An introduction to social network challenges

Arnaud Martin

Arnaud.Martin@univ-rennes1.fr

Université de Rennes 1 - IRISA - DRUID, Lannion, France

Paris, January, 22th 2018

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALÉATOIRES

Outline

1. What is a social network?
2. How to model a social network?
3. How to model information on social networks?
4. How to analyse social network?

What is a social network?



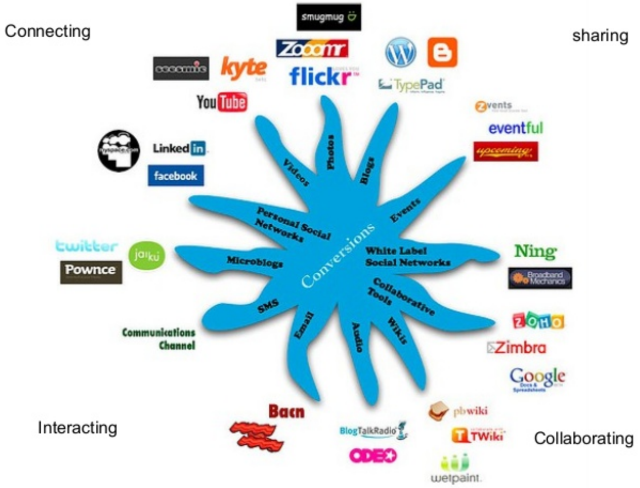


What is a social network?

What is a social network?

Collaborative platforms

What is a social network?



What is a social network?

A definition

A finite set of social actors (individual, organisations) with relations (collaboration, advice, control, influence, etc.) between them.

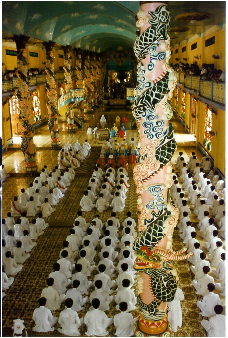
Remarks:

- Technical definition
- Is it really always finite?
- Relations and actors are never fixed
- Most of time not only one social network, not only one kind of group (community)

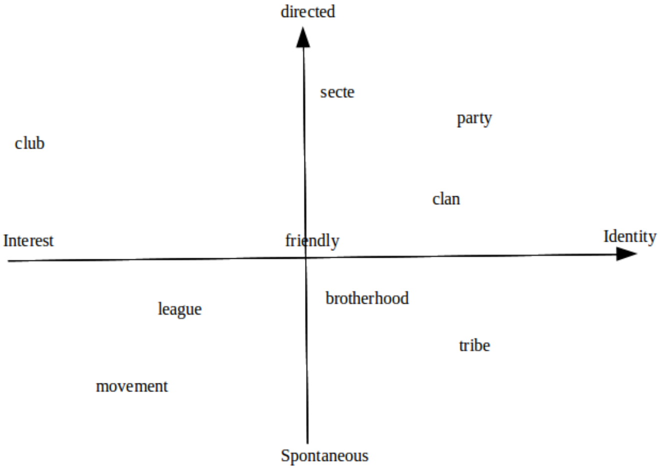
Application domains

- Sociology
- Ethnology
- Economy
- Demography
- Criminal networks
- Social media
- Literary
- Ecology
- etc.

Notion of communitiy



Notion of communitiy



Social network: a matter of scale

- 31% of world population connected on social network
- Facebook: 1,8 billions of users/month - 17.9 billions of $
- Qzone: 653 millions of users/month
- Instagram: 600 millions of users/month
- Twitter: 317 millions of users/month
- LinkedIn: 106 millions of users/month
- Snapchat: 150 millions of users/day

Social network: main challenges

- economical challenges:
  games, publicities, business image, marketing (viral marketing), etc.
- political challenges:
  social influence, e.g. Jasmin revolution, Obama elections, Trump tweets, etc.
- social challenges:
  share knowledge: all information at any time, communication (to find a job, a partner, etc.), etc.

Social network: main scientific challenges

- big data management:
  How to access to the data? How to make requests on the data? How to reduce complexity of processes?, etc.
- social mining:
  How to extract information from the data? How to characterize the data?, etc.
- privacy and security:
  How to protect people data? How to assure the security of people?, etc.

Outline

1. What is a social network?
2. How to model a social network?
3. How to model information on social networks?
4. How to analyse social network?

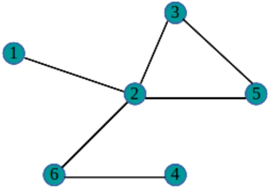
A graph: G
A set (V, E) with $V = \{v_1, \ldots, v_V\}$ a set of vertices/nodes and $E = \{e_1, \ldots, e_E\}$ a set of edges/links
$e_k \in E$ is couple of (v_i, v_j).

- $|V| = V$: order of the graph
- $|E| = E$ number of edges
- v_i and v_j are neighbour or adjacent if $\exists e_k \in E$ such as $e_k = (v_i, v_j)$
- $N(u) = \{v \in V, (u, v) \in E\}$: the neighbourhood of u
- Node degree: $d(u) = |N(u)|$ *i.e.* the number of edges from u.
- Centrality of a node: $\frac{d(u)}{E-1}$
- Link density: $D = \frac{2E}{V(V-1)}$

See Ernesto Estrada talk for more features on the graphs...

How to model a social network?

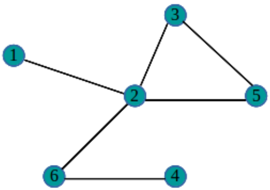
a graph:



an adjacent matrix:

$$
\begin{array}{c}
\\
1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6
\end{array}
\begin{array}{c}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}
\end{array}
$$

a graph:



a list of adjacent nodes:

1: 2

2: 1, 3, 5, 6

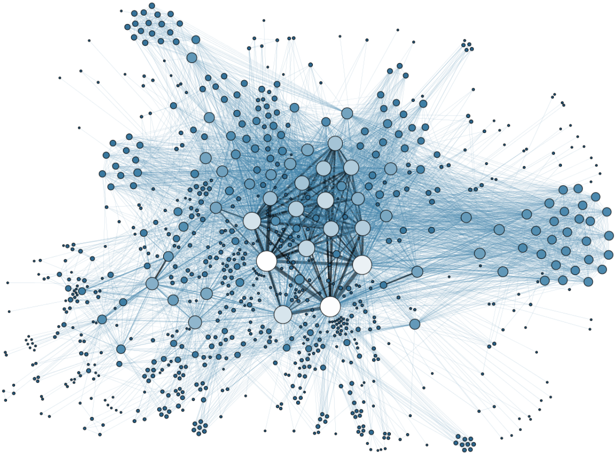
3: 2, 5

4: 6

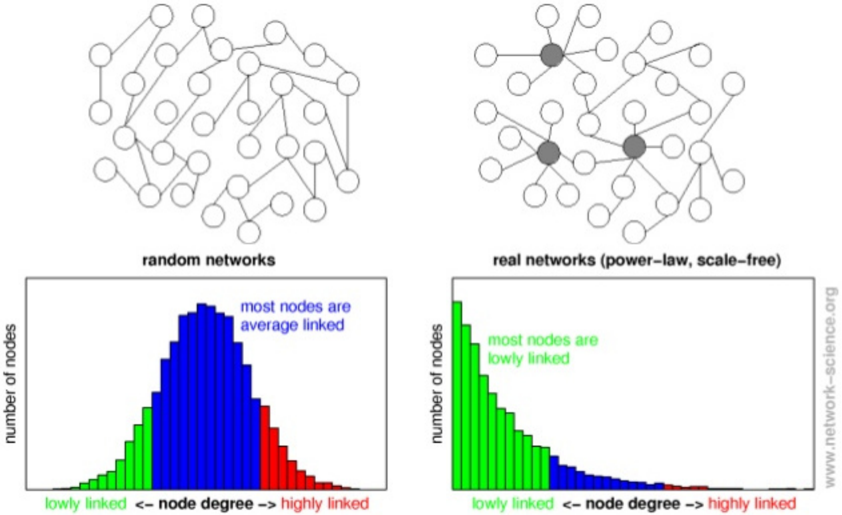
5: 2,3

6: 2, 4

Challenge: drawing large graphs



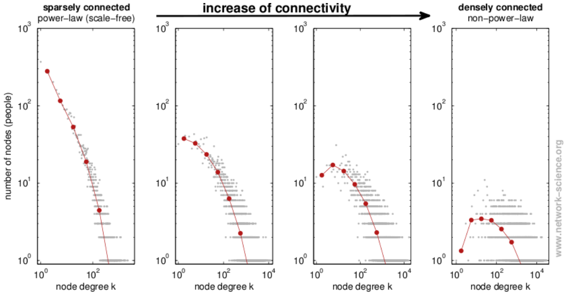
Specificity of social network



Main social networks are scale-free network and have a degree distribution given by a power distribution:

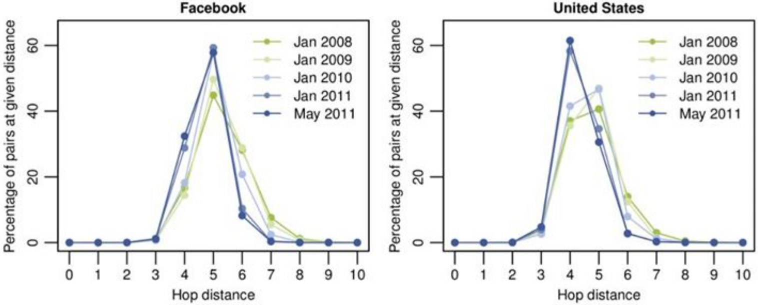
$$P(k) = Ck^{-\gamma}$$

$P(k)$ is the proportion of nodes with the degree k, in general $2 \leq \gamma \leq 3$ C a constant. The density of a graph depend on the application domain (Melançon, 2006)



Flikr social online network (Scholz 2015)

(Milgram 1967): In average, the number of links between two persons (nodes) is small (around 6).
(Facebook, 2011): Each person is linked to other by 4.74 relations (in average).



In social networks:

- The number of neighbours for a given node is approximately the same than the number of neighbours of its neighbours
- The distance L between two randomly chosen nodes is given by:

$$L \simeq \ln E$$

- geodesic distance: between two vertices is the shortest path (number of edges)
- eccentricity: $\epsilon(u)$ is the greatest geodesic distance between u and another vertex
- radius: $\min_{u \in V} = \epsilon(u)$
- graph diameter: $\max_{u \in V} = \epsilon(u)$

Problem: detection of cycles - NP-hard algorithms

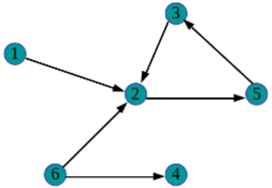
- intermediary centrality of a node:

$$IC(u) = \sum_{s \neq u, t \neq u, s \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}}$$

σ_{st}: number of shortest paths between s and t,
$\sigma_{st}(u)$: number of shortest paths between s and t passing by u

How to model a social network?

a directed graph: (*e.g.* followers in Tweeter)



an adjacent matrix:

$$\begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

(Fortunato, 2010) some properties for a community:

- Two neighbours in a same community are approximately the same
- Two neighbours in a same community must be near
- The nodes of a community have a high average degree
- A community contains a high proportion of triplets (high clustering coefficient)
- A community has a large embeddedness (ratio on internal and external degree)

(Fortunato, 2010) some properties for a community:

- Two neighbours in a same community are approximately the same
- Two neighbours in a same community must be near
- The nodes of a community have a high average degree
- A community contains a high proportion of triplets (high clustering coefficient)

$$C(u) = \frac{2|\{e_{ij} = (v_i, v_j) \in E : v_i, v_j \in N(u)\}|}{|N(u)(N(u) - 1)|}$$

$$C(G) = \frac{1}{V} \sum_{u \in V} C(u)$$

- A community has a large embeddedness (ratio on internal and external degree)

(Fortunato, 2010) some properties for a community:

- Two neighbours in a same community are approximately the same
- Two neighbours in a same community must be near
- The nodes of a community have a high average degree
- A community contains a high proportion of triplets (high clustering coefficient)
- A community has a large embeddedness (ratio on internal and external degree)

  For a given sub-graph G_c of G, A it adjacent matrix, $u \in G_c$:

$$k_u^{int} = \sum_{j \in G_c} A_{uj}$$

$$k_u^{ext} = \sum_{j \notin G_c} A_{uj}$$

(Fortunato, 2010) some properties for a community:

- Two neighbours in a same community are approximately the same
- Two neighbours in a same community must be near
- The nodes of a community have a high average degree
- A community contains a high proportion of triplets (high clustering coefficient)
- A community has a large embeddedness (ratio on internal and external degree)

  For a given sub-graph G_c of G, A it adjacent matrix, $u \in G_c$:

$$\xi_u = \frac{k_u^{int}}{k_u^{int} + k_u^{ext}}$$

(Fortunato, 2010) some properties for a community:

- Two neighbours in a same community are approximately the same
- Two neighbours in a same community must be near
- The nodes of a community have a high average degree
- A community contains a high proportion of triplets (high clustering coefficient)
- A community has a large embeddedness (ratio on internal and external degree)

Challenge: Give a definition of a community on a social network

See Mauro Sozio and Florence Sèdes talks...

Challenges: classical problems

- Travelling salesman problem: Find the shortest way to visit given nodes only one time (NP-hard) - equivalent to vehicle routing problem
- Graph labelling and colouring: give a label to all nodes (or links) (NP-hard)
- Maximum flow: in flow network (valued directed graph) find the largest possible total flow
- Large graph compression
- Maximal clique enumeration (NP-hard)
- Independent set problem: find the largest possible independent set (set of vertices with no two of which are adjacent) (NP-hard)

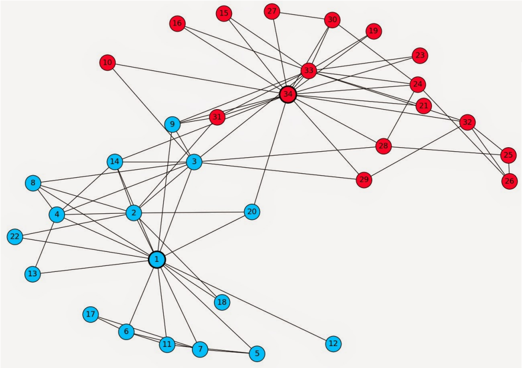
Most of problem on graph are equivalent to NP-hard optimisation problems. Some approximation algorithms are developed.

For different communities social network:
Lancichinetti-Fortunato-Radicchi LFR benchmark: based on power law distribution, need:

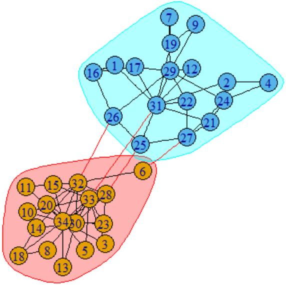
- number of nodes
- minimum and maximum for the community sizes
- average, maximum degree
- etc

defines:

- number of edges
- number of communities



Zachary's Karate club network



LFR generation

Challenge: realistic social network generation

(Largeron, et al, 2015), see Christine Largeron talk...

- Local preferential attachment: new link between vertices with high degree
- Small world
- Community structure: vertices are connected to vertices in a same group compared to other group (large embeddedness)
- Community homogeneity: similarity of vertices in a same group
- Homophily: vertices in a same group are more similar than with the other groups

allows:

- dynamical generation of social networks
- fix the number of vertices
- fix the number of communities

Outline

1. What is a social network?
2. How to model a social network?
3. How to model information on social networks?
4. How to analyse social network?

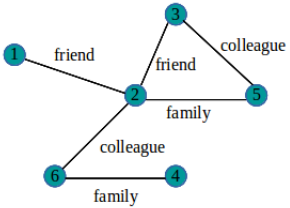
How to model information on social network?

On social network some information can be considered:

- information on the links: LinkedIn, etc.
- information on the nodes: Facebook, LinkedIn, etc.
- information (message) throw the network: Tweeter, collaborative platforms, etc.

Valued graphs

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

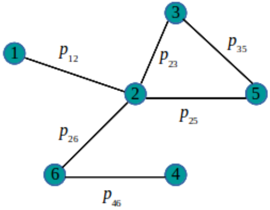


an adjacent matrix:

$$\begin{array}{c c} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} 0 & w_{12} & 0 & 0 & 0 & 0 \\ w_{12} & 0 & w_{23} & 0 & w_{25} & w_{26} \\ 0 & w_{23} & 0 & 0 & w_{35} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{46} \\ 0 & w_{25} & w_{35} & 0 & 0 & 0 \\ 0 & w_{26} & 0 & w_{46} & 0 & 0 \end{bmatrix} \end{array}$$

Valued graphs

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



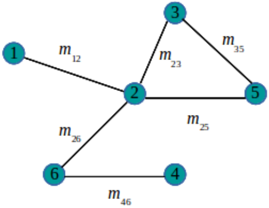
$p_{12}(\text{friend}) = 0.8$
$p_{12}(\text{family}) = 0.15$
$p_{12}(\text{colleague}) = 0.05$

an adjacent matrix:

$$
\begin{array}{c c}
 & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} &
\begin{bmatrix}
0 & p_{12} & 0 & 0 & 0 & 0 \\
p_{12} & 0 & p_{23} & 0 & p_{25} & p_{26} \\
0 & p_{23} & 0 & 0 & p_{35} & 0 \\
0 & 0 & 0 & 0 & 0 & p_{46} \\
0 & p_{25} & p_{35} & 0 & 0 & 0 \\
0 & p_{26} & 0 & p_{46} & 0 & 0
\end{bmatrix}
\end{array}
$$

Valued graphs

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



Veracity of information
Doubt
Reliability

an adjacent matrix:

$$
\begin{array}{c c}
 & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} &
\begin{bmatrix}
0 & m_{12} & 0 & 0 & 0 & 0 \\
m_{12} & 0 & m_{23} & 0 & m_{25} & m_{26} \\
0 & m_{23} & 0 & 0 & m_{35} & 0 \\
0 & 0 & 0 & 0 & 0 & m_{46} \\
0 & m_{25} & m_{35} & 0 & 0 & 0 \\
0 & m_{26} & 0 & m_{46} & 0 & 0
\end{bmatrix}
\end{array}
$$

Limits of the theory of probabilities

A probability is a positive and additive measure, p is defined on a σ-algebra of $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ and takes values in [0,1].
It verifies: $p(\emptyset) = 0$, $p(\Omega) = 1$, $\sum_{X \in \Omega} p(X) = 1$

- Difficulties to model the absence of knowledge (ex: Sirius)
- Constraint on the classes (exhaustive and exclusive)
- Constraint on the measures (additivity)

If one symptom f (for fiver) is always true when a patient get a illness A (flu) ($p(f|A) = 1$), and if we observe this symptom f, then the probability of the patient having A increases (because $p(A|f) = p(A)/p(f)$ so $p(A|f) \geq p(A)$).
The additivity constraint require then that the probability of the patient having not A decreases: $p(\overline{A}|f) = 1 - p(\overline{A}|f)$ so $p(\overline{A}|f) \leq p(\overline{A})$ While there is no reason if the symptom f can be also observe in some other diseases.

Bases on Belief functions

- Use of functions defined on sub-sets instead of singletons such as probabilities
- Discernment frame: $\Omega = \{\omega_1, \ldots, \omega_n\}$, with ω_i are exclusive and exhaustive classes
- Power set: $2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1 \cup \omega_2\}, \ldots, \Omega\}$.
- Several functions in one to one correspondence model uncertainty and imprecision: mass functions, belief functions, plausibility functions
- Extension of 2^Ω to D^Ω, hyper power set in order to model the conflicts
  - D^Ω closed set by union and intersection operators
  - D_r^Ω: reduced set with constraints $(\omega_2 \cap \omega_3 \equiv \emptyset)$

Mass functions

- The basic belief functions (bba or mass functions) are defined on 2^Ω and take values in $[0,1]$
- Normalisation condition: $\displaystyle\sum_{X\in 2^\Omega} m(X) = 1$
- A focal element is an element X of 2^Ω such as $m(X) > 0$
- Closed world: $m(\emptyset) = 0$
- We note m_j the mass function of the source S_j

Special cases:

- If only focal elements are ω_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

Discounting

From (Shafer, 1976):

$$m_j^\alpha(X) = \alpha_j m_j(X),\ \forall X \in 2^\Omega$$
$$m_j^\alpha(\Omega) = 1 - \alpha_j(1 - m_j(\Omega))$$

$\alpha_j \in [0,1]$ discounting coefficient can be seen as the reliability of the source S_j

If $\alpha_j = 0$ the source are completely unreliable, all the mass is transferred on Ω, the total ignorance

Fusion architecture

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

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \omega_1 & \ldots & \omega_i & \ldots & \omega_n \end{array} \\
\begin{array}{c} S_1 \\ \vdots \\ S_j \\ \vdots \\ S_s \end{array} &
\left[\begin{array}{ccccc}
m_1^1(x) & \ldots & m_i^1(x) & \ldots & m_n^1(x) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
m_1^j(x) & \ldots & m_i^j(x) & \ldots & m_n^j(x) \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
m_1^s(x) & \ldots & m_i^s(x) & \ldots & m_n^s(x)
\end{array}\right]
\end{array}
$$

Conjunctive rules

- Assume: two cognitively independent and reliable sources S_1 and S_2.
- The conjunctive rule is given for m_1 and m_2 bbas of S_1 and S_2, for all $X \in 2^\Omega$, with $X \neq \emptyset$ by:

$$m_{\text{Conj}}(X) = \sum_{Y_1 \cap Y_2 = X} m_1(Y_1) m_2(Y_2) \tag{1}$$

	$\emptyset$	ω_1	ω_2	ω_3	Ω
m_1	0	0.5	0.1	0	0.4
m_2	0	0.2	0	0.5	0.3
m	0.32	0.33	0.03	0.2	0.12

Dempster's rule

- Dempster's rule:

$$m_{\mathrm{D}}(X) = \frac{1}{1-\kappa} m_{\mathrm{Conj}}(X) \tag{2}$$

  where $\kappa = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$ is generally called conflict or *global conflict*. That is the sum of the *partial conflicts*.
- That is not a conflict measure.
- Conjunctive rules are not idempotent

Decision on Ω

- In general the decision is made on Ω and not on 2^Ω
  - Pessimist: $\max_{\omega\in\Omega} bel(\omega)$
  - Optimist: $\max_{\omega\in\Omega} pl(\omega)$
  - Compromise: $\max_{\omega\in\Omega} betP(\omega)$

Pignistic probability:

$$\mathrm{betP}(\omega) = \sum_{Y\in 2^\Omega, \omega\cap Y\neq\emptyset} \frac{1}{|Y|}\frac{m(Y)}{1-m(\emptyset)} \tag{3}$$

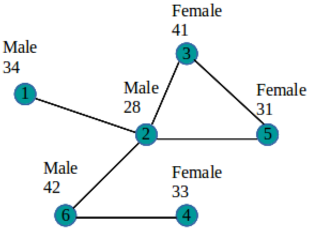
Information on social network

On social network some information can be considered:

- information on the links: LinkedIn, etc.
- information on the nodes: Facebook, LinkedIn, etc.
- information (message) throw the network: Tweeter, collaborative platforms, etc.

Node-attributed graphs

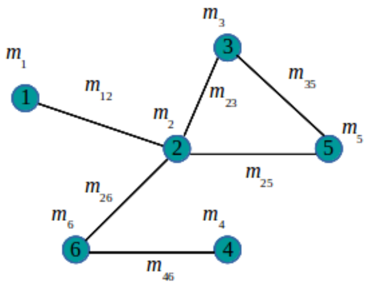
$G = (V, E, F)$ where $F : V \longrightarrow \mathcal{X}$
$F(v) = [f_1(v), \dots, f_a(v)]$



Attributes can be qualitative, quantitative (fuzzy, interval, probabilistic, belief, etc.).
see Christine Largeron talk...

Node and link-attributed graphs

$G = (V, E, m_u, m_e)$ where $m_u : V \longrightarrow \mathcal{X}$ and $m_e : e \in E \longrightarrow \mathcal{X}$
$m_u(v) = [m_1(v), \ldots, m_a(v)]$



(Ben Dhaou, 2014, 2017)

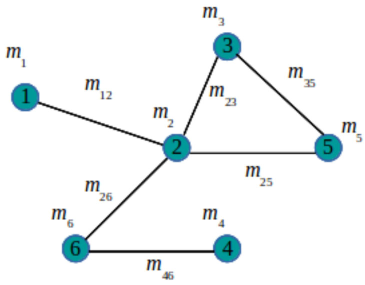
On social network some information can be considered:

- information on the links: LinkedIn, etc.
- information on the nodes: Facebook, LinkedIn, etc.
- information (message) throw the network: Tweeter, collaborative platforms, etc.

Characteristics of the messages:

- A message is a text (can be short text 140 characters on Twitter)
- That is not in general literature (many typos, errors, etc.)
- A message has an author
- A message can be send to some recipients
- A message has in general a date
- A message can have a label (type of message)
- A message can have an influence on the evolution of the network

Evolution of information on social network



Information on

- the existence of a node in the network
- the existence of a link between two nodes
- existence at time t can be model by a probability or a belief

Challenge: privacy-preserving data mining

How can we protect our personal data?
How do not send personal information?
What is personal, what is public?

- Cryptography and network security
- Watermarking (Gross-Amblard, 2003)
- Preference elicitation in Personal Information management Systems (Allard et al., 2017)

See Oana Goga talk.

Outline

1. What is a social network?
2. How to model a social network?
3. How to model information on social networks?
4. How to analyse social network?

Message mining

Challenges:

- Understand the messages
- Characterise emotion in the message
- Characterise the writer by the text (level of expertise, social level, etc.)
- Characterise positive/negative/neutral message
- Detect fake news
- Detect new topics, interest centres, etc.

Methods: coming from text mining must be lingual independent, robust to the form of the message, time dependent, etc.

Person identification

Challenges:

- Find criminals on a social network
- Find influencers for viral marketing
- Find spammers on participating platforms
- Find experts on participating platforms
- etc.

Expert identification in stackoverflow

15854 · 5355

I committed by accident the wrong files into Git, but I haven't pushed the commit to the server yet.

How can I undo those commits?

git · git-commit · git-reset · git-revert

share edit · edited Nov 19 at 16:51 · community wiki · 58 revs, 37 users 15% · Peter Mortensen

37 Warning: you should only do this if you have not yet pushed the commit to a remote, otherwise you will mess up the history of others who have already pulled the commit from the remote! – thSoft May 13 '15 at 21:18

12 Here's a very clear and thorough post about undoing things in git, straight from Github. – Nobita Jun 8 '15 at 19:39

2 This is a great resource straight from Github: How to undo (almost) anything with Git – jasonleonhard Feb 3 at 21:13

6 Before you post a new answer, consider there are already 65+ answers for this question. Make sure that your answer contributes what is not among existing answers. – Sazzad Hissain Khan Jun 15 at 15:26

show 3 more comments

71 Answers · active · oldest · votes

1 · 2 · 3 · next

16791

Undo a commit and redo

```
$ git commit -m "Something terribly misguided"             (1)
$ git reset HEAD~                                           (2)
<< edit files as necessary >>                               (3)
$ git add ...                                               (4)
$ git commit -c ORIG_HEAD                                   (5)
```

1. This is what you want to undo
2. This leaves your working tree (the state of your files on disk) unchanged but undoes the commit and leaves the changes you committed unstaged (so they'll appear as "Changes not staged for commit" in `git status`, and you'll need to add them again before committing). If you *only* want to *add* more changes to the previous commit, or change the commit message[1], you could use `git reset --soft HEAD~` instead, which is like `git reset HEAD~` but leaves your existing changes staged.
3. Make corrections to working tree files.
4. `git add` anything that you want to include in your new commit.
5. Commit the changes, reusing the old commit message. `reset` copied the old head to `.git/ORIG_HEAD`; `commit` with `-c ORIG_HEAD` will open an editor, which initially contains the log message from the old commit and allows you to edit it. If you do not need to edit the message, you could use the `-C` option.

Up Votes → UpV_i

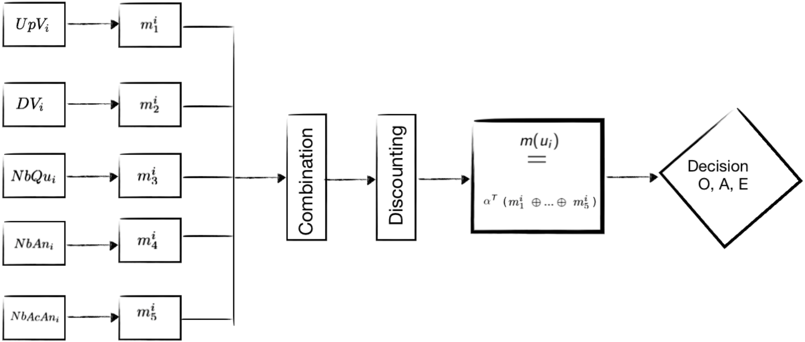
Down Votes → DV_i

Number of Questions → $NbQu_i$

Number of Answers → $NbAn_i$

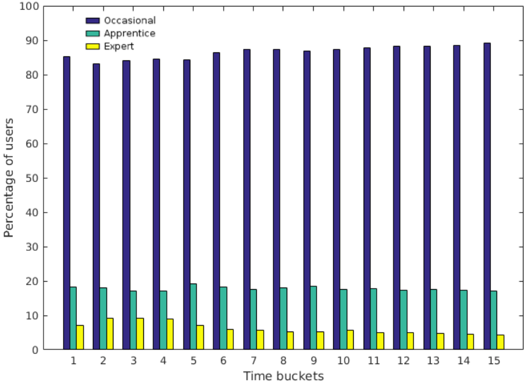
Number of Acc Answers → $NbAcAn_i$

Expert identification in stackoverflow



(Attiaoui, et al. 2017)

Expert identification in stackoverflow



Evolution of the percentage of each class over 15 months.

Data set: 37 Go, 2 Million users, 2.5 Million answers, 1.7 Million questions, Data from December 2013 to March 2015

Influencers identification

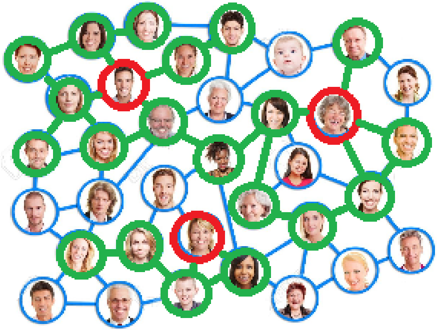
Problem: Given a social network, find a set of influencers that are able to trigger a large cascade.

Influencers identification

Problem: Given a social network, find a set of influencers that are able to trigger a large cascade.

Influencers identification

Problem: Given a social network, find a set of influencers that are able to trigger a large cascade.



Solution: Influencers on Twitter (Jendoubi, et al, 2016, 2017)

- Define an influence measure based on belief functions by:
  - $\Omega = \{I, P\}$ I for influencer, P for passive
  - Calculate belief weights on each edge (u, v)
  - Integrate opinion of tweet
  - Combine the mass functions
- Compute influence maximisation by CELF algorithm (Leskovec et al. 2007)

Influencers identification

Solution: Influencers on Twitter (Jendoubi, et al, 2016, 2017)

- Define an influence measure based on belief functions by:
  - $\Omega = \{I, P\}$ I for influencer, P for passive
  - Calculate belief weights on each edge (u, v) from numbers of common neighbours, number of tweets where u mentions v, number of tweets where v retweets from u
  - Integrate opinion of tweet
  - Combine the mass functions
- Compute influence maximisation by CELF algorithm (Leskovec et al. 2007)

Solution: Influencers on Twitter (Jendoubi, et al, 2016, 2017)

- Define an influence measure based on belief functions by:
  - $\Omega = \{I, P\}$ I for influencer, P for passive
  - Calculate belief weights on each edge (u, v)
  - Integrate opinion of tweet
    - Give a label to each word in the tweet using Stanford POS Tagger with the model GATE Twitter part-of-speech tagger,
    - Use the SentiWordNet dictionary to get the polarity of each word in the tweet
    - Build a belief function on $\Theta = \{Pos, Neg, Neut\}$
  - Combine the mass functions
- Compute influence maximisation by CELF algorithm (Leskovec et al. 2007)

Community detection

Define first type of communities expected:

- Hard communities: each node v belongs to one and only one community in $\Omega = \{C_1, \ldots, C_n\}$

$$\begin{cases} \mu_{vk} = 1 \text{ if } v \in C_k \\ \mu_{vk} = 0 \text{ otherwise} \end{cases}$$

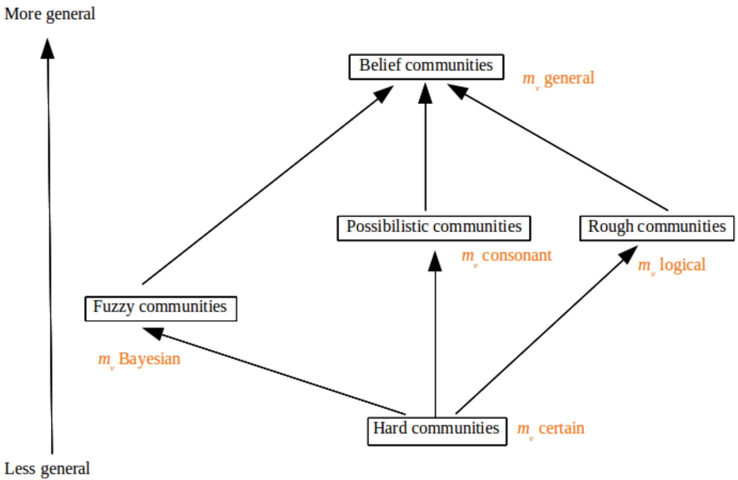
- Fuzzy communities: each node v has a degree of membership $\mu_{vk} \in [0, 1]$ to each community with $\sum_{k=1}^{n} \mu_{vk} = 1$

Community detection

Define first type of communities expected:

- Possibilistic communities: the condition $\sum_{k=1}^{n} \mu_{vk} = 1$ is relaxed. μ_{vk} can be interpreted as a degree of possibility that a node v belongs to the community C_k
- Rough communities: the membership of node v to community C_k is described by a pair $(\underline{\mu}_{vk}, \overline{\mu}_{vk}) \in \{0, 1\}^2$ indicating its membership to the lower and upper approximations of community C_k
- Belief communities: the membership of each node v is described by a belief function m_v over Ω.

Community detection



Community detection

Define first type of communities expected:

- Hard communities: each node v belongs to one and only one community in $\Omega = \{C_1, \ldots, C_n\}$
- Overlapped communities: each node v belongs to more than one community in Ω, $C_1, \ldots, C_n$ are not exclusive

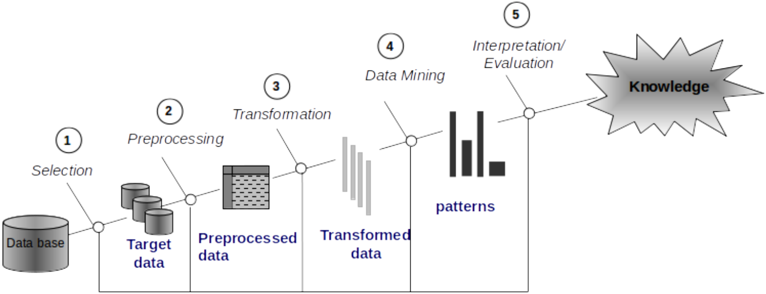
  With belief functions, work on D^Ω, hyper power set in order to model the overlapped communities:
  - D^Ω closed set by union and intersection operators
  - D_r^Ω: reduced set with constraints $(C_2 \cap C_3 \equiv \emptyset)$

See Rémy Cazabet talk...

Community detection

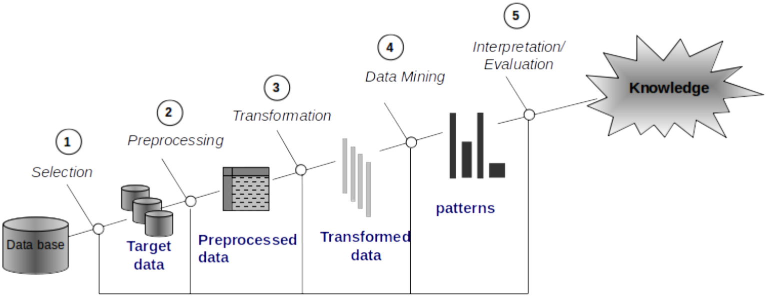
Methods: Depend on information in input and expected in output

1 Selection: can be from databases by requests, or by scanning the web



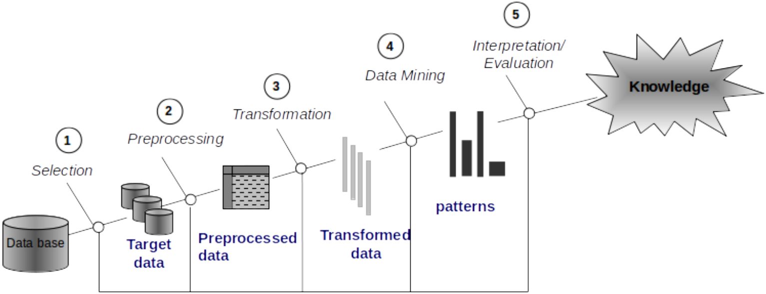
Methods: Depend on information in input and expected in output

2 Preprocessing: depend on the data, transform the data in graph, list of adjacent nodes, belief functions information, etc.



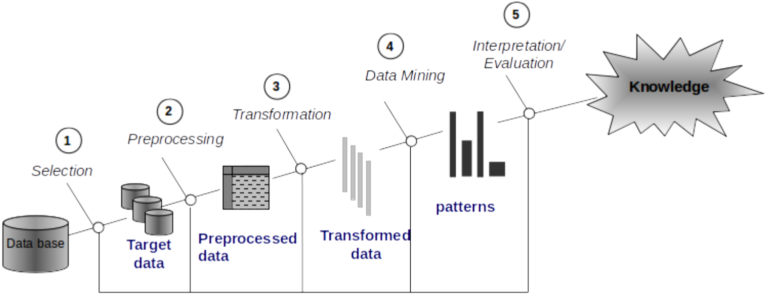
Methods: Depend on information in input and expected in output

3 Transformation: Calculate extracted feature (by supervised or unsupervised methods)



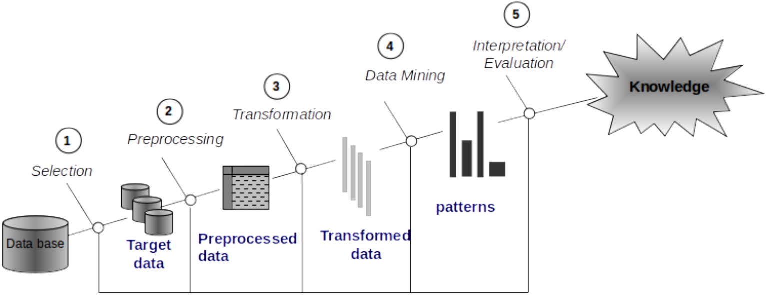
Methods: Depend on information in input and expected in output

4 Data Mining: Classify the data (by supervised or unsupervised methods)



Methods: Depend on information in input and expected in output

5 Evaluation: Calculate some measures on the obtained patterns



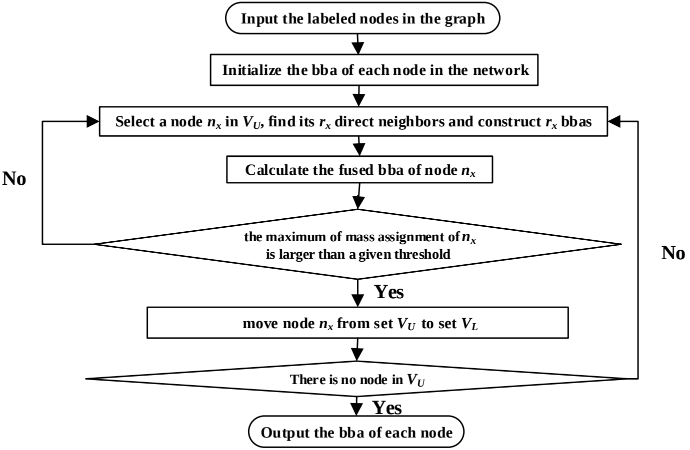
Characterisation of classical clustering methods (a challenge):

1. hierarchical methods by division or agglomeration build partitions
   Examples: Louvain algorithm, spectral approaches, etc.
2. partitioning methods:
   Examples: C-means, Fuzzy C-means, Evidential C-means (Zhou et al., 2015)
3. Label propagation methods

Need

- a distance (or similarity) on data (structure of the graph and information on the graph)
- an optimisation process

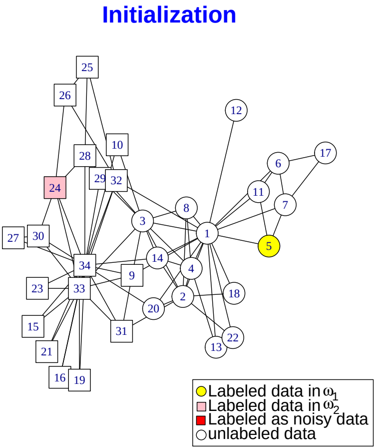
Community detection: SELP



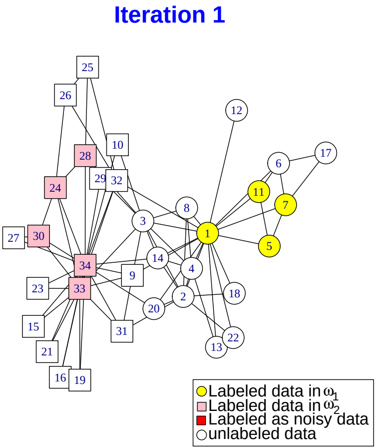
Semi-supervised Evidential Label Propagation algorithm (Zhou et al., 2018)

Community detection: SELP

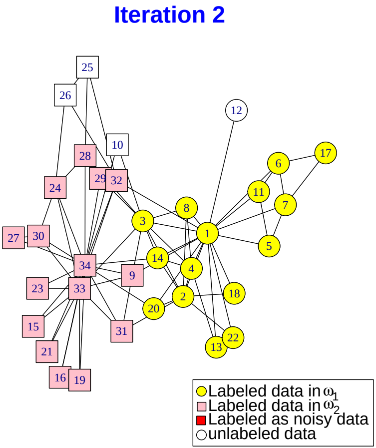
Example on Karate Club network



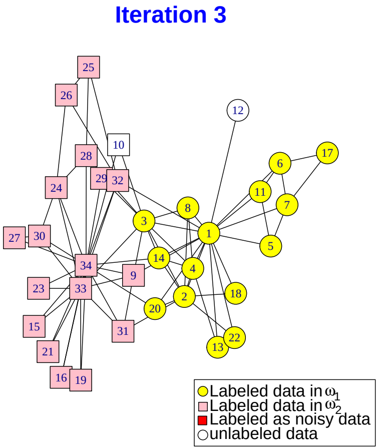
Example on Karate Club network



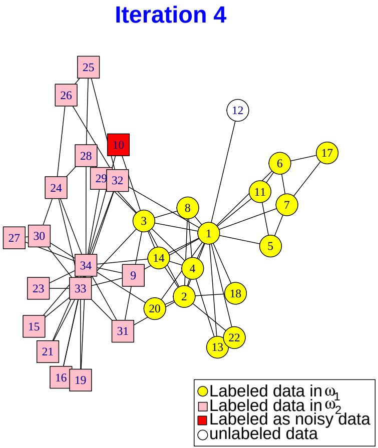
Example on Karate Club network



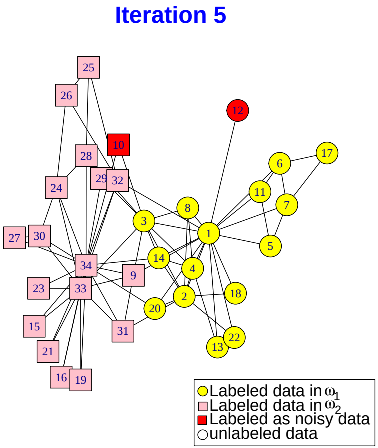
Example on Karate Club network



Example on Karate Club network



Example on Karate Club network



Challenges:

- How to learn information on graphs?
- How many communities? A difficult problem in clustering in general
- How to combine methods? Methods of information fusion can be used
- How to well consider the dynamic aspect of social network?
- How to reduce the time consuming of algorithms? Some algorithms can be parallelised
- How to evaluate the obtained communities? A difficult problem in clustering, more difficult when we don't know what is a community.
- etc.

To end

Many challenges around social networks

- We don't know exactly what is a social network
- We are not sure of given information on social network (veracity, precision, existence, etc.)
- We don't know exactly what is a community
- We have a lot of information
- Almost all our problems need a NP-hard algorithm

Next presentations during these two days will give you some answers.

To end

Many challenges around social networks

- We don't know exactly what is a social network
- We are not sure of given information on social network (veracity, precision, existence, etc.)
- We don't know exactly what is a community
- We have a lot of information
- Almost all our problems need a NP-hard algorithm

Next presentations during these two days will give you some answers.

My proposal: use the theory of belief functions in order to well model uncertainty and imprecision of information

References

- Stanford POSTagger: http://nlp.stanford.edu/software/tagger.shtml
- GATE Twitter part-of-speech tagger: https://gate.ac.uk/wiki/twitter-postagger.html
- SentiWordNet: http://sentiwordnet.isti.cnr.it/
- Santo Fortunato, Community detection in graphs. Physics Reports, 486(3):75-174, 2010
- Santo Fortunato, Darko Hric, Community detection in networks: A user guide, Physics Reports, 659, pp 1-44, 2016
- Guy Melancon, Just how dense are dense graphs in the real world?: a methodological note. In Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization, pp 1-7. ACM, 2006
- C. Largeron, P.N. Mougel, R. Rabbany, O.R. Zaiane, Generating attributed networks with communities, PloS one 10(4), 2015

References

- David Gross-Amblard, Query-preserving watermarking of relational databases and XML documents, Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2003
- Tristan Allard, Tassadit Bouadi, Joris Duguépéroux, Virginie Sans, From Self-Data to Self-Preferences: Towards Preference Elicitation in Personal Information Management Systems, International Workshop on Personal Analytics and Privacy (In conjunction with ECML PKDD 2017)
- Shafer, G. A mathematical theory of evidence. Princeton University Press, (1976)
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, Cost-effective outbreak detection in networks, KDD 2007

Personal references: with belief functions

- Kuang Zhou, Arnaud Martin, Quan Pan, Zhunga Liu, SELP: Semi-supervised evidential label propagation algorithm for graph data clustering, International Journal of Approximate Reasoning, Elsevier, 2018, 92, pp.139-154
- Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane, Belief Temporal Analysis of Expert Users: case study Stack Overflow, Big Data Analytics and Knowledge Discovery DAWAK, Aug 2017, Lyon, France
- Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane, Belief Measure of Expertise for Experts Detection in Question Answering Communities: case study Stack Overflow, 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Sep 2017, Marseille, France

Personal references: with belief functions

- Siwar Jendoubi, Arnaud Martin, Ludovic Liétard, Hend Ben Hadji, Boutheina Ben Yaghlane, Two Evidential Data Based Models for Influence Maximization in Twitter, Knowledge-Based Systems, 2017
- Salma Ben Dhaou, Kuang Zhou, Mouloud Kharoune, Arnaud Martin, Boutheina Ben Yaghlane, The Advantage of Evidential Attributes in Social Networks, 20th International Conference on Information Fusion, Jul 2017, Xi'an, China
- Kuang Zhou, Arnaud Martin, Quan Pan, Zhun-Ga Liu, Median evidential c-means algorithm and its application to community detection, Knowledge-Based Systems, 2015, 74, pp.69 - 88
- Kuang Zhou, Arnaud Martin, Quan Pan, A similarity-based community detection method with multiple prototype representation, Physica A: Statistical Mechanics and its Applications, Elsevier, 2015, pp.519-531

Personal references: www-druid.irisa.fr

- Imen Ouled Dlala, Dorra Attiaoui, Arnaud Martin, Boutheina Ben Yaghlane, Trolls Identification within an Uncertain Framework, International Conference on Tools with Artificial Intelligence - ICTAI, Nov 2014, Limassol, Cyprus
- Salma Ben Dhaou, Mouloud Kharoune, Arnaud Martin, Boutheina Ben Yaghlane, Belief Approach for Social Networks, Belief 2014, Oxford, United Kingdom
- Kuang Zhou, Arnaud Martin, Quan Pan, Evidential Communities for Complex Networks, 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Jul 2014