Increasing utility of economic statistical information

Steven Thomas

Section Chief – Statistics Canada

Presentation to

2021 Expert Meeting on Statistical Data Confidentiality

December, 2021

Delivering insight through data for a better Canada

Disclaimer

The views expressed in this presentation are those of the author alone and do not necessarily represent the position of Statistics Canada or the Government of Canada in these matters.

Access to informative data is more important than ever

- Researchers, policy makers and the general public require more data to make informed decisions about the economy and life in general
- Statistics Canada collects and publishes an immense amount of data for the various aspects of the Canadian economy
- This data is published in many aggregated tables released on a daily basis. However, many cells are suppressed because of disclosure risks
- Disaggregated data are made available to a limited set of researchers

How do we expand the analytical information that is released and allow expanded access to our data holdings while still ensuring that all information is kept confidential?

The risk versus utility conundrum

- **Utility**

Under the *Statistics Act*, Statistics Canada is required to collect, compile, analyse, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the people of Canada.

- **Risk**

Under that same Act, no person who has been sworn under section 6 shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in a manner that it is possible from the disclosure to relate the information obtained to any identifiable individual person, business or organization.


Reward
Risk

What is Disclosure?

Definition from the *Organization for Economic Co-operation and Development (OECD) Glossary of Statistical Terms*:

> **"Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure has two components: identification and attribution."**

Disclosure Model

1. Access: an attacker gains access to a Statistics Canada information

2. Re-Identification: Through some process, attacker is able to identify a respondent

3. Attribution: Attacker is able to attribute novel information to this respondent

Disclosure under an Attack Scenario for Economic Statistics

- Assuming that
  - Identities of contributors are known.
  - All contributions are known to some degree.
  - One contribution may be known exactly.
- Disclosure has occurred when an attacker can make an estimate of a contributor's value that is **too precise**

Identifying Disclosure in Economic Totals

The Prior–Posterior *PQ* rule (Willenborg and de Waal, 2001)

It is assumed that, prior to the publication of the table, every respondent can estimate the contribution of each other respondent to within ***q*** percent. A cell is considered sensitive if someone, e.g. one of the respondents, can estimate the contribution of an individual respondent to that cell within ***p*** percent after (i.e. posterior to) publication of the table.

Tabular Risk Mitigation - Cell Suppression

- Suppress sensitive cell values to control disclosure
- Suppress secondary (complementary) cells to avoid simple recalculation from disseminated tables
- Risk identification and complementary cell suppression with Statistics Canada's G-Confid software

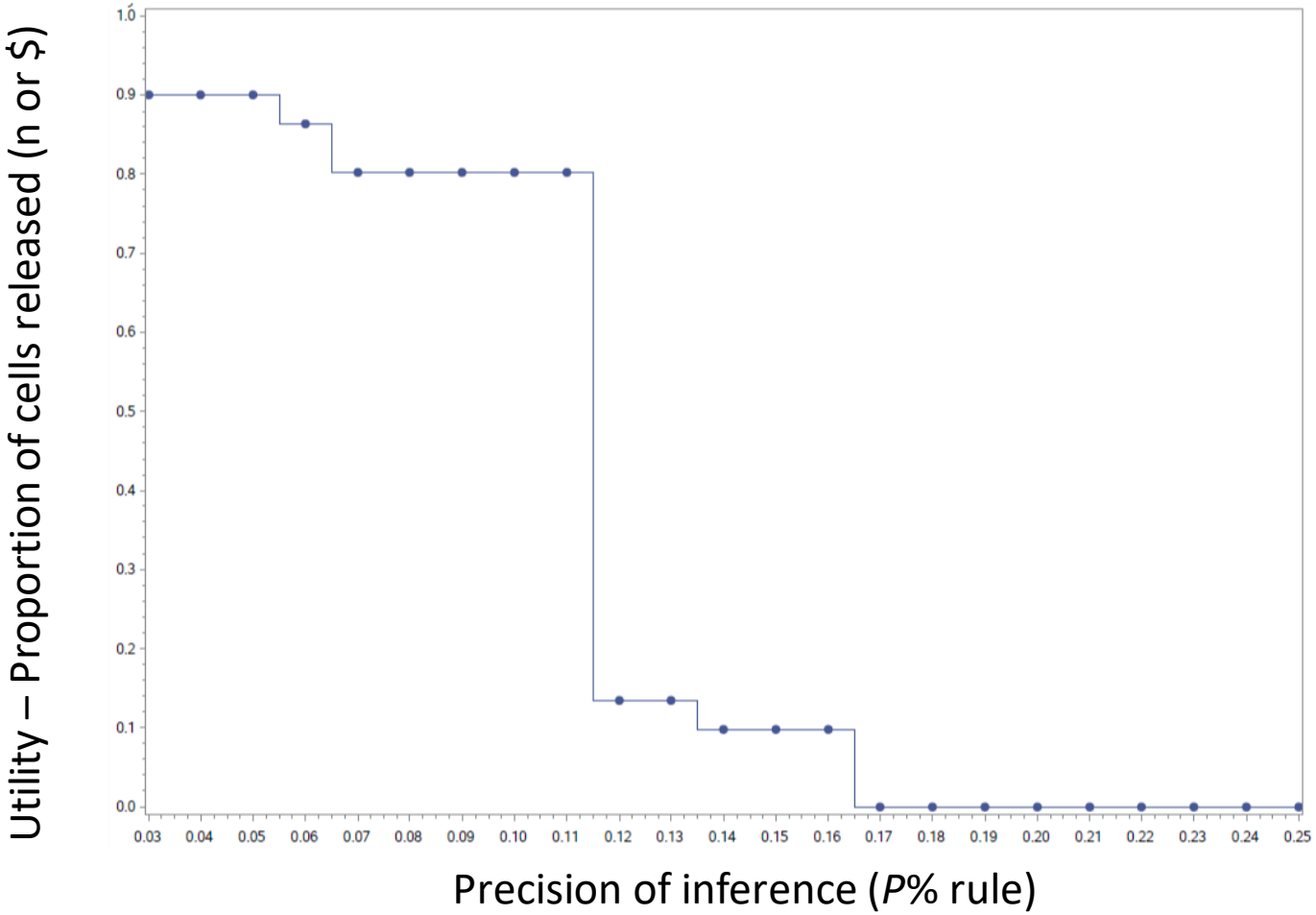
Illustration of suppression challenges

Petroleum and coal product manufacturing [324]					
Sales of goods manufactured (shipments)					
	Nov-20	Dec-20	Jan-21	Feb-21	Mar-21
Geography	Dollars				
Newfoundland and Labrador	x	x	x	x	x
Prince Edward Island	..	..	..	..	..
Nova Scotia	x	x	x	x	x
New Brunswick	x	x	x	x	x
Quebec	883,912	x	870,801	1,066,942	1,265,370
Ontario	933,614	1,022,768	1,008,672	973,491	1,337,444
Manitoba	4,149	x	x	x	x
Saskatchewan	x	x	x	x	x
Alberta	1,087,620	1,195,014	1,301,763	1,326,189	1,552,344
British Columbia	x	x	x	x	x

Increasing Utility

- Utility of a table can be measured in the form of the proportion of cells being released or the proportion of cell value being released
- Utility concerns may suggest taking more risk is appropriate
- Must still consider sensitivity of the information – Disclosure impact assessment
- Utility can be increased by taking more risk
  - By lowering the *P* in the *PQ* assessment fewer cells will be identified and will require less protection from complementary cells → Risk / utility balance

Risk vs Utility Graph



A Perturbation Alternative
Random Tabular Adjustment (Stinner, 2017)

- With the *PQ* assessment, the precision of inference is controlled through suppression. The protective effect is seen in the geometric risk assessment
- An alternative strategy is to consider the precision of inference through the distributions of inference. Prior knowledge distributions can be combined with information gained from tabular release to measure posterior inferences on individual contributions
- In cases where the posterior information is too precise, variability can be added in the form of random normally distributed noise in order to prevent disclosure

Simple Simulated Example

Perturbation Solution – RTA Confid

	Industry							
	A		B		C		All	
Region	Est	CV	Est	CV	Est	CV	Est	CV
1	$ 8,360	0%	$ 74,167	0%	$ 45,663	0%	**$ 128,190**	**0%**
2	$ 76,066	0%	$ 16,247	0%	~~$ 44,918~~ $ 49,106	~~0%~~ 5%	~~**$ 137,231**~~ **$ 141,419**	~~**0%**~~ **2%**
All	**$ 84,426**	**0%**	**$ 90,414**	**0%**	~~**$ 90,582**~~ **$ 94,769**	~~**0%**~~ **3%**	~~**$ 265,422**~~ $ 269,609	~~**0%**~~ 1%

- Loss of $175,008 with cell suppressions (66% of internal cells)

Secondary suppressions

Determined and Protected

Suppressed

Marginal Cells are adjusted accordingly

Suppression – G-Confid

	Industry							
	A		B		C		All	
Region	Est	CV	Est	CV	Est	CV	Est	CV
1	X	0%	$ 74,167	0%	X	0%	**$ 128,190**	**0%**
2	X	0%	$ 16,247	0%	X	0%	**$ 137,231**	**0%**
All	**$ 84,426**	**0%**	**$ 90,414**	**0%**	**$ 90,582**	**0%**	**$ 265,422**	**0%**

Perturbation- RTA

	Industry							
	A		B		C		All	
Region	Est	CV	Est	CV	Est	CV	Est	CV
1	$ 8,360	0%	$ 74,167	0%	$ 45,663	0%	**$ 128,190**	**0%**
2	$ 76,066	0%	$ 16,247	0%	$ 49,106	5%	**$ 141,419**	**2%**
All	**$ 84,426**	**0%**	**$ 90,414**	**0%**	**$ 94,769**	**3%**	**$ 269,609**	**1%**

A Micro-data Perspective
Public Business Data –
No Public-Use Files have ever been released.

- Creating public data usually includes some form of de-identification / anonymization
- Anonymization methods have been largely ineffective when applied to business files
    1. **The ease of obtaining business information**
    2. **The small size of business populations**
    3. **The distribution of businesses**

Alternative strategy

- Permit some probability of re-identification of businesses
- Instead, prevent disclosure as defined earlier by preventing attribution:

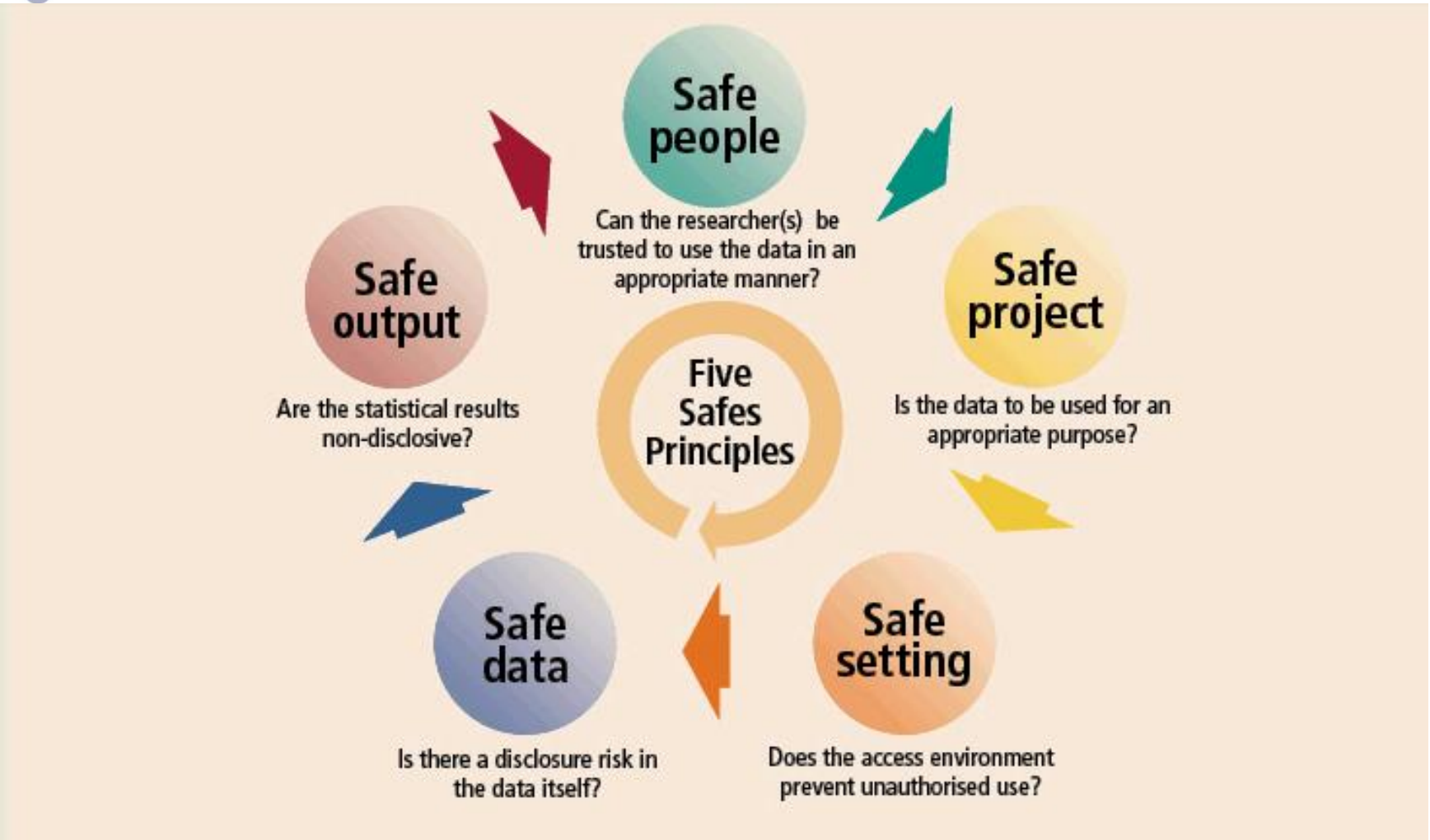
- Broad idea is that re-identification in the absence of attribution is not disclosure under the *Statistics Act*
- Noise addition can be used to prevent attribution (similar to RTA)
  - Noise-added values are no longer the true values – attribution fails because an attacker will not receive useful information about confidential targets

Challenges with adding noise

- How much noise will make these files safe enough?
- Communication - How can we illustrate the noise addition to the user and the data provider?
- What will the public's reception be and what should we consider from an ethical point of view? Perceived disclosure risks.

A way forward

- Developing a trust relationship with researchers and government organizations
- Based on a “Five Safes Framework”, access can be given to trusted researchers working in a safe environment with proper vetting controls
- For more information: www.fivesafes.org
- If the data is not safe enough, other controls can be used



Source: Consumers Health Forum of Australia

Research access to economic data – In development

- Access for trusted researchers to economic data
- A replacement for the Canadian Centre for Data Development and Economic Research – On site access for deemed employees
- An extension to Statistics Canada's successful Research Data Center (RDC) program – Remote secured access nationwide.
- Virtual access for trusted research partners
- Economic output vetting rules with cell suppression (Collapsing) strategy

Conclusion

- Increasing the amount of information available to researchers comes at a cost.
- This cost can be measured and calculated risks can be taken when appropriate.
- Alternative controlled access options are a way forward.

Thank you!
Merci!

Contact: Steven.Thomas@statcan.gc.ca