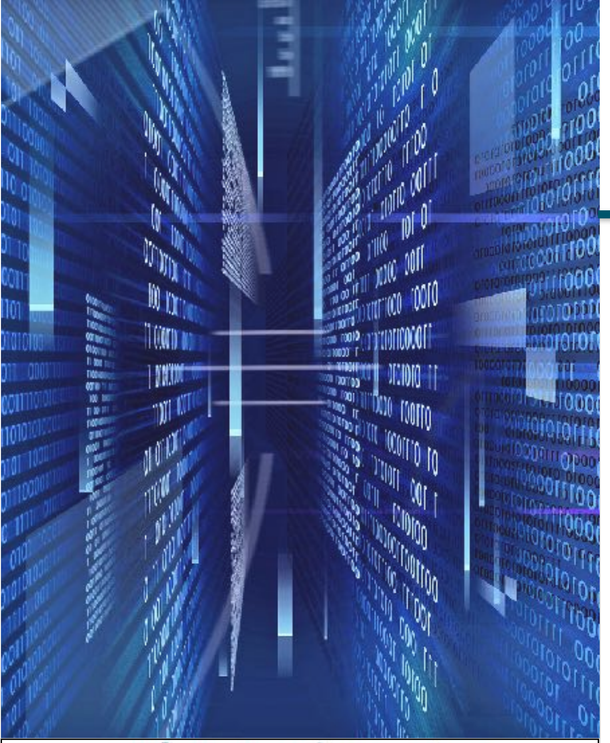


cybersecurity centre
GCHQ Academic Centre of Excellence EPSRC

UNIVERSITY OF Southampton

COMP6224(16)
week 3: Trust, Identity & Privacy

Vladimiro Sassone

Cyber Security Centre

University of Southampton

Cybersecurity Southampton

CyberSecuritySoton.org [w]

@CybSecSoton [fb & tw]

Part I

The identity and privacy challenge

trust, identity and privacy

Trust is crucial in pervasive global computing; yet its future is conditioned by the emergence of crime, scams, abuse, aggressive profiling, loss of privacy, etc

trust, identity and privacy

Trust is crucial in pervasive global computing; yet its future is conditioned by the emergence of crime, scams, abuse, aggressive profiling, loss of privacy, etc

Whilst we cannot give up ubiquitous computing, the trust and reliance we put on it comes under scrutiny. We (ought to) feel increasingly threatened by the future we are shaping.

trust, identity and privacy

Trust is crucial in pervasive global computing; yet its future is conditioned by the emergence of crime, scams, abuse, aggressive profiling, loss of privacy, etc

Whilst we cannot give up ubiquitous computing, the trust and reliance we put on it comes under scrutiny. We (ought to) feel increasingly threatened by the future we are shaping.

Let us focus on identity and privacy as key components of users' trust in future computing, as well as hard research topics.

computational trust

Trust is also a computational technique to achieve cooperation in highly decentralised environment, where no established authority can be relied upon.

Trust is also a computational technique to achieve cooperation in highly decentralised environment, where no established authority can be relied upon.

Eg, rating/reputation systems:

Many applications exist for (electronic) agents to build and use trust and reputation systems

Customer Reviews

1 Review

5 star: (0)
4 star: (1)
3 star: (0)
2 star: (0)
1 star: (0)

Average Customer Review
★★★★☆ (1 customer review)

Most Helpful Customer Reviews

2 of 2 people found the following review helpful:
★★★★☆ **Predictocracy Review**, February 27, 2008
By **Alexander Kirtland "UsableMarkets"** (Brooklyn, NY Unite
reviews
REAL NAME™
Michael Abrahmowicz, in his new book, Predictocracy: Market

Trust is also a computational technique to achieve cooperation in highly decentralised environment, where no established authority can be relied upon.

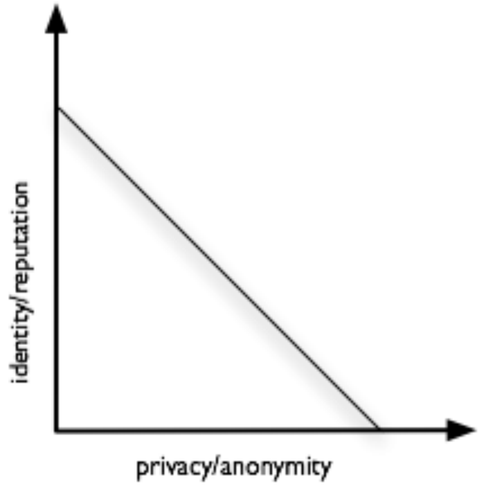
Eg, rating/reputation systems:

Many applications exist for (electronic) agents to build and use trust and reputation systems

So, why not to build a *global distributed trust & reputation* engine?
Identity and privacy are a challenge...

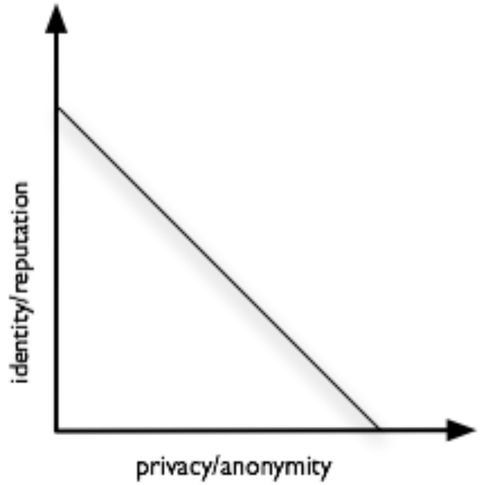
identity vs reputation

Reputation relies the observed behaviour associated to an identity, which is in direct tension with anonymity and privacy



identity vs reputation

Reputation relies the observed behaviour associated to an identity, which is in direct tension with anonymity and privacy



To use multiple pseudo identities to trade-off trust and privacy is not a solution.

Technical problem: to keep even two credible identities distinct is too high an effort, they'll soon be linked.

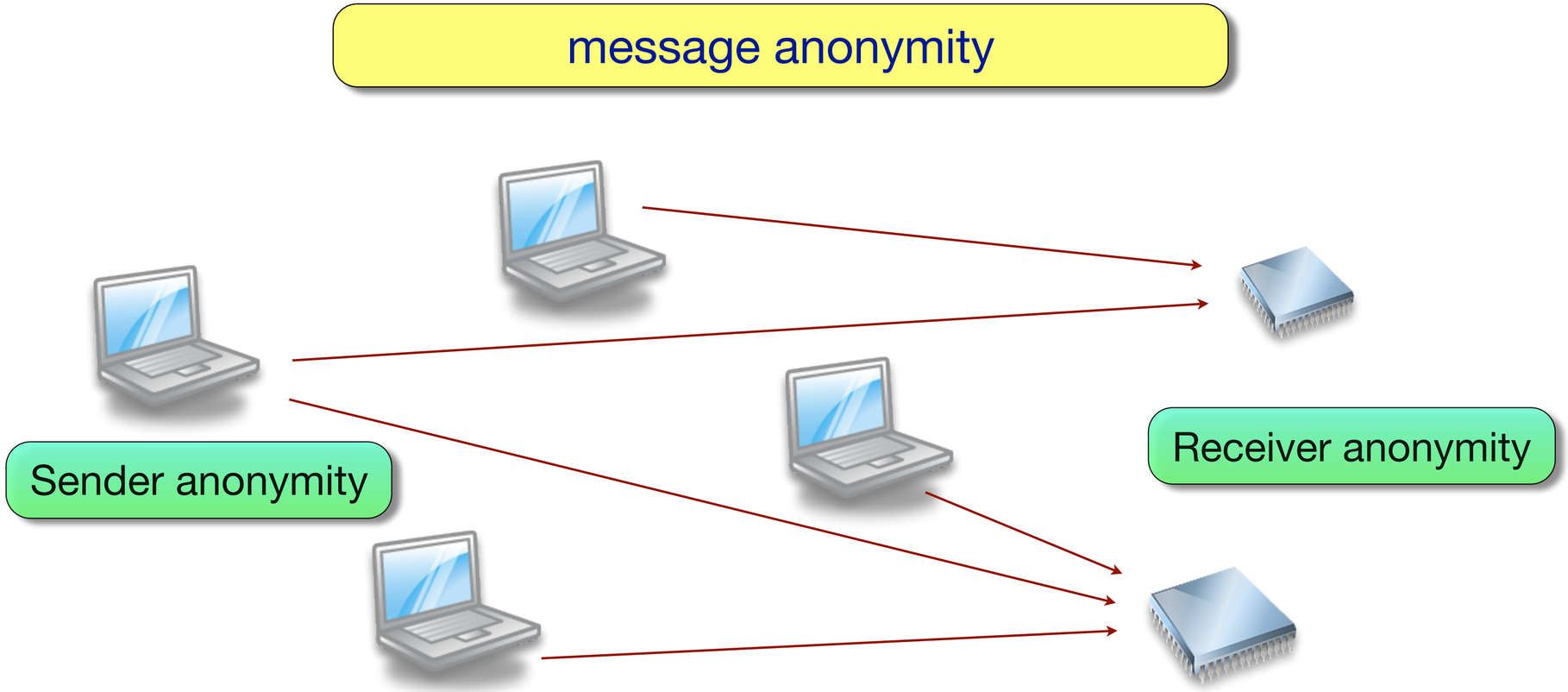
Social problem: multiple discardable ids are not conducive of accountable behaviour.

specialised privacy is possible

You can stay anonymous in some specific systems. Eg:

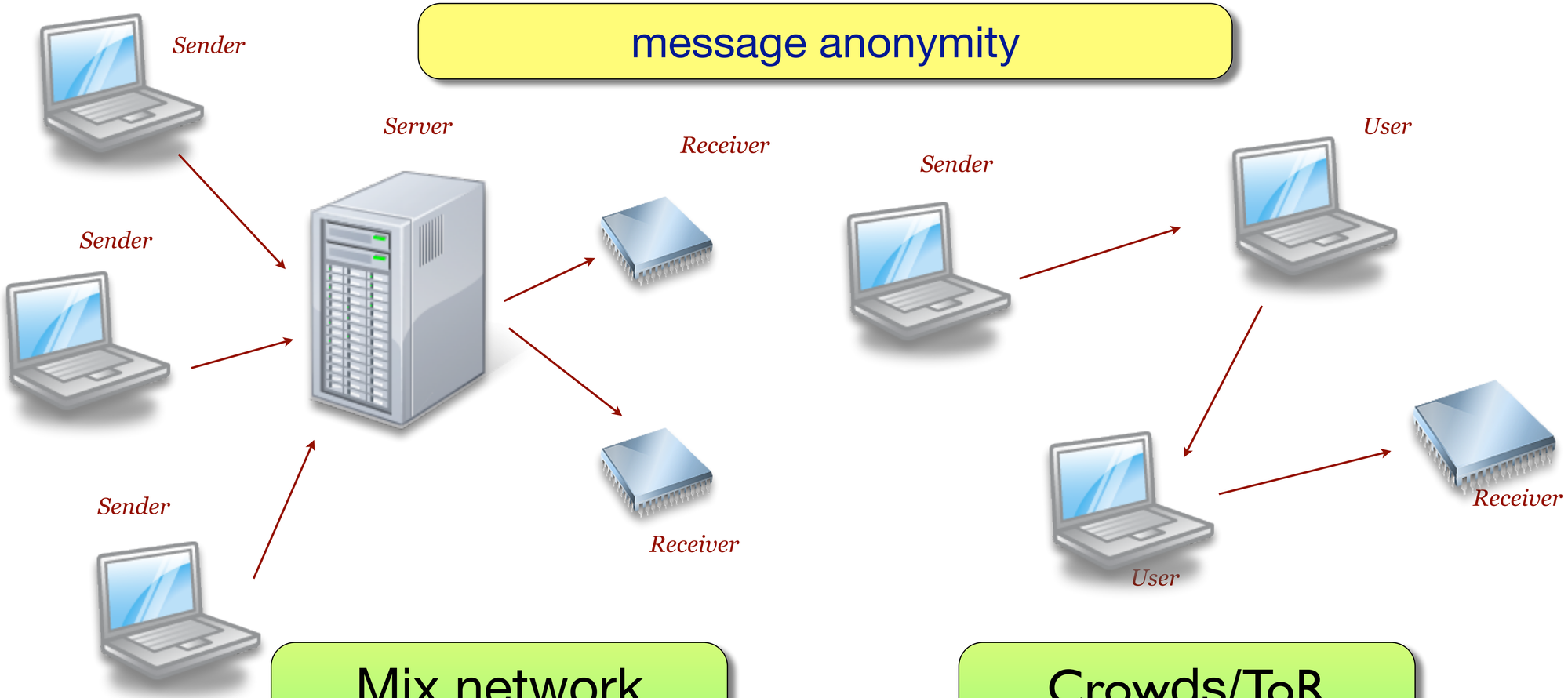
specialised privacy is possible

You can stay anonymous in some specific systems. Eg:



UNIVERSITY OF Southampton

You can stay anonymous in some specific systems. Eg:



specialised privacy is possible

You can stay anonymous in some specific systems. Eg:

zero-knowledge proof authentication

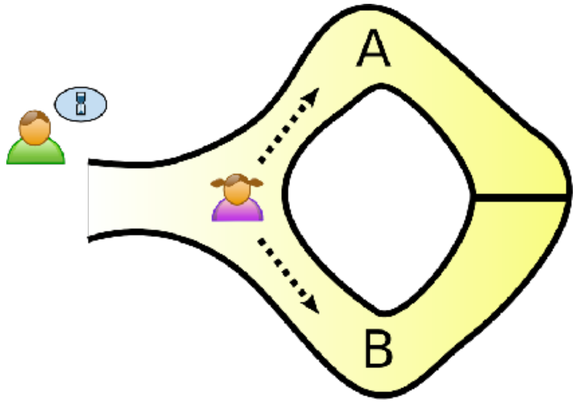
ZKP: crypto-technique that yield nothing but the validity of an assertion: eg "*i am a registered user of this system*"

specialised privacy is possible

You can stay anonymous in some specific systems. Eg:

zero-knowledge proof authentication

ZKP: crypto-technique that yield nothing but the validity of an assertion: eg "*i am a registered user of this system*"

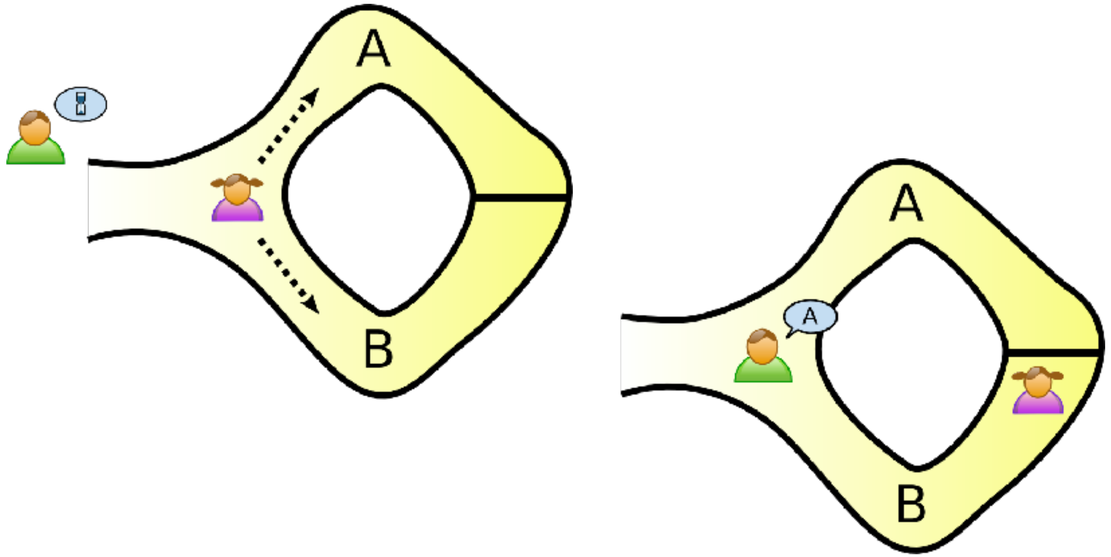


specialised privacy is possible

You can stay anonymous in some specific systems. Eg:

zero-knowledge proof authentication

ZKP: crypto-technique that yield nothing but the validity of an assertion: eg "*i am a registered user of this system*"

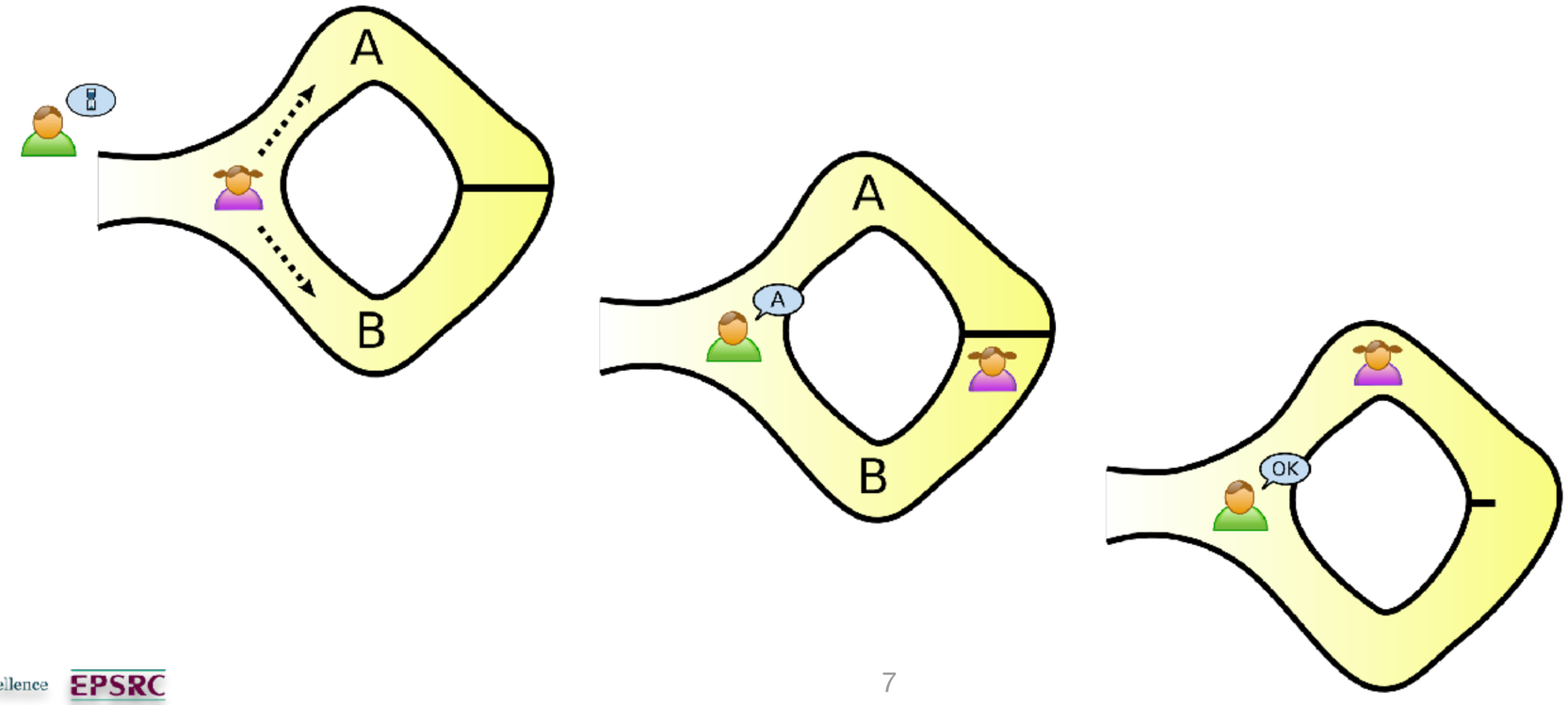


specialised privacy is possible

You can stay anonymous in some specific systems. Eg:

zero-knowledge proof authentication

ZKP: crypto-technique that yield nothing but the validity of an assertion: eg "*i am a registered user of this system*"



Part II

Privacy

threats to privacy

Web cookies have tracked your browser for a long time. Much worse, big data means that virtual and real identities are now being linked, and its essentially impossible to defend against it.

threats to privacy

Web cookies have tracked your browser for a long time. Much worse, big data means that virtual and real identities are now being linked, and its essentially impossible to defend against it.

Browsing is changing: websites know who you are and FB can see you on any page with a “like” button. All your activities (not just IDs) are linked, unless you have a spy-like discipline.

threats to privacy

Web cookies have tracked your browser for a long time. Much worse, big data means that virtual and real identities are now being linked, and its essentially impossible to defend against it.

Browsing is changing: websites know who you are and FB can see you on any page with a “like” button. All your activities (not just IDs) are linked, unless you have a spy-like discipline.

Profiling is worth $15b/year in USA only; BlueKai Exchange markets profiles of 300+m people; LeadPlease.com sold for $260 personal data on 1000 cancer patients (source FT).

threats to privacy

Web cookies have tracked your browser for a long time. Much worse, big data means that virtual and real identities are now being linked, and its essentially impossible to defend against it.

Browsing is changing: websites know who you are and FB can see you on any page with a “like” button. All your activities (not just IDs) are linked, unless you have a spy-like discipline.

Profiling is worth $15b/year in USA only; BlueKai Exchange markets profiles of 300+m people; LeadPlease.com sold for $260 personal data on 1000 cancer patients (source FT).

The explosion of high-quality, aggregate cheap data is making privacy a notion from the past...

recent privacy-related technical research

notorious case of NetFlix competition, opened the arena;

you can be pinpointed which accuracy by approx location of home and work: 5% can be even if approx just by county;

your future location can be predicted in the long term: impressively, with 80% accuracy where you'll be in 80 days;

4 data points about a phone's position can link it to a person;

smart meters with 2 readings/sec can be used to know what you're watching on TV (luckily UK plans 1 reading in 15mins);

crowd sourcing can locate people anywhere from a photo: US State Dept challenge: 3/5 targets found in 3:20, 7:20, 11 hours;

Bitcoin transactions de-anonymised linking together wallets;

Ongoing work on identifying people from free text.

approaches to privacy

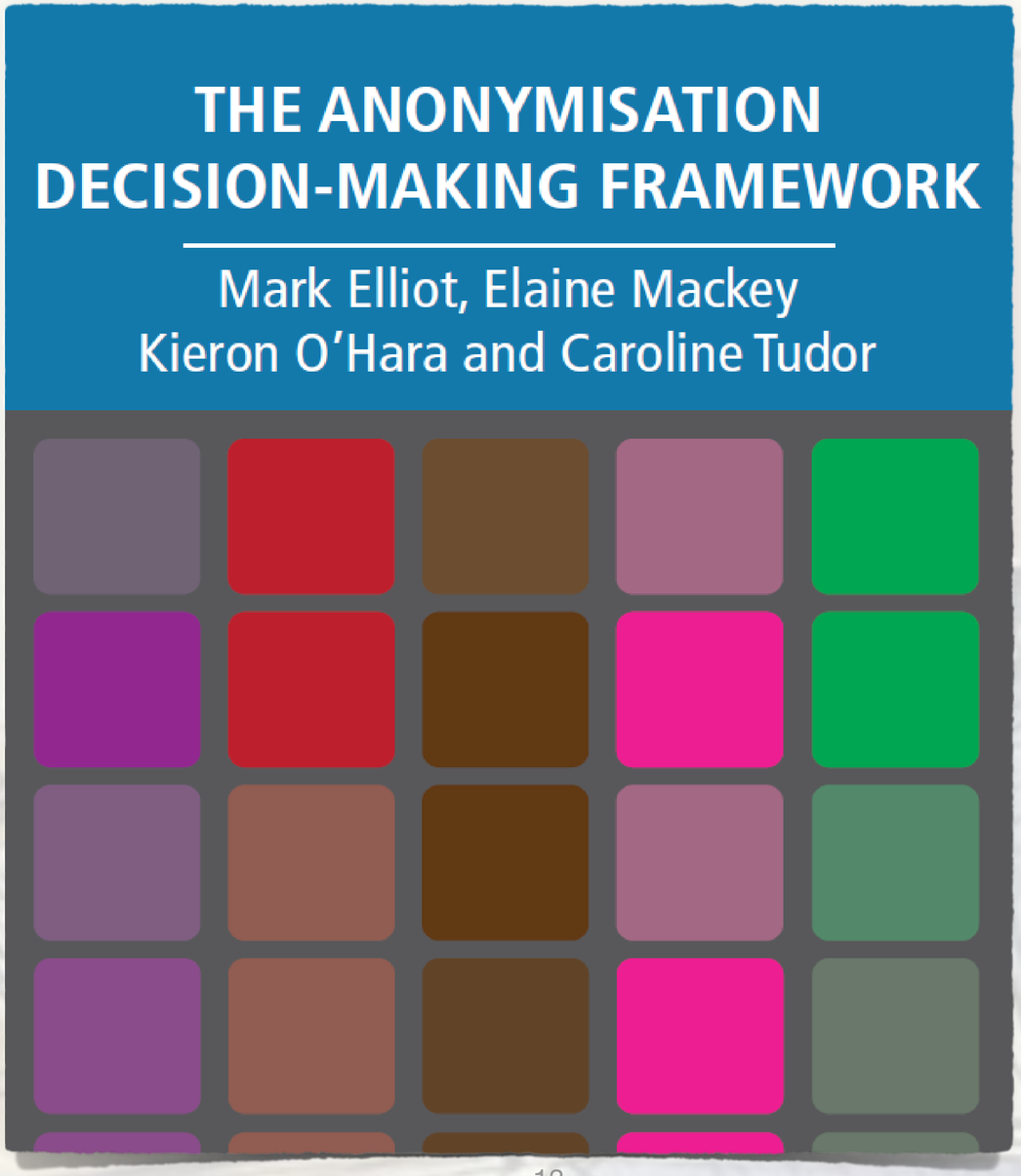
Hard: seek mathematical guarantees against attackers as powerful as (computational) nature allows. *This is hopeless!*

Soft: see problem as not *only* technical, but regulatory, legal and social too. Eg, FB privacy is expressive enough for users, point is they track you and sell your data. But this can be regulated rather than countered in purely technical terms.

This leaves the door open to powerful malicious parties, but defends against big players --those which hurt!-- casual and accidental offenders. Many interesting and effective ideas can be explored by multidisciplinary approaches: define meaning and ownership of personal data; introduce users' data reuse and expiration policies; ...

Part III

Anonymisation


THE ANONYMISATION
DECISION-MAKING FRAMEWORK
Mark Elliot, Elaine Mackey
Kieron O'Hara and Caroline Tudor

starting point: ADMF guidelines

1. know your data and its origins
2. understand the use case
3. understand the legal issues and pre-share/release governance
4. understand the issue of consent and your ethical obligations
5. know the processes you will need to go through to assess the risk of de-identification
6. know the processes you will need to go through to anonymise your data
7. understand the data environment
8. know your audience and how you will communicate
9. know what to do if things go wrong
10. what happens next once you have shared and or release data

anonymity

not just a property of the data

- it depends also on other information which may come into the possession of the attacker

it is then a relation between the data and a data environment

- providing context for the anonymised data
- consisting of infrastructure, processes, governance, agents (skills, motivations) and auxiliary data

that is, anonymity only makes sense within a context

hence, the risk of de-anonymisation is always > 0

- approach: raise the cost of re-identification above the its benefits

risk factors independent of the data

highly motivated attackers

consequences of disclosure (is this achievable with other means?)

data governance (who gets to see the data, under what conditions?)

provenance and other metadata

other available data

time series, open data, commercial data, data in the same domain

data quality

…

responsibilities of data controllers

understand how a privacy breach may occur

understand the possible consequences

address the risk of a breach occurring

- what do you do when it does? how do you manage incidents?

understand the environment

never *release-and-forget*

- anonymising is an ongoing commitment, a context-dependent process; anonymous now or here, does not imply anonymous tomorrow or there

1. you cannot decide whether data is anonymous by looking at the data
2. you cannot decide whether data is anonymous without looking at the data
3. anonymisation aims at producing data that is useful (as well as safe)
4. “zero risk” is not an option
5. anonymisation methods should be proportional to the risk

UNIVERSITY OF Southampton

Anonymisation: managing data protection risk code of practice

Annex 3 – Practical examples of some anonymisation techniques

– drawn up by Mu Yang, Vladimiro Sassone and Kieron O'Hara at the University of Southampton

ico. Information Commissioner's Office

data anonymisation techniques

- A — data reduction
- B — data perturbation
- C — non perturbative

UNIVERSITY OF Southampton

Removing variables

Example one: the removal of direct identifiers

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month	Ethnic
22	F	SO17	£20,000	£1,100	British
25	M	SO18	£22,000	£1,300	Irish
30	M	SO16	£32,000	£1,800	African
35	F	SO17	£31,500	£2,000	Chinese
40	F	SO15	£68,000	£3,500	Pakistani
50	M	SO14	£28,000	£1,200	British

Description: A variable is a characteristic or attribute of an individual – for each individual the variable will have a value (eg the values of the variable NAME for the three authors of this appendix are Mu, Vladimiro and Kieron). The simplest method of anonymisation is the removal of variables which provide direct or indirect identifiers from the data file. These need not necessarily be names; a variable should be removed when it is highly identifying in the context of the data and no other protection methods can be applied.

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month	Ethnic
22	F	SO17	£20,000	£1,100	British
25	M	SO18	£22,000	£1,300	Irish
30	M	SO16	£32,000	£1,800	African
35	F	SO17	£31,500	£2,000	Chinese
40	F	SO15	£68,000	£3,500	Pakistani
50	M	SO14	£28,000	£1,200	British

a2. removing records

Removing records

Example two: the removal of a particular record which is easy to identify

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Description: Removing records of particular units or individuals can be adopted as an extreme measure of data protection when the unit is identifiable in spite of the application of other protection techniques.

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
~~30~~	~~M~~	~~SO16~~	~~£32,000~~	~~£1,800~~
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

UNIVERSITY OF Southampton

Global recoding

Example three: aggregating the values observed in variables into pre-defined classes

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Description: This method makes variable values less specific, and the table correspondingly less informative. For a categorical variable (ie one that categorises the units), several categories are combined to form new (less specific) categories, thus resulting in a new variable. A continuous variable is replaced by another variable which aggregates ranges of the continuous variable. In other words, the global recoding method consists in aggregating the values observed in a variable into pre-defined classes. Every record in the table is recoded.

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17-19	low	low
25-29	M	SO17-19	low	low
30-34	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	M	SO14-16	medium	low

A more informative type of recoding involves recoding only the outliers (i.e. unusually high or unusually low values). For instance, incomes between, say £20,000 and £60,000 would be reproduced in the recoded table, but outside that range would be recoded as <£20,000 or >£60,000. This type of recoding leaves the vast majority of 'normal' values unchanged.

a4. suppression

Local suppression

Example four: replacing the observed value of one or more variables in a certain record with a missing value

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17	low	low
25-29	M	SO18	low	low
30-34	M	SO16	medium	medium
35-39	F	SO17	medium	medium
40-44	F	SO15	high	high
50-54	M	SO14	medium	low

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
missing	F	SO17	low	low
25-29	M	SO18	low	low
30-34	M	SO16	medium	medium
35-39	F	SO17	medium	medium
40-44	F	SO15	high	high
50-54	M	SO14	medium	low

Unique combination

Description: Local suppression consists of replacing the observed value of one or more variables in a certain record with a 'missing' value. This is particularly suitable with categorical key variables (a key variable is a variable that a researcher is particularly interested in). When combinations of such variables are problematic, local suppression consists of replacing an observed value with 'missing' or some other value which shows that the original value has been suppressed. The aim of the method is to reduce the information content of rare combinations. The result is an increase in the frequency count of records containing the modified combination.

b1. micro-aggregation

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
50	M	S014	£28,000	£1,200
35	F	SO17	£31,500	£2,000
30	M	SO16	£32,000	£1,800
40	F	S0O15	£68,000	£3,500

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/ month
22	F	SO17	£23,333	£1,100
25	M	SO18	£23,333	£1,300
50	M	S014	£23,333	£1,200
35	F	SO17	£43,833	£2,000
30	M	SO16	£43,833	£1,800
40	F	S0O15	£43,833	£3,500

k partition = 3

Description: The idea of micro-aggregation is to replace an observed value with the average computed on a small group of units. The units belonging to the same group will be represented in the released file by the same value. The groups contain a minimum predefined number *k* of units. Here *k* is a threshold value and the partition is called a *k-partition*. In order to obtain micro-aggregates from a dataset with a certain number of records, these records are combined (usually in a meaningful order, such as size order) to form groups of size at least *k*. We do this by computing the average value of the target variable over each group and then replacing the original values with this average value. The mean value for the whole population remains unchanged.

So, for example, if we had 100 individuals in the dataset and wished to form a 4-partition, then segment the dataset into 25 groups of 4. For each group, the average value of the variable is computed, and that average replaces the observed value in the dataset. If a group of 4 individuals had ages 31, 33, 33 and 34, the age for each individual in the published dataset would be 32.75.

b2. data swapping

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17-19	low	low
25-29	M	SO17-19	low	low
30-34	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	M	SO14-16	medium	low

value unique

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
35-39	F	SO17-19	low	low
25-29	M	SO17-19	low	low
25-29	M	SO14-16	medium	medium
20-24	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	F	SO14-16	medium	low

Swapped attribute is Age.

Swapping rate: r=33.3%*.

Constraints: only allow swaps of Age between records with the same value of Gender

* The rate r is typically in the range of 1-10%. We choose 33.3% because of the limited number of records

Description: Data swapping alters records in the data by switching values of variables across pairs of records in a fraction of the original data. The purpose is to introduce uncertainty for a data user or intruder as to whether records correspond to real data elements.

The variables that will be swapped are called *swapped attributes* or *swapping attributes* and the fraction of the total *n* records in the microdata that are initially marked to be swapped is called the *swap rate*, and is denoted by *r*. Typically, *r* is of the order of 1-10% (so that the fraction of attributes swapped will usually be less than one in ten).

Post-Randomisation Method (PRAM)

Example seven: producing a microdata file in which the scores on some categorical variables for certain records in the original file are changed into a different score according to a prescribed probability mechanism

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	M	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	F	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

target variable = Gender, the PRAM-matrix: p11=p22=0.9, p12=p21=0.1

Description: The Post-Randomisation Method is a probabilistic method to perturb categorical variables. In the released file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a probability mechanism called a *Markov matrix*. This is quite a complex method, which is somewhat difficult to describe in straightforward language.

Suppose we have a categorical variable *V* which we wish to perturb, and suppose that that variable has *K* categories (so, for example, 'sex' is a categorical variable with two categories). For that variable *V*, we can decide to change one of the *K* values to another with a certain probability fixed in advance; we can arrange these probabilities in a *K* x *K* matrix (the Markov matrix), where, say, the second cell in the fourth row is the transition probability that when we have a value in the fourth category in the observed data, we transform it into the value of the second category in the published data. We can then decide to transform or perturb the data or not, depending on a random process. So, for instance, if our categorical variable was 'sex', and all the probabilities in the 2 x 2 Markov matrix were 0.5, we could toss a coin each time to decide whether or not to alter the attribution of M or F to each individual in the data.

b4. adding noise

Adding noise

Example eight: adding a random value ε, with zero mean and predefined variance σ^2, to all values in the variable to be protected

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Standard Normal Distribution: mean=0, variance=1

-0.171932015
1.862281351
0.959896624
-2.543129085
-1.049088496
-0.308324388

x 1000

-£172
£1,862
£960
-£2,543
-£1,049
-£308

Income & Expenses Individual-level dataset

Age	Gender	Postcode	Income	Expenses/month
22	F	SO17	£19,828	£1,100
25	M	SO18	£23,862	£1,300
30	M	SO16	£32,960	£1,800
35	F	SO17	£28,957	£2,000
40	F	SO15	£66,951	£3,500
50	M	SO14	£27,692	£1,200

Description: Adding noise, a method applied to numerical data, consists of adding a random value ε to all values in the variable to be protected. The distribution of ε has mean zero and predefined variance σ^2. In other words, the expected value of ε is zero (sometimes the value will be positive, sometimes negative), so that given that noise is added to enough values the additions will cancel themselves out, leaving the mean of the distribution unchanged. The variance defines the range of the additional ε; a small variance means that ε is unlikely to be very far from 0 (and so the numerical change in the data unlikely to be large in any instance), while a larger variance will allow greater perturbations of individual data values. This type of distribution, a *normal distribution*, is the most standard type of distribution in statistics, very well-understood and often encountered in practice with real-world data.

b5. resampling

Example nine: drawing with replacement t samples of n values from the original data, sorting the sample and averaging the sampled values

Income Individual-level dataset

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£20,000	£23,000
25	M	SO18	£22,000	£22,000
30	M	SO16	£32,000	£30,000
35	F	SO17	£31,500	£35,000
40	F	SO15	£68,000	£58,000
50	M	SO14	£28,000	£29,000

Income Individual-level dataset

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£58,000	£58,000
25	M	SO18	£22,000	£20,000
30	M	SO16	£29,000	£20,000
35	F	SO17	£20,000	£68,000
40	F	SO15	£22,000	£30,000
50	M	SO14	£20,000	£31,500

Income Individual-level dataset with ordered mapping

Age	Gender	Postcode	Income (Jan)	Income (Feb)
22	F	SO17	£20,000	£20,000
25	M	SO18	£20,000	£20,000
30	M	SO16	£29,000	£31,500
35	F	SO17	£22,000	£58,000
40	F	SO15	£58,000	£68,000
50	M	SO14	£22,000	£30,000

Normal distributed resampling.
Hypothesis testing: Null Hypothesis

Grand mean: £33, 208

Description: Resampling is also designed for numerical data, and again requires understanding of statistical methods. It has three steps. First, we have to identify the way that the sensitive or key data variables vary across the whole population. This means deciding what the population will look like if put on a graph; typically, the answer will be a type of reasonably well-known type of distribution.

The second step is to generate a distorted sample artificially which has the same parameter values as our estimate. The sample should be the same size as the database.

The third step is to replace the confidential data in the database with the distorted sample. So, in the salary example, if we have 100 lines in our dataset, and having decided how salaries vary across the population, we generate an artificial distribution of 100 salaries that has the same mean and variance as the estimate for the whole population. We then substitute those 100 artificially generated salaries for the 100 observed salaries in the database.

c1. sampling

Description: Sampling is one of the non-perturbative methods in anonymisation techniques, suitable when the original data is in sufficient quantity to make a sample meaningful. Instead of publishing the original microdata file, we take a sample from it and publish that without identifiers. The resulting sample may contain information which is sensitive and which in other circumstances could be quite disclosive. However, because there is no way of knowing whether a particular individual's data is included in the sample, it is unlikely, though not impossible, that it would actually be disclosive.

Two common types of sampling are simple random sampling, where all possible subsets of specified size sample have an equal probability of selection, and Bernoulli sampling, where each record in the sample is selected independently with a certain probability.

The probability that a random sample preserves the basic statistical properties of the original dataset can be calculated.

Cross-tabulation of data

Example ten: generating the contingency table which does not contain the individual information.

Record NO.	Gender	Education Level
1	F	Undergrad Degree
2	M	Grad Degree
3	M	Doctorate
4	F	Doctorate
5	F	Doctorate
6	M	Undergrad Degree

	Undergrad Degree	Grad Degree	Doctorate	Total
Female	1	0	2	3
Male	1	1	1	3
Total	2	1	3	6

Description: When we have a table of data with two or more variables, we can create another table by tabulating the two variables against each other direction, in effect aggregating the data. The resulting table is called a contingency table. It can protect the confidentiality in microdata, especially for large numbers, and is non-perturbative.

Part IV

Differential Privacy

Naive desiderata (dalenius 1977): access to [anonymized data] should not enable to learn about an individual what could not be learned without access.

Naive desiderata (dalenius 1977): access to [anonymized data] should not enable to learn about an individual what could not be learned without access.

This cannot be done! Obstacle: external information

Famous example of the database of average national heights, proves impossibility and there is nothing you can do about it.

Naive desiderata (dalenius 1977): access to [anonymized data] should not enable to learn about an individual what could not be learned without access.

This cannot be done! Obstacle: external information

Famous example of the database of average national heights, proves impossibility and there is nothing you can do about it.

Differential Privacy (relativised disclosure)
any disclosure about an individual is "*almost equally*"
likely whether or not he/she is part of the dataset

Naive desiderata (dalenius 1977): access to [anonymized data] should not enable to learn about an individual what could not be learned without access.

This cannot be done! Obstacle: external information

Famous example of the database of average national heights, proves impossibility and there is nothing you can do about it.

Differential Privacy (relativised disclosure)
any disclosure about an individual is "*almost equally*"
likely whether or not he/she is part of the dataset

Observe that bad breaches can still happen, it's just that the risk of this happening is beyond the dataset's controllers.

Naive desiderata (dalenius 1977): access to [anonymized data] should not enable to learn about an individual what could not be learned without access.

This cannot be done! Obstacle: external information

Famous example of the database of average national heights, proves impossibility and there is nothing you can do about it.

Differential Privacy (relativised disclosure)
any disclosure about an individual is "*almost equally*"
likely whether or not he/she is part of the dataset

Observe that l[illegible]happen, it's just that the risk of this happening is beyond the dataset's controllers.

also watch this out!

implications for anonymisation & opendata

Research has taken to differential privacy, asking questions like *"can computation X be done with differential privacy?"*. Eg, recommender, reputation, DB query, location, itemised billing...

implications for anonymisation & opendata

Research has taken to differential privacy, asking questions like *"can computation X be done with differential privacy?"*. Eg, recommender, reputation, DB query, location, itemised billing...

Lots of questions still open. One thing however is quite clear:

"anonymise & release" does not work!

it is impossible to anonymise entire datasets without knowing what the data will be used for, and anyway in the long term...

implications for anonymisation & opendata

Research has taken to differential privacy, asking questions like *"can computation X be done with differential privacy?"*. Eg, recommender, reputation, DB query, location, itemised billing...

Lots of questions still open. One thing however is quite clear:

"anonymise & release" does not work!

it is impossible to anonymise entire datasets without knowing what the data will be used for, and anyway in the long term...

Research should focus on:

1) what can and cannot be published as open data?
2) can we do "*semi-open*" data with diff privacy?

And again, the solution cannot be purely technical.

differential privacy in statistical databases

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

Queries we would like to permit

How many people have the disease ?
Average age and weight of men who have the disease ?

aggregate

Queries that are dangerous for the privacy

Does Daniele Bianchi have the disease?
What is the name of the last record inserted in the database?
What is the age / weight of the last record inserted in the database?

individual

Problem

The restriction to aggregate queries is not sufficient: also these queries may leak information about individuals !

Differential privacy in statistical databases

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

Differential privacy in statistical databases

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

insertion of a new record

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

Differential privacy in statistical databases

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

How many men have the disease ? 2

What is the average age / weight of men who have the disease ? 50 / 101

insertion of a new record

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

Differential privacy in statistical databases

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

insertion of a new record

Name/Id	age	weight	sex	disease	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
Sergio Neri	20	140	M	yes	...

How many men have the disease ? 2

What is the average age / weight of men who have the disease ? 50 / 101

How many men have the disease ? 3

What is the average age / weight of men who have the disease ? 40 / 114

We can deduce the exact age / weight of the new record

Typical solution

Add some noise to the answer.

Namely:

given the set of possible databases, X

given a query with exact answer $f : X \to Y$,

the curator gives instead an approximate answer using some mechanism $K : X \to Z$

The idea is that a small perturbation of the global info produces a large perturbation of the individual infos

Privacy and Utility

The two main criteria to judge a randomised mechanism:

Privacy: how good is the protection against leakage of private information

Utility: how useful is the reported answer

Clearly there is a trade-off between privacy and utility, but they are not the exact opposites: privacy refers to the individual data, utility refers to the aggregated data.

The definition of differential privacy is the following:

A randomised function $K : X \rightarrow Z$ is ε-differentially private if for all adjacent databases x, x' (i.e. databases that differ only for one individual) and for all $S \subseteq Z$, we have

For discrete answers the following property is equivalent :

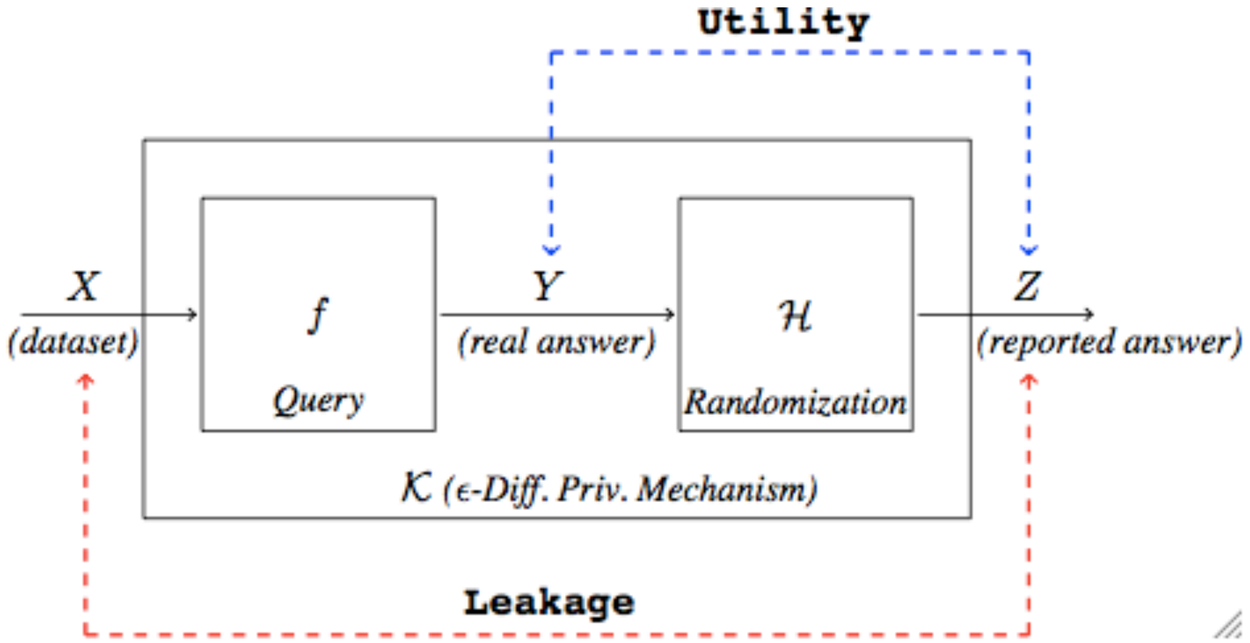
$$p(Z \in \mathcal{S}|X = x) \leq e^{\epsilon}\, p(Z \in \mathcal{S}|X = x')$$

$$\frac{p(Z = z|X = x)}{p(Z = z|X = x')} \leq e^{\epsilon}$$

oblivious mechanisms

Given $f : X \to Y$ and $K : X \to Z$, we say that K is oblivious if it depends only on Y (not on X)

If K is oblivious, it can be seen as the composition of f and a randomised mechanism H defined on the exact answers $K = H f$



Another reason why privacy and utility are not the exact opposite is that privacy concerns the information flow between the databases and the reported answers, while utility concerns the information flow between the correct answer and the reported answer

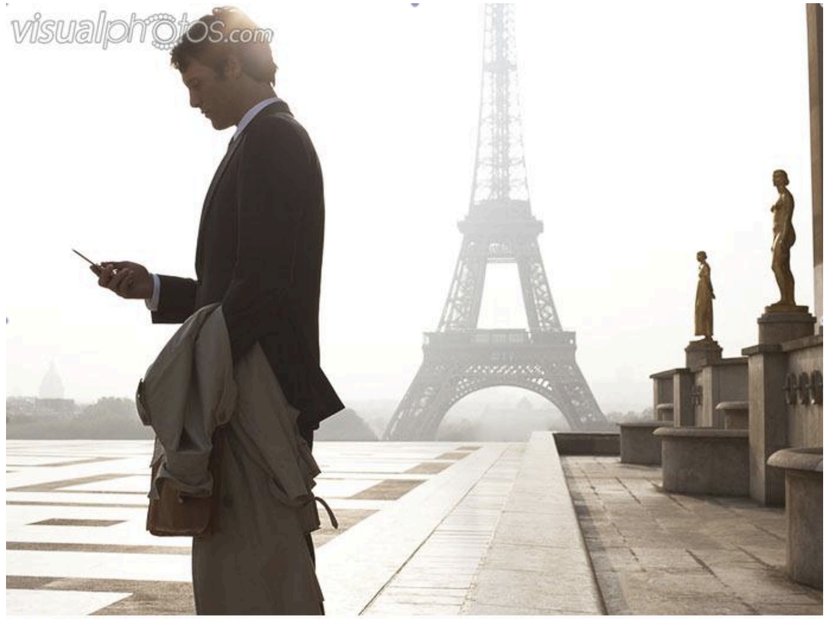
example: location based services

Use an LBS to find a restaurant

Without revealing the exact location

Revealing an approximate location is ok

geo-indistinguishability



d : the Euclidean distance

x : the exact location

z : the reported location

d-privacy:

$$\frac{p(x|z)}{p(x'|z)} \leq e^{\epsilon r} \frac{p(x)}{p(x')}$$