Social network and web data management

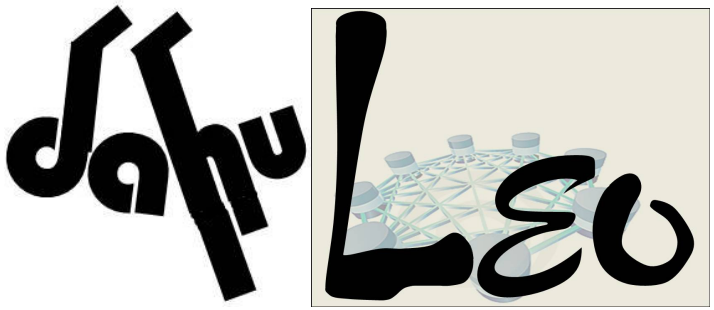
Alban Galland[1]

[1] INRIA Saclay (Leo Team [ex Gemo]) & ENS Cachan (LSV, Dahu axis)

June, 2010, *Barbizon Seminar*

Administrative context

- Thesis context:
  - My advisor: Serge Abiteboul
  - Date: July 2008-July 2011
  - Contract: "Ingenieur du Corps des Telecom/Mines", seconding at INRIA



- Teaching : vacation at Telecom ParisTech and Ecole Polytechnique

My thesis subject

- Social Network and Web data management
  - What kind of computation needed for Social Network?
  - How to find, sort and integrate knowledge?
  - What kind of data model needed for Social Networks?
  - How to manage data access and data distribution?

Outline

Outline

A motivating example

Alice want to know what are the grades of the climbing routes of roc14?

	Extrême gauche	Schtroumpfem--ent chaud pour Schtroumpf manchot	Alauda Arvensis	L'arete du saumon	Ultraviolet
Alban	5b	4b	4c	6b	6c
Luc	5b	4b	4c	6a	6b
Pierre	5b	4b	4b	6a	6a
Serge	3b	7c	6a	6b	6a

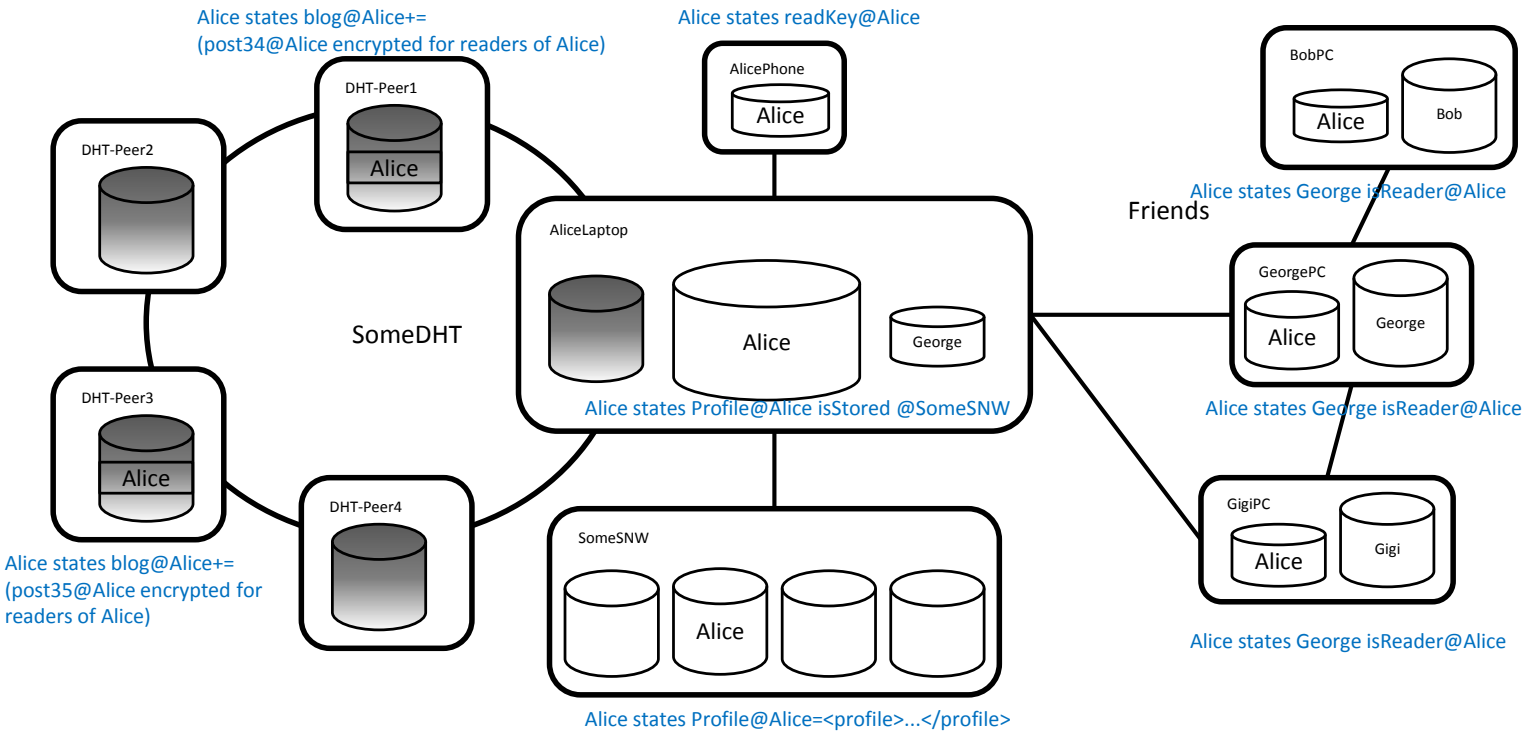
Corroboration

- **Problem:** sources assert facts, linked with functionnal dependencies and we want to guess which facts are true and which facts are false.
- **Idea:** try to compute confidence in truth of facts from confidences in sources, and confidence in sources from confidence in fact, following a probabilistic model of the data production.
- **Paper:** Alban Galland, Serge Abiteboul, Amelie Marian, Pierre Senellart, *Corroborating Information from Disagreeing Views*, WSDM Conference 2010

Outline

A motivating example

- The distributed knowledge base of Alice, the rockclimber:



Problem

- Problem
  - Describe all kinds of distribution schemes (centralized, structured and unstructured P2P)
  - Provide access control for reading and editing the data, and delegating this rights. Execute any valid and only valid instruction (read or edit) on the data
  - Enable reasoning on the knowledge (both on data and meta-data)
- Idea
  - A model of distributed data with access-control and provenance
  - Some constraint to guarantee properties of systems build on the model

Global view of the model

- Data and meta-data are all first class-citizen. They are represented as logical statement which are "valid" knowledge, enforcing read and edit rights
  - Two kinds of data statement: Document (read/write), Collection (read/append/remove)
  - Three kinds of meta-data statement: Access right, Key, Localization
- Instructions are used to request manipulation of data (get or update)

Access control and provenance

- The statements carry a large part of the access control mechanism: signatures for edit access rights, encryption for read access rights
  - Alice states news@roc14=T requester Bob at 2010/04/01 10:00:00GMT
- The exchange of knowledge keep the full trace of the previous exchange, by piling up signatures of the user which send the data
  - Bob says Alice says Alice states new@roc14=T [...] to Bob to George

Distribution schemes

- @Home: one trusted peer hosts all the data of the principal
- @Host: one untrusted peer hosts all the data of the principal, encrypted
  - @DHT: a set of untrusted peer hosts redundantly the data of the principal
- @Friends: each principal hosts his own knowledge and some data of other principals, he is interested in.

System properties

- We are interested by the following properties of system
  - Well-formedness: the data is syntactically correct
  - Soundness: only valid instructions (read or edit) are executed in the system.
  - Completeness: any valid instruction (read or edit) is correctly executed.
- Nothing prevents a participant to do something illegal such as giving a document to some unauthorized party. But then, the unauthorized party cannot prove that he obtained the information legally.

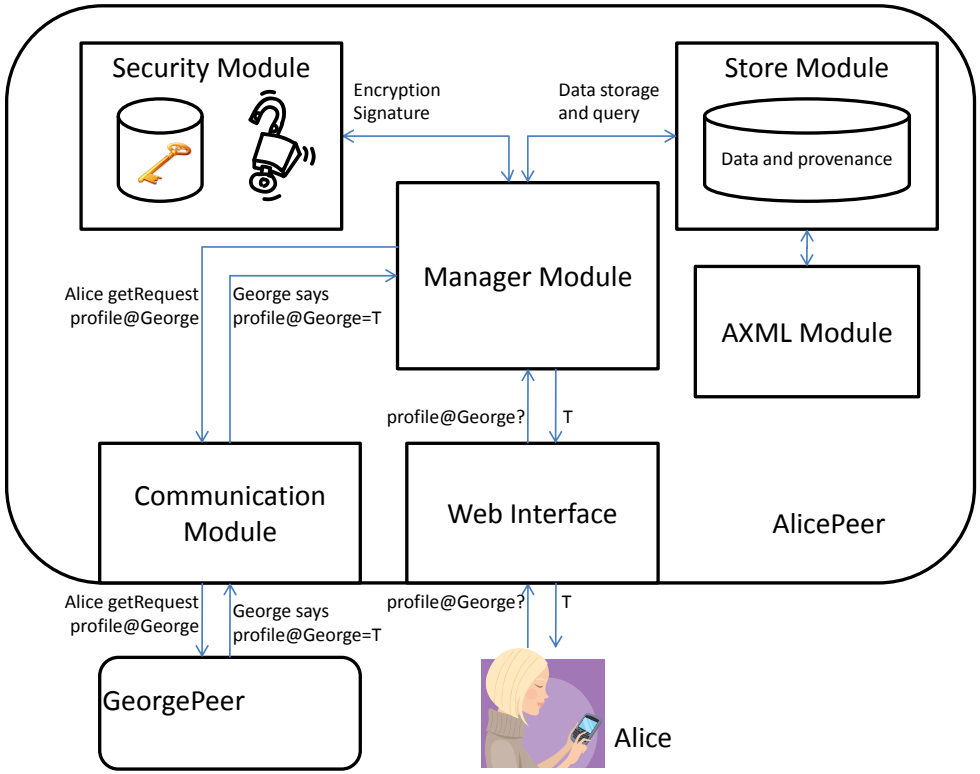
Query-based access control

- High level specification: *use queries (e.g. datalog) to define access control*
- Semantic problem: Does the evaluation of the access control by a central omniscient authority give the same result as the distributed one (Evaluation of the queries on the local data by each peers)?
  - In general case, undecidable (comparison of datalog programs)
  - Decidable case we know about are not very expressive

Outline

The Pastis system

- The architecture of the system:



Outline

Directions for future work

- Data integration on social tagging systems
- Building some wrapper (Facebook, OpenSocial...) to demonstrate private data-integration on the web
- Query processing based on distributed datalog evaluation
- Study of scenarios of distribution and verification of properties

Outline

Conclusion

- Some basis for social network and data management on the web
  - Data mining: corroboration
  - Data model: Pastis Model
  - System: Pastis System
- We believe Social Network is a cool way to motivate work on data integration, distribution, access control and verification