Classification of online disussions

Is tree structure sufficient enough?

Mattias MANO[1]

[1]Centre de Recherche en Gestion - CRG
École Polytechnique, France
Laboratoire de Recherche en Informatique - LRI
CentraleSupélec, France

mattias.mano[at]polytechnique.edu

Seminar LIP6 - Complex Networks, 30 sept. 2016

Outline

Outline

Thesis context

Financing and supervising:

- Open Online Problem-Solving - OOPS, Management Sciences

- université PARIS-SACLAY
- Jean-Michel DALLE, CRG, École Polytechnique et UPMC (Management sciences)
- Joanna TOMASIK, LRI, CentraleSupélec (Computer science)

Open Online Problem Solving

James Surowiecki (2004) - **Wisdom of the crowds**

Last several years:

- Development of online communities
- Arrival of "Q&A web sites": Yahoo! Answers, Stack Overflow, Bugzilla, Math Overflow, Reddit, ... $\rightarrow$ question $=$ problem
- Evolution of after-sale service management for companies (Velkovska, 2015)

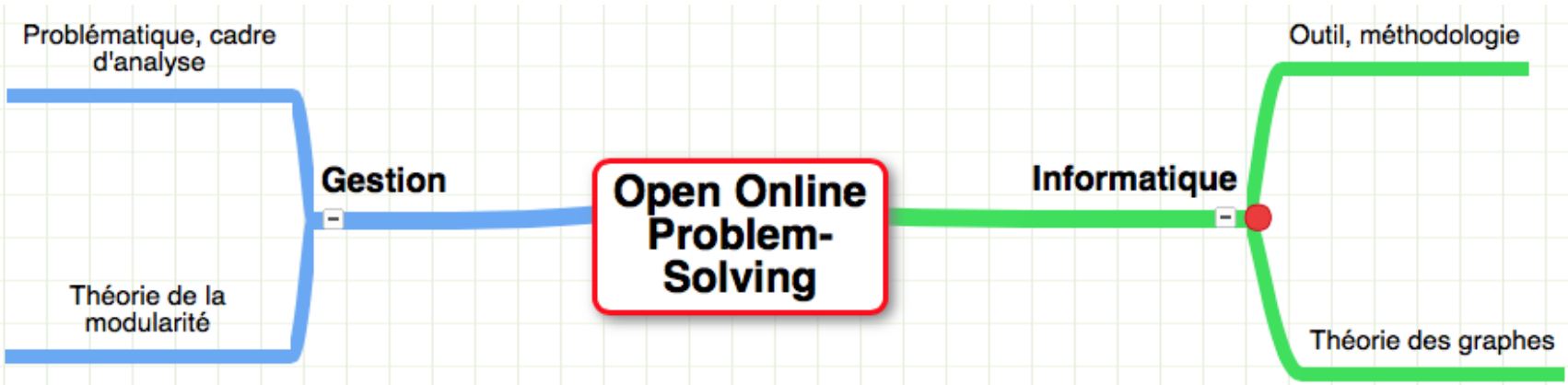
What has been done so far?

Important researches have been made concerning the problem solving:

- Test of **modularity theory** → McCormack et al. (2006)
- Notion of **congruence** → Cataldo et al. (2008)
- 'Distributed Problem Solving Networks' (DPSN) program led by P.A. David and W.H. Dutton from *Oxford Internet Institute* (2008)

SHS and STIC

This kind of researches needs a multidisciplinary approach between social science and computer science - **computational social science**.



What is the problem?

- What is the processus when a community, a group, try to solve a problem?
- Need to **characterize the problems** $\rightarrow$ understand such processus
- Charecterization of communities: Open Source (Dalle & David - 2003) or scientific community (Carayol & Dalle - 2006), "Core-Periphery" model (Halfaker et al. - 2012)

Issues to be addressed

1. Does a tree/forest shape exist? $\rightarrow$ Create a classification of the problems.
2. Could we determine the problem difficulty and the possibility to find a solution to it?

Outline

Literature review

Model for cascade evolution and pattern:

- Barabasi & Albert (1999): Preferential Attachment (PA) model
- Gómez, Kappen & Kaltenbrunner (2011): enhanced PA model
- Park & Barabasi (2007): (D,H)-phase diagram
- Tan, Luo & Peng (2012): enhanced Park-Barabasi model
- Milo, Shen-Orr et al. (2010): motif in networks

External variables:

- Dalle & David (2003): management within and among open source/free software (OS/FS) projects
- Carayol & Dalle (2006): problem choice within scientific communities
- Anderson, Huttenlocker et al. (2012): experience impactes time answering
- Tan, Niculae et al. (2016): entry time, back-and-forth, number of participants

Reddit - Change My View (CMV)

What is it? How does it work?

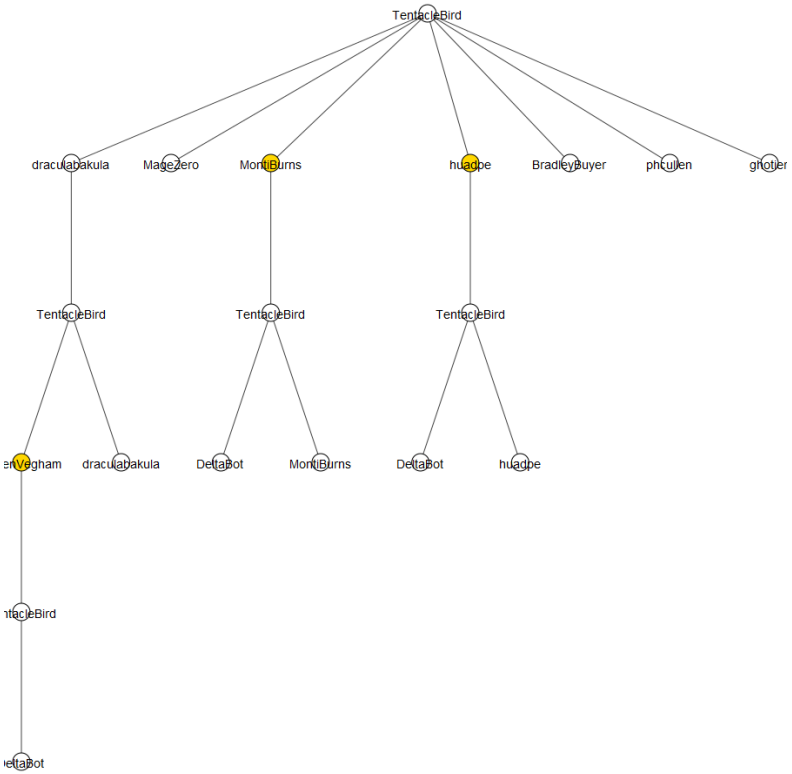
- Members argue on any subjects $\rightarrow$ *agora* in ancient Greece
- Important rules frame the discussion both on "Original Poster" and on challengers
- "Original Poster" (OP) opens a discussion, who "accept to have wrong and want help to change their view [...] within 3h"
- Members provide arguments to change the point of view of OP
- Anyone assignes a delta Δ, justifying a minimal change in the view $\rightarrow$ not the end of the discussion

Dataset 1/2

Where and when?

- Extraction from Reddit API by Tan et al. (2016)
- January 2013 (creation of CMV) to August 2015
- Sample from May to August 2015: 1'927 discussions, 111'811 posts, 1'606 Original Posters and 13'439 unique participants

Dataset 2/2



- *Internal data:* information on the shape of trees: depth, width, degree distribution, ...
- *External data:* information on the members of the community, on the posts: number of votes, experience in the forum, ...

Two paths

Follow two paths in parallel:

1. Characterisation of tree through its structure
2. Characterisation of tree through external information

Outline

Reddit - CMV: a typical PA-network? 1/2

Barabasi & Albert (1999): **Preferential-Attachment model** for the large networks:

- How a network evolve with time?
- Two common features of real networks:
  1. **Open world:** number of vertices N evolves with time
  2. **Preferential connectivity:** rich-gets-richer effect
- A new vertex appears and connect with probability Π:
- $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$, k_i: degree of vertex i
- $\Pi(k) \sim k^{\alpha} \rightarrow$ Power Law

Reddit - CMV: a typical PA-network? 2/2

Methodology: Gomez et al. (2014) testing enhanced PA-model, using a *maximum likelihood parameter estimation*

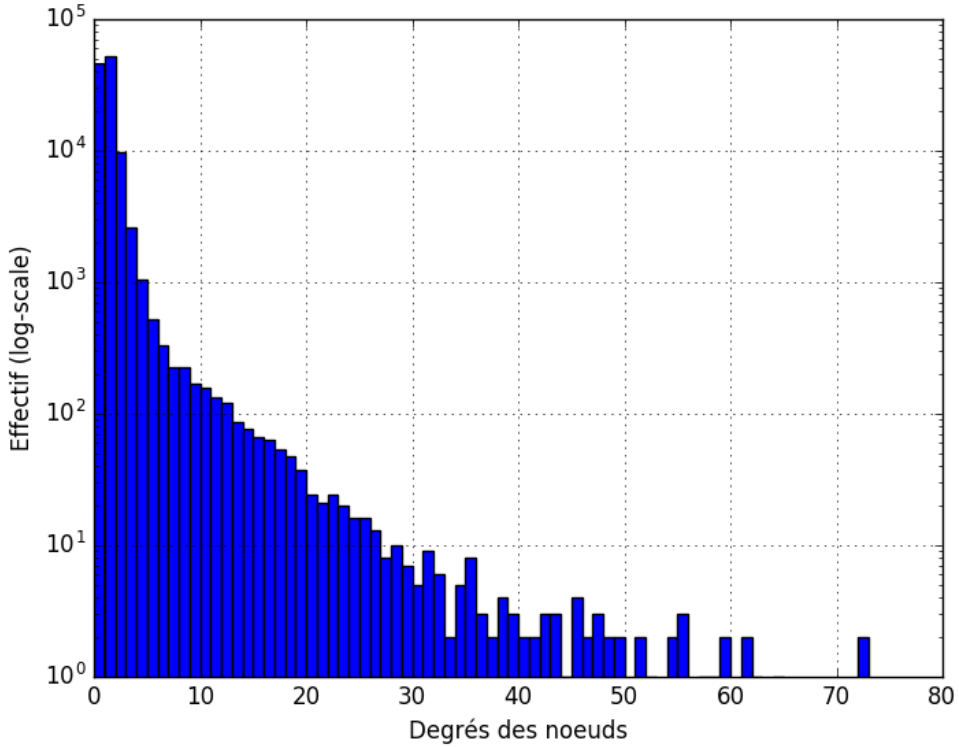


Figure: Degree distribution on the Reddit - CMV sample

Outline

Pattern inside the trees

Have a though about the rules: discussion within 3h, awarding Δ, gives up-vote $\rightarrow$ **influence on behaviors**

Pattern inside the trees

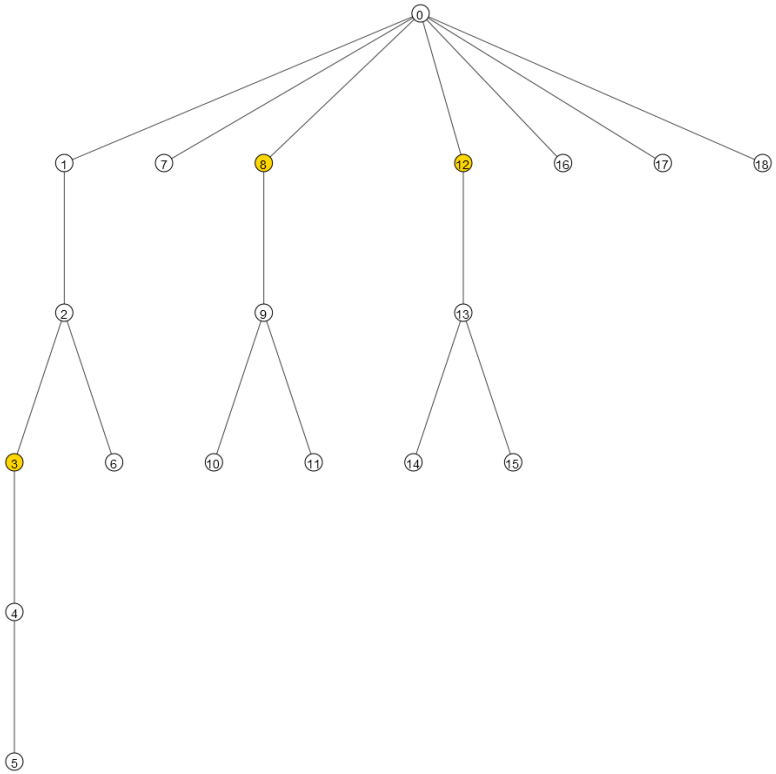
Have a though about the rules: discussion within 3h, awarding Δ, gives up-vote $\rightarrow$ **influence on behaviors**

Literature review gives us some important features when communities discuss:

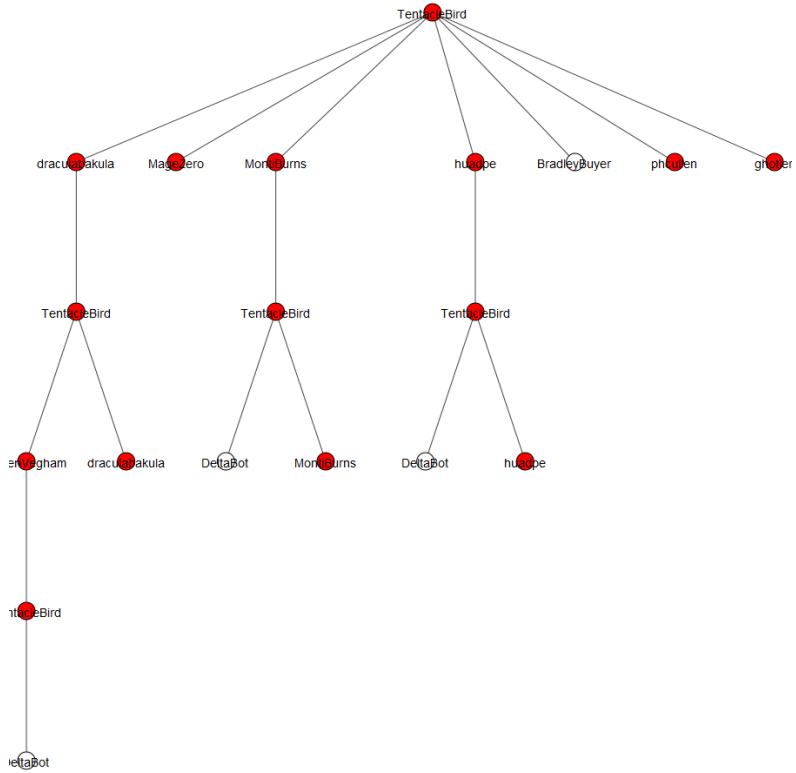
- **Experience:** how experienced a participant is? $\rightarrow$ number of Δ already received
- **Problem solved:** has the problem been solved? $\rightarrow$ a post gets a Δ when it convinces the OP
- **Vote:** members could vote for what they consider the "best" answer $\rightarrow$ better predictor than previous one
- **Time:** the sooner the better to get the reward

Value on the nodes?

Could these variables become intern to the tree structure?



Delta



Experience

Dyads 1/4

Park-Barabasi model for graph: how to characterise a network on a property?

- Property (P) takes only two values: 1 or 0 (P1 or P0)
- Graph: $N = n_1 + n_0$ & $M = m_{11} + m_{10} + m_{00}$
- P randomly distributed: $\bar{m}_{11} = \frac{n_1(n_1-1)}{2} p$, $\bar{m}_{10} = n_1 (N - n_1) p$ with *connectance* $p = \frac{2M}{N(N-1)}$ (how dense is the graph)

Dyads 2/4

If m_{11} (m_{10}) significantly deviate from $\bar{m}_{11}$ ($\bar{m}_{10}$) $\rightarrow$ **P is not distributed randomly!**

New indicators:

Dyads 2/4

If m_{11} (m_{10}) significantly deviate from $\bar{m}_{11}$ ($\bar{m}_{10}$) $\rightarrow$ **P is not distributed randomly!**

New indicators:

- **Dyadicity** $D = \frac{m_{11}}{\bar{m}_{11}}$

Dyads 2/4

If m_{11} (m_{10}) significantly deviate from $\bar{m}_{11}$ ($\bar{m}_{10}$) $\rightarrow$ **P is not distributed randomly!**

New indicators:

- **Dyadicity** $D = \frac{m_{11}}{\bar{m}_{11}} > 1 \rightarrow$ P1 is *dyadic*: nodes w/ P1 tend to connect more densely among themselves than expected for a random configuration
- **Heterophilicity** $H = \frac{m_{10}}{\bar{m}_{10}}$

Dyads 2/4

If m_{11} (m_{10}) significantly deviate from $\bar{m}_{11}$ ($\bar{m}_{10}$) $\rightarrow$ **P is not distributed randomly!**

New indicators:

- **Dyadicity** $D = \frac{m_{11}}{\bar{m}_{11}} > 1 \rightarrow$ P1 is *dyadic*: nodes w/ P1 tend to connect more densely among themselves than expected for a random configuration
- **Heterophilicity** $H = \frac{m_{10}}{\bar{m}_{10}} > 1 \rightarrow$ P1 is *heterophilic*: nodes w/ P1 have more connections to nodes w/ P0 than expected randomly

Dyads 2/4

If m_{11} (m_{10}) significantly deviate from $\bar{m}_{11}$ ($\bar{m}_{10}$) $\rightarrow$ **P is not distributed randomly!**

New indicators:

- **Dyadicity** $D = \frac{m_{11}}{\bar{m}_{11}} > 1 \rightarrow$ P1 is *dyadic*: nodes w/ P1 tend to connect more densely among themselves than expected for a random configuration
- **Heterophilicity** $H = \frac{m_{10}}{\bar{m}_{10}} > 1 \rightarrow$ P1 is *heterophilic*: nodes w/ P1 have more connections to nodes w/ P0 than expected randomly

We can calculate those indicators for each tree of the dataset and draw the phase diagram (Tan et al. (2012) applied to trees).

Dyads 3/4

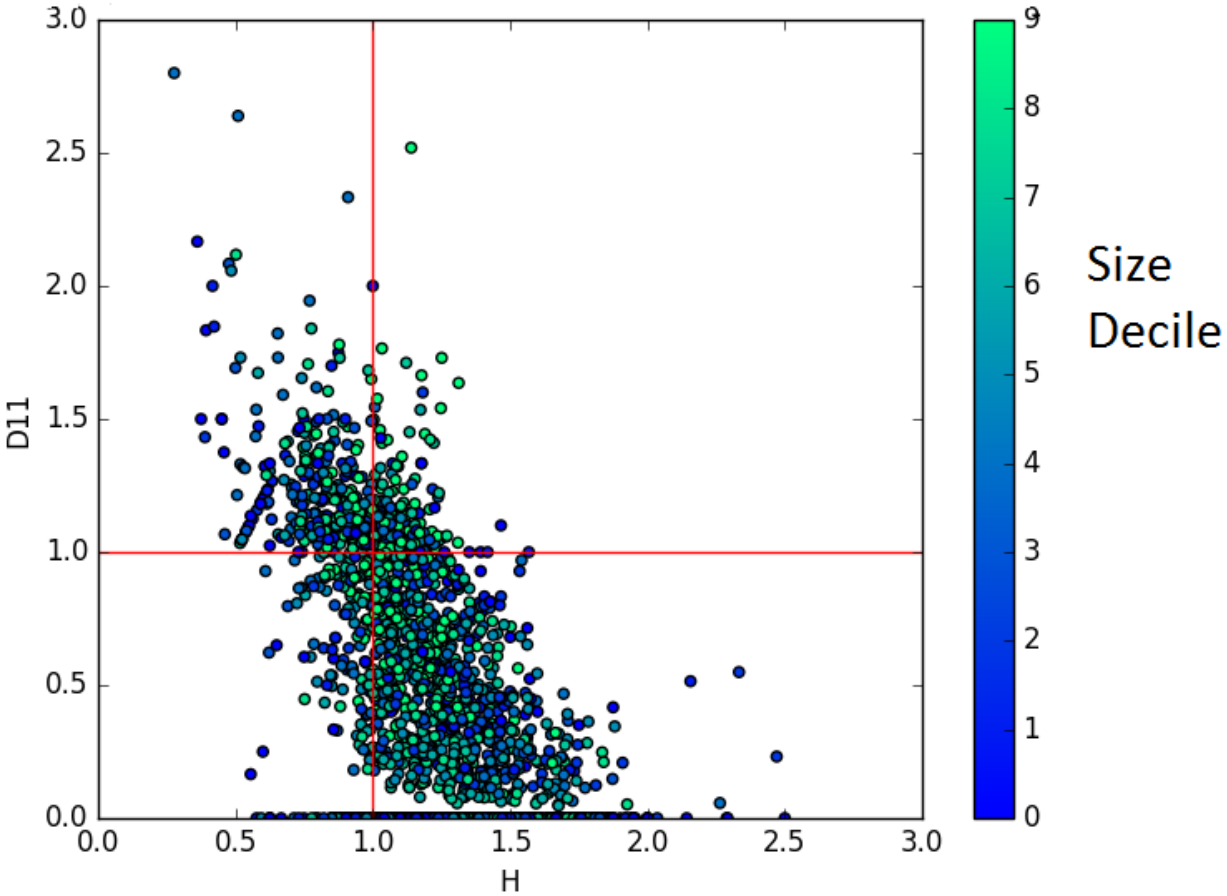


Figure: (D,H) - phase diagram for "experience" variable

Dyads 3/4

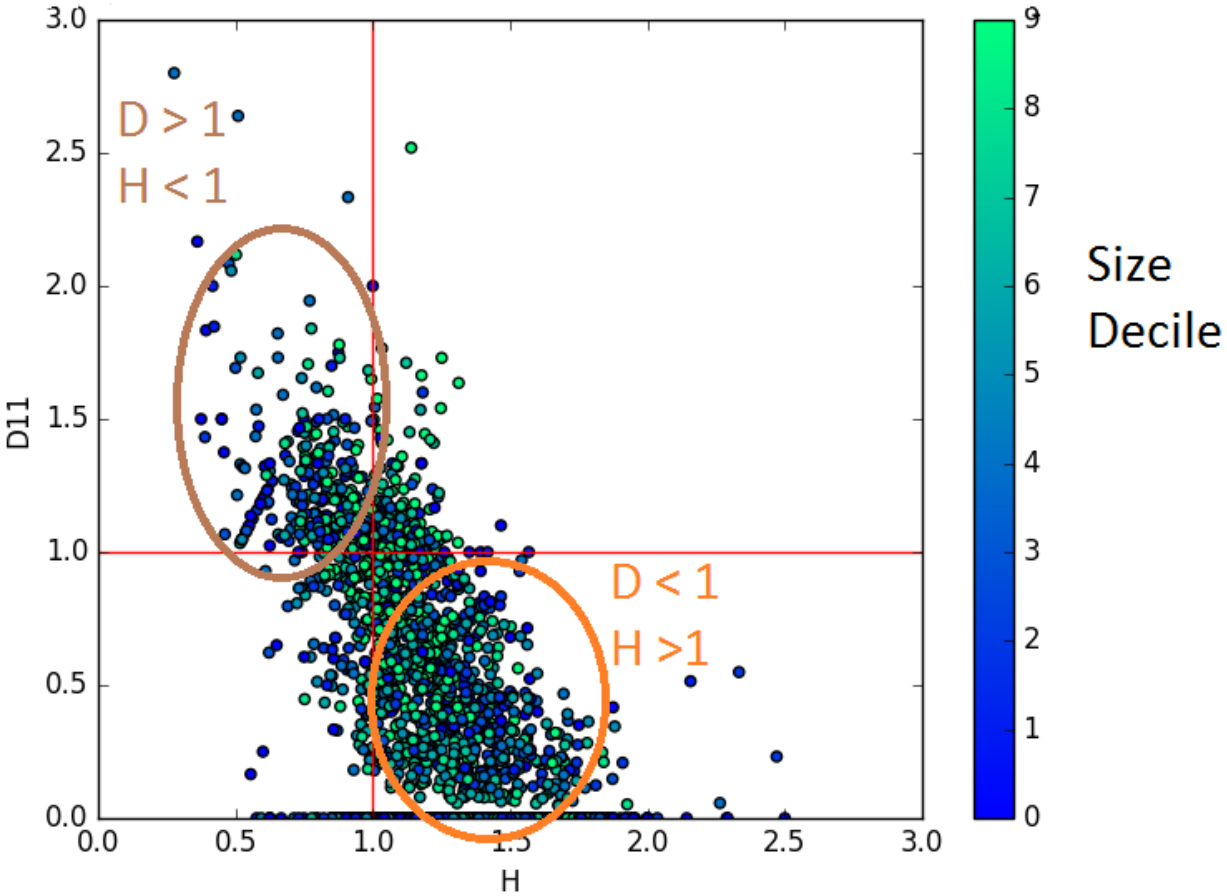


Figure: (D,H) - phase diagram for "experience" variable

Dyads 4/4

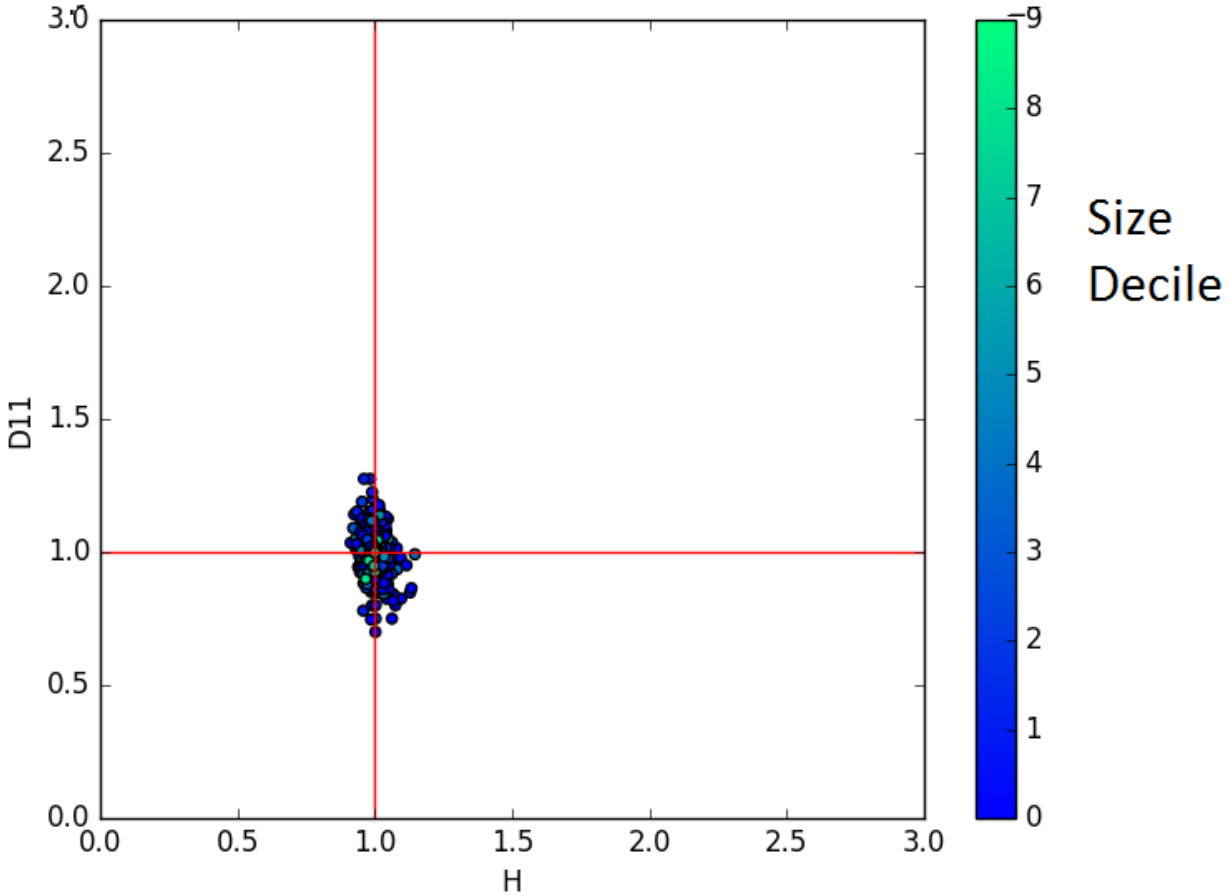


Figure: (D,H) - phase diagram for "experience" variable (simulated data)

Dyads 4/4

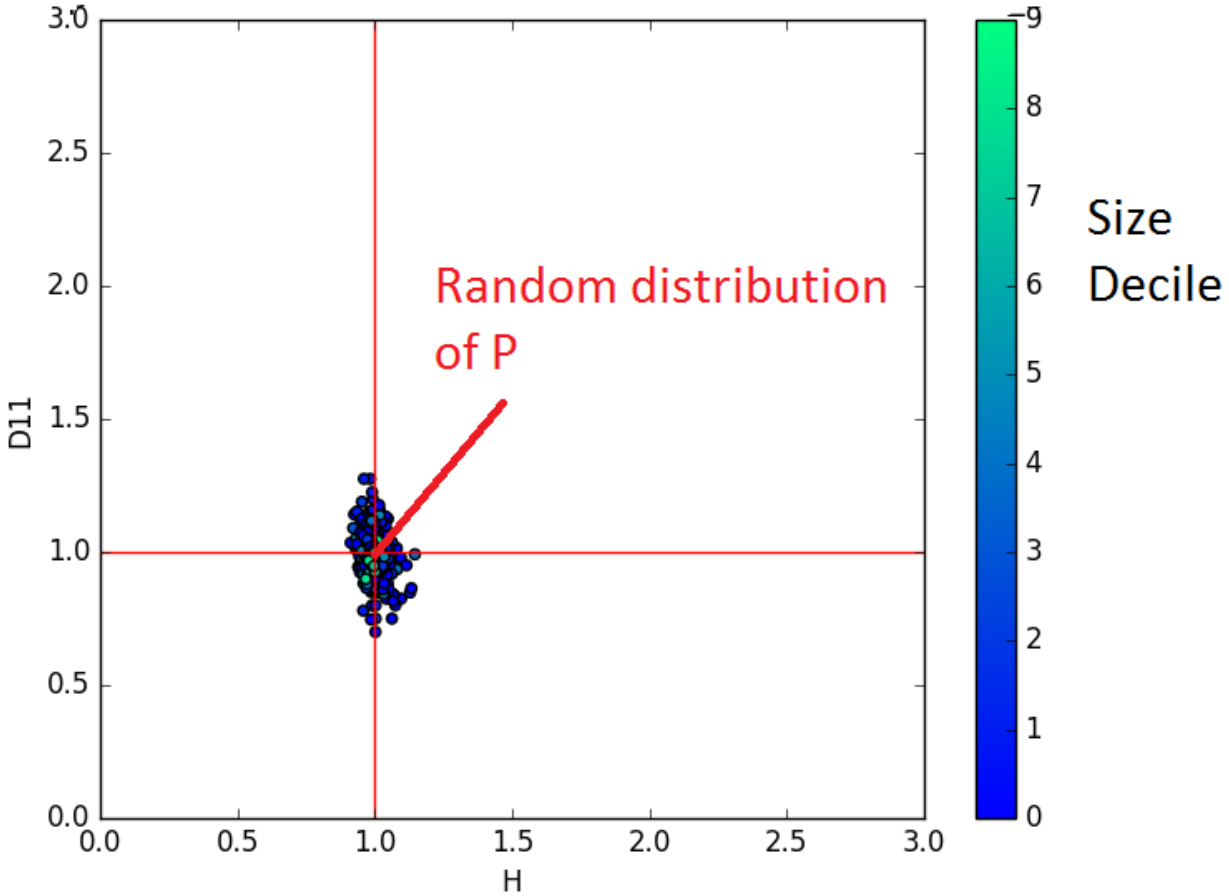
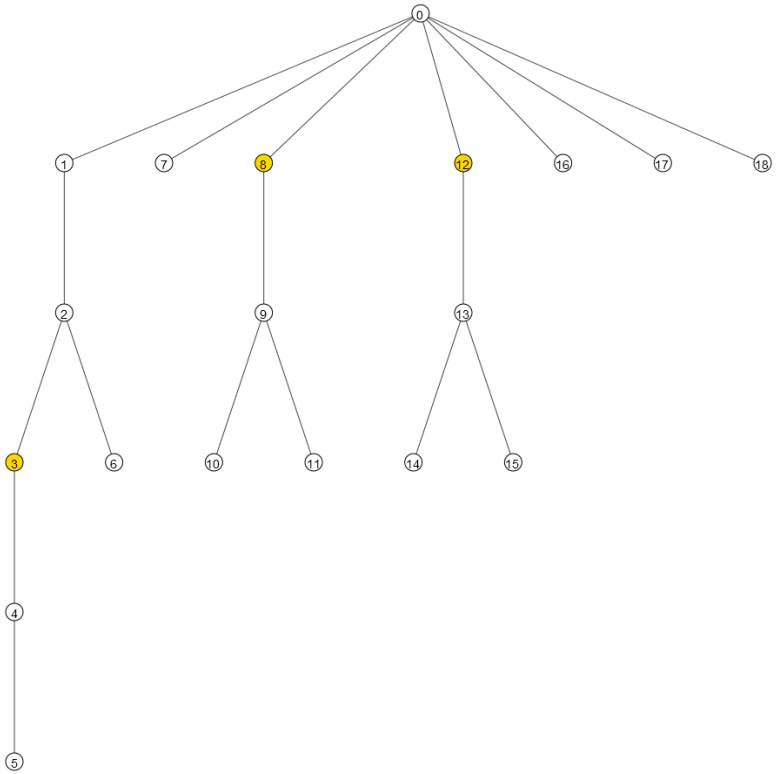


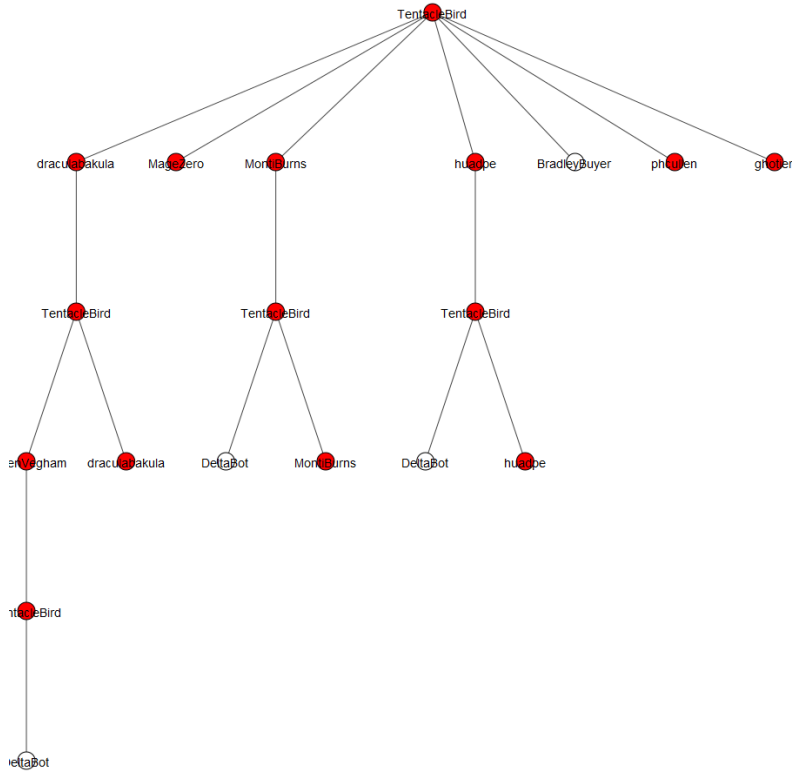
Figure: (D,H) - phase diagram for "experience" variable (simulated data)

Triads 1/2

What about triads?



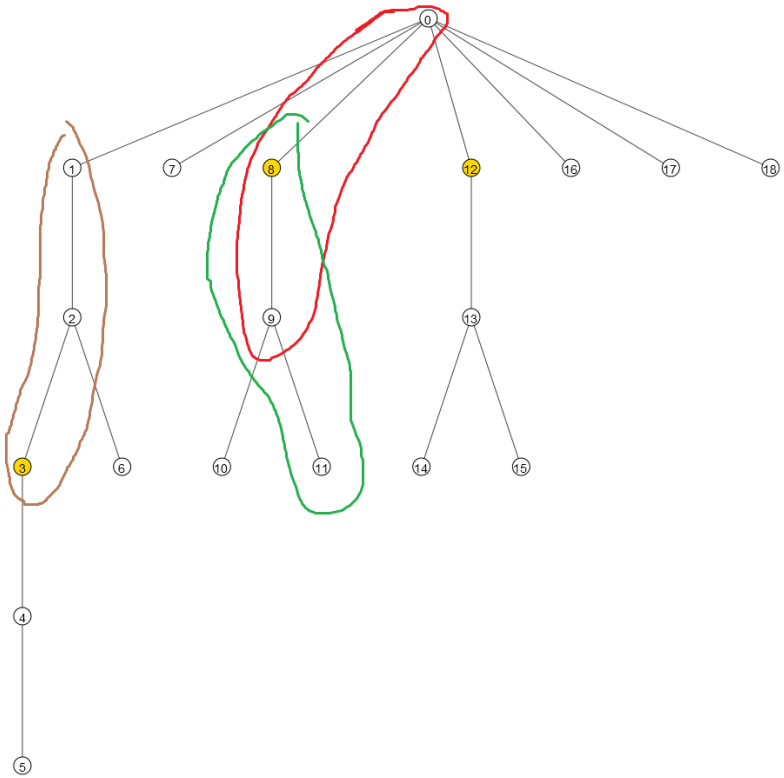
Delta



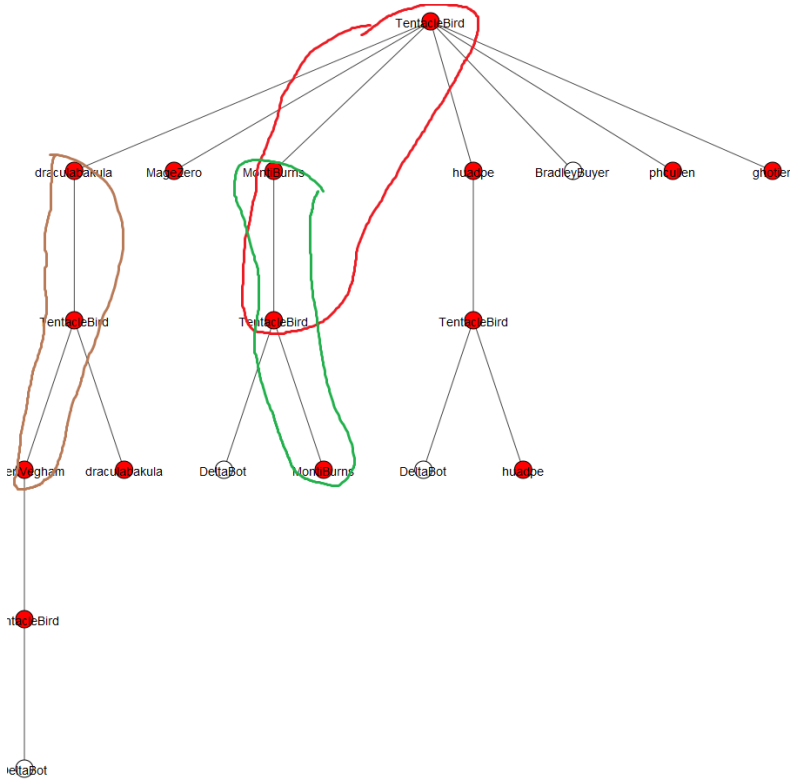
Experience

Triads 2/2

Highligt the 8 possible triads (Milo et al. 2010) : 000, 001, 010, 100, 101, 110, 011, 111.



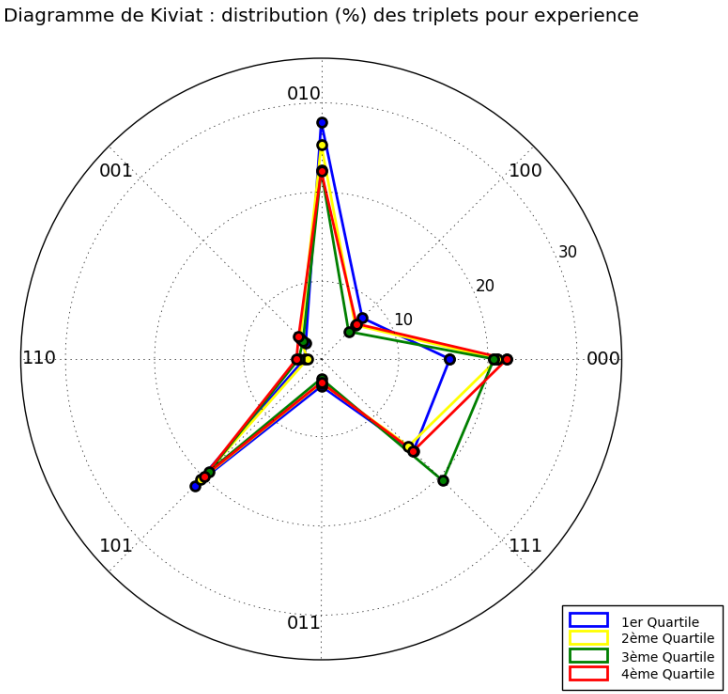
Delta



Experience

Any patterns?

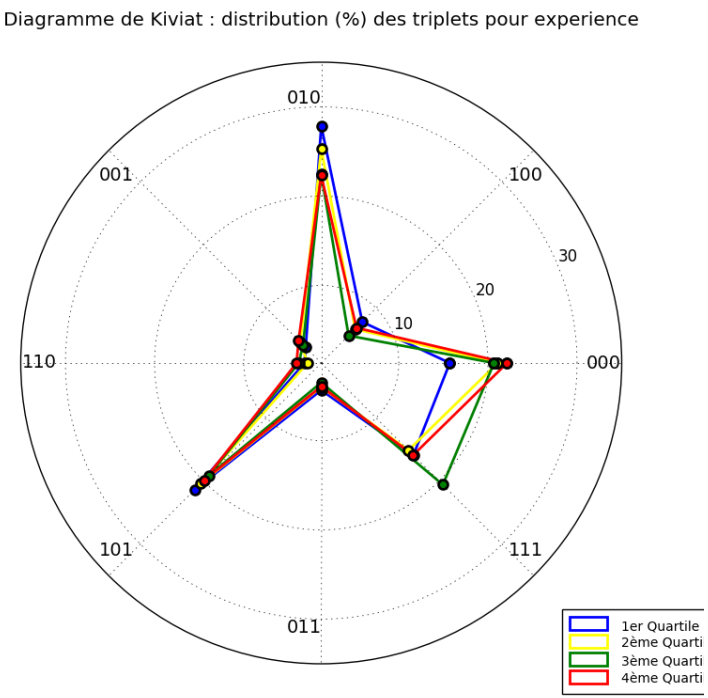
With the following property: "author has experience"



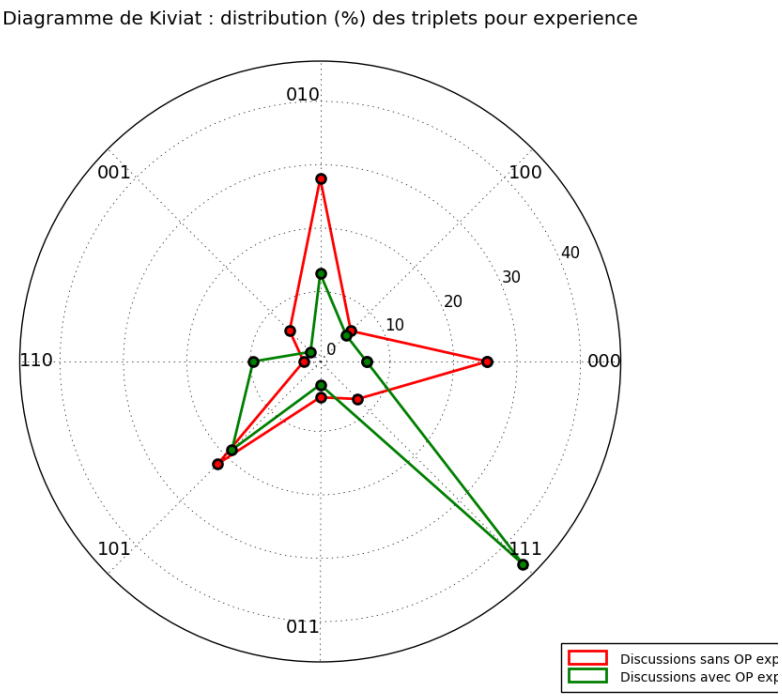
Size of discussions

Any patterns?

With the following property: "author has experience"



Size of discussions



(non) expert OP - green (red)

What next?

- Full implementation PA-model to add external variables → get a better fit
- Tree classification on a criterion → which one? Solving of the problem is not enough
- Extention to the 18'000 trees of the dataset
- Comparison with others forums (Coursera)

Classification of online disussions

Is tree structure sufficient enough?

Mattias MANO[1]

[1]Centre de Recherche en Gestion - CRG
École Polytechnique, France
Laboratoire de Recherche en Informatique - LRI
CentraleSupélec, France

mattias.mano[at]polytechnique.edu

Seminar LIP6 - Complex Networks, 30 sept. 2016