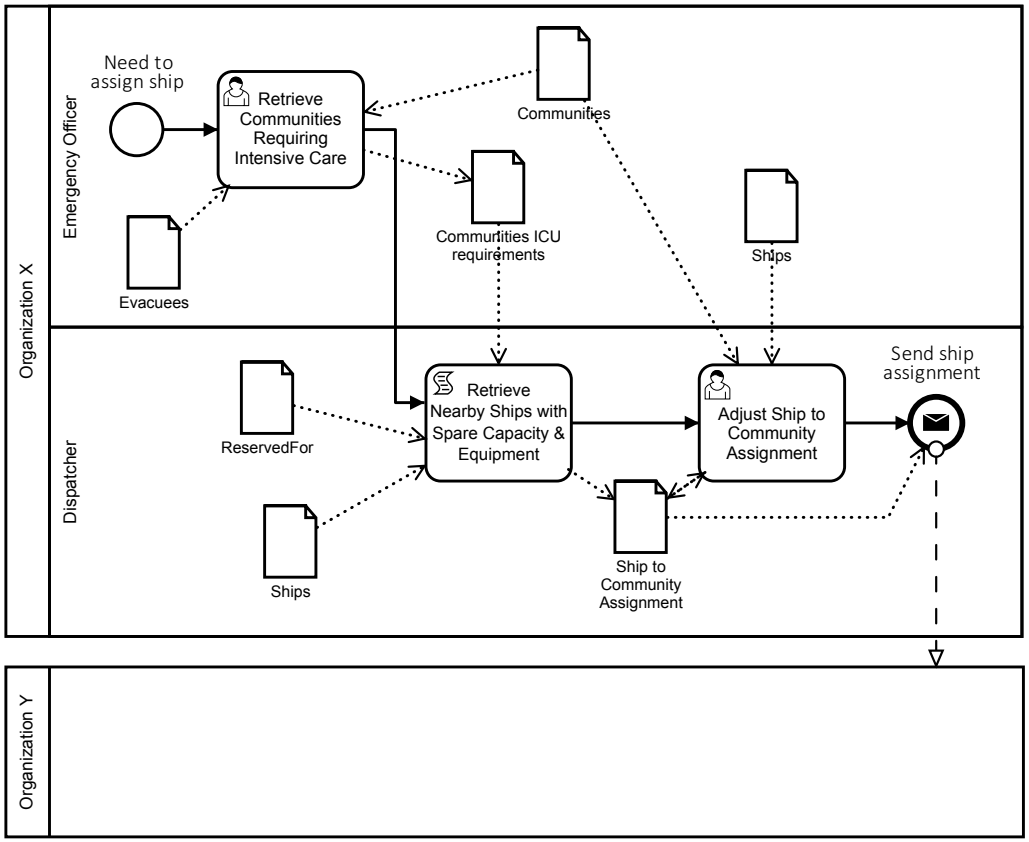
Differential Privacy Analysis of Data Processing Workflows

Marlon Dumas, Luciano García-Bañuelos, Peeter Laud

UNIVERSITY OF TARTU
CYBERNETICA

Motivating example



Conflicting goals: Privacy vs. Utility

We need to release **aggregate** information about data without leaking information about an **individual** involved in the incident

- Aggregate info: Number of crew members of nationality X in the ship
- Individual info: Is a particular crew member of nationality X?

Problem: Aggregate information may leak information on individuals

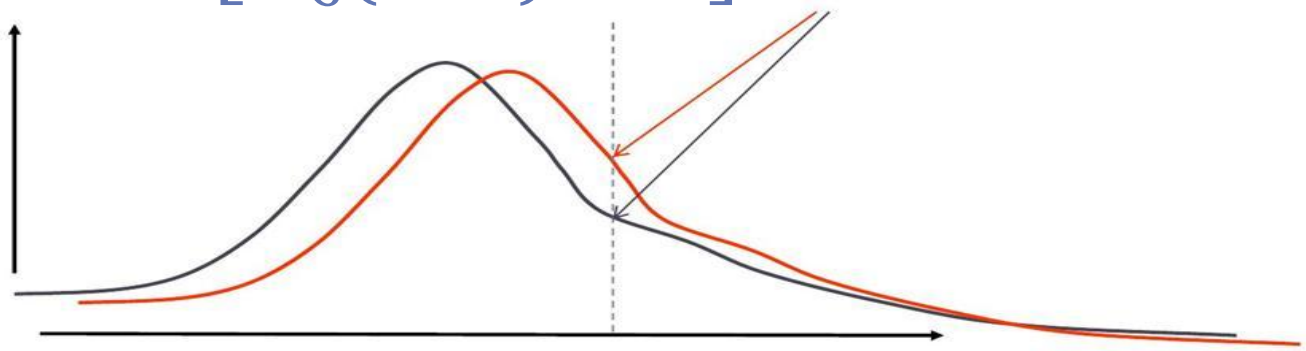
Number of crew members of nationality X in the ship,

Number of crew members of nationality X in the ship excluding Y

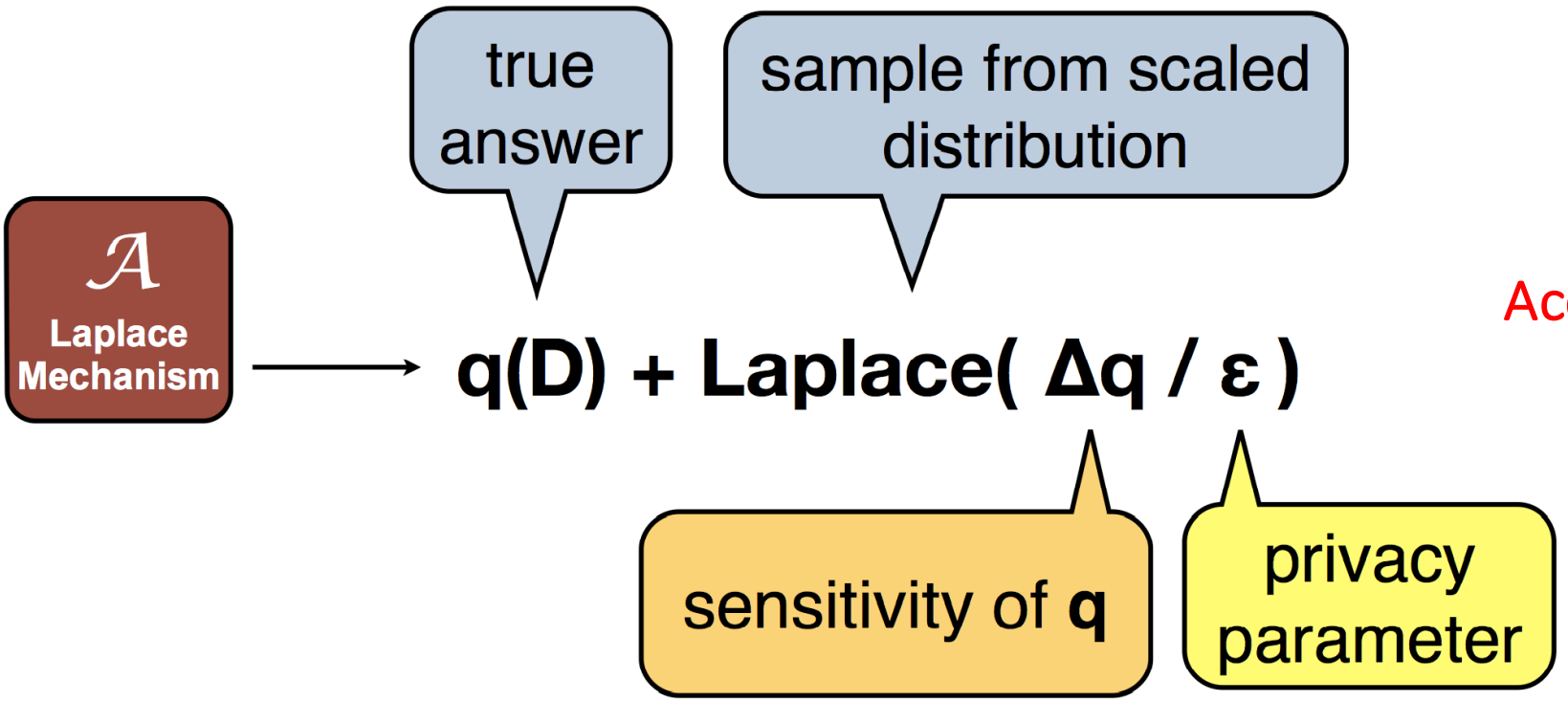
Differential privacy (Dwork 2006)

$\mathcal{K}$ gives ε–differential privacy if for all values of DB, DB' differing in a single element, and all $\mathbf{S}$ in $\mathbf{Range}(\mathcal{K})$

$$\frac{\Pr[\,\mathcal{K}(\mathrm{DB}) \text{ in } \mathbf{S}]}{\Pr[\,\mathcal{K}(\mathrm{DB'}) \text{ in } \mathbf{S}]} \leq e^{\varepsilon} \sim (1+\varepsilon)$$



Differential privacy (Laplacian noise)



Accuracy loss!

Source: Gerome Miklau and Michael Hay

NAPLES project's goal

Develop theoretical foundations for implementing tools that

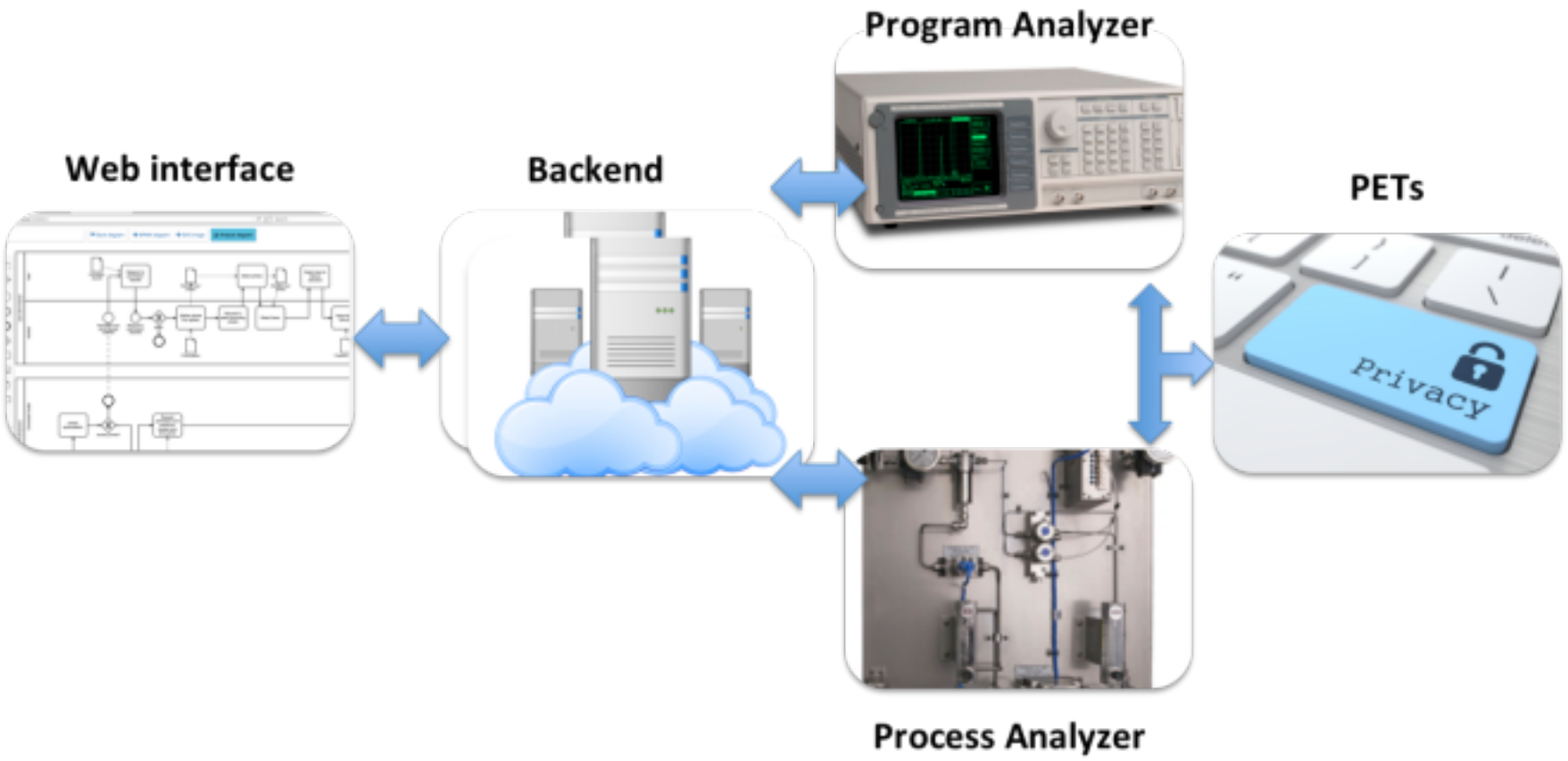
- Let one model stakeholders and flows in the Business Process Model and Notation (BPMN)
- Find data leaks in these process models, taking into account the Privacy-Enhancing Technologies used in the as-is models
- Quantify leakages using differential privacy
- Quantify accuracy loss
- Suggest relevant privacy-enhancing technologies to reduce privacy leaks

See http://pleak.io

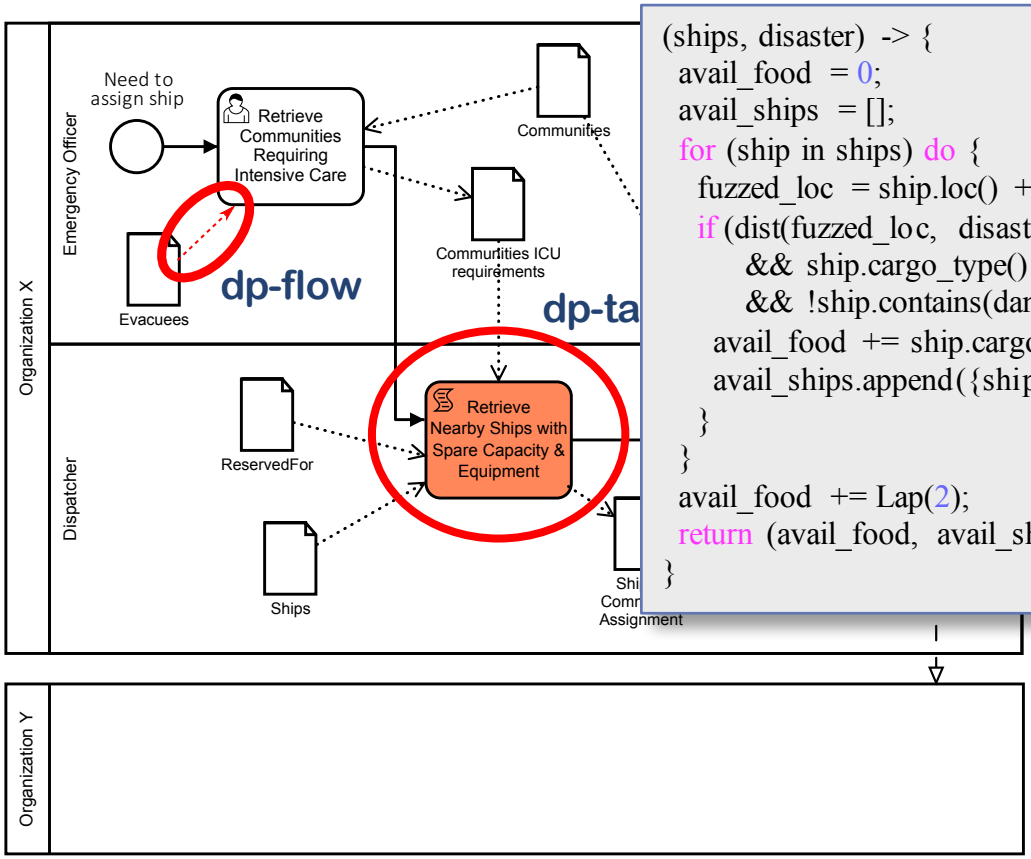
Usage scenarios

- Support privacy audit of existing system
  - What will each stakeholder of the System learn about a private data object? e.g. with respect to differential privacy
- Build a new privacy-aware system
  - What will each stakeholder of the System learn about each private data object?
  - Which Privacy Enhancing Technologies would help reducing the leakage?

Architecture

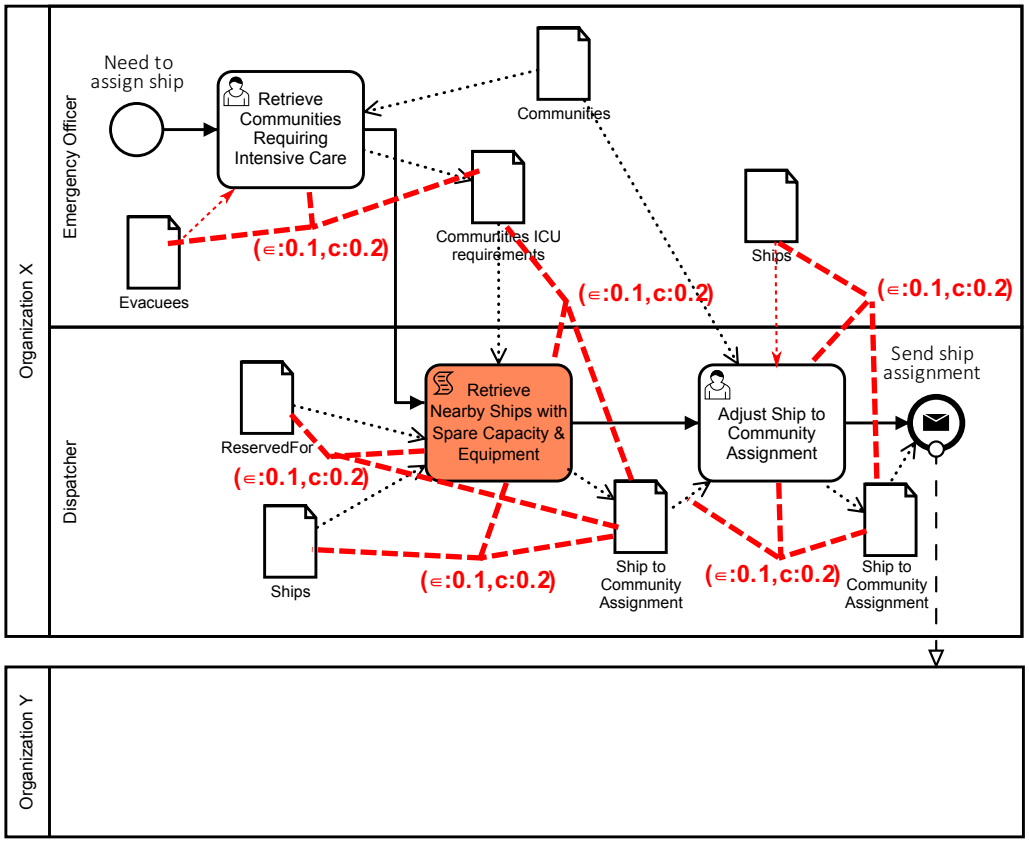


Adding privacy-enhancing technologies

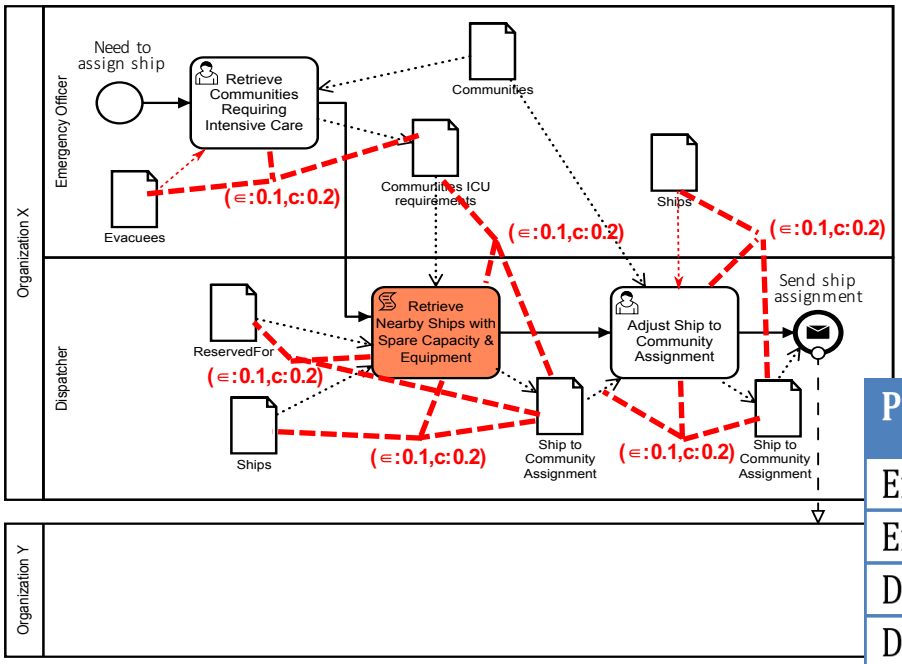



```
(ships, disaster) -> {
  avail_food = 0;
  avail_ships = [];
  for (ship in ships) do {
    fuzzed_loc = ship.loc() + Lap²(3);
    if (dist(fuzzed_loc, disaster.loc()) / ship.speed() <= 2
        && ship.cargo_type() == "food"
        && !ship.contains(dangerous_materials) ) {
      avail_food += ship.cargo();
      avail_ships.append({ship.name(), fuzzed_loc});
    }
  }
  avail_food += Lap(2);
  return (avail_food, avail_ships);
}
```

Annotating the model with DF and sensitivity bounds

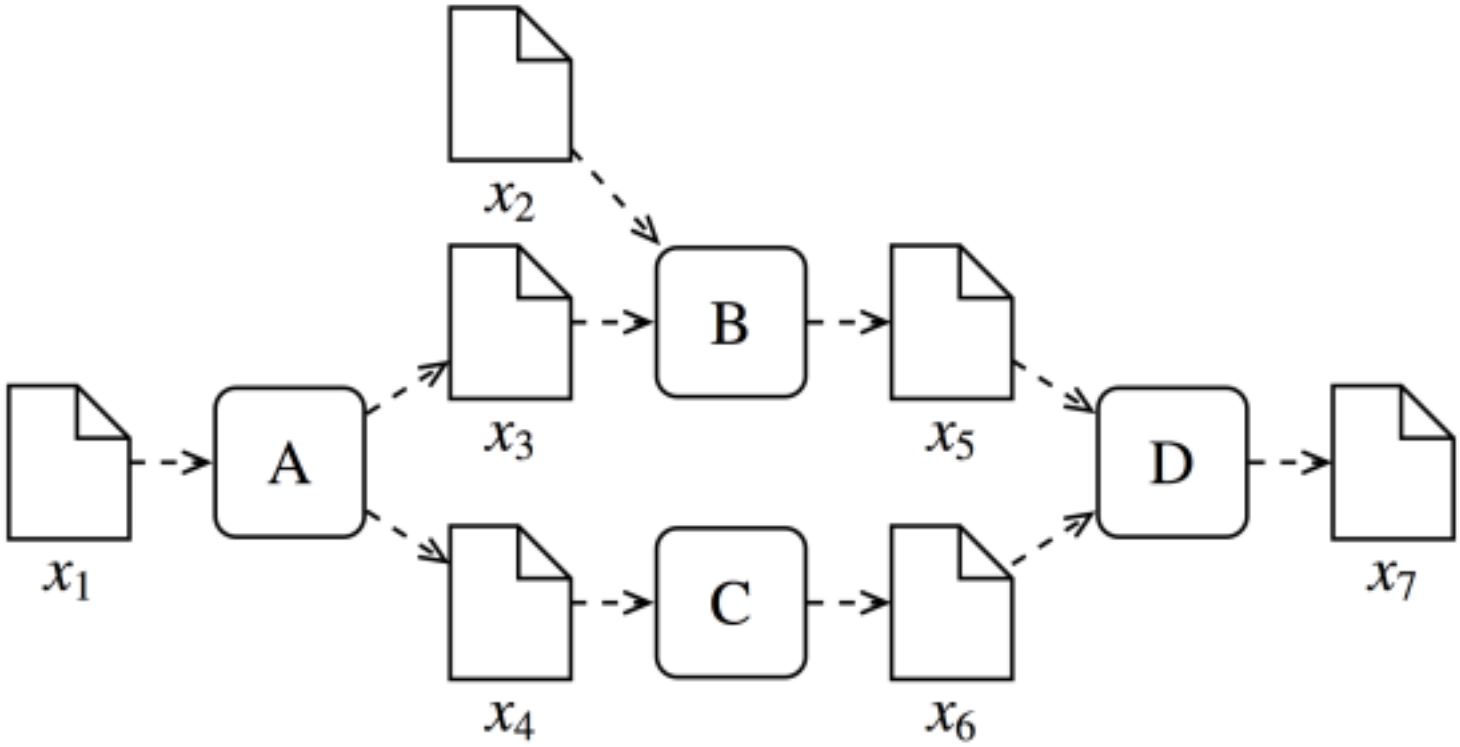


Differential Privacy Disclosure (Roles)

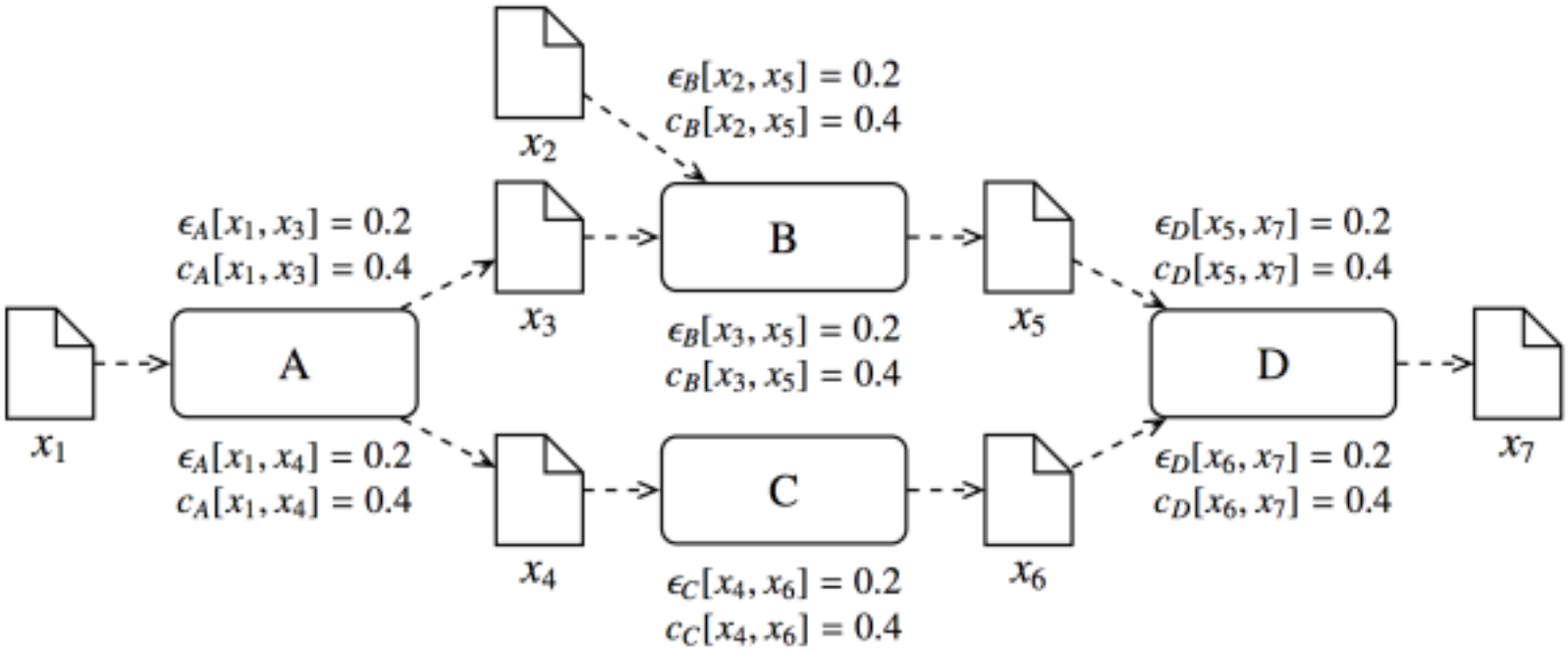


Party	Data Collection	Differential Privacy (ε)
Emergency Officer	Evacuees	0.1
Emergency Officer	Communities	Full disclosure
Dispatcher	Evacuees	min(0.1, 0.1*0.2) = 0.02
Dispatcher	ReservedFor	0.2
Dispatcher	Ships	0.1 + 0.1 = 0.2
Dispatcher	Communities	Full disclosure
Organization Y	Evacuees	min(0.1, 0.1*0.2) = 0.02
Organization Y	Ships	0.1 + 0.1 = 0.2
Organization Y	ReservedFor	0.2

Data processing workflows



Model with DF/sensitivity bounds



Generalized sensitivity

- Generalized distances – any partial order with addition and least element

$$d_X : X \times X \to V_X$$

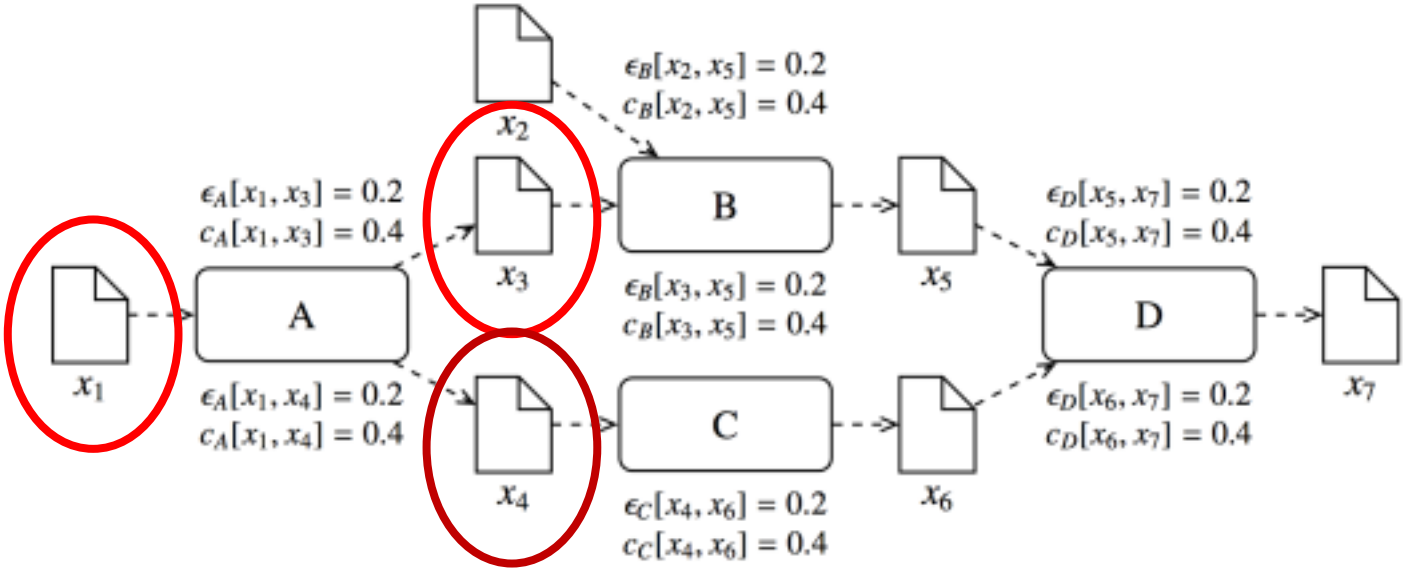
- $f : X \to Y$ has sensitivity $c_f : V_X \to V_Y$
- Differential privacy is a specific case of generalized sensitivity

$$d_{\mathrm{dp}}(\chi, \chi') = \sup_{y \in Y} \left| \ln(\chi(y)/\chi'(y)) \right|$$

- Generalized sensitivity is composable, i.e. $c_{f \circ g} = c_f \cdot c_g$

Proposition 2. *For $i \in \{1, \ldots, n\}$, let $f_i : X \to Y_i$ be c_i-sensitive with respect to the distances d_X on X and d_{Y_i} on Y_i. Let $f' : Y_1 \times \cdots \times Y_n \to Z$ be c'_i-sensitive with respect to the distances d_{Y_i} on Y_i and d_Z on Z (for all $i \in \{1, \ldots, n\}$). Then the mapping $g : X \to Z$, defined by $g(x) = f'(f_1(x), \ldots, f_n(x))$, is $\sum_{i=1}^{n} c_i c'_i$-sensitive with respect to the distances d_X on X and d_Z on Z.*

Model with DF/sensitivity bounds



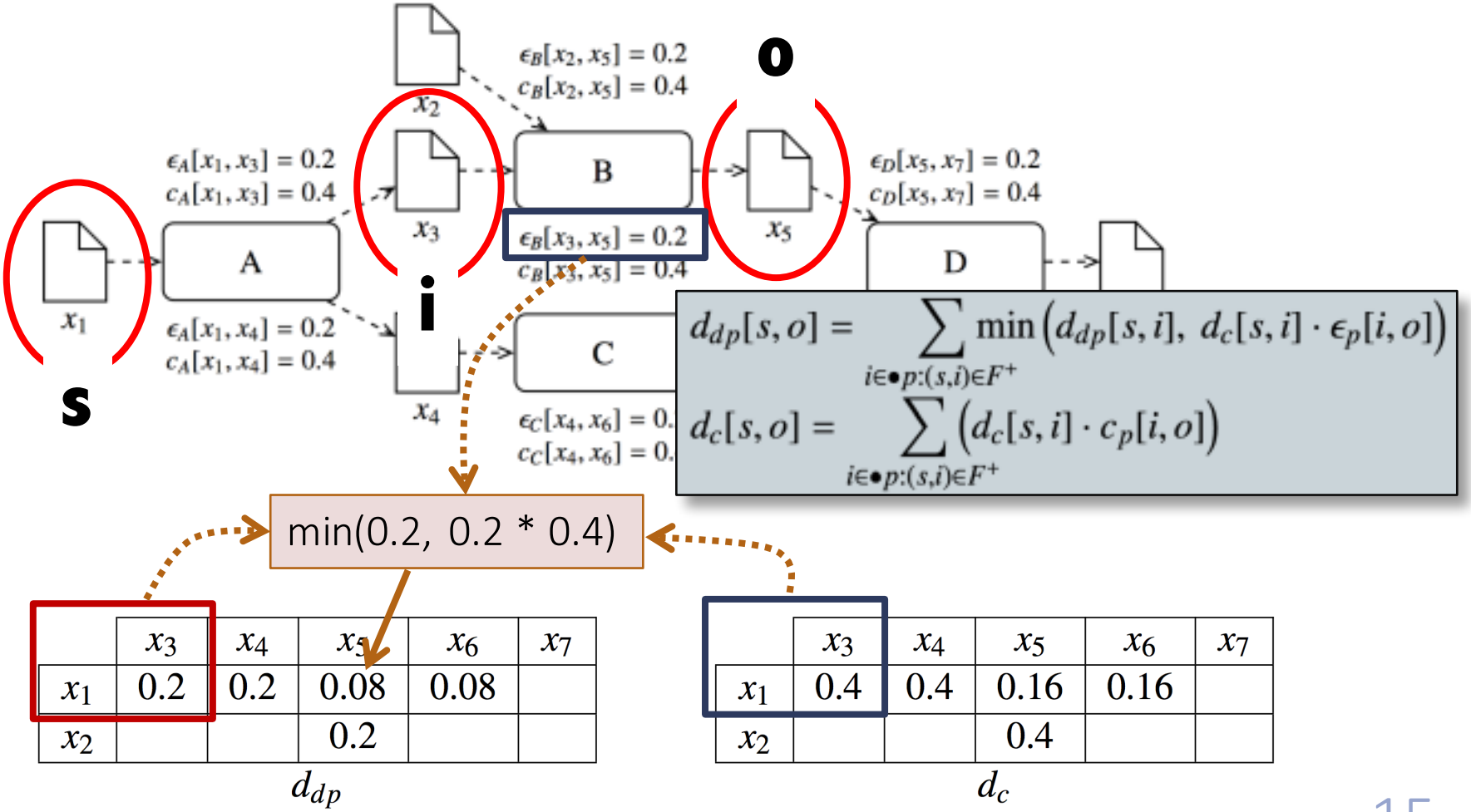
	x_3	x_4	x_5	x_6	x_7
x_1	$\epsilon_A[x_1, x_3] = 0.2$	$\epsilon_A[x_1, x_4] = 0.2$			
x_2					

d_{dp}

	x_3	x_4	x_5	x_6	x_7
x_1	$c_A[x_1, x_3] = 0.4$	$c_A[x_1, x_4] = 0.4$			
x_2					

d_c

Model with DF/sensitivity bounds



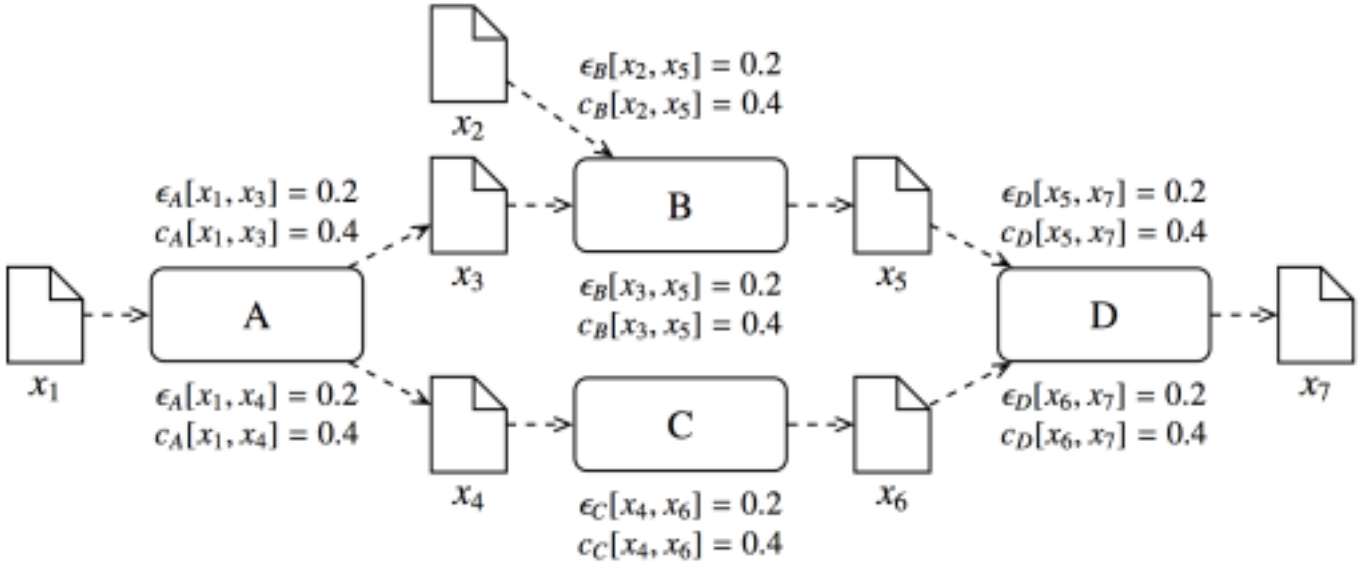
	x_3	x_4	x_5	x_6	x_7
x_1	0.2	0.2	0.08	0.08	
x_2			0.2		

d_{dp}

	x_3	x_4	x_5	x_6	x_7
x_1	0.4	0.4	0.16	0.16	
x_2			0.4		

d_c

Result of analysis



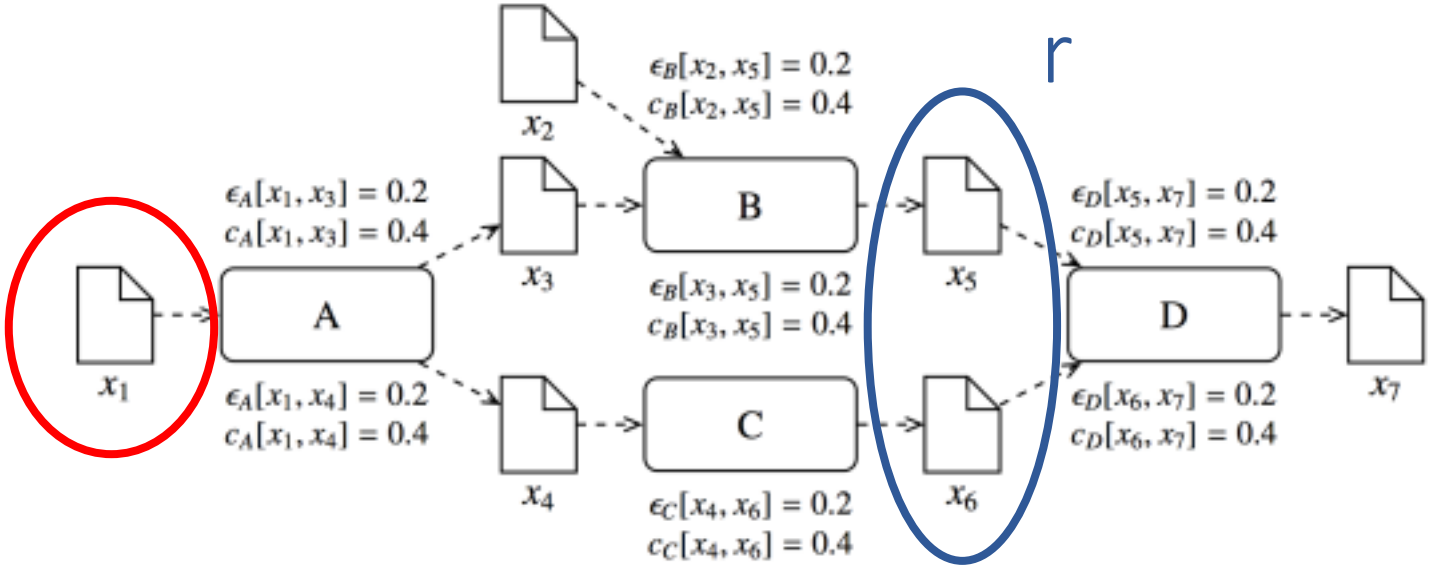
	x_3	x_4	x_5	x_6	x_7
x_1	0.2	0.2	0.08	0.08	0.064
x_2			0.2		0.16

d_{dp}

	x_3	x_4	x_5	x_6	x_7
x_1	0.4	0.4	0.16	0.16	0.128
x_2			0.4		0.08

d_c

Differential Privacy Disclosure of a Data Source to a Party



$$\epsilon_r(x_1) = d_{dp}[x_1, x_5] + d_{dp}[x_1, x_6]$$
$$= 0.08 + 0.08 = 0.16$$

Outlook

- Extend privacy analyzer to cover a broader class of BPMN process models
  - E.g. adding conditional branching
- Principles of program analysis for DP
  - For arbitrary generalized metrics and sensitivities
- Defining a super set of BPMN, with ad-hoc constructs to model privacy related concerns (i.e. PA-BPMN)
- Building the PETs library & extend PA-BPMN to cater for other PETs besides differential privacy

UNIVERSITY OF TARTU

Thanks!

Research funded by DARPA (Brandeis program 2015-2019)