Situated Information Flow

Sebastian Benthall
NYU – ILI & CCS

The problem

Privacy violations due to data reuse.

E.g. Using social media behavior to develop psychographic profiles for political ad targeting.

The problem

These tend to be cases where information flows “across contexts.”

Why?

- *Information gets its meaning from the context of its use.* [Implied in CI] A change in meaning can mean a change of context.
- *Contexts get their form from their purpose.*

There is a theoretical gap in Contextual Integrity about context clashes. This addresses it.. (Benthall, Gürses, and Nissenbaum, 2017)

How and why is Contextual Integrity used?

Contextual Integrity is the best theory of privacy!

This work is to develop the theory further. *Towards a CI 2.0?*

Contextual Integrity says there are five parameters of an *information norm*:

Sender, Receiver, Subject, **Topic**, and Transmission Principle.

[Patient, Doctor, Patient, **Health**, Confidentiality]

But... *information topics are indeterminate.* E.g.:

How and why is Contextual Integrity used?

The main challenge is to one of the Contextual Integrity norm parameters:

- Sender
- Receiver
- Subject
- ***Type or attribute***
- Transmission principle

- Social expectations around information type may be based in culturally understood social spheres.
- But the actual semantics of information is not bound by social expectation.
- Data does not “contain” its meaning.

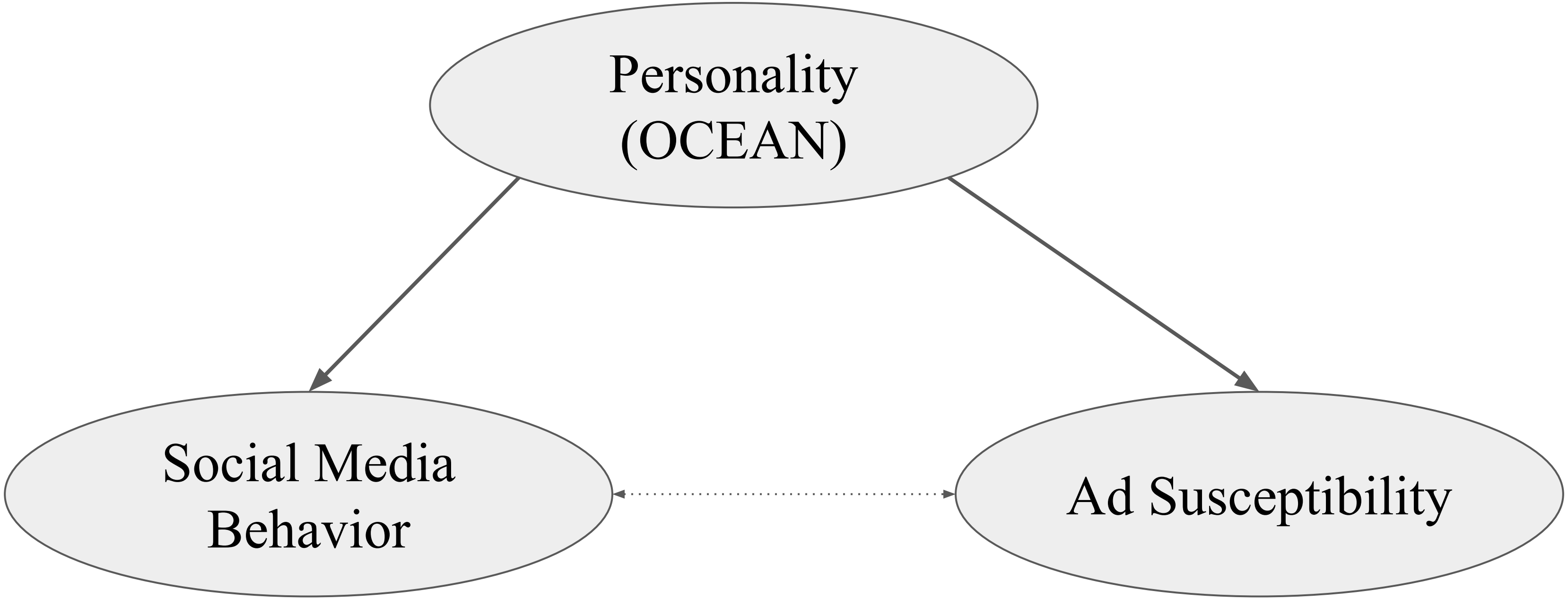
How and why is Contextual Integrity used?

We may *expect* that information flows like water or oil.

But it *doesn't*. It has a different physics.

Herein lies the problem.

Example


Personality
(OCEAN)
Social Media
Behavior
Ad Susceptibility

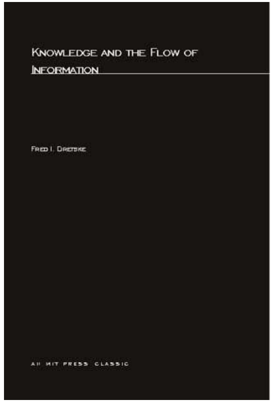
Current progress and results

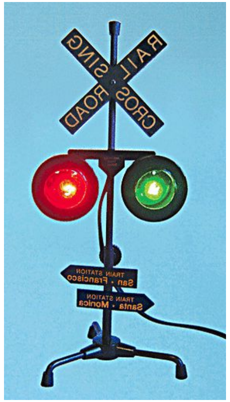
- The task is to understand what gives data its meaning.
- This is a philosophical question, but also a scientific one.
- *An effective theory of privacy must be built on the same sound science of information flows as is used in engineering.*

Current progress and results

- Judea Pearl’s theory of causation is a widely celebrated and applied theory of causal modeling.

What is *information flow*, really?





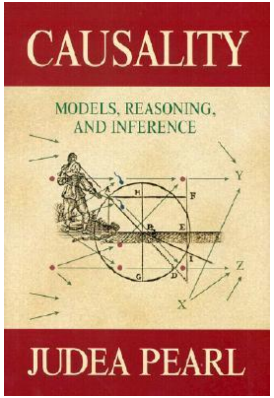
According to Dretske (1981) (epistemology, philosopher of mind) building on Shannon (1948), *information* is a naturalistic and causal property:

Information that P is the message/signal needed for a suitably equipped observer to learn P, due to the ***nomic associations*** *of the signal with P.*

Nomic means "law-like", as in scientific law.

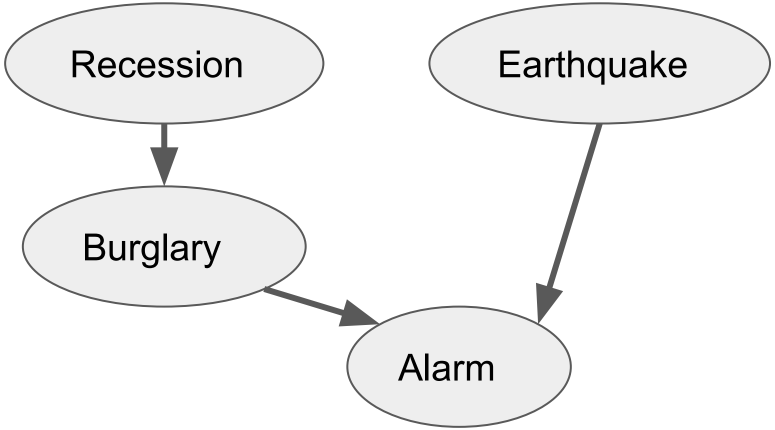
The red light carries the information that the train is coming because the ***(lawfully, regularly)*** *red is light* ***if and only if*** *the train is coming.*

What is *information flow*, really?

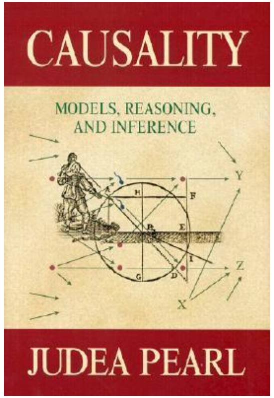


The **alarm** carries information about earthquakes, burglaries, and recessions. (Topics are indeterminate).

The **recession** does not carry information about **earthquakes** (in this model). They are conditionally independent.

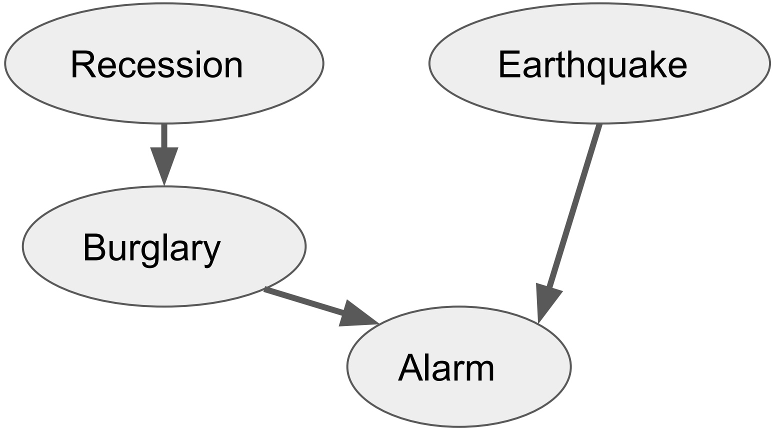


What is *information flow*, really?



Pearl's (2000) system for understanding causality is widely acknowledged and applied in *statistics, philosophy, machine learning, cognitive psychology, social science research methods,* ...

In Pearlian causality, events are part of a *causal structure* represented as a directed acyclic graph.



This structure determines the *conditional dependency* of events on each other; how they systematically covary.

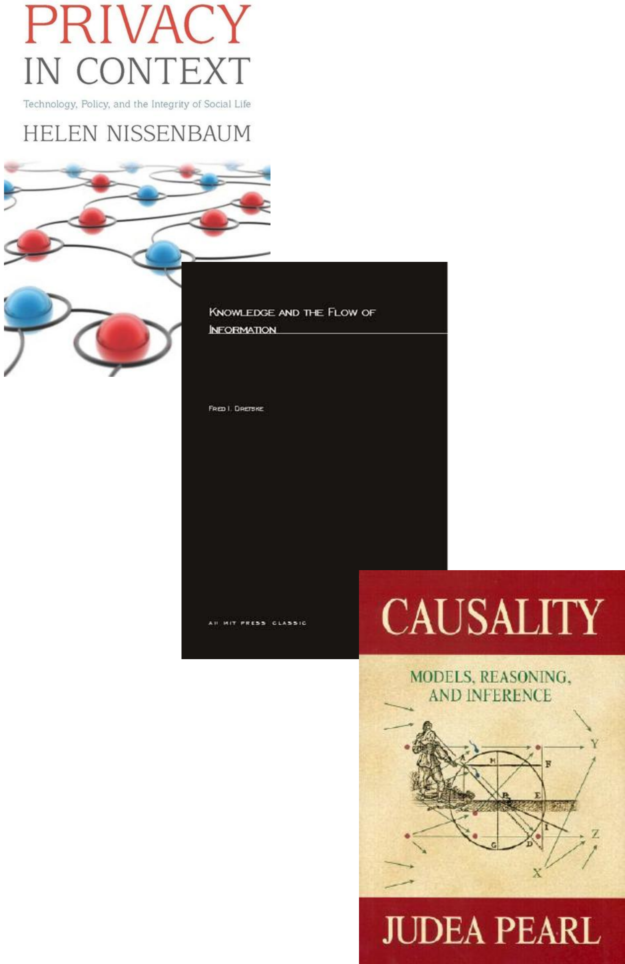
Situated Information flow

1. Privacy is appropriate information flow. (Nissenbaum)

2. Information flow is a message or signal from which something can be learned because of nomic association. (Dretske)

3. The nomic associations are the conditional dependencies derived from causal structure. (Pearl)

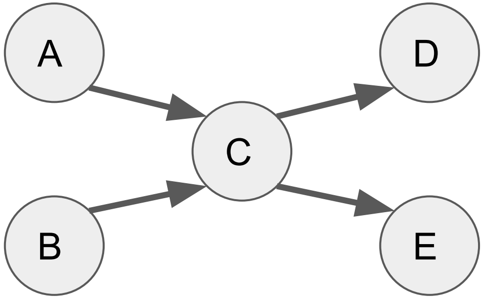
The meaning of data is a function of the processes that generated it, and their context.



Situated Information Flow

Def: A *situated information flow* is a causal flow situated in the context of other causal relations.

Bayesian Networks



Bayesian Networks (BN) are a formalism for representing the relationship between random events.

A BN has:

- A directed, acyclic graph of *nodes*, representing random variables, connected by edges
- A *conditional probability distribution* (CPD) for each node, which is the probability distribution of its random variables, conditional on its parent.

Together. these define a joint probability distribution over all the random variables, with some important independence relations qualitatively inferable from the graph.

Challenges encountered, lessons learned

Situated information flow theory raises deep questions about the fundamental nature of probability (i.e. Bayesian vs. frequentist interpretations) and causality.

We must distinguish between the *real causal relations* that generate the data and the *beliefs about causal structure* used by the observer/interpreter.

These deep questions can distract from its pragmatic value.

Future work

“So what?” Finding the pragmatic consequence.

A review of omnibus data protection laws reveals how they vary:

- Some refer to and regulate ‘categories’ of personal information (GDPR, CCPA)
- Others refer more vaguely to ‘personal data’ without categories (OPEC, APEC)

When is a general prevention of data transfer appropriate?

Future work

Maybe:

More requirements about revealing to data subjects *how* data is collected and use?

(Not just *what, when, where* and *why*.)

End

spb413@nyu.edu