Big Data and the Collective Turn in Biomedicine

How Should We Analyze Post-genomic Practices?

Alberto Cambrosio
McGill University

Pascale Bourret
Aix-Marseille Université, UMR SESSTIM

Vololona Rabeharisoa
Mines-ParisTech

Michel Callon
Mines-ParisTech

Abstract: We presently witness a profound transformation of the configuration of biomedical practices, as characterized by an increasingly collective dimension, and by a growing reliance on disruptive technologies that generate large amounts of data. We also witness a proliferation of biomedical databases, often freely accessible on the Web, which can be easily analyzed thanks to network analysis software. In this position paper we discuss how science and technology studies (S&TS) may cope with these developments. In particular, we examine a number of shortcomings of the notion of networks, namely those concerning: (a) the relation between agency and structural analysis; (b) the distinction between network clusters and collectives; (c) the (ac)counting strategies that fuel the networking approach; and (d) the privileged status ascribed to textual documents. This will lead us to reframe the question of the relations between S&TS and biomedical scientists, as big data offer an interesting opportunity for developing new modes of cooperation between the social and the life sciences, while avoiding the dichotomies – between the social and the cognitive, or between texts and practices – that S&TS has successfully managed to discard.

Keywords: big data; network analysis; post-genomic medicine; bio-clinical collectives; actor-network theory.

Corresponding author: Alberto Cambrosio, Department of Social Studies of Medicine, McGill University, 3647 Peel St., Montreal (QC) Canada H3A 1X1 – Email: alberto.cambrosio@mcgill.ca

1. Introduction

This is a position paper. It discusses how science and technology studies (S&TS), confronted with recent changes in the configuration of biomedical practices – in particular their increasingly collective dimension, and their reliance on disruptive technologies, such as microarrays and next-generation sequencing, that generate large amounts of data – may cope with these developments. Big data represent a (multifaceted) source of information for both S&TS scholars and health care practitioners, while also being the outcome of activities predicated upon the involvement of a large number of heterogeneous actors. As such, they are embedded in biomedical practices and have become key elements of knowledge production, especially in domains such as genomics, where they engender distinctive forms of evidence.

The dual nature of big data – they act as sources of information while also being the outcome of activities that are constitutive of biomedical practices – is not something new. Scientific texts (articles, books, reports) partake in scientific knowledge production, while simultaneously acting as a data repository for the natural scientists who produce and use them. As sources of evidence, they are also of use to social scientists who engage, for instance, in scientometric analyses of the socio-cognitive structure of science, or to historians of ideas investigating the dynamics of a given domain. S&TS scholars have successfully learned how to tame this multi-dimensional nature of scientific texts by displaying the links they entertain with scientific practices, without falling into the dichotomy between the social and cognitive dimensions of science. Big data, however, raise a novel and difficult challenge, for two main reasons. First, because we have only limited evidence concerning their actual use as part of research practices (but see, e.g., Leonelli 2012, 2013, 2014, and Edwards 2010 for noteworthy exceptions), which in turn leaves social scientists wondering how they should understand *and* use them. And second, because the "big" in big data refers not simply to the sheer quantity of data available, but also to their instability, heterogeneity, and proliferation into different domains. In other words, we presently face a dual task: on the one end, we need to better understand the research activities that rely on the production and analysis of big data, and, on the other hand, we need to figure out how science studies scholars can embed big data, and the configurations they generate, into their own practices, and what are the consequences of doing so. In particular, we should be wary of solutions that may end up reintroducing the dichotomies – between the social and the cognitive, or between texts and practices – that S&TS has successfully managed to dispense with. The present text explores a few of the issues and problems involved in such an endeavor.

Big data are everywhere, and thus the issues discussed in this text are not confined to S&TS. Rather, big data represent a more general challenge for the social sciences because they raise the following conundrums:

How should we, as social scientists, use them in our own investigations while taking into account the fact that they also partake in the activities of the actors we investigate, and cannot therefore be considered as unproblematic evidence? How can we revisit, in the light of the growing importance of big data, the traditional tension between qualitative and quantitative approaches, or between local ethnographies and cross-sectional studies? We are particularly interested in those situations in which both the social scientists and the actors they investigate attribute a strategic role to the notion of network, often generically defined. We will center this paper on the theoretical and visualization issues engendered by this notion. In a first section, we will briefly discuss the development of big data in the oncology domain, showing that they have become part and parcel of recent developments in this advanced biomedical domain. This will lead us, in a subsequent section, to examine how the notion of network plays a strategic role in this context. While this notion has, of course, enjoyed a staggering success within S&TS, we will focus on its shortcomings, and in particular on four thorny issues, namely: (a) the relation between agency and structural analysis; (b) the distinction between network clusters and collectives; (c) the (ac)counting strategies that fuel the networking approach; and (d) the privileged status ascribed to textual documents. We will explore how these shortcomings can be overcome, at least tentatively. In turn, this will lead us to reframe the question of the relations between the subjects and objects of observations, i.e., between S&TS and biomedical scientists. Big data, as it turns out, may offer an interesting opportunity for developing new modes of cooperation between social and life scientists.

2. 21st Century Biomedicine: Clinical Wards, Wet Labs, and Bytes

In his address to the 2011 meeting of the American Society of Clinical Oncology – with nearly 35,000 members, most likely the single largest professional organization in its domain – the Society's president, George Sledge, warned fellow oncologists about the upcoming "tsunami" of genomic information that was likely to result from a sharp decrease in the cost of sequencing tumors. He added: "When data are that cheap, every patient's cancer will be informative for tumor biology […] and things will get very, very complicated" (cited in Goldberg 2011). That same year, and along similar lines, in a promotional video for the European Multidisciplinary Cancer Congress, entitled: "Bench, bedside, 'bytes' and back" (also referred to as the three Bs), noted clinical researcher Anne-Lise Børreson Dale explained: "You start with the bed, you have the patients, and then you go to the bench, and then because we create so many […] huge amounts of data, you have bytes, as in gigabytes, and then you

go back to the bench to find out what is the right treatment for that patient, and then you go to the patient again […] it's like a spiral that goes up […] every patient is sort of an experiment for the next who will be coming in"[1]. These two quotations are far from uncommon. They illustrate recent themes and trends in oncology, namely the rise of translational research, closely combining biological and clinical investigations; the search for personalized treatments, whose horizon lies in the singularization (Callon 2012) of patients; and, finally, the premises upon which the previous two items are predicated, namely the availability of large sets of data, whose proliferation, accumulation and heterogeneity raises major interpretative challenges.

We will return in subsequent sections to the collective dimension of contemporary biomedicine, in particular translational research, as exemplified, for instance by large-scale genomic consortia – e.g. the "Breast Cancer Linkage Consortium" that mobilized approximately 100 centres[2], or the "Autism Genome Project" that mobilized "120 scientists from more than 50 institutions across 19 countries" (Szatmari et al. 2007) – or, perhaps more mundanely, the staging of large-scale, national and international clinical trials (Keating and Cambrosio 2012a), although we should hasten to add that, as we will see, the term "collective" does not refer simply to number and size. For now, let us examine the issue of big data that is related to, but not identical to the former topic. The generation and mining of large data sets is by no means an uncontroversial activity. For instance, MIT biologist Michael Yaffe (2013) recently claimed that while "the sequencing of human tumours [has] produced important data sets for the cancer biology community […] these studies have revealed very little new biology", further complaining that scientists were "addicted to the large amounts of data that can be relatively easily obtained [by genome sequencing], even though these data seem unlikely, on their own, to unveil new cancer treatment options or result in the ultimate goal of a cancer cure" (Yaffe 2013, 1). The important point, as far as we are concerned, is of course not whether Yaffe's criticism is warranted. Rather, our claim is that arguments both in favour and against the turn to big data confirm the fact that it has come to occupy a central place in contemporary biomedicine. The relevant issue, thus, is to examine what it involves in terms of rearranging the flow of biomedical practices.

This paper is part of a special issue entitled: *From Bench to Bed and Back*. The synecdoche in the title refers to translational research, as characterized by close relations between laboratory research (bench) and clinical work (bed). The "back" adverb marks a rejection of the unidirectional model of translation, as both the clinic and the laboratory

[1] Video retrieved on Feb 5, 2014 from: http://ecancer.org/conference/101-emcc-2011/video/891/bench--bedside-----bytes----and-back--a-virtuous-cycle-of-knowledge--1-5.php

[2] See http://www.humgen.nl/lab-devilee/bclchome.htm

can be the starting point of a successful translation. We go further and argue that rather than a relation or interface between two poles, translational research corresponds to a new, emerging site, characterized by the presence of distinctive activities. As argued in the previously quoted statement by Børreson Dale, in addition to benches and beds this site includes a third element, "bytes", or, in other words, a new kind of data and a new kind of practice, bioinformatics, needed to make sense of them. Bioinformatics is the "new kid on the block" of biomedical research[3], and, as described elsewhere (Keating and Cambrosio 2012b) has entertained somewhat controversial relations with the older data-processing specialty, biostatistics. For our present purpose the main issue is that by introducing bioinformatics, the rules of the game have changed. For bioinformatics cannot be reduced to the computerization of biology; rather, it involves a rearrangement of biological practices, a redefinition of what counts as valuable biomedical work (Yaffe's aforementioned criticism is a symptom of this process), and it shapes the kind of knowledge emerging from the translational research domain. As a bioinformatician put it: "We're not bioinformaticians who dabble in breast cancer". Instead, he and the members of his lab are "focused on understanding the disease"[4]. Understanding means reframing it, using the "new quantitative methods – the methods of the New Biology"[5].

Let us take as an example the development of a gene expression signature to predict clinical outcome in breast cancer (Finak *et al.* 2008). The researchers collected breast cancer tissue from 73 patients, used painstaking laboratory methods (laser capture micro-dissection) to pre-process the samples, and analyzed them with genomic tools in order to develop a candidate signature. For the subsequent stage, however, which involved the validation of the signature with independent samples, they no longer used local biological samples but, rather, resorted to publicly available data sets downloaded from institutions located in Amsterdam, Oxford, Rotterdam, and Uppsala. The development of the signature, in other words, was made possible by a hybrid approach that combined a "wet lab" analysis of local biospecimens with virtual testing using data sets available for download from the Internet. This is by no means an exceptional situation. If we take, for instance, MINDACT, a very large (several thousand patients), multi-center European breast cancer clinical trial testing another genomic signature, we find two parallel flows of material and data. Participating centres will ship different kinds of biological material (frozen and fixed tissue, RNA, and serum/blood) to central bioreposito-

[3] On the emergence and development of bioinformatics see McMeekin *et al.* (2002, 2004).

[4] Interview, February 14, 2011.

[5] Committee on a New Biology for the 21st Century: "Ensuring the United States Leads the Coming Biology Revolution", *A New Biology for the 21st Century*, (Washington: National Academies Press, 2009).

ries located at cancer institutes in Amsterdam and Milan, and at the biotech company that commercializes the signature. A parallel, web-based circuit will channel clinical and laboratory data from and to the participating centres, and store them in databanks located at the trial sponsor's secretariat, a Swiss bioinformatics institute, and the biotech company (in each case, with different rules for access). As recently forecasted by a leading French oncologist, the databases generated by the first generation of biomarker-driven clinical trials should lead the production of algorithms propelling the design of a second generation of trials, which will in turn generate databases, and so on (André n.d.). In the meantime, this kind of data is becoming increasingly available, as shown, for instance, by the recent announcement that the International Cancer Genome Consortium has made publicly available data from thousands of cancer genomes.

Bioinfomatics is not confined to the handling of data produced by the new genomic technologies: it is constitutive of them. Let us take the example of gene expression profiling (GEP) mentioned in the previous paragraph. One of the key technologies of post-genomic oncology, gene expression profiling, has generated new entities, such as multi-gene "signatures", that have simultaneously been developed in clinical, laboratory and commercial biotech settings (Kohli-Laven *et al.* 2011). Figure 1 reprinted from an article that analyses the development of this field (Cointet *et al.* 2012) uses a modified version of a scientometric technique called "co-citation analysis".

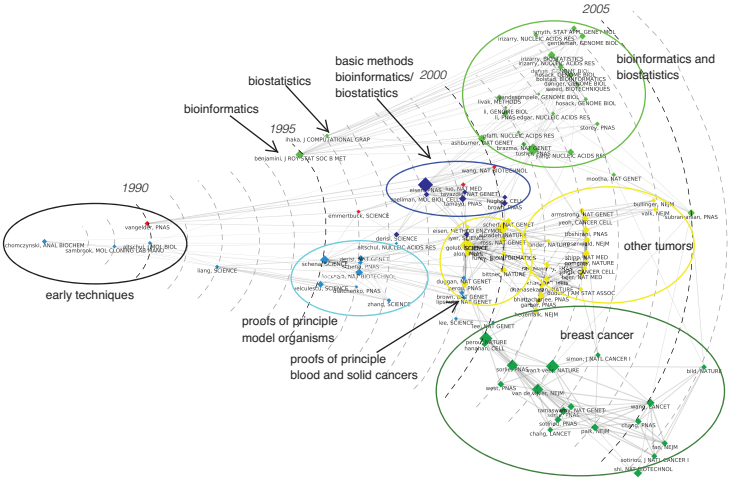

Fig. 1 – Co-citation analysis of the development of gene expression profiling: see text for explanations. Source: modified version of Figure 2 in Cointet *et al.* (2012).

Briefly, two articles are co-cited if they appear together in the list of bibliographic references of another article. Networks of highly co-cited articles display key contributions to a field, and can be equated to its cognitive substructure. After downloading over 16,000 GEP references from the biomedical database *PubMed*, the authors used the software platform *CorTexT* (www.cortetxt.net) to generate a map of the most frequently co-cited references. Each node of the network corresponds to a reference (labeled by the first author's name and journal abbreviation), the size of the node being proportional to the number of citations. The network is arranged chronologically, with time flowing from left to right. Rather than a professional historical narrative, it provides an account of the development of the field as perceived by the authors of articles at a given point in time (in the present case, during the 1990-2010 time window). Using different time windows, the resulting map would be different, as actors will redefine the foundations of their domain: the "historicity" of chronological sequences, in other words, will be displaced by the "historiality" of science reshaping its past (Rheinberger 1997). Clusters of closely associated references organize themselves into specific subdomains that are automatically detected by a clustering algorithm and color-coded accordingly. For further clarification we have added to the original map a number of tags identifying the nature of the activities of each cluster.

Here is a quick summary of the most relevant features of the map (readers can refer to the original article for more information). The oldest references correspond to the basic molecular biology techniques that are held to provide a basis for the subsequent development of GEP. They lead to two clusters of "proof of principle" articles, i.e. demonstrations that GEP did actually work: this was done first with non-medical model organisms, and then with human tumor specimens, thus entering the clinical domain. At approximately the same time we notice a cluster of articles corresponding to biostatistical and bioinformatic methods, in particular heat maps and hierarchical clustering techniques (Wilkinson and Friendly 2009), which are needed to analyze the large data sets produced by GEP. In the case of GEP as with other recent biomedical techniques, there is no such thing as "raw data", strictly speaking, as the data generated by the instruments are already highly processed, while meaningful (i.e., interpretable) results necessitate further statistical and visual manipulations (Cambrosio and Keating 2000). Hence the mutually constitutive relation entertained by wet-lab and data analysis tools. In the most recent period we see the deployment of GEP in the oncology domain, with a strong presence of breast cancer as a distinctive cluster, concurrent with the further development of robust bioinformatic and biostatistical methods. Interestingly enough, references included in this latter cluster refer back to two founding articles, one in biostatistics (on false discovery rates) and one in bioinformatics (on the R language), cor-

responding to the hybrid (and, as previously mentioned, somewhat controversial) nature of this emergent domain.

Mimicking, at a far smaller scale, the collaborative dynamics we saw in the GEP domain between clinical and bioinformatics researchers, the Cointet *et al.* (2012) article exemplifies a collaborative endeavor between social scientists and informatics specialists, in the present case the developers of the *CorTexT* platform. This is why, to cite our own (admittedly anecdotal) evidence, while more traditional social science audiences often experience difficulties in understanding the network slides we present at talks and conferences, natural scientists can readily relate to them, in particular when, as part of our fieldwork, we ask them to comment on the maps corresponding to their activities (Bourret *et al.* 2006). We can now apply Yaffe's aforementioned critical questions to ourselves: are S&TS analysts also becoming addicted to big data? To what extent does the motley of newly available data sources contribute to a renewal of the S&TS research agenda?

3. Problematizing Network Analysis

At this point readers will have noticed that we are entering reflexivity's territory, as the techniques used to produce a map like the one displayed by Figure 1 overlap with those used in the bioinformatics references displayed on the map. While social network analysis has been around for long time, network analysis has been recently transformed by an inflow of mathematical and modeling approaches originating from the physical and life sciences (Watts 2004). Supported by a staggering increase in computer power, these new approaches have found a privileged domain of application in the scientometric analysis of the scientific literature, in particular co-authorship patterns (e.g. Newman 2004), thanks to a parallel development, namely the increasing availability on the Internet of large databases of scientific publications such as Medline (and its search engine *PubMed* freely accessible since 1997), *Web of Science*, *Scopus*, and *Google Scholar*. Traditionally, social network analysis examined social ties between a relatively small number of actors, often derived from *ad hoc* procedures such as interviewing selected actors about their connections or resorting to sampling (Scott 2000). Large-scale bibliographic databases now allow, at the click of a mouse, to obtain information about relational patterns, such as co-authorship, between millions of actors. But these new possibilities come at a price. The fact that a reflexivity loop seems to exist at the level of tools does not necessarily imply that a similar loop should necessarily obtain in terms of conceptual framing. Put otherwise: the fact that scientists can easily relate to maps created by network sociologists can be a positive aspect, but also a symptom of looming problems.

The large databases, the search engines that have been developed to

exploit them, and the data-mining, text-mining, and network-analysis tools that S&TS scholars use to process the resulting data, do indeed give access to unprecedented amounts of information and lead to stunning visuals (Lima 2011). We should not forget, however, that they have not been conceived primarily for sociological analysis. As they emerge from the physical and life sciences – sometimes transiting through the newly established specialty of "information science" (Börner 2010) – they come with built-in epistemological assumptions and models that are seamlessly carried over into the social sciences when they are recycled for use by S&TS scholars. Faced with the sterile alternative of either embracing these new approaches without too many qualms because of their striking effectiveness, or of rejecting them for fear of contamination, we prefer a third alternative, namely to explore the issue of the adequacy between these newly available tools and S&TS research agendas.

The notion of network has provided a key heuristic tool for developing a research program that rejects both technological and sociological determinism, and can thus be put to fruitful use for the analysis of biomedical activities, but this notion is now a victim of its own success. We find it everywhere, within and outside biomedicine, as the term is used for every purpose, from the mundane to the specialized. The expansion of its semantic field, in parallel with the steady increase in the offer of affordable data-mining software and network visualization tools, has resulted in the development of a "network lingo" and of standardized interpretations that are indistinctly applied to substantive, methodological and conceptual issues. To further complicate the situation, the adoption and deployment of network analysis tools have by and large taken place within quantitative domains such as scientometrics and, most recently, information science and informetrics, whose development, in spite of their focus on scientific and technical activities, has only occasionally intersected with conceptual developments in S&TS. Only rarely have these quantitative approaches been interfaced with ethnographic methods (for exceptions see Velden and Lagoze 2013; Navon and Shwed 2012; Bourret *et al.* 2006; Cambrosio *et al.* 2004), but, most often, their production within self-contained professional circles of information specialists has resulted in the offer of tools in search of possible uses (for a recent example, see Skupin *et al.* 2013)[6].

As argued by Michel Callon (2001), thick ethnographic descriptions of individual field sites are ill suited to deal with large-scale collaborative endeavors such as the ones discussed in the previous section. The alternative of reducing such endeavors to a few quantitative indicators is equally unsatisfactory, insofar as it destroys for all practical purposes the very phenomena under investigation. The newly available network analysis tools, in combination with more traditional fieldwork methods, seem to

[6] For earlier examples see the special issue of "Proceedings of the National Academy of Sciences" on "Mapping knowledge domains" (2004; 101, suppl 1).

offer a partial response to this predicament, provided they avoid the limitations of traditional social network analysis. These limitations include an exclusive focus on human actors, and the assumption of the existence of a unified social space within which social ties can be properly measured and described. In a subsequent English version of the 2001 paper, Callon (2006) revisited this issue by postulating that network analysis tools should avoid two pitfalls. First, the aforementioned assumption that actors' interactions take place within a unified space; this assumption belies the existence of a multiplicity of regimes of engagement deployed in different, more or less overlapping spaces (Boltanski and Thévenot 2006; Moreira 2012). A second pitfall lies in the a priori categorization of entities according to a number of pre-set, analyst-defined attributes. In contrast with this approach stands a focus on the emergent categories generated by the relational ties that human and non-human entities establish between each other. By taking into account the heterogeneity of networks (both in the sense of consisting of different entities *and* of corresponding to different regimes of engagement) social scientists can enter in a reflexive relation with the entities they analyze. Such a reflexive relation can itself be of different kinds. It has a substantive dimension, as actor-generated categories and, more generally, the framing they produce, will often question the analyst's assumptions about the proper categories that constitute the world, and his/her epistemological privilege to define them. It also has a methodological dimension, because of the aforementioned, increasing overlap between the network analysis tools developed by natural scientists and those used in the social sciences.

Still, while one should not mistake the co-authorship "network" generated by a few clicks on the Internet for the "network" of actor-network theory (ANT) (Latour 2011), the new tools offer interesting opportunities for the empirical exploration of new techno-scientific configurations, using the conceptual avenues opened-up by ANT. It should be noted, in this respect, that the founders of ANT were among the pioneers of mapping approaches, in particular co-word analysis (Callon *et al.* 1986). These initial attempts have been criticized for their alleged reductionism with regards to the issue of agency, and for lending themselves to structural interpretations. In the meantime, several versions of ANT have been developed that are not always mutually compatible. On the one hand, in response to the aforementioned criticism, there have been attempts to revisit the processes previously analyzed solely in terms of networks by using notions such as regimes and assemblages, or collectives and arrangements. From this perspective, visualization tools can become problematic, and do in fact partake of the emergence of new regimes of innovation that S&TS should investigate rather than adopt blindly (Callon 2012; Rabeharisoa *et al.* 2014). On the other hand, and in spite of their acknowledged limitations and shortcomings, navigational practices that are made possible by the availability of large databases and software tools initially devised to investigate complex systems, are seen as creating

the conditions of possibility for a new kind of generalized social theory, one that could dispense with the opposition between individuals and aggregates (Latour 2011; Latour et al. 2012).

In the present paper we adopt a position closer to the first alternative in order to explore some of the problems raised by the new visualization tools and to discuss, using examples from recent studies of biomedical practices, how we can partly address them. These problems fall in at least four different categories:

- As previously mentioned, while network analysis algorithms are in principle well adapted to the kind of relational sociology embraced, among others, by ANT, they tend to reify the notion of network and to convey structural or strategic interpretations of specific network configurations. Typical examples include analyses in terms of structural holes, obligatory passage points, centrality, etc. The issue thus becomes: Is it possible, and if so how, to interpret maps without resorting to a vocabulary that is derived from structural and strategic analysis? A major obstacle, in this respect, is that 'structure' is embedded into the very production of maps; for instance, the algorithms used to position nodes rely on structural properties, such as symmetry, structural equivalence of points, centrality and 'betweenness' of nodes. Put otherwise: does network analysis allow us to make inferences about the dynamics of a given domain without reducing it to changes in the morphology of the network? Or should we rather opt for a hybrid approach, whereby networks will no longer represent the ultimate analytical horizon, but a tool to better investigate assemblages, or, to use a term that avoids mechanical implications and reintroduces agency, *agencements* (Callon 2013; see also Rheinberger 2009 for the case of biomedicine)? While shifting the conceptual and substantive focus from networks to *agencements*, such a move would still leave room for networks, as they add flexibility, dynamics, but also some amount of ordering to *agencements*.
- In order to make sense of a network, as already hinted in the case of Figure 1, analysts (or the algorithms that replace them) trace boundaries around clusters of closely connected nodes. The sociological relevance of these (formally defined) clusters is itself open to interpretation, as they do not necessarily correspond to taken-for-granted groups or institutions: in fact, if and when they do (which is probably more often the case with homogeneous social networks than with heterogeneous ones), the heuristic interest of tracing a network decreases correspondingly, as it transmutes from being an investigational tool able to produce surprises to a redundant illustration of well-known arrangements. If they do not, we then face the issue of the collective agency of the heterogeneous clusters displayed on maps. When adopting a structural interpretation, this issue is most often swept under the carpet. A closely related issue, similarly overlooked by structural interpretations, has to do with situations in

which the transformation of the entities making up a heterogeneous collective are not the consequence but, rather, the cause of the dynamics of these collectives. Here again, the path forward may necessitate a shift in focus from networks per se to the processes involved in producing specific *agencements* that account for the heterogeneous and distributed nature of collective agency.

- As already mentioned, network analysis, because of its figurational dimension, can be seen as a healthy alternative to the statistical reductionism of quantitative indicators. It also partakes, however, of the quantitative domain, as networks are firmly embedded in a metrological infrastructure. The point is not to contrast qualitative with quantitative analysis, as in the longstanding conflict within professional sociology, but to signal that the modality of action that underlies networks is to "add up", to be "counted in". Other modalities are possible, such as qualifying links instead of accumulating them. The "adding up" strategy, as exemplified most obviously by citation counts, is embedded in a number of databases whose goal is precisely to make things (ac)countable in this specific way. The seamless production of networks derived from these databases brackets the very infrastructure that makes those data, and their relational nature, available and witnessable. From this point of view, networks have no epistemological privilege, as they are one among possible forms of interpretation and enactment of 'the social'. How, then, to integrate this aspect in our analysis? The maps we produce bear the invisible traces of the strategies deployed by data providers: how can we make them visible and, most importantly, take them into consideration when interpreting our results?
- Most often than not, the components of a network are obtained by analyzing bibliographic databases (articles, patents, etc.), repositories of full-text articles, blogs, and other textual documents. While, given its focus on the materiality of practices, non-textual elements, in combination with textual ones, play a key role in ANT analyses, only the latter, or at least entities mediated through inscriptions, end up in the maps. How, then, to convey the heterogeneity of networks when we can only produce and access them via textual inscriptions?

In what follows we revisit these issues – the reductionist understanding of agency resulting from strategic/structural interpretations of networks, their limited capacity to account for the dynamics of collectives, their actuarial nature that privileges quantity over content, and their exclusive reliance on texts. We focus on the first two elements using a few concrete examples.

4. Network Dynamics

Both from a methodological and theoretical point of view, accounting for network dynamics has been one of the major stumbling blocks of this kind of analysis. Change has mostly been interpreted as structural change. A notion such as 'obligatory passage point' equates a given position within a network with processes of circulation, displacement or movement. Dynamics is thus reduced to the distribution of points and their relations in a (virtual) space. The agency of the entities represented in a network is mechanically conflated with their structural/strategic positioning, and since the capacity to act strategically and reflexively is generally ascribed solely to humans, it is not surprising that *social* network analysis still occupies center-stage. Methodologically speaking, attempts to account for dynamical processes often rely on the structural comparison of the 'same' network at different times, pointing to the elements that are held responsible for the observed changes. Algorithms can be used to identify the entities (actors or groups thereof) that are at the origin of structural transformations.

A possible, although not entirely satisfactory way out of this predicament is to opt for interpretations focusing on events, i.e. to 'play' with the content of maps[7] by taking into account the heterogeneous roots of a network's dynamics. A structural reading, when comparing maps corresponding to different periods (say: t1 and t2), focuses on networks characterized by the presence of the same kind or category of entities, e.g., academic researchers, clinicians, biotech or pharmaceutical companies, either individually or as members of homogenous subdomains. To account for change, analysts will for instance point to the role of biotech companies that while only playing a marginal role at t1, have become key intermediaries between public and large private organizations at t2. This kind of account is characterized by the presence of a strong and sophisticated human agency: observers easily acknowledge the key role of biotech companies (or, rather, the entrepreneurial skills of their managers), but are less keen to attribute a similar role to cells and molecules. A non-structural reading will opt for a different approach: to account for the difference between t1 and t2 we should consider the role of entities that were absent from the original t1 and t2 maps, i.e. produce complementary maps that include cells, instruments, molecules or diseases. In other words, the passage from a homogeneous network at t1 to a homogeneous t2 network can in fact be accounted for by the presence of a number of heterogeneous entities that did not appear on the initial maps: the emergence (or disappearance) of connections between two groups of researchers is not reducible to the sole agency of other researchers; it involves the

[7] As the very notion of a 'map' lends itself to structural interpretations, we should opt for a term with different undertones.

simultaneous agency of biomedical entities such as mutations, antibodies, or cells.

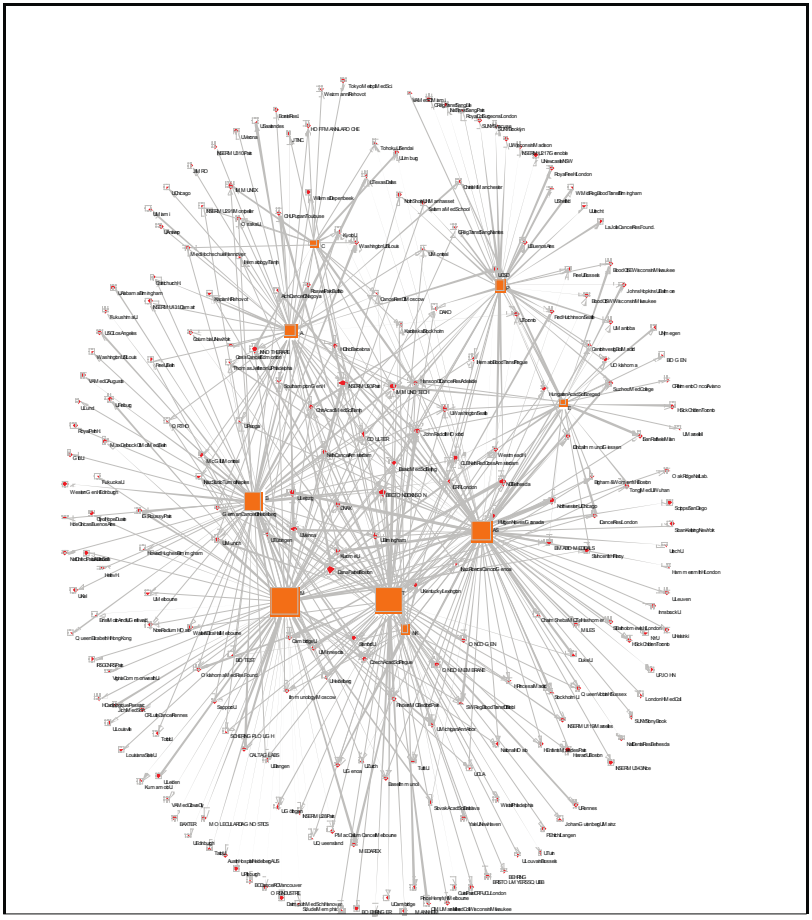

Fig. 2 – Map of laboratories producing monoclonal antibodies targeting different categories of cells: see text for explanations. Source: modified version of Figure 7 in Cambrosio *et al.* (2004)

The following example is taken from a paper (Cambrosio *et al.* 2004) that, in the wake of ethnographic fieldwork on the emergence and circu-

lation of a new kind of reagents known as monoclonal antibodies, attempted to visualize the regulatory infrastructure that resulted in their generalized use. This infrastructure emerge from the establishment of equivalences between individual antibodies produced by different laboratories around the world: antibodies that were held to be the same, in spite of their different institutional or geographical origin, were assigned a same CD (cluster designation) number and could be used interchangeably. In the present case, the authors used an *ad hoc* database of substances and laboratories that they found on the Web, rather than a bibliographic database such as *Medline*. Figure 2 considers two kinds of entities: individual laboratories or companies (round red nodes), and the general category of cells (T-cells, B-cells, etc.) targeted by antibodies (square orange nodes: their size is proportional to the number of antibodies available for that category). A structural interpretation will focus on the positioning of the laboratories vis-à-vis these general cell categories, as the latter correspond to specific biomedical domains (of varying importance as shown by the size of the nodes). The organizations at the center of the map (including all major commercial companies in that field) position themselves strategically, in order to ensure their presence throughout the spectrum of biomedical activities, whereas organizations at the periphery of the map, while aiming to profit from the scientific and/or commercial opportunities offered by this new technology, have adopted a specialization or niche strategy. The original article included maps corresponding to different points in time, thus arguably allowing readers to follow the evolution of these strategies.

Figure 3, in contrast, disaggregates, so to speak, the previous figure by including the same institutions (square orange nodes) and the specific CD antibodies they had developed (round red nodes): the size of the nodes corresponds to the number of antibodies produced by a given organization or included in the same CD. Figure 3 can no doubt also be interpreted structurally (e.g., large vs. specialized producers of widely used vs. esoteric CD antibodies), but a non-structural interpretation will insist on the evolution of the links between researchers and entities in this rapidly unfolding domain. For instance, it appears that some CDs are very robust, as their existence is supported by several laboratories, whereas others are weak, as their existence is ensured by the presence of only one laboratory. Moreover, maps from different periods (not shown here: see original article) document the emergence of novel categories of cells in conjunction with the proliferation of antibodies targeting them, or the transformation (splitting, redefinition, etc.) of individual CDs.

Admittedly, the alternative illustrated by this example still conveys aspects and elements of a structural interpretation, only alleviating its worst shortcomings. This is due in large part to the limits of the database that only listed a limited number of different entities. Moreover, the database did not provide indications about the informational content of the antibodies, i.e. the domains, tests or diseases for which they were deemed to

be relevant. Combining data from different databases could circumvent this difficulty, an approach exemplified in practice (but with a quite different intent) by Boyack *et al.* (2004) who in the case of melanoma research analyzed a data set consisting of papers from *Medline*, genes derived from the *Entrez Gene* database, and proteins from the *UniProt* database. Similarly, but using different techniques and with a different perspective, Mogoutov *et al.* (2008) explored the development of microarrays by combining data derived from *Web of Science* articles, with those from the *CRISP* database of research grants awarded by the US National Institutes of Health (NIH), and patents from the US Patent office and the *Derwent Innovation Index*.

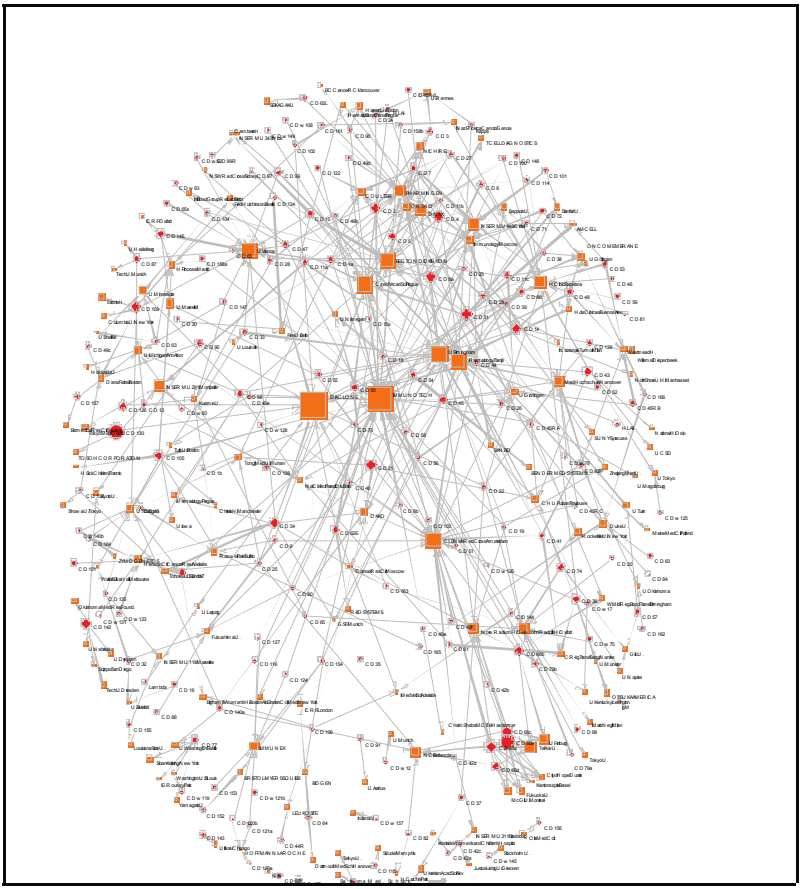

Fig. 3 – Map of laboratories and equivalent categories of monoclonal antibodies. Source: modified version of Figure 12 in Cambrosio *et al.* (2004).

Readers may wonder why a combination of data from different databases is at all necessary, since one could extract those heterogeneous actants from articles alone. But there are two main reasons for pursuing this strategy. First of all, text-mining article databases for these different kinds of entities runs into a number of technical problems (such as identifying the nature of those entities) that can at least in part be obviated by the combination approach. Second, and most importantly, each database corresponds to different regimes of engagement: the modality of engagement of the 'same' gene in a patent vs. an article or a grant proposal will vary in significant ways. The analytical strategy, then, amounts to diversifying the 'entry points': one can start with a set of human actors, as identified by fieldwork, or, alternatively, with a variety of bio-clinical entities that can be found in publications, but also in specialized databases devoted to genes and mutations, biomarkers and tests, or microarray experiments. Information can also be retrieved from websites, such as medical blogs or patient organization websites. Other (but expensive) opportunities to diversify entry points are offered by databases such as *RECAP* (http://www.recap.com/) that provide information about commercial deals in the biopharmaceutical domain. Multiple maps may destabilize conventional readings, generate a feeling of analytical strangeness, and record unexpected events, in a way similar to how new objects, accounts, and relations redefine and displace the boundaries of emerging domains.

We mentioned these examples as possible, uncertain avenues for further investigation, as they have so far not been exploited in the perspective we are advocating here (but see the next section for steps in this direction). This is partly due to the fact that laborious technical bridges need to be established between the different databases; these calculations and manipulations stand in contrast with the seamless association of heterogeneous entities that underlies translations and mediations between different regimes of engagement, as captured by (multi-site) fieldwork. In the biomedical translational research domain, a promising development is the establishment of the *CinicalTrials.gov* database by the NIH. The creation of this database is itself part of policy initiatives aiming at regulating the controversial domain of clinical research, marred by accusations of conflicts of interest, publication bias, etc. Unsurprisingly, the database itself has run into trouble, due to criticism about its incomplete coverage, failure to include relevant information, and lack of standardization, which in turn has led to additional policy initiatives (compulsory registration of trials if results are to be published, etc.) (Zarin *et al.* 2011). In spite of all these problems that complicate its appropriation for our own purposes, the database offers the advantage of assembling in a single virtual space entities such as clinical researchers, molecules (drugs), the institutions performing the trial, public organizations (oncology networks), commercial organizations (pharmaceutical and biotech companies), diseases, technologies, and publications. Bridges with other databases with a different take on those 'same' entities can then be built. Other databases,

such as *Orphanet* on rare diseases similarly offer opportunities for the kind of heterogeneous analysis we advocate.

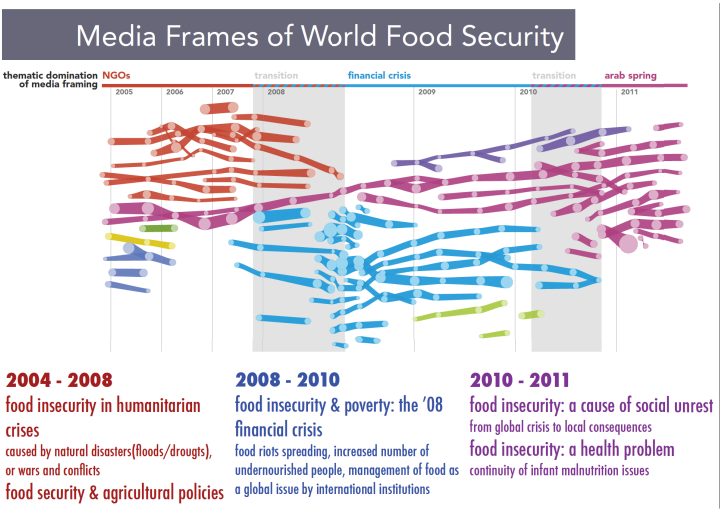

Fig. 4 – Security streams: see text for explanations. Source: Chavalarias *et al.* (2011).

Before closing this section, we would like to briefly introduce a recent attempt to tackle the issue of dynamics. The example below is taken from a report on food security based on the analysis of around 20,000 press articles published between 2004 and 2011 and listed in the database *Factiva* (Chavalarias *et al.* 2011). A somewhat similar approach, albeit with far more primitive tools, was introduced 20 years earlier by the developers of cop-word analysis (Callon *et al.* 1991), and applied to the biomedical domain shortly afterwards (Cambrosio *et al.* 1993). The authors of the 2011 article divided the corpus into 20 subsets, text-mined them, and produced for each of them a semantic network that included clusters of closely associated terms, each corresponding in principle to a topic. Instead of analyzing individual maps separately, they produced a single map with *streams* of clusters, according to the following principle: clusters from a given point in time are linked to previous or subsequent clusters through a stream if they have terms in common. As shown in Figure 4, a stream can split, merge, grow, emerge, decay etc. In spite of a common designation – food security – the domain in 2005 bears little resemblance to the domain in 2011, as new entities have emerged and redefined how

this issue is problematized. Stream analysis amounts to observing the digital traces left by evolving associations in a dynamic landscape, whereby innovation derives from the emergence of new "concerned" entities (Callon and Rabeharisoa 2008), rather than from relational shifts between a predefined list of entities describing a stable state of the world. Parallel instances of "overflows" (Callon 2002) can be associated with these dynamic streams, as indicated on Figure 4.

5. Clusters and Collectives

As previously suggested, the most common interpretation of network maps hinges, first, on the identification of clusters of closely connected entities, and, subsequently, on the analysis of the relations each of these subsets entertains with the others. The tracing of cluster boundaries used to be done manually, by visual inspection of the maps, but cluster detection algorithms, some of which include a fuzzy approach whereby a node can belong to more than one cluster, now increasingly perform this task. Insofar as these algorithms are based on purely structural calculations, they do not necessarily lead to sociologically meaningful units, if by the latter we refer to collective forms of organization and their associated practices, programs, and bodies of knowledge; in short, *agencements* characterized by coordinated (if not homogeneous) ways of problematizing issues. From this point of view, visual inspection, whereby one could deploy his or her sociological imagination, might at first appear as a better alternative, were it not for the following two counter-arguments. First, clustering algorithms are not inflexible tools: one can vary their parameters depending on whether one wants to emphasize, for instance, continuities or discontinuities between clusters, thus playing with variable boundaries. Far from dictating their will, clustering algorithms can thus be used as interactive tools for exploring the associations deployed on a map, the latter becoming an experimental device that can be used to explore alternative configurations in connection to working hypotheses and fieldwork observations. Second, it is far from obvious that the analyst's presuppositions about the proper constitution of the world (which can moreover vary from observer to observer) should have priority over the surprises generated by unexpected network configurations, especially when elicited by interactive tools. Here too maps can function as devices for exploring the variable geometry of the world, rather than as final statements about its ontology. The relevant components of social ontology are in any event open to debate, as shown by the not always mutually consistent attempts to capture them through different notions, such as "communities of practice" (Wenger 1998), "epistemic communities" (Akrich 2010), "collaborative communities" (Adler *et al.* 2008), and the like.

For maps to play an optimal role in this respect we can resort to a trick similar to the one discussed in the previous section, namely to produce a number of maps displaying different categories of actants, i.e. human actors such as researchers and clinicians, infrastructural components such as journals, techniques and models, and notions or concepts. Adding or subtracting some of these components in different combinations could lead to more dense or fragmented situations, helping analysts to put forward hypotheses about the elements that lead to new associations or result in disjunctions. Could, for instance, the densification of a network following the introduction of conceptual components or, alternatively, of certain kinds of tools and techniques be used to differentiate between epistemic communities and communities of practice? While, for a variety of reasons, this seems unlikely, we mention this possibility as a thought experiment to illustrate the kind of analytical approaches we would like to deploy. Actual examples of these approaches do not fully correspond to an ideal translation into practice of this research agenda, but are still worth examining.

Navon and Shwed (2012) analyzed 1400 articles to investigate how a genetic mutation (a microdeletion) transformed biomedical understandings of several rare clinical syndromes, unifying a set of previously independent clinical entities on the basis of molecular analysis. The microdeletion, in other words, was a key actant in "foster[ing] enduring ties between several small, previously disjunct fields of medical research, creating a densely connected literature that brought together an otherwise incoherent set of patients, expertise and clinical observations" (Navon and Shwed 2012, 1640). Their demonstration relies on generating networks derived from citation links between three decades of papers, identifying research communities interested in the older conditions with the help of a modularity algorithm, and showing how the microdeletion progressively unified them, turning a previously invisible collection of conditions into a visible field of coordinated knowledge production. They tell this story by using a set of four maps corresponding to four distinct periods during the last 30 years of the 20th century, and a set of two maps depicting the situation at the beginning of the 21st century.[8] They describe their approach as a way to overcome the limitations of our existing social scientific toolkit that is unable to grapple with non-human entities, such as genetic mutations, that are presently reconfiguring the biomedical field.

Navon and Shwed's (2012) article, which, it should be added, relies on concurrent fieldwork, is a fine-grained investigation of a specific biomedical domain. The reconfiguration of biomedical work by new bioclinical entities can be observed at a higher level of aggregation, where one can examine how translational research has emerged as a distinctive

[8] Given the number of figures, we refer readers to the original publication instead of reprinting them.

site of biomedical activities that cannot be conflated with fundamental and clinical research. These categories are not a priori categories defined by the analyst: they can be derived by the self-organizational properties of maps[9], for instance by observing how journals organize themselves into distinct clusters by spinning a dense web of inter-citations; these clusters can subsequently be qualified by an algorithm that distributes them along a translational continuum on the basis of the terms that appear in the titles of the journals' articles. Cambrosio *et al.* (2006) have adopted this approach to examine a large set of oncology journals: while early periods were characterized by the presence of two major clusters that correspond to activities taking place in either the laboratory or in the clinic, at the turn of the century a third, in-between cluster has become apparent. To further investigate the nature of that cluster, the authors text-mined the titles and abstracts of the articles published by a number of journals in each of the three clusters. Figure 5 shows the resulting map, with the translational space associated with a number of recent bio-clinical entities, such as oncogenes and mutations. Here again, this interpretation was supported by extensive, concurrent fieldwork.

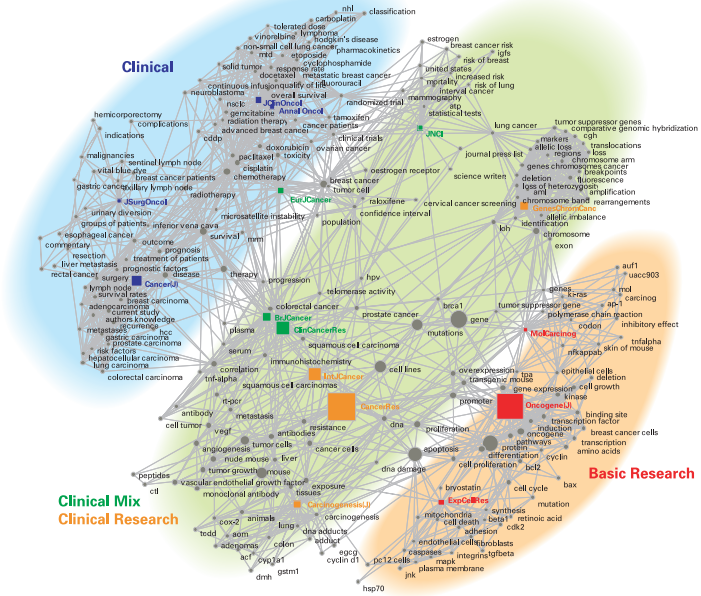

Fig. 5 – Heterogeneous network of journals and bio-clinical entities in the oncology domain: see text for explanations. Source: Cambrosio *et al.* (2006).

[9] Methods have been recently developed for generating maps of biomedicine based on self-organizing algorithms; see Skupin et al. (2013).

As noted at the beginning of this paper, recent biomedical work, in particular translational research, is characterized by a collective turn, which situates it firmly within the scope of inter-laboratory arrangements. This is one of the reasons why local ethnographic observations show their limitations, and the resort to cartography has been suggested as an important complement for investigating contemporary biomedicine, even if ethnography maintains its relevance, in particular for interpreting the maps. As shown by work on French cancer genetics (Bourret 2005) and psychiatric genetics (Rabeharisoa and Bourret 2009), this collective turn is not to be confused with a mere increase in the number of authors co-signing a paper. It is better captured by the notion of "new bio-clinical collective", rather than understood as a network, because it corresponds to a configuration centered on a specific activity, namely the simultaneous development of cancer genetics as a research field and as a domain of clinical intervention. One must start with this activity in order to define the collective. The human components of the collective include a variety of healthcare professionals, whose direct or indirect collaborations and interactions are a *sine qua non* for the development of this hybrid domain. The non-human components include a number of emerging bio-clinical entities, in particular different kinds of mutations, whose uncertain status needs to be managed, re-adjusted, and stabilized as part of the emergence of a "clinic of mutations" (Rabeharisoa and Bourret 2009). The focus of the collective lies precisely in the necessarily temporary qualification of these bio-clinical entities, which explains why the structure and nature of the collective modifies itself on an ongoing basis in relation to the emerging entities that need to be domesticated and mastered for the activity to continue. While the activities of the collective center on building more robust bio-clinical entities, they also involve producing knowledge about what should count as uncertain and unstable: the known unknown. Instead of a passage from local to extended networks, as typically described by early ANT analysts, we face here a situation characterized by the presence of an open-ended list of problematic entities. This is why in order to mobilize these entities they need to be often re-qualified and re-specified. As a result, the collective evolves by incorporating new actors, technologies, entities, and by opening up new fields of investigation.

As an attempt to capture at least a few elements of this dynamics, Bourret *et al.* (2006) collected a comprehensive set of publications by French cancer geneticists over more than three decades[10], and divided them into four periods as defined by major turning points in the history of the field. These data were then used to produce two kinds of maps. First, a set of more traditional co-authorship maps that displayed the progressive constitution of a social network, from an initially fragmented sit-

[10] The procedure involved combining references from *Medline* with those obtained from individual CVs.

uation with a number of local, regional sites, to a fully integrated situation defined by the presence of a single major component. For the second kind of maps the authors opted for an approach displaying the relations between researchers and the bio-clinical entities derived from text-mining titles and abstracts. And here something interesting became visible. Figures 5 and 6 show improved versions (obtained using more sophisticated text-mining software) of the maps used in the original publication. Figure 5 corresponds to the initial period (1970s and early 1980s) of French cancer genetics. As can be seen, the map is organized around a few key researchers: although relations between these researchers are mediated by non-human entities, the distribution of these entities espouses the polarity defined by human actors. Figure 6, corresponding to the turn of the century period, shows a reversal of this situation, as non-human entities, such as mutations, exons, chromosomes, and cell lines, appear to play a key role in organizing the map. The initial maps (not shown here) did not correspond to a given field or specialty, but to the early activities of researchers who subsequently converged on cancer genetics. In other words, the maps do not display structural positions in a scientific field or social world; rather, they follow the movement of researchers and bio-clinical entities leading to the establishment of a collective, even when individual researchers might not conceive of themselves as members of that collective.

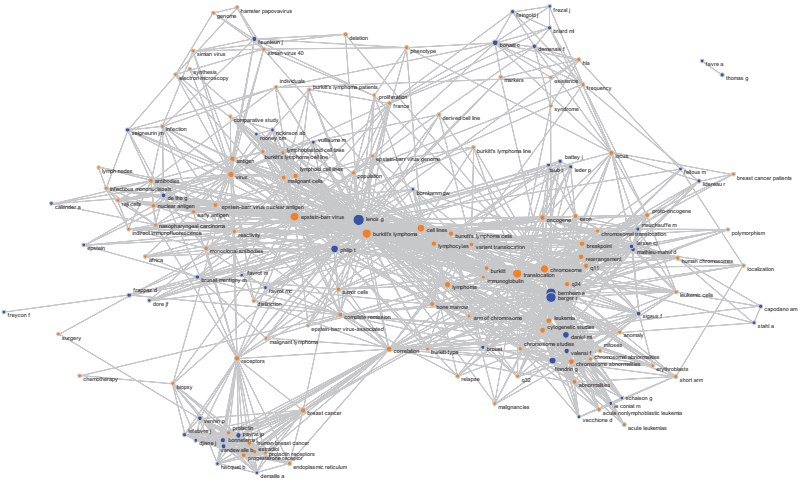

Fig. 6 – Heterogeneous network of early French cancer genetics. Humans: blue nodes; non-humans: orange nodes. Source: revised version of Figure 4 in Bourret *et al.* (2006).

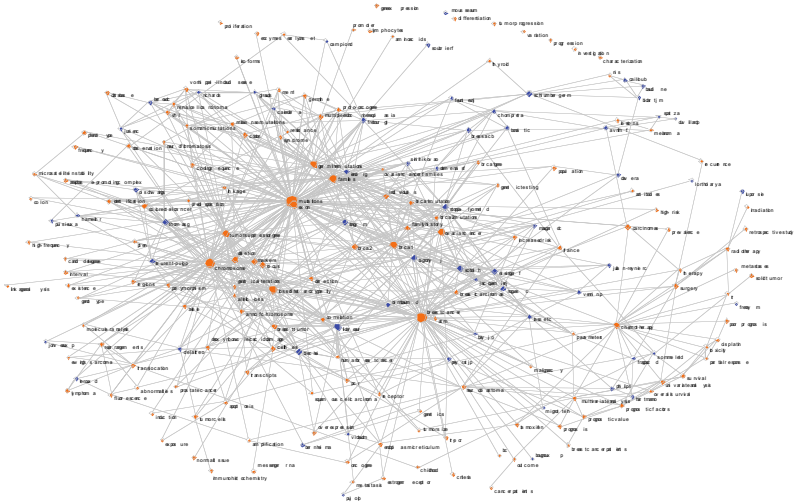

Fig. 7 – Heterogeneous network of turn of the century French cancer genetics. Humans: blue nodes; non-humans: orange nodes. Source: revised version of Figure 9 in Bourret *et al.* (2006).

It can thus be argued, almost paradoxically, that these maps allow one to (pragmatically) distinguish networks from collectives, as the emergence of the collective coincides with the activity of the emerging bio-clinical entities that led to the ongoing readjustment of its internal connections. A collective, thus, amounts not merely to a set of collaborative ties but to a configuration where collaborative work takes place and has been reorganized around these entities – in other words, what we have referred to as an *agencement*. The developmental trajectory of the collective cuts across the initial distinctions between different specialties (cytogenetics, hematology, oncology or medical pediatrics), and reaches a stage where it displays collective agency. The point is not to investigate how networks relate to collectives, but to use network analysis to produce something different from networks. To do so, we need to connect what we see on the maps – the organization of the collective around a number of entities – with what happens in the field, i.e. with the disparate, yet mutually constitutive activities of the collective, including the production, qualification, regulation and circulation of the new entities; in short, all that is needed for these objects to achieve a clinical existence. It is worth repeating that for this to happen actants do not need to be directly acquainted with each other, as long as they work on the same biomedical platforms

(Keating and Cambrosio 2003) that establish transitive relations between, for instance, mutations, diagnostic categories, drugs, and diseases. This also means that in order to capture this dynamics we need to go beyond texts, and take into account wet lab and clinical activities, the circulation of material entities (test kits, samples etc.), and, most recently, the algorithms and codes of bioinformatics.

6. Conclusion: Back to Reflexivity

We presently witness a proliferation of data and databases, often freely accessible on the Web, that can be easily searched and analyzed thanks to a mounting offer of dedicated software platforms, including network analysis software. S&TS scholars, even those with little understanding of quantitative approaches, can now easily perform (semi)quantitative analyses. This is a positive development, but it raises the issue of how S&TS analysts have come to accept these opportunities without asking too many questions about the sociotechnical scripts embedded in the databases they so happily use. Indeed, while S&TS scholars have had much to say about infrastructure, in particular information infrastructure (Star and Ruhleder 1996; Star 1999; Bowker 2006), they have so far not quite succeeded in reflexively incorporating these insights into their own work with (rather than on) information databases. Those of us who work on biomedicine consult almost daily the *Medline* database, and yet we rarely investigate how it has established – thanks to its peculiar structure, format, outreach and universal access – a network-like, worldwide space that multiplies inter-textual relations and favors strategies based on the accumulation of references, citations and co-authorship links. Critical observers have focused their attention on a database like *Web of Science*, holding it responsible for the rise of a whole industry of citation counts and evaluations through controversial tools such as impact factors. Fewer analysts have looked at how *Medline* and its search engine *PubMed*, which are regularly reformatted in response to new information-retrieval needs to whose emergence they contribute, have led to the constitution of a collaborative space by multiplying socio-semantic networks. The aforementioned debates and controversies surrounding the establishment of a database such as *ClinicalTrials.gov* provide a clear indication of how much is at stake in developing this kind of initiatives.

Databases are not restricted to bibliographic databases. Genomics, as noted at the beginning of this text, is generating its own avalanche of big data stored in a number of databases. In order to become "actionable" (Nelson *et al.* 2013) these data need to be interpreted (Leonelli 2014), and part of this interpretation process involves establishing connections between the information provided by the articles and the bio-clinical data stored in the genomics databases. Private companies have invested in this market niche. For instance, *Linguamatics* (http://www.linguamatics.com)

offers text-mining software and services to extract and combine information from the life sciences literature, electronic medical records, clinical pathology documents, clinical trial data and patents. Notice how text- and data-mining tools allow researchers to navigate a seamless web of heterogeneous documents, by the same token moving across the material basis of specialties and disciplines. Nor is this circulation limited to tools: it also involves conceptual transfers, as when information scientists borrow a molecular biology notion – DNA transcriptional bursting – to design and designate algorithms that track word and topic bursts in documents (e.g., Mane and Börner 2004).

As previously noted, in S&TS we are mostly on the receiving end of these processes, both with respect to the tools used to investigate existing databases, and to the establishment of the databases carrying the information that is then retrieved and processed by those tools. We need to investigate these processes in order to understand how these new ways of producing, storing, interpreting, and disseminating data are reframing biomedical activities and configurations. Work by Sabina Leonelli (2012, 2013, 2014) is particularly useful in this respect. But, as just noted, we also need to find ways to reflexively integrate these analyses into our own work with data- and text-mining tools, which in turn means repositioning ourselves both vis-à-vis network analysis approaches and the conceptual and analytical scripts and frames they embed. The point is not simply that we urgently require visualization tools that are better adapted to our theoretical and conceptual framings. A discussion of the shortcomings of existing tools should also lead us to re-examine some key aspects of our conceptual and methodological approaches, especially when they tend to mistake one-click network structures for more complex, rhizome-like arrangements, or to replace agency with structure.

The notion of network is, of course, central to this line of questioning, especially when the actors we investigate reason in terms of networks and extended circulation spaces. But this reflexivity loop, while offering new opportunities for collaboration between S&TS and biomedical researchers, could lead to serious difficulties if insufficiently problematized. As far as opportunities are concerned, we can think of jointly exploring the dynamics (and thus also the forms of agency) characterizing a given domain, or the nature of collectives involved in specific endeavors. Biomedical colleagues are in a good position to replace the few, selective connections displayed on a map with accounts that better correspond to what Strathern (1999) calls the "proliferation of the social", and at the same time our position vis-à-vis their activities is no longer one of externality, as a granting agency such as Genome Canada strongly supports the integration into biomedical projects of ancillary studies on Genomics and its Ethical, Economic, Environmental and Social aspects (GE^{3}LS). As for difficulties, the main one, as we hope to have shown, lies in the notion of network itself, which needs to be theoretically repositioned, because what is relevant in this new context are collective agencies and processes of

agencement, rather than bundles of relations. Such a requirement does not simply express a theoretical preference; it also derives from a close observation of the development of biomedicine in the last half-century (Rheinberger 2009).

As readers who have followed us so far will have realized, we only offer partial solutions, mostly based on tinkering. In fact, we suggested these temporary work-arounds more as a way of exemplifying our questions than as a solution to the conundrums mentioned in this paper. One way of weakening a too strong reliance on structural network interpretations is to multiply the networks, by including different kinds of entities and diversifying entry points. A more intriguing suggestion concerns illegible maps: very often, maps produced in the early stages of a research project do not seem to offer any interpretative handle, as nodes and ties either form a dense, shapeless network, or seem to be randomly distributed. A lot of algorithmic work is then deployed to make those maps legible, to uncover network patterns that were not there at the outset. But what if the lack of a network is indeed the relevant result, and what if instead of using algorithms to turn illegible into legible maps we were to develop tools to explore and account for that illegibility? The point is not to add mess to mess, as some would wish (Law 2004), but to explore the work needed to make maps readable as part of an experimental setting that includes other devices and forms of investigation, not necessarily only interviews, observations, or other traditional forms of fieldwork, but also membership, however temporary, in the collectives we investigate. Beyond what at first we mistook as illegibility we might discover the vanishing points of collective agency.

Acknowledgments

Research for this paper was made possible by grants from the Canadian Institutes for Health Research (MOP-93553) and the Fonds de recherche du Québec – Société et culture (SE-164195). The roots of this paper go back to a September 2007 presentation at the meeting celebrating the 40th anniversary of the "Centre de Sociologie de l'Innovation" (Mines-ParisTech). More recently, preliminary versions of this paper were presented at the "STS Italia Workshop" (Università di Padova, 19 April 2013) and at the "4ème Congrès du Réseau international francophone de la recherche qualitative" (Fribourg, Switzerland, 19-21 June 2013). Many thanks to our colleagues Peter Keating and Nicole Nelson for numerous and ongoing discussions of the issues raised in this paper.

References

Adler, P.S., Seok-Woo, K. and Heckscher C. (2008) *Professional Work: The Emergence Of Collaborative Community*, in "Organization Science", 19 (2),

pp. 359-376.

Akrich, M. (2010) *From Communities of Practice to Epistemic Communities: Health Mobilizations on the Internet*, in "Sociological Research Online" 15 (2).

André, F. (n.d.) *Precision medicine: from stratified therapies to personalized therapies* (PowerPoint presentation), from: http://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CC0QFjAA&url=http%3A%2F%2Fwww.aviesan.fr%2Fen%2Fcontent%2Fdownload%2F7012%2F59872%2Ffile%2Fparlement%2Beuropeen.ppt&ei=qU4BU_r1JseU2wXV_oDQBw&usg=AFQjCNFxsvBGtd8xbXjzlRfmy4YltRiBQ&bvm=bv.61535280,d.b2I, retrieved on February 16, 2014.

Boltanski, L. and Thévenot, L. (2006) *On Justification: Economies of Worth*, Princeton, Princeton University Press.

Börner, K. (2010) *Atlas of Science*, Cambridge (MA), MIT Press.

Bourret, P. (2005) *BRCA Patients and Clinical Collectives. New Configurations of Action in Cancer Genetics Practices*, in "Social Studies of Science", 35 (1), pp. 41-68.

Bourret, P., Mogoutov, A., Julian-Reynier, C. and Cambrosio A. (2006) *A New Clinical Collective for French Cancer Genetics: A Heterogeneous Mapping Analysis*, in "Science, Technology, & Human Values", 31 (4), pp. 431- 464.

Boyack, K.W., Mane, K. and Börner, K. (2004) *Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research*, in "Proceedings. Eight International Conference on Information Visualisation: IV2004", London (UK), pp. 965-971.

Bowker, G.C. (2006) *Memory Practices in the Sciences*, Cambridge, MA, MIT Press.

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

Callon, M. (2002) *From Science as an Economic Activity to Socioeconomics of Scientific Research: The Dynamics of Emergent and Consolidated Techno-Economic Networks*, in P. Mirowski and E.M. Sent (eds.), *Science Bought and Sold. Essays in the Economics of Science*, Chicago, University of Chicago Press, pp. 277-317.

Callon, M. (2006) *Can Methods for Analyzing Large Numbers Organize a Productive Dialogue with the Actors They Study?*, in "European Management Review", 3, pp. 7-16.

Callon, M. (2012) *Quel rôle pour les sciences sociales face à l'emprise grandissante du régime de l'innovation intensive?,* in "Cahiers de recherche sociologique",

53, pp. 121-165.

Callon, M. (2013) *Qu'est-ce qu'un agencement marchand?*, in M. Callon, M. Akrich, A. Dubuisson-Quellier, C. Grandclément, A. Hennion, B. Latour, A. Mallard, C. Méadel, F. Muniesa and V. Rabeharisoa (eds.), *Sociologie des agencements marchands: textes choisis*, Paris, Presses des Mines, pp. 325-440.

Callon, M., Law, J. and Rip, A. (eds.) (1986) *Mapping the Dynamics of Science and Technology*, Houndmills, Macmillan.

Callon, M., Courtial, J.P. and Laville, F. (1991) *Co-word Analysis as a Tool for Describing the Network of Interactions Between Basic and Technological Research: The Case of Polymer Chemistry*, in "Scientometrics", 22 (1), pp. 155-205.

Callon, M. and Rabeharisoa, V. (2008) *The Growing Engagement of Emergent Concerned Groups in Political and Economic Life: Lessons from the French Association of Neuromuscular Disease Patients*, in "Science, Technology & Human Values", 33 (2), pp. 230-261.

Cambrosio, A., Limoges, C., Courtial, J.P. and Laville, F. (1993) *Historical Scientometrics? Mapping Over 70 Years of Biological Safety Research with Co-Word Analysis*, in "Scientometrics", 27 (2), pp. 119-143.

Cambrosio, A. and Keating, P. (2000) *Of Lymphocytes and Pixels: The Techno-Visual Production of Cell Populations*, in "Studies in History and Philosophy of Biological and Biomedical Sciences", 31 (2), pp. 233-270.

Cambrosio, A., Keating, P. and Mogoutov, A. (2004) *Mapping Collaborative Work and Innovation in Biomedicine: A Computer-Assisted Analysis of Antibody Reagent Workshops*, in "Social Studies of Science", 34 (3), pp. 325-364.

Cambrosio, A., Keating, P., Mercier, S., Lewison, G. and Mogoutov, A. (2006) *Mapping the emergence and development of translational cancer research*, in "European Journal of Cancer", 42 (18), pp. 3140-3148.

Chavalarias, D., Cointet, J.P.L., Duong, T.K., Mogoutov, A., Villard, L., Savy, T. and Roth, C. (2011), *Streams of Media Issues, Monitoring World Food Security*, Technical report, Global Pulse – United Nations.

Cointet, J.P., Mogoutov, A., Bourret, P., El-Abed, R. and Cambrosio, A. (2012) *Les réseaux de l'expression génique : émergence et développement d'un domaine clé de la génomique*, in "Médecine/Sciences", 28, pp. 7-13.

Edwards, P.N. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, Cambridge (MA), MIT Press.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M. and Park, M. (2008) *Stromal Gene Expression Predicts Clinical Outcome in Breast Cancer*, in "Nature Medicine", 14 (5), pp. 518-527.

Goldberg, P. (2011) *Prepare for "Tsunami" of Genomic Information, Sledge Urges*

In ASCO Presidential Address, in "The Cancer Letter", 37 (23), pp. 1-7.

Keating, P. and Cambrosio, A. (2003) *Biomedical Platforms. Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine*, Cambridge, MA, MIT Press.

Keating, P. and Cambrosio, A. (2012a), *Cancer on Trial. Oncology as a New Style of Practice*, Chicago, The University of Chicago Press.

Keating, P. and Cambrosio, A. (2012b) *Too Many Numbers: Microarrays in Clinical Cancer Research*, in "Studies in History and Philosophy of Biological and Biomedical Sciences", 43 (1), pp. 37-51.

Kohli-Laven, N., Bourret, P., Keating, P. and Cambrosio, A. (2011) *Cancer Clinical Trials in the Era of Genomic Signatures: Biomedical Innovation, Clinical Utility, And Regulatory-Scientific Hybrids*, in "Social Studies of Science", 41, pp. 487-513.

Latour, B. (2011) *Networks, Societies, Spheres: Reflections of an Actor-Network Theorist*, in "International Journal of Communication", 5, pp. 796–810.

Latour, B., Jensen, P.,Venturini, T., Grauwin, S. and Boullier, D. (2012) "*The Whole Is Always Smaller Than Its Parts": A Digital Test Of Gabriel Tarde's Monads*, in "British Journal of Sociology", 63 (4), pp. 590–615.

Law, J. (2004) *After Method: Mess in Social Science Research*, London, Routledge.

Leonelli S. (2012) *When Humans Are the Exception: Cross-Species Databases at the Interface of Biological and Clinical Research*, in "Social Studies of Science", 42 (2), pp. 214-236.

Leonelli S. (2013) *Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research*, in "BioSocieties", 8, pp. 449-465.

Leonelli S. (2014) *Data Interpretation in the Digital Age*, in "Perspectives on Science", 22, pp. 397-417.

Lima, M. (2011) *Visual Complexity: Mapping Patterns of Information*, New York, Princeton Architectural Press.

Mane, K.K. and Börner, K. (2004) *Mapping Topics and Topic Bursts in PNAS*, in "Proceedings of the National Academy of Sciences", 101 (Suppl. 1), pp. 5287–5290.

McMeekin, A. and Harvey, M. (2002) *The Formation of Bioinformatics Knowledge Markets: An 'Economies Of Knowledge' Approach*, in "Revue d'Économie Industrielle", 101 (1), pp. 47-64.

McMeekin, A., Harvey, M. and Gee, S. (2004) *Emergent Bioinformatics and Newly Distributed Innovation Processes*, in M. McKelvey, J. Laage-Hellman and A. Rickne (eds.), *The Economic Dynamics of Modern Biotechnology*, Oxford, Ox-

ford University Press, pp. 236-261.

Mogoutov, A., Cambrosio, A., Keating, P. and Mustar P. (2008) *Biomedical Innovation at the Laboratory, Clinical and Commercial Interface: A New Method for Mapping Research Projects, Publications And Patents in the Field of Microarrays*, in "Journal of Informetrics", 2, pp. 341-353.

Moreira, T. (2012) *The Transformation of Contemporary Health Care. The Market, the Laboratory, and the Forum*, New York, Routledge.

Navon, D. and Shwed, U. (2012) *The Chromosome 22q11.2 Deletion: From the Unification of Biomedical Fields to a New Kind of Genetic Condition*, in "Social Science & Medicine", 75 (9), pp. 1633-1641.

Nelson, N., Keating, P. and Cambrosio, A. (2013) *On Being 'Actionable': Clinical Sequencing and the Emerging Contours of a Regime of Genomic Medicine in Oncology*, in "New Genetics & Society", 32 (4), pp. 405-428.

Newman, M.E.J. (2004) *Coauthorship Networks and Patterns of Scientific Collaboration*, in "Proceedings of the National Academy of Sciences", 101 (Suppl. 1), pp. 5200–5205.

Rabeharisoa, V. and Bourret, P. (2009) *Staging and Weighting Evidence in Biomedicine: Comparing Clinical Practices in Cancer Genetics and Psychiatric Genetics*, in "Social Studies of Science", 39 (5), pp. 691-715

Rabeharisoa, V., Callon, M., Felipe, A.M., Nunes, J.A., Patterson, F. and Vergnaud, F. (2014) *From 'Politics Of Numbers' to 'Politics Of Singularisation'. Patients's Activism and Engagement in Research on Rare Diseases in France and Portugal*, in "Biosocieties", 9 (2), pp. 194-217.

Rheinberger, H.J. (1997) *Towards a History of Epistemic Things: Synthesizing Proteins in the Test Tube*, Stanford, CA, Stanford University Press.

Rheinberger, H.J. (2009) *Recent Science and Its Exploration: The Case of Molecular Biology*, in "Studies in History and Philosophy of Biological and Biomedical Sciences", 40 (1), pp. 6-12.

Scott J. (2000) *Social Network Analysis. A Handbook*, London, Sage.

Skupin, A., Biberstine, J.R. and Börner, K. (2013) *Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach*, in "PLoS One", 8 (3), e58779.

Star, S.L. (1999) *The Ethnography of Infrastructure*, in "American Behavioral Scientist", 43 (3), pp. 377-391

Star, S.L. and Ruhleder, K. (1996) *Steps toward an Ecology of Infrastructure: Design and Access for Large Information Space*, in "Information Systems Research", 7 (1), pp. 111-134.

Strathern, M. (1999) *What Is Intellectual Property After?*, in J. Law and J. Hassard

(eds.), *Actor Network Theory and After*, Oxford, Blackwell, pp. 156-180.

Szatmari, P., Paterson, A.D., Zwaigenbaum, L., Roberts, W., Brian, J., et al. (2007) *Mapping autism risk loci using genetic linkage and chromosomal rearrangements*, in "Nature Genetics", 39 (3), 319–328.

Velden, T. and Lagoze C. (2013) *The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication*, in "Journal of the American Society for Information Science and Technology", 64 (12), pp. 2405-2427.

Watts, D.J. (2004) *The "New" Science of Networks*, in "Annual Review of Sociology", 30, pp. 243–270.

Wenger, E. (1998) *Communities of Practice: Learning, Meaning, and Identity*, Cambridge, MA, Harvard University Press.

Wilkinson, L. and Friendly, M. (2009), *The History of the Cluster Heat Map*, in "The American Statistician", 63 (2), pp. 179–184.

Yaffe, M.B. (2013) *The Scientific Drunk and the Lamppost: Massive Sequencing Efforts in Cancer Discovery and Treatment*, in "Science Signaling", 6 (269), pe13.

Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M. and Ide, N.C. (2011), *The ClinicalTrials.Gov Results Database: Update and Key Issues*, in "The New England Journal of Medicine", 364 (9), pp. 852–860.