Open access • Journal Article • DOI:10.1142/S0219525913500136

Socio-semantic frameworks — Source link

Camille Roth, Camille Roth

Institutions: School for Advanced Studies in the Social Sciences, Centre national de la recherche scientifique

Published on: 20 Oct 2013 - Advances in Complex Systems (World Scientific Publishing Company)

Socio-Semantic Frameworks

Camille Roth

Advances in Complex Systems

SOCIO-SEMANTIC FRAMEWORKS

CAMILLE ROTH
CNRS
Centre d'Analyse et de Mathématique Sociales (CAMS),
CNRS/EHESS, 190 avenue de France, Paris, F-75013, France
and
Centre Marc Bloch (CMB),
CNRS/HU/MAE, Friedrichstrasse 191, Berlin, D-10117, Germany
roth@cmb.hu-berlin.de

Received (received date)
Revised (revised date)

Socio-technical systems involve agents who create and process knowledge, exchange information and create ties between ideas in a distributed and networked manner: webloggers, communities of scientists, software developers and wiki contributors are, among others, examples of such networks. The state-of-the-art in this regard focuses on two main issues which are generally addressed in an independent manner: the description of content dynamics and the study of social network characteristics and evolution. This paper relies on recent endeavors to merge both types of dynamics into co-evolutionary, multi-level modeling frameworks, where social and semantic aspects are being jointly appraised. Case studies featuring socio-semantic graphs, socio-semantic hypergraphs and socio-semantic lattices are notably discussed.

Keywords: Socio-technical systems; socio-semantic networks; socio-semantic hypergraphs; socio-semantic lattices; coevolution.

1. Introduction

Contexts where agents collectively interact to exchange information, create ties between ideas, process and produce knowledge have recently been proliferating: blogs, wikis, open-source development platforms, social networking sites, tagging platforms and, more broadly, user-generated-content platforms.

These social systems shall be referred to by the term "*social-technical systems*", for they all share one core feature: interactions between agents occur in a relatively decentralized and autonomous manner and, to some extent, rely on information and communication technologies which eventually host, organize and facilitate large-scale processes of *social cognition*.[a]

[a] "Social cognition" should be understood here as socially-distributed information production and processing within a system of a generally large number of individuals (i.e. system-level cognitive processes relying on interacting humans), rather than the traditional cognitive psychology inter-

In this regard, science and its various subcommunities have long been the only large socio-technical system whose dynamics could be empirically appraised, and to which a whole discipline, namely scientometrics, has been devoted. For decades, it featured a seemingly archaic socio-technical apparatus revolving around cultural artifacts called "books", synchronized repositories called "libraries" and physical gatherings called "conferences". In this context, data collection and processing as well as result dissemination are being done in an asynchronous and distributed manner by scientific teams operating locally, on subproblems, with no central plan in mind.

The major and recent development of various socio-technical systems has been supported by a wide range of digital platforms, many of them entirely new, and has simultaneously led to the availability of large amounts of detailed data on their users' behavior. Social cognition not only flourished, it also became prone to systematic *in vivo* observation, which paved the way to empirically-founded agent-based modeling.

Two elements of these systems are essential to social cognition: on one side, agents interacting in diverse manners and, on the other side, a mesh of information "items" (texts, opinions, tags, and more broadly digital content). Additionally, the dynamics of information production/manipulation and the dynamics of interactions appear to obey to similar time-scales: virtually by design, content manipulation indeed involves interactions which contribute to shape future content creation which, in turn, influences the evolution of the social fabric; and so on. It is rather difficult to think of actual exceptions to this rule — that is to say, cases where the social structure could be considered constant meanwhile content evolves or, conversely, where the social structure evolves while the distribution of content remains static. Yet, this co-evolution remains rarely taken into account explicitly in the literature on descriptive (empirical) models of socio-technical systems.[b]

This paper will advocate the notion that the full appraisal of processes occuring within socio-technical systems, top and foremost social cognition phenomena, requires (agent-based) modeling frameworks which jointly feature social structure and semantic characteristics; that is, *socio-semantic* frameworks. After introducing some of the most relevant research streams in that regard (Secs. 2.1 & 2.2), I will therefore review the often distinct efforts aimed at empirically understanding either the social or the semantic dimensions of these socio-technical systems (Sec. 2.3), before eventually describing and synthesizing several recent endeavors to actually develop operational socio-semantic frameworks (Sec. 3).

pretation of social cognition describing individual-level cognitive processes in the context of human interactions (i.e. psychology of interactions).

[b]This shall not be the case for normative models, which we do not aim at addressing here (see for instance the network-based literature on cooperation where agents may rewire their neighborhood according to other agents' properties and past actions [31, 55]).

2. Understanding social cognition processes

Several research fields are dealing with issues touching to social cognition. I shall yet concentrate on three specific fields where it constitutes a focal point and which are thus highly relevant to the study of socio-technical systems.

2.1. *Social epistemology*

First comes social epistemology — perhaps the only area of research to explicitly and almost exclusively address the conditions of the collective production and foundation of knowledge. Its most theoretical ramifications deal with the characterization of collective knowledge (defining for instance a given proposition p as "community knowledge" *iff* agents know p and know that others know p and trust them [36]) and typically overlap with epistemic logic (which involves, however, little empirical modeling – see nevertheless [62], for instance, for a normative application to agent-based models). The more sociological ramifications focus on the social factors behind the construction and adoption of knowledge, regarding for example the origin of consensual or authoritative statements [40]. These works basically question the influence of the bias induced by agents, voluntarily or not, on the processing of information and its evaluation by a given social group. Here, science appears to be a natural prototype [37], with early sociological studies dealing with the joint dynamics of knowledge and social organization [39].

This sociological stance leads in particular to the study of the social procedures pertaining to the organization of cognitive labor. This is more closely connected to socio-technical systems because of the specific emphasis on the role of the technological environment: Hutchins [34], for one, exemplified the notion of "distributed cognition" by showing that the successful piloting of a ship to seaport requires a distributed effort where all parties, agents and devices alike, have to play a local role — science also represents a typical case, again, and is interestingly illustrated in the so-called "actor-network theory" ontology [12], where scientific agents and artifacts are indistinctly gathered into a hybrid network.

2.2. *Cultural anthropology*

The area of cultural anthropology shares similar high-level goals with social epistemology, yet with a specific focus on the emergence of culture and cultural similarity: in other words, "explaining the capacity of some representations to propagate until becoming precisely cultural, that is, revealing the reasons of their contagiosity" [42]. The articulation with social cognition appears even more evidently when culture is defined as "acquired information, such as knowledge, beliefs, and values, that is inherited through social learning, and expressed in behaviors and artifacts" [49].

Clearly, as it addresses the conditions surrounding the propagation and reproduction of knowledge and representations, this research program is at the same time more precise than the issue of social cognition, and it applies to more general contexts than just socio-technical systems. On the modeling side, its implications are

however more focused. Memetics [22] in particular has long been seen by modelers as an efficient and naturalistic framework for understanding cultural convergence (see for instance [18]), but it raised doubts from the side of cultural anthropology itself, concerning in particular the assumption that there exist atomic (cultural) representations and high-fidelity replication. The theory of cultural epidemiology/contagion defended by Sperber [65] received a wider anthropological support, by clarifying the underlying cognitive processes and, notably, emphasizing both the role of reformulation and the existence of aggregates of cultural representations rather than cultural "atoms" (thereby extending the Levi-Straussian notion that a "myth" is "the set of all of its versions"). In any case, here again, modeling efforts, although convincing, have generally remained less descriptive than normative (see e.g. [15]) — and, perhaps more importantly, they usually gave more importance to the dynamics of content/representations than to the underlying social structure, which is yet at the core of socio-technical systems.

On the whole, while the two above disciplines shall be able to provide conceptual guidance on the dynamics of socio-technical systems, they seem yet to have yielded a relatively limited literature on empirical models, as they remain essentially focused on the theoretical aspects or the qualitative understanding of the aforementioned processes.

2.3. *Social complex systems and social cognition*

We finally turn to the more recent stream of research on social complex systems, and in particular to an area increasingly denoted as "social computing" [70]. This domain possibly sustains one of the strongest connections with the previous fields. It is also generally much more empirically-minded and sometimes essentially fueled by the availability of large datasets detailing the *in vivo* traces of human behavior — fashionably called "big data", and stemming from sources as varied as government agencies (such as public health, economic or bibliographical records), companies (regarding consumer behavior, including merchant, transit network or cell phone data) or online services in various contexts (discussion forums, wikis, blogs, etc.). As such, social computing naturally overlaps with the empirical study of social cognition, yet as a by-product of a more general aim of understanding human dynamics in the broad sense: it indeed addresses a variety of issues including social sensing,

- either by considering agents as a distributed set/network of sensors informing us on the state of a given social system (Google FluTrends being a simple yet iconic example [28]) or enabling us to predict its future state [5],
- or by offering the opportunity to uncover human behavior through social dynamics both at the global (see e.g. human mobility [30]) and local level (see e.g. the description of voting behavior from Wikipedia election histories [44], or the preliminary observation of the bias induced in the copy-pasting of quotations within blogspace [63]).

This effort obviously relates to a larger body of knowledge on social complex systems, dealing on one hand with social interactions at large, and with content dynamics analysis on the other hand.

Characterizing social interactions. Scholars familiar with the literature on complex systems would probably be already aware of the pretty large amount of knowledge that has been assembled in this area. Social networks have had a long history of research, starting with the pioneer mathematical sociology endeavors extending over the 1940s-1990s. This period was rather focused on "small" case-studies, with datasets describing social interactions of groups of less than a hundred people, more often a few dozens, and introducing most of the key formal (algebraic) frameworks of today (centralities, random networks comparisons, behavioral inference, community detection) — the classic book of Wasserman & Faust [71] reviews the already rich state-of-the-art of this field as of the mid-1990s. The later years have witnessed a growing interest in large-scale studies (see e.g. the following overviews [4, 8, 25, 54]), as part of a larger effort on "complex networks" mobilizing disciplines traditionally labeled as "hard" or "natural" sciences, especially statistical physics and computer science — fueled by the initial observations that empirical networks were rather heterogeneous (with keywords such as "scale-free", "power-law", etc.) and structured (with keywords such as "clustered", "small-world") and revisiting and improving the earlier mathematical sociology concepts on much larger datasets.

On the whole, it is reasonable to claim that these two overlapping (and now merging) streams of research have achieved today a good characterization of empirical social networks, both statically and dynamically. Classical stylized facts are well-described in a large number of different contexts (connectivity, transitivity, patterns, topological communities, *inter alia*) and, after the initial all-purpose, universal models targeting the reconstruction of the ubiquitous heterogeneous degree distribution observed in almost all systems, realistic morphogenesis models are successfully being proposed in increasingly specific case studies in order to explain increasingly specific patterns (such as, for instance, agent-based models based on blog posting behavior in order to reproduct the temporal features of diffusion cascades [29]).

Characterizing informational dynamics. The quantitative description of content or representation dynamics is a relatively newer endeavor. This field still tends to appraise social cognition in its simplest form — the spatio-temporal usage of terms or aggregates of terms — but nonetheless made significant progress in the last few years when it turned to large datasets originating from predominantly online socio-technical systems. Studies focusing on terms or n-grams are not far from signal analysis, distinguishing for instance spikes vs. chatter [32] or differentiating source type or location [45], describing vocabulary dynamics [13] and eventually predicting usage by exploiting behavioral regularities over time [6, 51]. A few studies make use of more sophisticated notions of information, beyond n-grams. For

instance, [43] looked at the existence of aggregates of slightly-varying sentences by describing clusters of relatively similar quotations stemming from the same original utterance by a public figure in a large corpus of blog posts. Shortly thereafter, the same dataset made it possible to characterize the underlying low-level social cognition processes of sentence reformulation [63] — a result which would probably fulfill some of the preliminary objectives of cultural epidemiologists. Note here that studies on science, once again, were early in proposing automatic history reconstruction methods based on clusters of terms, for instance through the now-classical co-word analysis [11].

3. Socio-semantic frameworks

The modeling of socio-technical systems is at the junction of these disciplinary efforts. Such social systems usually involve semantic interactions. Modeling-wise, as underlined above, we have a relatively established body of research describing either the topology of interactions or the dynamics of information. Yet, at this stage, a strictly structural viewpoint may overlook a large part of the drivers of interactions. In parallel, when there is a focus on representations, i.e. on the rather semantic side, there is little on the underlying relational structure.

Admittedly, when it comes to ICT-mediated systems, the technical arrangement of the underlying interaction platforms, both in terms of social engineering (interaction modes and possibilities) and regarding the design of the conceptual ontology (channeling more or less sophisticated representations, from opinions and likes/dislikes, to tags, sentences, or documents), does have an impact on the subsequent social cognition processes. Equally important, we contend, is the joint appraisal of these dimensions.

To make this point we shall present recent situations where we introduced a socio-semantic framework to understand social cognition processes, detailing altogether the micro-, meso- and macro-level dynamics of socio-technical systems — focusing in particular on cases where the perspective is being enriched, sometimes even changed, by the introduction of a socio-semantic framework.

3.1. *Micro: Socio-semantic networks*

The issue of a co-evolution, or at least a correlation, between social and semantic aspects has been posed by mathematical sociology a couple of decades ago, which already emphasized the fact that semantic aspects in interactional processes become expressly pertinent when knowledge and relationships evolve at a similar pace [24, 41]. Of particular interest is the statistical modeling framework of Snijders et al. on the empirical evaluation of the coevolution of structure and behavior [64], where the contribution of both behavioral and structural properties in the formation of new links is being estimated within a single model.

On the complex system modeling side, attempts are more recent and follow the observation that features in one of the realms could be correlated with features from

the other realm. This more precisely means,

- uncovering the social signatures of semantic differences or similarities. For instance, [1] shows that the social network of US political blogs is structurally segregated according to a simple binary semantic indicator — Democrat- or Republican-leaning bloggers — while [46] demonstrates that patterns of citation cascades differ from a given topical community to another. More broadly, the literature on homophily at large substantiates the existence of an impact from the semantic to the social (see [47] for a standard overview, or [61] for a more recent and social computing-oriented example).
- or uncovering the semantic signatures of social differences. This could consist in showing that users with a more focused semantic profile would get more citations/incoming links on Twitter [14] or that they would produce better quality output, where "quality" is an evaluation aggregated from the opinion or votes of their fellows on a system-wide scale [2].

On the whole, these social-to-semantic or semantic-to-social studies show that partial information could be gained from one dimension about the other, albeit this would not entirely qualify as socio-semantic morphology per se, where both aspects would be jointly appraised.

The empirical modeling of socio-semantic morphogenesis seems to have started only recently: for instance, [19] empirically describes social and semantic coevolution by examining the semantic patterns of user-to-user interaction in Wikipedia through discussion pages. More precisely, this work finds a compact yet enlightening description of joint socio-semantic evolution: they specifically show that after a first interaction, profiles are getting semantically closer, given a proper semantic similarity measure, following an exponential pace around the time of the interaction. They also show that the same pattern occurs *before* the interaction. Here, the sigmoid function is an original socio-semantic pattern, beyond homophily-like predictions on interaction as a function of similarity.

Socio-semantic blog networks.

Going further into this direction, I first wish to relate a recent series of socio-semantic studies by us, focused on blogspace [16, 17, 58], where agents interact through blog posts, discuss a variety of topics and cite other blogspace agents. The corresponding ontology is a hybrid socio-semantic network gathering agents and semantic "items", where links are being made everytime an agent cites another agent (directed links) or mentions an issue (undirected link between agent and issue). More formally, the agent set $\mathbf{A} = \{a_1, a_2, .., a_n\}$ describes distinct individual blogs of the dataset. The data itself is based on dated blog posts collected by the social web content analysis company "Linkfluence" on a fixed perimeter of a librarian-curated set of 1,066 blogs essentially commenting the US presidential

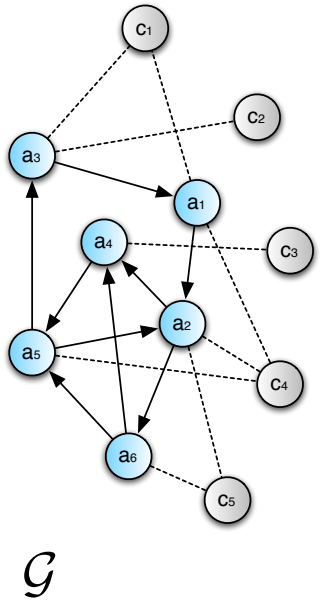


Fig. 1: Ontology of the socio-semantic network $\mathcal{G}$: "a"-nodes ($\in \mathbf{A}$) denote agents (blogs), while "c"-nodes ($\in \mathbf{C}$) denote concepts; links correspond either to citation (directed) or usage (undirected).

election and being active on a period of 4 months starting on Nov 1st, 2008. The relational network is built from citation links. Issues, or concepts, are collected following a basic Natural Language Processing (NLP) task aimed at isolating the most frequent of the meaningful n-grams in post contents, i.e. excluding stop-words (such as "and", "or", "then") and detecting likely nominal groups made of one or two words. A selection of terms among the most frequent yet meaningful groups of terms eventually yields a concept set $\mathbf{C} = \{c_1, c_2, .., c_{n'}\}$ of 80 items. In the context of political blogs, this means keeping concepts such as "climate change", "national security", "immigrati" (as a lemma), while discarding "top issue" or "important debate". This finally defines a socio-semantic network or graph $\mathcal{G}$ whose node and edge sets are respectively $V_{\mathcal{G}} = \mathbf{A} \cup \mathbf{C}$ and $E_{\mathcal{G}} \subseteq (\mathbf{A} \times \mathbf{A}) \cup (\mathbf{A} \times \mathbf{C})$ — see Fig. 1. Nodes are either agents or concepts, and edges correspond to directed connections either from a citing agent to a cited agent, or from an agent to a concept she used. The network is additionally dynamic, in the sense that $\mathcal{G}_t$ denotes the cumulated state of the network for all links appearing up to t (as a result, $E_{\mathcal{G}_t} \subseteq E_{\mathcal{G}_{t'>t}}$). Unless otherwise noted, the dataset has concretely been divided into 8 successive periods of 14 days, so that $t \in \{1, .., 8\}$.

Some traditional observations may first be basically extended from the social to

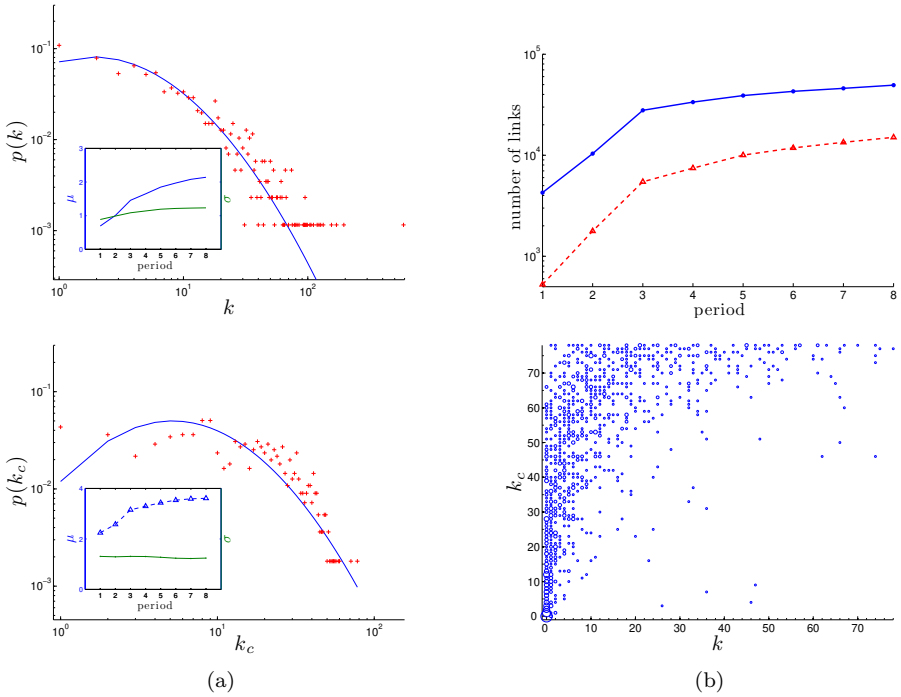


Fig. 2: **(a)** Distribution of the number of agents having a social degree k (*top*) and semantic degree $k_{\mathbf{C}}$ (*bottom*) during the last observation period. Insets represent the evolution of the mean μ and standard deviation σ over the 8 periods. **(b)** On *top*, growth of the network in terms of agent nodes and agent-to-concept edges. *Bottom*, correlation between social and socio-semantic capitals (degrees), respectively denoted k and k_c. From [58].

the socio-semantic network. Such is the case of the hierarchical configuration [58]. Distributions of the quantity of edges per node exhibit indeed a strong heterogeneity in both the social and socio-semantic networks: few have a capital of many links and use many concepts, many have few links and use few concepts. This is depicted on Fig 2a-top for the distribution of the social capital k (or social degree, i.e. number of neighbors in $\mathbf{A}$) and on Fig 2a-bottom for the semantic capital $k_{\mathbf{C}}$ (or semantic degree, i.e. neighbors in $\mathbf{C}$). Both insets describe the evolution of the parameters of the distribution, which is equally well fitted over each period t and has a relatively constant standard deviation σ, in spite of an increasing mean number of links μ. In other words, the shape of this stratification is dynamically stable in both networks, in spite of a vigorous low-level dynamics consisting of the appearance of a large number of new nodes and links (see Fig. 2b, top). Mutual constraints between the two aspects may however already be noticed: for instance, it seems to be hardly

possible for a blogger to have a large social capital without using many different concepts (see Fig. 2b, bottom). Cohesion too — i.e., the existence of local aggregates — may be described, in both networks, as being remarkably strong, relatively to a typical random case. In particular, the proportion of closed triads around a_i is:

$$c_{\mathbf{A}}(a_i) = \frac{|\{(i', i'') \text{ s.t. } \{(a_i, a_{i'}), (a_i, a_{i''}), (a_{i'}, a_{i''})\} \subset E_{\mathcal{G}}\}|}{|\{(i', i'') \text{ s.t. } \{(a_i, a_{i'}), (a_i, a_{i''}))\} \subset E_{\mathcal{G}}\}|} \tag{1a}$$

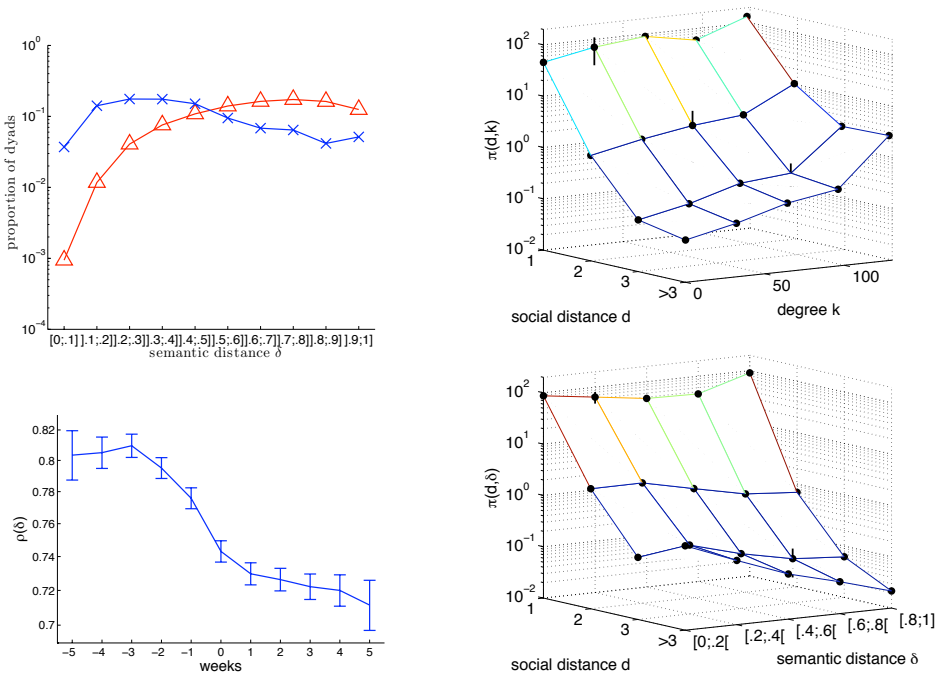
It is found to be on average equal to 15.8%, in general one order of magnitude larger than in a uniform random network with the same number of agents and social links. Yet, the bipartite characterization of socio-semantic aggregates demands the definition of cohesion patterns of a slightly higher order: namely, by going from triads of connected agents (or concepts) $\{a_i, a_{i'}, a_{i''}\}$ to quaternions of pairs of agents (or concepts) who are jointly linked to pairs of concepts (or agents), $\{a_i, a_{i'}, c_j, c_{j'}\}$. Such quaternions thereby characterize how much two agents connected to a same concept are likely to share other concepts (or dually, how concepts used by a same agent are likely to be jointly used by other agents). For a given agent a_i, the quaternion proportion is:

$$c_{\mathcal{G}}(a_i) = \frac{|\{(i', j, j') \text{ s.t. } \{(a_i, c_j), (a_{i'}, c_j), (a_i, c_{j'}), (a_{i'}, c_{j'})\} \subset E_{\mathcal{G}}\}|}{|\{(i', j, j') \text{ s.t. } \{(a_i, c_j), (a_{i'}, c_j), (a_i, c_{j'})\} \subset E_{\mathcal{G}}\}|} \tag{1b}$$

Except for the first two observation periods, this coefficient is found to be between 50 and 75%, and thus around one order of magnitude higher than in the uniform random case as well (same number of nodes and links of both types).

Beyond pairs of agents and concepts, one may consider socio-semantic cohesion at a higher level than triads-quaternions by comparing the whole conceptual neighborhoods of given pairs of agents. This comes down to describing semantic homophily in the social network. We therefore introduce measures of semantic dissimilarity between agents. To meaningfully do this type of comparison, we need to factor in the varying term occurrence frequencies. In other words, we shall assume that use frequency matters to describe semantic profiles of bloggers: less used terms (in the corpus) should potentially help discriminate two agents to the same extent as more used terms, rather than being negligible in comparison. Formally, our definition for the semantic dissimilarity δ relies on the classical *tf.idf* framework [60], where concept occurrence frequency (term frequency *tf*) in a blogger's post production is divided by its occurrence frequency in the whole corpus (inverse document frequency *idf*). This eventually defines for each agent a (weighted) vector of *tf.idf* scores on all concepts. Then, the semantic dissimilarity $\delta(a_i, a_j)$ between agents a_i and a_j is simply 1 minus the cosine of the angle between the *tf.idf* vectors of a_i and a_j (it varies from 0 for agents with identical vectors to 1 for strictly orthogonal vectors).[c] We observe that in blogspace, socially connected agents are more

[c]In more detail, at a given time t, $\delta_t(a_i, a_j) = 1 - \cos(\widehat{W_t(a_i), W_t(a_j)})$ where $W_t(a_i)$ is the *tf.idf* vector of a_i at t. In turn, the mth-coefficient of the *tf.idf* vector $W_t(a_i)_m$ is equal to



(a) (b)

Fig. 3: **(a)** Semantic proximity of connected dyads: on *top*, comparison between the semantic dissimilarity of connected dyads (blue crosses) and pairs of nodes in the whole network (red triangles); *bottom*, evolution of the average relative semantic dissimilarity $\rho(\delta)$ over all pairs of nodes getting connected for the first time at week 0. **(b)** Joint socio-semantic interaction propensity with respect to topological distance d and social capital k (*top*) or semantic dissimilarity δ (*bottom*). From [17].

often semantically similar, as illustrated on Fig. 3a-top by the comparison of the distribution of semantic dissimilarity for connected vs. all dyads.

More interesting is the examination of the low-level socio-semantic behavior behind these *a posteriori* descriptions, once networks have been formed. In other words, what are the *a priori* drivers of their evolution? To this end, we introduce the notion of interaction propensity: put simply, propensity measures the preference bias towards certain kinds of interactions, with respect to a baseline where all interactions would be equally likely. Formally, propensity should describe the

$\frac{w_t(a_i)_m}{\sum_{m'=1}^{n'} w_t(a_i)_{m'}} \cdot \log(\frac{n}{|\{j, w_t(a_j)_m > 0\}|})$. The "log" part relates to the inverse ratio of the number of blogs where term m appears over the total number of blogs. Note that using a Jaccard coefficient yields the same qualitative results.

relative likelihood of appearance of links between some types of dyads. For a dyadic property "d" (e.g. a distance), we thus compute the following quantity based on conditional probabilities of "L" (appearance of an interaction) given a certain value of d, over a particular period of time:

$$\pi(d) = \frac{P(\mathsf{L}|d)}{P(\mathsf{L})} = \frac{P(d|\mathsf{L})}{P(d)} = \frac{\nu(d)}{\nu} \cdot \frac{n}{n(d)} \tag{2}$$

where $\nu(d)$ describes the number of new interactions appearing between dyads of type "d" over that period of time (ν being the total number of new interactions), while $n(d)$ describes the number of such potential dyads ($n = n_{\mathbf{A}}$ being the number of nodes). Variables may be combined, for instance when looking at the joint propensity $\pi(k, d)$ for a topological distance d and a social capital k by considering dyads at distance d whose target has degree k.

Interactions, for one, appear to be shaped by the social structure: the combined propensity graph on Fig. 3b-top shows that more connected agents benefit proportionally from new connections, as is traditionally observed and plainly denoted as preferential attachment (social capital, denoted by social in-degree k, matters); besides, topologically closer agents are exponentially more likely to attract new connections (topological distance between agents, d, matters). Interestingly, social capital appears to matter more when social distance is higher: in the local neighborhood of repeated or "friend-of-friend" interactions, capital matters less. Interactions are additionally shaped by the semantic structure, as more similar agents are more likely to establish a relationship: see Fig. 3b-bottom which describes the magnitude of the interaction propensity with respect to both social distance d and semantic dissimilarity δ. As underlined before, we see that the propensity is much smaller for agents located in distant areas of the network. Moreover, propensity decreases with semantic dissimilarity whenever agents have not interacted before (i.e. for $d > 1$). Yet, the joint computation of propensity with respect to social and semantic features reveals that, in the case of repeated interactions ($d = 1$), semantic homophily plays a much weaker role. This points to the existence of two types of interaction modes (local and distant), with distinct socio-semantic processes [17]. Besides and more broadly, this suggests that the perspective is not only enriched, but also changed by the introduction of this socio-semantic framework, presenting the online world as a local rather than small world.

This phenomenon which may be interpreted as a joint social and semantic contraction of the network also exhibits a specific timeline: it goes on several weeks *after* interaction and, surprisingly, it also appears to start several weeks *before*, as Fig. 3a-bottom shows. This graph corresponds to the evolution of the *average relative* semantic dissimilarity $\rho(\delta)$ around the appearance of a social link (temporally translated as week "0" here, irrespective of the actual underlying period — note that we also exceptionally use a granularity of one week instead of two, to allow for a more precise description). It is measured *relatively* to the semantic dissimilarity in the network: a value of $\rho(\delta) = 0.8$ indicates that the observed dyad is 20% more

similar than all other pairs of network nodes. It is further *averaged* on all actual interactions (in the new temporal referential). In effect, the social and semantic closeness of interacting agents increases after *and* before they become connected.[d] In parallel, although not detailed further here because of space constraints, it is possible to use the same data to show that the distribution and dissemination of content appears to be influenced by the prior presence of similar content and by the social structure [16].[e] On the whole, blogspace is thereby appraised as a socio-technical system where content distribution is affected by content and structure, while structure is affected by content and structure as well.

3.2. *Meso: Socio-semantic collectives*

Several socio-technical systems feature horizontal, self-organized team work. In the particular context of intellectual production, individuals more or less freely decide to gather in teams to produce knowledge (science, open-source software, wikis, etc.): social cognition occurs not only at the macro-level of the whole system, but also at the meso-level of teams. People aspire at best choosing a team of collaborators or partners to achieve a creative project (paper writing, software development, etc.) — at building up quality collectives, whose characteristics and, especially, quality is a complex mix of the skills of the underlying collective, of its social arrangement and cognitive affinities.

We touch here the limits of network-based frameworks which focus on the individual or dyadic level: some characteristics are expressable at the meso-level of the team only and team formation processes are not a sum of individual rationalities. There is currently an increasingly developing body of work whose goal is to describe and model teams explicitly. This field has been loosely gathered under the term of "team science" [66]. The seminal study of Ruef [59] shows how several factors including gender, status, or ethnicity, influence the propensity to compose a team of entrepreneurs founding companies.

Here, hypergraphs appear to be an appropriate modeling framework: they generalize graphs in that hyperlinks gather an arbitrary number of nodes/agents, and not just two, by design, as is the case for graphs. Team work in socio-technical

[d]Note also the work of Aiello et al. [3] on aNobii (a book rating/tagging social network site) and principally focused on homophily, in terms of book library similarity (and geographic proximity) with respect to structure (principally topological distance in the social network). They essentially find that topologically closer users are more similar and, relatively surprisingly for a virtual community, that they live in closer geographical areas. They also notice that the semantic similarity gets reinforced after two users get linked.

[e]More precisely, content is more easily disseminated when it originates from socially better-connected agents (yet in a non-linear fashion: an agent with few connections has the same influence as one with none, while strongly connected or extremely connected agents also have the same strong influence – in other words, there are plateaus on both ends of the connectivity values) [16]. Content is also more easily disseminated when it goes through relatively "intermediate" users who connect remote (yet not too remote) areas of the social network, as measured by an index similar to betweenness centrality: a medium centrality value is optimal for content dissemination.

systems also depends on cognitive properties: teams are formed according to both social and semantic features. Elaborating on the previous section, I will now suggest that socio-semantic hypergraphs represent a very natural framework to appraise the joint socio-semantic dynamics of collectives.

Socio-semantic scientific hypernetworks.

In the specific case of academic teams [23, 48], quantitative and formal frameworks have traditionally been based on multi-dimensional surveys [20] and graphs [52, 53]. In a recent study [68], by contrast, we assert the relevance of socio-semantic hypergraphs.

Let us elaborate on its key points. Formally, we define a socio-semantic hypergraph $\mathcal{H}$ whose node set is, again, $V_{\mathcal{G}} = V_{\mathcal{H}} = \mathbf{A} \cup \mathbf{C}$ and whose hyperedge set is $H_{\mathcal{H}} \subseteq \mathcal{P}(V_{\mathcal{H}})$, the power set of $V_{\mathcal{H}}$. The empirical work is based on a large dataset of bibliographical records stemming from specific (biological) scientific fields — namely, papers from the publicly-available *Medline* database and dealing with "rabies", the model animal "zebrafish", or based on members of two joint FAO/WHO expert groups, "JEMRA" and "JEFCA", over the period 1985-2007 (yielding between $4,648$ and $8,685$ papers each). Records are in line with the socio-semantic hypergraphic ontology: they provide a set of individual author names $\mathbf{A}$ (from $9,684$ to $21,195$ distinct items), associated with abstracts which may be processed using simple NLP techniques similar to the ones used to extract data from blogs, yielding a set $\mathbf{C}$ of around 69–85 relevant concepts for each dataset. The dynamic hypergraph $\mathcal{H}_t$ is growing through the cumulative addition of hyperlinks $h \in H_{\mathcal{H}_t}$ describing authors and scientific concepts which "participated" in the same collaboration event, that is, a published paper at time t (see Fig. 4 — here again, $H_{\mathcal{H}_t} \subseteq H_{\mathcal{H}_{t'>t}}$). Time is discrete and corresponds to publication years (from $t = 0$ for 1985 to $t = 22$ for 2007).

Simple hypergraphic measures may be defined, at any time, depending on the past arrangement of socio-cognitive teams. For instance, the *socio-semantic expertise ratio* of a hyperlink h in a given concept c at time t, noted $\xi_{c,t}(h)$, denotes the number of agents of h who already appeared in at least one past hyperlink containing c before t. That is,

$$\xi_{c,t}(h) = \frac{|\{a \in h \cap \mathbf{A} \text{ s.t. } \exists h' \in H_{\mathcal{H}_{t-1}},\, \{a, c\} \subseteq h'\}|}{|\{h \cap \mathbf{A}\}|} \tag{3a}$$

Going further, the degree of social originality of a hyperlink h at time t, or *social hypergraphic repetition ratio*, may be measured by counting the proportion of subsets of agents of h which were already included in a past hyperlink at $t' < t$: it goes from 1 when all agents were previously all together in at least one collaboration, to 0 when the team does not even feature a single pair of previously interacting scientists. In other words,

$$r_t^{\mathbf{A}}(h) = \frac{|\{h' \in \mathcal{P}(h \cap \mathbf{A}) \text{ s.t. } \exists h'' \in H_{\mathcal{H}_{t-1}}, h' \subseteq h''\}|}{|\mathcal{P}(h \cap \mathbf{A})|} \tag{3b}$$

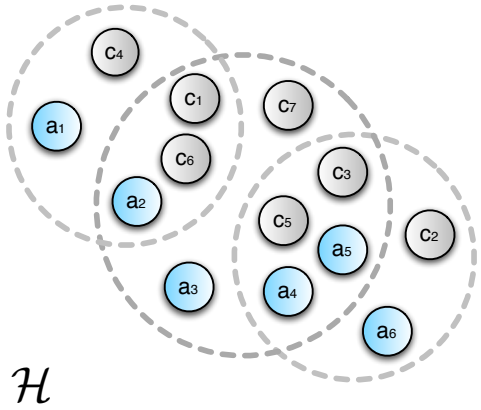


Fig. 4: Toy example of a socio-semantic hypergraph $\mathcal{H}$. Nodes represent either agents from $\mathbf{A}$ or concepts from $\mathbf{C}$. The boundaries of three partially-overlapping socio-semantic hyperlinks are figured by thick dashes: the top-left hyperlink, for instance, gathers $\{a_1, a_2, c_1, c_4, c_6\}$.

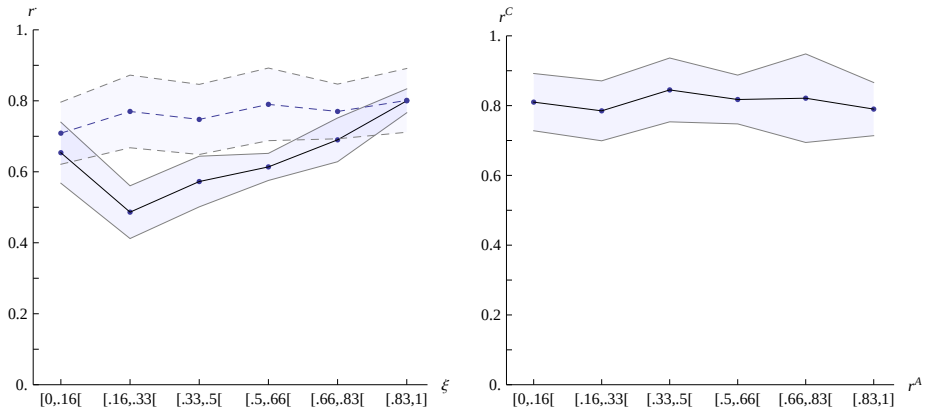


Fig. 5: Average relationship: **(a)** between expertise ratio (x-axis, ξ) and social ($r^{\mathbf{A}}$) or semantic ($r^{\mathbf{C}}$) ratios (y-axis, respectively solid and dashed lines); or **(b)** between social hypergraphic repetition ratio $r^{\mathbf{A}}$ (x-axis) and average semantic repetition ratios $r^{\mathbf{C}}$ (y-axis). Results are averaged over all four datasets, tubes indicate standard deviations.

Conceptual originality may be dually measured by a *semantic hypergraphic repetition ratio* $r_t^{\mathbf{C}}$ based on concepts, i.e. replacing $\mathbf{A}$ with $\mathbf{C}$ in the previous formula.

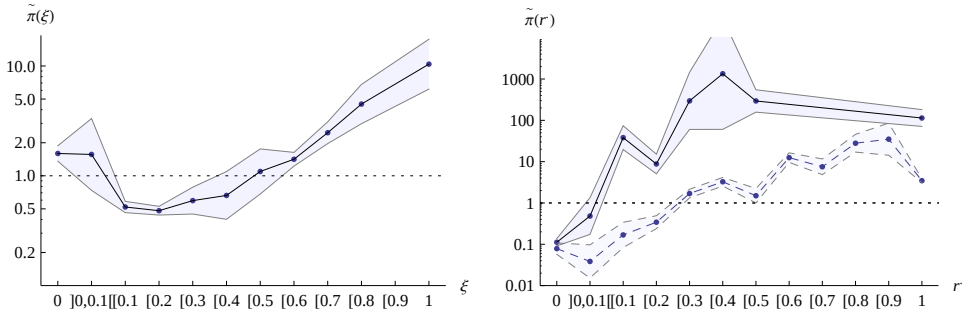


Fig. 6: Hypergraphic propensity (team bias): **(a)** for the proportion of experts per article; **(b)** for hypergraphic repetition ratios (social in solid line, semantic in dashed line). (Geometrically averaged behavior based on all four datasets.)

These indices make it possible to describe the *a posteriori* composition of teams, in terms of the raw distribution of teams exhibiting a given expertise ratio ξ or given social or semantic hypergraphic repetition ratios $r^{\mathbf{A}}$ and $r^{\mathbf{C}}$. More interesting is the correlation between these properties, as they may indicate whether the level of expertise in a team is related to a certain conceptual or social originality. On Fig. 5a, we first see (solid line) that social originality is lower on both ends of expertise, i.e. for teams made of experts only or made of non-experts only ($r^{\mathbf{A}}$ is closer to 1 when ξ is closer to either 1 or 0); hyperlinks of a mixed level of expertise correspond on average to a more original gathering of individuals. On the other hand (dashed line), there seems to be no correlation between the expertise ratio and semantic originality. Additionally, and perhaps contrarily to intuition, the absence of visible correlation on Fig. 5b indicates that new semantic associations (lower $r^{\mathbf{C}}$) do not correspond more to original teams (lower $r^{\mathbf{A}}$) than to repeated teams — in other words, conceptual originality does not seem to be related to an original social composition of the underlying team.

However, they also provide a key insight on (academic) team assembly mechanisms when compared with a random baseline. The simplest null-model consists of an evolving hypergraph featuring artificial hyperlinks conserving the same number of agents and concepts, but arranged in an arbitrary manner. More precisely, the empirical hypergraph is growing from $\mathcal{H}_t$ to $\mathcal{H}_{t+1}$ through the addition of a subset of hyperlinks $\Delta_{\mathcal{H}_t}$ describing the teams formed during time step t. To measure the possible bias in the formation of these teams, we could in principle use a formula similar to Eq. 2 adapted to a hypergraphic setting. However, the computation of the distribution of the hypergraphic characteristics of the $2^{n_{\mathbf{A}}+n_{\mathbf{C}}}$ potential hyperlinks is generally intractable. We thus recourse to a simulation-based model based on a synthetic subset of additional hyperlinks $\widetilde{\Delta_{\mathcal{H}_t}}$ at t such that it contains the

same number of hyperlinks, with the same size in terms of agents and concepts as is empirically observed in $\Delta_{\mathcal{H}_t}$. In other words, the joint social/semantic distribution of subset sizes $(|h \cap \mathbf{A}|, |h \cap \mathbf{C}|)$ for all teams h in $\Delta_{\mathcal{H}_t}$ is conserved in $\widetilde{\Delta_{\mathcal{H}_t}}$, while the exact composition of the synthetic hyperlinks is being uniformly randomized, in terms of specific agents or concepts.

Then, the discrepancy between what empirically happens and what the model predicts yields an estimate of the propensity of team formation, pretty much like the interaction propensity in blogspace. More precisely, the hypergraphic bias or propensity $\tilde{\pi}$ for a given hypergraphic property x may be measured at each simulated period t as:

$$\tilde{\pi}_t(x) = \frac{|h \in \Delta_{\mathcal{H}_t} \text{ s.t. } h \text{ is of type } x|}{|h \in \widetilde{\Delta_{\mathcal{H}_t}} \text{ s.t. } h \text{ is of type } x|} \tag{4}$$

With this in mind, we first plot the hypergraphic propensity with respect to the expertise ratio. Figure 6a describes the team formation bias averaged over all concepts, and then, because propensity is a ratio, geometrically averaged over all corpuses (assuming that we are measuring a similar underlying scientific behavior). We observe a strong socio-semantic preference for groups made of a high proportion of experts (ξ close to 1), as well as a preference for teams made on non-experts (ξ close to 0, i.e. teams where all or almost all members started working on a given concept for the first time). By plotting the bias with respect to social and semantic hypergraphic repetition ratios (Fig. 6b), we further observe a significant social and semantic confinement: there is a high likelihood to repeat previous collaborations patterns (high propensity for high $r^{\mathbf{A}}$), while the hypergraphic arrangement of concepts by a given team depends largely on the repetition of previous associations (globally growing propensity with respect to $r^{\mathbf{C}}$). On the whole, the hypergraphic micro-dynamics are tilted towards repetition. Knowing this not only opens the way to descriptive models of the meso-level evolution of such socio-technical systems, but also enables the development of normative models suggesting incentives for the formation of this or that type of teams, assuming that some types of teams should be favored over some others because they appear to produce better quality output (for instance by correlating those socio-semantic hypergraphic characteristics with quality measures such as citation counts).

3.3. *Macro: Socio-semantic phylogenies*

In general, meso-level features remain relatively unexplored, even outside of a socio-technical perspective. On the contrary, there is a long history of research on macro-level characterizations, especially when it comes to community structure detection. Here again, studies focusing on scientific dynamics have paved the way, co-mapping individuals or journals and fields of knowledge (see e.g. [38] or more recently [56], among many others). From a broader viewpoint, determining the success of community detection from a given social structure may generally be roughly assimilated to

a socio-semantic mapping operation. Social aggregates, indeed, are deemed to correspond to underlying semantic boundaries: a good global map would for instance successfully differentiate oncologists from embryologists in academic collaboration networks; or workmates from schoolmates, in friendship networks.

On the semantic side, beyond the very fertile scientometric field, we may mention the "social semantic web" as the first explicit formulation of the integration of social aspects into the so-called "semantic web" (which was itself an already-thriving field at the time). This idea is supported in [50] by an empirical example where communities of concepts are found through user-made associations. In this work, weighted links between concepts correspond to joint mention by users; communities of strongly (socially-)connected concepts are then detected using a standard community finding method. The resulting macro-level structure is essentially a knowledge map, where agents are implicit connectors.

Socio-semantic epistemic phylogenies.

In all these cases, however, the point is to find aggregates on one side first, and then project them back on the other side: either the social or the semantic aspect is primary. Knowledge communities may yet have to be described simultaneously as possibly overlapping groups of individuals dealing with groups of similar issues, sharing similar beliefs, interested in similar things. This type of issue extends over the study of affiliation in mathematical sociology. Affiliation networks are bipartite structures describing the membership of actors in groups and as such constitute a rather basic socio-semantic structure.

In a bipartite network, one of the simplest group pattern might consist of bicliques of agents affiliated with the same attributes, i.e. maximal sets of agents linked to a maximal set of attributes [9, 10, 26]. This approach may be further developed in the case of social cognition, building upon the notion of epistemic community (EC). An EC refers *a minima* to groups of agents sharing the same concepts and epistemic goals [33]: solving a given socio-technical problem, advancing science in a given field, etc. In an earlier study by ours [57], ECs have been formalized as a dual set of agents altogether using the same concepts. In other words, we focus here on the strictly socio-semantic part of $\mathcal{G}$, i.e. with an edge set restricted to $E_{\mathcal{G}} \cap (\mathbf{A} \times \mathbf{C})$. Here, a biclique is a maximal set of nodes $C \subseteq V_{\mathcal{G}}$ such that $\forall (a,c) \in (C \cap \mathbf{A}) \times (C \cap \mathbf{C})$, $\{(a,c)\} \subset E_{\mathcal{G}} \cap (\mathbf{A} \times \mathbf{C})$. Such a set C is called an EC: all its agents are connected to all its concepts, and there exists no superset of C where the same property holds. Note that $\mathbf{A}$ and $\mathbf{C}$ are by definition bicliques and bound the lattice (top and bottom).

All the ECs of a given socio-semantic graph $\mathcal{G}$ may be represented in a socio-technical lattice figuring maximal groups of agents & concepts, ordered by a set inclusion relationship: in other words, an EC C is said to be more general than another EC C' when the agent set of C contains that of C' (and thus, dually, the concept set of C is included in that of C'). Note that such lattice is a particular

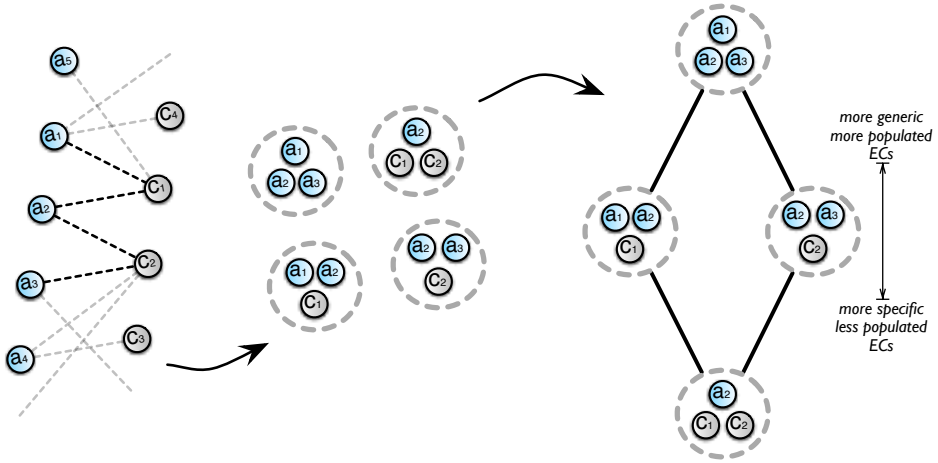


Fig. 7: Illustration of the construction of a socio-technical lattice $\mathcal{L}$, from left to right: (i) a bipartite graph features usage of concepts c_1, c_2, etc. by agents a_1, a_2, a_3 (part of $\mathcal{G}$); (ii) maximal groups of agents using the same concepts are then extracted (they are meso-level socio-semantic hyperlinks and thus form a type of socio-semantic hypergraph $\mathcal{H}$) and finally arranged into a hierarchical socio-technical lattice $\mathcal{L}$ (being a high-level socio-semantic hypergraph with the partial order $\geq_{\text{EC}}$).

instance of a Galois lattice [7, 26], a structure which is also the main focus of the "Formal Concept Analysis" (FCA) community [27].

C will then be the "parent" of C' in the lattice. Formally, we define the partial order $\geq_{\text{EC}}$ on bicliques such that for two ECs C and C',

$$C \geq_{\text{EC}} C' \iff (C \cap \mathbf{A}) \supseteq (C' \cap \mathbf{A}) \iff (C \cap \mathbf{C}) \subseteq (C' \cap \mathbf{C}) \tag{5}$$

Eventually, the (finite) socio-technical lattice $\mathcal{L}$ is based on $V_{\mathcal{G}} = \mathbf{A} \cup \mathbf{C}$ (the same set of nodes as $\mathcal{G}$ and $\mathcal{H}$), the order relation $\geq_{\text{EC}}$, and the set of all socio-semantic bicliques/ECs of $\mathcal{G}$ (note that this last set is itself a type of socio-semantic hypergraph $\mathcal{H}$). See a more concrete illustration on Fig. 7. As a result, navigating the lattice from top to bottom is equivalent to exploring socio-semantic communities from the most generic to the most specific ones. Moreover, since ECs may have more than one parent and more than one descendant, they are well-suited to the representation of non-Aristotelian taxonomies, allowing for the membership of an item to several categories.

An application of this procedure is given on Fig. 8, following [57]. As in Sec. 3.2, bibliographical data had been automatically collected from *MedLine* for all records mentioning the word "zebrafish", in order to capture the whole scientific community interested in this model animal. This typically yields a socio-semantic network $\mathcal{G}$ whose social boundaries are semantically defined, i.e. both extending to and limited

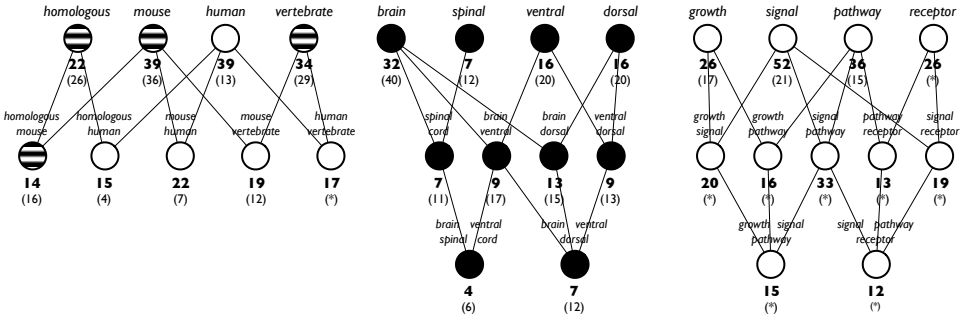


Fig. 8: Excerpt of a socio-technical lattice in the case of the zebrafish community, focused on the top (most generic) ECs. Each node is an EC, labels correspond to its concept set, while figures indicate the underlying population size, as a percentage of the total population (it totals more than 100% because of the overlap between ECs): the main figures in bold correspond to the period 1998-2003, the figures in italics inside brackets to 1990-1995, showing how the lattice population evolved over the period. Black ECs experienced a significant population decrease (more than 15%) while dashed ECs stagnated (growth within ±15%).

to scientists of this field. The "zebrafish" field is particularly suited to this task, in that individuals publishing on this animal are strongly likely to mention its name in the abstract, to the contrary of e.g. oncologists where a term like "cancer" may not necessarily be used. To allow for a longitudinal analysis, we rely on field experts to select two temporal periods of equal size corresponding respectively to the early beginnings of the zebrafish community (1990-1995) and to an institutionalized stage featuring the attributes of a normal science field, with e.g. well-established yearly, large-scale conference cycles (1998-2003). Over each time period, we build a network aggregating all links between scientists and concepts. Final networks gather 70 concepts, as in Sec. 3.2, and respectively $1,094$ and $9,691$ authors.

Because of the combinatorial complexity of the computation of bicliques, and in order to enable the longitudinal comparison between lattices over the two periods, we extract socio-semantic subgraphs of identical size. To do so, we randomly select a sample of 250 agents for each period, assuming that these uniformly random samples are representative of the main features of the structure of socio-semantic communities in the field. Admittedly, even with 250 agents and 70 concepts, resulting lattices are still huge: the first period lattice contains for instance more than $214,000$ ECs. This effect calls for the additional use of pruning heuristics, as is traditional in the study of such Galois lattices. In our case, we rely on an unsophisticated pruning technique based on scores applied to nodes of the lattice. More precisely, we assign to each EC a score consisting of the ratio between its population size (in terms of number of agents, in order to favor communities which gather a sizable portion of the whole field) and its distance to the lattice top (in order to favor more

generic communities). We then select the 20 top-ranking such ECs for each lattice, thus forming a partially-ordered set which may be vertically represented as a Hasse diagram [21].[f]

Figure 8 principally represents an excerpt of the top-ranking ECs of the final period lattice (1998-2003). It features the most generic ECs — admittedly the main topics of the community — denoted by their concepts and the size of their population (in percentage of the total community size), in order to grasp the respective size of groups of people interested in the same groups of concepts. Furthermore, population proportions inside brackets represent the size of ECs with the same concept set in the 1990-1995 lattice, if they existed (or a star, if some EC of the last period did not exist in the first period). In other words, the figure shows the evolution of the socio-semantic macro-structure of the zebrafish community from the period 1990-95 to 1998-2003.

More broadly, the comparison of such lattices at different points in time makes it possible to describe the high-level distribution of social cognition processes within a given socio-technical system. Here, the socio-technical lattice describes three main areas of research, organized around three subsets of concepts and corresponding scientists: (i) the study of biochemical signaling mechanisms, involving pathways and receptors; (ii) comparative studies focusing on similarities and differences between humans, mice, zebrafish as vertebrates; (iii) the examination of the nervous system and brain development. The first and, to a lesser extent, the second subfields grew in importance within the community at the expense of the last field: research on brain and spinal cord decreased and its relationship with ventral and dorsal aspects became weaker. On the other hand, the community started to venture into signaling issues; which is partly explained by the emergence of a more general background trend in molecular biology. These static and dynamic maps are all validated by our expert of the field, who confirms the existence and content of these three main trends.

Beyond this example, the macroscopic exploration of intrinsically socio-semantic structures remains a challenging field where little is currently known.[g] In the case of knowledge community mapping, socio-semantic hypergraphs could here again be part of the solution towards developing a unified formalism for relational and topical communities.

[f]The top EC is trivial: it gathers the whole community and no concept, as there is no universally-shared concept (note here that "zebrafish" is evidently not part of the 70 selected concepts). The top EC is therefore not featured on this diagram.

[g]Note that the more sophisticated description of *folksonomies* using data stemming from online socio-technical communities as discussed in [35] relies on similar methods, this time applied to tri-partite relationships connecting agents, concepts (attributes) and artifacts (digital items).

4. Concluding remarks

Socio-technical systems feature a co-evolutionary dynamics between social networks of interaction or collaboration, and so-called "semantic" networks of term, topic and issue associations. As such, they are particularly prone to the experimental and *in vivo* observation of social cognition processes: in all generality, they raise broad classes of research questions ranging from the peculiar structure of socio-semantic communities to the generalized understanding of multi-level socio-semantic cultural dynamics — i.e., co-evolution phenomena between social and semantic networks, between various communities/territories, and between the various scales of social cognition dynamics: macro-, micro-, and meso-. Here, we aimed to illustrate how these various levels could be modeled within a formalism where interactional and conceptual dynamics may jointly be appraised, be it at the level of a socio-semantic graph $\mathcal{G}$, a socio-semantic hypergraph $\mathcal{H}$ or a socio-semantic lattice $\mathcal{L}$.

This type of understanding is not only necessary for the sake of engineering better socio-technical platforms: societal applications follow naturally and make this question a crucial one for civic debates and policy-making. For one, understanding the construction and differentiation of the social groups and actors behind specific broad topics and particular sets of issues (media, citizens, organizations, etc.) through the emergence, melding, scission, decline of socio-topical communities and, more locally, the dynamics behind the structural and semantic embedding of actors into given socio-semantic communities (as well as, symmetrically, the way topics may become closer or merge as a result of an alignment of their respective underlying social base). Formally, describing the co-evolving nature of the alignment between actors and alignment between topics induces the binding of both the social & semantic aspects, and the macro & micro levels — by exhibiting (macro-level) issue dynamics, (meso-level) transmission paths and (micro-level) key relaying actors or catalyst concepts.

This prospect is also likely to be key in more pragmatic debates regarding the *"balkanization"* of the public space, in particular online: are the new digital public spaces facilitating the confrontation of antagonistic and competing opinions coming from varied social circles, or are they reinforcing and, sometimes, isolating groups of individuals sharing similar views [67, 69]? Empirically verifying the hypothesis of a polycentric public space, made of a multi-layered structure of topically-focused and interconnected web communities, and from which local authorities may emerge, would have significant political side effects. Going further, the binding of both socio-semantic and macro-micro interactions could help understanding the "*reification*" of topical communities, by observing how they are progressively being denoted by the actors as, indeed, communities (i.e. understand how and when actors start to reflexively acknowledge the existence of some communities, or, more formally, when macro-structures become apparent to and designated by actors).

For the moment, still, the above case studies seem to demonstrate that, while we are starting to understand some empirical socio-semantic phenomena, we also

have essentially superficial knowledge of the social cognition process as a whole: the connection between the various levels of agency (micro / meso / macro) remains quite unexplored. To reach an integrated understanding of social cognition in socio-technical systems, the type of socio-semantic dynamics that we presented here needs certainly to be developed further into a multi-level framework; it also needs to be enriched on the side of information description — our way of appraising mental representations is at best sketchy (n-grams), at worst erroneous (for instance, by generally assuming some sort of perfect copying process in studies focused on contagion). More broadly, while we now have good knowledge of social network processes, we still need to enhance our description of local cognition processes. This would also constitute a first step towards the possibility of a broad program of empirical description and modeling of the theories of social epistemology and cultural anthropology.

Acknowledgments

I am particularly grateful to Jean-Philippe Cointet for follow-up discussions and comments on the present article. This paper partially relies on cited work done in collaboration with Jean-Philippe Cointet, Carla Taramasco and Paul Bourgine. I also thank two anonymous reviewers for their constructive remarks. This work was additionally partially supported by the European Commission Future and Emerging Technologies programme FP7-COSI-ICT through project QLectives (grant no.: 231200).

References

[1] Adamic, L. A. and Glance, N., The political blogosphere and the 2004 U.S. election: divided they blog, in *LinkKDD '05: Proc. 3rd Intl. Workshop on Link discovery* (ACM Press, New York, NY, USA, 2005), ISBN 1-59593-215-1, pp. 36–43.

[2] Adamic, L. A., Wei, X., Yang, J., Gerrish, S., Nam, K. K., and Clarkson, G. S., Individual focus and knowledge contribution, *First Monday* **15** (2010) 1.

[3] Aiello, L. M., Barrat, A., Cattuto, C., Ruffo, G., and Schifanella, R., Link creation and profile alignment in the anobii social network, in *Proc. of SocialCom'10 2nd IEEE Intl Conf on Social Computing* (2010), pp. 249–256.

[4] Albert, R. and Barabási, A.-L., Statistical mechanics of complex networks, *Reviews of Modern Physics* **74** (2002) 47–97.

[5] Asur, S. and Huberman, B. A., Predicting the future with social media (2010), arXiv.org e-print archive: 1003.5699.

[6] Balog, K., Mishne, G., and de Rijke, M., Why are they excited? identifying and explaining spikes in blog mood levels, *11th Meeting EACL* (2006).

[7] Barbut, M. and Monjardet, B., *Algèbre et Combinatoire*, Vol. II (Hachette, Paris, 1970).

[8] Barthélemy, M., Spatial networks, *Physics Reports* **499** (2011) 1–101.

[9] Boeck, P. D. and Rosenberg, S., Hierarchical classes: Model and data analysis, *Psychometrika* **53** (1988) 361–381.

[10] Breiger, R. L., The duality of persons and groups, *Social Forces* **53** (1974) 181–190.

[11] Callon, M., Courtial, J.-P., and Laville, F., Co-word analysis as a tool for describing

the network of interactions between basic and technological research: The case of polymer chemistry, *Scientometrics* **22** (1991) 155–205.

[12] Callon, M., Law, J., and Rip, A., *Mapping the dynamics of science and technology* (MacMillan Press, London, 1986).

[13] Cattuto, C., Loreto, V., and Pietronero, L., Semiotic dynamics and collaborative tagging, *PNAS* **104** (2007).

[14] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P., Measuring user influence in twitter: The million follower fallacy, in *Proc. AAAI ICWSM Intl. Conf. Weblogs and Social Media 2010* (2010), pp. 10–17.

[15] Claidière, N. and Sperber, D., Commentary: The role of attraction in cultural evolution, *Journal of Cognition and Culture* **7** (2007) 89–111.

[16] Cointet, J.-P. and Roth, C., Socio-semantic dynamics in a blog network, in *Proc. IEEE 4th Intl. Conf. Social Computing* (2009), pp. 114–121.

[17] Cointet, J.-P. and Roth, C., Local networks, local topics: Structural and semantic proximity in blogspace, in *Proc. 4th ICWSM AAAI Intl. Conf. on Weblogs and Social Media* (AAAI, 2010), pp. 223–226.

[18] Conte, R., Memes through (social) minds, in *Darwinizing Culture: The Status of Memetics as a Science*, ed. Aunger, R. (Oxford: Oxford University Press, 2000), pp. 83–120.

[19] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S., Feedback effects between similarity and social influence in online communities, in *Proc. of KDD '08 14th ACM SIGKDD Intl Conf on Knowledge Discovery and Data mining* (2008), pp. 160–168.

[20] Cummings, J. and Kiesler, S., Coordination costs and project outcomes in multi-university collaborations, *Research Policy* **36** (2007) 1620–1634.

[21] Davey, B. A. and Priestley, H. A., *Introduction to Lattices and Order*, 2nd edn. (Cambridge, UK: Cambridge University Press, 2002).

[22] Dawkins, R., *The Selfish Gene*, chapter 11: Memes, The New Replicator (Oxford: Oxford University Press, 1976), pp. 189–201.

[23] deB. Beaver, D., Collaboration and teamwork in physics, *Czech Journal of Physics B* **36** (1986).

[24] Emirbayer, M. and Goodwin, J., Network analysis, culture, and the problem of agency, *American Journal of Sociology* **99** (1994) 1411–1454.

[25] Fortunato, S., Community detection in graphs, *Physics Reports* **486** (2010) 75—174.

[26] Freeman, L. C. and White, D. R., Using Galois lattices to represent network data, *Sociological Methodology* **23** (1993) 127–146.

[27] Ganter, B. and Wille, R., *Formal Concept Analysis: Mathematical Foundations* (Springer, Berlin, 1999).

[28] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L., Detecting influenza epidemics using search engine query data, *Nature* **457** (2009) 1012–1014.

[29] Goetz, M., Leskovec, J., McGlohon, M., and Faloutsos, C., Modeling blog dynamics, in *ICWSM 2009 Proc. 3rd International AAAI Conference on Weblogs and Social Media* (2009).

[30] González, M. C., Hidalgo, C. A., and Barabási, A.-L., Understanding individual human mobility patterns, *Nature* **453** (2008) 779–782.

[31] Goyal, S., *Connections: An Introduction to the Economics of Networks* (Princeton University Press, Princeton, NJ, 2009).

[32] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A., Information diffusion through blogspace, in *WWW2004: Proceedings of the 13th Intl Conf on World Wide*

Web (NYC, NY, USA, 2004), pp. 491–501.
[33] Haas, P., Introduction: epistemic communities and international policy coordination, *International Organization* **46** (1992) 1–35.
[34] Hutchins, E., Distributed cognition, in *International Encyclopedia of the Social and Behavioral Sciences*, eds. Smelser, N. J. and Baltes, P. B. (Elsevier, 2001), pp. 2068–2072.
[35] Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., and Stumme, G., Discovering shared conceptualizations in folksonomies, *Web Semantics: Science, Services and Agents on the World Wide Web* **6** (2008) 38–53.
[36] Kitcher, P., Contrasting conceptions of social epistemology, in *Socializing Epistemology: The Social Dimensions of Knowledge*, ed. Schmitt, F. (Rowman and Littlefield, Lanham, MD, 1995), pp. 111–134.
[37] Knorr-Cetina, K., *Epistemic Cultures: How the Sciences Make Knowledge* (Harvard University Press, Cambridge, MA, 1999).
[38] Kreuzman, H., A co-citation analysis of representative authors in philosophy: Examining the relationship between epistemologists and philosophers of science, *Scientometrics* **51** (2001) 525–539.
[39] Latour, B. and Woolgar, S., *Laboratory Life: The Social Construction of Scientific Facts* (Sage Publications, Beverly Hills, 1979).
[40] Lazega, E., *Micropolitics of Knowledge: Communication and Indirect Control in Workgroups* (Aldine de Gruyter, New York, NY, 1992).
[41] Leenders, R. T., Longitudinal behavior of network structure and actor attributes: Modeling interdependence of contagion and selection, in *Evolution of social networks*, eds. Doreian, E. and Stokman, E. N. (Gordon and Breach, Amsterdam, 1997), pp. 165–184.
[42] Lenclud, G., La culture s'attrape-t-elle ?, *Communications, EHESS, Centre d'études transdisciplinaires* **66** (1998) 165–183.
[43] Leskovec, J., Backstrom, L., and Kleinberg, J., Meme-tracking and the dynamics of the news cycle, in *Proc. ACM SIGKDD'09 15th Intl. Conf. on Knowledge Discovery and Data Mining* (2009), pp. 497–506.
[44] Leskovec, J., Huttenlocher, D., and Kleinberg, J., Governance in social media: A case study of the wikipedia promotion process, in *Proc. AAAI ICWSM Intl. Conf. Weblogs and Social Media 2010* (2010), pp. 98–105.
[45] Lloyd, L., Kaulgud, P., and Skiena, S., Newspapers vs. blogs: Who gets the scoop?, in *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs(AAAI-CAAW), Palo Alto, California, USA* (2006), pp. 117–124.
[46] McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M., and Glance, N., Finding patterns in blog shapes and blog evolution, in *International Conference on Weblogs and Social Media* (2007).
[47] McPherson, M., Smith-Lovin, L., and Cook, J. M., Birds of a feather: Homophily in social networks, *Annual Review of Sociology* **27** (2001) 415–444.
[48] Melin, G. and Persson, O., Studying research collaboration usign co-authorships, *Scientometrics* **36** (1996) 363–377.
[49] Mesoudi, A., Whiten, A., and Laland, K. N., Perspective: Is human cultural evolution darwinian? Evidence reviewed from the perspective of the origin of species, *Evolution* **58** (2004) 1–11.
[50] Mika, P., Ontologies are us: A unified model of social networks and semantics, *Journal of Web Semantics* **5** (2007) 5–15.
[51] Mishne, G. and Glance, N., Leave a reply : An analysis of weblog comments, in *Proc. 3rd annual workshop on the Weblogging Ecosystem: aggregation, Analysis and*

Dynamics, Edimburgh, WWW06 (2006).
[52] Mullins, N. C., The development of a scientific specialty: The phage group and the origins of molecular biology, *Minerva* **10** (1972) 51–82.
[53] Newman, M. E. J., Scientific collaboration networks. I. Network construction and fundamental results, and II. Shortest paths, weighted networks, and centrality, *Physical Review E* **64** (2001) 016131 & 016132.
[54] Newman, M. E. J., The structure of scientific collaboration networks, *PNAS* **98** (2001) 404–409.
[55] Roca, C. P. and Helbing, D., Emergence of social cohesion in a model society of greedy, mobile individuals, *PNAS* **108** (2011) 11370–11374.
[56] Rosvall, M. and Bergstrom, C. T., Mapping change in large networks, *PLoS One* **5** (2010) e8694.
[57] Roth, C. and Bourgine, P., Epistemic communities: Description and hierarchic categorization, *Mathematical Population Studies* **12** (2005) 107–130.
[58] Roth, C. and Cointet, J.-P., Social and semantic coevolution in knowledge networks, *Social Networks* **32** (2010) 16–29.
[59] Ruef, M., A structural event approach to the analysis of group composition, *Social Networks* **24** (2002) 135–160.
[60] Salton, G., Wong, A., and Yang, C. S., Vector space model for automatic indexing, *Communications of the ACM* **18** (1975) 613–620.
[61] Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F., Folks in folksonomies: Social link prediction from shared metadata, in *Proc. WSDM'10 ACM 3rd Intl Conf on Web Search and Data Mining* (ACM, New York, NY, 2010), pp. 271–280.
[62] Shoham, Y. and Leyton-Brown, K., *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, chapter 13, Logics of Knowledge and Belief (Cambridge University Press, 2008), pp. 393–420.
[63] Simmons, M. P., Adamic, L. A., and Adar, E., Memes online: Extracted, substracted, injected and recollected, in *Proc. 5th ICWSM - AAAI Intl Conf Weblogs & Social Media* (2011), pp. 353–360.
[64] Snijders, T. A. B., Steglich, C., and Schweinberger, M., Modeling the co-evolution of networks and behavior, in *Longitudinal models in the behavioral and related sciences*, eds. van Montfort, K., Oud, H., and Satorra, A. (Mahwah, NJ: Lawrence Erlbaum, 2007), pp. 41–71.
[65] Sperber, D., *Explaining Culture: A Naturalistic Approach* (Oxford: Blackwell Publishers, 1996).
[66] Stokols, D., Hall, K. L., Taylor, B. K., and Moser, R. P., The science of team science, *American Journal of Preventive Medicine* **35** (2008) S78–S89.
[67] Sunstein, C., Architecture of serendipity, *Harvard Crimson* (2008).
[68] Taramasco, C. A., Cointet, J.-P., and Roth, C., Academic team formation as evolving hypergraphs, *Scientometrics* **85** (2010) 721–774.
[69] Van Alstyne, M. and Brynjolfsson, E., Electronic Communities: Global Village or Cyberbalkans?, in *Proc. 17th Intl Conf Information Systems (ICIS 1996), Cleveland, OH*, eds. DeGross, J., Jarvenpaa, S., and Srinivasan, A. (1996), pp. 80—98.
[70] Wang, F.-Y., Carley, K. M., Zeng, D., and Mao, W., Social computing: From social informatics to social intelligence, *Intelligent Systems* **22** (2007) 79–83.
[71] Wasserman, S. and Faust, K., *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).