Social and Semantic Coevolution in Knowledge Networks

Camille Roth[*, †] & Jean-Philippe Cointet[‡, †]
roth@ehess.fr, jean-philippe.cointet@polytechnique.edu

Published by Elsevier in **Social Networks**, *32(1):16-29, 2010.*
Publisher version available at `http://dx.doi.org/10.1016/j.socnet.2009.04.005`

Abstract

Socio-semantic networks involve agents creating and processing information: communities of scientists, software developers, wiki contributors and webloggers are, among others, examples of such knowledge networks. We aim at demonstrating that the dynamics of these communities can be adequately described as the coevolution of a social and a socio-semantic network. More precisely, we will first introduce a theoretical framework based on a social network and a socio-semantic network, i.e. an epistemic network featuring agents, concepts and links between agents and between agents and concepts. Adopting a relevant empirical protocol, we will then describe the joint dynamics of social and socio-semantic structures, at both macroscopic and microscopic scales, emphasizing the remarkable stability of these macroscopic properties in spite of a vivid local, agent-based network dynamics.

Keywords: knowledge networks, socio-semantic networks, hierarchies, dynamics, cohesiveness, epistemic communities, bicliques, Galois lattices, blogs, scientific networks.

1 Introduction

Socio-semantic networks involve agents who produce, manipulate and exchange knowledge or information: communities of scientists, free software developers, "wiki" contributors and webloggers are such instances, among others, of groups of distributed knowledge creation and processing — or knowledge communities. We aim here at understanding the *morphogenesis* of these networks, and particularly what is proper to knowledge networks. This, in turn, would be likely to provide a specific insight and perspective on several underlying social and cognitive processes, such as cultural epidemiology phenomena (Valente, 1995; Sperber, 1996), consensus, relevance and authority judgments (Bourdieu, 1991; Lazega, 1992), or even the emergence of various types of stratification (Cole and Cole, 1973; Freeman, 1989; Cohendet et al., 2003).

In this respect, interactions occurring in such socio-semantic complex systems are determined, at least partially, by the structure of past interactions and by conceptual affinities. Simultaneously, new interactions shape and modify both the social network structure and

[*]CAMS (CNRS/EHESS, France) & CRESS (U. Surrey, UK), 54 bd Raspail, F-75006 Paris, France.
[†]ISC-PIF (Institut des Systèmes Complexes de Paris-Île-de-France), 57-59 rue Lhomond, F-75005 Paris, France.
[‡]CREA (CNRS/Ecole Polytechnique, France) & INRA-SenS (INRA, France), Bois de l'Etang - 5, Bd Descartes, F-77454 Marne-la-Vallée.

the distribution of semantic characteristics and interests within the network, predictably influencing future interactions (Emirbayer and Goodwin, 1994). More precisely, taking into account semantic aspects in interactional processus becomes especially relevant when knowledge and relationships evolve at similar timescales — i.e., when semantic features are likely to coevolve with their social environment (Leenders, 1997).

We focus here on two particular kinds of knowledge networks: scientific collaboration networks and blogger citation networks, which are both settings where link creation between agents and use of semantic items are plausibly occurring jointly, at a similar pace. Our study will be empirically based on moderately sizable and well-delimited networks: on one hand, a community of several thousands of embryologists studying an animal model, the "*zebrafish*", and, on the other hand, a network of about a thousand webloggers posting articles regarding the US presidential elections of 2008.

Beyond the numerous behavioral studies devoted to processes of link formation in such knowledge networks (Latour and Woolgar, 1988; Katz and Martin, 1997; Cardon and Delaunay-Teterel, 2006), *large-scale structural studies* on this kind of communities have essentially focused on citations (de Solla Price, 1965, 1976; McGlohon et al., 2007) and interactions (Newman, 2004; Ali-Hasan and Adamic, 2007). While the social structure has generally been the cornerstone of these previous works, its potential intertwinement with semantic features has mostly remained unaddressed (Pattison, 1994).

This paper aims at demonstrating that the dynamics of these knowledge communities can be adequately described as the coevolution of a social and a socio-semantic network. More precisely, we will first introduce a theoretical framework based on a social network and a socio-semantic network, i.e. an epistemic network featuring agents, concepts and links between agents and between agents and concepts (Sec. 2). Adopting a corresponding empirical protocol (Sec. 3), we will then describe the coevolution of social and socio-semantic structures, at both macroscopic and microscopic scales, emphasizing the remarkable stability of these macroscopic properties in spite of a vivid local, agent-based network dynamics. We will in particular exhibit several relevant structural patterns and properties:

(i) in terms of hierarchies (Sec. 4), a strong heterogeneity in the connectivity or usage, respectively, of certain agents or concepts, along with significant dependencies between social and semantic aspects of these hierarchies;

(ii) in terms of aggregates (Sec. 5), a strong social cohesion (transitivity), echoed by a socio-semantic homogeneity, present at both the local level (conceptual resemblance within the social neighborhood) and the global level (presence of large groups of agents manipulating identical concepts, traditionally denoting "epistemic communities").

Eventually, empirical estimations of the non-uniform link creation processes, including semantic homophily, will constitute plausible underpinnings of the observed structures and of their stability. The paper will essentially be organized in a dual manner, progressively introducing static and dynamic observables linked to social networks only (which are oftentimes classical) altogether with their socio-semantic counterparts.

2 Epistemic networks

The distinctiveness of epistemic networks. Although graph theory indifferently applies

to any kind of network –social or not– it may appear debatable to consider that mechanisms proper to social networks can be likened to those operating in other classes of networks, beyond some universal phenomena whose reach could seem relatively limited.[1]

An identical argument could be proposed in the case of knowledge networks: social and semantic networks are often studied separately. Social network analysis indeed rarely focuses, practically, on the relationships between social structures and semantic configurations (Emirbayer and Goodwin, 1994) while the structure of the social network constitutes, if not the sole structure, at least the reference frame. Yet, knowledge networks feature interaction where semantic content is decisive, thus underlining the relevance of considering structural patterns or interaction mechanisms which are not strictly social (Callon, 2001), i.e. only based on inter-agent relationships.

We therefore aim at providing a theoretical framework binding both networks, by suggesting that the analysis of knowledge communities and the underlying agent behaviors (notably relational behaviors) must take into account the reciprocal and joint influence of both social and semantic features. The rationale behind this approach is double-minded: introducing semantic aspects to the traditional social network ontology makes it possible to both (i) precisely characterize the structure of knowledge communities and (ii) understand what defines and determines interactions depending on semantic objects, i.e. consider simultaneously phenomena of selection and influence (Leenders, 1997; Robins et al., 2001b,a) through the *co-evolution* of social and semantic configurations. This will also require, as we shall see, the introduction of descriptions proper to this object or, in other terms, propose a class of "epistemic" patterns rather than just social patterns. In particular, we will for instance assume that the selection of similar agents —homophily— plays a significant role in socio-semantic pattern formations (McPherson et al., 2001). This argument has rarely been involved in a coevolving modeling framework. In our empirical setting, we will accordingly endeavor at understanding collaborations among scientists sharing similar concerns, or link formation between bloggers dealing with identical topics.

Formal framework. We first distinguish the *social network*, whose nodes are agents and links indicate observed relationships. Relationships may a priori refer indifferently either (i) to *interactions* (for instance, a scientist collaborates with another scientist or an individual comments a post on a blog) or (ii) to *authority attributions* (e.g. a scientist cites the works of another scientist, a blogger cites a post of another blogger).

In any case, the social network is denoted by $\mathbf{G} = (\mathbf{S}, \mathcal{R}^{\mathbf{S}})$ where $\mathbf{S}$ is the agent set and $\mathcal{R}^{\mathbf{S}} = \mathcal{R} \subset \mathbf{S} \times \mathbf{S} \times \mathbb{N}$ denotes the set of dated links: a link $l = (s, s', t) \in \mathcal{R}^{\mathbf{S}}$ means that s is related to s' at t. Links can be directed (s cites s') or non-directed, as is the case in the scientific collaboration network where, if s interacts with s', then s' interacts with s, indifferently (in which case $(s, s', t) \in \mathcal{R}^{\mathbf{S}} \Leftrightarrow (s', s, t) \in \mathcal{R}^{\mathbf{S}}$).

We then introduce semantic objects which we call "concepts" and which correspond here to terms or noun phrases considered as atomic units — $\mathbf{C}$ denotes the concept set. This enables us to define a second network binding agents and concepts: the *socio-semantic* network $\mathbf{G}^{\mathbf{C}}$, made of agents of $\mathbf{S}$, concepts of $\mathbf{C}$ and links between these elements: $\mathcal{R}^{\mathbf{C}}$ thus denotes

[1] This constrasts with a not infrequent opinion that network study may be universal across all disciplines (for a review, see Dorogovtsev and Mendes, 2003, *inter alia)* and underlining universal phenomena focused on the global connectivity structure, which several authors have nevertheless demonstrated to be sensibly diverse across both social and non-social networks (e.g. **?**).

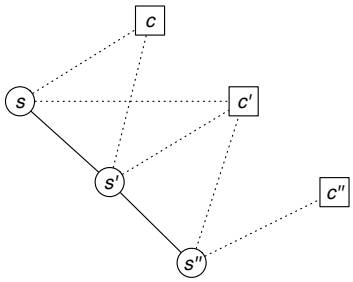

Figure 1: Illustration of an epistemic network made of three agents $\mathbf{S} = \{s, s', s''\}$, three concepts $\mathbf{C} = \{c, c', c''\}$ and two kinds of links: between agents $\mathcal{R}^{\mathbf{S}}$ (straight lines), and between agents and concepts $\mathcal{R}^{\mathbf{C}}$ (dashed lines).

the use of concepts by agents: an agent is linked to concepts he mentioned (in a paper, in a post). Thus, $\mathcal{R}^{\mathbf{C}} \subset \mathbf{S} \times \mathbf{C} \times \mathbb{N}$, and a link $l^{\mathbf{C}} = (s, c, t) \in \mathcal{R}^{\mathbf{C}}$ means that s used c at t.

Note that both networks thus correspond to two distinct ontologies, because social relations and cognitive properties, in the broad sense, admittedly refer to two different kinds of settings, even if they are constructed from a common source (papers and posts). While the social network is indeed appraised as a monopartite graph, as it is essentially a network of social relations *between agents*, the socio-semantic network is a bipartite graph, as it is essentially a straightforward representation of *agent affiliations* to attributes, as is obvious from Fig. 1.

3 Empirical protocol and methods

3.1 Case description

An *epistemic network* is thus defined by these two networks — or, rather, these two kinds of nodes and two kinds of links — and will constitute the cornerstone of the representation of the coevolution of agents & concepts. Without restraining the generality of this framework, we empirically focus on one hand on an interaction network, made of research collaborations involving scientific concepts, and on the other hand on a citation network consisting of bloggers producing posts and citing other bloggers' posts.

Scientific networks: the *zebrafish* community

We first focus on a scientific network, which corresponds to embryologists working on the zebrafish, or "*brachydanio rerio*". The zebrafish is a small fish which has quickly become a model animal due to its exceptional biological characteristics, including very rapid and translucid development (Bradbury, 2004). For this reason, the community has known an exponential growth during the last 20 years. Our dataset is more precisely built from papers published from 1999 to the end of 2006 in one of the $20\,000$ journals indexed by the US *Library of Congress* — we therefore collected the data from the free and public bibliographic platform *Medline* provided by the US *National Library of Medicine*. Over the period, our database includes $6,641$ papers featuring $15,204$ authors, increasing the number of overall authors and

published articles in the field by about 5 times (from ca. $1,500$ papers to over $8,000$, and from $3,000$ authors to about $18,000$).

The corresponding epistemic network is a collaboration network, where published papers are collaboration events. Authors are thus the actors of the social network, while concepts refer to topics used in papers. The bibliographic data mentions *which agents* collaborated on *which topic* at *which time*, therefore describing all dated social and socio-semantic links.

Blogger networks: a portion of the US political blogosphere

The blog dataset is built upon the content of posts published by a selection of $1,066$ political blogs in the context of the US presidential elections of 2008. Data was collected from November 1, 2007 to February 29, 2008 by LINKFLUENCE[2] and essentially consists of blog entries basically made of a title, a full text content, associated with a list of hyperlinks.

As such, the corresponding epistemic network is fundamentally a citation network. We consider that each hyperlink found in the content of an entry is deemed as a citation of the blog corresponding to this hyperlink. Similarly to the scientific network, post authors are the nodes of the social network and a selected set of topics makes the concept set. Besides, although this network exhibits very high rates of links creation (see Fig. 2), most blogs from the dataset are active from the start and over all periods, contrarily to the scientific network.

3.2 Data processing

Agents are uniquely identified from the data by their original name in the zebrafish case and by their unique blog URL in the blog network. The social network is thus easily created by taking into account joint collaborations between two scientists or citation of a blogger by another blogger.

Delineating concepts in the raw semantic data, which is made of abstracts for scientists and entries for blogs, is a more delicate process. Adopting a straightforward approach based on the retrieval of keywords and tags specified by authors themselves could be questionable: first, this data is seldom available and, second, when it is, it depends on very subjective and individual taxonomies. We regard the use of article contents as safer, following the hypotheses of the distributional program in NLP (Jones and Kay, 1973) where the atomic semantic unit is a term or a noun phrase (such as "*brain*", "*spinal cord*") — which is commonplace for instance in scientometrics (Callon et al., 1986).

We therefore simply achieve a basic correspondence between *lemmatized terms* and *concepts*, i.e., we apply to both datasets the following simple linguistic processing. We first gather all *terms* and lemmatize them, i.e. we build classes of terms sharing the same base form, or *lemma*; thereby aggregating the various forms of an identical term in a single class. We then exclude meaningless words (or "stop-words") with respect to the context, such as "*convincing*", "*example*", "*now*", etc. With the help of external experts on each dataset, we eventually consider a relatively small selection of discriminating and distinct concepts,

[2]http://linkfluence.net — LINKFLUENCE is "a research institute specializing in the conversations of the social web". Originally, this data was being used to feed *Presidential Watch '08* (http://presidentialwatch08.com), a monitoring system for the blogosphere focused on the upcoming elections.

among the most frequent in the database. This first list of lemmas, later called "concepts", remains unchanged for the whole analysis (we concretely defined 65 concepts for the zebrafish community and 80 for the political blogosphere). We finally create the socio-semantic network by linking each agent to concepts s/he previously used.

3.3 Community boundaries

The boundaries of our epistemic networks, in the sense of Laumann et al. (1983), are principally defined in semantic terms rather than in a structural way (as is done, e.g., in Doreian and Woodard, 1994). In the zebrafish case, the network is made of all agents who mentioned the term "*zebrafish*" in at least one abstract, concepts are then later retrieved from these very articles. In the blog case, the set of bloggers has been created from a first pool of candidates dealing with the American presidential election of 2008, selected by experts of LINKFLUENCE, then extended on structural grounds to neighbors who were also explicitly dealing with the election — blogs not dealing with this topic were thus discarded.

Network growth. Our epistemic networks are also *growing* networks: agents and links appear once and for all, since the starting date of our sample (i.e. 1999 for scientists, Nov 2007 for bloggers). This assumption makes sense for several reasons. It would first be methodologically difficult to talk about disconnection because, while it is possible to show the exact time when a link is created as it corresponds directly to a positive event (a collaboration or a citation), the data does not describe negative events (cessation of a link, withdrawal of an actor, etc.). Second, it is quantitatively sound: in the case of bloggers, it seems particulary difficult to qualify the obsolescence of citation links or the loss of use of concepts over a relatively short timespan of a few weeks; in the case of scientists, because of the exponential growth of the number of articles which accounts for a tremendous growth of the underlying community, network activity is significantly smaller in the first years and, *comparatively*, it may nonetheless plausibly be neglected in latter periods of a growing network.

In the remainder, the networks we consider at time t are thus formed of the aggregation of all links present until t: the dynamic social network at time t is $\mathbf{G}_t = (\mathbf{S}, \mathcal{R}^{\mathbf{S}} \cap \mathbf{S} \times \mathbf{S} \times \{0, ..., t\})$, while the dynamic socio-semantic network is denoted by $\mathbf{G^C}_t = (\mathbf{S} \cup \mathbf{C}, \mathcal{R}^{\mathbf{C}} \cap \mathbf{S} \times \mathbf{C} \times \{0, ..., t\})$. When it does not result in an ambiguity, we omit the mention of time for the sake of clarity.

Observation periods. Additionally, the temporal observation span for our datasets is of 8 years for scientists and 4 months for bloggers, which both exhibit a considerable growth over the whole period. To observe the evolution of the networks, we define eight time points for each dataset, equally spaced in time: for scientists, the time points are set at 1, 2, 3, etc. years; for bloggers, they are 15, 30, 45, etc. days, subsequently defining *observation periods*. Put simply, in the remainder, measures labelled as e.g. "period 4" will refer to the state of the network aggregated over the first 4 years in the case of scientists, or over the first 2 months in the case of bloggers.

This number of *eight* periods originally comes from the granularity of the bibliographical database, which is of one year. For matters of comparison, we therefore also defined the same number of periods for bloggers, corresponding in this case to two weeks. This choice

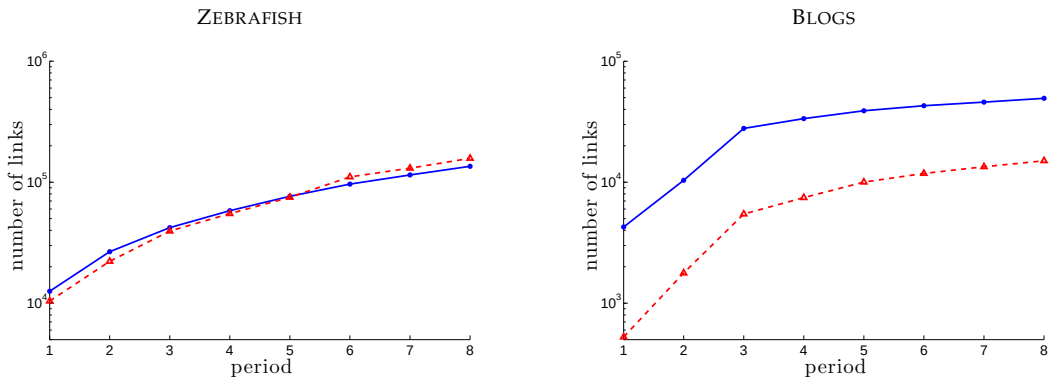

Figure 2: Total number of links in the social and socio-semantic networks (*red triangles*: social links, *blue squares*: socio-semantic links).

is also relevant in that both networks grow in comparable proportion between the first and the last period thus defined, where links increased by a factor 10 to 15 as shown on Fig. 2. In other words, over 90% of the links of the final network were not present during the first period. As we shall see, such vigorous dynamics surprisingly lead to a noticeable stability of the structural characteristics of these communities.

3.4 Qualitative description and quantitative estimation

SNA matches formal structures with sociologically relevant descriptions. In particular, we broadly elaborate upon role and position analysis (Faust and Wasserman, 1992) to exhibit connections between individual configurations and global structures (Freeman, 1989), between qualitative features and algebraic properties as "patterns" or "stylized facts" characterizing the network (Pattison and Wasserman, 1995; Anderson et al., 1992). Instead of detailing the specificities of every sub-community of the network, we will thus adopt a naturalist rather than ethnographic approach by endeavoring at exhibiting systematic and *quantitative* structural and behavioral patterns (Callon, 2001). Therefore, as we will be mostly interested in the dynamical structure of these epistemic networks, we will not investigate thoroughly the qualitative context of link creation (content of the collaboration, conditions of a citation, actual exchange of knowledge, etc.).

We will therefore describe, firstly, a series of simple stylized facts —some appropriate to social networks in general, others specific to epistemic networks— which will notably inform us about the existence of hierarchies and aggregation phenomena. More generally, we will be interested in characterizing heterogeneous or homogeneous features in each network, at the macro *and* micro level, on both social (interactional) criteria and semantic, and both in a static *and* dynamic setting. We will also endeavor at choosing stylized facts among the simplest possible.

Adopting a more micro-level perspective, focused on individual behaviors, we will secondly exhibit non-uniform interaction mechanisms —again, some being generic, others being proper to epistemic networks— which are dynamically linked to the observed stylized facts. In more details, when analyzing the role of agents we will systematically estimate to what extent the empirical behavior of link creation diverges from a uniformly random

setting; because e.g. of homophilic behavior, structural constraints, etc. Since we will systematically apply this approach throughout the paper, we technically present herebelow the very toolbox which we will be using.

Measuring agent behavior through interaction propensities. There are indeed several traditional methods to estimate *quantitatively* interaction preferences, through regression models aiming at statistically estimating structural and non-structural parameters which influence describe the diverse contributions of varied types of preferences; most notably, either:

- by assuming that the probability of dyad formation directly depends on various parameters proper to agents or to the network structure (Holland and Leinhardt, 1981; Wasserman and Weaver, 1985; Lazega and van Duijn, 1997; Liben-Nowell and Kleinberg, 2003; Powell et al., 2005),
- by using "Markov-chain"-based models (Wasserman, 1980), for instance by assuming that agents are maximizing an objective function depending on these very parameters (Snijders, 2001),
- or, very simply, by computing the proportion of links preferentially created towards some kind of agents, relatively to the proportion of these agents in the whole network (Barabási et al., 2002); in other words, *in fine*, relatively to a *uniformly random* network evolution model (*a la* Snijders, 1981, for instance).

We conform here to the latter framework, which is the most basic method, our aim being mainly to describe in a simple manner behavioral disparities with respect to particular properties, while being able to distinguish the disparities in function of the various values of these properties. Put simply, we wish to draw histograms of propensities for each value of the given property: we thus make no assumption on the shape of propensity functions (linear, quadratic, exponential, monotonous, etc.). On the other hand, this framework is too elementary to easily account for complex correlations between variables. In other words, our approach is not holistic since we ignore the simultaneous effect of parameters with respect to each other; contrarily to several above-mentioned methods such as (Snijders, 2001).

To sum up, we will simply focus on the *ratio* between (i) links which effectively appeared between some kinds of actors or dyads and (ii) links which could have appeared in a uniformly random setting, *ceteris paribus*. We assume that the interaction propensity of actors with respect to a given property m can be formally described by a function f of m: $f(m)$ represents the conditional probability $P(\mathrm{L}|m)$ that an agent of type m receives a link L (resp. that a dyad of type m appears). It is thus $f(m)/f(m')$ times more likely that an agent of type m participates in an interaction than an agent of type m' (resp. that a dyad of type m rather than of type m' appears). It is possible to estimate simply this preferential propensity through $\hat{f}$ such that $\hat{f}(m) = \frac{\nu(m)}{N(m)}$ if $N(m) > 0$, 0 otherwise, where $\nu(m)$ denotes the number of new link extremities which are pointing to agents of type m (resp. the number of dyads of type m which are created) during a time period, and $N(m)$ typically denotes the number of agents (resp. of dyads) of type m.

This will enable us to determine the *interaction propensity* in function of actor or dyad properties. Because we assume that the network is growing, we only consider *entirely new* links, i.e. appearing between dyads which were not previously linked. Proceeding in this

direction, it will in particular be possible to appraise the notion of homophily, which describes the propensity of an agent to interact preferentially with another agent because s/he is similar. Heterophily, on the other hand, describes the opposite phenomenon and, more broadly, it is possible to formally apprehend these processes as the preferential interaction of some kind of agent with some other kind of agent (Degenne, 2004).

4 Hierarchies

4.1 Heterogeneity of social & semantic capitals

Degree centrality as "capital". Degree centrality, or "degree", is a simple measure of agent connectivity (Freeman, 1978) and may *in fine* account for a more or less dominant position within a network. In an epistemic network, degree centrality may be interpreted diversely depending on whether the social or the socio-semantic network is being observed. Similarly, degree centrality bears distinct meanings in directed settings depending on whether links are received (in-bound) or given (out-bound).

In the social network, which is growing by definition here, we define the neighborhood of a node i with $\mathcal{V}(i)_t = \{j \mid \exists t' \leq t,\ (j, i, t') \in \mathcal{R}^{\mathbf{S}}\}$, the social degree of i denoted by $k(i)_t$ is $|\mathcal{V}(i)_t|$: degree exactly corresponds to the total number of past interactors or referrals during the whole observation period.

In this sense, degree may be apprehended as *social capital* in a very minimal manner, if it simply accounts for a structural capital linked to past interactions in a collaboration network or to an indirect measure of authority in a citation context (Cole and Cole, 1973). In both cases, and at least partially, it thus provides information about a kind of social stratification at work within the community. More broadly, since our intention is not to carry a detailed study of the various aspects of social capital – including its management or representation by actors — we adopt a basic understanding of the term "capital", while keeping in mind its eminently structural aspects: we therefore endeavor at *measuring structural capital in that it "facilitates some forms of social capital"* (Coleman, 1988).

Equivalently, degrees of agents in terms of socio-semantic relations may be loosely interpreted as semantic capital: in other words, the number of concepts an actor has previously used is likely to render the variety of topics s/he has dealt with in the epistemic network. The socio-semantic degree of i at t therefore measures the number of concepts i used at t and is denoted by:

$$k_{\mathbf{C}}(i)_t = |\mathcal{V}(i)_t^{\mathbf{C}} = \{c \text{ such that } \exists t' \leq t,\ (i, c, t') \in \mathcal{R}^{\mathbf{C}}\}|$$

Contrasting degree values across agents and notably exhibiting a hierarchical structure in the macroscopic distribution of $k_{\mathbf{C}}$ may thus inform us about the configuration of cultural capital.

Heterogeneity. The distribution of social degree centrality has benefited from a strong interest in the literature, especially in the very case of scientific networks; first, indirectly, by Lotka (1926) through the distribution of the number of published papers and by de Solla Price (1965) through the number of cited paper (Subramanyam, 1983) — then, more recently, by Barabási et al. (2002), Newman (2004) and Redner (2005). Hindman et al. (2003) and Adamic and Glance (2005) have carried similar studies in the case of political opinion websites.

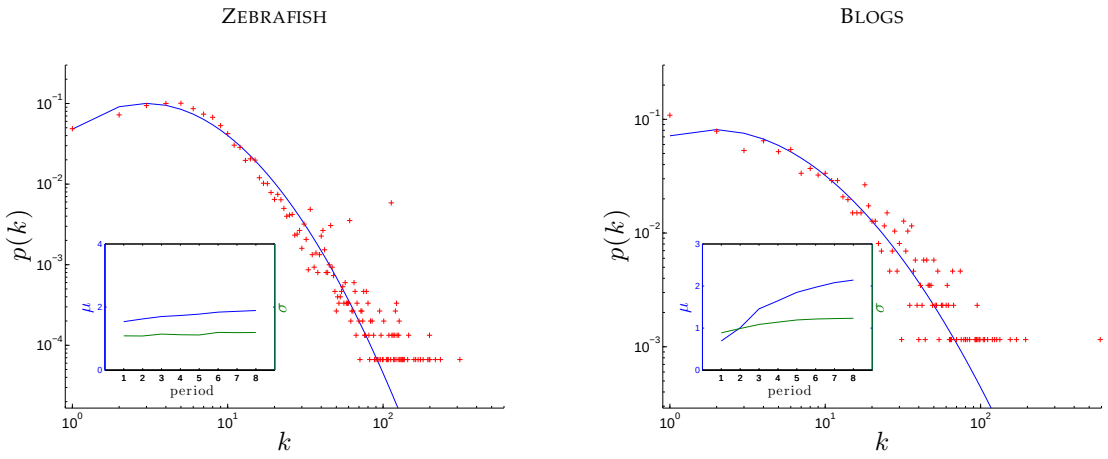

Figure 3: Distributions of "social capital", i.e. degree centrality in the social network. *Points:* $N(k)$, number of agents of degree "k" at the last period. *Continuous line:* best log-normal fit. *Inset:* evolution of μ and σ over 8 periods.

The study of semantic capital, on the other hand, has remained mostly overlooked. We notice on Figs. 3 & 4 that semantic and social capital distributions are similar, in the sense that they exhibit the same kind of heterogeneity: a small yet non-negligible number of agents have used many concepts. Empirically, this kind of distribution is traditionally said to be approximated by a "power-law", or "Zipf law" (or "Pareto law") even if, quite often, this fit only accounts for the right part of the distribution (higher values). Frequently, distributions actually exhibit a roughly flat shape for lower values of the variable, followed by a sharp decrease for higher values, which asymptotically appears to tend towards a straight line in a log-log setting — in which case, a "log-normal" distribution appears to be much more adequate. On Figs. 3 & 4 empirical data is thus fitted using this reference distribution.

Without debating further which analytical probability distribution would be the most relevant to describe the empirical data, it is worth noting that, in all generality, these distributions are spread on several orders of magnitude, which is typical of a strong heterogeneity between agents. It is additionally mostly asymmetrical: there is a non-negligible number of agents with a high degree, while more and more agents have lower and lower degrees. These two features confirm the hierarchical structure of both social and semantic capital in both networks.

Hierarchical homogeneity. This stratification is, however, not deprived of various forms of homogeneity. Scientists who already had a high number of collaborations are likely to have links with similarly "rich" scientists, while agents with a lower social capital are here generally linked to equally "poor" collaborators. As shown on Fig. 5, zebrafish embryologists are assortative, like other kinds of scientists (Newman, 2002). As such, connections among rich agents are homogeneous in the sense that similar agents flock together. On the other hand, no such trend is observable on bloggers: rich agents are equally likely of being cited by all types of agents, while they also cite rich agents no more often than poor agents (because of the directedness of the network we distinguished these two cases). In this case, connections are indifferently homogeneous in the sense that all agents have similarly rich neighborhoods

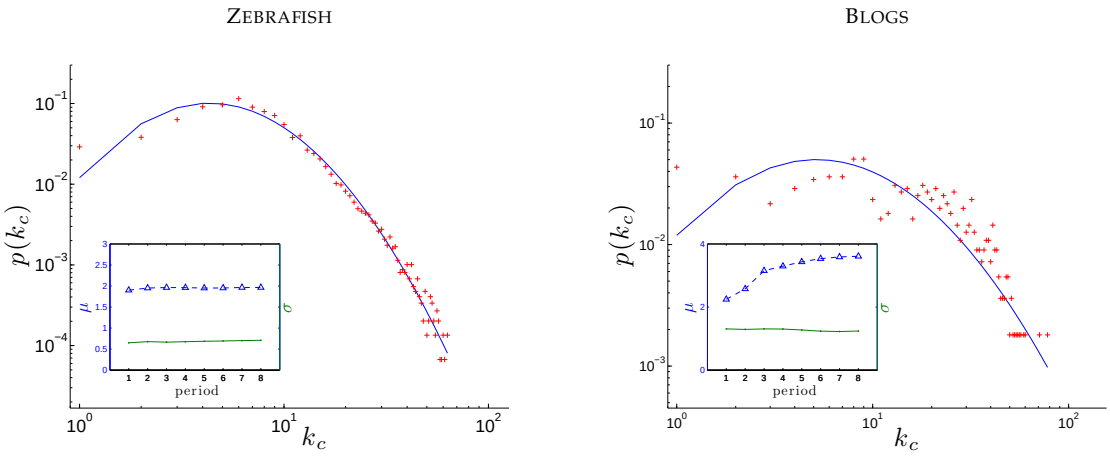

Figure 4: Distributions of semantic capital. *Points:* $N(k_{\mathbf{C}})$ at the last period. *Continuous line:* best log-normal fit. *Inset:* evolution of μ and σ over 8 periods.

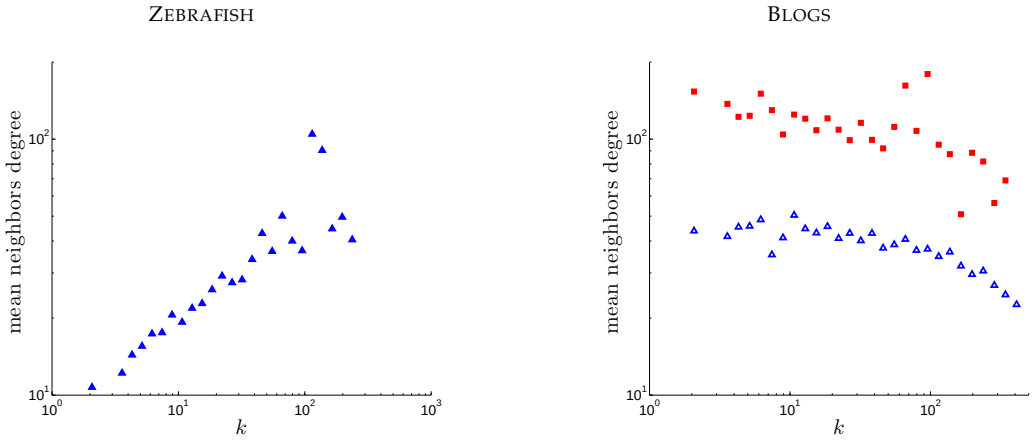

Figure 5: Assortativity in terms of social capital k, estimated through averages of mean neighbor degrees. In the directed case of the blogosphere, at right, squares describe mean neighbor degrees of blogs cited by ego whereas triangles correspond to blogs citing ego.

(both in- and out-bound).

In the two cases anyhow, the resemblance of the heterogeneity observed for both types of capital incites to look for some possible correlation between the two variables. Formally, Pearson's correlation coefficient for scientists is 0.67, which traditionally renders a relatively sensible correlation, while it is 0.38 for bloggers, a much less significant value. These facts have to be contrasted with the density map of the joint values of each capital (Fig. 6). This map indeed confirms that there exists a wide range of possible combinations of joint values of semantic and social capital. For instance, while scientists with a small social capital are bound to have a limited semantic capital, this does not seem to be the case in the blogosphere — admittedly, rarely-cited bloggers may address a large range of topics, whereas weakly-linked researchers are generally on narrow issues. On the other hand, it seems to be hardly possible for a blogger to have a large social degree without having a large semantic capital, to the contrary of scientists who may be socially rich but semantically focused.

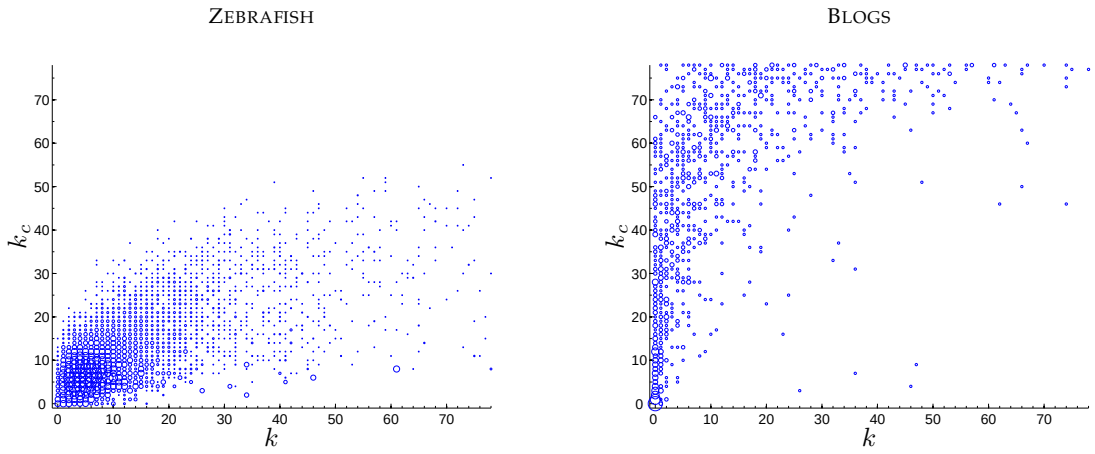

Figure 6: Joint distributions of social and semantic capitals: circle diameters are proportional to the number of agents having a couple of degrees $(k, k_{\mathbf{C}})$.

On the whole, denser areas on these maps have very distinct positions in the two cases: most scientists are spread in the bottom-left part of the diagram (where no striking correlation exists between social and semantic capitals), while bloggers are mostly located in an upper-left triangle (where it is impossible to correlate a high semantic capital with any particular level of social capital or, symmetrically, to correlate a low social capital with a particular value of semantic capital). Beyond the straightforward observation that both types of capital are not necessarily bound to match, this phenomenon suggests that processes underlying the appearance of social and socio-semantic links may obey distinct rationale. This supports the relevance of a disjoint study of their dynamics, as will indeed be confirmed herebelow.

4.2 Dynamic hierarchies

At a macroscopic level, the very hierarchical structure that is observed for higher degrees is *stable* during time: all degree distributions roughly exhibit the same trend for all periods, and the value of the σ parameter of the log-normal fit, which describes the tail, is indeed stable. This stability is generally surprising, since both networks are growing at a remarkable pace, as mentioned in Sec. 3.3: most agents and/or links present at the last period were not initially active. As such, assuming a correspondence between centrality distributions and hierarchical structures, this process can be likened to "*spinning-top models*" (Lazega et al., 2006) where top position configurations are temporally stable, even when their members are replaced at a sustained rate. Here, social and semantic hierarchies are thus *dynamically stable*.

Furthermore, these hierarchies are obviously not typical of networks where links would form in an uniformly random manner (as is the case with the model of Erdős and Rényi, 1959, often considered as a null-model of social network morphogenesis). This supports a further investigation of the shape of interaction processes and preferences (as will also be carried throughout the paper: see therefore §5.1 & §5.2.1). Put differently, is there a dynamic relation between relational and cultural wealth at a local level which underlies, or even echoes, these macro-scale observations?

To briefly sum up, we principally notice two types of regularities and correlations regarding socially "rich" agents, i.e. *those who have received most links*: (i) they are in a stable

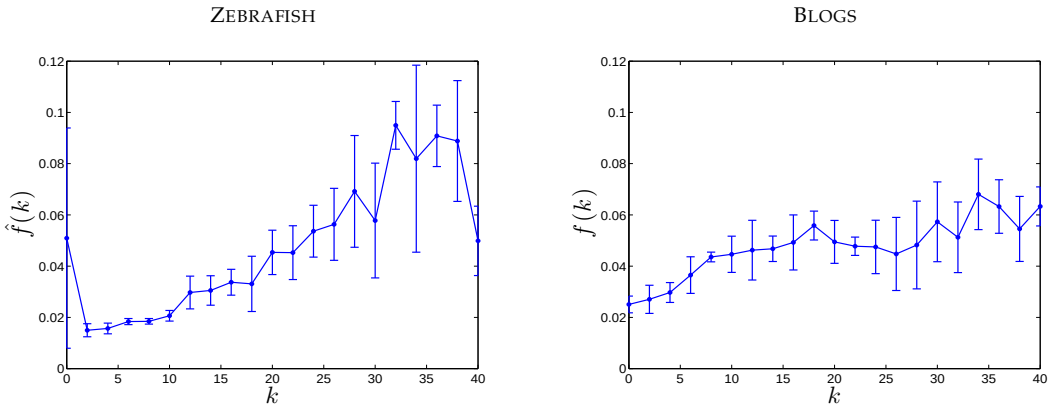

Figure 7: Interaction propensity $\hat{f}$ with respect to social capital k. Average on 8 periods, confidence intervals at 95%. (Note: to accommodate to data scarcity, graphs have been smoothed by binning (bin size: 2) and truncation ($k < 40$).)

relative proportion (Figs. 3 & 4) and (ii) they are rarely poor semantically (Fig. 6) — the only divergence between the two empirical cases relates to the fact that "rich" scientists are usually linked to past collaborators having a high social degree (Newman, 2002) while the configuration of citations is much more uniform for bloggers who, whichever their degree, are globally cited by similarly rich agents. Agent preferences in link creation could therefore relevantly be appraised in order to determine whether a possible reinforcement of these similarities is at work: is relational or semantic wealth a robust predictor of new collaborations or citations?

Using the methodology presented in §3.4, we estimate interaction propensities f and $f_{\mathbf{C}}$ with respect to social and semantic capitals k and $k_{\mathbf{C}}$, respectively. We approximately confirm the assertion that propensities are roughly proportional to the degree (Figs. 7 & 8), i.e. social links preferentially go to agents having higher degrees *both* in the social and semantic dimensions. In the very case of social capital in scientific networks, this result also partially corroborates previous works by Jeong et al. (2003) (let us mention that an identical phenomenon has also been described in scientific *citation* networks by de Solla Price (1976); these findings provide a quantitative sketch of capital accumulation dynamics in scientific communities in terms of both interactions and authority attributions). Yet, more interestingly, comparing propensities between the two cases reveals that they are sensibly different: indeed, linking propensities are much flatter for bloggers, and poorer bloggers tend to be less disadvantaged in receiving links. This in turn echoes the relatively more pronounced unassortativeness of the blogger network, as said above.

Additionally, these increasing propensities can be interpreted indifferently (i) as a roughly increasing *preferential attachment* to agents with higher degrees or, as well, (ii) as a stronger *activity* from agents with higher capital: more active scientists participate in more authoring events, thus creating/receiving mechanically more links, whereas more active bloggers post more, thus providing opportunities for being cited more, irrespective of their present degree. At this point, these two interpretations are both consistent with observed propensities.
Yet, we could notice that the activity of scientists correlates almost perfectly with social and semantic capitals, while the activity of bloggers is much less correlated with capitals — on

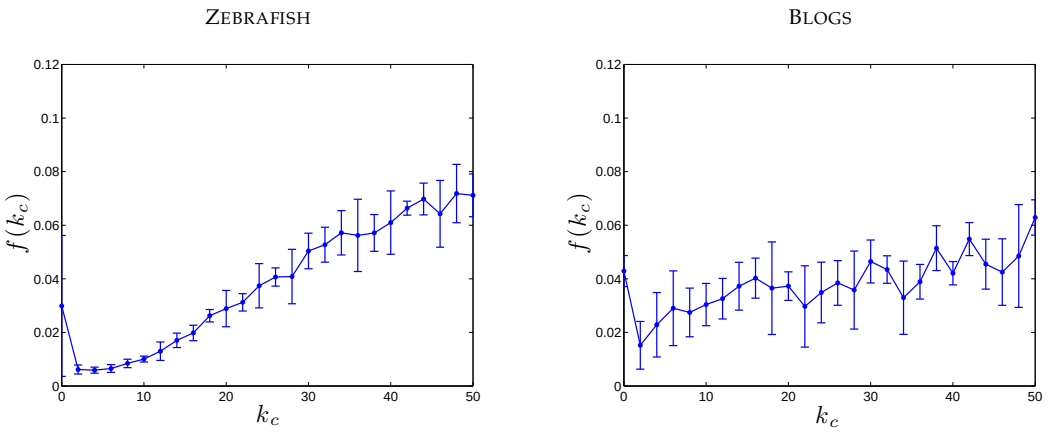

Figure 8: Interaction propensity $\hat{f_{\mathbf{C}}}$ with respect to semantic capital $k_{\mathbf{C}}$.

the whole, activity graphs in both cases imitate propensity trends. This latter interpretation puts the focus on activity rather than attraction and diverges from what is usually proposed in studies dealing with the notion of "preferential attachment" (see e.g. Barabási and Albert, 1999, "(...) *a new actor is casted most likely in a supporting role, with more established, well-known actors* (...)"). By contrast, it is more consistent with traditional sociological interpretations where, for instance in scientific communities, the number of papers, thus collaboration events, are simply proportional to research activity (Gordon, 1980). More broadly, these observation cast doubt on usual assertions regarding increasing propensities as a universal phenomenon related to attractivity, rather than a context-dependent process correlated with activity.

5 Communities and neighborhoods

Beyond the observation of hierarchical features it is possible to investigate cohesion between agents, in a broad sense. We start with a strictly social and somewhat usual point of view (§5.1) to extend the formalism to a socio-semantic perspective, introducing both local and large-scale motifs (§5.2).

5.1 Social cohesiveness

Local cohesiveness may be appraised in a very basic manner through the population of triads (Holland and Leinhardt, 1976; Snijders and Stokman, 1987) which more precisely exhibits how and how much neighbors of an agent are also neighbors (or not). In particular, this notion may refer to two different kinds of topological feature and underlying behaviors: *clustering*, or the proportion of neighbors of ego who are also direct neighbors, and *transitivity*, or the fact that a neighbor of a neighbor of ego becomes a neighbor of ego. Several kinds of triads have been found to be significantly frequent in many social networks (Watts and Strogatz, 1998; Milo et al., 2004) while this topological feature, along with degree centrality, has also been the target of many recent models (Pattison et al., 2000; Jin et al., 2001).

Here, beyond appraising the social cohesiveness of our epistemic networks, we are also interested in the way the social and semantic capital of an agent may be correlated to its

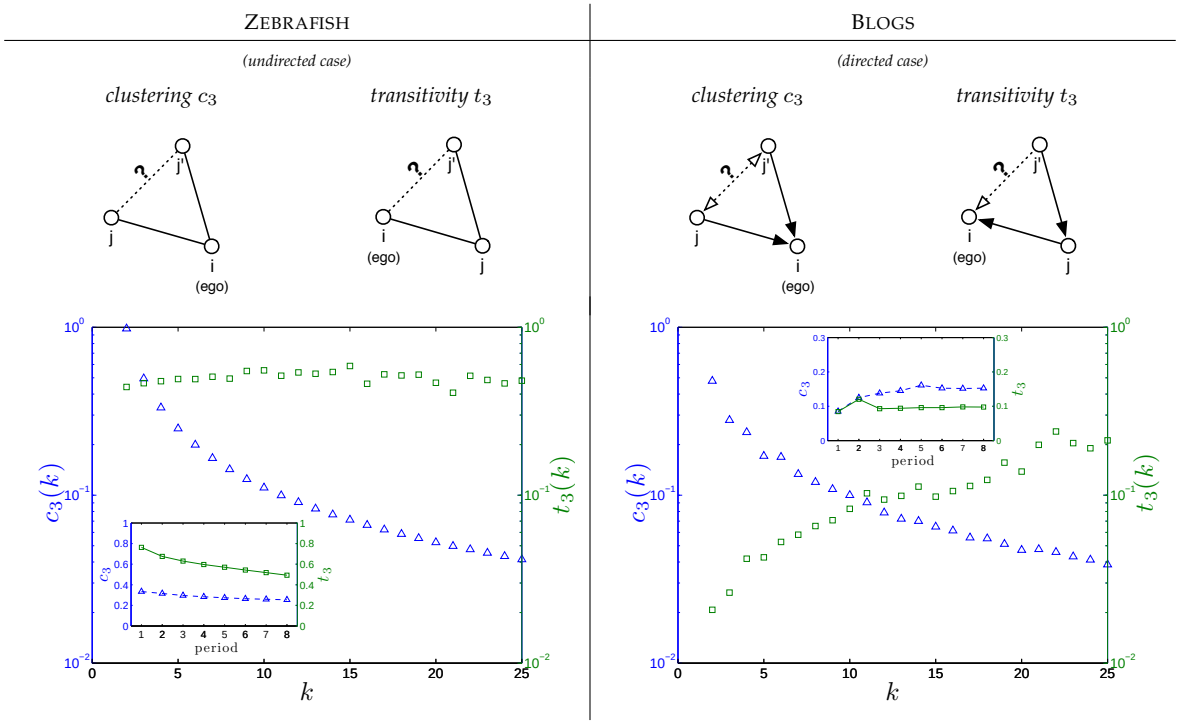

Figure 9: Clustering c_3 (blue triangles) and transitivity t_3 (red squares). *Top:* decriptions of patterns. *Below:* values with respect to social capital k (*insets:* evolution of average values).

tendency to belong to closed triads. More precisely, we first define, indifferently for both directed and undirected networks:

- the clustering coefficient c_3 of an agent i as the proportion of links between pairs of agents who cited or collaborated with i: $c_3(i) = \dfrac{|(j,j') \in \mathcal{V}(i)^2 \text{ such that } j' \in \mathcal{V}(j)|}{k(i)\,(k(i)-1)}$
- the transitivity coefficient t_3 of i as the proportion, among paths of length 2 to i (i.e. from neighbors of neighbors of i), of paths of length 1 (i.e. direct neighbors): $t_3(i) = \dfrac{|(j,j') \in \mathcal{V}(i)^2 \text{ such that } j' \in \mathcal{V}(j)|}{|(j,j') \text{ such that } j \in \mathcal{V}(i) \text{ and } j' \in \mathcal{V}(j)|}$

These patterns are schematized on Fig. 9, with the corresponding empirical results for both networks. We plotted the clustering and transitivity coefficients against social capital at the last period of our datasets as well as the evolution of the average over the whole network of these coefficients: $c_3 = \langle c_3(i) \rangle$, $t_3 = \langle t_3(i) \rangle$. To the contrary of several other results of this paper, we did not find any significant relationship between the semantic capital and social cohesiveness.

We found however two dimensions of constrast with respect to social capital: on one hand, c_3 vs. t_3, on the other hand, scientists vs. bloggers. In all cases, all values are well above what would usually be expected in a network exhibiting the same density. We verified this through series of simulations on uniformly random networks (Erdős and Rényi, 1959). The collaboration network feature very high c_3 and t_3 values, with respectively 2 and

3 orders of magnitude above the random case; the blog network is approximately one order of magnitude above this null model. First, c_3 is related to the density of the immediate surroundings of ego. It is decreasing with the social degree in both networks, indicating that the most connected agents tend to have a less clustered neighborhood. This decreasing shape also corroborates previous studies and seems to be a classical bias of local clustering measurement (for an extensive discussion, see Soffer and Vázquez, 2005).

This is in stark contrast with the behavior of t_3, almost constant in the scientific case, and significantly increasing for bloggers; t_3 apparently does not suffer from the above-mentioned bias concerning c_3. The trend of t_3 in the blog case shows that agents with higher social capital tend to have attracted comparatively more links originating from their neighbors of neighbors than lower social capital nodes. This effect does not seem to hold in the zebrafish network, which is undirected — this property indeed induces some uncertainty in the interpretation of the behavior of ego, as it is obviously impossible to tell from the data if ego is the target or the initiator of a transitive triad, admittedly undirected.

Transitive processes. Admittedly, both measures capture significantly different ego-centered properties, as well as probably distinct underlying behaviors. Nonetheless, their average values seem pretty stable over time on both networks, as evidenced by insets on Fig. 9 which describe the evolution of averages of both clustering and transitivity coefficients, and even if, again, a massive number of new links (for blogs and scientists) and new agents (for scientists only) are added during time (Fig. 2).

In a dynamic setting, and with no ambition to fully decipher the underpinning of these peculiar triadic landscapes, we can partially investigate the shape of the behaviors of local triad formation. We do this by measuring to what extent agents create links with neighbors of neighbors (i.e., how have neighbors of neighbors of an agent become direct neighbors?). We achieve this by estimating the propensity of link formation with respect to the social distance d, which is the smallest number of steps one has to navigate from an agent to a given agent. In a broader perspective, it is altogether possible to examine to what extent links are forming towards "longer-distance neighbors", at distances larger than 2. Exceptionally, and without losing in generality, we also computed the propensity for link reiterations, that is, repeated citations or interactions, which corresponds to $d = 1$. To reframe the corresponding propensities, we first show the global distribution of distances between all possible actor pairs, for each network, on Fig. 10. As can be seen, there are many more couples of actors at a higher distance.

Propensity results are next gathered on Fig. 11 — note that we grouped all values for distances strictly above 4 as they were roughly identical. When a link at distance two is formed, a triad appears; and we can first notice that there is empirically a much higher propensity for this kind of links to form. There is, comparatively, an exponentially lower likeliness to form links at a longer distance, i.e. with more remote neighbors. In short, most new interactions are of a triadic nature and tend to reinforce the existing cohesiveness; hence shedding light on the particularly high empirical values of c_3 and t_3. Yet, we can also notice that interaction repetition is an even more significant source of link creation, signalling that, on the whole, only a small proportion of links are created between agents which were not previously connected. Although apparently unconnected to semantic issues, we will demonstrate below that this phenomenon also helps clarifying the existence of a strong homophily.

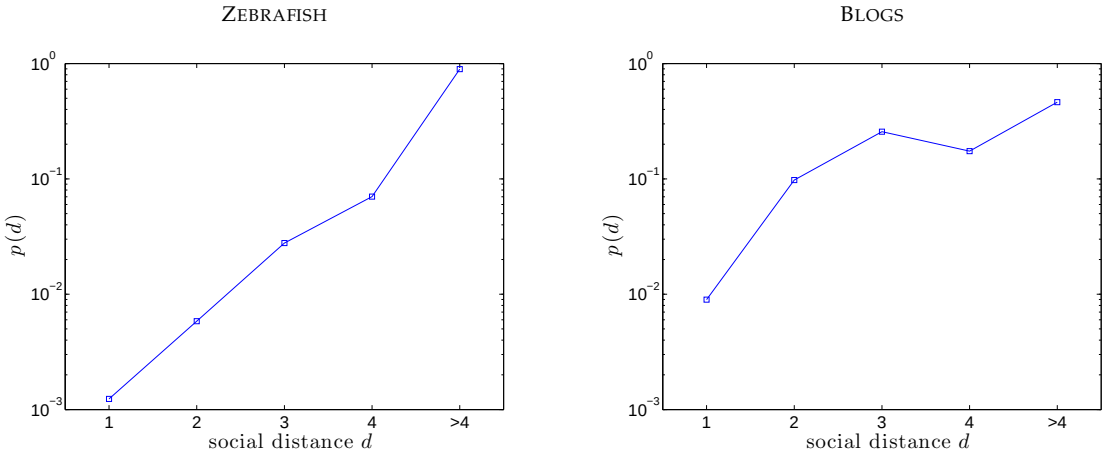

Figure 10: Distribution of actors at social distance d.

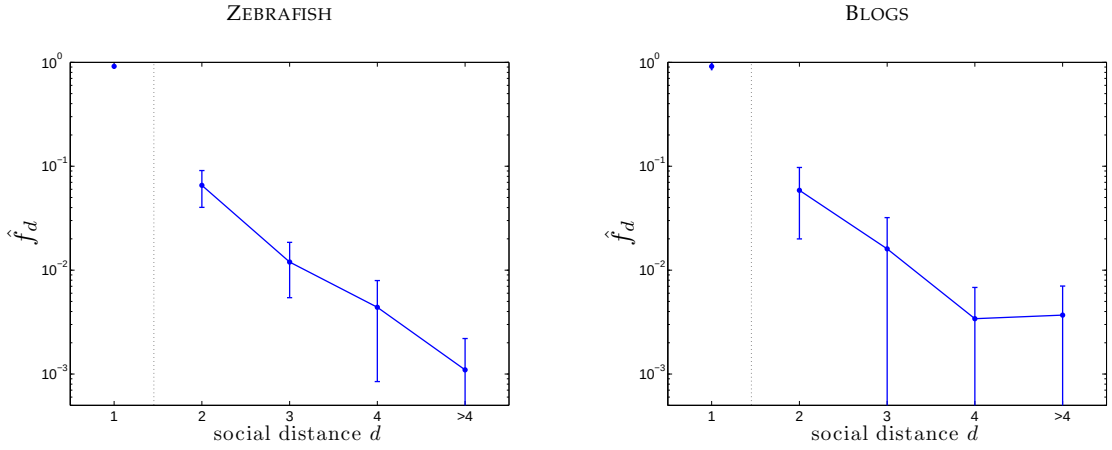

Figure 11: Interaction propensity $\hat{f}_d$ with respect to social distance d.

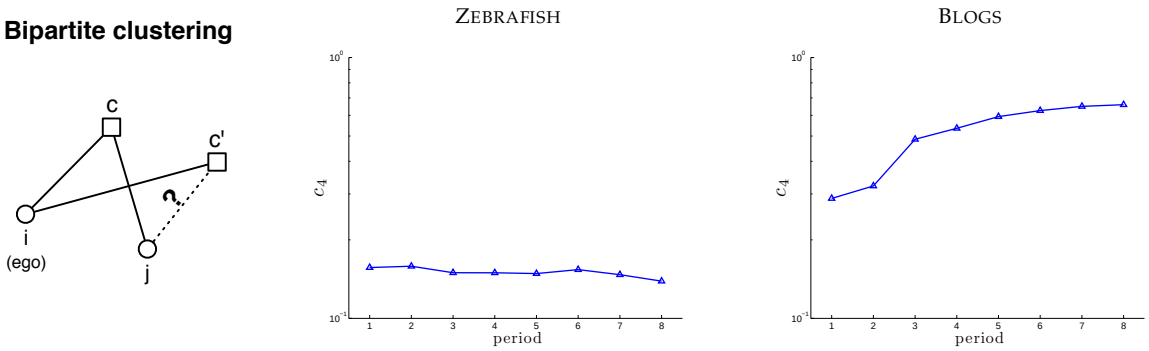

Figure 12: Socio-semantic clustering c_4: schema (left) and evolution of average values (right).

5.2 Socio-semantic aggregates

5.2.1 Local semantic cohesiveness

Could there be a socio-semantic equivalent to these strictly social cohesiveness and transitiveness? More broadly, are there socio-semantic aggregation patterns and processes proper to these epistemic networks? For instance, we may first investigate whether two agents sharing one concept are likely to share more. To this end, we define a *bipartite clustering coefficient* c_4 (Robins and Alexander, 2004) for the socio-semantic network, as the ratio of diamonds around an agent (i.e. the probability that pairs of concepts used by an agent are jointly used by another agent), as sketched on Fig. 12. As such, it is the most basic equivalent to the above-mentioned strictly social cohesiveness coefficients. Put formally, for agent i, we have:[3]

$$c_4(i) = \frac{\sum_{\{c,c'\} \subseteq \mathcal{V}^{\mathbf{C}}(i)} [\kappa(c,c') - 1]}{\sum_{\{c,c'\} \subseteq \mathcal{V}^{\mathbf{C}}(i)} [k_{\mathbf{C}}(c) + k_{\mathbf{C}}(c') - \kappa(c,c') - 1]}$$

where $\kappa(c,c')$ is the number of agents linked to same same pair of concepts (c,c'), i.e. $\kappa(c,c') = |j \in \mathbf{S} \text{ such that } \{c,c'\} \subseteq \mathcal{V}^{\mathbf{C}}(j)|$.

Average values for both networks on all periods are shown on Fig. 12. This semantic cohesiveness is high (between about 15 to 75%), and again, much higher than in equivalent socio-semantic networks having the same density (two to ten times higher). No dependency was found, however, with respect to social nor semantic capital.

Proximity and neighborhoods. This high local overlap between pairs of agents and concepts leads us to examine more thoroughly whether social and semantic neighborhoods coincide. In other words, to what extent are agents semantically close to each other, in the network and, more specifically, in their social neighborhood? To this end, we first need to

[3] Alternative definitions of bipartite clustering, as previously proposed (Robins and Alexander, 2004; Lind et al., 2005; **?**), may diverge quantitatively from this measure, because they e.g. describe the ratio of closed cycles of length 4 over open cycles of length 4 (Zhang et al., 2007), or compute the ratio of diamonds directly over the whole network. We checked that our qualitative results hold with these various formulas.

define a notion of semantic distance δ between pairs of agents. This distance should be such that it increases from 0 to 1 with a decreasing proportion of shared concepts. We choose a cosine-based distance δ based on the classical *tf.idf* framework (Salton et al., 1975) which assigns to each agent a semantic profile based on usage weights (rarer terms weigh more). Note that distances based on the Jaccard coefficient (Batagelj and Bren, 1995) yield similar results (**?**).

Distributions of semantic distances in both networks as plotted on Fig. 13 reveal that there are on the whole few semantically similar agents, especially scientists. We can however notice that neighbors (blue squares) are at a much smaller semantic distance, especially scientists, again. As for scientists, this phenomenon should not be surprising *per se*, since collaborations induce links towards the same concepts, which mechanically produces similarity among ex-coauthors[4] — yet, even in that case, the semantic proximity of neighbors is extremely strong when compared with the rest of the network (with a discrepancy of about three orders of magnitude for the smallest distances).

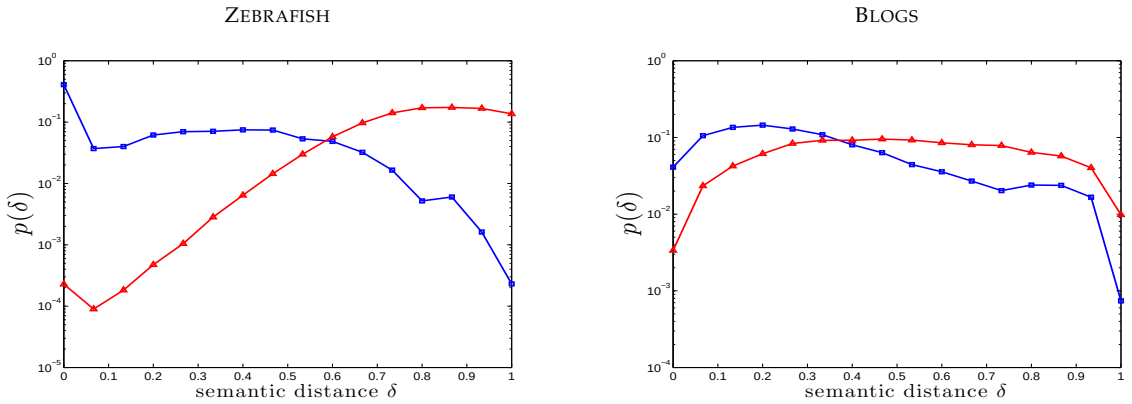

Figure 13: Semantic distance in the neighborhood (blue squares) and in the whole network (red triangles).

Semantic homophily. In this sensibly homophilic landscape, insofar as social and semantic capitals influence agent behavior, we ought to examine as well the effect of the (static) semantic proximity context on future interactions. As agents prefer to establish relationships with similar fellows (for a review, see McPherson et al., 2001), it is indeed quite probable that the topology of interactions is also modified by semantic profiles. To what extent, quantitatively, could semantic proximity predict, dynamically, the likelihood of coauthoring papers, or citing others? This comes to quantitatively determine homophilic processes (as in Fienberg and Wasserman, 1981; Lazega and van Duijn, 1997, for instance), yet in a co-evolving framework: here, semantic features jointly evolve with the successive reconfigurations of the social structure and agent semantic profiles (Roth, 2005; Crandall et al., 2008).

Using the same methodology as in the previous sections, we appraise semantic affinity-

[4]Hence, note that this effect derives in part from an artifact of the protocol, or, plausibly, from an artifact proper to this kind of community if we assume that this property of the empirical protocol accounts for a real phenomenon (which is equivalent to saying that all collaborators have effectively adopted all concepts involved in the interaction — without elaborating further on this latter interpretation).

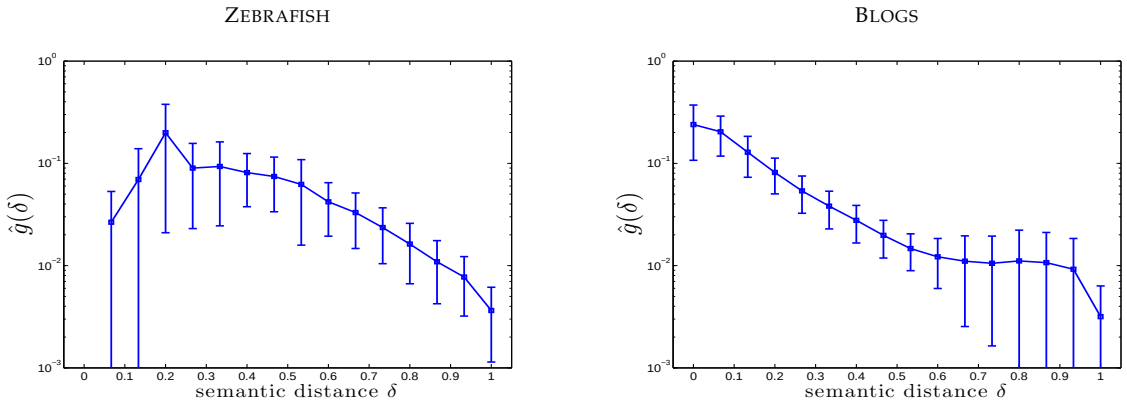

Figure 14: Homophilic propensity g with respect to semantic distance δ.

based interaction propensities g using the above-mentioned semantic distance δ. The trend of both graphs of $\hat{g}$ on Fig. 14 renders an overall behavior massively favoring link creation between agents displaying similar semantic profiles (recall that we only consider *new* link creations, between agents not previously linked). Yet, regarding scientific *collaboration*, there is a slight preference for dissimilarity; i.e., in a strongly homophilic landscape as it is for the zebrafish, we observe a moderate heterophily. Scientists favor interaction with scientists working on similar topics, but *not too much*, granting a bit of diversity — by contrast, bloggers display a much simpler homophilic behavior: $\hat{g}$ is just decreasing.

While Fig. 13 describes a static homophily, i.e. *ex-post*, Fig. 14 describes a dynamic homophily, i.e. *ex-ante* (that is, before agents become neighbors in the social network). On the whole, these results suggest that, while bloggers are more strictly homophilic when citing than scientists when collaborating, the absence of an occasion to explicitly share topics does not make them sensibly similar *a posteriori* — they are less homophilic *ex post*, which is likely to induce a larger spread of semantic profiles over the network, even within the very neighborhood. On the contrary researchers, who are putting together concepts when collaborating, get therefore much closer in comparison with the rest of the scientific network. Put differently, even if scientists who start a collaboration are not necessarily extremely close semantically beforehand, they tend to become much closer afterwards[5] — especially given the high rate of interaction repetition as demonstrated on Fig. 11.

5.2.2 Epistemic communities

Extending this perspective, it is possible to describe agent groups typical of knowledge networks, on a large-scale basis: in particular, the present formalism can cast light on the existence and configuration of epistemic communities. The term of "epistemic community" (EC) traditionally refers to the collaboration of agents who work (i) within the same epistemic framework and (ii) towards common and collective goals of knowledge production or information validation (Haas, 1992). In a scientific context, ECs therefore classically describe

[5]This admittedly happens partly by construction of the network, since writing papers implies concept sharing. Nevertheless, it cannot be dismissed that the protocol also accounts for a realistic interaction process: after a collaboration, *all* authors indeed supposedly become acquainted with *all* concepts mentioned in a paper abstract.

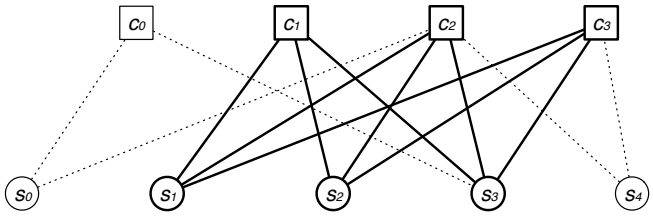

Figure 15: An epistemic community as a biclique of the socio-semantic network : $\{(s_1, s_2, s_3), (c_1, c_2, c_3)\}$.

groups of researchers advancing a field while recognizing a common set of conceptual tools and shared representations (Kitcher, 1995), within a paradigm. By analogy, in the context of online communities ECs may loosely refer to subgroups of individuals who are addressing identical issues or refering to similar topics — such as bloggers interested in similar political matters (Hindman et al., 2003).

In order to identify these communities in a concise manner and relate them to simple socio-semantic patterns, we adopt a strictly descriptive approach aiming at inventorying sets of topics and their actors. More precisely, we do not explicitly pay attention to how agents behave with respect to deference, authority, or even knowledge transfer matters, as is often the case in more qualitative works (Bourdieu, 1991; Lazega, 1992). Rather, we adopt a fundamentally structural notion which corresponds to configurations where groups of agents share common groups of concerns, within an identical conceptual framework Haas (1992). As such and for instance, the present study considers the whole "zebrafish" community and the whole political blogosphere as ECs of reference. In this respect, while our understanding of the notion of "epistemic community" seems to be admittedly restricted, it already constitutes a primary step in characterizing the limits of actual ECs – ECs within which actors elaborate locally knowledge, relevance judgments, etc.

Bicliques and ECs. We thus formally define an EC as a pair of a set of agents and a set of concepts, such that all agents share all concepts (or, dually, such that all concepts are shared by all agents). This pair of sets is *maximal*: it is not possible to find more agents sharing the same concepts, or more concepts share by all these agents. This definition exactly corresponds to a *biclique* in the socio-semantic network, as a maximal set of agents linked to a maximal set of concepts — see Fig. 15. Bicliques may appear as a generalization of the above-mentioned socio-semantic clustering (c_4) and, more broadly, to a loose understanding of the notion of structural equivalence (Lorrain and White, 1971). The EC pattern also defines communities at various levels of generality, encompassing a variable number of agents and topics, and such that agents, like concepts, may simultaneously belong to several, possibly overlapping epistemic communities (Roth and Bourgine, 2005).

The identification of this kind of dual structure in networks has been the focus of several qualitative and quantitative(Breiger, 1974; Wille, 1992; Freeman, 1996; White and Duquenne, 1996; Falzon, 2000; Roth and Bourgine, 2005; Lehmann et al., 2008) studies within the framework of bipartite graphs. Freeman and White (1993) have notably shown how to jointly group agents and events they participate in. On the whole however, qualitative approaches

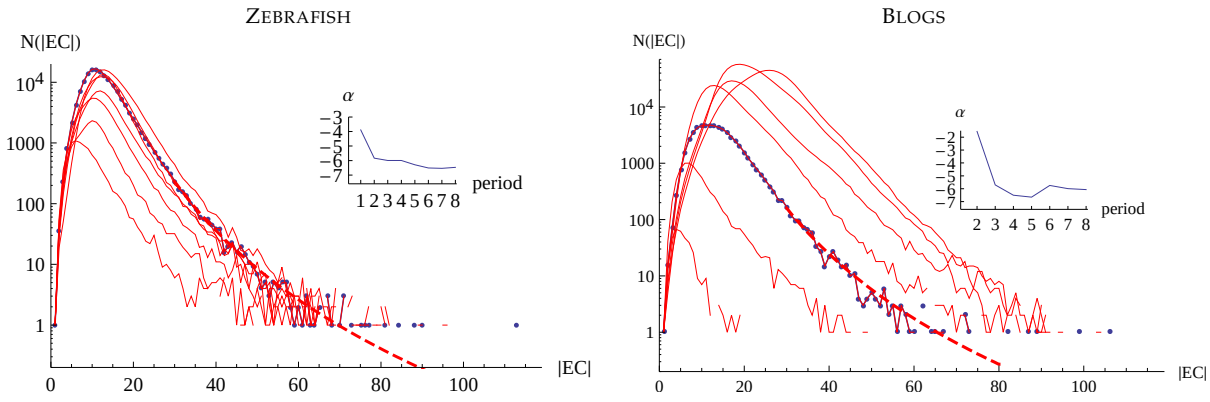

Figure 16: Distribution of EC populations (number of agents), from period 1 (bottom curve) to period 8 (top curve). The zebrafish dataset is outlined on period 7, and fitted by a power-law tail (dashed thick line), it is period 3 for blogs. Insets correspond to the evolution of the exponent of the power-law fit, i.e. the structure of the set of most populated ECs.

in SNA focus on social aspects and, often, on relatively local aspects — leaders, peripheral members, cooperation process within and/or between groups. Here, we use these epistemic structures not necessarily to focus on the role of particular individuals but, rather, in order to appraise which semantic groupings are the most salient quantitatively, in terms of population size. In other words, we carry a demographic study of the bicliques and compare the relative weight of each epistemic community (and corresponding topic groupings) within the whole population. To sum up and more to the point, thanks to the distribution of populations of the various ECs it is possible to have an overview of the global landscape the dissemination of topics over agents in the epistemic network.

Demographics of ECs. Additionally, focusing on larger groups yields a better insight of the large-scale epistemic structure of the network. Noticing indeed that more complex and longer lists of concepts correspond to smaller agent groups, ECs gathering few agents are likely to be very specific and specialized, sometimes typical of single agents. On one hand, given a threshold on population sizes, a hierarchically overlapping representation of these larger ECs could then give us a quick and compact insight on which main topic groups can be found in the whole epistemic network and, more importantly, on their relative importance — for more details on this taxonomical approach, we refer to Roth and Bourgine (2006); **?**. On the other hand, and beyond this global epistemological picture, two elements are of interest to us here:

(i) There is a particular structure of populated ECs. A significant number of groups of topics gathers a significant share of the whole community, as demonstrated on the right side of the demographic graphs on Fig. 16 (because of computational complexity issues, we carried computations on a limited random sample of 150 agents from each empirical dataset). In other words, a relatively small yet substantial number of associations of topics are especially popular, while many concept sets gather smaller groups. Put shortly, this structure is heterogeneous: it is spread on several orders of magnitude

and there are also significantly more small ECs than populated ECs. Additionally, the tail of the distribution has a similar shape in both epistemic networks, for zebrafish and blogs. This tail partly characterizes the structure of the set of most populated ECs.

(ii) Focusing on these larger ECs, thus on the distribution tail, we moreover notice that it exhibits a stable shape, in spite of significant micro-level variations. Excepted the very first period(s), power-law fits indeed remain within the same range for all periods, as shown on the insets on Fig. 16.While out of scope of the present paper, an historical study of the underlying taxonomy, in terms of ECs and sub-ECs, would additionally demonstrate that the main thematics of zebrafish embryologists globally gather a stable proportion of the whole (growing) population (Roth and Bourgine, 2006; **?**), even if some new topics emerge. In other words, despite the fact that distributed knowledge production relies on a significantly increasing number of new agents, its organization is only slightly modified — which indicates the functional stability in time of the community (Pattison, 1993).

Besides, the main ECs attract proportionally more agents with time, as evidenced by the top-right progression of EC distributions on Fig. 16. Yet, when estimating the propensity to choose concepts at the core of the largest ECs relatively to less used concepts, we can eventually notice that there is indeed a marked tendency, in both networks, to make proportionally more socio-semantic links towards more popular topics. This low-level behavior, in turn, is plausibly likely to underlie the temporal reinforcement of the main ECs.

6 Conclusion

We aimed at introducing epistemic networks as a framework wherein knowledge communities could be studied in a dual manner: social structural aspects on one side, echoed by socio-semantic features on the other side. In this respect, our purpose was to convert several strictly social indicators — which might have appeared familiar to the reader — into simple socio-semantic analogs (see Tab. 1 for a summary). We have thus more broadly designed stylized facts proper to knowledge networks and revealed interaction processes which depend on the epistemic network *as a whole*.

Independently of the peculiarities of each dataset, this epistemic framework renders (i) heterogeneities in both social and semantic dimensions (ii) which support hierarchies between agents and which, in turn, are diversely homogeneous; and (iii) social and semantic cohesiveness, attested at both a local and a more global scale. Further, we could describe the behavioral (ego-centered) counterpart of each of these observations by exhibiting higher propensities of link creation towards richer and semantically similar agents, in both cases. In a dynamic perspective, this approach enabled us to characterize the coevolution of social structures and semantic features by exhibiting the joint and reciprocal dependence of social linkages on the socio-semantic network. Semantic homophily, for instance, as well as socio-semantic bicliques, could hardly be *reduced* to the strict social network. Additionally, noticing that the vigorous dynamics of the networks did not prevent the existence of temporally stable patterns, we suggested that some properties of the low-level behavior tended moreover to foster and reinforce the above-mentioned patterns.

properties	ZEBRAFISH	BLOGS
social capital k	*heterogeneous distribution, temporally stable*	
	assortative	*slightly disassortative*
semantic capital k_c	*heterogeneous distribution, temporally stable*	
	positively correlated with k	*diversely correlated with k*
propensity f to social capital	*increasing*	*slightly increasing*
propensity $f_{\mathbf{C}}$ to semantic capital	*increasing*	*slightly increasing*
clustering (c_3)	*very high*	*high*
	decreasing with k	
	temporally stable	
transitivity (t_3)	*very high*	*high*
	stable with k	*growing with k*
	temporally stable	
social distance distribution	*higher proportion of pairs at a long distance*	
semantic distance distribution	*neighbors semantically closer*	
propensity f_d to social distance	*strongly decreasing*	
propensity g to semantic distance	*slightly increasing for small δ before decreasing strongly*	*strongly decreasing*
socio-semantic clustering (c_4)	*high, temporally stable*	
epistemic communities	*heterogeneous distribution, temporally stable*	

Table 1: Qualitative summary of the measures.

It would plausibly be useful to extend this framework to other communities than bloggers and scientists in order to check the presence of the same kind of epistemic patterns. Are there, also, some regularities in the behavior of agents which could be generic of knowledge networks, or at least observed in other kinds of epistemic networks — in particular between interaction and citation networks? Beyond these results, once a description of these networks is available at both the macro- and micro-levels, a broader aspiration would then later consist in reconstructing the observed epistemic structures by simulating a dynamic epistemic network, using assumptions designed after actual empirical measurements of agent-based behavior. This kind of exercise would be useful in showing that a co-evolutionary morphogenesis of epistemic networks constitutes a plausible explanation for the observed patterns, as shown in the case of zebrafish scientists by **?**.

Computations of ECs were achieved using *galois*, an open-source program freely downloadable from *http://code.google.com/p/networks-tb–galois*

Acknowledgments. We are extremely grateful to RTGI SAS/LINKFLUENCE for graciously letting us use and analyze their data. We particularly thank Nadine Peyriéras and Guilhem Fouetillou for their expertise on, respectively, the zebrafish community and the US political blogosphere dataset. We are also grateful to the anonymous referees for their constructive suggestions. This research has been partly financed by an ANR grant 'Webflunce' #SYSC-009-03.

Note: methods and qualitative results presented here on the "zebrafish" community case study are partially based on a *portion* of a previous *French* paper by one of us (CR) which has been published recently in the Revue Française de Sociologie (RFS), although we use here a slightly different dataset ranging over 1999–2006. Results from the political blogosphere case study are totally original (as obviously is the subsequent comparison), while the general theoretical framework has been thoroughly reshaped in order to take into account and refocus on the conceptual and structural similarities and differences induced by the two cases.

References

Adamic, L.A., Glance, N., 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In: LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery. ACM Press, New York, NY, USA, pp. 36–43.

Ali-Hasan, N., Adamic, L.A., 2007. Expressing Social Relationships on the Blog through Links and Comments. In: International Conference on Weblogs and Social Media.

Anderson, C.J., Wasserman, S., Faust, K., 1992. Building stochastic blockmodels. Social Networks 14, 137–161.

Barabási, A.L., Albert, R., 1999. Emergence of Scaling in Random Networks. Science 286, 509–512.

Barabási, A.L., Jeong, H., Ravasz, E., Neda, Z., Vicsek, T., Schubert, A., 2002. Evolution of the social network of scientific collaborations. Physica A 311, 590–614.

Batagelj, V., Bren, M., 1995. Comparing Resemblance Measures. Journal of Classification 12, 73–90.

Bourdieu, P., 1991. The Peculiar History of Scientific Reason. Sociological Forum 6, 3–26.

Bradbury, J., 2004. Small Fish, Big Science. PLoS Biology 2, 568–572.

Breiger, R.L., 1974. The Duality of Persons and Groups. Social Forces 53, 181–190.

Callon, M., 2001. Les méthodes d'analyse des grands nombres. Contribuent-elles à l'enrichissement de la sociologie du travail? In: A. Pouchet (Ed.), Sociologies du travail: quarante ans après, Elsevier, Paris. pp. 335–354.

Callon, M., Law, J., Rip, A., 1986. Mapping the dynamics of science and technology. MacMillan Press, London.

Cardon, D., Delaunay-Teterel, H., 2006. La production de soi comme technique relationnelle. Un essai de typologie des blogs par leurs publics. Réseaux .

Cohendet, P., Kirman, A., Zimmermann, J.B., 2003. Émergence, Formation et Dynamique des Réseaux – Modèles de la morphogenèse. Revue d'Economie Industrielle 103, 15–42.

Cole, J.R., Cole, S., 1973. Stratification in Science. University of Chicago Press.

Coleman, J., 1988. Social Capital in the Creation of Human Capital. American Journal of Sociology 94, S95–S120.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S., 2008. Feedback Effects between Similarity and Social Influence in Online Communities. In: KDD '08: Proc. of the 14th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining.

de Solla Price, D.J., 1965. Networks of Scientific Papers. Science 149, 510–515.

de Solla Price, D.J., 1976. A General Theory of Bibliometric and Other Cumulative Advantage Processes. Journal of the American Society for Information Science 27, 292–306.

Degenne, A., 2004. Les réseaux sociaux. Mathématiques et sciences humaines 168, 5–9.

Doreian, P., Woodard, K.L., 1994. Defining and locating cores and boundaries of social networks. Social Networks 16, 267–293.

Dorogovtsev, S.N., Mendes, J.F.F., 2003. Evolution of Networks — From Biological Nets to the Internet and WWW. Oxford University Press, Oxford.

Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of agency. American Journal of Sociology 99, 1411–1454.

Erdős, P., Rényi, A., 1959. On random graphs. Publicationes Mathematicae 6, 290–297.

Falzon, L., 2000. Determining Groups from the Clique Structure in Large Social Networks. Social Networks 22, 159–172.

Faust, K., Wasserman, S., 1992. Blockmodels: Interpretation and evaluation. Social Networks 14, 5–61.

Fienberg, S.E., Wasserman, S., 1981. Categorical data analysis of single sociometric relations. In: S. Leinhardt (Ed.), Sociological Methodology, Jossey-Bass, San Francisco. pp. 156–192.

Freeman, L.C., 1978. Centrality in Social Networks – Conceptual Clarification. Social Networks 1, 215–239.

Freeman, L.C., 1989. Social Networks and the Structure Experiment. In: L.C. Freeman, D.R. White, A.K. Romney (Eds.), Research Methods in Social Network Analysis, George Mason University Press, Fairfax, Va. pp. 11–40.

Freeman, L.C., 1996. Cliques, Galois Lattices, and the Structure of Human Social Groups. Social Networks 18, 173–187.

Freeman, L.C., White, D.R., 1993. Using Galois Lattices to Represent Network Data. Sociological Methodology 23, 127–146.

Gordon, M.D., 1980. A critical reassessment of inferred relations between multiple authorship, scientific collaboration, the production of papers and their acceptance for publications. Scientometrics 2, 193–201.

Haas, P., 1992. Introduction: epistemic communities and international policy coordination. International Organization 46, 1–35.

Hindman, M., Tsioutsiouliklis, K., Johnson, J.A., 2003. Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web. In: Annual Meeting of the Midwest Political Science Association.

Holland, P.W., Leinhardt, S., 1976. Local Structure in Social Networks. Sociological Methodology 7, 1–45.

Holland, P.W., Leinhardt, S., 1981. An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association 76, 33–65.

Jeong, H., Néda, Z., Barabási, A.L., 2003. Measuring Preferential Attachment for Evolving Networks. Europhysics Letters 61, 567–572.

Jin, E.M., Girvan, M., Newman, M.E.J., 2001. The structure of growing social networks. Physical Review E 64, 046132.

Jones, K.S., Kay, M., 1973. Linguistics and information science. Academic Press, New York.

Katz, J.S., Martin, B.R., 1997. What is research collaboration? Research Policy 26, 1–18.

Kitcher, P., 1995. Contrasting Conceptions of Social Epistemology. In: F. Schmitt (Ed.), Socializing Epistemology: The Social Dimensions of Knowledge, Rowman and Littlefield, Lanham, MD. pp. 111–134.

Latour, B., Woolgar, S., 1988. La vie de laboratoire – La production des faits scientifiques. Sciences humaines et sociales. La Découverte.

Laumann, E.O., Marsden, P.V., Prensky, D., 1983. The boundary specification problem in network analysis. In: R.S. Burt, M.J. Minor (Eds.), Applied Network Analysis, Sage, Beverly Hills. pp. 18–34.

Lazega, E., 1992. Micropolitics of Knowledge: Communication and Indirect Control in Workgroups. Aldine de Gruyter, New York, NY.

Lazega, E., Lemercier, C., Mounier, L., 2006. A Spinning top model of formal organization and informal behavior: dynamics of advice networks among judges in a commercial court. European Management Review 3, 113–122.

Lazega, E., van Duijn, M., 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: a logistic regression model for dyadic network data. Social Networks 19, 375–397.

Leenders, R.T., 1997. Longitudinal behavior of network structure and actor attributes: Modeling interdependence of contagion and selection. In: E. Doreian, E.N. Stokman (Eds.), Evolution of social networks, Gordon and Breach, Amsterdam. pp. 165–184.

Lehmann, S., Schwartz, M., Hansen, L.K., 2008. Biclique communities. Physical Review E 78.

Liben-Nowell, D., Kleinberg, J., 2003. The link prediction problem for social networks. In: CIKM '03: Proceedings of the 12th international conference on Information and knowledge management. ACM Press, New York, NY, USA, pp. 556–559.

Lind, P.G., Gonzalez, M.C., Herrmann, H.J., 2005. Cycles and clustering in bipartite networks. Physical Review E 72, 056127.

Lorrain, F., White, H.C., 1971. Structural Equivalence of Individuals in Social Networks. Journal of Mathematical Sociology 1.

Lotka, A.J., 1926. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16, 317–323.

McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M., Glance, N., 2007. Finding Patterns in Blog Shapes and Blog Evolution. In: International Conference on Weblogs and Social Media.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27, 415–444.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenhtat, I., Sheffer, M., Alon, U., 2004. Superfamilies of evolved and designed networks. Science 303, 1538–1542.

Newman, M.E.J., 2002. Assortative mixing in networks. Physical Review Letters 89, 208701.

Newman, M.E.J., 2004. Coauthorship networks and patterns of scientific collaboration. PNAS 101, 5200–5205.

Pattison, P., 1993. Algebraic Models for Social Networks. Cambridge University Press, New York.

Pattison, P., 1994. Social Cognition in Context: Some Applications of Social Network Analysis. In: S. Wasserman, J. Galaskiewicz (Eds.), Advances in Social Network Analysis: Research in the Social and Behavioral Sciences, Sage Publications. pp. 79–109.

Pattison, P., Wasserman, S., 1995. Constructing algebraic models for local social networks using statistical methods. Journal of Mathematical Psychology 39, 57–72.

Pattison, P., Wasserman, S., Robins, G., Kanfer, A.M., 2000. Statistical Evaluation of Algebraic Constraints for Social Networks. Journal of Mathematical Psychology 44, 536–568.

Powell, W.W., White, D.R., Koput, K.W., Owen-Smith, J., 2005. Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. American Journal of Sociology 110, 1132–1205.

Redner, S., 2005. Citation Statistics from 110 Years of Physical Review. Physics Today 58, 49–54.

Robins, G., Alexander, M., 2004. Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. Computational and Mathematical Organization Theory 10, 69–94.

Robins, G., Elliott, P., Pattison, P., 2001a. Network models for social selection processes. Social Networks 23, 1–30.

Robins, G., Pattison, P., Elliott, P., 2001b. Network Models for Social Influence Processes. Psychometrika 66, 161–190.

Roth, C., 2005. Generalized Preferential Attachment: Towards Realistic Socio-Semantic Network Models. In: ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis. Galway, Ireland, vol. 171 of *CEUR-WS Series (ISSN 1613-0073)*, pp. 29–42.

Roth, C., Bourgine, P., 2005. Epistemic Communities: Description and Hierarchic Categorization. Mathematical Population Studies 12, 107–130.

Roth, C., Bourgine, P., 2006. Lattice-based dynamic and overlapping taxonomies: The case of epistemic communities. Scientometrics 69, 429–447.

Salton, G., Wong, A., Yang, C.S., 1975. Vector Space Model for Automatic Indexing. Communications of the ACM 18, 613–620.

Snijders, T.A.B., 1981. The degree variance: an index of graph heterogeneity. Social Networks 3.

Snijders, T.A.B., 2001. The Statistical Evaluation of Social Networks Dynamics. Sociological Methodology 31, 361–395.

Snijders, T.A.B., Stokman, F.N., 1987. Extensions of triad counts to networks with different subsets of points and testing underlying random graph distributions. Social Networks 9, 249–275.

Soffer, S.N., Vázquez, A., 2005. Network clustering coefficient without degree-correlation biases. Phys. Rev. E 71, 4.

Sperber, D., 1996. La contagion des idées. Odile Jacob, Paris.

Subramanyam, K., 1983. Bibliometric studies of research collaboration: A review. Journal of Information Science 6, 33–38.

Valente, T.W., 1995. Network Models of the Diffusion of Innovations. Hampton Press.

Wasserman, S., 1980. Analyzing Social Networks As Stochastic Processes. Journal of the American Statistical Association 75, 280–294.

Wasserman, S., Weaver, S.O., 1985. Statistical Analysis of Binary Relational Data: Parameter Estimation. Journal of Mathematical Psychology 29, 406–427.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

White, D.R., Duquenne, V., 1996. Social Network & Discrete Structure Analysis: Introduction to a Special Issue. Social Networks 18, 169–172.

Wille, R., 1992. Concept lattices and conceptual knowledge systems. Computers Mathematics and Applications 23, 493.

Zhang, P., Wang, J., Li, X., Di, Z., Fan, Y., 2007. The clustering coefficient and community structure of bipartite networks. Arxiv preprint arXiv:0710.0117 .