Morphogenesis of epistemic networks: a case study

Camille Roth

CRESS, Department of Sociology, University of Surrey, Guildford, GU2 7XH, United Kingdom
c.roth@surrey.ac.uk

published: *Proceedings of ESSA European Social Simulation Association 4th Conf., 09/2007*

Abstract. Agents producing and exchanging knowledge are forming as a whole a socio-semantic complex system. We argue that several significant aspects of the structure of a knowledge community are primarily produced by the co-evolution between agents and concepts, i.e. the evolution of an epistemic network. Focusing on a particular community of scientists working on a well-defined topic, we micro-found various stylized facts regarding its structure by exhibiting processes at the level of agents accounting for the emergence of epistemic community structure. After assessing the empirical interaction and growth processes, and assuming that agents and concepts are co-evolving, we successfully propose a morphogenesis model rebuilding relevant high-level stylized facts.

Introduction

Agents producing, manipulating, exchanging knowledge are forming as a whole a socio-semantic complex system: they are fully immerged in flows of information on which they can have an impact and leave their footprints at the same time. The massive availability of informational content and the potential for extensive interactivity has recently made the focus slip from single "groups of knowledge" to the entire "society of knowledge", in a *networked* fashion, calling for the use of new methods and the characterization of new phenomena, with knowledge being distributed and appraised on a more horizontal basis.

Understanding the structural aspects of these communities relate more broadly to a recent issue in social science, social network formation modeling, involving several disciplines from graph theory (computer science and statistical physics), mathematical sociology to economics [1–3]. Most of the recent interest has stemmed from the empirical observation that real social network structure strongly differs from that of uniform random graphs *a la* Erdős-Rényi (ER) [4], suggesting that agents interact non-randomly with respect to heterogenous preferences for interacting with other agents. While this fact was already well-documented in social science [5, 6], general network models had been limited for long to ER-like random graphs [7–9]. Subsequently, much work has been devoted to determining novel non-uniform interaction and growth mechanisms reconstructing complex network structures consistent with the real world, through a rich set of statistical parameters [10]. On the whole, this amounts to find the solution of a reverse problem: given such an evolving system, what kind of (possibly minimal) dynamics rebuild its structure? In other words, we look for a valid network morphogenesis model for the real-world structure.

We focus here on a particular socio-semantic complex system, a scientific community working on a well-defined topic, and we make the following assumption: *modeling interactions at the level of agents who co-evolve with the concepts they manipulate is sufficient to carry the micro-founded reconstruction of this complex system.* More precisely, we will rebuild several aspects of the structure of such a community by introducing a co-evolutionary framework based on a social network, a semantic network and a socio-semantic network; as such an epistemic network made of agents, concepts, and relationships between all of them. We will then show that dynamics at the level of this epistemic network are sufficient to reproduce several stylized facts of interest. To achieve such a morphogenesis model, we first build tools enabling the estimation of interaction and growth mechanisms from past empirical data. Then, assuming that agents and concepts are co-evolving, we successfully reconstruct a real-world scientific community structure for a relevant *selection* of features.

1 Networks

(Social) network morphogenesis models. Networks (or graphs) are omnipresent in the real world: from the lowest levels of physical interaction to higher levels of description such as biology, sociology, economics and linguistics. For long however, network appraisal had been restricted to theoretical approaches in graph theory and small scale empirical studies on a case-by-case basis; while network models were mostly limited to the seminal work of Erdős-Rényi [4] (ER), which was assumed to be realistic for most purposes. In this respect, the recent availability of increasingly larger computational capabilities has made possible the use of quantitative methods on large networks, which yielded surprising results, often contradictory with those provided by ER models. This consequently precipited an unprecedented interest in networks [10–12]. Three statistical parameters in particular appeared to provide an enormous insight on the topological structure of networks: (i) clustering coefficient (the proportion of neighbors of a node who are also connected to each other, averaged over the whole network), (ii) average distance (i.e. the length of the shortest path between two nodes, averaged over all pairs of nodes), (iii) degree distribution (the degree (or the connectivity) of a node is basically the number of nodes this node is connected to).

Several recent works suggested morphogenesis models matching empirical data on these statistical parameters, contradicting and eventually replacing the ER model [13–16]. More specifically, Barabasi & Albert (BA) [16] insisted on the point that such topology could be due to two very particular phenomena that models were so far unable to take into account: network growth, and preferential attachment of nodes to other nodes. They thus pioneered the use of these two features to successfully rebuild a scale-free degree distribution. In their network formation model, new nodes arrive at a constant rate and attach to already-existing nodes with a likeliness linearly proportional to their degree. This model was a great success and has been widely spread and reused. As a consequence, the term "preferential attachment" has been often understood as degree-related preferential attachment only, in reference to BA's work. Since then, many other authors introduced network morphogenesis models with diverse modes of preferential link creation depending on various node properties (attractiveness [17], common

neighbors [18], fitness [19], centrality [20], hidden variables and "types" [21], bipartite structure [22], etc.) and various linking mechanisms (stochastic copying of links [23], competitive trade-off and optimization heuristics [20, 24], payoff-biased network reconfiguration [25], group formation [26], to cite a few). On the other side, growth processes (if any) were often reduced to the regular addition of nodes which attach to older nodes — sometimes growth is absent and studies are focused on the evolution of links only.

The idea is usually to exhibit high-level statistical parameters and suggest low-level network processes, in order to deduce the former from the latter. Obviously, after selecting a set of relevant stylized facts to be explained, model design consists of two subtasks: defining the way agents are bound to interact with each other, as well as specifying how the network grows. However, even in recent papers, hypotheses on such mechanisms are often arbitrary and rarely empirically checked. This attitude is still convenient for normative models but is rather questionable for descriptive models. Here, we therefore endeavor to (i) exhibit high-level stylized facts characteristic of epistemic networks, (ii) point out relevant low-level features that may account for these high-level facts, (iii) design measurement tools to appraise these low-level features, and (iv) design a reconstruction model based on the *observed* low-level dynamics.

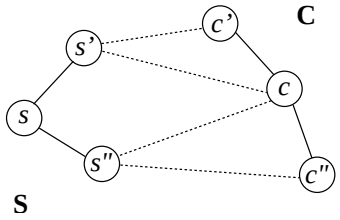

Fig. 1. Sample epistemic network with $\mathbf{S} = \{s, s', s''\}$, $\mathbf{C} = \{c, c', c''\}$, and relations $\mathcal{R}^{\mathbf{S}}$, $\mathcal{R}^{\mathbf{C}}$ (solid lines) and $\mathcal{R}$ (dashed lines).

Epistemic networks and empirical setting. We first introduce the objects we deal with: we distinguish a *social network* (linking agents), a *semantic network* (linking concepts) and a socio-semantic network (linking agents to concepts). Nodes in the social network $\mathcal{S}$ are agents, and links represent co-occurrence of two agents in an event. Thus $\mathcal{S} = (\mathbf{S}, \mathcal{R}^{\mathbf{S}})$, $\mathbf{S}$ is the set of agents and $\mathcal{R}^{\mathbf{S}}$ the set of undirected links. The semantic network $\mathcal{C} = (\mathbf{C}, \mathcal{R}^{\mathbf{C}})$ is the network of co-occurrences of concepts within events: $\mathbf{C}$ denotes the concept set, $\mathcal{R}^{\mathbf{C}}$ denotes links between concepts. The socio-semantic network $\mathbf{G_{SC}}$ is made of agents of $\mathbf{S}$, concepts of $\mathbf{C}$, and links between them, $\mathcal{R}^{\mathbf{SC}} = \mathcal{R}$, denoting usage of concepts by agents (undirected links for co-occurrence of agents and concepts). An *epistemic network* is thus given by these three networks, key for providing an account of the reciprocal influence and co-evolution of authors and concepts (Fig. 1). This is not to be confused with a bipartite network and its various projections: while the socio-semantic graph is indeed bipartite, social and semantic networks are not projections of the latter.

Translated in this framework, events are articles, agents are their authors, and concepts are made of expert-selected abstract words. We considered the community of

embryologists working on the model animal *"zebrafish,"* during the period 1997–2004. Our main source of data is `MedLine`, a US National Library of Medicine reference database. The dataset contains around $10,000$ authors, $6,000$ articles and 70 concepts, adopting a weak linguistic assumptions by assuming that a lemmatized term corresponds to a concept. We restrict the dictionary to the 70 most used and significant words in the community selected with the help of a domain expert, in order to avoid rhetorical and neutral terms ("stop-words"). These concepts are given *a priori*: in the semantic network, only links appear, not nodes.

2 High-level features

We first endeavor to describe a few high-level statistical parameters particularly appropriate for epistemic networks. While we could have looked at many single-network parameters (such as assortativity [27], giant component size [28], single-network communities [29, 30], etc.), we focused instead on features specific to this epistemic network (thus, mostly bipartite parameters) — many results and models are already available for most traditional statistical features.

Degree distributions. In an epistemic network, ties appear in the social, semantic, and socio-semantic networks; hence, four degree distributions are of interest (Fig. 5):

1. *degrees k for the social network*: this distribution has been extensively studied in the literature [31, 32], and is traditionally said to follow a power-law, although only the tail actually does; some may suggest that this distribution follows a log-normal [33] or q-exponential law [34].
2. *degrees k_c for the semantic network*: since there are only 70 concepts the data are very sparse, we considered cumulated distributions (see exact definitions on Fig. 5) — all concepts are being progressively connected to each other.
3. *degrees from agents to concepts ($k_{a\to c}$)*: following a power-law; few agents use many concepts, many agents use few concepts.
4. *degrees from concepts to agents ($k_{c\to a}$)*: few concepts being used by a lot of agents, and most concepts being used by an average number of agents.

Clustering. The clustering coefficient is another valuable parameter [13]. It is basically a measure of the transitivity in one-mode networks, expressing how neighbors of a given node are connected to each other ("friends of friends are friends"). This coefficient is usually found to be very high in empirical social networks when compared to typical random networks such as those produced by ER, BA models. Along with degree distribution, this stylized fact has been the target of many more recent models [18, 35].

We use the local clustering coefficient, $c_3(i)$, measuring the proportion of neighbors of node i who are connected together: $c_3(i) = \frac{\text{[number of pairs of connected neighbors]}}{k_i \cdot (k_i - 1)/2}$ with k_i degree of i. This coefficient is close to 1 here and decreases rather slowly with node degree (Fig. 6). Yet, networks built with an underlying event structure are bound to exhibit a high coefficient [36, 37], thus a poorly informative criterion. By contrast, a

bipartite clustering coefficient counting the proportion of diamonds [38] is a meaningful measure of how two agents connected to a same concept are likely to be connected to other concepts (as such a very local kind of structural equivalence): c_4 is the proportion of common neighbors among the neighbors of a node — in other words, are two agents connected to a same concept likely to be connected to other concepts? This coefficient appears to be one order of magnitude larger compared to that measured in scale-free random networks: pairs of agents linking together to certain concepts share other concepts abnormally often (Fig. 6).

Epistemic community structure. A key high-level stylized fact characteristic of epistemic networks is the particular distribution of epistemic communities (ECs) as *groups of agents using jointly the same concepts*, or maximal bipartite cliques in the socio-semantic network [39, 40]. An adequate epistemic network model should ultimately yield the same EC profile as in the real-world, which shows a significantly larger proportion of high-size ECs — see Fig. 7.

Besides, just as we observed the bipartite clustering between agents and concepts, we may want to know whether agents in the network are semantically close to each other. Likewise, and more specifically, in which manner are they semantically close to their social neighborhood? To this end, we need to introduce a semantic distance, i.e. a function of a dyad of agents which (i) decreases with the number of shared concepts, (ii) increases with the number of distinct concepts, (iii) equals 1 when there is no concept in common, 0 when all concepts are identical. Given $(s, s') \in \mathbf{S}^2$, and $s^\wedge$ the set of concepts s is linked to, we suggest the following metric semantic distance $\delta(s, s') \in [0; 1]$, based on the classical Jaccard coefficient [41], such that $\delta(s, s') = \frac{|(s^\wedge \setminus s'^\wedge) \cup (s'^\wedge \setminus s^\wedge)|}{|s^\wedge \cup s'^\wedge|}$. As δ takes real values in $[0,\ 1]$ we discretize δ, using a uniform partition of $[0,\ 1[$ in $I - 1$ intervals, to which we add the singleton $\{1\}$. We thus define a new discrete distance d taking values in $\mathcal{D} = \{d_1, d_2, ..., d_I\}$ such that: $\mathcal{D} = \{[0, \frac{1}{I-1}[,\ [\frac{1}{I-1}, \frac{2}{I-1}[, ...[\frac{I-2}{I-1}, 1[, \{1\}\}$. Then, we look at the distribution of semantic distances in the network, both on a global scale (by computing the distribution for all pairs of agents) and on a more local scale (by carrying the computation for pairs of already-connected agents only). Results on Fig. 7 suggest that while similar nodes are usually rare in the network, the picture is radically different when considering the social neighborhood: acquaintances are at a strongly closer distance.

3 Low-level dynamics

3.1 Measuring interaction behavior

Designing a credible social network morphogenesis model requires to understand both low-level interaction and growing mechanisms. We therefore first show how to design such low-level dynamics from empirical data. If the observed empirical structure diverges from the ER uniform random model, this suggests that interactions are not occurring totally at random but, rather, are directed by preferences on agents. In other words, there is prefential attachment from some agents to some other kinds of agents.

Formally, preferential attachment (PA) is the likeliness for a node to be involved in an interaction with another node with respect to node properties. Existing *quantitative* estimations of PA and subsequent validations of modeling assumptions are quite rare, and are often either related to the classical degree-related PA [32, 33], or considering PA as a scalar quantity, using direct mean calculation, econometric estimation approaches or Markovian models [28, 42, 43]. We here use a unified framework where properties are neither strictly based on social network topology nor reduced to single scalar quantities, while appraising how distinct properties correlatively influence PA.

We distinguish (i) single node properties, or *monadic* properties (such as degree, age, etc.) from (ii) dyadic properties (social distance, dissimilarity, etc.). When dealing with monadic properties indeed, we seek to know the propension of some kinds of nodes to be involved in an interaction. On the contrary when dealing with dyads, we seek to know the propension for an interaction to occur preferentially with some kinds of couples. We assume the influence on PA of a given monadic property m can be described by a function f of m, the *interaction propension*, independent of the distribution of agents of kind m: $f(m)$ is simply the conditional probability $P(\mathrm{L}|m)$ that an agent of kind m receives a link L. Thus, it is $f(m)$ times more probable that an agent of kind m is involved into an interaction. For instance, the classical degree-based PA used in BA and subsequent models is an assumption on f equivalent to $f(k) \propto k$. We may estimate f through $\hat{f}(m) = \frac{\nu(m)}{P(m)}$ if $P(m) > 0$, 0 otherwise, where $\nu(m)$ is the expectancy of new link extremities attached to nodes of property m along a period, and $P(m)$ typically denotes the distribution of nodes of type m. We adopt a dyadic viewpoint whenever a property has no meaning for a single node, such as proximity, similarity — or distances in general. Similarly, we assume the existence of an essential dyadic interaction behavior embedded into $g(d)$ for a given dyadic property d defined on couples of agents, corresponding to the conditional probability $P(\mathrm{L}|d)$. Again, g is estimated with $\hat{g}(d) = \frac{\nu(d)}{P(d)}$.

The PA behavior embedded in $\hat{f}$ (or $\hat{g}$) can be used to shape modeling hypotheses, either by taking the empirically estimated function, or by stylizing the trend of $\hat{f}$ (or $\hat{g}$) to allow analytic solutions. When considering a property which enjoys an underlying natural order, it is also useful to examine the cumulative propension $\hat{F}(m_i) = \sum_{m'=m_1}^{m_i} \hat{f}(m')$ as an estimation of the integral of f, especially with noisy data. Besides, when considering a collection of properties one must make sure that they are uncorrelated: for instance, node degrees may depend on age. If two distinct properties p and p' are independent, the distribution of nodes of kind p in the subset of nodes of kind p' does not depend on p', i.e. the quantity $\frac{P(p|p')}{P(p)}$ theoretically equals 1.

3.2 Empirical PA

Using these tools, we examine PA based on (i) a monadic property (node degree) and (ii) on a dyadic property (semantic distance d, rendering homophily). We first consider the node degree k as property m: we compute the real slope $\hat{f}(k)$ of the degree-related PA and empirically roughly verify the classical assumption "$f(k) \propto k$" (Fig. 2). This precise result is not new and tallies with existing studies on degree-related PA [44, 45].

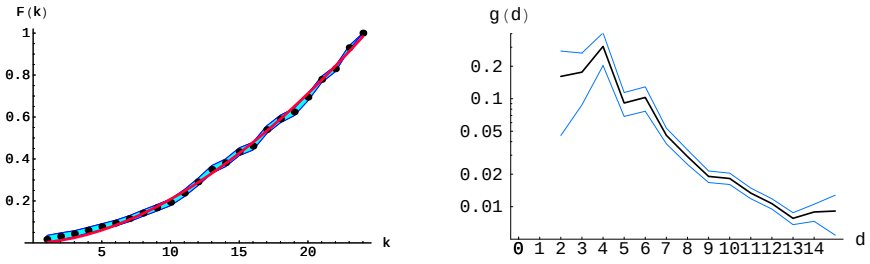

Fig. 2. *Left:* Cumulated propension $\hat{F}$. Dots represent empirical values, the solid line is the best non-linear fit for $\hat{F} \sim k^{1.83}$ (i.e. $\hat{f} \sim k^{0.83}$), and the gray area is the confidence interval. ***Right:*** Homophilic interaction propension $\hat{g}$ with respect to a semantic distance $d \in \{0, ..., 15\}$ (thick solid line) and confidence interval for $p < .05$ (thin lines).

We also assess the extent to which agents are "homophilic" (they prefer to interact with similar agents) by using the semantic distance introduced in Sec. 2. Empirical results on Fig. 2 show that while agents favor interactions with slightly different agents, they still very strongly prefer similar agents. Besides, the exponential trend of $\hat{g}$ suggests that homophily is even more influent than connectedness. This fiercely advocates the use of semantic content for modeling such networks, while showing that simple non-structural properties may strongly shape interaction behavior in some networks. As underlined above, we check if the two properties are independent, i.e. whether or not a node of low degree is more or less likely to be at a larger semantic distance of other nodes. Here, there is no correlation between degree and semantic distance.

We finally examine if concepts are preferentially chosen: are well-connected concepts used more often, thus 'interacting' with even more authors? It turns out that concepts appear proportionally to their socio-semantic degree (i.e. the number of agents who use them) which reflects their popularity.

3.3 Growth- and event-related parameters

These features yield an essential insight on how local interactions occur. Now, in order to complete the description of the way the network grows, studying how events are structured in terms of both authors and concepts is also a crucial information. Regularly, new articles are produced, involving on one side a certain number of authors who have already authored a paper (old nodes) and possibly a fraction of new authors (new nodes), and on the other side, concepts that the authors bring in as well as new concepts.

Network growth. The first step is to determine the raw network growth, in terms of new nodes. How many new events appear, how many new articles are written during each period? Articles gather existing authors as well as new authors around concepts. Since we consider the set of concepts to be fixed *a priori*, new nodes appear in the social network only. The evolution of the size of the social network N_t depends on the number of new nodes per period ΔN_t, with $N_{t+1} = N_t + \Delta N_t$. In turn, there is a

strong link between ΔN_t and the number of articles n_t, depending on the fraction of new authors per article. The growth of both ΔN_t and n_t is roughly linear with time: we can approximate the evolution of n by $n_{t+1} = n_t + n_+$, for a given arithmetic growth rate of n_+; every period the number of new articles increases by n_+. In our case, $n_+ \simeq 96$ ($\sigma \simeq 28$). ΔN and n seem to be linearly correlated, suggesting that the proportion of new authors in all articles is stable across periods.

Size of events. This leads us to study how articles are structured: in particular, how many agents are gathered in an event, and how many of them are new nodes? As shown on Fig. 3, the distribution of the number of agents per article appears to follow roughly a geometric distribution. On the other hand, the weight of new authors within articles obeys a distribution centered around three modes $\{0, 0.5, 1\}$, suggesting that in most cases either (i) authors are all new, (ii) they are all old, or (iii) half are new & half are old. Since this proportion is stable across periods, n_t is a good indicator of network growth: new articles appear and pull new authors into the network — on average, articles gather 4.4 authors, among which 55% are new, thus $.55 \times 4.4 = 2.42$ new authors, which is close to the coefficient of the best linear fit of ΔN with respect to n: $\Delta N \sim 2.25n$.

Since the size of the network is increased by ΔN in a period, and ΔN here shows a linear behavior, N exhibits a quadratic growth. The fact that the number of articles per period linearly increases is however proper to the evolution of *this* empirical situation. The evolution of n and N is a consequence of this — this is obviously not the case for all networks: if for instance this field of research were to be abandoned, we would have a decrease of articles, not a linear growth.

Exchange of concepts. Knowing the structure of articles, and how authors are gathered, we now investigate how concepts are used. The distribution of the number of concepts is plotted on Fig. 3, and could be accurately approximated by a geometric distribution. Besides, while old authors bring a certain proportion of their concepts, some concepts are used for the first time: they do not belong to the intension of authors. The distribution of the proportion of new concepts — *new to the authors* — also shown on Fig. 3, makes it possible to distinguish concepts chosen within the intension of authors, from new, unused ones. It has a single mode 0, but is on the whole relatively flat.

4 Epistemic network morphogenesis model

Design. Using empirically-measured low-level parameters (article composition, interaction preferences) we design a model that rebuilds a high-level structure compatible with real-world stylized facts (degree and semantic distance distributions, bipartite clustering, EC structure). Three key modeling features are implemented: (i) event-based network growth, (ii) co-evolution between agents and concepts, and (iii) realistic low-level descriptions, especially for interactions. Events are articles, made of agents (more or less active depending on their degree k, and gathering preferentially with respect to their interests) and concepts (more or less popular, depending on their degree $k_{\mathrm{c}\to\mathrm{a}}$). Our low-level dynamics, or model of a coevolving epistemic network, consists in:

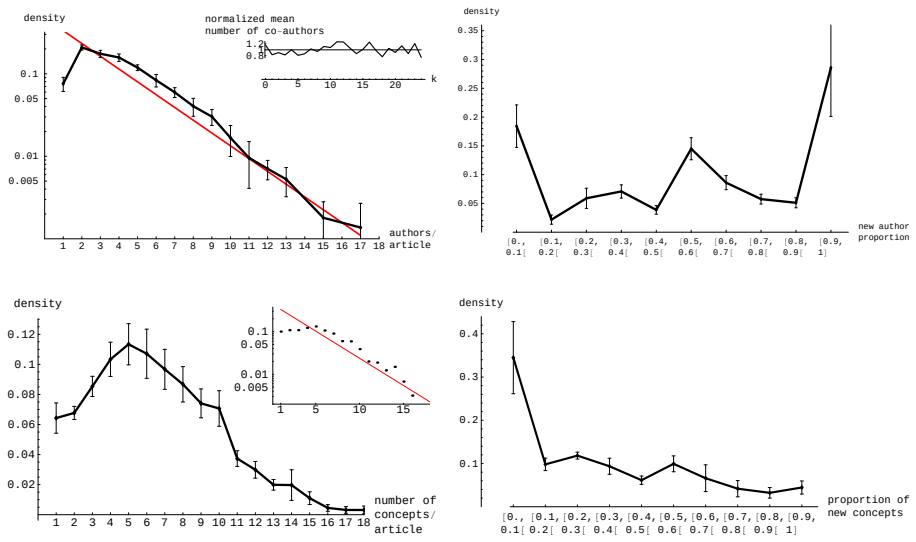

Fig. 3. *Top, left:* Distribution of the size of events, averaged on 8 periods 97-04, with confidence intervals for $p < .05$. The mean number of authors is 4.4 ($\sigma = 3.1$), and the best non-linear fit is $\propto \exp^{-\mu n}$ with $\mu = .36 \pm .06$ (straight line). The inset shows the mean number of coauthors with respect to degree k, relatively to the global mean number of co-authors: in case of independence, this ratio equals 1. *Top, right:* Proportion of new authors with respect to total authors, averaged on 7 periods (98–04) — the mean proportion is 0.55, but $\sigma = .33$ because of the tri-modal distribution. *Bottom, left:* Distributions of concepts per article — mean: 6.5, $\sigma = 3.6$. In the inset, the solid line represent the best exponential fit, $\propto e^{-\mu n}$ with $\mu = 0.29$. *Bottom, right:* Distribution of the proportion of new concepts that none of the agents anteriorly used — only for articles where there is at least one old agent. The mean is .32, with $\sigma = .28$.

1. *Creating and defining events*. n_t articles are created at each period: $n_{t+1} = n_t + n_+$. Author set and concept set sizes follow geometric laws, with observed means.
2. *Choosing authors.* Because of the tri-modal distribution articles feature either only new authors, either only old authors, or equally old and new authors, all equiprobably. If there is at least one old agent, an 'initiator' is randomly chosen proportionally to her social network degree k; then, other old agents of degree k' are picked according to $P(\mathrm{L}|k', d) = P(\mathrm{L}|k')P(\mathrm{L}|d)$, d is the semantic distance to the initiator. Finally new nodes are created.
3. *Choosing concepts.* New concepts (i.e. such that no old agent uses) are a fixed proportion of the article concept set. Other concepts are chosen from the concept set of authors. All are chosen randomly proportionally to their degree $k_{\mathrm{c}\to\mathrm{a}}$.
4. *Updating the network,* once author and concept sets are defined (Fig. 4).

Results. We ran the model for 8 periods $t \in \{1, \cdots, 8\}$, starting with an empty epistemic network — in other words, the morphogenesis starts *from scratch*. Obviously,

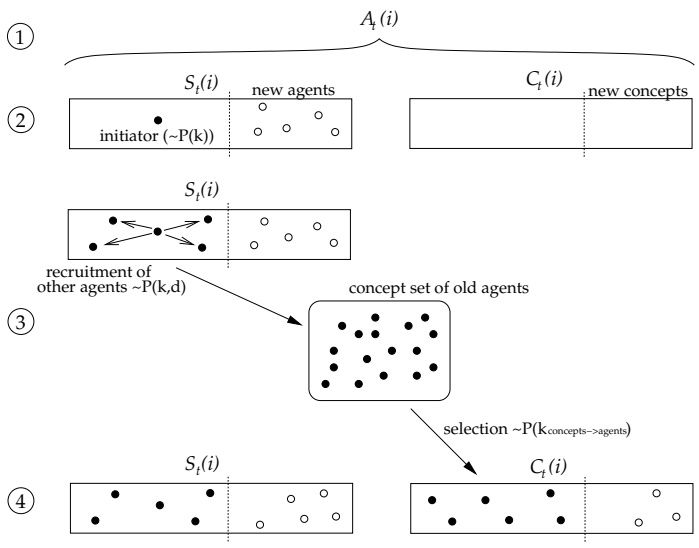

Fig. 4. Modeling an event by specifying contents of article i, $A_t(i) = (S_t(i), C_t(i))$, author and concept sets. The numbered steps indicated here follow those of the model description in Sec. 4.

periods correspond to years. One hundred new articles were to appear during the first period, with a growth rate of 100 articles per period per period: $n_1 = 100$, $n_+ = 100$. We focus on networks obtained after simulations are completed for 8 periods, and we have a satisfying adequation for every stylized fact, both in shape and in magnitude:

- *Rebuilding network size.* Simulated networks contain 10982 agents on average ($\sigma = 215$, for fifteen runs), agreeing with empirical data.
- *Rebuilding degree distributions.* Results for all four degree distributions are shown on Fig. 5, indicating a very good fit — in particular, power-law tails have a similar exponent, with a shape which fits a log-normal distribution similar to that of the empirical case.
- *Rebuilding clustering coefficients.* Clustering coefficients are accurately reproduced, as shown on Fig. 6.
- *Rebuilding epistemic community structure.* ECs have been computed (see Fig. 7) and distributions of EC sizes are close to those of the real network. Semantic distances are also correctly rebuilt, see Fig. 7.

Discussion. Hence, epistemic communities are produced by the co-evolution of agents and concepts. Not only is the high-level structure accurately reconstructed by our model, but low-level dynamics are consistent as well — this is a not a minor point: rebuilding high-level phenomena remains dubious if the low-level dynamics is incorrect. Truthfulness of descriptions must reach the higher level *as well as* the lower level. In any case, we may still wonder what weight some of our hypotheses bear towards the apparition

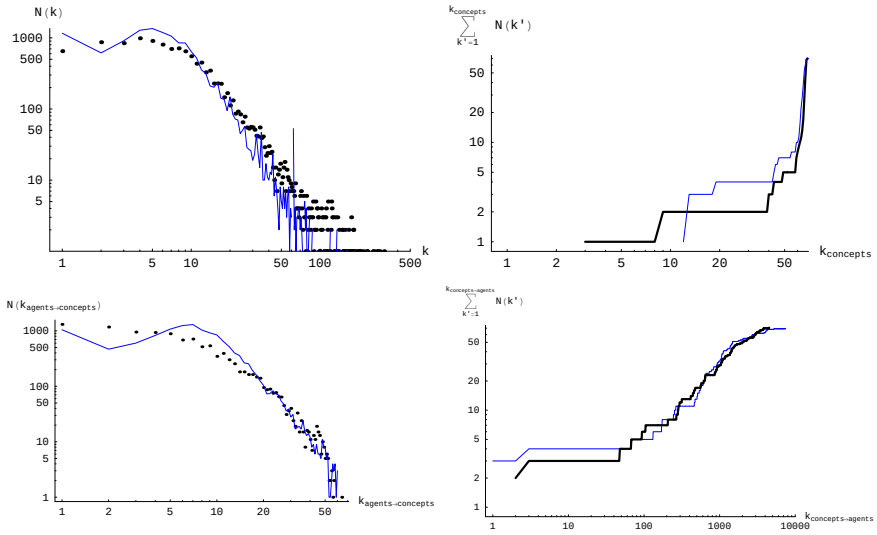

Fig. 5. Social (top-left), semantic (top-right) and socio-semantic (bottom) degree distributions. Simulation results (dots or thick line) acceptably fit empirical data (thin line). *x-axes* correspond to degrees in a given network, while *y-axes* represent the number or cumulated number of nodes having a given degree — both are log-scale.

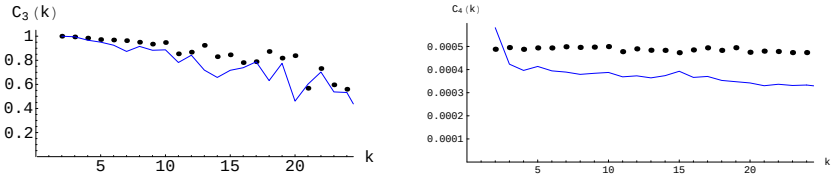

Fig. 6. $c_3(k)$ (left) and $c_4(k)$, simulated (dots) vs. empirical values (solid line). See Sec. 2 for details on these coefficients.

of high-level phenomena: is our model a minimal model *as regards the stylized facts we selected*? In particular, consider basic event-based models for social networks — which have become popular very recently among a few other authors as well [22, 26, 28] — that simply rest on n-adic events instead of dyadic interactions and that do not even specify any kind of PA. Yet, these models lead to scale-free distributions and high one-mode clustering coefficients. These results suggest that PA is not required to rebuild degree distributions and c_3, by contrast to dyadic-interaction-based models (such as BA model).

Recall that our model features (i) event-based modeling, (ii-a) degree-related preferential attachment (or activity) for the choice of agents and (ii-b) for concepts, and (iii) homophily of agents. Are the high-level stylized facts still reproduced if we loosen

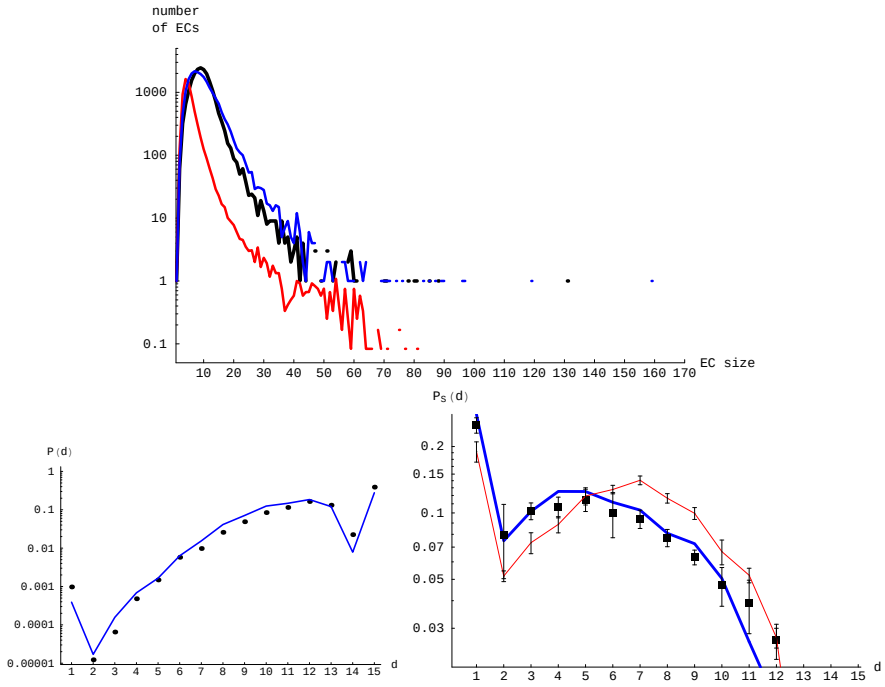

Fig. 7. *Top:* Number of ECs with respect to agent set sizes, in GLs computed for samples of 250 agents. Simulation results (thick black line, above) fit the empirical data (thin blue line, above), compared to a random "rewired" cases where degree distributions on from agents to concepts and from concepts to agents are conserved: as expected, they contain significantly less ECs, by one order of magnitude (thin red line, below). *Bottom, left:* Simulated mean distribution of semantic distances on the whole graph (dots) compared to original empirical data (line). *Bottom, right:* Same quantities, but computed only for the social neighborhood of each agent. Note the thin solid line, representing simulations not using homophily.

some of these hypotheses? Since many combinations of simplified models are envisageable, we only examine what happens when relaxing one hypothesis at a time. Yet, at least one high-level fact is not accurately reproduced when relaxing any feature of our model (event-based modeling, degree-related preferential attachment (or activity) for the choice of agents and for concepts, or homophily of agents).[1]

Conclusion

We investigated the formation of the emerging "zebrafish" scientific community and assumed that we could micro-found the evolution of the structure of this social complex system by modeling agents co-evolving with concepts. Therefore, we introduced tools

[1] Comprehensive details about these results are omitted because of length restrictions.

to estimate low-level interaction and growth processes from past data. Only thereafter could we hope for a *realistic*, *descriptive* model. The final success of the reconstruction gives credit to our hypothesis that the structure of knowledge communities is at least produced by the co-evolution of agents and concepts. *In fine*, we argued for an empirical stance when designing models: even if the model reproduces the desired stylized facts, it is essential to know whether the alleged low-level dynamics is empirically grounded.

References

1. Skyrms, B., Pemantle, R.: A dynamic model of social network formation. PNAS **97**(16) (2000) 9340–9346
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics **74** (2002) 47–97
3. Cohendet, P., Kirman, A., Zimmermann, J.B.: Emergence, formation et dynamique des réseaux – modèles de la morphogenèse. Revue d'Economie Industrielle **103**(2-3) (2003) 15–42
4. Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae **6** (1959) 290–297
5. Touhey, J.C.: Situated identities, attitude similarity, and interpersonal attraction. Sociometry **37** (1974) 363–374
6. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology **27** (2001) 415–444
7. May, R.K.: Will a large complex system be stable? Nature **238**(413–414) (1972)
8. Barbour, A., Mollison, D.: Epidemics and random graphs. In Gabriel, J.P., Lefevre, C., Picard, P., eds.: Stochastic Processes in Epidemic Theory. Lecture Notes in Biomaths, 86. Springer (1990) 86–89
9. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
10. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks — From Biological Nets to the Internet and WWW. Oxford: Oxford University Press (2003)
11. Barabási, A.L.: Linked: The new science of Networks. Cambridge, Mass.: Perseus Publishing (2002)
12. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45**(2) (2003) 167–256
13. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
14. Redner, S.: How popular is your paper? An empirical study of the citation distribution. European Phys. Journal B **4**(131–134) (1998)
15. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the Internet topology. Computer Communication Review **29**(4) (1999) 251–262
16. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286** (1999) 509–512
17. Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N.: Structure of growing networks with preferential linking. Physical Review Letters **85**(21) (2000) 4633–4636
18. Jin, E.M., Girvan, M., Newman, M.E.J.: The structure of growing social networks. Physical Review E **64**(4) (2001) 046132
19. Caldarelli, G., Capocci, A., Rios, P.D.L., Munoz, M.A.: Scale-free networks from varying vertex intrinsic fitness. Physical Review Letters **89**(25) (2002) 258702
20. Fabrikant, A., Koutsoupias, E., Papadimitriou, C.H.: Heuristically optimized tradeoffs: A new paradigm for power laws in the internet. In: ICALP '02: Proceedings of the 29th International Colloquium on Automata, Languages and Programming, London, UK, Springer-Verlag (2002) 110–122
21. Boguna, M., Pastor-Satorras, R.: Class of correlated random networks with hidden variables. Physical Review E **68** (2003) 036112

22. Peltomaki, M., Alava, M.: Correlations in bipartite collaboration networks. arXiv e-print archive **physics** (2005) 0508027
23. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: Stochastic models for the web graph. In: IEEE 41st Annual Symposium on Foundations of Computer Science (FOCS). (2000) 57
24. Berger, N., Borgs, C., Chayes, J., D'Souza, R., Kleinberg, R.: Competition-induced preferential attachment. In: Proceedings of the 31st International Colloquium on Automata, Languages and Programming. (2004) 208–221
25. Carayol, N., Roux, P.: Micro-grounded models of complex network formation. Cahiers d'Interactions Localisées **1** (2004) 49–69
26. Ramasco, J.J., Dorogovtsev, S.N., Pastor-Satorras, R.: Self-organization of collaboration networks. Physical Review E **70** (2004) 036106
27. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Physical Review E **68**(036122) (2003)
28. Guimera, R., Uzzi, B., Spiro, J., Amaral, L.A.N.: Team assembly mechanisms determine collaboration network structure and team performance. Science **308** (2005) 697–702
29. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS **99** (2002) 7821–7826
30. Latapy, M., Pons, P.: Computing communities in large networks using random walks. arXiv e-print archive (2004) 0412568
31. Newman, M.E.J.: Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E **64** (2001) 016131
32. Barabási, A.L., Jeong, H., Ravasz, E., Neda, Z., Vicsek, T., Schubert, A.: Evolution of the social network of scientific collaborations. Physica A **311** (2002) 590–614
33. Redner, S.: Citation statistics from 110 years of physical review. Physics Today **58** (2005) 49–54
34. White, D.R., Kejzar, N., Tsallis, C., Farmer, D., White, S.D.: A generative model for feedback networks. Physical Review E **73** (2006) 016119
35. Ravasz, E., Barabási, A.L.: Hierarchical organization in complex networks. Physical Review E **67** (2003) 026112
36. Newman, M.E.J., Strogatz, S., Watts, D.: Random graphs with arbitrary degree distributions and their applications. Physical Review E **64**(026118) (2001)
37. Guillaume, J.L., Latapy, M.: Bipartite structure of all complex networks. Information Processing Letters **90**(5) (2004) 215–221
38. Lind, P.G., Gonzalez, M.C., Herrmann, H.J.: Cycles and clustering in bipartite networks. Physical Review E **72** (2005) 056127
39. Roth, C., Bourgine, P.: Epistemic communities: Description and hierarchic categorization. Mathematical Population Studies **12**(2) (2005) 107–130
40. Roth, C., Bourgine, P.: Lattice-based dynamic and overlapping taxonomies: The case of epistemic communities. Scientometrics **69**(2) (2006) 429–447
41. Batagelj, V., Bren, M.: Comparing resemblance measures. Journal of Classification **12**(1) (1995) 73–90
42. Snijders, T.A.: The statistical evaluation of social networks dynamics. Sociological Methodology **31** (2001) 361–395
43. Powell, W.W., White, D.R., Koput, K.W., Owen-Smith, J.: Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. American Journal of Sociology **110**(4) (2005) 1132–1205
44. Newman, M.E.J.: Clustering and preferential attachment in growing networks. Physical Review Letters E **64** (2001) 025102
45. Jeong, H., Néda, Z., Barabási, A.L.: Measuring preferential attachment for evolving networks. Europhysics Letters **61**(4) (2003) 567–572