Confidentiality protection and physical safeguards

Lars Vilhuber
Cornell University

Funding acknowledged under NSF-#1131848 (NCRN) and a grant from the Alfred P. Sloan Foundation
publication trade-offs

Ease of access

Privacy loss

 Loss of detail

Tabulations

Public-use microdata

Raw microdata
How to provide easy and convenient access to data with more detail than public-use microdata, less privacy loss than direct publication of raw data?
public use data

Data provider/custodian

Data user/researcher
confidential data

What type of access device?
What type of person?
What type of room?
How do results leave the room?
Where is the data?
basic paradigm

“Data Enclave” or “Secure Room”

Where is the data?

What type of room?
making things virtual

“What type of access device?”

“What type of room?”

Where is the data?

“Virtual Data Enclave”
virtual data enclaves

Synonyms:

VDI
(virtual desktop infrastructure)

Thin clients

Remote desktop
Examples in 1990s

**Physical data enclaves**
- BLS HQ
- BJS data access
- Department of Education data
- Census Bureau RDCs
- Canadian RDCs
- HRS restricted-access data
- and many more

**Virtual data enclaves**
(data remains in secure data center)
Examples in 2017

Physical data enclaves

• BLS HQ
• BJS data access
• Department of Education data
  • Census Bureau RDCs
• Canadian RDCs
  • HRS restricted access data
• and many more

Virtual data enclaves
(data remains in secure data center)

• Census Bureau/Federal Statistical RDCs (since early 2000s)
• German IAB RDCs (since mid 2000s)
• French CASD (since late 2000s)
• Cornell’s CRADC, NORC (early 2000s)
• HRS restricted access data (2015)
• and many many more
basic levers

What type of access device?

What type of room?
basic levers

Where?

How?
access methods: enclaves

Ease of use

Privacy loss

Loss of detail

Tabulations
Public-use microdata

RDC

Physically secure room housing access devices and/or data

← Loss of detail
access methods: enclaves

Ease of use

- Tabulations
- Public-use microdata

Remote desktop
- Thin client
- Remote execution
- RDC

Software on your own PC giving a view onto secure data environment
Secondary secure PC giving a view onto secure data environment
Submitting analysis programs by email or through website (possibly combined with synthetic microdata)

Loss of detail
What type of room?
## Access matrix for confidential data

<table>
<thead>
<tr>
<th></th>
<th># access points</th>
<th>Access computers</th>
<th>Access rooms</th>
<th>Avail. analysis methods</th>
<th>Type disclosure avoidance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FSRDC researcher</strong></td>
<td>24 sites (~700 users)</td>
<td>Full</td>
<td>Full (badge access)</td>
<td>Some (choice of software)</td>
<td>Manual/ variety of rules</td>
</tr>
<tr>
<td>Census staff</td>
<td>n.d.</td>
<td>None (VDI)</td>
<td>None (VDI)</td>
<td>Some (choice of software)</td>
<td>Manual/ self/ variety of rules</td>
</tr>
<tr>
<td><strong>IAB: JoSuA researcher</strong></td>
<td>414 users</td>
<td>None (Web application)</td>
<td>None (Web application)</td>
<td>Smaller (software, whitelist commands)</td>
<td>Manual/ variety of rules</td>
</tr>
<tr>
<td><strong>CASD researcher</strong></td>
<td>371 sites (1471 users)</td>
<td>Extra Full (custom-built hardware)</td>
<td>Some (university office, EU)</td>
<td>Some (choice of software)</td>
<td>Manual/ variety of rules €300/ pack of 10</td>
</tr>
<tr>
<td>Stat.Denmark (typical EU)</td>
<td>?</td>
<td>None (VDI) - Some (host institution)</td>
<td>None (VDI) - Some (host institution)</td>
<td>Some (choice of software)</td>
<td>Manual/ self/ variety of rules</td>
</tr>
</tbody>
</table>
How do results leave the room?
Typically, the researcher asks an authorized agent of the data provider to review the results for risks of disclosure, and he will then send them to the researcher.
What if the “authorized agent” were the researcher?
self-controlled release of results

• Researcher controls release of results
  • Prepares results herself
  • According to certain prescribed rules
  • Sends them through a system
  • Automatically receives results typically per email

• Used
  • Most often by contractually-controlled non-enclave data
  • Data in some university- or faculty-controlled enclaves (HRS, Dept. of Ed)
  • Danish researcher access system
access methods: enclaves

Ease of use

Tabulations

Public-use microdata

Self-controlled release

Remote desktop

Thin client

Remote execution

RDC

Raw microdata

← Loss of detail
access methods: enclaves

Tabulations

Public-use microdata

Ease of use

Loss of detail

Raw microdata

Remote desktop
Thin client
Remote execution
RDC

Self-controlled release

← Loss of detail
## Access matrix for confidential data

<table>
<thead>
<tr>
<th>Control by data provider of:</th>
<th># access points</th>
<th>Access computers</th>
<th>Access rooms</th>
<th>Avail. analysis methods</th>
<th>Type disclosure avoidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSRDC researcher</td>
<td>24 sites (~700 users)</td>
<td>Full</td>
<td>Full (badge access)</td>
<td>Some (choice of software)</td>
<td>Manual/ variety of rules</td>
</tr>
<tr>
<td>Census staff researcher</td>
<td>n.d.</td>
<td>None (VDI)</td>
<td>None (VDI)</td>
<td>Some (choice of software)</td>
<td>Manual/ self/ variety of rules</td>
</tr>
<tr>
<td>IAB: JoSuA researcher</td>
<td>414 users</td>
<td>None (Web application)</td>
<td>None (Web application)</td>
<td>Smaller (software, whitelist commands)</td>
<td>Manual/ variety of rules</td>
</tr>
<tr>
<td>CASD researcher</td>
<td>371 sites (1471 users)</td>
<td>Extra Full (custom-built hardware)</td>
<td>Some (university office, EU)</td>
<td>Some (choice of software)</td>
<td>Manual/ variety of rules €300/ pack of 10</td>
</tr>
<tr>
<td>Stat.Denmark (typical EU)</td>
<td>?</td>
<td>None (VDI) - Some (host institution)</td>
<td>None (VDI) - Some (host institution)</td>
<td>Some (choice of software)</td>
<td>Manual/ self/ variety of rules</td>
</tr>
</tbody>
</table>
penalties
• FSRDC and federal employee:
  • federal prison sentence of up to **five (5)** years, a fine of up to **$250,000**, or both.

• France:
  • prison sentence of up to **one (1)** year, a fine of up to **€15,000**, or both.
penalties

• IAB:
  • Loss of data access for up to **two (2)** years for researcher and institution
  • Contractual penalty up to **€60,000** paid by the **institution**

• Denmark:
  • Researcher: Loss of data access **for life**, or up to **three (3)** years for “minor breaches”
  • **Institution**: Loss of access for a positive but limited (undefined) period
  • No financial or penal penalties

**Of Note**: the FSRDC contract explicitly **excludes** a responsibility of the university for the actions of its employees, though university remains bound by FWA/IRB.
penalties

• Does ease of application matter (penal vs. contractual rules)?

• Is it conducive to more strongly engage the researcher’s employer (typically but not exclusively a university)?
trust and access
What type of person?
hypothesis: culture matters

• Researchers and agencies create the communities in which rules are applied and enforced
  • Training and “indoctrination”:
    • Training of FSRDC researchers (short, decentralized) vs. FedStat employees (≥1 day on-site)
    • 1 full day on-site (in Paris) training for French researchers
  • Common forums:
    • Conferences: Canadian, US (FSRDC, NCHS) yearly RDC conferences
    • Discussion, local groups: users of FSRDC share a common physical space

• More or less tight binding of researchers into a community is important
virtual enclave = centralization
Concerns about centralized compute infrastructure

• Scope
  • FSRDC infrastructure dwarfed by other federal research investments (e.g. XSEDE) that cannot be utilized

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cores</th>
<th>Tflops</th>
<th>As a multiple of FSRDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSRDC</td>
<td>240</td>
<td>4.36</td>
<td>1x</td>
</tr>
<tr>
<td>Wrangler (TACC)</td>
<td>2304</td>
<td>62</td>
<td>14x</td>
</tr>
<tr>
<td>Stampede (TACC)</td>
<td>102400</td>
<td>9600</td>
<td>2202x</td>
</tr>
</tbody>
</table>
summary
some concluding thoughts

• How to enable a scalable and secure system
  • Does it require changes in the legal framework?
  • How to build a culture of responsible and secure data access among researchers?
  • What kind of devices or access mechanisms do we want to enable?
  • Who gets to hold the data that researchers actually access?
thank you

lars.vilhuber@cornell.edu
Thanks

• Stefan Bender (formerly IAB and now Bundesbank, Germany)
• Jörg Heining (IAB, Germany)
• Roxanne Silberman (CASD, France)
• Kamel Gadouche (CASD, France)
• Jean Poirier (CIQSS, Canada)
Some References


