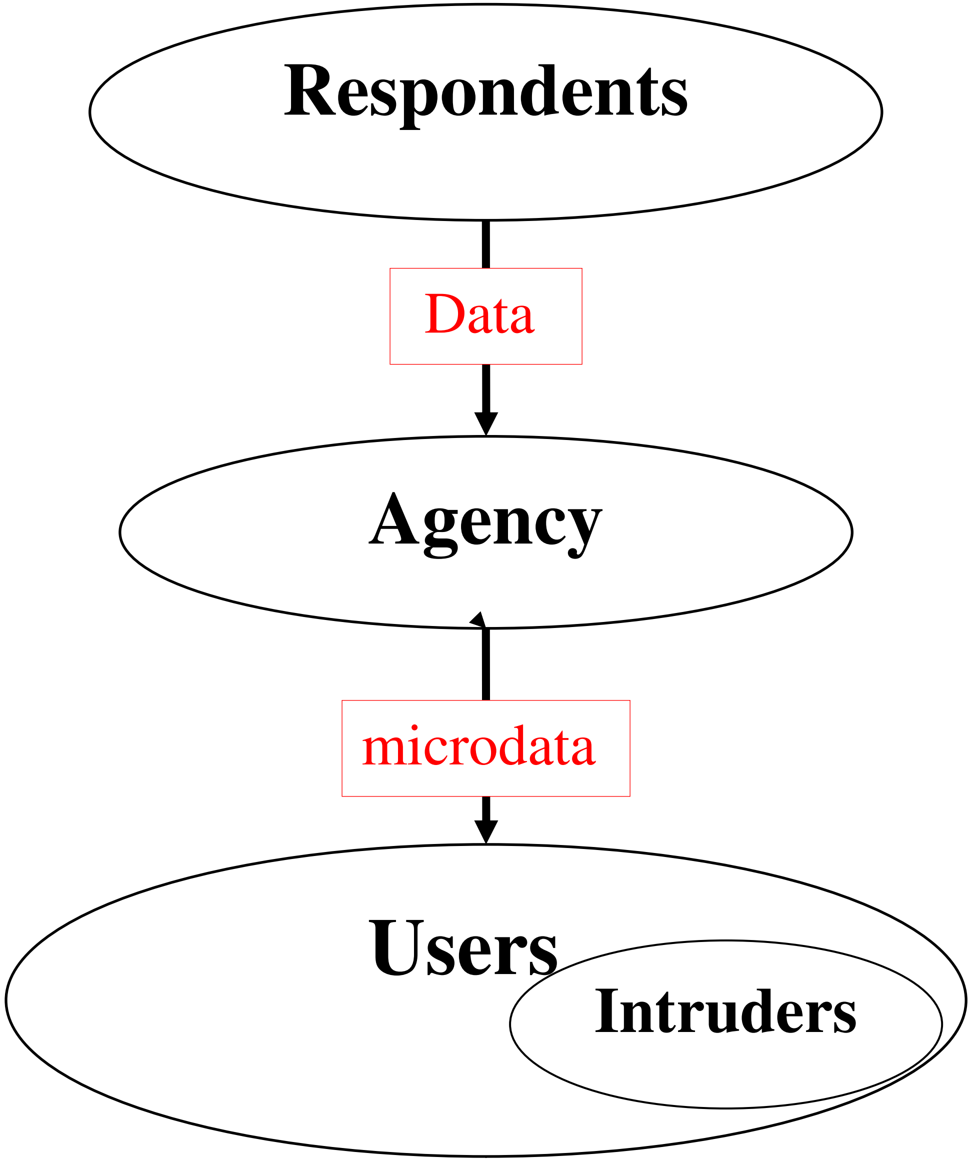
Disclosure Risk, Decision Theory and Predictive Distributions

Chris Skinner

University of Southampton



Respondents

Data

Agency

microdata

Users

Intruders

Multiple Potential Decision Makers (Perspectives)

- **Agency**
- **Intruder**
- **User**
- **…**

The Agency as Decision Maker – Example

- **Action** - publish or suppress cell

- **Loss** = loss of information if suppress
  = potential disclosure if publish

Zaslavsky and Horton (1998, *JOS*)

Decision Framework for Agency - Standard

Actions:

- *Statistical Disclosure Control* (SDC) methods

Loss Criteria:

- *Disclosure Risk*
- *Utility*

Risk-Utility analysis (Duncan et al., 2001a, 2001b)

Decision Framework for Agency - More General

Actions:

- *SDC* methods (modification of outputs)
- *Disclosure management strategy* (disincentives to intrusion or misuse – access licenses, penalties for misuse, training,…)

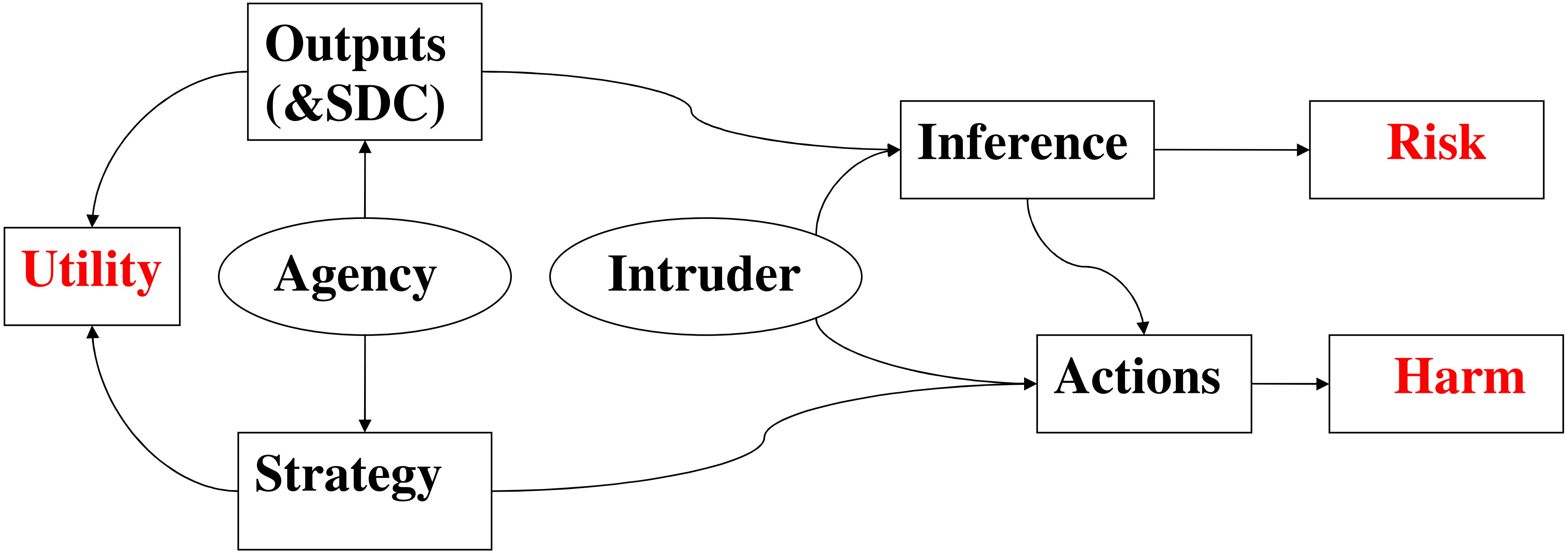
Loss Criteria:

- *Disclosure Risk* (ability to infer)
- *Disclosure Harm* (harmful intruder actions)
- *Utility*

Rationale for Definition of Disclosure Risk…

UK National Statistics code of practice “no statistics [i.e. outputs] will be produced that are likely to identify an individual”

US Code Title 13 prohibits US Census Bureau from producing outputs “whereby the data furnished by any particular establishment or individual under this title can be identified”


Outputs (&SDC)

Utility

Agency

Intruder

Inference

Risk

Actions

Harm

Strategy

Intruder Perspective

1. Intruder's predictive distribution $p(y)$, where y is true value of target (identity or attribute disclosure)
   Bayesian inference – prior→ data=output → posterior

2. $a =$ action of intruder
   intruder's loss function $L(y,a)$.
   optimal a minimises $E[L(y,a)] = \sum_x L(y,a)\,p(y)$

Duncan and Lambert (1986, *JASA*; 1989, *JBES*)

Disclosure Risk or Harm?

Decision theory models intruder behaviour

> "….disclosure is limited only to the extent that the intruder is discouraged from making any inferences, correct or incorrect, about a particular target respondent"
>
> Lambert (1993, *JOS*)

Expected loss relates to harm not risk

$p(y)$ *risk of perceived identification*

Extending Decision Theory

- agency, intruder and user decision makers.
- agency’s loss function “quantifies, from the agency’s perspective, the harm that the intruder’s action… produces to the agency and the data providers” for a given ‘state of the world’.
- ‘disclosure risk’ is the expected value of agency’s loss function with respect to the agency’s posterior probability distribution about the actions of the intruder and the state of the world.

Dobra, Fienberg and Trottini (2003, *Bayesian Stat. 7*).

Game Theory

- intruder as adversary of the agency and user
- focus on actions and thus harm

Keller-McNulty, Nakhleh and Singpurwalla (2005, *ISR*)

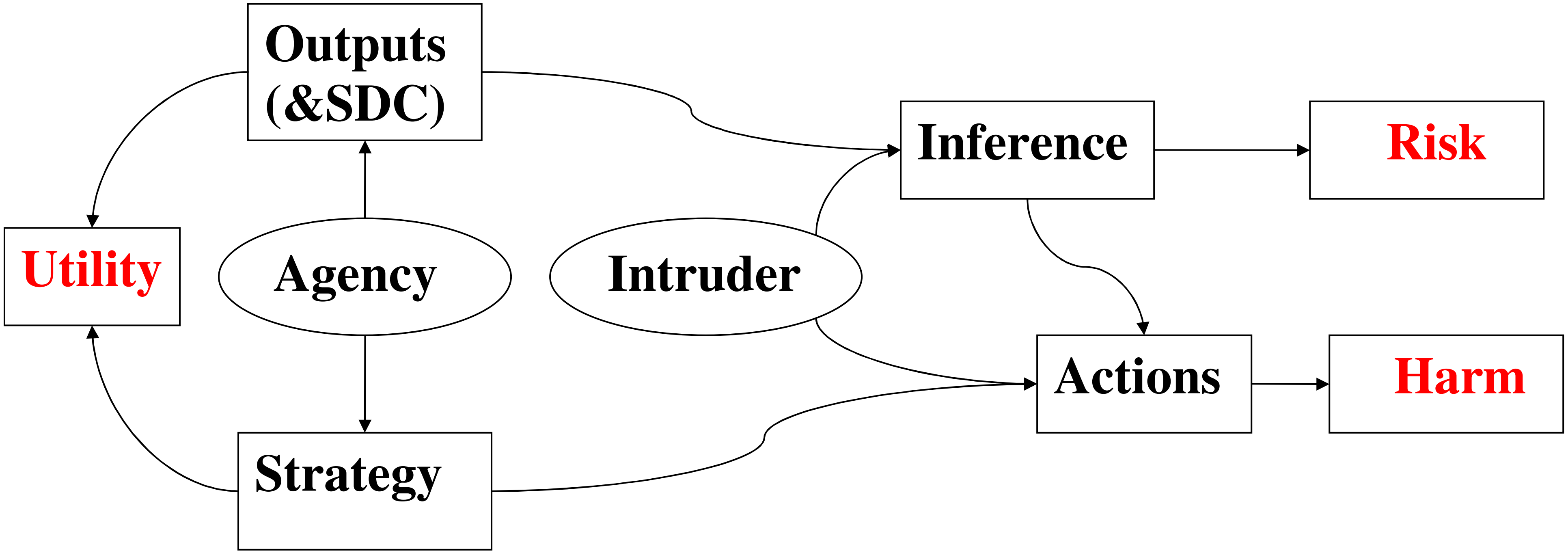
Controlling Disclosure Harm

‘Politics’

- Measuring harm – dependence on perspectives of respondent ‘stakeholders’ & agencies, variation over time
- Managing perceptions

‘Social science’

- Intruder behaviour – minimising *E*(loss) models, dependence on disclosure management strategy


Outputs (&SDC)
Inference
Risk
Utility
Agency
Intruder
Actions
Harm
Strategy

Disclosure risk as predictive distribution: dependence on method of attack

Method of attack – intruder's actions before disclosure

$$p(y) = \sum_k p(y \mid attack\ \ method\ \ k)\Pr(attack\ \ method\ \ k)$$

Disclosure risk = $p(y \mid attack\ \ method\ \ k)$

- deals with multiple intruders
- avoids dependence of disclosure risk on management strategy

Handling Many Possible Attack Methods

- ‘fundamental difficulty’ (Frank, 1986)
- consider worst case (Duncan et al., 2001)
- exclude extreme scenarios, e.g. ONS - “ statistical disclosure control methods…will be judged sufficient when the guarantee of confidentiality can be maintained, taking account of information *likely* to be available to third parties, either from other sources or as previously released National Statistics outputs”.

- ONS - scenarios excluded if:
  ‘it would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain’.
- *de facto anonymisation* of business microdata in Germany
  - scenarios excluded if
    the intruder’s “costs of trying to reidentify records in the dataset” are deemed to be “higher than the benefit gained by the disclosed information” (Brandt et al., 2008).

    c.f. Lambert (1993)

Disclosure risk as posterior predictive distribution: dependence on prior

Bayesian

Fienberg, Makov and Sanil (1997, *JOS*)
Reiter (2005, *JASA*)

Intruder or agency priors?

Frequentist model-based

Fuller (1993, *JOS*),
Skinner & Shlomo (2008, *JASA*)

Non-ignorability of Attack Method

Exact match of microdata record using categorical key variables without measurement error

$$p(y\,|\,data, attack\,method) \neq p(y\,|\,data)$$

Examples of attack methods:

Fishing: start with microdata record and search population for match

Directed search: start with known population unit and search microdata for match

Skinner (2007, *JRSS'A'*)

(Identity) Disclosure Risk for Record Linkage

Intruder matches records in microdata to external file

identity disclosure = correct match

$y = 1$ if match is correct
$y = 0$ if match is incorrect

What is $p(y)$?

Probabilistic Record Linkage

$\tilde{X}_a$ key variables for record a in microdata

X_b key variables for record b in external database

$\gamma(\tilde{X}_a, X_b)$ comparison vector

$$m(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in M\,]$$

$$u(\gamma) = \Pr[\gamma(\tilde{X}_a, X_b) = \gamma \mid (a,b) \in U\,]$$

M matched pairs ($a = b$), U unmatched pairs ($a \neq b$)

likelihood ratio $m(\gamma)/u(\gamma)$

Identity Disclosure Risk

$$p(1) = \Pr(y = 1) = \Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = \gamma]$$

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = \gamma] = m(\gamma)\theta / [m(\gamma)\theta + u(\gamma)(1-\theta)]$$

where $\theta = \Pr[(a,b) \in M]$

Example: exact match on categorical key variables

$\gamma(\tilde{X} = X) = j$ if $\tilde{X} = X = j, \quad j = 1,\ldots,K$

$\gamma(\tilde{X} = X) = K + 1$ otherwise

If no misclassification ($\tilde{X}_a = X_b$ if $(a,b) \in M$)
$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] = 1/F_j$

where F_j is number in population with $X = j$

Example (continued)

F_j usually unknown to intruder

worst cases $F_j = 1, 2, 3, \ldots$, $1/F_j = 1.0, 0.5, 0.333, \ldots$

F_j cannot be estimated consistently

$p(1) = E(1/F_j \mid data)$, where $E(. \mid data)$ is with respect to predictive distribution of F_j

e.g. based on log-linear model for key variables (Skinner and Shlomo, 2008) tends to shrink (reduce) risk for worst cases

Example with misclassification

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = j] \doteq \frac{\phi_{jj}}{\tilde{F}_j}$$

$$\text{where } \phi_{jj} = \Pr(\tilde{X}_a = j \mid X_a = j)$$

Estimation of predictive distribution

- assume no training data (intruder has no validating information)

- ML or Bayes estimation under mixture model

- or sensitivity analysis (little information in data about some parameters)
  e.g. ML estimation for given values of ϕ_{jj}

Example: Estimation of θ

$$\Pr[(a,b) \in M \mid \gamma(\tilde{X}_a, X_b) = \gamma] = m(\gamma)\theta / [m(\gamma)\theta + u(\gamma)(1-\theta)]$$

Can estimate θ consistently using design-based sampling theory if assume:

- microdata sample obtained by probability sampling
- can determine microdata inclusion probabilities for units in external database
- size of overlap ‘large’

Concluding Remarks

- intruder-based decision theory not relevant to assessment of disclosure risk
- limited role for intruder-based priors when assessing disclosure risk
- need for empirical comparison of model-based estimation of disclosure risk from record linkage with empirical estimation from re-identification experiments
- record-level vs. file level measures of risk