

SUNBELT XXX

**Evolving hypergraphs to appraise
academic team formation processes**

Carla TARAMASCO, Jean-Philippe COINTET and Camille ROTH

CREA (CNRS), INRA-SenS (INRA), CAMS (EHESS, CNRS) & ISCFIF

July 2, 2010

Mechanisms of academic collaboration

Academic collaborations — a long tradition of research :

- collaborative activity growth (*M. Smith, 1958*)
- International collaboration growth (*Wagner, C. S., Leydesdorff, L., 2005*)
- co-authorship network as a complex evolving network, (*Moody, 2004*), (*Newman, M. E. J., 2004*)

Sources of collaborations — it's all about proximity

- Spatial/Physical (*Kraut, R.E., Fussell, S.R., Brennan, S.E., Siegel, J., 2002*), (*Katz, 2002*)
- Social distance (*WO Hagstrom, 1965*)
- Intellectual (*Cowan, R., Jonard, N., Zimmermann, J.-B, 2002*)

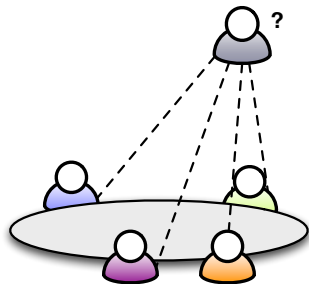
The Team Level & Networks

Limits when focusing on the level of individual

- only dyads, overlook the influence of characteristics expressable at the meso-level of the team itself
- team formation processes \neq sum of individual rationalities

Toward *meso-level* approach

- focus on *teams* rather than pairs of agents interacting together



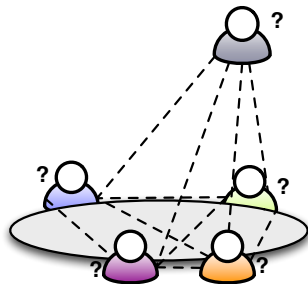
The Team Level & Networks

Limits when focusing on the level of individual

- only dyads, overlook the influence of characteristics expressable at the meso-level of the team itself
- team formation processes \neq sum of individual rationalities

Toward *meso-level* approach

- focus on *teams* rather than pairs of agents interacting together
- hyperlinks not cliques



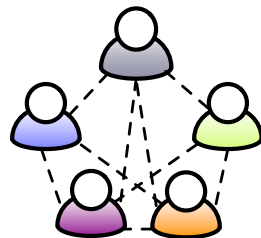
The Team Level & Networks

Limits when focusing on the level of individual

- only dyads, overlook the influence of characteristics expressable at the meso-level of the team itself
- team formation processes \neq sum of individual rationalities

Toward *meso-level* approach

- focus on *teams* rather than pairs of agents interacting together
- hyperlinks not cliques



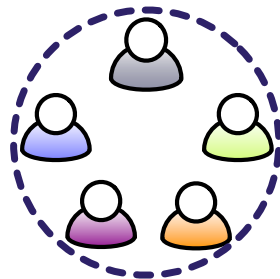
The Team Level & Networks

Limits when focusing on the level of individual

- only dyads, overlook the influence of characteristics expressable at the meso-level of the team itself
- team formation processes \neq sum of individual rationalities

Toward *meso-level* approach

- focus on *teams* rather than pairs of agents interacting together
- hyperlinks not cliques

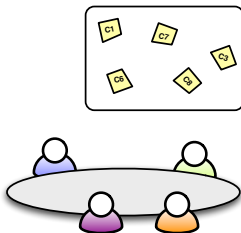


Hybrid Networks : Actors and Concepts

Collaboration also depends on cognitive properties

epistemic dynamics = reconfiguration of collectives made of :

- actors,
- concepts



Question :

How new teams are made given **past teams composition** ?

Datasets

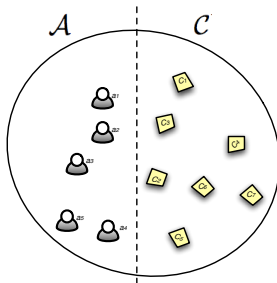
Experimental protocol

- 4 different datasets describing research production over ~ 20 years

Datasets

Experimental protocol

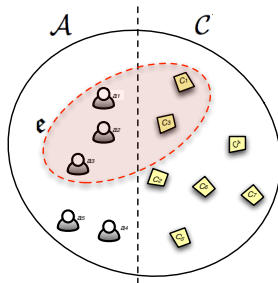
- 4 different datasets describing research production over ~ 20 years
- extract the set of agents \mathcal{A} and a set of pertinent concepts \mathcal{C}



Datasets

Experimental protocol

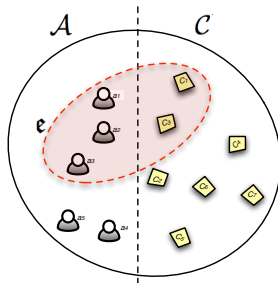
- 4 different datasets describing research production over ~ 20 years
- extract the set of agents \mathcal{A} and a set of pertinent concepts \mathcal{C}
- each paper is defined as a hyperlink : $e \in \mathcal{P}(\mathcal{A} \cup \mathcal{C})$, that is the joint grouping of both agents and concepts



Datasets

Experimental protocol

- 4 different datasets describing research production over ~ 20 years
- extract the set of agents \mathcal{A} and a set of pertinent concepts \mathcal{C}
- each paper is defined as a hyperlink : $\epsilon \in \mathcal{P}(\mathcal{A} \cup \mathcal{C})$, that is the joint grouping of both agents and concepts

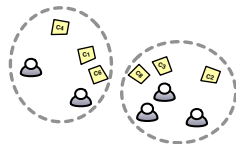


Projection operator

A hyperlink ϵ can be decomposed on any subset of $\mathcal{A} \cup \mathcal{C}$ with a projection operator.

$\epsilon^{\mathcal{A}} = \epsilon \cap \mathcal{A}$ is the set of co-authors of article ϵ (in our example : $\{a_1, a_2, a_3\}$),
 $\epsilon^{\mathcal{C}} = \epsilon \cap \mathcal{C}$ describes concepts used in ϵ (in this example : $\{c_1, c_3\}$)

A dynamic epistemic hypergraph



 \mathcal{E}_{t-1}

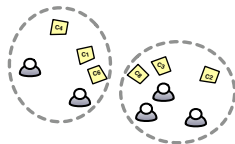
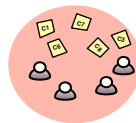
Epistemic Hypergraph

- *epistemic hypergraph* = triple $(\mathcal{A}, \mathcal{C}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{P}(\mathcal{A} \cup \mathcal{C})$

A dynamic epistemic hypergraph

Epistemic Hypergraph

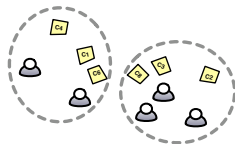
- *epistemic hypergraph* = triple $(\mathcal{A}, \mathcal{C}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{P}(\mathcal{A} \cup \mathcal{C})$

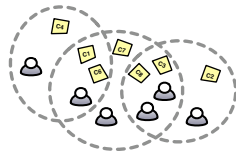

 \mathcal{E}_{t-1}

 $\cup \Delta \mathcal{E}_t$

A dynamic epistemic hypergraph

Epistemic Hypergraph

- *epistemic hypergraph* = triple $(\mathcal{A}, \mathcal{C}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{P}(\mathcal{A} \cup \mathcal{C})$
- The epistemic hypergraph is growing with time : \mathcal{E}_t


 \mathcal{E}_{t-1}

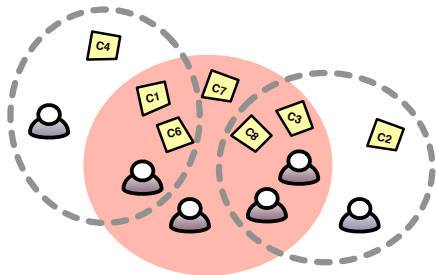
 $\cup \Delta \mathcal{E}_t$

 $= \mathcal{E}_t$

Definitions

Homogeneity of teams and expertise ratio

$\xi_c(e)$ expertise ratio of an article e given a concept $c \in e^c$:

$$\xi_c(e) = \frac{|\{a \in e^A \mid a \text{ expert in } c\}|}{|\{a \in e^A\}|}$$



examples :

$$\xi_{C_1}(e) = \frac{1}{4}$$

$$\xi_{C_6}(e) = \frac{1}{4}$$

$$\xi_{C_7}(e) = 0$$

$$\xi_{C_8}(e) = \frac{2}{4}$$

$$\xi_{C_3}(e) = \frac{2}{4}$$

Definitions

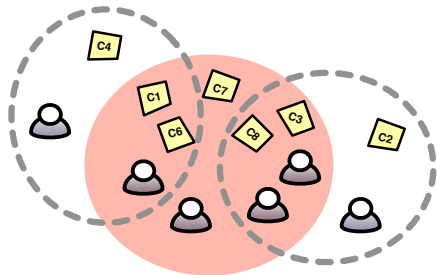
Hypergraphic repetition

- Originality of the composition of a team : *social originality* and *conceptual originality*
- Set of nodes repetition : is there at least one previously existing hyperlink including this set ?

$$\rho_t(\epsilon) = \begin{cases} 1 & \text{if } \exists e' \in \mathcal{E}_{t-1}, \epsilon \subseteq e' \\ 0 & \text{otherwise.} \end{cases}$$

- Hypergraphic repetition = proportion of subsets of ϵ that are repeated :

$$r_t(\epsilon) = \frac{1}{2^{|\epsilon|} - |\epsilon| - 1} \sum_{\substack{e' \subseteq \epsilon \\ |e'| \geq 2}} \rho_t(e')$$



examples :

social hypergraphic repetition rate $r_t(\epsilon^A) = \frac{1}{4}$

conceptual hypergraphic repetition rate $r_t(\epsilon^C) = \frac{2}{11}$

Definitions

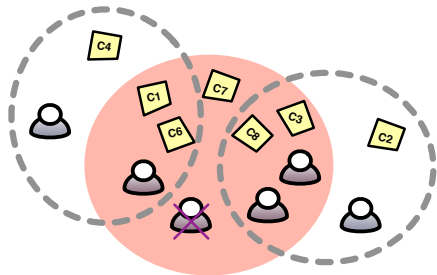
Hypergraphic repetition

- Originality of the composition of a team : *social originality* and *conceptual originality*
- Set of nodes repetition : is there at least one previously existing hyperlink including this set ?

$$\rho_t(\epsilon) = \begin{cases} 1 & \text{if } \exists e' \in \mathcal{E}_{t-1}, \epsilon \subseteq e' \\ 0 & \text{otherwise.} \end{cases}$$

- Hypergraphic repetition = proportion of subsets of ϵ that are repeated :

$$r_t(\epsilon) = \frac{1}{2^{|\epsilon|} - |\epsilon| - 1} \sum_{\substack{e' \subseteq \epsilon \\ |e'| \geq 2}} \rho_t(e')$$



examples :

social hypergraphic repetition rate $r_t(\epsilon^A) = \frac{1}{4}$

conceptual hypergraphic repetition rate $r_t(\epsilon^C) = \frac{2}{11}$

Definitions

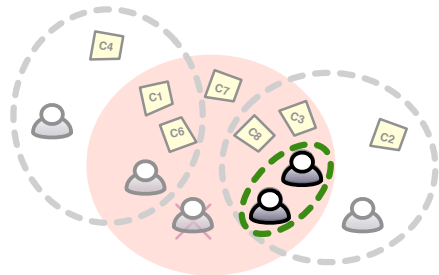
Hypergraphic repetition

- Originality of the composition of a team : *social originality* and *conceptual originality*
- Set of nodes repetition : is there at least one previously existing hyperlink including this set ?

$$\rho_t(\epsilon) = \begin{cases} 1 & \text{if } \exists \epsilon' \in \mathcal{E}_{t-1}, \epsilon \subseteq \epsilon' \\ 0 & \text{otherwise.} \end{cases}$$

- Hypergraphic repetition = proportion of subsets of ϵ that are repeated :

$$r_t(\epsilon) = \frac{1}{2^{|\epsilon|} - |\epsilon| - 1} \sum_{\substack{\epsilon' \subseteq \epsilon \\ |\epsilon'| \geq 2}} \rho_t(\epsilon')$$



examples :

social hypergraphic repetition rate $r_t(\epsilon^A) = \frac{1}{4}$

conceptual hypergraphic repetition rate $r_t(\epsilon^C) = \frac{2}{11}$

Definitions

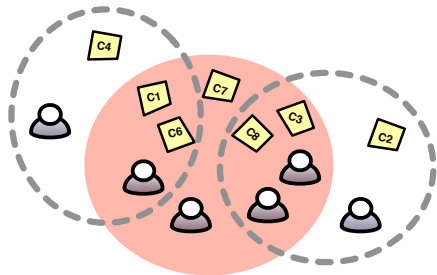
Hypergraphic repetition

- Originality of the composition of a team : *social originality* and *conceptual originality*
- Set of nodes repetition : is there at least one previously existing hyperlink including this set ?

$$\rho_t(\epsilon) = \begin{cases} 1 & \text{if } \exists \epsilon' \in \mathcal{E}_{t-1}, \epsilon \subseteq \epsilon' \\ 0 & \text{otherwise.} \end{cases}$$

- Hypergraphic repetition = proportion of subsets of ϵ that are repeated :

$$r_t(\epsilon) = \frac{1}{2^{|\epsilon|} - |\epsilon| - 1} \sum_{\substack{\epsilon' \subseteq \epsilon \\ |\epsilon'| \geq 2}} \rho_t(\epsilon')$$



examples :

social hypergraphic repetition rate $r_t(\epsilon^A) = \frac{1}{4}$

conceptual hypergraphic repetition rate $r_t(\epsilon^C) = \frac{2}{11}$

Estimating Propensities of team formation

Null-model of hypergraph

we generate at each time step a set of new teams $\widetilde{\Delta\mathcal{E}_t}$ which respects the following distributions :

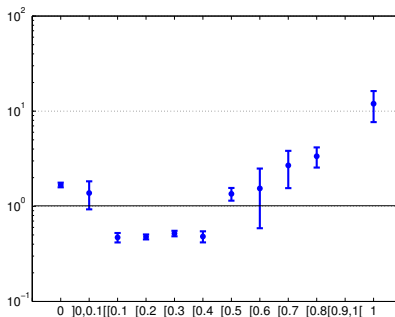
- same distribution of sizes of new hyperlinks (same dist. $|e^A|$ and $|e^C|$ for $e \in \Delta\mathcal{E}_t$)
- same distribution of participations of elements in these new hyperlinks.

Propensity

Given a measure f (e.g. hypergraphic repetition) on a hyperlink, we compute the likeliness that for a new team e , $f(e) = x$.

$$\Pi_t(x) = \frac{|\{e \in \Delta\mathcal{E}_t \text{ such that } f(e) = x\}|}{|\{e \in \widetilde{\Delta\mathcal{E}_t} \text{ such that } f(e) = x\}|}$$

Teams expertise ratio



The curve is **U-shaped** :

teams are more likely to be mainly composed with all-non-experts or all-experts,

mixed teams are less frequent than expected from our null-model

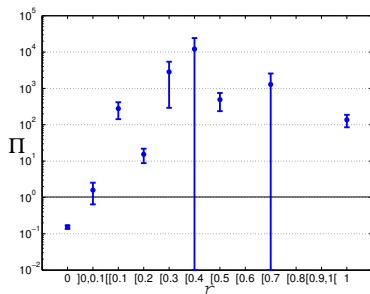
Propensity that team have a given expertise ratio

computed over 12 bins and shown on one dataset

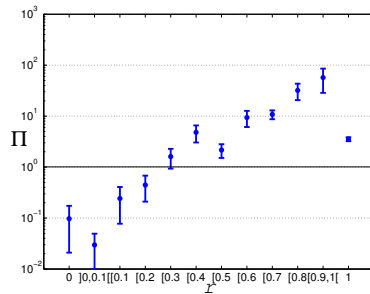
Teams hypergraphic repetition rates

Teams hypergraphic repetition rates propensity

Likelihood to produce teams with a given social (*left*) and conceptual (*right*) hypergraphic rate of repetition (computed over 12 bins and shown on one dataset)

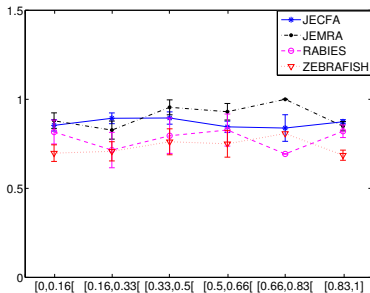


higher proportion of **interaction repetitions**



bias towards **gathering groups of concepts** which were previously associated

Are hypergraphic repetition rates correlated ?



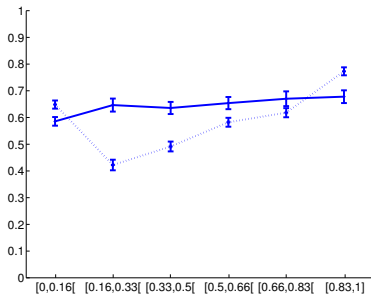
We observe no correlation

contrarily to intuition, new semantic associations **do not** stem more from original teams than from repeated teams

Correlation between *agents repetition rate* and *average semantic repetition rate*

Average semantic hypergraphic repetition rate (*y-axis*) for a given range of social hypergraphic repetition rate (*x-axis*), computed on 6 bins and shown for every datasets

Are hypergraphic repetition rates correlated with expertise ratio ?



We observe no correlation between expertise ratio and semantic originality

Social originality is increased when there is a mixed proportion of experts, but not semantic originality

Correlation between *expertise ratio* and *hypergraphic repetition rates*

Average hypergraphic repetition rates (*y-axis*) with respect to expertise ratios (*x-axis*) : social (*dashed line*) and semantic (*plain line*) cases, computed on 6 bins and shown for one dataset

Conclusion & Perspectives

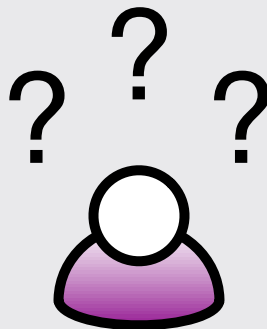
Main results

- formal framework to appraise the underpinnings of collaboration formation with a hypergraphic approach which encompasses both the **meso-level** of teams and the **joint dynamics of social and semantic features**.
- (i) high likeliness to **repeat previous collaborations patterns**, along with a polarization between groups made of experts only or made of non-experts only
(ii) similarly, sensible **semantic confinement** where associations of concepts depend largely on the repetition of previous associations.
(ii) the **originality** of a paper does not seem to stem from an original composition of the underlying team

Perspectives

- Realistic **reconstruction models** of academic collaboration taking into account an implicit group structure, both at a social and at a socio-semantic levels
- Which features, at the level-team, favor better collaborations — not only in terms of semantic originality, but also in terms of **quality and creativity** of output ?

Questions



Reference

Academic team formation as evolving hypergraphs

Taramasco, C., Cointet, J.P. and Roth, C., *Scientometrics*, 2010