

# Lattice-based dynamic & overlapping taxonomies: the case of epistemic communities

Camille Roth\*, Paul Bourguine\*

final version appeared in *Scientometrics*, 69(2):429-447, 2006

The original publication is available at <http://www.springerlink.com/content/08210r4u63947573>

## Abstract

We present a method for describing taxonomy evolution. We focus on the structure of epistemic communities (ECs), or groups of agents sharing common knowledge concerns. Introducing a formal framework based on Galois lattices, we categorize ECs in an automated and hierarchically structured way and propose criteria for selecting the most relevant epistemic communities — for instance, ECs gathering a certain proportion of agents and thus prototypical of major fields. This process produces a manageable, insightful taxonomy of the community. Then, the longitudinal study of these static pictures makes possible an historical description. In particular, we capture stylized facts such as field progress, decline, specialization, interaction (merging or splitting), and paradigm emergence. The detection of such patterns in epistemic networks could fruitfully be applied to other contexts.

*Keywords:* Social complex systems, Scientometrics, Categorization and Evolving taxonomies, Galois lattices, Applied epistemology, Knowledge discovery in databases.

## Introduction

Several formal frameworks and automated processes have been proposed to analyze knowledge community structure and find groups of agents

or documents related by common notions or concerns. The most important source of interest has come from knowledge discovery in databases (KDD) (Jain *et al.*, 1999), along with the massive development of informational content (in particular scientific data); and scientometrics (e.g. Leydesdorff, 1991), which have developed a whole set of methods for characterizing specifically such communities. Working on both articles, their authors and the concepts they use, the goal is to track the evolution of paradigms (Callon *et al.*, 1986; McCain, 1986); using *inter alia* multidimensional scaling in association with co-citation data (Kreuzman, 2001) or other co-occurrence data (Noyons & van Raan, 1998), in order to produce two-dimensional cluster mappings. Nonetheless, many approaches are either based on social relationships, with community extraction methods stemming from graph theory applied to social networks (Wasserman & Faust, 1994), or on semantic similarity, namely clustering methods applied to document databases where each document is considered as a vector in a semantic space (Salton *et al.*, 1975). There have been few attempts to link social and semantic aspects, although an epistemic community is a dual notion; on one side a group of agents who, on the other side, share common concepts.

Together with this profusion of community-finding methods, often leaning towards AI-oriented clustering, an interesting issue concerns the representation of communities in an ordered fashion. On the whole, many different techniques have been proposed for producing and representing categorical structures including, to cite a few, hierarchical clustering (Johnson, 1967), for-

---

\*CREA (Center for Research in Applied Epistemology), CNRS/Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France. Corresponding author: [camille.roth@polytechnique.edu](mailto:camille.roth@polytechnique.edu)

mal concept analysis (Wille, 1982), blockmodeling (White *et al.*, 1976; Doreian *et al.*, 2005), graph theory-based techniques (Newman, 2004), neural networks (Kohonen, 2000). Here, the notion of *taxonomy* — an ordered set of categories, or *taxons* — is particularly relevant with respect to communities of knowledge, while it has been widely used in biology (Whittaker, 1969), cognitive psychology (Rosch & Lloyd, 1978), as well as in ethnography (Berlin, 1992) and anthropology (Lopez *et al.*, 1997). Alongside, while taxonomies have initially been built using a subjective approach, the focus has moved to formal and statistical methods (Sokal & Sneath, 1963). However, taxonomy building itself is generally poorly investigated and often limited to dendrograms and tree-based structures which cannot efficiently deal with inter- and multi-disciplinarity. Arguably, taxonomy evolution during time has been fairly neglected. Our intent is thus to address both topics: build a taxonomy pertinent for epistemic community description, then monitor its evolution.

We therefore propose a method based on Galois lattices (Birkhoff, 1948; Barbut & Monjardet, 1970) to represent a relevantly reduced view of such a taxonomy. Then, we describe the community taxonomy in an historical perspective by studying the evolution of these static pictures. In particular, we rebuild stylized facts relating to epistemic evolution. These facts consist of field progress or decline, field scope enrichment or impoverishment, and field interaction (merging or splitting). This would be most useful for disciplines such as history of science and scientometrics. It would also provide agents with automated methods to know the structure of the community they are evolving in.

In section 1 we introduce the formal framework needed for representing epistemic community taxonomy using Galois lattices. Section 2 describes how to build reduced taxonomies and to monitor their evolution. A case study is investigated in section 3, followed by a general discussion and comparison with existing approaches in section 4.

## 1 Formal framework

**Definitions** We introduce the notion of *epistemic community*. In the literature (Cowan *et al.*, 2000;

Dupouet *et al.*, 2001; Haas, 1992), an epistemic community is a group of agents sharing a common set of subjects, topics, concerns, sharing a common goal of knowledge creation. In order to use this notion, we first have to bind agents to semantic items, or *concepts*.

To this end, we consider a binary relation  $\mathcal{R}$  between an agent set  $S$  and a concept set  $C$ .  $\mathcal{R}$  expresses any kind of relationship between an agent  $s$  and a concept  $c$ . The nature of the relationship depends on the hypotheses and the empirical data. In our case, the relationship represents the fact that  $s$  used  $c$  in some article. We may thus introduce two fundamental notions: the *intent* and the *extent*. The *intent*  $S^\wedge$  of an agent set  $S$  is the set of concepts that is being used by every agent in  $S$ . Similarly, the *extent*  $C^\star$  of a concept set  $C$  is the set of agents using every concept in  $C$ .

**Epistemic community** We then adopt the following definition: *an epistemic community (EC) is the largest set of agents sharing a given concept set*. Accordingly, an *EC based on an agent set* is the EC of its concept set. Such EC is the largest agent set having in common the same concepts as the initial agent set. In other words, taking the EC of a given agent set extends it to the largest community sharing its concepts. This notion strongly relates to structural equivalence (Lorrain & White, 1971), with ECs being groups of agents linking equivalently to some concepts.

The EC based on an agent set  $S$  is therefore the largest agent set with the same intent as  $S$ . It is then obvious that it is the extent of the intent of  $S$ , or  $S^{\wedge\star}$ . Thus, the operator " $\wedge\star$ " yields the EC of any agent set. Notice that we can similarly define an *EC based on a concept set* as the largest set of concepts sharing a given agent set. Here, one starts with a concept set and seeks to know its corresponding EC using the operator " $\star\wedge$ ". The EC based on a concept set  $C$  is the same as that based on an agent set  $S = C^\wedge$ . Hence, in the remainder of the paper we equivalently denote an EC by its agent set  $S$ , its concept set  $C$  or the couple  $(S, C)$ .

**Taxonomies and lattices** A relationship between agents and concepts is thus sufficient to capture the underlying epistemic communities of a given scientific field, yet we still need to hierarchize the raw set of all ECs to build a taxonomy. The canonical (Aristotelian) approach for

ordering categories consists of trees: categories are nodes, and sub-categories are child nodes of their unique parent category. In this case, we cannot deal with objects belonging to multiple categories or weak at representing paradigmatic categories. A straightforward way to improve a tree-based structure is a lattice-based structure, which allows *category overlap* representation (taxons may have more than one ascendant). To represent ECs hierarchically in a lattice-based taxonomy, we first provide a *partial order* " $\sqsubset$ " between ECs. An EC  $(S, S^\wedge)$  is a *subfield* of a field  $(S', S'^\wedge)$  if its intent is more precise than that of the field:  $(S, S^\wedge) \sqsubset (S', S'^\wedge) \Leftrightarrow S \subset S'$ . We can thus render both generalization and specification of closed couples (Wille, 1992), because  $(S, S^\wedge)$  can be seen as a specification of  $(S', S'^\wedge)$  (larger concept set, less agents) and conversely  $(S', S'^\wedge)$  is a "*superfield*" or a generalization of  $(S, S^\wedge)$ . Now, the *Galois lattice* is exactly the ordered set of ECs built from  $S, C$  and  $\mathcal{R}$ :

**Definition 1** (Galois lattice). *The Galois lattice  $\mathcal{G}_{S,C,\mathcal{R}}$  is the set of every closed couple  $(S, C) \subseteq \mathbf{S} \times \mathbf{C}$  under relation  $\mathcal{R}$ :  $\mathcal{G}_{S,C,\mathcal{R}} = \{(S^\wedge, S) \mid S \subseteq \mathbf{S}\}$ , partially ordered with  $\sqsubset$ .*

A graphical representation of a GL is drawn on Fig. 1.

## 2 Building and monitoring lattice-based taxonomies

### 2.1 Constructing a knowledge community taxonomy

Thus GLs, as hierarchically ordered structures, are both a categorization tool and a taxonomy building method. In this respect, GLs have been widely used in conceptual knowledge systems, formal concept classification, as well as mathematical social science (Wille, 1982; Freeman & White, 1993; Monjardet, 2003). But why would GLs be able to capture a *meaningful* structure in the case of knowledge communities? Here, we assume that a knowledge field can be divided into many subfields, themselves possibly having subcategories or belonging to various general fields, being respectively *multi-disciplinary* or *inter-disciplinary* (Klein, 1990). For instance, some scientists are linguists, and some among them

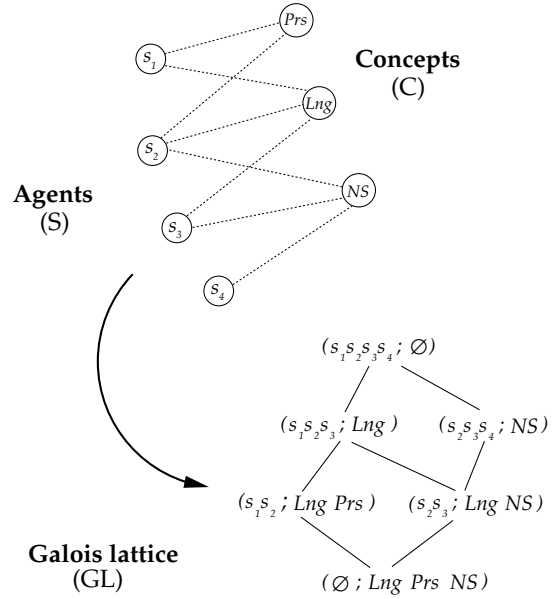


Figure 1: Example of binary relation with 4 agents and 3 concepts, *prosody* (Prs), *linguistics* (Lng) and *neuroscience* (NS) — on the right, the corresponding Galois lattice (6 ECs); lines indicate hierarchic relationships: from top (most general) to bottom (most specific); ECs are represented as a pair (extent, intent) =  $(S, C)$  with  $S^\wedge = C$  and  $C^* = S$ .

deal with a given aspect, say prosody; some deal with neuroscience, while a few of them are interdisciplinary and use both concepts. Since we also suppose that knowledge fields can be described by concept lists, and are being implemented by sets of agents, GL relevance as regards these alleged properties should result from the fact that (i) knowledge fields and corresponding agent sets are ECs, which are precisely what GLs consist of, (ii) GL natural partial order  $\sqsubset$  reflects a generalization/specialization relationship between fields and subfields, and also exhibits multidisciplinary and interdisciplinarity.

**Trimming the lattice** A serious caveat of GLs is that they may grow extremely large, with significantly more than several thousands of ECs, even for few agents and concepts. GLs contain all ECs, and among those many do not correspond to an existing or relevant field of knowledge: how to produce a *useful* and *usable* representation? Put differently, we wish to select and extract signifi-

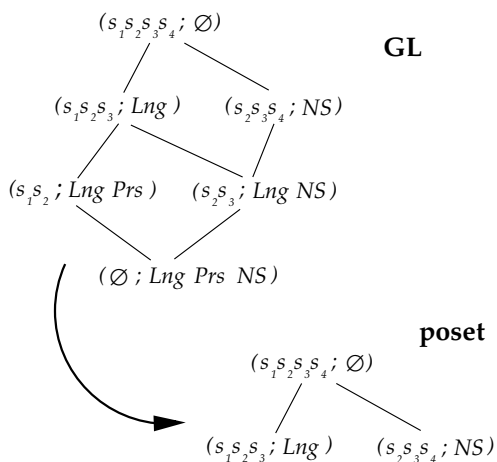


Figure 2: From the original GL to a selected partially-ordered set (poset).

cant ECs from a possibly huge GL, while excluding irrelevant ECs; in order to get closer to expert-based taxonomies. Formally, the set of extracted ECs is not a lattice anymore but is a partially-ordered set, or *poset*, that overlays on the lattice structure and still enjoys the taxonomical properties we are interested in. From now, we equivalently refer to a partial taxonomy or to a poset of ECs.

We need to design criteria to distinguish ECs relevant for a concise taxonomy description. This selection process has so far been underestimated in the study of GLs: an important part of the effort has focused on computation and representation (Godin *et al.*, 1998; Kuznetsov & Obiedkov, 2002), while few authors insist on the need for semantic interpretations and approximation theories to cope with GL combinatorial complexity (Stumme *et al.*, 2002; Duquenne *et al.*, 2003). At first, we would certainly keep most populated ECs: since fields tend to be made of larger groups of agents, and also because a GL mostly consists of small communities, agent set size is a segregating and efficient criterion, categorizing a large portion of the whole community (Roth & Bourguine, 2005). One should not pay attention to very small closed sets, for instance those of size one or two, which cannot be considered representative of a real EC. Yet, some large ECs are too specific, while small ECs close to the top are also likely to be relevant (as a minority field), and there is thus a pertinent

threshold for the size criterion.

Some other criteria might therefore apply as well: (i) “minority” ECs indeed, while being small, are unlikely to be subsets of other ECs and are more likely to be located in the surroundings of the lattice top; (ii) alternatively, they may be unusually specific with respect to their position in the lattice; (iii) finally, being outside the mainstream may make them less likely to mix with other ECs, thus having fewer descendants. Similarly, large yet noisy ECs may augment the GL uselessly. To keep small meaningful ECs and to exclude large insignificant ones, we may suggest the following criteria: (i) agent set size,  $|S|$ ; (ii) level (shortest distance to the top<sup>1</sup>),  $d$ ; (iii) specificity (concept set size),  $|C|$ ; (iv) sub-communities (number of descendants),  $n_d$ .

Then, we design several simple *selection heuristics* attributing a score to each EC by combining these criteria, so that we only keep the top scoring ECs: for instance, heuristics favoring (i) large ECs ( $|S|$ ), (ii) close to the top ( $\frac{|S|}{d}$ ), (iii) unusually specific ( $|S|\frac{|C|}{d}$ ), (iv) close to the top and having few descendants ( $\frac{|S|}{d \cdot n_d}$ ). We may not necessarily be able to express all preferences through a unique heuristic. Therefore, one heuristic could select large communities, while another is best suited for minority communities.<sup>2</sup> Depending on scientometric objectives, fine tuning and combining these functions eventually requires active feedback from empirical data. One could prefer to focus on taxonomies including only large, populated, dominant ECs, and accordingly consider the first heuristics only. Exploring further the adequacy and optimality of the choice and design of these heuristics would also be an interesting task — heuristics yielding e.g. a maximum number of agents for a minimal number of ECs — however unfortunately far beyond the scope of this paper. We will thus authoritatively keep and combine

<sup>1</sup>We take here the shortest length of all paths leading to the top EC ( $S, \emptyset$ ) (the whole community) — paths from a node to the top are not unique in a lattice.

<sup>2</sup>Notice that  $|S|$  remains a major criterion: a heuristic which ignores size could assign the same score, for example, to a very small EC with few descendants (like those at the lattice bottom) and to a larger EC with as many few descendants (possibly a worthy heterodox community). Besides, in general we need heuristics that keep the significant upper part of the lattice, so  $d$  is important as well.

these few heuristics to build a partial taxonomy from the original GL, as shown on Fig. 2. In any case, the agreement between taxonomies extracted using GLs and those explicitly given by domain experts will acknowledge the validity of this choice.

## 2.2 Taxonomy evolution

To monitor taxonomy evolution we follow the evolution of partial taxonomies. To this end, we create a series of posets from each GL corresponding to each period, and we capture some *patterns* reflecting epistemic evolution by comparing successive static pictures. In other words, we proceed to a longitudinal study of this series. Interesting patterns include:

- *progress or decline of a field*: a burst or a lack of interest in a given field;
- *enrichment or impoverishment of a field*: the reduction or the extension of the set of concepts related to a field;
- *reunion or scission of fields*: the merging of several existing fields into a more specific subfield or the scission of various fields previously mixed.

In terms of changes between successive posets, the first pattern simply translates into a variation in the population of a given EC: the agent set size increases or decreases (see Fig. 3–top-left). The second pattern reduces in fact to the same phenomenon. Indeed, suppose “*linguistics*” is enriched by “*prosody*”, i.e.  $\{Lng\}$  is enriched by  $\{Prs\}$ , thus becoming  $\{Lng, Prs\}$ . This means that the population of  $\{Lng, Prs\}$  is increasing. Since this EC is still a subfield of  $\{Lng\}$ , the enrichment of  $\{Lng\}$  by  $\{Prs\}$  translates into a subfield increase. Similarly, the decrease of  $\{Lng, Prs\}$  would indicate an impoverishment of the superfield  $\{Lng\}$ .<sup>3</sup> Finally, the union of various fields into an interdisciplinary subfield as

<sup>3</sup>More formally, say a field  $(S, C_1)$  is enriched by a concept  $c$ , “becoming”  $(S', C_1 \cup c)$ . This means that the subfield  $(S', C_1 \cup c)$  is increasing — as it is still a subfield of  $(S, C_1)$ , it is a subfield increase. In the limit case, when all agents working on  $C_1$  are also working on  $c$ , the superfield  $(S, C_1)$  becomes exactly  $(S, C_1 \cup c)$ . In all other cases, it is  $(S', C_1 \cup c)$ , a strictly smaller subfield of  $(S, C_1)$ , with  $S' \subset S$ . Conversely, if a field  $(S', C_1 \cup c)$  is to lose a specific concept  $c$ , the subcategory  $(S', C_1 \cup c)$  is going to decrease relatively to  $(S, C_1)$ .

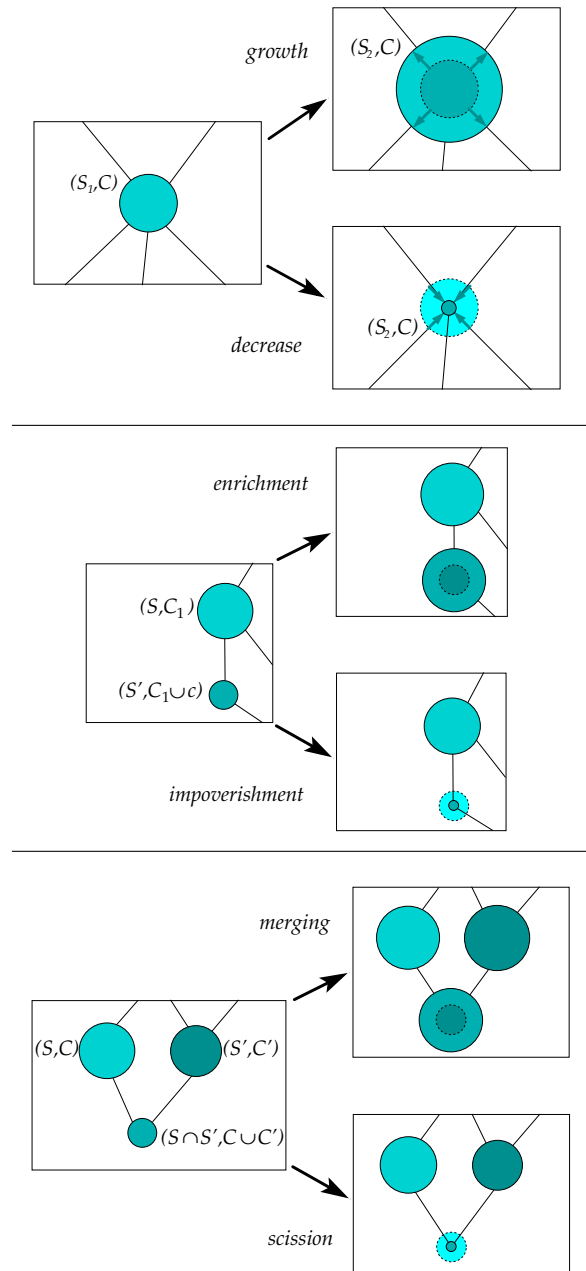


Figure 3: *Top*: progress or decline of a given EC  $(S_1, C)$ , whose agent set is growing (above) or decreasing (below) to  $S_2$ . *Middle*: enrichment or impoverishment of  $(S, C_1)$  by a concept  $c$ , through a population change of the subfield  $(S', C_1 \cup c)$ . *Bottom*: emergence or disappearance of a joint community (diamond bottom) based on two more general ECs,  $(S, C)$  and  $(S', C')$ . Disk sizes represent agent set sizes.

well as the scission of this interdisciplinary field comes in fact to an increase or a decrease of a joint subfield — geometrically, this means that a diamond bottom is emerging or disappearing (see Fig. 3–bottom). Obviously a merging (respectively a scission) is also an enrichment (resp. impoverishment) of each of the superfields.

Hence, each of these three kinds of patterns corresponds to a growth or a decrease in agent set size. The interpretation of the population change ultimately depends on the EC position in the partial taxonomy, and should vary according to whether (i) there is simply a change in population, (ii) the change occurs for a subfield and (iii) this subfield is in fact a joint subfield. These patterns, summarized on Fig. 3, describe epistemic evolution with an increasing precision. More precise patterns could naturally be proposed, but as we shall see, these ones are nevertheless sufficiently relevant for the purpose of our case study.

### 3 Case study

We now give an empirical protocol for this method and present our findings on a case study.

#### 3.1 Empirical protocol

To describe the community evolution over several periods of time, we need data telling us *when* an agent  $s$  uses a concept  $c$ . To this end, assuming articles to be a faithful account of what their authors deal with, we use a database of dated articles.

Accordingly, we divide the database into several time-slices, and build a series of relation matrices aggregating all events of each corresponding period. Before doing so, we need to specify the way we choose the *time-slice width* (size of a period), the *time-step* (increment of time between two periods) and how we *attribute a concept* to an agent, thus to an article.

**Time-slice width** We must choose a sufficiently wide time-slice in order to take into account minority communities (who publish less) and to get enough information for each author (especially those who publish in multiple fields).<sup>4</sup> Doing so

<sup>4</sup>For instance, extremely few authors publish more than one paper during a 6-month period, so obviously 6-month time-slices are not sufficient.

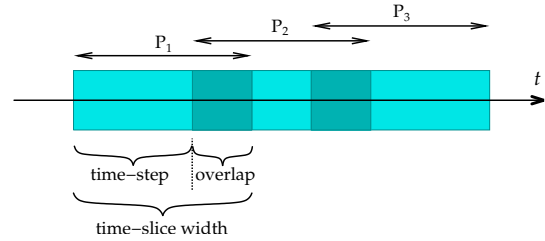


Figure 4: Series of overlapping periods  $P_1$ ,  $P_2$  and  $P_3$ .

also smoothes the data by reducing noise and singularities due to small sample sizes. However, when taking a longer sample size, we take the risk of merging several periods of evolution into a single time-slice. There is arguably a tradeoff between short but too insignificant time-slices, and long but too aggregating ones. This parameter must be empirically adapted to the data: depending on the case, it might be relevant to talk in terms of months, years or decades.

**Time-step** The time-step is the increment between two time-slices, so it defines the pace of observation. We need to consider overlapping time-slices, since we do not want to miss developments and events covering the end of a period and the beginning of the next one. Therefore, we have to choose a time-step strictly shorter than the time-slice width, as shown on Fig. 4. Moreover, the time-step is strongly related to the community *time-scale*: seeing almost no change between two periods would indicate that we are below this time-scale. We need to pick out a time-step such that successive periods exhibit sensible changes.

**Concept attribution** We attribute to each author the concepts he used in his articles. What kind of concepts should we extract from articles? At first, article keywords seem a good candidate, but they are actually very heterogenous data since authors often omit important keywords or choose randomly a keyword instead of another. Therefore, we consider *each word or nominal group as a concept*, and dismiss more complicated linguistic phenomena such as homonymy, polysemia or

synonymy.<sup>5</sup> We also proceed with title and abstract only, because complete contents are seldom available. While apparently and arguably rough (Leydesdorff, 1997), these minimal assumptions do not prevent us from building significant taxonomies.

### 3.2 Case and dataset description

We considered the community of embryologists working on the model animal “zebrafish”. This is a clearly defined group, with a decent size. We focused on publicly available bibliographic data from the MedLine database, covering the years 1990-2003. This timespan corresponds to a recent and important period of expansion for this community, which gathered approximately 1,000 agents at the end of 1995, and reached nearly 10,000 people by end-2003. We chose a time-slice width of 6 years, with a time-step of 4 years — that is, a 2 years overlap between two successive periods. We thus splitted the database in three periods: 1990-1995, 1994-1999 and 1998-2003.

To limit computation costs, we limited the dictionary to the 70 most used and significant words in the community, selected with the help of our expert. We also considered for each period a random sample of 255 authors. Besides, we used a fixed-size author sample so as to distinguish taxonomic evolutions from the trend of the whole community. Indeed, as the community was growing extremely fast, an EC could become more populated because of the community growth, while it was in fact becoming less attractive. With a fixed-sized sample, we could compare the relative importance of each field with respect to others within the evolving taxonomy.

### 3.3 Rebuilding history

Few changes occurred between the first and the second period, and between the second and the third period: the second period is a transitory period between the two extreme periods. This seems to indicate that a 4-year time-step is slightly below the time-scale of the community,

<sup>5</sup>More technically, we only consider words belonging to an expert-made selection of the most frequent words, excluding common and rhetorical words (empty words) as well as non-words (figures, percentages, dates, etc.). Additionally, we do not distinguish morphological variants such as plural, etc.

while 8 years can be considered a more significant time-scale.<sup>6</sup> We hence focus on two periods: the first one, 1990-1995, and the third one, 1998-2003. The two corresponding partial taxonomies are drawn on Fig. 6 (page 13). We observe that:

- *First period (1990-1995), first poset:*  $\{develop\}$  and  $\{pattern\}$  strongly structure the field: they are both large communities and present in many subfields. Then, slightly to the right of the poset, a large field is structured around *brain*<sup>7</sup> and *ventral* along with *dorsal*. Excepted one agent, the terms *spinal* and *cord* form a community with *brain*; this dependance suggests that the EC  $\{spinal, cord\}$  is necessarily linked to the study of *brain*. Subfields of  $\{brain\}$  also involve *ventral* and *dorsal*. Similarly,  $\{brain, ventral\}$  has a common subfield with  $\{spinal, cord\}$ . To the left, another set of ECs is structured around  $\{homologous\}$ ,  $\{mouse\}$  and  $\{vertebrate\}$ , and  $\{human\}$ , but significantly less.
- *Third period (1998-2003), second poset:* We still observe a strong structuration around  $\{develop\}$  and  $\{pattern\}$ , suggesting that the core topics of the field did not evolve. However, we notice the strong emergence of three communities,  $\{signal\}$ ,  $\{pathway\}$  and  $\{growth\}$ , and the appearance of a new EC,  $\{receptor\}$ . These communities form many joint subcommunities together, as we can see on the right of this lattice, indicating a convergence of interests. Also, there is a slight decrease of  $\{brain\}$ . More interestingly, there is no joint community anymore with  $\{ventral\}$  nor  $\{dorsal\}$ . The interest in  $\{spinal, cord\}$  has decreased too, in a larger proportion.

Finally,  $\{human\}$  has grown a lot, not  $\{mouse\}$ . These two communities are both linked to  $\{homologous\}$  on one side,  $\{vertebrate\}$  on the other. While the importance of  $\{homologous\}$  is roughly the same, the joint community with  $\{human\}$

<sup>6</sup>Kuhn (1970) asserts that old ideas die with old scientists — equivalently new ideas rise with new scientists. In this community, 8 years could represent the time required for a new generation of scientists to appear and define new topics; e.g. the time between an agent graduation and his first students graduation.

<sup>7</sup>We actually grouped *brain*, *nerve*, *neural* and *neuron* under this term.

has increased a lot. The same goes with *{vertebrate}*: this EC, which is almost stable in size, has a significantly increased role with *{mouse}* and especially *{human}* (a new EC *{vertebrate, human}* just appeared).

To summarize in terms of patterns: some communities were stable (e.g. *{pattern}*, *{develop}*, *{vertebrate, develop}*, *{homologous, mouse}*, etc.), some enjoyed a burst of interest (*{growth}*, *{signal}*, *{pathway}*, *{receptor}*, *{human}*) or suffered less interest (*{brain}* and *{spinal cord}*). Also, some ECs merged (*{signal}*, *{pathway}*, *{receptor}* and *{growth}* altogether; and *{human}* both with *{vertebrate}* and *{homologous}*), some splitted (*{ventral-dorsal}* separated from *{brain}*). We did not see any *strict* enrichment or impoverishment — even if, as we noted earlier, merging and splitting can be interpreted as such.

We can consequently suggest the following story: (i) research on brain and spinal cord depreciated, weakened their link with ventral/dorsal aspects (in particular the relationship between ventral aspects and the spinal cord), (ii) the community started to enquire relationships between signal, pathway, and receptors (all actually related to biochemical messaging), together with growth (suggesting a messaging oriented towards growth processes), indicating new very interrelated concepts prototypical of an emerging field, and finally (iii) while mouse-related research is stable, there has been a significant stress on human-related topics, together with a new relationship to the study of homologous genes and vertebrates, underlining the increasing role of *{human}* in these differential studies and their growing focus on human-zebrafish comparisons (leading to a new “interdisciplinary” field). Point (ii) entails more than the mere emergence of numerous joint subcommunities: all pairs of concepts in the set *{growth, pathway, receptor, signal}* are involved in a joint subfield. Put differently these concepts form a clique of joint communities, a pattern which may be interpreted as *paradigm emergence* (see Fig. 6–bottom).

**Comparison with real taxonomies** We compared these findings with empirical taxonomical data, coming both from:

1. Expert feedback: Our expert, Nadine Peyri ras, confirms points (i), (ii) and (iii)

in the previous paragraph. For instance, according to her, the human genome sequencing in the early 2000s (International Human Genome Sequencing Consortium, 2001) opened the path to zebrafish genome sequencing, enabling systematic comparison between zebrafish and human, and consequently led to the development described in point (iii). In addition, the existence of a subcommunity with *brain, spinal cord* and *ventral* but not *dorsal* reminded her the initial curiosity around the ventral aspects of the spinal cord study, due to the linking of the ventral spinal cord to the mesoderm (notochord), i.e. the rest of the body.

2. Literature: The only article yet dealing specifically with the history of this field is that of Grunwald & Eisen (2002). This paper presents a detailed chronology of the major breakthroughs and steps of the field, from the early beginnings in the late 1960s up to 2002. While it is hard to infer a comprehensive taxonomic evolution, part of their investigation confirms some of our most salient patterns: “Late 1990s to early 2000s: Mutations are cloned and several genes that affect common processes are woven into molecular pathways” — here, point (ii). Note that some other papers address and underline specific concerns of the third period, such as the development of comparative studies (Dooley & Zon, 2000; Bradbury, 2004).
3. Conference proceedings: Finally, some insight could be gained from analyzing the evolution of the session breakdown for the major conference of this community, “Zebrafish Development & Genetics” (Cold Spring Harbor Laboratory, 1994). Contributions reflect the current community interests, yet it may be uneasy for organizers to label sessions with a faithful and comprehensive name — “organogenesis” for instance covers many diverse subjects (it works like a keyword). Reviewing the proceedings roughly suggests that comparative and sequencing-related studies are an emerging novelty starting in 1998, at the beginning of the third period, which agrees with our analysis. On the contrary, the importance of issues related to the brain & the nervous system, as well as

signaling, seem to be constant between the first and the third period, which diverges from our conclusions.

The expert feedback here is obviously the most detailed and valuable feedback — other sources of empirical validation are more subject to interpretation and therefore more questionable. A more comprehensive empirical protocol would consist in a larger set of experts providing more details as well as a more intersubjective viewpoint, thus objective.

## 4 Discussion

In the first place, this method can be very helpful to historians of science, in domains where historical studies are lacking — notably when examining the recent past. Papers such as the history of the “zebrafish” community written by scientists originating from this very community (Grunwald & Eisen, 2002) could profit from such non-subjective analysis. In this particular case the present article might be considered the second historical study of this community. Also, using this method is possible in at least any practical case involving a relationship between agents and semantic items. As stated by Cohendet *et al.* (2003), “*a representation of the organization as a community of communities, through a system of collective beliefs (...), makes it possible to understand how a global order (organization) emerges from diverging interests (individuals and communities).*”<sup>8</sup> Note that one must distinguish (i) scientific field taxonomy building through GL extraction; and (ii) scientific field dynamics mapping, through dynamic patterns detailed in Sec. 2.2. In our opinion, it is the interplay of these two features that makes the whole method relevant for science dynamics monitoring. In addition to epistemology, scientometrics and sociology, other fields of application include economics (start-ups dealing with technologies, through contracts), linguistics (words and their context, through co-appearance within a corpus), marketing (companies dealing with ethical values, through customers cross-preferences), and history in general (e.g. evolu-

<sup>8</sup>“*Une représentation de l’organisation comme une communauté des communautés, à travers un système de croyances collectives (...), permet (...) de comprendre comment émerge un ordre global (organisation) à partir d’intérêts divergents (individus et communautés).*”

tion of industrial patterns linked to urban centers (White & Spufford, 2006)).

**Comparison with existing approaches** Existing approaches in knowledge community mapping and monitoring are often focused on co-word and co-author analysis — see Callon *et al.* (1986); Persson & Beckmann (1995); Noyons & van Raan (1998), *inter alia*, as well as an extensive review made by Archambault & Gagne (2004). First, methods relying only on single networks of social relationships (e.g. co-authorship) or word co-occurrence networks may be unreliable for finding epistemic communities: to cite a few problems, such communities are not necessarily socially bound, and dyadic relationships between words are blind to larger groups of structurally equivalent words (Leydesdorff, 1991). On the whole, one-mode data — be it the projection of two-mode data or not — entails a loss of crucial structural information (see Fig. 5). Quite to the contrary, the duality of the reciprocal linkage of agents to concepts and the corresponding symmetry between agent-based and concept-based notions is well rendered by a GL — ECs are referring to a structurally equivalent usage of terms, independent on relations between terms themselves.

On the other hand, several methods avoid the kind of loss of information induced by one-mode data. Co-citations for example can be employed to create categories based on similar patterns of citations (McCain, 1986; Kreuzman, 2001; McCain *et al.*, 2005), and as such to some extent make use of duality (it should even be noted that the present GL-based method could be tried on authors linked to citations, instead of concepts). However, and more broadly, most studies do not depart from a dendrogram-based representation. Yet, a dendrogram is a cluster tree, and ascendancies cannot be multiple: a community is bound to be embedded into a lineage of increasing communities, and an agent cannot be part of many non-embedded, overlapping ECs. Abandoning dendrograms is thus crucial to deal with inter-disciplinarity, which in turn is fundamental to further understand why disciplines merge or cross-fertilize — to our knowledge, works dealing with overlapping categories are very rare.

Finally, GLs make it easy to define meaningful dynamic patterns on *interactions* between differ-

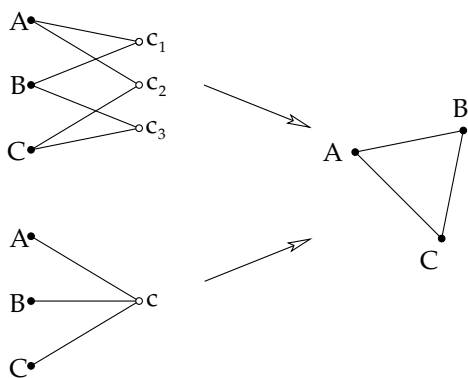


Figure 5: Two significantly different two-mode datasets yield an identical one-mode projection (*right*), when linking pairs of agents sharing at least one concept. In the projection there is no way to distinguish a triangle of authors  $s_1, s_2, s_3$  sharing the same concept  $c$ , from a triangle of authors linked by pairs to distinct concepts  $c_1, c_2, c_3$ . This feature is however key, and the argument is identical when inverting agents and concepts.

ent epistemic communities. Of course, it could also be useful to characterize and identify the actors who are driving the paradigm change — although unaddressed here, monitoring the evolution of particular groups of agents would be simple.<sup>9</sup> Additionally, it is painless to develop the partial taxonomy on-the-fly and in an interactive fashion, in order to have further details on a particular sub-community. Unlike various community mapping techniques, labelling here would still be straightforward, as agent groups are automatically bound to a semantic content.

**Lattice manipulation** On the other hand, our method could enjoy several improvements. Primarily, computing the whole GL then extracting a partial taxonomy is certainly not the most ef-

<sup>9</sup>Indeed, we ignore whether individual agents have fixed roles or not: the stability of the size of an EC does not imply the stability of its agent set. Fortunately, even if random agent samples are not consistent across periods, rebuilding the whole community taxonomy by filling the partial ECs with their corresponding full agent sets is effortless. In this case, field scope enrichment or impoverishment could be described in a better way: by monitoring an identical agent set, and by watching whether its intent increases or not.

ficient option. Computing only the lattice part most likely to contain basic-level taxa could perform better — using e.g. an algorithm computing the “valuable” upper part only. Similarly, selection heuristics must allow for significant child nodes to appear. Indeed, when two fields do not seem to form a joint subfield in the partial taxonomy, it is hard to know whether they actually form a joint subfield, yet are below the threshold. In the second poset for instance the EC  $\{\textit{brain}, \textit{spinal cord}\}$ , although of similar importance as  $\{\textit{spinal cord}\}$  (17 vs. 18 agents), is excluded by the selection threshold and does not appear; possibly leading us to wrongly deduce that  $\{\textit{brain}\}$  does not mix at all with  $\{\textit{spinal cord}\}$ . Also, considering that some authors are more or less strongly related to some concepts, a binary relationship may seem too restrictive: here, we could use a weighted relation matrix together with fuzzy GLs (Belohlavek, 2000). In the same direction, we could endeavor to exclude false positives and merge clusters of ECs into single multidisciplinary ECs (like for instance “*signal*”, “*pathway*”, “*receptor*”). Questions arise however regarding the best way to define a cluster of ECs without destroying overlapping communities, one of the most interesting feature of GLs. Accordingly, it could also be profitable to disambiguate and regroup terms using Natural Language Processing (NLP) tools (Ide & Véronis, 1998): certainly not everyone assigns the same meaning to “*pattern*”; we would thus have to introduce “*pattern-1*”, “*pattern-2*”, etc.

More broadly, we could consider the lattice *dynamically*, instead of adopting a *longitudinal* approach: this would yield a better representation of field evolution at smaller scales; transforming the empirical discussion about the right time-step into a methodological question: how to detect low-level dynamic features which correspond to an epistemic or paradigmatic change?

## Conclusion

We presented a method for building a manageable taxonomy, and describing its evolution. We focused on the structure of epistemic communities, and introduced a formal framework based on Galois lattices to categorize ECs in an automated and hierarchically structured way. Since the resulting lattice is often unwieldy, we pro-

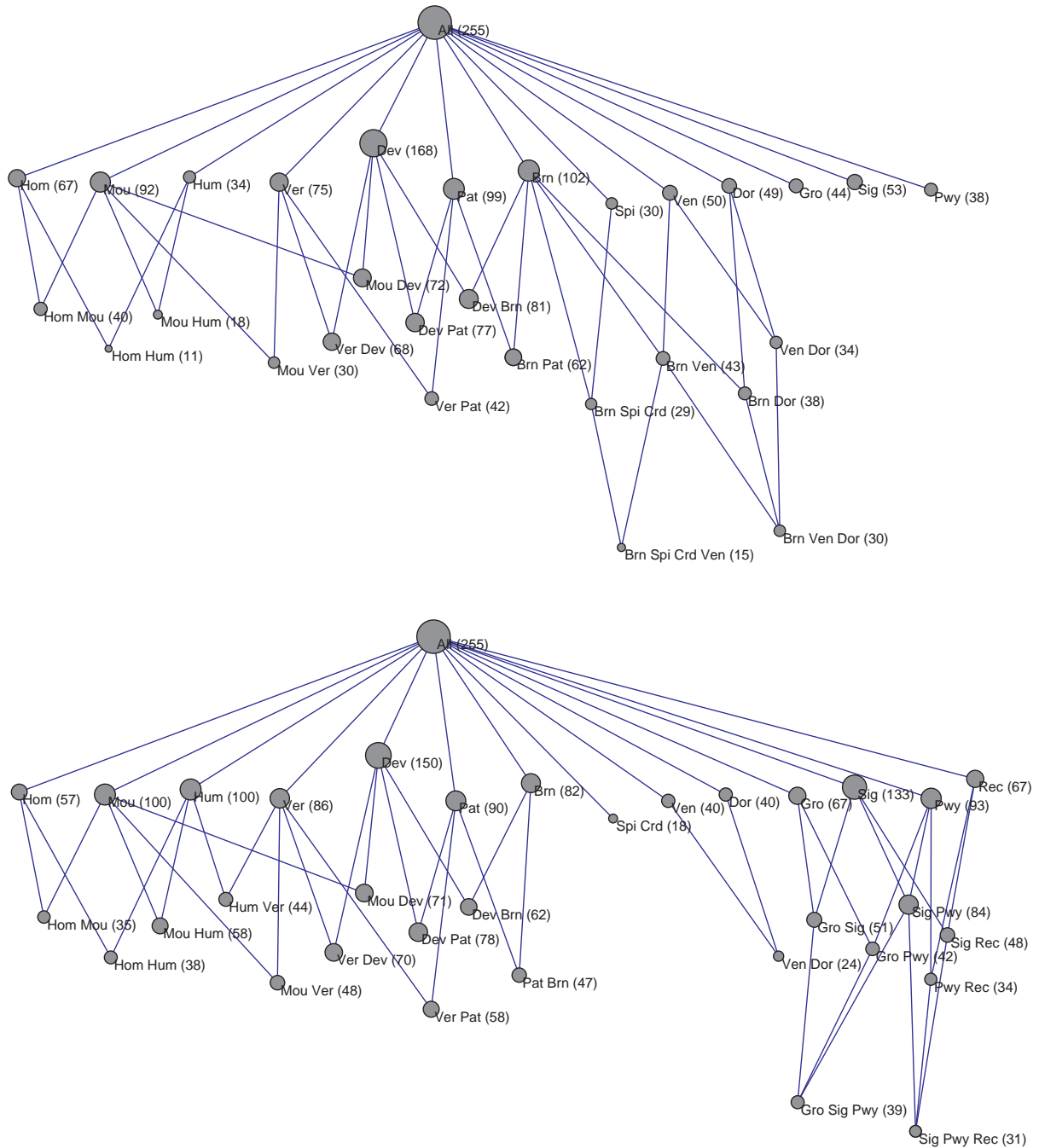
posed selection criteria for building a partial taxonomy under the form of a poset gathering the most relevant ECs and overlaying the original lattice-based structure. Consequently, the longitudinal study of such posets made possible an historical description. In particular, we proposed to capture stylized facts related to epistemic evolution such as field progress, decline and interaction (merging or splitting). We ultimately applied our method to the subcommunity of embryologists working on the “zebrafish” between 1990 and 2003, and successfully compared the results with taxonomies given by domain experts. Yet, this method can also be easily improved and fruitfully ported to other domains.

**Acknowledgements** The authors wish to express the warmest thanks to Nadine Peyri eras for the numerous and essential discussions and advices. We are also grateful of useful comments made by Sergei Obiedkov, Douglas White, an anonymous reviewer and participants of the Sunbelt XXV International Conference on Social Networks. This work has been partly funded by the CNRS.

## References

- E. Archambault and E. V. Gagne (2004). *L'utilisation de la bibliom trie dans les sciences sociales et les humanit s*. Tech. rept. Conseil de recherches en sciences humaines du Canada.
- M. Barbut and B. Monjardet (1970). *Alg bre et combinatoire*. Vol. II. Paris: Hachette.
- R. Belohlavek (2000). Fuzzy Galois connections and fuzzy concept lattices: From binary relations to conceptual structures. *Pages 462–494 of: V. Novak and I. Perfilova (eds), Discovering the world with fuzzy logic*. Heidelberg: Physica-Verlag.
- B. Berlin (1992). *Ethnobiological classification - principles of categorization of plants and animals in traditional societies*. Princeton: Princeton University Press.
- G. Birkhoff (1948). *Lattice theory*. Providence, RI: American Mathematical Society.
- J. Bradbury (2004). Small fish, big science. *PLoS Biology*, 2(5), 568–572.
- M. Callon, J. Law, and A. Rip (1986). *Mapping the dynamics of science and technology*. London: MacMillan Press.
- P. Cohendet, A. Kirman, and J.-B. Zimmermann (2003). Emergence, formation et dynamique des r seaux – mod les de la morphog nese. *Revue d'Economie Industrielle*, 103(2-3), 15–42.
- Cold Spring Harbor Laboratory (1994, 1996, 1998, 2000, 2001, 2002, 2003). *Zebrafish development & genetics*. Cold Spring Harbor, NY.
- R. Cowan, P. A. David, and D. Foray (2000). The explicit economics of knowledge codification and tacitness. *Industrial & corporate change*, 9(2), 212–253.
- K. Dooley and L. I. Zon (2000). Zebrafish: a model system for the study of human disease. *Current opinion in genetics & development*, 10(3), 252–256.
- P. Doreian, V. Bategelj, and A. Ferligoj (2005). *Generalized blockmodelling*. Cambridge: Cambridge University Press.
- O. Dupouet, P. Cohendet, and F. Creplet (2001). *Economics with heterogenous agents*. Berlin: Springer. Chap. Organisational innovation, communities of practice and epistemic communities: the case of Linux.
- V. Duquenne, C. Chabert, A. Cherfouh, A.-L. Doyen, J.-M. Delabar, and D. Pickering (2003). Structuration of phenotypes and genotypes through Galois lattices and implications. *Applied artificial intelligence*, 17(3), 243–256.
- L. C. Freeman and D. R. White (1993). Using Galois lattices to represent network data. *Sociological methodology*, 23, 127–146.
- R. Godin, H. Mili, G. W. Mineau, R. Missaoui, A. Arfi, and T.-T. Chau (1998). Design of class hierarchies based on concept (Galois) lattices. *Theory and practice of object systems (TAPOS)*, 4(2), 117–134.
- D. J. Grunwald and J. S. Eisen (2002). Headwaters of the zebrafish – emergence of a new model vertebrate. *Nature rev. genetics*, 3(9), 717–724.
- P. Haas (1992). Introduction: epistemic communities and international policy coordination. *International organization*, 46(1), 1–35.
- N. Ide and J. V ronis (1998). Word sense disambiguation: The state of the art. *Computational linguistics*, 24(1), 1–40.
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
- A. K. Jain, M. N. Murty, and P. J. Flynn (1999). Data clustering: a review. *ACM computing surveys*, 31(3), 264–323.

- S. C. Johnson (1967). Hierarchical clustering schemes. *Psychometrika*, **2**, 241–254.
- J. T. Klein (1990). *Interdisciplinarity: History, theory, and practice*. Detroit, MI: Wayne State University Press.
- T. Kohonen (2000). *Self-organizing maps*. 3rd edn. Berlin: Springer.
- H. Kreuzman (2001). A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science. *Scientometrics*, **51**(3), 525–539.
- T. S. Kuhn (1970). *The structure of scientific revolutions*. 2nd edn. Chicago, IL: University of Chicago Press.
- S. O. Kuznetsov and S. A. Obiedkov (2002). Comparing performance of algorithms for generating concept lattices. *Journal of experimental and theoretical artificial intelligence*, **14**(2-3), 189–216.
- L. Leydesdorff (1991). In search of epistemic networks. *Social studies of science*, **21**, 75–110.
- L. Leydesdorff (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American society for information science*, **48**(5), 418–427.
- A. Lopez, S. Atran, J. D. Coley, D. L. Medin, and E. E. Smith (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive psychology*, **32**(3), 251–295.
- F. Lorrain and H. C. White (1971). Structural equivalence of individuals in social networks. *Journal of mathematical sociology*, **1**(49–80).
- K. W. McCain, J. M. Verner, G. W. Hislop, W. Evanco, and V. Cole (2005). The use of bibliometric and Knowledge Elicitation techniques to map a knowledge domain: Software Engineering in the 1990s. *Scientometrics*, **65**(1), 131–144.
- K. W. McCain (1986). Cocited author mapping as a valid representation of intellectual structure. *Journal of the American society for information science*, **37**(3), 111–122.
- B. Monjardet (2003). The presence of lattice theory in discrete problems of mathematical social sciences. Why. *Mathematical social sciences*, **46**(2), 103–144.
- M. E. J. Newman (2004). Detecting community structure in networks. *European physical journal B*, **38**, 321–330.
- E. C. M. Noyons and A. F. J. van Raan (1998). Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. *Journal of the American society for information science*, **49**(1), 68–81.
- O. Persson and M. Beckmann (1995). Locating the network of interacting authors in scientific specialties. *Scientometrics*, **33**(3), 351–366.
- E. Rosch and B. Lloyd (1978). Cognition and categorization. *American psychologist*, **44**(12), 1468–1481.
- C. Roth and P. Bourgin (2005). Epistemic communities: Description and hierarchic categorization. *Mathematical population studies*, **12**(2), 107–130.
- G. Salton, A. Wong, and C. S. Yang (1975). Vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.
- R. R. Sokal and P. H. A. Sneath (1963). *Principles of numerical taxonomy*. San Francisco, CA: W.H. Freeman.
- G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal (2002). Computing iceberg concept lattices with TITANIC. *Data and knowledge engineering*, **42**, 189–222.
- S. Wasserman and K. Faust (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
- D. R. White and P. Spufford (2006). *Medieval to modern: Civilizations as dynamic networks*. Book Ms.
- H. C. White, S. A. Boorman, and R. L. Breiger (1976). Social-structure from multiple networks. I: Block-models of roles and positions. *American journal of sociology*, **81**, 730–780.
- R. H. Whittaker (1969). New concepts of kingdoms of organisms. *Science*, **163**, 150–160.
- R. Wille (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. *Pages 445–470 of: I. Rival (ed), Ordered sets*. Dordrecht-Boston: Reidel.
- R. Wille (1992). Concept lattices and conceptual knowledge systems. *Computers mathematics and applications*, **23**, 493.



Legend: All: the whole community, Hom: *homologue/homologous*, Mou: *mouse*, Hum: *human*, Ver: *vertebrate*, Dev: *development*, Pat: *pattern*, Brn: *brain/neural/nervous/neuron*, Spi: *spinal*, Crd: *cord*, Ven: *ventral*, Dor: *dorsal*, Gro: *growth*, Sig: *signal*, Pwy: *pathway*, Rec: *receptor*.

Figure 6: Two partial taxonomies of the zebrafish community at the end of 1995 (*top*) and at the end of 2003 (*bottom*). Figures in parentheses indicate the number of agents per EC. Lattices established from a sample of 255 agents (out of 1, 000 for the first period vs. 9, 700 for the third one).