to S, a disclosure has taken place.
  - Dalenius, T. 1977. Towards a methodology for statistical disclosure control. *Statistik Tidskrift 15*, 429-444, 2–1.

Perfect Privacy?

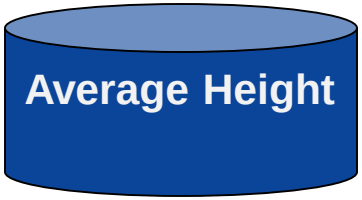
- "Perfect privacy": Anything that can be learned about a respondent from the published dataset (or statistical database) can be learned without access to the database.

- Unfortunately, this is NOT achievable.

Alice is 7 inches taller than average



+



= **Alice's height**

Perfect Privacy?

- “Perfect privacy”?
- I will participate in the survey only if my participation will not change the output.
  - Privacy vs. Utility?
- Published result will not disclose any information about me.
  - It will, whether you are in the survey or not.

- Whether I submit my information or not, the attacker’s information gain from the dataset (through whatever statistical query) stays ALMOST the same.

Perfect Privacy?

- "Perfect privacy"?
  - We have a database of all the employees in the company
  - Users could only issue aggregate queries (e.g., sum, average) on the *salary* attribute
  - Two queries: before and after the CEO's resignation
  - Now you know the CEO's salary
  - What if the database only gives you an approximate result?
- A protocol that has a probability distribution over outputs such that if person i changed her input from x_i to any other allowed x_i', the relative probabilities of any output do not change ***by much***.

Differential Privacy 不考

- A theoretical model
  - Dataset D
  - D_1 and D_2 are two versions of D, differ on at most one record
  - M is a statistical query or a data mining algorithm

M is ε-differential private, if

$$\Pr[M(D_1) = \mathrm{R}] \leq e^{\varepsilon} \times \Pr[M(D_2) = \mathrm{R}]$$

  - Whether you (or anyone) are in the dataset or not, no outputs (and consequences of outputs) would become significantly more or less likely.

Differential Privacy

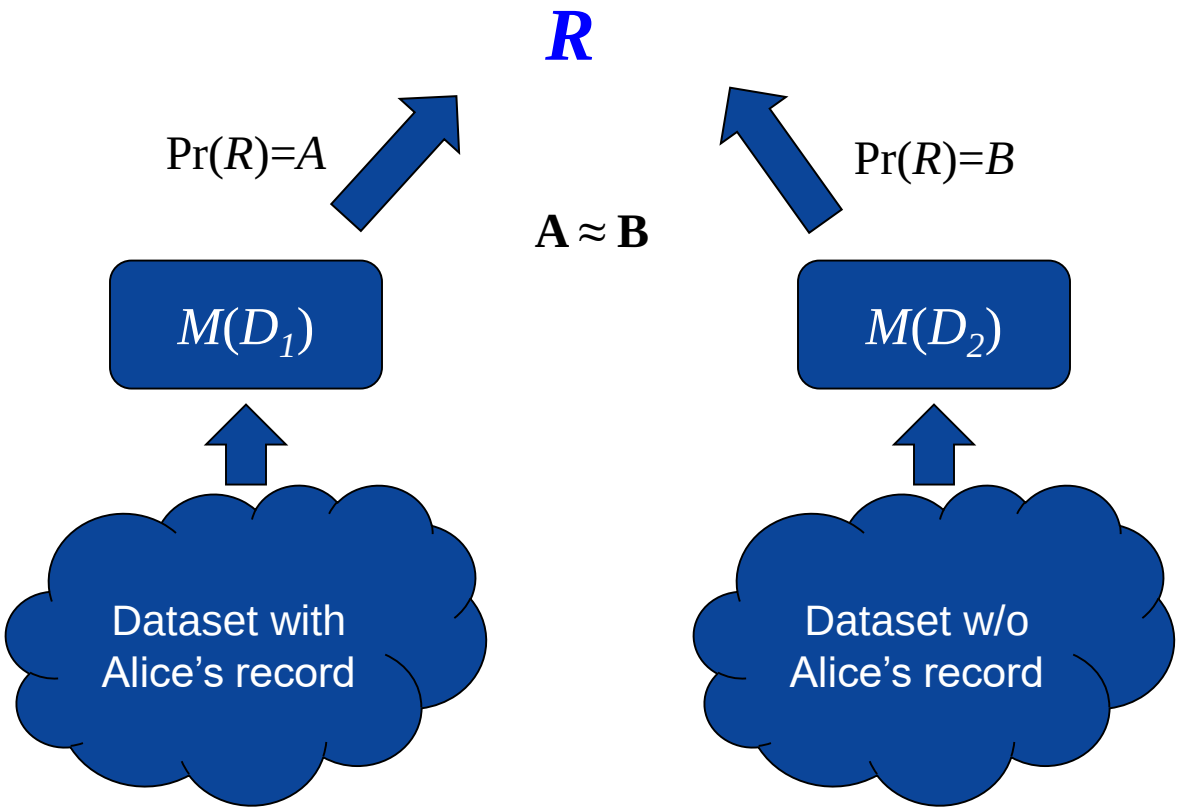
- A theoretical model
  - Dataset D
  - D_1 and D_2 are two versions of D, differ on at most one record
  - M is a statistical query or a data mining algorithm

M is ε-differential private, if

$$\Pr[M(D_1) = \mathrm{R}] \leq e^{\varepsilon} \times \Pr[M(D_2) = \mathrm{R}]$$

  - D_1 and D_2 are *neighboring datasets* (or *adjacent datasets*)

Differential Privacy



How to Achieve Differential Privacy?

- Output perturbation
  - Add noise to the output
  - But, how much?
  - Consider these two functions:
  1. Return the average salary
  2. Return the total summary
- The scale of the added noise
  - be proportional to the maximum difference between two neighboring datasets

How to Achieve Differential Privacy?

- Sensitivity of a function
  - How much one record could affect the output?
  - D_1 and D_2 are two *neighboring datasets*, differ in at most one record.

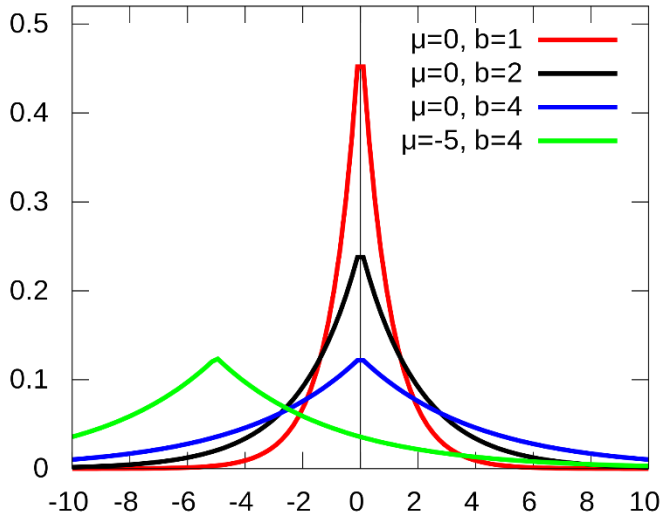
$$\Delta f = \max_{D_1, D_2} | M(D_1) - M(D_2) |$$

  - Example: count() have sensitivity 1
  - Sensitivity of sum(): largest value of the attribute.

How to Achieve Differential Privacy?

- Laplace distribution
  - Lap(μ, b): μ is the position, b is the scale (also called spread)

$$P(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$



How to Achieve Differential Privacy?

- Add Laplacian noise.

> On query f, to achieve ε-differential privacy, use scaled symmetric noise $Lap(b)$ with $b = \frac{\Delta f}{\varepsilon}$.

  - Thus, the distribution of the returned value will be:

$$Pr(R = x|\text{D is the true world}) = \frac{\varepsilon}{2\Delta f}\, e^{-\frac{|x-F(D)|\varepsilon}{\Delta f}}$$

How to Achieve Differential Privacy?

- Add Laplacian noise.

> On query f, to achieve ε-differential privacy, use scaled symmetric noise $Lap(b)$ with $b = \frac{\Delta f}{\varepsilon}$.

  - Can you prove that this achieves ε-differential privacy?

Differential Privacy

- Good for low-sensitivity functions/queries
  - Good: count()
  - Not so good: sum()
  - Even worse: max(), min()
- Statistical inference
  - Repeated queries
- Implementation
- Still trade-off between utility and privacy

Differential Privacy

Name	School year	Absence days
Chris	1	1
Kelly	2	2
Pat	3	3
Terry	4	10

Attack model: The adversary knows the true world X. His goal is to figure out absence of a victim individual in X' by using knowledge of X.

Query: mean absence days vs. mean school year

Possible world(ω)	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.01$
{1, 2, 3}	0.9705	0.5519	0.4596	0.3477	0.2328	0.2482
{1, 2, 10}	0.0159	0.1859	0.2019	0.2305	0.2527	0.2503
{1, 3, 10}	0.0087	0.1463	0.1791	0.2171	0.2558	0.2506
{2, 3, 10}	0.0049	0.1159	0.1594	0.2048	0.2588	0.2509

"There always exists a distribution that is more likely than others given the query response." -- Lee 2011

Lee, Jaewoo, and Chris Clifton. "How much is enough? choosing ε for differential privacy." *International Conference on Information Security*. Springer Berlin Heidelberg, 2011.

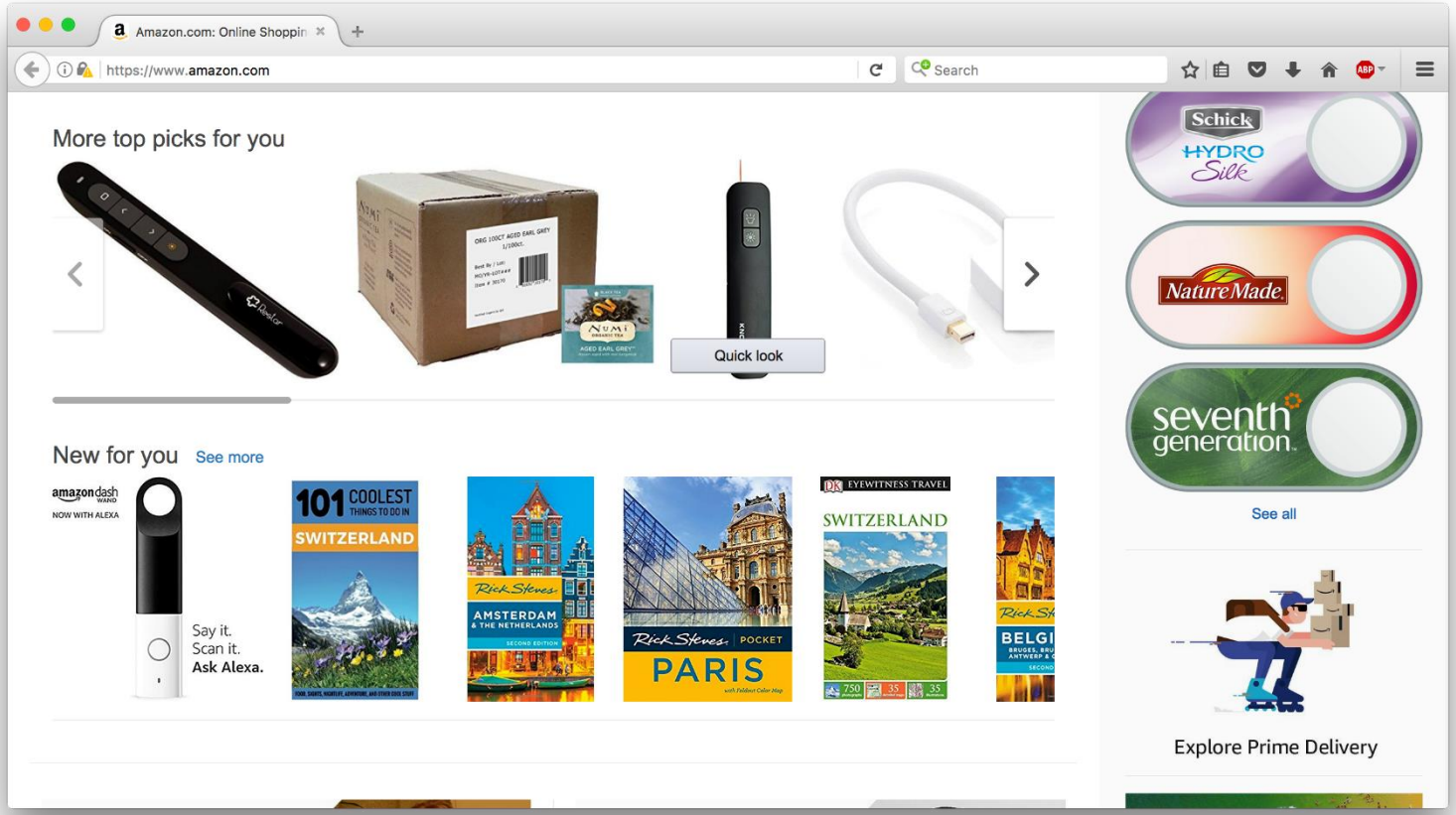
Privacy

- Plausible deniability: Plausible deniability is the ability for persons (typically senior officials in a formal or informal chain of command) to deny knowledge of or responsibility for any damnable actions committed by others (usually subordinates in an organizational hierarchy) because of a lack of evidence that can confirm their participation, even if they were personally involved in or at least willfully ignorant of the actions.

-- Wikipedia

Plausible deniability

- Personalized pages reveal private information
  - Recommendations are based on inferred attributes



Pól Mac Aonghusa and Douglas J. Leith. 2016. Don't Let Google Know I'm Lonely. ACM Trans. Priv. Secur. 19, 1, Article 3 (August 2016), 25 pages.

Plausible deniability

- Personalized pages reveal private information
  - Recommendations are based on inferred attributes
  - Search engines
  - Social networks
  - Online shopping/recommendation
  - Advertising
  - Media: videos, books, news, etc.

Pól Mac Aonghusa and Douglas J. Leith. 2016. Don't Let Google Know I'm Lonely. ACM Trans. Priv. Secur. 19, 1, Article 3 (August 2016), 25 pages.

Plausible deniability

- Threat model
  - *Distinguishability* instead of individual identifiability
    - Does not seek to *identify* the user as an individual
    - Seeks to determine the user's likely interest in commercially valuable topics
    - Privacy becomes an issue when any of the topics match subjects deemed sensitive by the user
- Research Challenges
  - How to quantify (model) plausible deniability?
  - How to enforce?
    - It's not easy to confuse the search engines!

Pól Mac Aonghusa and Douglas J. Leith. 2016. Don't Let Google Know I'm Lonely. ACM Trans. Priv. Secur. 19, 1, Article 3 (August 2016), 25 pages.