UNIVERSITÉ DE RENNES 1

DRUID

# DRUID: Declarative & Reliable management of Uncertain, user-generated & Interlinked Data

DRUID Team - DKM department

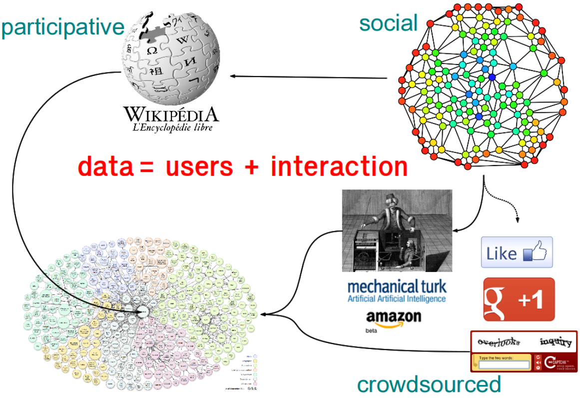
IRISA, Lannion-Rennes

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALÉATOIRES

UMR IRISA

# Context

- Huge amount of data available
- e.g. Linked / Open Data
- But who are the sources of these data?



# Context

Humans behind the data

## Challenges

- How to profile users, analyze their relationships?
- How to interact with them efficiently to solve data acquisition tasks, in a reliable way?

## Team

10 teacher-researchers: Rennes-Lannion

# Context: Useful

- Social network analytics
  - Tools for User Profiling, User Targetting, User influence, User preferences
  - Supporting Social Sciences
- Crowdsourcing for complex tasks
  - 300,000 users available anytime on AMT
  - Participative sciences (FoldIt success, GalaxyZoo,...)

# Crowdsourcing: Natural Intelligence

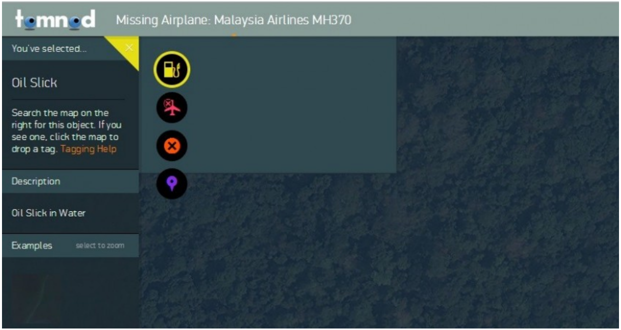
- Goal of crowdsourcing: "obtain needed services, ideas, or content by soliciting contributions from a large group of people"
- Human fallback: obtain an answer when machine learning is not mature enough
- Many crowdsourcing platforms solicit on-line crowd.
- Micro-tasks
  - audio transcription, text translation, image tagging, citizen science, audio or image quality perception
  - implicit collaboration
  - consensus usually achieved with majority voting: Information fusion more adapted

# Some crowdsourcing problems

- How to extract ground truth? IA: obtain data for training
- Answers could be imprecise and uncertain: How to ask the questions? IA: Knowledge representation
- How to fuse the information? IA: Information fusion
- How to obtain knowledge on workers? IA: Knowledge representation, learning (supervised, unsupervised)
  Such as the reliability of a worker:
  - to be honest
  - to be expert in a domain
- How to assign/recommend tasks to workers according to their profile? IA: learning, prevision
- How to ask questions according to previous answers of the workers? IA: Reinforcement learning, active learning

# Crowdsourcing: MH370 example

- Where Malaysia airlines flight MH370 disappeared without a trace in March 2014?
- DigitalGlobe and tomnod.com offer their satellite photos of ocean in crowdsourcing effort
- 3 million have joined the platform



# What is information?



: Many debris are on the image.

Imprecise proposition

: Ten debris are on the image.

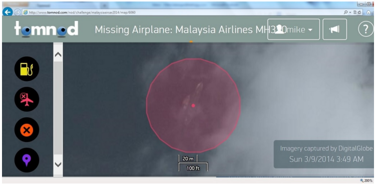
Precise proposition

Imprecision is a kind of imperfection of information

# What is information?

 : the airplane is at the position S 8°22' E 71°46'

## Bad weather




Uncertain proposition

## Good weather

Certain proposition

Uncertainty is another kind of imperfection of information

**Goal**

To combine information coming from many imperfect sources in order to improve the decision making taking into account of imprecisions and uncertainties

To model imperfections: Artificial Intelligence Reasoning by uncertainty theories:
Probability theory (Bayesian approach) or possibility theory or the theory of belief functions

# Fusion architecture for classifiers fusion



$s$ sources $S_1$, $S_2$, ..., $S_s$ that must take a decision on an observation $x$ in a set of $n$ classes $x \in \Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$

$$
\begin{array}{c c}
 & \begin{array}{ccccc} \omega_1 & \ldots & \omega_i & \ldots & \omega_n \end{array} \\
\begin{array}{c} S_1 \\ \vdots \\ S_j \\ \vdots \\ S_s \end{array} &
\left[ \begin{array}{ccccc}
M_1^1(x) & \ldots & M_i^1(x) & \ldots & M_n^1(x) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
M_1^j(x) & \ldots & M_i^j(x) & \ldots & M_n^j(x) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
M_1^s(x) & \ldots & M_i^s(x) & \ldots & M_n^s(x)
\end{array} \right]
\end{array}
$$

4 steps

1. Modeling
2. Estimation
3. Combination
4. Decision

# Probability theory for information fusion

**Modeling:** A probability is a positive and additive measure, $p$ is defined on a $\sigma$-algebra of $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ and takes values in [0,1].

It verifies: $p(\emptyset) = 0$, $p(\Omega) = 1$, $\sum_{X \in \Omega} p(X) = 1$

**Estimation:** Choice of the distribution, and/or estimation of parameters

**Combination:** Bayes rule

$$p(x \in \omega_i / S_1, \ldots, S_s) = \frac{p(S_1, \ldots, S_s / x \in \omega_i) p(x \in \omega_i)}{p(S_1, \ldots, S_s)} \qquad (1)$$

Independence assumption must of the time necessary

**Decision:** *a posteriori* maximum, likelihood maximum, mean maximum, *etc.*

# Limits of the theory of probabilities

- Difficulties to model the absence of knowledge
  ex: Sirius: ignorance on life $p(\text{life}) = p(\overline{\text{life}}) = \frac{1}{2}$, but also $p(\text{animal}) = p(\text{vegetate}) = p(\overline{\text{life}}) = \frac{1}{3}$ so $p(\text{life}) = \frac{2}{3}$
- Constraint on the classes (exhaustive and exclusive)
- Constraint on the measures (additivity)
  Knowing information such as $p(f|A) = 1$ transfers information on $p(\overline{A}|f)$

# Theory of belief functions for fusion

**Modeling:** The basic belief functions (bba or mass functions) are defined on $2^\Omega$ and take values in $[0,1]$ with

- Discernment frame: $\Omega = \{\omega_1, \dots, \omega_n\}$, with $\omega_i$ are exclusive and exhaustive classes
- Power set: $2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1 \cup \omega_2\}, \dots, \Omega\}$.

It verifies: $\sum_{X \in 2^\Omega} m(X) = 1$

**Estimation:** Learning

**Combination:** Conjunctive rule

$$m_{\text{Conj}}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2)$$

Assume: cognitively independence of sources

**Decision:** maximum of belief, plausibility, pignistic probability - possible decision on $2^\Omega$

# Mass functions for imperfection model



**Special cases:**

- If only positive masses are $\omega_i$ then $m_j$ is a probability
- $m_j(\Omega) = 1$: total **ignorance** of $S_j$
- categorical mass function: $m_j(X) = 1$ (noted $m_X$): $S_j$ has an imprecise knowledge
- $m_j(\omega_i) = 1$: $S_j$ has a precise knowledge
- simple mass functions $X^w$:
  $m_j(X) = w$ and $m_j(\Omega) = 1 - w$: $S_j$ has an **uncertain and imprecise** knowledge


DRUID, Arnaud Martin - Arnaud.Martin@irisa.fr

# Outline

1. Introduction to information fusion
2. Theory of belief functions for information fusion
3. Applications
   - crowdsourcing
   - social network

# Clustering on exactitude and precision degrees

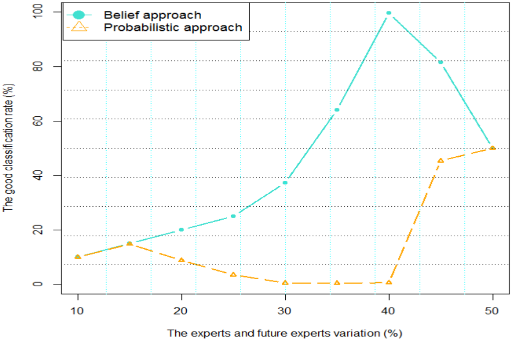
1. Step 1: Calculate an exactitude degree based on the distance between $m_{U_j}^{\Omega_k}$ and the average of the responses proposed by the $s-1$ participants $(m_{U_{\varepsilon_{s-1}}}^{\Omega_k})$
2. Step 2: Calculate a precision degree from the specificity degree based on the assumption "the majority has right"
3. Step 3: Calculate a global degree and applied a clustering on it.

Comparison with a probabilistic approach: Just an exactitude degree, no precision degree with probability.
(A. Ben Rjad, et al., 2016)

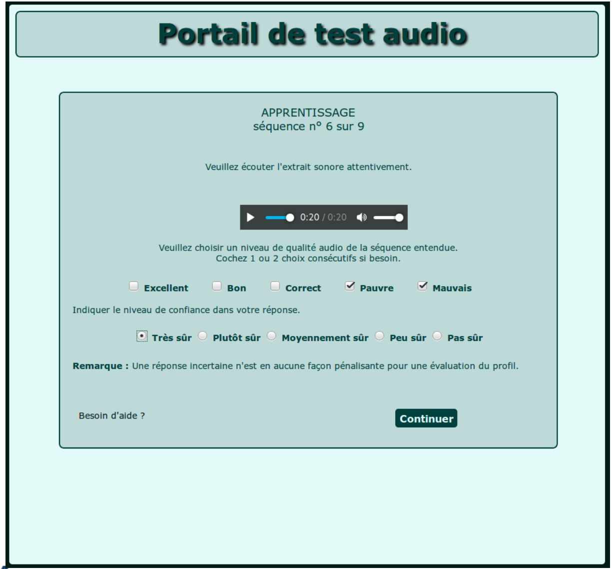
# Results experiment

Goal: prove the interest of the use of the theory of belief functions instead of probability on generated data

- expert and imprecise expert with the same percentage from 10% to 50%



# Questions for imprecise and uncertain answer

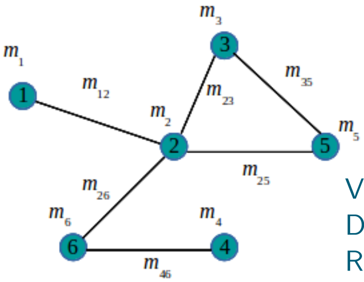


# Social network: an uncertain valued graph

Node and link-attributed graphs

$G = (V, E, m_u, m_e)$ where $m_u : V \longrightarrow \mathcal{X}$ and $m_e : e \in E \longrightarrow \mathcal{X}$

$m_u(v) = [m_1(v), \ldots, m_a(v)]$




Veracity of information

Doubt

Reliability

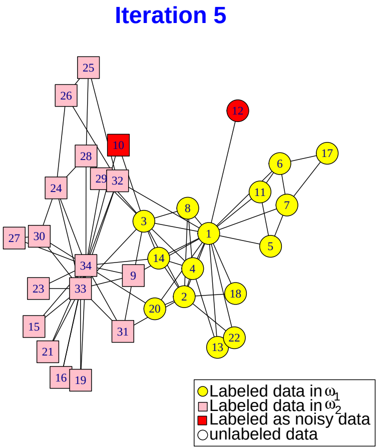
(Ben Dhaou, 2014, 2017)

# Community detection: SELP

Semi-supervised Evidential Label Propagation algorithm

(Zhou et al., 2018)

Example on Karate Club network



# Druid: http://www-druid.irisa.fr/

Druid team has many connections with AI methods/problems

- Social networks
  - Preferences model and fusion: see Yiru Zhang poster
  - Word embeddings: ANR EPIQUE: see Ian Jeantet poster
- Crowdsourcing platforms (ANR HEADWORK)
- Sensor fusion (CIFRE TOTAL)
- Privacy and related problems (ANR CROWDGUARD):
  - Privacy of the individuals involved in personal-data-centic applications (*e.g.* crowdsourcing, social networks, open data)
  - Transparency of black box personalization algorithms (*e.g.* predictions of risk recidivism, web recommendations)

Implication of the team in AFIA: http://afia.asso.fr/