



Mapping scientific institutions

Sebastian Grauwin, Pablo Jensen

► To cite this version:

Sebastian Grauwin, Pablo Jensen. Mapping scientific institutions. *Scientometrics*, Springer Verlag, 2011, 89 (3), pp.943-954. <10.1007/s11192-011-0482-y>. <hal-00650267>

HAL Id: hal-00650267

<https://hal.archives-ouvertes.fr/hal-00650267>

Submitted on 9 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mapping scientific institutions

Sebastian Grauwin ^{1,2,3} Pablo Jensen ^{1,2,3,4}

June 21, 2011

Abstract

We have developed a set of routines that allows to draw, in a few hours, different maps of the research carried out in a scientific institution. Our toolkit uses OpenSource elements to analyze bibliometric data gathered from the Web Of Science. We take the example of our institution, *ENS de Lyon*, to show how different maps, using co-occurrence (of authors, keywords, institutions...) and bibliographic coupling can be built. These maps may become a valuable tool for institutions' directors, as they offer different views on the institution at a global scale.

1 Introduction

Scientometrics has proved a valuable tool to understand the organization of scientific fields [1] and their evolution [2, 3]. Global science maps [1, 4, 5, 6, 7, 8, 9] have become **feasible** recently, offering a tentative overall view of scientific fields and fostering dreams of a “science of science” [6]. In this article, we propose a more modest but less explored mapping, that of single scientific institutions. The scope is to achieve a global point of view on the institutions that no individual can have, in order to understand their organization, their strong and weak points, the papers or authors that link different Departments or disciplines... Such maps may become important **as policy tools as few directors have such a global view of their institution**.

Recently, Rafols and Leydesdorff have suggested a simple way to picture the disciplinary weight of an institution [8]. This method is rapid and can be carried out online. As it uses Web of Science [10] “subject categories” as relevant sub-disciplines to project the data, it has the advantage of enabling a comparison across different institutions or years. The drawback of this rigid projection skeleton is that it preselects, without local information, the relevant communities. As acknowledged by Rafols and Leydesdorff [8]: “The two characteristics that

¹Université de Lyon

²Institut des Systèmes Complexes Rhône-Alpes (IXXI)

³Laboratoire de Physique, École Normale Supérieure de Lyon and UMR CNRS 5672, 69007 Lyon, FRANCE

⁴Laboratoire d’Économie des Transports, Université Lyon 2 and UMR CNRS 5593, 69007 Lyon, FRANCE

make overlay maps so useful for comparison, their fixed positional and cognitive categories, are also inevitably, their major limitations and a possible source of misreadings. Since the position in the map is only given by the attribution in the disciplinary classification, it does not say anything about the direct linkages between the nodes.”

Here we propose different ways of mapping scientific institutions based on the articles published with that address (and not the journals as in [8]). We do not propose a real methodological innovation, but rather a toolbox that allows to draw, in a few hours, several maps of the chosen scientific institution. More specifically, we show four different ways of mapping our institution, ENS de Lyon, and show how each of these gives different information. Our scope is to display - in an accessible (but not too simplistic) way - the institution’s complexity thus helping to generate discussions on its policy among its scientists.

2 Methodology

2.1 Data Extraction

The “Ecole normale supérieure de Lyon” (ENS de Lyon), focused on Natural sciences, was created in Lyon in 1987 after a move from Saint-Cloud in the suburbs of Paris. In 2010, it merged with the “Social and Human Sciences” Ecole Normale Supérieure. Today, it gathers 350 researchers, 270 professors, 390 administrative and technical personnel and a budget of more than 110 million Euros. A simple query (performed in January 2011) in the ISI Web of Knowledge database [10] yields 7584 papers containing an *ENS de Lyon* address (mostly under the form “Ecole Normal Super Lyon”, but also “ENS-LYON” and “ENS de Lyon”). We save the “Full records” of all these articles, the records containing authors, journal, year of publication, title, keywords (given by the authors and/or ISI Web of Science), subjects, addresses (institutions, cities and countries), and the list of references of the articles. It is well-known that Social and Human sciences (especially French ones) are not well represented in Web of Science. Therefore, our maps mainly deal with the natural sciences at ENS de Lyon.

Records are parsed and gathered in MySQL tables, which renders the handling of the data more straightforward. Simple frequency analysis of the records allows to get a first global representation of the institution. Our method uses the relations present in the data [11] to display different perspectives on the inner structure of an institution.

2.2 Bibliographic coupling

Links between articles are calculated through their common references. The bibliographic coupling similarity between two articles i and j is defined as [12]:

$$\omega_{ij} = \frac{|\mathcal{R}_i \cap \mathcal{R}_j|}{\sqrt{|\mathcal{R}_i||\mathcal{R}_j|}} \quad (1)$$

where \mathcal{R}_i is the set of references of article i .

In comparison to co-citation link (which is the more usually used measure of articles similarity), bibliographic coupling (BC) offers two advantages: it allows to map recent papers (which have not yet been cited) and since it deals with all published papers (whether cited or not), it represents “normal science”. The reason why weighted links are used is that they reinforce the dense (in terms of links per article) regions of the BC networks. This reinforcement facilitates the partition of the network into meaningful groups of cohesive articles, or communities. A widely used criterion to measure the quality of a partition is the modularity function [13, 14], which is roughly is the number of edges inside communities (as opposed to crossing between communities), minus the expected number of such edges if the network were randomly produced. We compute the graph partition using the efficient heuristic algorithm presented in [15].

Applying the Louvain algorithm yields a partition of the network into communities (see Figure 2). Simple frequency analysis then allows to characterise each community through its more frequent items (keywords, authors, etc...). The significativity σ of the presence of a given item into a community is computed by comparing its frequency f in the community to its frequency f_0 within the whole database. More precisely, we use the normalized deviation

$$\sigma = \sqrt{N} \frac{f - f_0}{\sqrt{f_0(1 - f_0)}} \quad (2)$$

where N is the total number of article in the database. The links between two communities I and J can also be characterized qualitatively by analyzing their shared references and quantitatively by computing the mean weight $\omega_{IJ} = <\omega_{ij}>_{i \in I, j \in J}$.

The final step in order to create a representation of the BC communities network is to choose a visualization algorithm. We use the Gephi software [16]. Gephi is a intuitive and interactive software allowing, in which force-directed layout algorithms are implemented. These algorithms produce a graph by simulating the dynamics of the network as if it were a physical system (the nodes being charged particules and the edges springs). The simulation is run until the system comes to an equilibrium state.

2.3 Copublication coupling

The data can also be analyzed through more common approaches, such as coauthoring or co-keyword analysis [11]. For this, a list of all items (authors, keywords, addresses) are taken from the records to obtain the nodes of our maps, whose size are proportionnal to the number of articles in which they appear. Two nodes (items) i and j are linked whenever the number n_{ij} of articles in

Table 1: Most frequent *ENS de Lyon*'s Subjects.

Subject	Prop of articles (%)
Biochemistry & Molecular Biology	8.25
Physics, Multidisciplinary	7.85
Computer Science, Theory & Methods	7.5
Mathematics	7.49
Geochemistry & Geophysics	7.34
Chemistry, Physical	6.41
Physics, Mathematical	6.18
Astronomy & Astrophysics	5.47
Mathematics, Applied	4.73
Cell Biology	4.43
Chemistry, Multidisciplinary	4.34
Physics, Condensed Matter	4.17
Physics, Atomic, Molecular & Chemical	3.9
Physics, Fluids & Plasmas	3.57
Genetics & Heredity	3.23

which they both appear is non-zero. More specifically, we use weighted links, where the co-occurrence normalized weight is chosen as

$$w_{ij} = \frac{n_{ij}}{\sqrt{n_i n_j}} \quad (3)$$

The visualization step of the produced maps is once again achieved through to gephi and its force-based layout algorithms.

2.4 Software available

We have developed a “Biblio Toolbox” which allows to draw the different maps presented here in a few hours. The toolbox needs access to Web of Science database but otherwise relies on OpenSource software. It is available at our website (<http://www.sebastian-grauwin.com/>).

3 Gaining perspective on the *ENS de Lyon*

3.1 Statistical analysis

ENS de Lyon gathers a broad spectrum of scientific subjects (Table 1), mostly in the natural sciences as discussed above. The institution has significantly grown over the last 20 years, as shown by its increasing production of papers (Fig 1). Our data gathers 12398 distinct authors, among which 952 have authored more than 5 papers. By construction of the database, at least one author of each article is a member of ENS de Lyon but this number also takes into account

all the authors of the papers among whom some may not be members of the ENS. ENS de Lyon collaborates with a broad range of institutions of different countries as shown below.

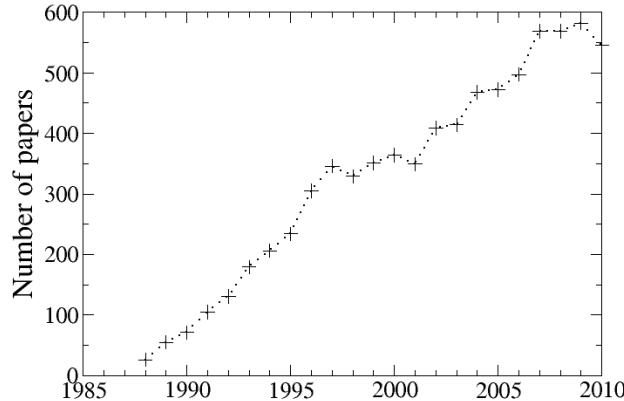


Figure 1: Number of paper with an *ENS de Lyon* address published by year, according to WoS, January 2011.

3.2 Bibliographic Coupling communities map

Figure 2 shows the map obtained with bibliographic coupling of articles and their grouping into "natural" subfields through modularity maximization. Each community is characterized by its more frequent author and keyword. Table 2 displays an "ID card" for the community labelled *Hansen JP/MOLEC-DYNAMICS*. This community gathers physicists interested in the understanding of condensed matter using molecular dynamics simulations. The "ID Cards" of the other communities are available online on <http://www.sebastian-grauwin.com/>.

What do we learn from this first map? First, note that the spatial organization of the communities fits well with the scientific organization of ENS de Lyon in different departments (different colors in Figure 2). This confirms that bibliographic coupling can recover the scientific organization of institutions. Interestingly, the precise community structure does not match the inner administrative/scientific subdivision of departments. For example, the physics lab is administratively divided into four groups, while our map distinguishes seven teams. This raises interesting questions on the structuration of the groups and their interactions. Two physics' communities (Oswald P/LiquidCrystals and Peyrard M/DNA) belong to the "soft-matter and biological systems" group but

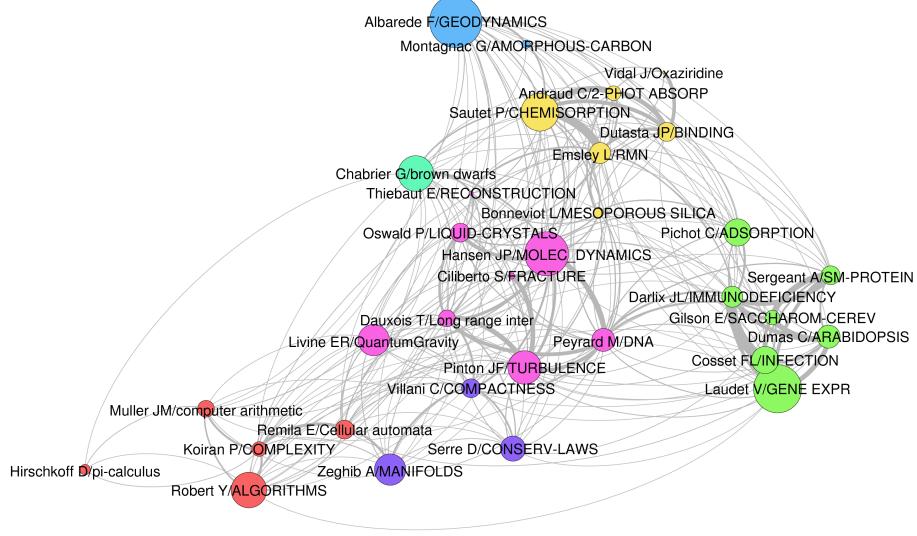


Figure 2: “Bibliographic Coupling” community structure of the *ENS de Lyon*. The surface of a community I is proportional to its number of articles N_I and the width of the link between two communities I and J is proportional to the mean bibliographic coupling. For the sake of clarity, communities with less than 10 articles are not displayed. Labels are obtained thanks to a frequency analysis of the authors and keywords. Each color corresponds to one of the ENS de Lyon scientific departments : biology (green), chemistry (yellow), physics (pink), computer science (red), mathematics (violet), earth sciences (blue) and astrophysics (turquoise). The belonging of a community to a department is determined through the proportion of community’s articles that use the department (for an example, Table 2 shows that more than 50% of “HansenJP/Molec-Dynamics” articles’ display the Physics Lab in the address).

our map shows that they are quite distant, which means that they do not share many references. The difference between the map and the physics lab organization is one example of the discussions that our work can generate.

Another example is given by the overall spatial structure. Our map clearly places physics at the scientific center of the ENS de Lyon, a fact that was used by its director to suggest the importance of his lab within the institution. The question is then : how much does this central position depend on the precise visualization algorithm used? Is it robust enough to allow for an interpretation and possibly orient governance? The forthcoming maps will comment on this issue, but let us already note that the central position of the Physics Lab *within*

this representation is quite robust. The reason is quite simple : the Physics Lab is the only one to have strong links to the other labs. Indeed, different physics' communities are linked to all other labs (for example Mathematics and Computer science (through Livine E/Quantum gravity), to Biology (through Peyrard M/DNA)...). The other labs are strongly linked only to one or two other labs (for example, Biology is only linked to Chemistry, through Pichot C/Adsorption, in addition to its link with the Physics lab), which explains their more peripheral position in the map. Therefore, the central position of the Physics lab can tentatively be interpreted as its central position in terms of modelling tools (molecular simulations tools shared with chemists for example), experimental tools (on "frictional mechanics" with the geophysics lab for example) or theoretical concepts (spin glass theory also studied by mathematicians). All these shared tools generate common references which lead to the links that structure our map.

3.3 International collaborations

It is straightforward to use the communities of the preceding map to include the international collaborations of the different teams (Fig. 3). We simply define links as given by the frequency of appearance of a foreign country in the community's articles addresses. For example, the strongest link is obtained for the Astrophysics papers, for which 41% of the papers are written in collaboration with a USA institution. The map shows that some groups rely heavily on many international collaborations (Emsley L/RMN has strong links with England, Italy and USA), while others are strongly linked to a single country (Dauxois T/Long range inter, to Italy) and others have mainly French collaborations (Oswald P/Liquid crystals).

3.4 Co-keywords, co-authors and heterogeneous maps

We now turn to more traditional maps, obtained by co-occurrence of keywords or authors in articles. Figure 4 shows the co-keywords map obtained by using Web of Science and authors' keywords. One should be cautious since some terms are clearly polysemic ("evolution", "particles"...) and create links between subdisciplines which are not very relevant. However, it is clear that physics is no longer at the center of the map. Instead, "crystal-structure" links chemistry (top left) with biology (right), "growth" links biology to physics and "transition" and "dynamics" link chemistry to physics (left). Another significant difference : what appeared to be a coherent whole when investigated through bibliographic coupling (the "Albarede F/Geodynamics" community) turns out to split into geochemistry (bottom of Figure 4) and geophysics (just up of the latter, close to physics, with keywords as "high-temperatures" or "high-pressure").

Figure 5 displays a co-author map. This represents an accessible way of showing data to the institutions' scientists, since names are usually well-known by the community. It also represents a good way to tap into directors' previous knowledge of the institution. However, coauthorship indicates a different

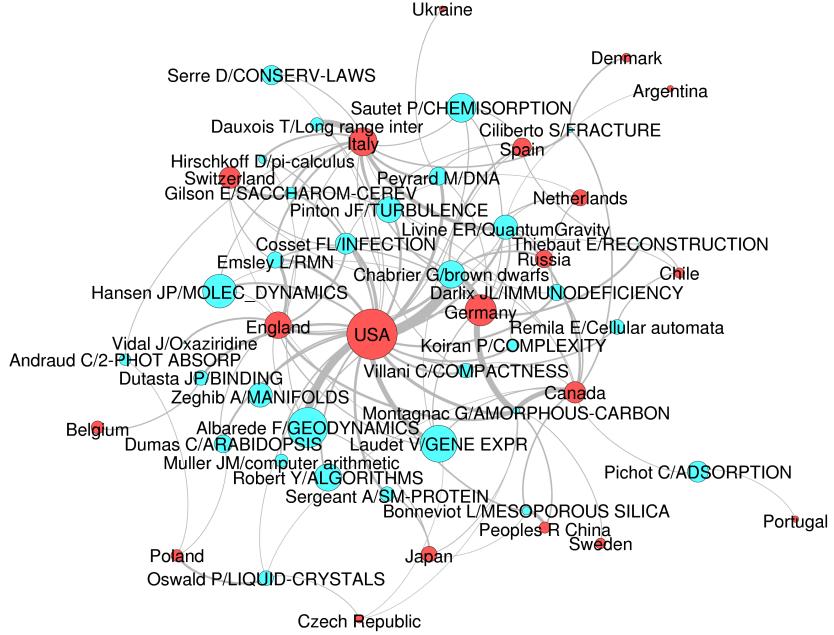


Figure 3: **International collaborations of the communities.** The size of the nodes correspond to the number of articles in each community which imply a collaboration with a foreign country. We only keep countries appearing in more than 10 articles and links corresponding to more than 3% of the articles implying a collaboration with the linked country. The width of the links is proportionnal to the proportion of linked articles.

(and stronger) link from the link established by sharing references (as in bibliographic coupling). This is visible in Figure 5 which does not show many links across disciplines (and some of the links are actually homonyms, such as Bertin E). The main co-publication link arises from collaborations between a biophysics lab and computer simulations of biological molecules (Peyrard/Bouvet/Gilson).

To improve over the limitations of both co-keyword and co-author analysis and gather most of the available information in a single map, it is possible to include all the co-occurrences between keywords, authors and institutions. Fig. 6 shows the map obtained for the ENS de Lyon. It displays the connecting rôle of a physics-biology interdisciplinary lab (Lab Joliot Curie, center right). One can also see that, while the CNRS plays an important an central rôle, other institutions collaborate on more specialized subfields (for example Univ California, Berkeley, lower left).

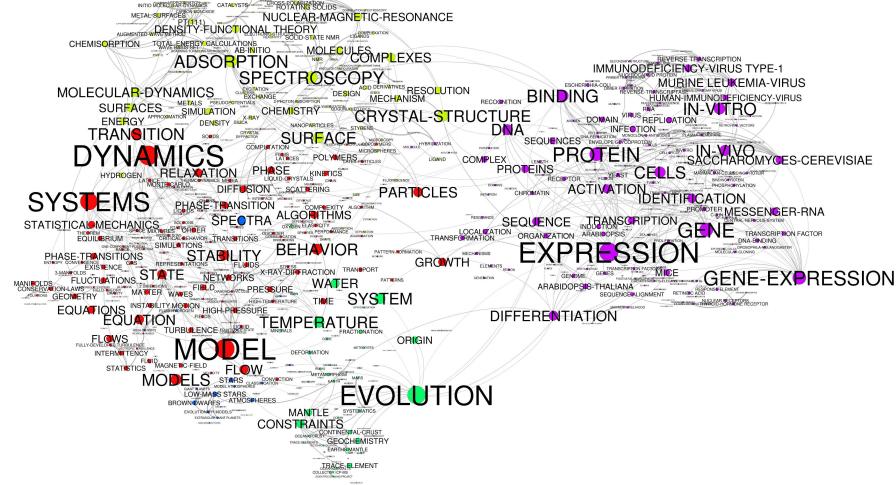


Figure 4: **Co-Keywords Network.** The size of the nodes is proportional to the number of times a keyword is used in our database. The width of the links indicates the cooccurrence weight between two keywords in the same article. We keep only keywords used in more than 10 publications. Colors correspond to a community analysis performed by gephi based on the same Louvain algorithm used for the Bibliographic Coupling analysis.

4 Discussion, Conclusions

Our aim in this paper was to present a toolbox to make institution mapping easy and rapid and to show on the example of our institution, ENS de Lyon, what kind of insights can be derived from these maps. It should be clear by now that there is not a unique (or a “best”) map of a scientific institutions, but rather many possible representations, each map containing a projection from a specific perspective [7, 17, 18].

We are now experimenting with ENS scientists’ and direction. ENS heads are enthusiastic about this global vision and five posters representing this papers’ figures are now displayed in the building. We hope that scientists at ENS de Lyon will test these maps against their own knowledge of the institution, will argue with us when what we picture does not fit and join the public discussion by offering alternative interpretations. As Nietzsche said [19]: “the more affects we allow to speak about one thing, the more eyes, different eyes, we can use to observe one thing, the more complete will our ‘concept’ of this thing, our

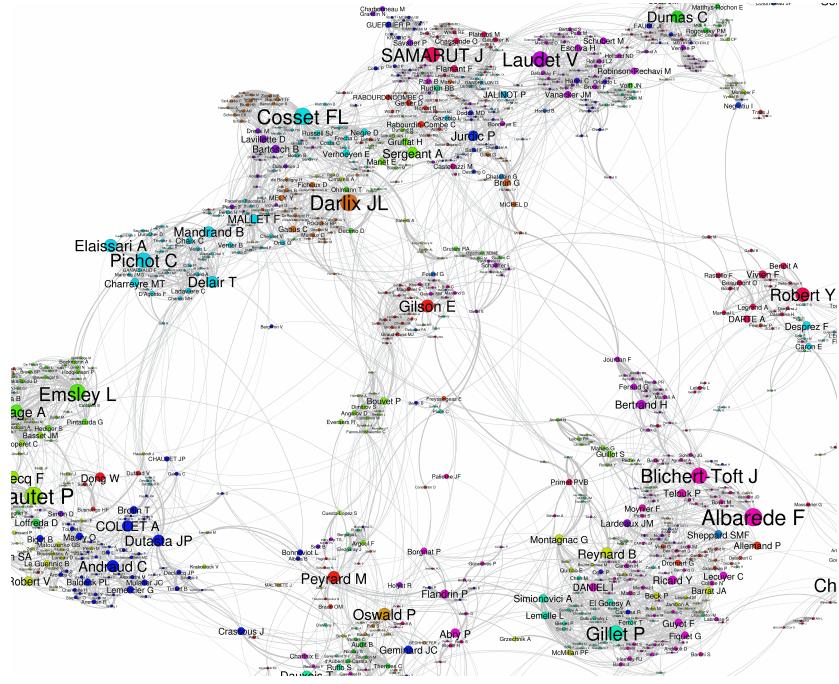


Figure 5: **Co-Authors Network (detail)**. The size of the nodes is proportional to the number of articles of our database authored by the author. The width of the links indicates the cooccurrence weight between two co-authors. We keep only authors used with more than 5 publications. Colors correspond to a community analysis performed by gephi based on the same Louvain algorithm we used for the bibliographic coupling analysis.

‘objectivity’ be”. The point is that although everybody acknowledges that maps are only representations and not the real thing, maps affect how we think about the institution [20].

We hope that our toolbox will lead other scientists to build maps of their own institutions, thus fostering ongoing dialogue and praxis in the institution. Future work includes preparing different maps for successive time periods, in order to grasp the evolution of the institution, and collaboration with other institutions (such as CNRS and CEMAGREF) which are interested in such global maps.

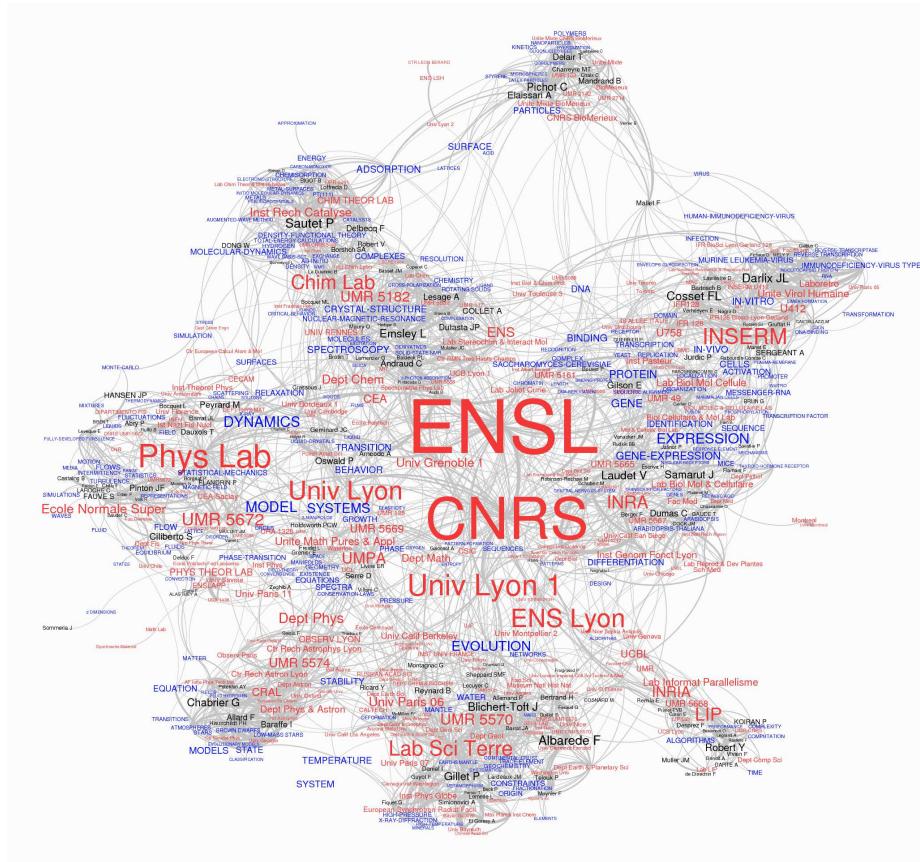


Figure 6: **Heterogeneous Network, mixing authors, keywords and institutions.** The size of the labels is proportional to the number of articles in which an item appear (we keep only items used in more than 20 publications). The width of the links indicates the cooccurrence weight between two items (we kept only links with a co-occurrence weight $\omega > 0.1$). Colors correspond to the type of the items (authors in black, keywords in blue and institutions in red).

References

- [1] Small H, (1999) Visualizing Science by Citation Mapping, *J Am Soc Inform Sci*, 50(9):799-813.
- [2] Chavaliarias D, Cointet J-P (2009) The Reconstruction of Science Phylogeny, arXiv:0904.3154v3
- [3] Cambrosio A, Peter Keating, Simon Mercier, Grant Lewison, Andrei

Mogoutov (2006), Mapping the emergence and development of translational cancer research, European Journal of Cancer 42, 3140–3148

- [4] Klavans R, Boyack KW (2009) Toward a consensus map of science, *J Am Soc Inform Sci Tech*, 60(3), 455-476.
- [5] Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents, *J Am Soc Inform Sci*, 24, 265-269.
- [6] Börner K, Scharnhorst A (2009), Visual conceptualizations and models of science. *Journal of Informetrics* 3:161-172.
- [7] Leydesdorff L, Rafols I (2009) A Global Map of Science Based on the ISI Subject Categories, *J Am Soc Inform Sci Tech*, 60(2).
- [8] Rafols I, Leydesdorff L (2010) Science Overlay Maps: A NewTool for Research Policy and Library Management, *J Am Soc Inform Sci Tech*, 61(9):1871-1887.
- [9] Agarwal P, Skupin A (2008) Self-organising maps: applications in geographic information science, Ed. John Wiley & Sons Inc
- [10] <http://apps.isiknowledge.com/>
- [11] Börner K, Chen Ch and Boyack K (2003), Visualizing Knowledge Domains, *Annual Review of Information Science and Technology*, Volume 37, 179–255
- [12] Kessler MM (1963) Bibliographic coupling between scientific papers. *American Documentation* 24:123-131.
- [13] Girvan M, Newman MEJ (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113.
- [14] Fortunato S Barthélémy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci USA* 104:36-41.
- [15] Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech P10008*.
- [16] Bastian M, Heymann S, Jacomy M (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proc 3rd Intl ICWSM Conf*
- [17] Roessner D (2000) Quantitative and qualitative methods and measures in the evaluation of research, *Research Evaluation* 9(2):125-132.
- [18] Stirling A (2008) “Opening up” And “Closing down”: Power, participation, and pluralism in the social appraisal of technology. *Science, Technology & Human Values* 33(2):262-294.
- [19] Nietzsche F (1969), On the genealogy of morals, New York, Vintage Books, cited by Flyvbjerg B, Making social science matter, Cambridge University Press (2001).

- [20] Denis Wood and John Fels (2008) *The Natures of Maps: Cartographic Constructions of the Natural World*, Univ Chicago Press.

Table 2: **Community “ID Card”.** The community *Hansen JP/MOLEC-DYNAMICS* contains $N = 547$ articles. Its average internal link weight is $\langle \omega_{in} \rangle \simeq 1/223$ (roughly, two random articles within the community share 1 reference over 223).

Institution	prop	σ	Authors	Nb Authored Paper
Ecole Normale Super Lyon	0.766	0.19	Hansen JP	55
Phys Lab	0.543	26.29	Barrat JL	40
CNRS	0.508	1.95	Bocquet L	38
UMR 5672	0.133	10.52	Ciliberto S	38
Univ Lyon	0.111	1.33	Geminard JC	24
Dept Phys	0.076	6.73	Holdsworth PCW	22
Univ Lyon 1	0.075	-3.41	Alastuey A	21
ENS Lyon	0.073	-1.47	Charlaix E	20
Phys Theor Lab	0.073	12.21	Dong W	20
CECAM	0.065	15.67	Cornu F	19
Subject	prop	σ	Countries	Nb Coll Paper
Physics, Multidisciplinary	0.27	16.7	France	704
Physics, Mathematical	0.254	18.67	USA	89
Physics, Fluids & Plasmas	0.16	15.77	Italy	47
Physics, Condensed Matter	0.128	10.09	England	37
Physics, Atomic, Molecular & Chemical	0.117	9.42	Germany	35
Chemistry, Physical	0.098	3.3	Netherlands	25
Materials Science, Multidisciplinary	0.069	6.25	Poland	25
Mechanics	0.065	6.7	Switzerland	23
Physics, Applied	0.064	9.03	Japan	20
Polymer Science	0.042	3.15	Chile	15
Keyword	prop	σ	Refs	Times used
DYNAMICS	0.135	14.03	HANSEN JP, 1986, THEORY SIMPLE LIQUID	60
SYSTEMS	0.117	13.65	CUGLIANDOLO LF, 1997, PHYS REV E	37
MODEL	0.104	9.9	CUGLIANDOLO LF, 1993, PHYS REV LETT	34
MOLECULAR-DYNAMICS	0.053	10.41	KOSTERLITZ JM, 1973, J PHYS C SOLID STATE	25
BEHAVIOR	0.043	6.65	GOTZE W, 1992, REP PROG PHYS	22
TRANSITION	0.042	6.12	BOUCHAUD JP, 1998, SPIN GLASSES RANDOM	22
FLUIDS	0.04	12.14	JAEGER HM, 1996, REV MOD PHYS	22
RELAXATION	0.038	7.27	ALASTUEY A, 1989, PHYS REV A	21
FLOW	0.038	6.96	FRENKEL D, 2002, UNDERSTANDING MOL SI	18
MONTE-CARLO	0.036	12.65	GRIGERA TS, 1999, PHYS REV LETT	18
Journal	prop	σ	Refs (journals)	Times used
PHYSICAL REVIEW E	0.155	19.32	PHYS REV LETT	1709
PHYSICAL REVIEW LETTERS	0.095	11.28	J CHEM PHYS	1406
JOURNAL OF CHEMICAL PHYSICS	0.067	13.92	PHYS REV B	480
EUROPHYSICS LETTERS	0.058	11.24	PHYS REV E	408
JOURNAL OF PHYSICS-CONDENSED MATTER	0.043	11.22	PHYS REV A	399
JOURNAL OF STATISTICAL PHYSICS	0.038	9.18	EUROPHYS LETT	353
EUROPEAN PHYSICAL JOURNAL B	0.029	6.78	PHYS REV E 1	343
JOURNAL OF STATISTICAL MECHANICS-THEORY AND EXPERIMENT	0.027	8.39	J STAT PHYS	310
PHYSICAL REVIEW B	0.023	2.1	NATURE	286
JOURNAL OF PHYSICAL CHEMISTRY B	0.021	4.36	PHYSICA A	276