AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

ICAISC Zakopane
International Conference on Artificial Intelligence and Soft Computing
June 1-5, 2014

Belief propagation during data integration in a P2P network

Piotr Szwed

AGH University of Science and Technology

Department of Applied Computer Science

e-mail: pszwed@agh.edu.pl

Agenda

1. Introduction and motivation
2. Communication within P2P integration platform
3. Linear algebra model
    1. State and communication matrix
    2. Closure of communication graph
    3. Operators
4. Example
5. Conclusions

Introduction (1)

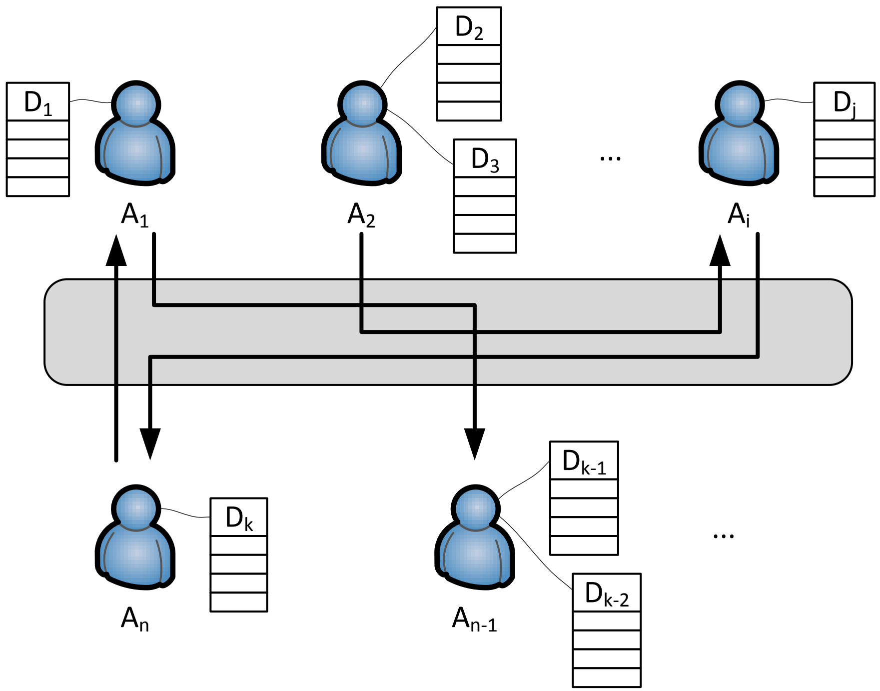
- Use cases for data integration :
  - company mergers or acquisitions
  - bioinformatics (Caragea et al, 2005)
  - coordination of military systems (Tolk, Muguira, 2003)
  - crime and intelligence analysis (Chen, Wang 2005)
- Integration architectures (Cruz, Xiao 2009)
  - Central repository or data warehouse
  - Software platform:
    - Centralized (using a mediator service)
    - P2P – each peer represents autonomous information system

Introduction (2)

- Schema mappings
  - Centralized (Lenzerini, 2002)
    - GaV (Global as View)
    - LaV (Local as View)
  - P2P: mappings between pairs of agents or a global ontology (Arenas et al. 2003, Calvanese 2004)
- Epistemic logic can be used to describe states (beliefs) of communicating agents
  - Semantics of P2P data integration systems (Calvanese et al. 2004)
  - Reason about communication graphs (Pacuit, Parikh, 2007)
  - Linear algebra models (Liau 2004; Tojo 2013)

Motivation

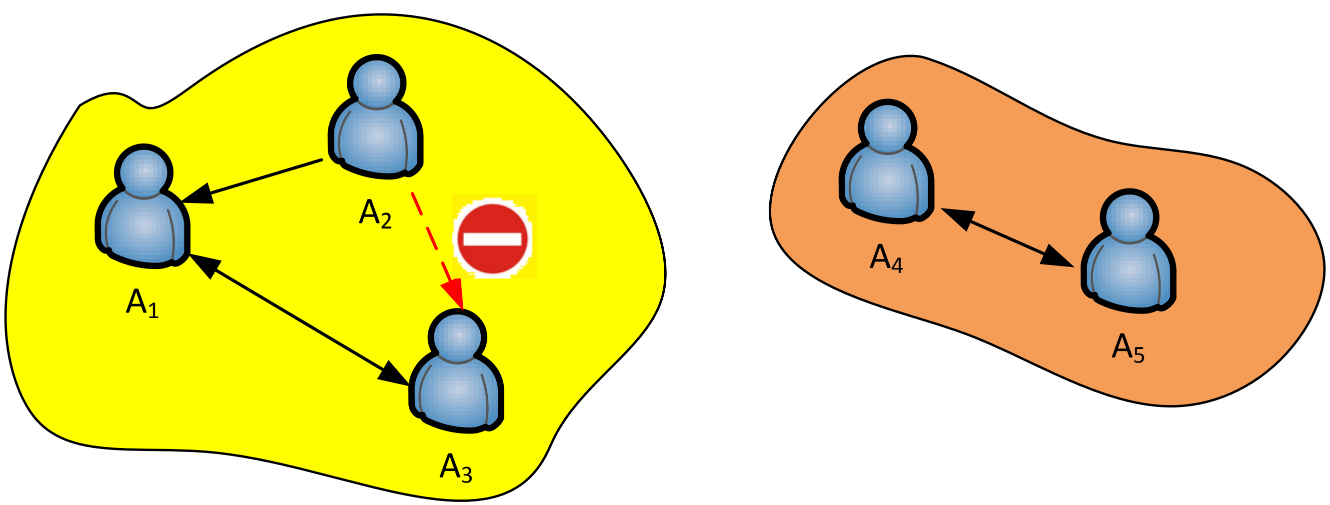
Specification and design of a platform enabling data integration between various security and law enforcement agencies.



- Several organizations ($A_1 \ldots A_n$) are responsible for collecting data and keeping them in local repositories.
- Agents $A_1 \ldots A_n$ form a P2P network.
- Restrictions on information exchange (law regulations or bilateral contracts).
- Security and confidentiality requirements.

Problem statement

- We start with specifications of communication channels: **who sends what to whom** (locally).
- **Problem 1**: Which *data types* should an agent be aware of to implement correctly communiaction interfaces?
- **Problem 2**: Is it possible to detect:
  - unintended information *leakage*
  - unintended *silos or islands* of information

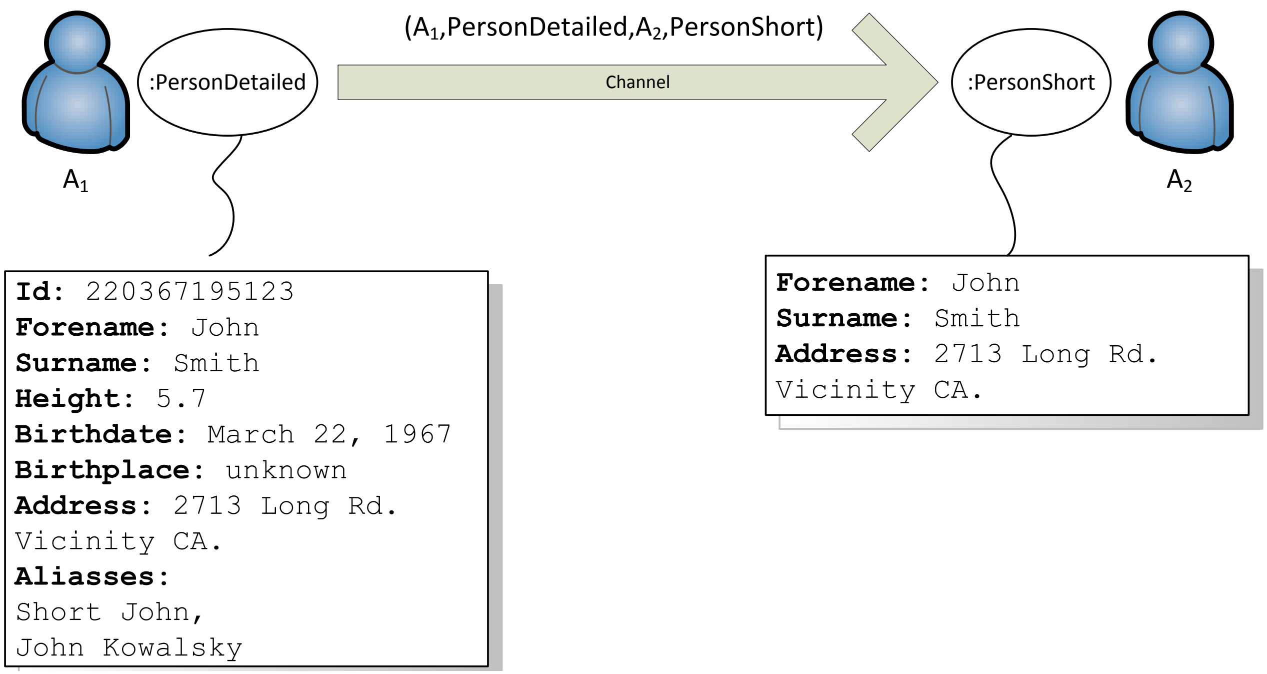


Solution outline

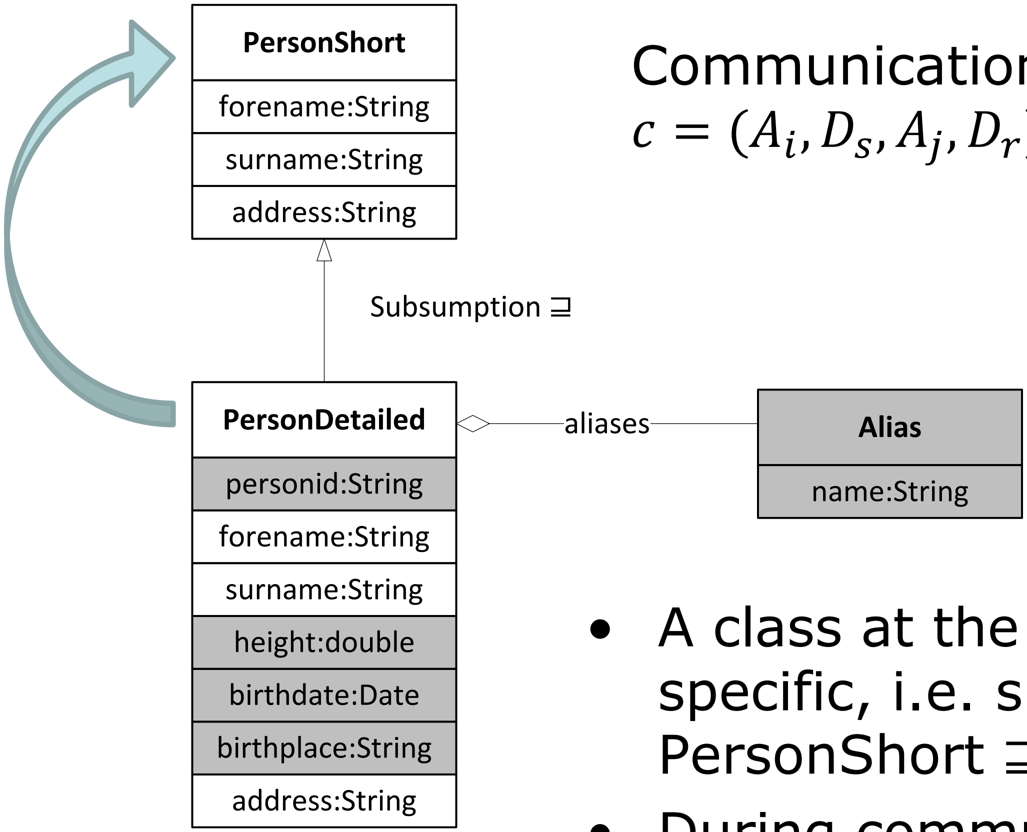
- Focus on exchanged data types (classes)
- Assumption that all data belongs to a global schema
- Statement: agent A_i **knows class** D_j is a part of global belief state
- The belief may be changed due to defined information flows.
- We use a linear algebra model for belief states and their updates (an extension to Stoshi Tojo’s model of epistemic logic).

Comunication

- Agents $A = \{A_1 \dots A_n\}$ are linked by communication channels $c_1, \dots, c_m$ and exchange data of types (classes) $D = \{D_1, \dots D_k\}$
- During communication agents expose only parts of data objects, e.g. **PersonDetailed** is converetd to **PersonShort**



Comunication - upcasting

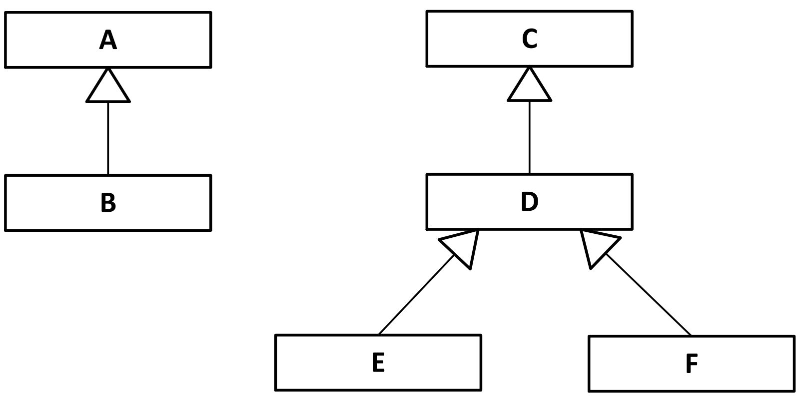


Communication channels are tuples
$c = (A_i, D_s, A_j, D_r)$

- A class at the channel output is less specific, i.e. subsumes the input class: PersonShort ⊒ PersonDetailed
- During communication data objects are **upcast** from D_s to D_r

Upcast matrix

Closure of subsumption relation $\sqsupseteq$ is repesented by $|D| \times |D|$ *upcast matrix* U of boolean values.



$$U = \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} T & T & F & F & F & F \\ F & T & F & F & F & F \\ F & F & T & T & T & T \\ F & F & F & T & T & T \\ F & F & F & F & T & F \\ F & F & F & F & F & T \end{pmatrix} \end{array} \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}$$

Linear algebra model

System state is as an assignment of sets of classes to agents

- Encoded as a $|D| \times |A|$ matrix $S = \left[s_j^i\right]$ of boolean values
- $s_j^i = T$ if an agent A_j is aware of the D_i class existence

Communication matrix

- The set of channels $C \subset A \times D \times A \times D$ is encoded as 4-dimensional communication matrix: $E = [e_{ki}^{lj}]$.

$$e_{ki}^{lj} = \begin{cases} T, & \text{if } (A_i, D_j, A_k, D_l) \in C \\ T, & \text{if } l = i \text{ and } k = j \\ F, & \text{otherwise} \end{cases}$$

Linear algebra model - state equation

Belief updates are described by the state equation:

$$S(m+1) = E \circ S(m),$$

where

$$s_k^l(m+1) = \bigvee_i \bigwedge_j e_{ki}^{lj} \, s_j^i(m)$$

Proposition 1.

The sequence $\sigma = S(0), S(1), \dots, S(n), \dots$, where $S(i+1) = E \circ S(i)$ converges.

Consequence: If we assume, what an agent knows, i.e. which types of data it stores, we may conclude how far this information can be propagated throughout the network.

This allows for detecting information silos or islands of belief.

Linear algebra model - closure of the communication graph

The state equation can be expressed as

$S(i) = E^i \circ S(0)$, where $\mathrm{E}^i = E \otimes E \otimes \cdots E$ (i-times)

Operator $\otimes$ multiplying two communication matrices E and G is given by the formula:

$$f_{mn}^{kl} = e_{ij}^{kl} \, g_{mn}^{ji}$$

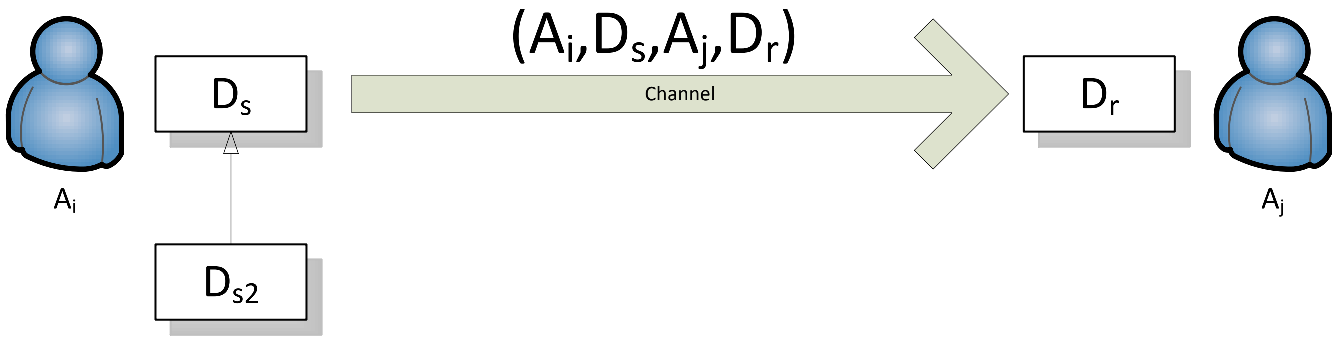
Proposition 2:

The sequence $\epsilon = E, E^2, \dots E^i \dots$ converges.

Consequence: The matrix E^* is a closure of the communication graph. It describes derived channels:

- Shortcuts improving cooperation are possible
- Unintended leaks can be detected

Linear algebra model – upcasting on input



- Agents A_i and A_j are linked by a channel $n = (A_i, D_s, A_j, D_r)$
- Agent A_i is aware of a class D_{s2} satisfying $D_s \sqsupseteq D_{s2}$.
- Agent A_i can upcast object of D_{s2} to D_s and transmit it through the channel n.

Reformulation

- Modified state equation:

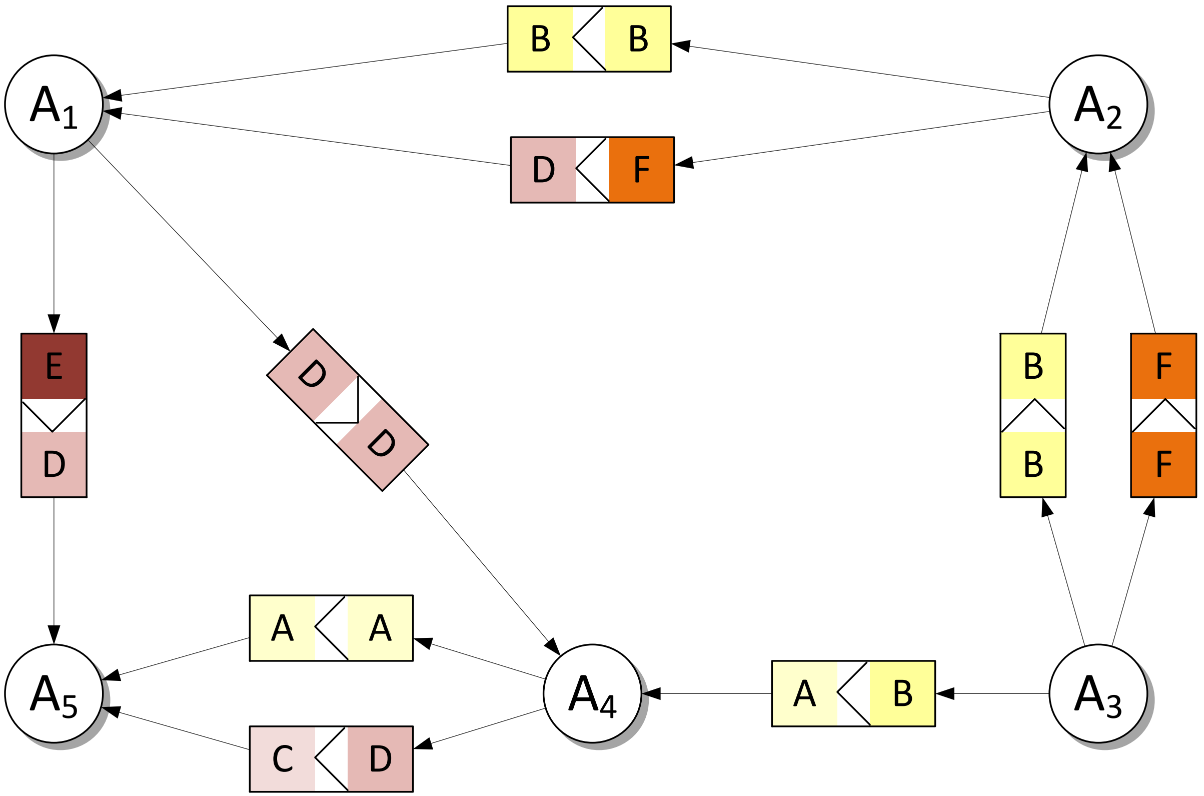
$$S(m+1) = (E \odot U) \circ S(m)$$

- The operator $\odot$ is defined as:

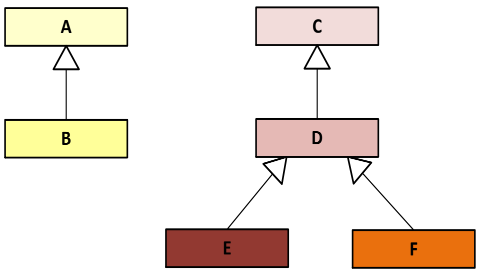
$$f_{mj}^{kl} = e_{mi}^{kl} \, u_j^i$$

Example

Communication system



Class hierarchy

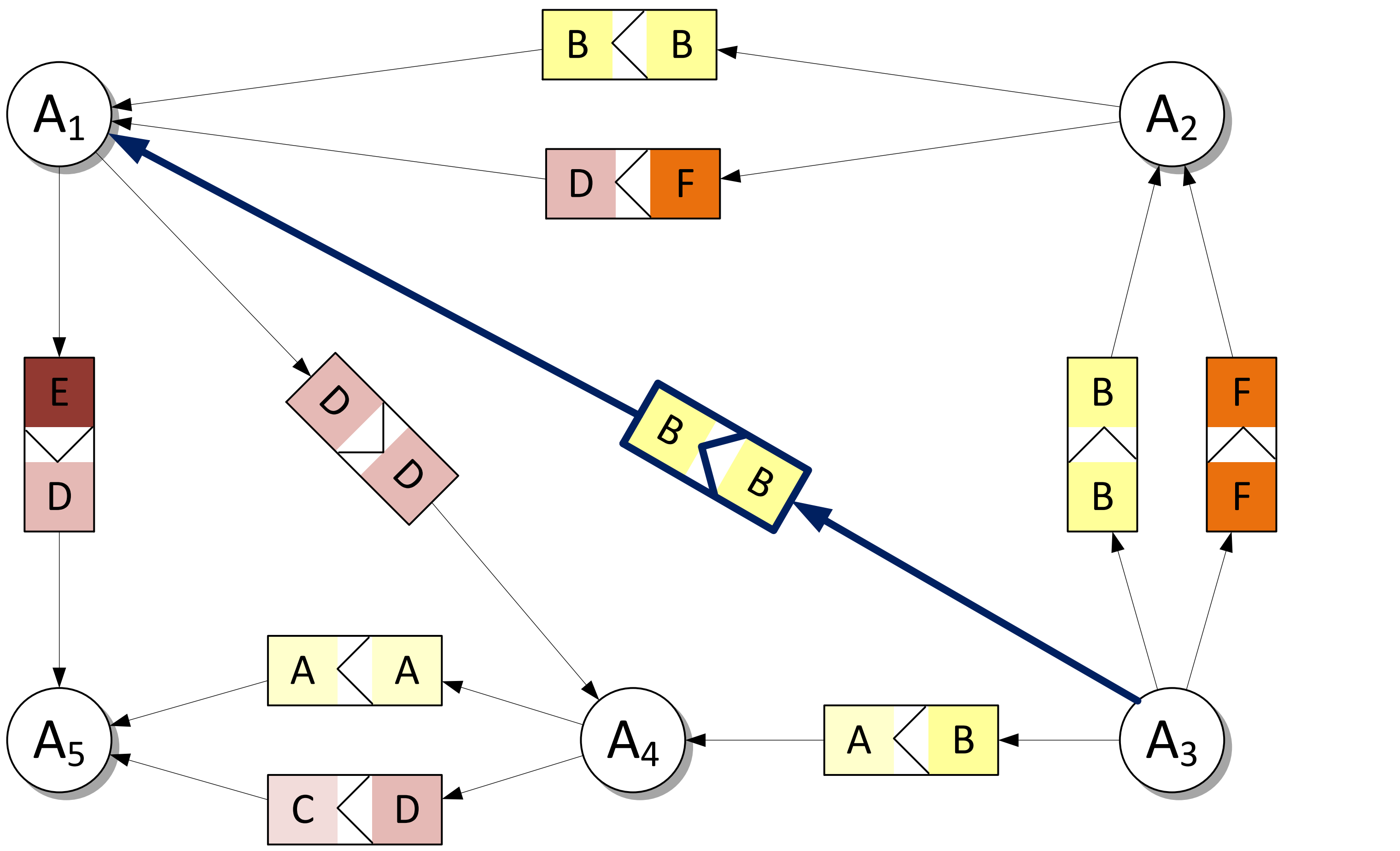


Closures of communication graph 1

E^*	E^*U	UE^*	$(EU)^*$
1. $A_1 \to (D) \to (D) \to A_4$	1. $A_1 \to (D) \to (D) \to A_4$	1. $A_1 \to (D) \to (C) \to A_4$	1. $A_1 \to (D) \to (C) \to A_4$
2. $A_1 \to (D) \to (C) \to A_5$	2. $A_1 \to (D) \to (C) \to A_5$	2. $A_1 \to (D) \to (D) \to A_4$	2. $A_1 \to (D) \to (D) \to A_4$
3. $A_1 \to (E) \to (D) \to A_5$	3. $A_1 \to (E) \to (D) \to A_4$	3. $A_1 \to (D) \to (C) \to A_5$	3. $A_1 \to (D) \to (C) \to A_5$
4. $A_2 \to (B) \to (B) \to A_1$	4. $A_1 \to (E) \to (C) \to A_5$	4. $A_1 \to (E) \to (C) \to A_5$	4. $A_1 \to (E) \to (C) \to A_4$
5. $A_2 \to (F) \to (D) \to A_1$	5. $A_1 \to (E) \to (D) \to A_5$	5. $A_1 \to (E) \to (D) \to A_5$	5. $A_1 \to (E) \to (D) \to A_4$
6. $A_2 \to (F) \to (D) \to A_4$	6. $A_1 \to (F) \to (D) \to A_4$	6. $A_2 \to (B) \to (A) \to A_1$	6. $A_1 \to (E) \to (C) \to A_5$
7. $A_2 \to (F) \to (C) \to A_5$	7. $A_1 \to (F) \to (C) \to A_5$	7. $A_2 \to (B) \to (B) \to A_1$	7. $A_1 \to (E) \to (D) \to A_5$
8. $A_3 \to (B) \to (B) \to A_1$	8. $A_2 \to (B) \to (B) \to A_1$	8. $A_2 \to (F) \to (C) \to A_1$	8. $A_1 \to (F) \to (C) \to A_4$
9. $A_3 \to (B) \to (B) \to A_2$	9. $A_2 \to (F) \to (D) \to A_1$	9. $A_2 \to (F) \to (D) \to A_1$	9. $A_1 \to (F) \to (D) \to A_4$
10. $A_3 \to (B) \to (A) \to A_4$	10. $A_2 \to (F) \to (D) \to A_4$	10. $A_2 \to (F) \to (C) \to A_4$	10. $A_1 \to (F) \to (C) \to A_5$
11. $A_3 \to (B) \to (A) \to A_5$	11. $A_2 \to (F) \to (C) \to A_5$	11. $A_2 \to (F) \to (D) \to A_4$	11. $A_2 \to (B) \to (A) \to A_1$
12. $A_3 \to (F) \to (D) \to A_1$	12. $A_3 \to (B) \to (B) \to A_1$	12. $A_2 \to (F) \to (C) \to A_5$	12. $A_2 \to (B) \to (B) \to A_1$
13. $A_3 \to (F) \to (F) \to A_2$	13. $A_3 \to (B) \to (B) \to A_2$	13. $A_3 \to (B) \to (A) \to A_1$	13. $A_2 \to (F) \to (C) \to A_1$
14. $A_3 \to (F) \to (D) \to A_4$	14. $A_3 \to (B) \to (A) \to A_4$	14. $A_3 \to (B) \to (B) \to A_1$	14. $A_2 \to (F) \to (D) \to A_1$
15. $A_3 \to (F) \to (C) \to A_5$	15. $A_3 \to (B) \to (A) \to A_5$	15. $A_3 \to (B) \to (A) \to A_2$	15. $A_2 \to (F) \to (C) \to A_4$
16. $A_4 \to (A) \to (A) \to A_5$	16. $A_3 \to (F) \to (D) \to A_1$	16. $A_3 \to (B) \to (B) \to A_2$	16. $A_2 \to (F) \to (D) \to A_4$
17. $A_4 \to (D) \to (C) \to A_5$	17. $A_3 \to (F) \to (F) \to A_2$	17. $A_3 \to (B) \to (A) \to A_4$	17. $A_2 \to (F) \to (C) \to A_5$
	18. $A_3 \to (F) \to (D) \to A_4$	18. $A_3 \to (B) \to (A) \to A_5$	18. $A_3 \to (B) \to (A) \to A_1$
	19. $A_3 \to (F) \to (C) \to A_5$	19. $A_3 \to (F) \to (C) \to A_1$	19. $A_3 \to (B) \to (B) \to A_1$
	20. $A_4 \to (A) \to (A) \to A_5$	20. $A_3 \to (F) \to (D) \to A_1$	20. $A_3 \to (B) \to (A) \to A_2$
	21. $A_4 \to (B) \to (A) \to A_5$	21. $A_3 \to (F) \to (C) \to A_2$	21. $A_3 \to (B) \to (B) \to A_2$
	22. $A_4 \to (D) \to (C) \to A_5$	22. $A_3 \to (F) \to (D) \to A_2$	22. $A_3 \to (B) \to (A) \to A_4$
	23. $A_4 \to (E) \to (C) \to A_5$	23. $A_3 \to (F) \to (F) \to A_2$	23. $A_3 \to (B) \to (A) \to A_5$
	24. $A_4 \to (F) \to (C) \to A_5$	24. $A_3 \to (F) \to (C) \to A_4$	24. $A_3 \to (F) \to (C) \to A_1$
		25. $A_3 \to (F) \to (D) \to A_4$	25. $A_3 \to (F) \to (D) \to A_1$
		26. $A_3 \to (F) \to (C) \to A_5$	26. $A_3 \to (F) \to (C) \to A_2$
		27. $A_4 \to (A) \to (A) \to A_5$	27. $A_3 \to (F) \to (D) \to A_2$
		28. $A_4 \to (D) \to (C) \to A_5$	28. $A_3 \to (F) \to (F) \to A_2$
			29. $A_3 \to (F) \to (C) \to A_4$
			30. $A_3 \to (F) \to (D) \to A_4$
			31. $A_3 \to (F) \to (C) \to A_5$
			32. $A_4 \to (A) \to (A) \to A_5$
			33. $A_4 \to (B) \to (A) \to A_5$
			34. $A_4 \to (D) \to (C) \to A_5$
			35. $A_4 \to (E) \to (C) \to A_5$
			36. $A_4 \to (F) \to (C) \to A_5$

Prototype software was used to compute the closures

Closures of communication graph 2
Possible shortcuts…



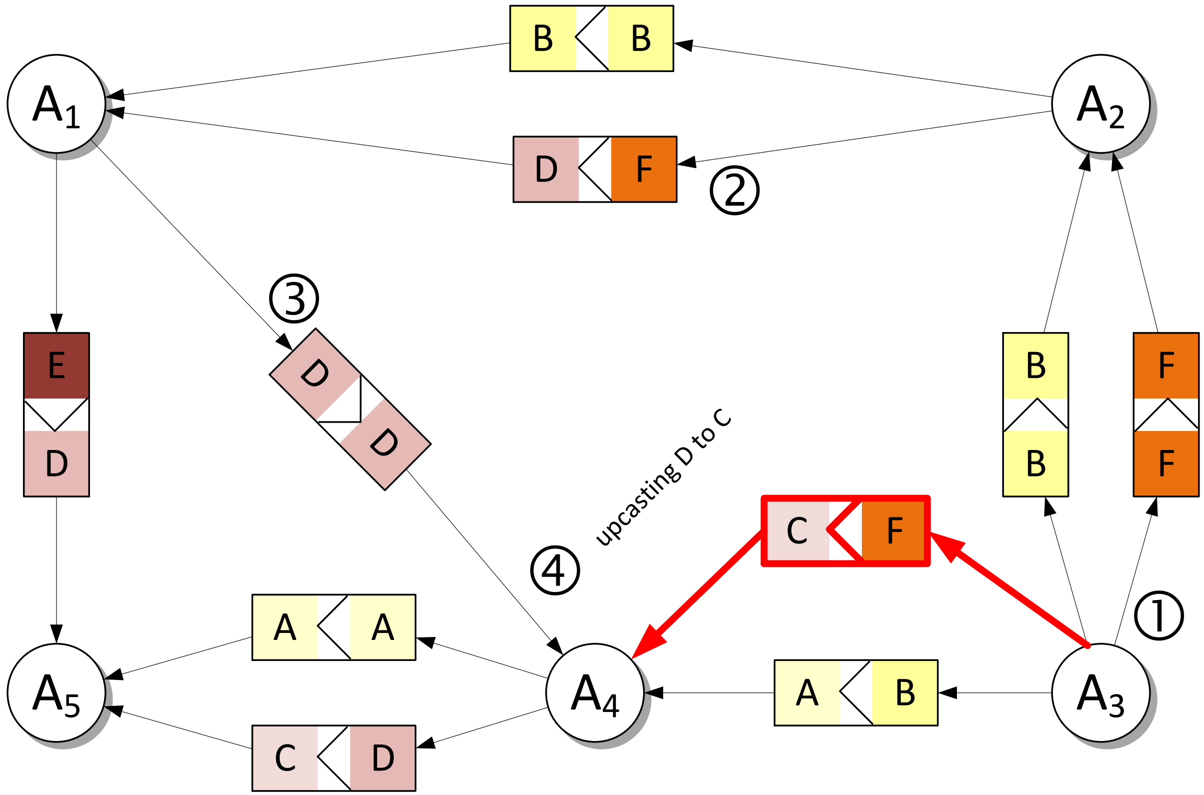
8. $A_3 \rightarrow (B) \rightarrow (B) \rightarrow A_1$

12. $A_3 \rightarrow (B) \rightarrow (B) \rightarrow A_1$

14. $A_3 \rightarrow (B) \rightarrow (B) \rightarrow A_1$

19. $A_3 \rightarrow (B) \rightarrow (B) \rightarrow A_1$

Closures of communication graph 3
Forbidden channel circumvented…

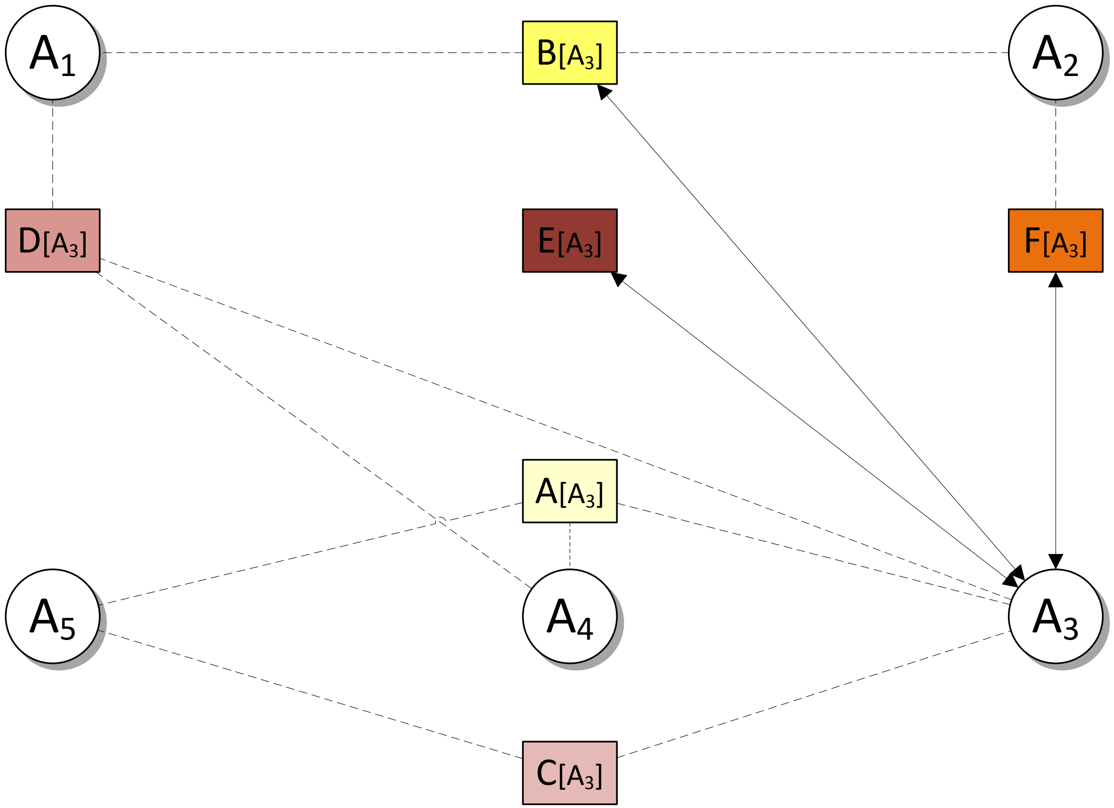


- Forbidden channel (A_3, F, C, A_4) can be circumvented by (A_3, F, F, A_2), (A_2, F, D, A_1), (A_1, D, D, A_4) and finally upcasting.
- (A_3, F, C, A_4) appears in the closures UE^* and $(EU)^*$

Channel (A_3, F, C, A_4)

E^*	E^*U	UE^*	$(EU)^*$
1. $A_1 \to (D) \to (D) \to A_4$	1. $A_1 \to (D) \to (D) \to A_4$	1. $A_1 \to (D) \to (C) \to A_4$	1. $A_1 \to (D) \to (C) \to A_4$
2. $A_1 \to (D) \to (C) \to A_5$	2. $A_1 \to (D) \to (C) \to A_5$	2. $A_1 \to (D) \to (D) \to A_4$	2. $A_1 \to (D) \to (D) \to A_4$
3. $A_1 \to (E) \to (D) \to A_5$	3. $A_1 \to (E) \to (D) \to A_4$	3. $A_1 \to (D) \to (C) \to A_5$	3. $A_1 \to (D) \to (C) \to A_5$
4. $A_2 \to (B) \to (B) \to A_1$	4. $A_1 \to (E) \to (C) \to A_5$	4. $A_1 \to (E) \to (C) \to A_5$	4. $A_1 \to (E) \to (C) \to A_4$
5. $A_2 \to (F) \to (D) \to A_1$	5. $A_1 \to (E) \to (D) \to A_5$	5. $A_1 \to (E) \to (D) \to A_5$	5. $A_1 \to (E) \to (D) \to A_4$
6. $A_2 \to (F) \to (D) \to A_4$	6. $A_1 \to (F) \to (D) \to A_4$	6. $A_2 \to (B) \to (A) \to A_1$	6. $A_1 \to (E) \to (C) \to A_5$
7. $A_2 \to (F) \to (C) \to A_5$	7. $A_1 \to (F) \to (C) \to A_5$	7. $A_2 \to (B) \to (B) \to A_1$	7. $A_1 \to (E) \to (D) \to A_5$
8. $A_3 \to (B) \to (B) \to A_1$	8. $A_2 \to (B) \to (B) \to A_1$	8. $A_2 \to (F) \to (C) \to A_1$	8. $A_1 \to (F) \to (C) \to A_4$
9. $A_3 \to (B) \to (B) \to A_2$	9. $A_2 \to (F) \to (D) \to A_1$	9. $A_2 \to (F) \to (D) \to A_1$	9. $A_1 \to (F) \to (D) \to A_4$
10. $A_3 \to (B) \to (A) \to A_4$	10. $A_2 \to (F) \to (D) \to A_4$	10. $A_2 \to (F) \to (C) \to A_4$	10. $A_1 \to (F) \to (C) \to A_5$
11. $A_3 \to (B) \to (A) \to A_5$	11. $A_2 \to (F) \to (C) \to A_5$	11. $A_2 \to (F) \to (D) \to A_4$	11. $A_2 \to (B) \to (A) \to A_1$
12. $A_3 \to (F) \to (D) \to A_1$	12. $A_3 \to (B) \to (B) \to A_1$	12. $A_2 \to (F) \to (C) \to A_5$	12. $A_2 \to (B) \to (B) \to A_1$
13. $A_3 \to (F) \to (F) \to A_2$	13. $A_3 \to (B) \to (B) \to A_2$	13. $A_3 \to (B) \to (A) \to A_1$	13. $A_2 \to (F) \to (C) \to A_1$
14. $A_3 \to (F) \to (D) \to A_4$	14. $A_3 \to (B) \to (A) \to A_4$	14. $A_3 \to (B) \to (B) \to A_1$	14. $A_2 \to (F) \to (D) \to A_1$
15. $A_3 \to (F) \to (C) \to A_5$	15. $A_3 \to (B) \to (A) \to A_5$	15. $A_3 \to (B) \to (A) \to A_2$	15. $A_2 \to (F) \to (C) \to A_4$
16. $A_4 \to (A) \to (A) \to A_5$	16. $A_3 \to (F) \to (D) \to A_1$	16. $A_3 \to (B) \to (B) \to A_2$	16. $A_2 \to (F) \to (D) \to A_4$
17. $A_4 \to (D) \to (C) \to A_5$	17. $A_3 \to (F) \to (F) \to A_2$	17. $A_3 \to (B) \to (A) \to A_4$	17. $A_2 \to (F) \to (C) \to A_5$
	18. $A_3 \to (F) \to (D) \to A_4$	18. $A_3 \to (B) \to (A) \to A_5$	18. $A_3 \to (B) \to (A) \to A_1$
	19. $A_3 \to (F) \to (C) \to A_5$	19. $A_3 \to (F) \to (C) \to A_1$	19. $A_3 \to (B) \to (B) \to A_1$
	20. $A_4 \to (A) \to (A) \to A_5$	20. $A_3 \to (F) \to (D) \to A_1$	20. $A_3 \to (B) \to (A) \to A_2$
	21. $A_4 \to (B) \to (A) \to A_5$	21. $A_3 \to (F) \to (C) \to A_2$	21. $A_3 \to (B) \to (B) \to A_2$
	22. $A_4 \to (D) \to (C) \to A_5$	22. $A_3 \to (F) \to (D) \to A_2$	22. $A_3 \to (B) \to (A) \to A_4$
	23. $A_4 \to (E) \to (C) \to A_5$	23. $A_3 \to (F) \to (F) \to A_2$	23. $A_3 \to (B) \to (A) \to A_5$
	24. $A_4 \to (F) \to (C) \to A_5$	24. $A_3 \to (F) \to (C) \to A_4$	24. $A_3 \to (F) \to (C) \to A_1$
		25. $A_3 \to (F) \to (D) \to A_4$	25. $A_3 \to (F) \to (D) \to A_1$
		26. $A_3 \to (F) \to (C) \to A_5$	26. $A_3 \to (F) \to (C) \to A_2$
		27. $A_4 \to (A) \to (A) \to A_5$	27. $A_3 \to (F) \to (D) \to A_2$
		28. $A_4 \to (D) \to (C) \to A_5$	28. $A_3 \to (F) \to (F) \to A_2$
			29. $A_3 \to (F) \to (C) \to A_4$
			30. $A_3 \to (F) \to (D) \to A_4$
			31. $A_3 \to (F) \to (C) \to A_5$
			32. $A_4 \to (A) \to (A) \to A_5$
			33. $A_4 \to (B) \to (A) \to A_5$
			34. $A_4 \to (D) \to (C) \to A_5$
			35. $A_4 \to (E) \to (C) \to A_5$
			36. $A_4 \to (F) \to (C) \to A_5$

State reachability
How far agent's data may go?



Agent A_3 initially knows B, E and F.

Applying $(EU)^*$:

- A can reach A_4 and A_5
- B can reach A_1 and A_2
- C can reach A_5
- D can reach A_1 and A_4
- E is not shared
- F can reach A_2

Conclusions

- Application of linear algebra model for epistemic logic to a P2P integration platform within the security domain
- Extension of Tojo’s model with upcasting operations
  - Upcasting channels required to model partial information hidding
  - Modeling upcasting resulted in 4D (instead of 3D) communication matrices
- Expected problem size: few dozen agents and about 100 classes
- At present supported by a small prototype software

Thank you