Tiers of Quality for Tiers of Access in Privacy-Protected Data Products

Joshua Snoke
RAND

FCSM 2024
October 24, 2024

What is quality and why do we care?

Data sharing in the context of privacy risks

- We must manage the many tradeoffs between:
  - Minimizing the privacy loss (risk)
  - Minimizing the error in the data (accuracy/utility)
  - Maximizing the amount of information shared (quantity)
  - Minimizing administrative burden (cost)
  - Maximizing access (equity)

Quality

Quality measures the usefulness for end-users

- Starting questions to ask when evaluating quality
  - What is the goal of the data sharing?
  - Who are the users?
  - What will users do with the data/output?
  - What needs to be shared to facilitate this?

Some example answers to these questions

- What is the goal of the data sharing?
  - → E.g., Expand access, increase transparency, inform policy
- Who are the users?
  - → E.g., Researchers, students, state and local gov., community organizers
- What will users do with the data/output?
  - → E.g., Generate estimates, run statistical models
- What needs to be shared to facilitate this?
  - → Tables, microdata, other specific statistics, metadata

What impacts quality and how do we measure it?

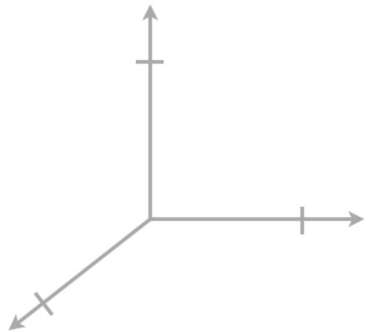
Privacy error comes in a variety of forms

- Restrictions on the domain of the input data
  - E.g., top-coding, aggregation, variable suppression
- Restrictions on the range of valid statistics
  - E.g., selective tables, simulated/synthetic models
- Bias or increased uncertainty from noise-infusion
  - E.g., record suppression, swapping, differentially private mechanisms

A three dimensional view of quality

- **Granularity**: What data are released?
  - Do the released data contain the relevant variables and granularity?
- **Validity**: Is it possible to perform valid inference?
  - Was the noise added in such a way that you can account for it?
- **Error**: Is the error acceptable?
  - After accounting for privacy error, are the estimates meaningful?

A wholistic view of quality considers all three dimensions



Valid inference depends on the analysis target

- Accuracy of inference to the true parameter
  - $(\bar{X}_p \leftrightarrow \mu)$
  - E.g., Coverage, asymptotic convergence rates
- Accuracy of difference with respect to the confidential parameter
  - $(\bar{X}_p \leftrightarrow \bar{X}_c)$
  - E.g., Absolute/relative/std. difference
- Similarity of inference with the confidential data
  - $((\bar{X}_p \leftrightarrow \mu) \leftrightarrow \overline{(X_c \leftrightarrow \mu)})$
  - E.g., Confidence interval overlap, sign/significance/overlap (SSO)

Recall the importance of understanding the users' needs

Common challenges in determining quality and connections with tiers of access

Various challenges arise in practice

- Common challenges:
  1. Forecasting the users
  2. Producing generally useful products versus specifically useful products
  3. Meeting quality demands amidst disclosure risks
  4. How good is good enough?

Tiers of access can help facilitate tiers of quality

Common tiers of access models

Data use agreements, access restrictions

- Most limited, highest quality

Remote access systems, validation servers

- Less limited, enables targeted analysis

PUFs, tables

- Potentially least limited, most general

Applying our concepts of quality to tiers of access

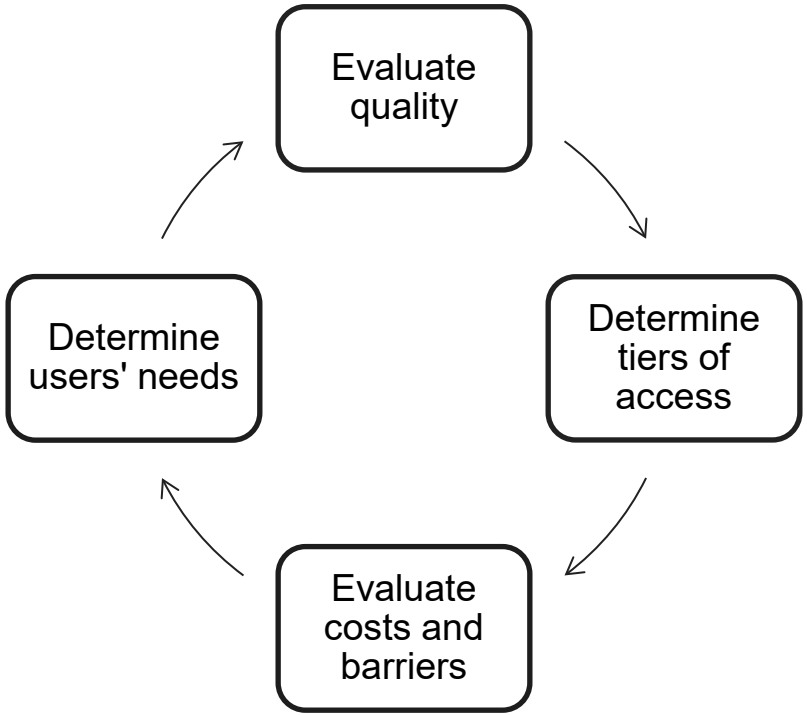
Access Tier	Granularity	Validity	Error	Target
Restricted	Highest	Yes	Least	True parameters
Validation Server	Varied	Likely	More	True parameters, similarity with confidential inference
PUFs	Less	Sometimes	Varied	Similarity to confidential estimates

Facilitating tiers of quality with tiers of access

- Tiers of access enable:
  - Multiple modes of access for different users
  - Multiple types of privacy-preserving methodologies
  - Different tradeoffs for our three-dimensional view of quality
- But tiers introduce tradeoffs between quality and cost/equity:
  - More restrictive tiers enable greater breadth and accuracy of analyses
  - They also entail greater administrative burdens and access barriers

Some levels of quality will only be possible with certain access restrictions

Developing data products requires a feedback loop



Closing thoughts on quality

- Quality depends on the users and their needs
- Quality should be measured on multiple dimensions
  - Granularity, validity, error
- Tiers of access can facilitate different tiers of quality
  - But must be balanced against costs and access barriers
- Consider an iterative process for developing data products

Closing thoughts on quality

- Quality depends on the users and their needs
- Quality should be measured on multiple dimensions
  - Granularity, validity, error
- Tiers of access can facilitate different tiers of quality
  - But must be balanced against costs and access barriers
- Consider an iterative process for developing data products

Thank you!

Comments/complaints/criticisms: jsnoke@rand.org