Data Mining & Differential Privacy

Mihai Maruseac

June 4, 2014

Context

- big data, stream of information

Context

- big data, stream of information
- important applications publishing sensitive data about individuals

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health
    - patterns of disease spreading

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health
    - patterns of disease spreading
  - web search

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health
    - patterns of disease spreading
  - web search
    - better search results

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health
    - patterns of disease spreading
  - web search
    - better search results
    - better recommendations

Context

- big data, stream of information
- important applications publishing sensitive data about individuals
  - medical research
    - best treatments
    - better diagnosis
    - mapping drugs to phenotypes
  - public health
    - patterns of disease spreading
  - web search
    - better search results
    - better recommendations
    - authoritative answers

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic
    - relationships between people (Facebook, etc.)

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic
    - relationships between people (Facebook, etc.)
    - census data, society evolution

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic
    - relationships between people (Facebook, etc.)
    - census data, society evolution
  - industrial data

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic
    - relationships between people (Facebook, etc.)
    - census data, society evolution
  - industrial data
    - details about sales, income, customers, costs, etc.

Context (2)

- big data, stream of information
- important applications publishing sensitive data about individuals
  - urban planning
    - home, workplace, leisure – tracked by GPS
    - travel patterns, experience, cost (time and money)
  - energy conservation
    - patterns of usage
    - changing the behaviour to better via smart appliances / on demand energy sources
  - networking
    - pattern of infrastructure usage
    - handle exceptional traffic
    - relationships between people (Facebook, etc.)
    - census data, society evolution
  - industrial data
    - details about sales, income, customers, costs, etc.
  - workflows

Context :: Access

Access strictly controlled

- only inside company/agency which collected the data

Context :: Access

Access strictly controlled

- only inside company/agency which collected the data
- only after signing a special contract (taxi, click streams)

Context :: Access

Access strictly controlled

- only inside company/agency which collected the data
- only after signing a special contract (taxi, click streams)
- only in coarse-grained summaries (health)

Context :: Access

Access strictly controlled

- only inside company/agency which collected the data
- only after signing a special contract (taxi, click streams)
- only in coarse-grained summaries (health)
- only after a long wait (census)

Context :: Access

Access strictly controlled

- only inside company/agency which collected the data
- only after signing a special contract (taxi, click streams)
- only in coarse-grained summaries (health)
- only after a long wait (census)
- only with 3-letters-organisations approval

Context :: Issues

- access to data strictly controlled
- data released with privacy issues (AOL click stream)

Society would benefit if we could publish useful data without worrying about privacy and access issues.

Privacy

- naïve solution: remove sensitive columns

Privacy

- naïve solution: remove sensitive columns
- cross table references

Privacy

- naïve solution: remove sensitive columns
- cross table references
- 87% Americans uniquely identified by `zip`, `gender`, `birthdate`

Privacy

- naïve solution: remove sensitive columns
- cross table references
- 87% Americans uniquely identified by `zip`, `gender`, `birthdate`
- 2002, medical records of Governor of MA

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA
- population 11,000

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA
- population 11,000
- some queries for Jarret Arnold

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA
- population 11,000
- some queries for Jarret Arnold
- 14 people with this name in Lilburn

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA
- population 11,000
- some queries for Jarret Arnold
- 14 people with this name in Lilburn
- contact each of them / social engineering

Privacy (2)

- query logs: useful for CS researchers, system admins, etc.
- find all AOL logs of user 4417749
- multiple queries for services in Lilburn, GA
- population 11,000
- some queries for Jarret Arnold
- 14 people with this name in Lilburn
- contact each of them / social engineering
- AOL user 4417749 = Thelma Arnold

Privacy (3)

- Netflix prize (2009, 1M$)

Privacy (3)

- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers

Privacy (3)

- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed

Privacy (3)

- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed
- use blogs, FB posts, twitters post, IMDB profiles to identify users

Privacy (3)

- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed
- use blogs, FB posts, twitters post, IMDB profiles to identify users
  - 8 ratings, dates within 2 weeks – 99% of raters

Privacy (3)

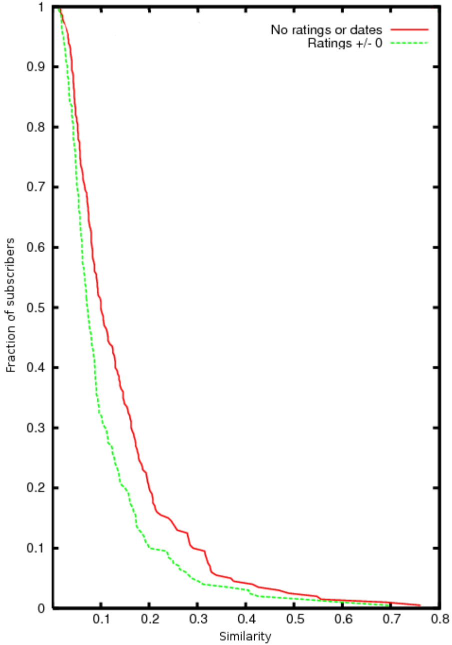
- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed
- use blogs, FB posts, twitters post, IMDB profiles to identify users
  - 8 ratings, dates within 2 weeks – 99% of raters
  - 2 ratings, dates within 3 days – 68% of raters

Privacy (3)

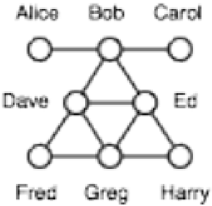
- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed
- use blogs, FB posts, twitters post, IMDB profiles to identify users
  - 8 ratings, dates within 2 weeks – 99% of raters
  - 2 ratings, dates within 3 days – 68% of raters
  - all raters of top 100 movies

Privacy (3)

- Netflix prize (2009, 1M$)
- training data: 100M ratings from 18K movies of 500K customers
- 10% of data slightly disturbed
- use blogs, FB posts, twitters post, IMDB profiles to identify users
  - 8 ratings, dates within 2 weeks – 99% of raters
  - 2 ratings, dates within 3 days – 68% of raters
  - all raters of top 100 movies
  - IMDB comments – Netflix reidentification


No ratings or dates
Ratings +/- 0
Fraction of subscribers
Similarity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

One customer . . . sued Netflix, saying she thought her rental history could reveal that she was a lesbian before she was ready to tell everyone.



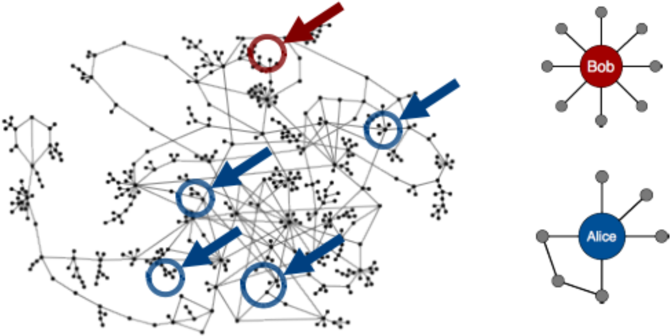
Edges from call & email logs: what did they know and when did they know it?

Nodes

ID	Age	HIV
Alice	25	Pos
Bob	19	Neg
Carol	34	Pos
Dave	45	Pos
Ed	32	Neg
Fred	28	Neg
Greg	54	Pos
Harry	49	Neg

Edges

ID1	ID2
Alice	Bob
Bob	Carol
Bob	Dave
Bob	Ed
Dave	Ed
Dave	Fred
Dave	Greg
Ed	Greg
Ed	Harry
Fred	Greg
Greg	Harry


Bob
Alice

Important note

Just because data looks hard to re-identify, doesn't mean it *is*.

Solution

Publish a distorted version of the data.

Solution

Publish a distorted version of the data.

privacy
: privacy "adequately" protected

utility
: information is useful for its intended purpose

Solution

Publish a distorted version of the data.

privacy privacy “adequately” protected

utility information is useful for its intended purpose

privacy $\nearrow$, utility $\searrow$

Privacy protection needs

- membership disclosure: is X in Xs?
- sensitive attribute disclosure: has X a?
- identity disclosure: does i belong to X? are x and y the same?

k-anonymity

Your quasi-identifiers are indistinguishable from at least other k people's.

k-anonymity

Your quasi-identifiers are indistinguishable from at least other k people's.

- easy to understand
- easy to attack
  - doesn't say anything about operations done on data
  - join on other columns
  - no protection against background knowledge
  - updates (age) destroy protection

Other approaches

l-diversity : each group must have at least l distinct values

probabilistic l-diversity : frequency of the most frequent value in a class is bounded by $1/l$

entropy l-diversity : entropy of distribution of values inside a class is at least $\log(l)$

recursive (c, l)-diversity

$\ldots$ ($>$ 100 related approaches)

Other approaches

l-diversity : each group must have at least l distinct values

probabilistic l-diversity : frequency of the most frequent value in a class is bounded by $1/l$

entropy l-diversity : entropy of distribution of values inside a class is at least $\log(l)$

recursive (c, l)-diversity

$\ldots$ ($>$ 100 related approaches)

- hard to achieve
- underkill/overkill

Fatal flaws of privacy by syntactic transformation of data

- insecure against attackers with too much background info
- no composition
- no meaningful definitions for privacy and utility
- no mathematic guarantees of protection.

Fatal flaws of privacy by syntactic transformation of data

- insecure against attackers with too much background info
- no composition
- no meaningful definitions for privacy and utility
- no mathematic guarantees of protection.

Privacy is **not** a property of the data.

- privacy depends on the analysis done on the data
- identity transformation

Differential Privacy

An analysis result should not change much when adding/removing a single tuple.

Differential Privacy

An analysis result should not change much when adding/removing a single tuple.

Each user should not be worse off by having its record in the database.

Differential Privacy

An analysis result should not change much when adding/removing a single tuple.

Each user should not be worse off by having its record in the database.

$$e^{-\epsilon} \leq \frac{Pr(\mathcal{A}(Q, D_1) = R)}{Pr(\mathcal{A}(Q, D_2) = R)} \leq e^{\epsilon}$$

Differential Privacy (2)

Add noise to analysis result.

Differential Privacy (2)

Add noise to analysis result.

- sensibility of result (query)
- the more sensible the result, the more noise needs to be added
- sensibility is *worst-case* measure
- sensibility is independent of data in database
- sensibility of *how many people have this disease?* is 1
- sensibility of *what's the average salary of employees* is very high (sum, max, min, . . .)

Differential Privacy (3)

How to define sensibility?

Differential Privacy (3)

How to define sensibility?

- what can we publish?

Differential Privacy (3)

How to define sensibility?

- what can we publish?

  YES average height
  NO individual height

Differential Privacy (3)

How to define sensibility?

- what can we publish?

YES	average height
NO	individual height

- add 1m to height of one person: what changes?

Differential Privacy (3)

How to define sensibility?

- what can we publish?

  YES average height

  NO individual height

- add 1m to height of one person: what changes?

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

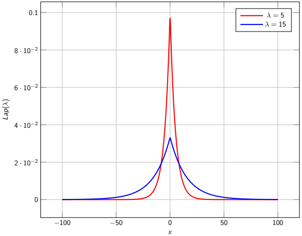
Differential Privacy (4)

Laplace mechanism

Differential Privacy (4)

Laplace mechanism

$$\begin{aligned} \tilde{x} &= x + Lap(\lambda) \\ Lap(\lambda) &= \frac{1}{2\lambda}\exp(-\frac{|x|}{\lambda}) \\ \lambda &= \frac{\Delta(f)}{\epsilon} \end{aligned}$$



Laplace mechanism :: example

- How many users viewed more than 10 movies?

Laplace mechanism :: example

- How many users viewed more than 10 movies?
- sensibility: $\Delta(f) = 1$

Laplace mechanism :: example

- How many users viewed more than 10 movies?
- sensibility: $\Delta(f) = 1$
- actual result: $x = 42$

Laplace mechanism :: example

- How many users viewed more than 10 movies?
- sensibility: $\Delta(f) = 1$
- actual result: $x = 42$
- $\epsilon = 0.1$, $\lambda = 10$

Laplace mechanism :: example

- How many users viewed more than 10 movies?
- sensibility: $\Delta(f) = 1$
- actual result: $x = 42$
- $\epsilon = 0.1$, $\lambda = 10$
- (possible) output: $\tilde{x} = 37$ (noise -5)

Differential Privacy (5)

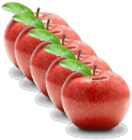
Exponential mechanism

Differential Privacy (5)

Exponential mechanism

- Laplace mechanism works for numerical data
- Exponential mechanism works for categorical data
- each item has a quality function $q(x)$
- randomly output item with probability $\sim \exp(\frac{q(x)}{\lambda})$
- $\lambda = \frac{2\Delta(q)}{\epsilon}$

Exponential mechanism :: example



Could set the price of apples at $1.00 for profit: $4.00

Could set the price of apples at $4.01 for profit $4.01

Best price: $4.01
2nd best price: $1.00
Profit if you set the price at $4.02: $0
Profit if you set the price at $1.01: $1.01

Differential Privacy (6)

Composability

Differential Privacy (6)

Composability

sequential composition $\epsilon_t = \epsilon_1 + \epsilon_2 + \ldots + \epsilon_k$

parallel composition $\epsilon_t = \max\{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$

Differential Privacy (7)

How to use the mechanisms?

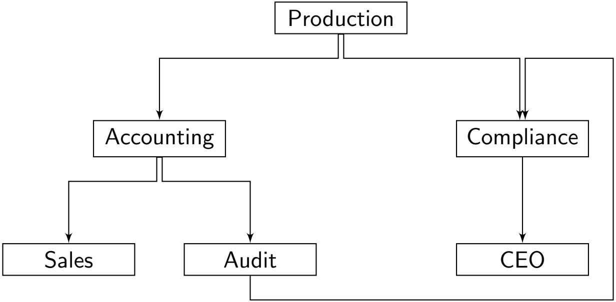
Differential Privacy (7)

How to use the mechanisms?

- using them directly gives not so good results
- composability properties help
- we can generate synthetic data and apply all algorithms on that
- we can interleave dp mechanisms with the original data-mining algorithm
- optimization problems

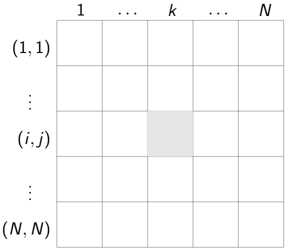
Workflow

Paths in organisation.



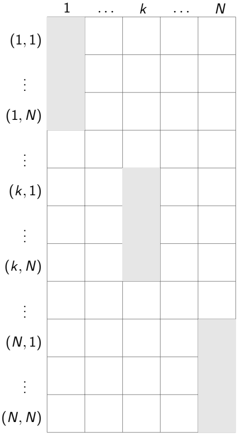
Workflow (2)

How many docs go through $i \to j \to k$



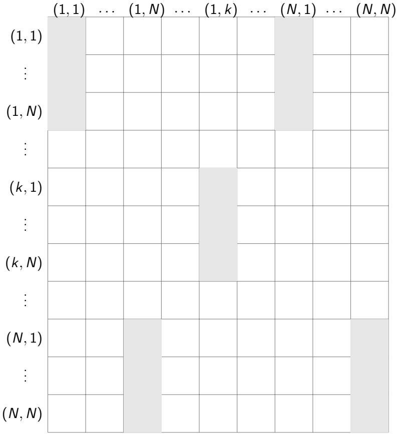
Workflow (3)

How many triangle paths $(i \to j \to i)$



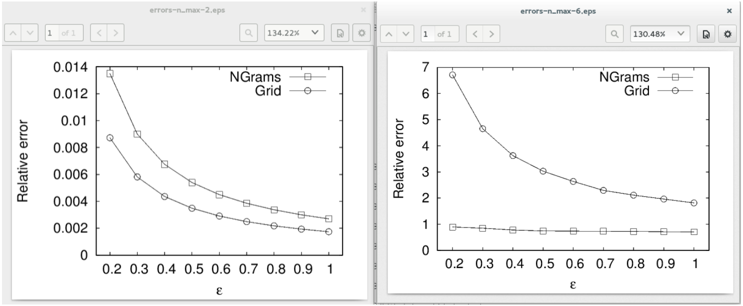
Workflow (4)

How many returned docs $(i \to \ldots j \to \ldots i)$



Workflow (5)

- n-gram model for higher dimensionality
- integrity constraints



Differential Privacy (8)

Limitations

- results are worse for higly-corellated data
- no extensions for complex models
- how to properly set ϵ
- expensive computations
- error bounds
- no direct relationship between utility and privacy
- inconsistencies