Speech Acts and Tokens for Access Control and Provenance Tracking

Fabian Neuhaus (NCOR) & Bill Andersen (Highfleet)
STIDS 2011
Sorry

Fabian can’t be with us today ...
Problem Statement

- Stock semantic technology works well where:
  - There is trust among producers and consumers of data
  - There is little or no noise in the data
- Often these conditions don’t hold
  - Intelligence, LE, engineering, health care, E-science
- Such applications are characterized by
  - Multiple data sources
  - Need to protect sources and methods
  - Need to protect privacy
  - Need to control release of sensitive information
  - Need to support consumer confidence in integrated data
Motivating example

(T) Asset says AQ has nuke at location L
(S) AQ has nuke at location L
(T) If AQ has nuke then it’s Pakistani
(T) Pakistan controls all their nukes

(U) NYT reports Azhar claims AQ has nuke

Provide all independent records (from A, B, & C) that support that Al Qaeda has WMD
What is the right answer?

- It is not the “information” (propositions) that are protected, but the records (tokens) in each system
  - The right answer depends on access controls
- Source systems understand where their records came from to avoid duplication and false corroboration
  - The right answer depends on provenance
- We addressed these issues at STIDS 2009
- Here is what’s new:
  - Hearsay (source of data external to provenance control)
  - Logical inconsistency
  - More about hearsay later
(Onto)logical approach

- **Ontology**
  - Formal languages
  - Propositions
  - Sentence types
  - Sentence tokens
  - Speech acts

- **Formal System**
  - Proof calculus
  - Axiomatization of `supportedBy` relation
Ontology

- Basic unit is the sentence token (of a formal language)
  - Need not be overtly logical
  - DBs, for example, will do
- We consider only formal languages here
  - ...not restricted to overtly logical languages
- Sentence types encode propositions
- Sentence tokens are material objects
  - ...or constituted of material objects (ask me offline)
- Sentence tokens instantiate sentence types
- Speech acts
Kinds of provenance

- **IT processing tracking**
  - Records history of operations on tokens inherent to operation of the information system
  - Examples
    - Copying
    - Algorithmic transformations
    - Automated theorem proving

- **Hearsay tracking**
  - Records history of tokens as originating from agents’ communicative speech acts
  - Intention is essential to this view
About speech acts

- With typical deductive systems, we simply “insert” data without any further consideration of the act of assertion
  - Except perhaps recording transaction time, etc
- Speech acts are **intentional acts** by which linguistic tokens are brought into existence for achieving some type of communication
- We consider two types of speech acts
  - Assertive (updates)
  - Interrogative (queries)
- Speech acts provide the ontological foundation for individuation of assertions and queries
The formal system – preliminaries

- Deductive system that supports
  - Discretionary and mandatory access control
  - IT and “hearsay” provenance tracking
  - Privacy (not discussed in this work)

- Note
  - Direct implementation of this formal system not required to enjoy the benefits
  - Best to think of it as a specification of correct behavior that can be implemented in multiple ways
The formal system - details

- FOL system (Common Logic) as basis
- IKL-like extensions for **proposition names**
  - (that (likes fabian cookies))
  - Φ iff ((that Φ)) … for all formulas Φ
- **Addition of two modal operators**
  - ◊ Φ – Φ is logically possibly true
  - □ Φ – Φ is logically necessarily true
- **A form of paraconsistent logic**
  - Contradictions don’t introduce chaos
Key non-logical vocabulary

- **Record**
  - A unary relation that ranges over tokens
- **ResidesIn**
  - A one-place function from tokens to systems
- **PropositionalContent (PC)**
  - A one-place function from tokens to propositions
- **SupportedBy**
  - A binary relation on propositions and token sequences
- **BasedOn**
  - A binary relation between tokens
- **AssertionAct**
  - A unary relation that ranges over assertive speech acts
**Motivating example – refresher**

(T) Asset says AQ has nuke at location L

(S) AQ has nuke at location L

(T) If AQ has nuke then it’s Pakistani

(T) Pakistan controls all their nukes

(U) NYT reports Azhar claims AQ has nuke

Provide all independent records (from A, B, & C) that support that Al Qaeda has WMD
Record(token001) &
ResidesIn(token001) = repository_A &
ClassifiedAs(token001, TS) &
Compartment(token001, alQaeda_cmpt) &
Compartment(token001, proliferation_cmpt) &
CreatedBy(token001) = agent1234 &
PropositionalContent(token001) =
(that ( ∃x (AssertionAct(x) &
    Speaker(x, source007) &
    Date(x) = 20.10.2011) &
    PropositionalContent(x) =
    (that ( ∃y (Owns(alQaeda,y) & NuclWeap(y)))))))

Assertive speech act records the claim of 007 that Al Qaeda has a nuclear weapon
Repository B

Record(token002) &
ResidesIn(token002) = repository_B &
ClassifiedAs(token002, S) &
BasedOn(token002, token001) &
ResidesIn(token001) = repository_A &
PropositionalContent(token002) =
  (that ( ∃ x (AssertionAct(x) &
    PropositionalContent(x) =
      (that( ∃ y (Owns (alQaeda y) & NuclWeap(y))))))
Record(token003) &
ResidesIn(token003) = repository_C &
ClassifiedAs(token003, U) &
PropositionalContent(token003) =
  (that ( ∃ x (AssertionAct(x) &
           Speaker(x, nyt) &
           Date(x) = 23.10.2011 &
           PropositionalContent(x) =
             (that ( ∃ y (AssertionAct(y) &
                      Speaker(y, Azhar) &
                      Date(y) = 22.10.2011 &
                      PropositionalContent(y) =
                        (that ( ∃ z (Owns(alQaeda, z) &
                             NuclWeap(z)))))))

Three levels of “hearsay”
Axiomatization of SupportedBy

- **Reflexivity**
  - \((\text{Record}(x) \& \boxdot \text{PC}(x)) \rightarrow \text{PC}(x)[x]\)

- **Axiomhood**
  - \(A \rightarrow A[]\)

- **And-introduction**
  - \((A[s_1] \& B[s_2] \& \boxdot (A\&B)) \rightarrow (A\&B)[s_1s_2]\)

- **Modus Ponens**
  - \((A[s] \& \Box(A\rightarrow B)) \rightarrow B[s]\)

- **Hearsay**
  - \((\boxdot A \& (\exists x(AA(x) \& \text{PC}(x)=(\text{that } A)))[s]) \rightarrow A[s]\)

**A[s] is shorthand for SupportedBy(A,s) where s is a sequence of tokens**
A query and one answer

Query:
\[ \exists x(\text{Owns(alQaeda} x) \& \text{WMD}(x))[?s] \]

Proof:
1) \[ \forall x(\text{NuclWeap}(x) \rightarrow \text{WMD}(x))[ ] \]
2) \[ \exists x(\text{AssertionAct}(x) \& \text{Speaker}(x \text{ source007}) \& \text{Date}(x) = 20.10.2011) \& \]
   \[ \text{PropositionalContent}(x) = (\text{that}(\exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y))))[\text{token001}] \]
3) \[ \square((\exists x(A \& B \& C \& D)) \rightarrow \exists x(A \& D)) \]
4) \[ \exists x(\text{AssertionAct}(x) \& \]
   \[ \text{PropositionalContent}(x) = (\text{that}(\exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y))))[\text{token001}] \]
5) \[ \Diamond(\exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y))) \]
6) \[ \exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y))[\text{token001}] \]
7) \[ \Diamond((\forall x(\text{NuclWeap}(x) \rightarrow \text{WMD}(x)) \& \exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y)))) \]
8) \[ (\forall x(\text{NuclWeap}(x) \rightarrow \text{WMD}(x)) \& \exists y(\text{Owns(alQaeda} y) \& \text{NuclWeap}(y)))[\text{token001}] \]
9) \[ \square((\forall x(\text{NuclWeap}(x) \rightarrow \text{WMD}(x)) \& \]
   \[ \exists y(\text{Owns(alQaeda} y) \& \text{WMD}(y))) \rightarrow \exists x(\text{Owns(alQaeda} x) \& \text{NuclWeap}(x))) \]
10) \[ \exists x(\text{Owns(alQaeda} x) \& \text{NuclWeap}(x))[\text{token001}] \]

The query is true and token001 says so!
Our approach focuses on tokens in a formal language
- Tokens provide ontological and logical foundation for access control and provenance tracking in deductive information systems
- Speech acts provide the justification for individuation of assertions and query responses as linguistic tokens
  - Also enabling the tracking of “hearsay” provenance
- A form of Labeled Deductive System (Gabbay)
  - We are exploring application of these results to our work
- Approach is extensible to access control and provenance of non-linguistic artifacts (e.g. images, video, paper)
Implementation

- In case you’re wondering, this actually works!
- Highfleets has implemented a version of this approach
  - Currently limited to atomic sentences
  - Integration with XKS (CL-based) and triple store (RDF-based) products – stay tuned
- Used for provenance tracking in US DoD system
  - Does not (but could) incorporate hearsay tracking

Grab me after the talk and I’ll show you a demo!
Questions?

andersen@highfleetc.com