Decentralising Data Collection and Anonymisation

Steven Murdoch
University College London
An anonymisation system has architecture, security definition and mechanism

**Architecture**
How is data collected, stored, processed and distributed?

**Security definition**
What security and privacy properties must the system preserve, and what level of confidence is required?

**Mechanism**
How is access to data restricted or data altered in order to meet security definition?

In this presentation I’m going to show an alternative architecture to the traditional centralized data curator allowing the relaxation of the “plausible deniability” security property
A trusted, centralised architecture is common but creates risk and expense.
A trusted, centralised architecture is common but creates risk and expense

**Trusted**
In a position to violate security properties

**Data submitted to central authority is inherently sensitive**

To give data subjects confidence that their data will be protected, strong procedures and technical controls need to be in place, which is expensive in terms of effort, lost opportunity, and monetary cost
A trusted, centralised architecture is common but creates risk and expense. Census Bureau Sensitive data from data subjects. Anonymous data Requests from analysts. Trusted in a position to violate security properties. Data submitted to central authority is inherently sensitive. To give data subjects confidence that their data will be protected, strong procedures and technical controls need to be in place, which is expensive in terms of effort, lost opportunity, and monetary cost.
Within the centralised architecture, security definition and mechanism may vary

Data released could be noninteractive (anonymised version of original data, aggregate statistics), or interactive (results of a query interface)
To me, differential privacy may be as authoritarian in its conceptual underpinnings as [Identity Based Cryptography]

The Moral Character of Cryptographic Work, Phillip Rogaway

Architecture, security definition and mechanism are linked but often conflated. Authoritarian tendencies come from trusted central authority, not security definition. Criticism above applies mainly to interactive centralized architecture.
Confidence in institutions is at an all-time historical low

% who trust the govt in Washington always or most of the time

Pew Research Center
If data is anonymised by data subject, the central authority is no longer trusted.
Google RAPPOR uses randomised response to prevent submissions being sensitive

**Architecture**
Untrusted central authority

**Security definition**
Differential privacy

**Mechanism**
Randomised response

Are you a member of the Communist Party?
1. Toss a fair coin
2. If heads, say “Yes”
3. If no, tell the truth

Gives *plausible deniability for individuals* but allows accurate aggregate statistics, and the uncertainty, to be calculated

RAPPOR extends this approach to arbitrary values, and shows that it fulfils differential privacy, subject to some caveats

Submissions are linked to individual identity though IP address, but are not sensitive even if linked to identity
RAPPOR offers strong security but is limited in terms of where it can be applied.

Homepage URL can be identified only if it is set by at least about 14,000 users which limits the type of organisations that can gain useful insights.

Security definition assumes consecutive responses are uncorrelated.
Anonymisation can be extended to include not only data but communications too.
Anonymisation can be extended to include not only data but communications too.
If data is only sensitive if linked to identity and communication is anonymous, the central authority is no longer trusted.
Short location traces can be anonymous if unlinked from other traces and identity

- An individual’s location trace can be easily linked to an individual through auxiliary data
- Shown by Neustar Research using NYC Taxi dataset

Splitting the location trace into small segments greatly reduces the risk of re-identification provided individual traces cannot be linked either by the trace or by the IP address the trace is submitted from
Intrusion detection logs can be anonymous if unlinked from other logs and identity. Organisations may wish to share information on attacks they are subject to but still hide their identity. Real example of a consortium of large companies, competing with each other but subject to same threats.

Techniques like prefix-preserving anonymization can hide details from logs but still need to submit data without disclosing identity.
Network nodes are run by volunteers, and changes to structure are made by consensus decision of 8 directory authority operators.
Each node knows the prior step in the circuit, and the next step, but unless there is collusion nobody can link source to destination.
Encryption prevents data flowing into a node from being linked to data flowing out, preserving unlinkability property.

Tor and other common anonymity systems do not hide traffic patterns.
There are useful alternatives between the extremes of fully centralised and full anonymisation by data subject

- Fully centralised approaches are versatile but have disadvantages
  - Risk of data compromise, expensive to set up and maintain security
  - Public lacks confidence in institutions
  - Centralisation of power lends itself to authoritarian tendencies
- Local anonymisation such as RAPPOR has limitations
  - Requires large amounts of data to get meaningful results
  - Care needed to preserve security if there is correlated data
- Data can be anonymised such that it is not sensitive in itself but does not offer plausible deniability either
  - In which case, submitting data over an anonymous communication system such as Tor can protect privacy without having to trust a central authority to anonymise data