On metrics and models for multiplex networks

Gemmetto, V.

Author: Gemmetto, V.
Title: On metrics and models for multiplex networks
Issue Date: 2018-01-16

Chapter 5

Scientific publications network

Community detection techniques are widely used to infer hidden structures within interconnected systems. Despite demonstrating high accuracy on benchmarks, they reproduce the external classification for many real-world systems with a significant level of discrepancy. A widely accepted reason behind such an outcome is the unavoidable loss of non-topological information (such as node attributes) encountered when the original complex system is converted into a network. In this chapter we systematically show that the observed discrepancies may also be caused by a different reason: the external classification itself. For this end we use scientific publication data which i) exhibit a well defined modular structure and ii) hold an expert-made classification of research articles. Having represented the articles and the extracted scientific concepts both as a bipartite network and as its unipartite projection, we applied modularity optimization to uncover the inner thematic structure. The resulting clusters are shown to partly reflect the author-made classification, although some significant discrepancies are observed. A detailed analysis of these discrepancies shows that they may carry essential information about the system, mainly related to the use of similar techniques and methods across different (sub)disciplines, that is otherwise omitted when only the external classification is considered.

The results presented in this chapter have been published in the following reference:
V. Palchykov, V. Gemmetto, A. Boyarsky, D. Garlaschelli, *EPJ Data Science*, **5**, 28 (2016).

5.1 Introduction

A conflict between two members of a relatively small university organization that happened more than 40 years ago [1] has attracted a lot of attention in the scientific community so far [2]. A confrontation during the conflict resulted in a fission of the organization, known as Zachary's karate club, into two smaller groups, gathered around the president and the instructor of the club, respectively. Predicting the sizes and compositions of the resulting factions, given the structure of the social interaction network before the split, attracted a lot of attention. This puzzle, supplemented by the known outcome, makes this system among the best studied benchmarks to test community detection algorithms [3]. Having verified a high level performance on the aforementioned system and on other benchmarks [4], community detection algorithms have then been massively applied to uncover tightly connected modules within large real-world systems. This allowed scientists to identify, for instance, Flemish- and French-speaking communities in Belgium using mobile phone communication networks [5], detect functional regions in the human or animal brain from neural connectivity [6], observe the emergence of scientific disciplines [7] and investigate the evolution of science using citation patterns and article metadata [8, 9, 10].

A bird's eye view on the identified clusters in real-world systems certifies their meaningfulness. However, an in-depth quantitative validation of the community structure requires its comparison with an external classification of the nodes, which is accessible only for a limited number of large systems. Examples include crowd-sourced tag assignments for software packages [11], product categories for Amazon copurchasing networks [12], declared group membership for various online social networks [13, 14] and publication venues for coauthorship networks in the computer science literature [13]. Surprisingly, significant discrepancies have been identified between the extracted grouping of nodes and their external classification for these systems [11, 15]. This message remains robust independently of the system under investigation and the technique used to uncover its community structure, and calls for a detailed inspection of such discrepancies in order to understand the reasons behind them.

One of the possible reasons concerns the strong simplification that occurs during the projection of the original complex system into a network. This projection may omit some crucial information that cannot be encoded into the structural connection pattern [11]. The missing information may correspond to age or gender of individuals in social networks [16, 17] or geographical position of the nodes within spatially embedded systems [18]. Following this direction, several algorithms [19, 20] have been developed in order to handle specific nodes attributes, beside the usual connectivity patterns. Such approaches have been shown to identify groups of nodes that more closely reproduce the external classification in real-world systems [20] than the techniques that rely on the connectivity patterns only.

In this chapter we argue that, independently of the aforementioned issue, the supposedly poor performance of community detection algorithms may be caused by the external classification itself and its misinterpretation. For instance, a system may possess several alternative classification schemes, such as thematic and methodological groupings in a system of scientific publications or in academic coauthorship networks [21]. In such situation, the discrepancies between the community detection results and a single accessible classification (e.g. based on thematic similarity) may carry, instead, meaningful information (e.g. about methodological similarity), therefore providing an added value to the system understanding.

Here we explore this idea by performing a detailed analysis of a scientific publication record system. This system may be simplified into a structural network representation, where the nodes correspond to scientific articles, and the links represent the relationship between them. There are various possibilities to map these relationships: direct citation [22], cocitation and bibliographic coupling [23] or content related similarities [24, 25]. In this chapter we focus on the latter, considering scientific terms or concepts that appear within the articles. Performing community detection on the corresponding network, we compare the results with an expert made classification of these articles, considering both similarities and discrepancies between the two different partitions. Then we investigate the main reasons causing the most notable deviations.

This chapter is organized as follows. In the section 5.2 we present the dataset used; in sec. 5.3 we introduce the methodology used to build the networks, extract the partitions and compare them with the external classification. Finally, in sections 5.4 and 5.5 we present our findings and discuss them.

5.2 Data

We investigate a collection of scientific manuscripts submitted to e-print repository `arXiv` [26] during the years 2013 and 2014. During the submission process, the authors were requested to classify the manuscript according to the `arXiv` classification scheme by assigning at least one category to it. In our analysis we are focussed only on the articles that have been assigned to a single category, restricting ourself to the field of physics. Moreover, the collections of manuscripts submitted during the years 2013 and 2014 will be considered separately, eliminating the possible issues related to the temporal evolution of research disciplines. The resulting datasets consist of 36386 articles submitted during 2013 and 41848 articles submitted during 2014, and will be referred below (together with the extracted contents) as the `arxivPhys2013` and `arxivPhys2014` datasets, respectively. The numbers of articles belonging to each category are shown in Tab. 5.1.

Each article is represented by a set of scientific concepts that characterize its content, i.e. specific words or combinations of them. The concepts have been identified within the full text by the `ScienceWISE.info` platform (`SW`). `SW`

category	n^{s}_{2013}	n^{m}_{2013}	n^{s}_{2014}	n^{m}_{2014}
`nucl-th`	648	1628	766	1210
`nucl-ex`	315	924	324	736
`hep-ph`	2625	3935	3116	2885
`hep-ex`	602	1726	706	1225
`hep-lat`	352	695	419	417
`hep-th`	1787	3717	2316	2960
`gr-qc`	1118	2782	1527	2204
`astro-ph`	10984	3023	11445	2437
`physics`	4452	6479	5711	4880
`cond-mat`	10549	4609	11397	3538
`nlin`	392	327	522	905
`quant-ph`	2558	3240	3187	2471
`math-ph`	0	3789	412	2668

Table 5.1: **Distribution of articles among categories.** The number of manuscript submitted during the year y that have been assigned to a given category only ($n^{\mathrm{s}}_{\mathrm{y}}$) or to the category and at least one another ($n^{\mathrm{m}}_{\mathrm{y}}$). List of categories: theoretical and experimental nuclear physics (`nucl-th` and `nucl-ex`, respectively), four branches of high energy physics (`hep-ph`: phenomenology, `hep-ex`: experiment, `hep-lat`: lattice and `hep-th`: theory), general relativity and quantum cosmology (`gr-qc`), astrophysics (`astro-ph`), physics (`physics`), condensed matter physics (`cond-mat`), nonlinear science (`nlin`), quantum physics (`quant-ph`) and mathematical physics (`math-ph`).

is a web service connected to the main online repositories such as `arXiv`, whose peculiarity is a bottom-up approach in the management of scientific concepts [27]. The initially created scientific ontology was followed by a continuous editing by the users, for instance by adding new concepts, definitions and relationships. This crowd-sourced procedure leads to the most comprehensive vocabulary of scientific concepts in the domain of physics. Such vocabulary takes care of synonyms that refer to the same concepts and it includes physics concepts explicitely labeled as generic like `mass` or `energy`, or more specific ones like `community detection`. Both are the results of crowd-sourcing by the registered expert-users.

The number k of concepts significantly vary among the manuscripts, reaching up to $k_{\mathrm{max}} \sim 400$ for review articles. The average number of identified concepts $\langle k \rangle$ per article, together with some other characteristics of the datasets `arxivPhys2013` and `arxivPhys2014`, are shown in Tab. 5.2.

	N	V	V_{gen}	$\langle k\rangle$	L_{idf}	L_{bp}
`arxivPhys2013`	36386	12200	347	37	3.3×10^{8}	1.3×10^{6}
`arxivPhys2014`	41848	12728	344	38	4.5×10^{8}	1.6×10^{6}

Table 5.2: **Basic characteristics of the datasets.** Total number of articles (N), total number of identified concepts (V) and the number of generic ones (V_{gen}) among them; $\langle k\rangle$ gives the average number of non-generic concepts within arbitrary chosen article. The number of links in a unipartite network (provided that the generic concepts are excluded) L_{idf} is two orders of magnitude larger than the corresponding number of links in bipartite networks (L_{bp})[1]. This results in significant differences in computational resources needed to perform community detection analysis.

5.3 Methods

The dataset may be represented as a network, whose nodes correspond to articles. Two nodes i and j are connected by a link if the corresponding articles share at least a single common concept. The resulting networks are extremely dense, covering almost 90% of all possible network connections; this number may be reduced to 50% if the generic concepts are ignored (see Tab.5.2). Below, to save the computational resources, we will ignore the generic concepts in our analysis. The weight of the link between two manuscripts is designed to reflect the level of content similarity between two articles, i.e. the overlap between the respective lists of concepts. Different concepts, however, may contribute differently to the similarity among two articles. Indeed, sharing a widely used concept should affect the similarity between two articles differently than sharing a specific one, suggesting that specific concepts should have a higher impact on the similarity. Each concept c in the dataset is therefore weighted according to its occurrence, which may be accounted for by the so-called *idf*(c) factor [28]:

$$\mathit{idf}(c) = \log\frac{N}{N(c)}. \tag{5.1}$$

Here N is the total number of articles and $N(c)$ is the number of articles that contain concept c. As mentioned above, among the V concepts identified by `SW`, we will consider only the specific ones, discarding the V_{gen} generic concepts. The content of each article can be therefore expressed by means of a $(V - V_{\text{gen}})$-dimensional concept vector $\vec{v}_i$. The element v_{ic} of the concept vector of the article i has non-zero value equal to $\mathit{idf}(c)$ only if the concept c appears within the article i and equals zero otherwise.

The similarity between the contents of two articles i and j, and the link weight w_{ij} between the corresponding nodes, may then be estimated by the cosine simi-

[1] These represent, in all the cases, roughly the 60% of all the links, i.e. including also the contribution given by the generic concepts.

larity between the two concept vectors $\vec{v}_i$ and $\vec{v}_j$ as follows:

$$w_{ij} = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i||\vec{v}_j|}. \tag{5.2}$$

The resulting network will be referred below as the `idf` representation of the data.

Alternatively to `idf` representation, the dataset may be mapped into a bipartite network. Such network consists of the nodes of two types that correspond to manuscripts and scientific concepts, respectively. The unweighted links in the simplest case reflect the appearance of a concept within the article. This network will be referred below to as a `bp` representation of the data, and the usage of the two alternative representation will serve the robustness of our results. The number of links (L_{idf}, L_{bp}) of these networks are shown in Tab. 5.2. As one may see, the number of links in `bp` representation is about two orders of magnitude smaller than the number of links in the corresponding `idf` representation. This has significant consequences on the run-time and memory used to analyse the networks.

Indeed, the run-time t of the Louvain algorithm scales with the number of links L of the considered network. Since empirically in the bipartite representation $L_{bp} \sim O(N)$ while in the unipartite case $L_{idf} \sim O(N^2)$, this reflects in much different computational resources required to perform the community detection. Moreover, here we point out that the bipartite representation is the most natural and suitable characterization of the dataset, since the null model behind such representation of the data is definitely more correct. In fact, the bipartite null model is consistent with the constraints on both the types of node (number of papers per concept and concepts per article). This feature is instead lost when the system is projected into a unipartite network, since the previous constraints are not matched any more. Furthermore, the bipartite representation and null model already take into account the presence of more frequent concepts, sparing us the use of any `idf` factor. In this context, we therefore propose the use of the bipartite representation as a possible alternative to the more widespread `idf` (or `tf-idf`) unipartite representation.

In order to find a unipartite network partition, we will maximize a modularity function [29]. To deal with bipartite networks, we adopt a co-clustering approach [30] and Barber's generalization of modularity [31].

In both cases, we assume that each article may belong to a single cluster only, hence exploiting the notion of non-overlapping communities. Furthermore, the co-clustering approach makes stronger restrictions on a bipartite partition, compared to a unipartite one. Indeed, the resulting clusters of a bipartite partition consist of both articles and related concepts, and we assume that each concept belongs to a single cluster as well. Such restriction may be relaxed, for instance by using alternative ways to generalize modularity for bipartite network [32] or by employing stochastic block model techniques [33]. However, we will consider co-clustering of bipartite networks since it allows us to straightforwardly employ the same greedy optimization algorithm [5] for the networks of both types.

The restriction towards a single algorithm is also caused by the result [11] that i) the selected algorithm is among the ones that perform best on real-world networks and ii) the major influence on the accuracy is related to the dataset itself rather than the algorithm. Due to the stochastic origin of this algorithm, it has been applied 100 times for unipartite networks and 1000 times for bipartite ones (due to the significantly different number of links and, therefore, the required computational resources). Among the detected partitions, for each network we will select the single partition that corresponds to the highest value of modularity; this partition will be referred below as the optimal partition for each network.

5.4 Results

A partition of a bipartite network consists of clusters that contain both articles and scientific terms (concepts), while clusters of a unipartite network partition consist of articles only. To compare both unipartite and bipartite partitions with the external article classification, we will be focussed only on the articles that fall into each cluster. Thus, by referring below to a cluster of bipartite partition we mean the set of articles that belong to the specified cluster. In this perspective, the external classification of the articles is represented by the `arXiv` standard split into different subject classes or categories (`astro-ph`, `cond-mat`, etc.).

Then, given two partitions P and Q of the same network (for instance a detected network partition and the `arXiv` classification), an initial comparison between them has been performed using an information-based symmetrically normalized mutual information:

$$I_{\mathrm{N}}(P,Q) = \frac{2I(P,Q)}{H(P)+H(Q)}. \tag{5.3}$$

Here $I(P,Q)$ is the mutual information [34] between two partitions P and Q, and $H(P)$ is the entropy of partition P. The normalized mutual information $I_{\mathrm{N}}(P,Q)$ may vary between 0 and 1. A value of 0 indicates that the two partitions have no information in common, while a value of 1 corresponds to identical partitions. In Tab. 5.3 we show the level of similarity between each optimal partition and the `arXiv` classification ones. The reported values of normalized mutual information indicate the existence of some common information between automatically identified clusters of articles (both in the bipartite and unipartite cases) and the author based classification. However, the values being quite far from the possible maximum of 1 reflect evidence for some discrepancies between the partitions. Below we perform a detailed analysis of these discrepancies. Here we will show the results for the `arxivPhys2013` dataset; similar findings can be observed in the `arxivPhys2014` case and are shown in the following appendix.

The first difference is observed in the numbers of detected clusters and of `arXiv` subject classes: while the number of categories in the `arXiv` classification

	idf	bp
arxivPhys2013	0.60 ± 0.02	0.56 ± 0.03
arxivPhys2014	0.55 ± 0.00	0.54 ± 0.02

Table 5.3: **Similarity between network partitions and external classification.** Average value of the normalized mutual information I_{N} (5.3) between a partition of each network representation and arXiv classification of the articles and the corresponding standard deviations. Both bp and idf partitions demonstrate similar value of closeness to arXiv classification.

scheme is 12 [2], the number of clusters in our partitions is only equal to 4 in the idf and to 6 in the bp network representations, respectively[3]. Indeed, the articles of some different arXiv categories tend to belong to a single cluster. This may be clearly observed in Fig. 5.1 that shows the fraction of articles of each arXiv category belonging to each cluster in the resulting partitions. This merger is especially visible for different high energy physics (hep) categories (hep-ph, hep-ex, hep-lat and hep-th): in the idf partition, almost 99% of all these articles fell into a single cluster, independently of the sub-field. This result, despite deviating from the arXiv classification scheme, is reasonable since we observe a union of almost all papers about high energy physics, no matter if they deal with experimental or theoretical issues.

Instead, in the bp partition the articles of the four hep categories are almost entirely distributed among two clusters, focussed on experimental and theoretical issues, respectively. The first of them joins 95% of all articles that belong to experimental categories (hep-ph, hep-ex or hep-lat), while the second one contains 94% of all theoretical (hep-th) articles. Thus, the presence of more clusters within the bipartite network partition allows us to identify methodologically different clusters of articles within the hep categories, in particular dividing theoretical papers from experimental ones.

Even though the split of hep articles into two groups may be simply explained by the different approaches used to study the phenomena, a further result can be observed from Fig. 5.1: in the bipartite network partition, hep-th articles tend to form a single cluster with the articles that belong to general relativity and quantum cosmology (category gr-qc) rather than with the other high energy physics articles, thus appearing to be more similar to gr-qc papers rather than to the other hep ones. Intuitively, indeed, we know that both hep-th and gr-qc both focus mostly on general relativity, while the other hep categories focus on particle physics [4].

[2] In fact, there are 13 physics categories in arXiv classification scheme, but there is no single article in arxivPhys2013 dataset that belong to math-ph category only.

[3] By performing a detailed comparison we ignore all single-node clusters, which contain the articles for which no concept has been identified.

[4] Indeed, it is very likely that nowadays the hep- categories would be split in multiple subcategories (namely hep-th, hep-lat, etc.). However, here we point out that our study (in particular in the

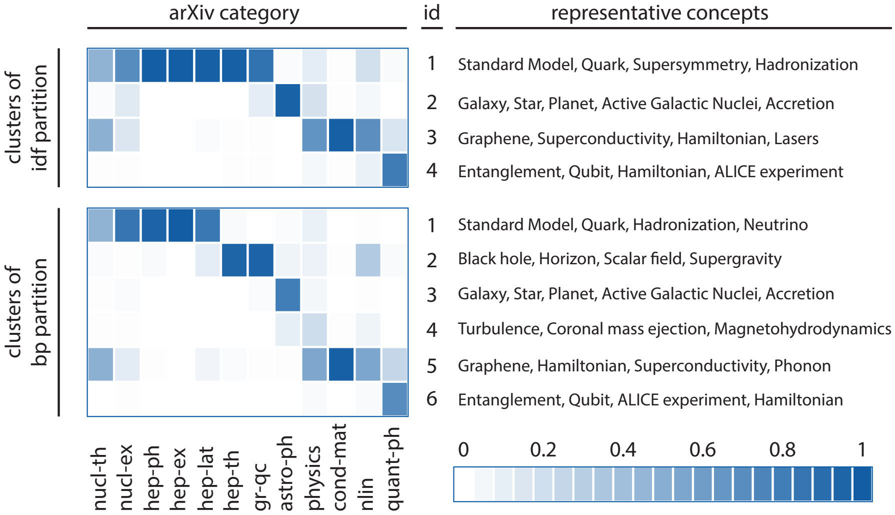

Figure 5.1: **Inner composition of arxivPhysics2013 partitions.** The color of each cell accounts for the fraction of articles of a given category belonging to a cluster (each column sums to 1). The articles of the same categories tend to incorporate into single clusters as justified by the clearly visible block-diagonal structure of both `idf` and `bp` partitions. Nevertheless, the split of some categories into distinct clusters may be observed. For instance, the articles of `nucl-th` category are roughly equally split among `hep-` and `cond-mat`-dominated categories. On the right, the most representative concepts for each cluster are shown.

Such a relatedness between the articles of the two theoretical physics categories (`hep-th` and `gr-qc`) may be verified independently by a category co-occurrence analysis. To show this, we will use the complementary part of the investigated dataset. This set consists of all articles that have been submitted to `arXiv` during the same 2013 year, but for which the authors have assigned at least two different categories. Thus, no article of this set overlaps with the clustered `arxivPhys2013` collection. Irrespective of the details of the decision-making process through which authors assign multiple categories, this multiplicity reflects the authors' decision that the scope of the article can not be properly covered by a single category of a given classification scheme. Whilst several categories may cover the scope of a single research article, the co-occurrence of the same two categories in a significant fraction of articles may reflect some hidden relationships between them. The corresponding empirical co-occurrence matrix is shown in Fig. 5.2 and indicates

bipartite case) shows that `hep-th` looks actually more similar to `gr-qc` than to the other `hep-` classes. This therefore seems to strengthen the apparently counterintuitive choice of dividing the high energy articles in different primary classes.

the fraction of articles of a given category that have been co-submitted to the other categories. The diagonal elements of this matrix indicate the fraction of articles of each category that have been assigned to a single category by the author(s), i.e. the articles of the arxivPhys2013 dataset. A normalization procedure has been performed such that each column of the matrix sums to 1.

Fig. 5.2 confirms that the hep-th subject class is indeed more related to the gr-qc class than to the other hep categories: hep-th co-occurred with gr-qc in 1721 articles, and with all other hep categories in only 1286 articles, even though the number of the corresponding hep papers (hep-ph, hep-ex, hep-lat) exceeds the number of gr-qc ones threefold. This high level of relatedness between hep-th and gr-qc categories justifies the merging of the articles of these categories into a single cluster and indicates the meaningful deviation from the arXiv classification scheme. It is worth to mention that in the idf partition, where all hep category articles tend to belong to a single cluster, the same cluster is supplemented by 87% of all gr-qc articles, in agreement with the result observed above. Moreover such a tendency is not restricted to the dataset for the selected year: it has also been observed for the arxivPhys2014 one (as shown in the appendix).

The same approach explains the presence of a significant fraction of physics, non-linear (nlin) and quantum physics (quant-ph) articles into the cond-mat clusters. It also allows us to understand a possible reason why nuclear physics articles (both theory and experiment) occur significantly within the hep clusters. However, it cannot explain the presence of roughly one half of nucl-th articles into the condensed matter cluster (cluster No. 3 in idf and No. 5 in bp partitions) in both network representations. The latter deviation from the article classification, which is not explained by category co-occurrence, does not exclude that similarities between these topics exist but are considered not strong enough by the authors to label the articles with both subject classes. To uncover the possible essence of these similarities, we examine the top representative concepts that characterize the nucl-th articles that belong to the two different clusters, see Table 5.4. In both cases, the top representative concepts contain the ones that characterize the object of investigation within theoretical nuclear physics, such as Isotope, Isospin or Nuclear matter. However, one may clearly identify method-related concepts, such as Hartree-Fock, Hamiltonian, Mean field and Random phase approximation, among the top representative concepts of articles in the cond-mat cluster. These concepts clearly characterize methods that are widely used in condensed matter physics research, and that have not been identified among top concepts in any other cluster. This result emphasizes the ability of scientific concepts found within research articles to highlight not only topics focussed on the same objects, but also methodologically similar research directions.

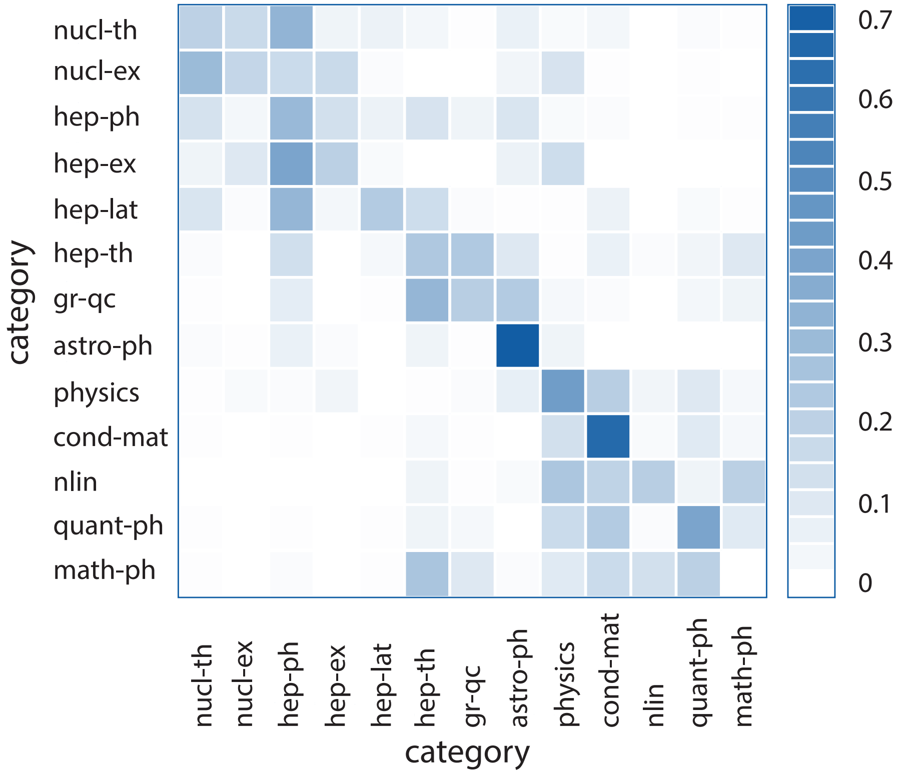

Figure 5.2: **Co-occurrence matrix of `arXiv` categories during year 2013.** Built on the complementary dataset to `arxivPhys2013`, this matrix reflects the relationships between `arXiv` categories and allows to justify the meaningfulness of some remarkable discrepancies, like the merger of `hep-th` and `gr-qc` articles. Each non-diagonal element reflects the fraction of articles in which two specified categories have co-occurred. The diagonal cells represent the fractions of articles that have been assigned to a single category, i.e. they concern the articles of the `arxivPhys2013` dataset. A normalization procedure has been performed such that each row of the matrix sums to 1. Thus, the aforementioned fractions correspond to the fractions of manuscripts that have been labeled with a given category.

%	Concept (cluster no. 1)	%	Concept (cluster no. 3)
43	Hadronization	55	Isotope
39	Isospin	53	Hamiltonian
37	Pion	39	Hartree-Fock
33	Degree of freedom	36	Quadrupole
32	Heavy ion collision	34	Isospin
31	Quark	31	Nuclear matter
29	Chirality	30	Degree of freedom
29	Hamiltonian	28	Mean field
29	Nuclear matter	26	Harmonic oscillator
26	Coupling constant	25	Spin orbit

Table 5.4: **Representative concepts of two groups of articles categorized as `nucl-th`.** The left side of the table represents the group of articles that fell into the `hep` dominated cluster (no. 1) in `idf` partition. The right side – the other group: the `nucl-th` articles that fell into the `cond-mat` dominated cluster (no. 3). For each group, the numbers next to the concepts give the percentage of articles in which the concept has been identified. The table allows us to make a suggestion that the two groups of articles significantly differ by the methods used to investigate nuclear matter.

5.5 Conclusions

The differences between the outcomes of community detection algorithms and possible external classifications may have various reasons. The most notable of them concern a possible failure of the considered algorithm or the unavoidable loss of data about real complex systems determined by their representation as networks. To deal with the first issue, algorithms are heavily tested on benchmarks, while the second issue is still under investigation [20]. In this chapter, we emphasize a third possible reason behind such discrepancies, i.e. the fact that the external classification itself may possess its own limitations. For this reason we performed a detailed investigation of a scientific publication system which i) may be naturally represented as a network and ii) owns an external author-made classification of scientific articles. While, indeed, some discrepancies are caused by the lack of data (for instance in the case of the articles for which no concept has been identified), we argue that the most remarkable of them may reflect real commonalities across different subject classes. Academic publications are traditionally categorized and classified[5] according to objects or phenomena under investigation. The same phenomena, however, may be explored using various approaches, experimental observation and theoretical modeling being among them.

[5] Document classification and categorization are different processes: classification refers to the assignment of one or more predefined categories to a document, while categorization refers to the process of dividing the set of documents into priory unknown groups whose members are in some way similar to each other [35].

On the other hand, the phenomena that belong to different research topics may be investigated using the same methods, composing the core of the interdisciplinary research. Thus, a more comprehensive classification or research articles may be represented by a two layer categorization scheme, where one layer reflects phenomena or objects while the other one stands for the methods of investigation. Usually, these two layers are not taken equally into account. The expert made classification may include rather a strong bias towards the object layer. The reasons involve the classification scheme itself and the limited knowledge about all other research disciplines that employ the same methods. Instead, automatic concept-based categorization has no direct preference for any of the layers: the extracted concepts correspond both to phenomena and methods, and the algorithm has no information about the possible division of the concepts. Thus, the observed discrepancies may reflect the dominance of the methodological layer over the other one, which corresponds to phenomena or objects. Similar results have been previously observed within the collaboration network of scientists at Santa Fe Institute [21], where, besides the expected grouping around common topics, some methodologically driven clusters have been observed.

This shows that the failure in reproducing an external classification may indicate a genuinely more complicated organization within the system, in addition to the lack of data or algorithmic mistakes. Besides developing sophisticated algorithms to deal with real systems, we should therefore keep in mind that some observed discrepancies may go beyond the standard classification and carry important information about the system under study. We believe that similar results may be observed in other systems. Indeed, the ground truth necessarily follows from a given classification criterion; however, the considered data may contain more than that single type of information (perhaps in conflict one with each other). In general, therefore, it may happen that what we consider as the ground truth is just one of the possible reference points, rather than some absolute truth. Understanding the information employed to define the so-called ground truth is therefore crucial in order to perform a proper comparison between external classification and automatically retrieved communities.

Appendix

5.A Scientific publications network in 2014

Here we show the results of the community detection algorithm to the so-called `arxivPhys2014` dataset, representing the content-relations between 41848 scientific articles that have been assigned to a single physics category, submitted to `arXiv` in 2014; our findings are reported in Figure 5.3 (top panel). The partitions obtained through the Louvain algorithm are very similar to those observed for the `arxivPhys2013` dataset: we see that, also in this case, the manuscripts belonging to the same category tend to merge into single clusters as illustrated by the block-diagonal structure of both `idf` and `bp` clusterings. Still, the split of some categories into different communities may be observed, such as `nucl-th` and `math-ph`.

Furthermore, we can justify our results based on the co-occurrence matrix reported in Figure 5.3 (bottom panel). This matrix, built on the complementary dataset of `arxivPhys2014`, namely the set of articles showing more than one physics category, reflects the relations between the various `arXiv` categories in 2014 and can therefore explain the reason of some of the observed discrepancies, such as the union of `hep-th` and `gr-qc` manuscripts.

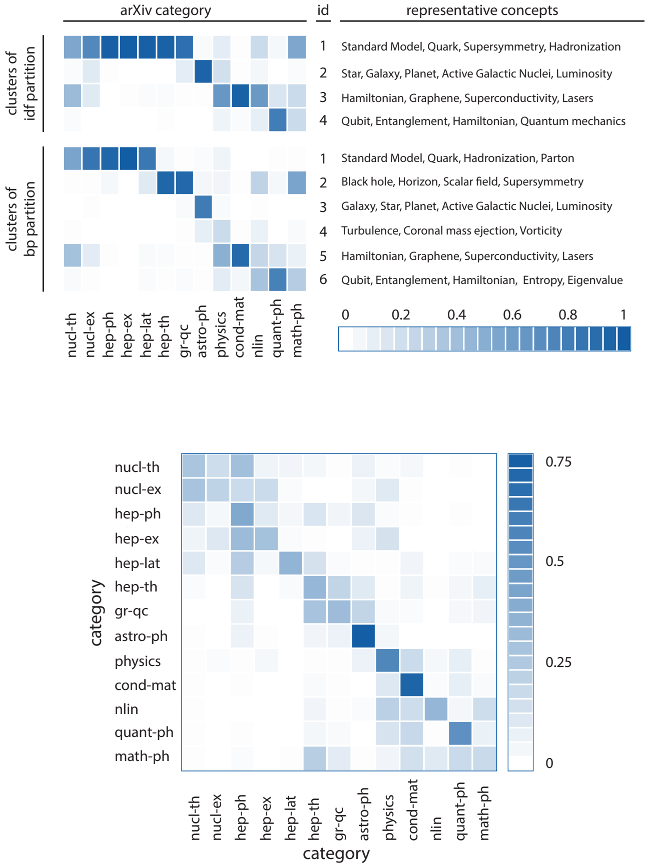

Figure 5.3: **Results of the analysis for the `arxivPhys2014` dataset.** Top: inner composition of the obtained partitions. The color of each cell accounts for the fraction of articles of a given category belonging to a cluster (each column sums to 1); the articles of the same categories tend to incorporate into single clusters as justified by the clearly visible block-diagonal structure of both `idf` and `bp` partitions. Bottom: co-occurrence matrix of `arXiv` categories during year 2014. Each non-diagonal element reflects the fraction of articles in which two specified categories have co-occurred; the diagonal cells represent the fractions of articles that have been assigned a single category, i.e. they concern the articles of the `arxivPhys2014` dataset. A normalization procedure has been performed such that each row of the matrix sums to 1.

Bibliography

[1] W. W. Zachary (1977) 'An information flow model for conflict and fission in small groups', *Journal of Anthropological research*, 452

[2] M. E. J. Newman (2012) 'Communities, modules and large-scale structure in networks', *Nature Physics* **8** (1), 25

[3] S. Fortunato (2010) 'Community detection in graphs', *Physics Reports* **486** (3), 75

[4] A. Lancichinetti, S. Fortunato, F. Radicchi (2008) 'Bechmark graphs for testing community detection algorithms', *Physical Review E* **78** (4), 046110

[5] V. Blondel, J.-L. Guillame, R. Lambiotte, E. Lefebvre (2008) 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment* **2008** (10), P10008

[6] E. Bullmore, O. Sporns (2009) 'Complex brain networks: graph theoretical analysis of structural and functional systems', *Nature Reviews Neuroscience* **10**, 186

[7] N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima (2008) 'Detecting emerging research fronts based on topological measures in citation networks of scientific publications', *Technovation* **28** (11), 758

[8] M. Herrera, D. C. Roberts, N. Gulbahce (2010) 'Mapping the evolution of scientific fields', *PloS ONE* **5** (5), e10355

[9] M. Rosvall, C. T. Bergstrom (2010) 'Mapping change in large networks', *PloS ONE* **5** (1), e8694

[10] P. Chen, S. Redner (2010) 'Community structure of the physical review citation network', *Journal of Informetrics* **4** (3), 278

[11] D. Hric, R. K. Darst, S. Fortunato (2014) 'Community detection in networks: structural communities versus ground truth', *Physical Review E* **90** (6), 062805

[12] J. Leskovec, L. A. Adamic, B. A. Huberman (2007) 'The dynamics of viral marketing', *ACM Transactions on the Web (TWEB)* **1** (1), 5

[13] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan (2006) 'Group formation in large social networks: membership, growth, and evolution', *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 44

[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee (2007) 'Measurements and analysis of online social networks', *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 29

[15] J. Yang, J. Leskovec (2015) 'Defining and evaluating network communities based on ground-truth', *Knowledge and Information Systems* **42** (1), 181

[16] V. Palchykov, K. Kaski, J. Kertész, A. L. Barabási, R. I. M. Dunbar (2012) 'Sex differences in intimate relationships', *Scientific Reports* **2**, 370

[17] L. Kovanen, K. Kaski, J. Kertész, J. Saramäki (2013) 'Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences', *Proceedings of the National Academy of Sciences USA* **110** (45), 18070

[18] P. Expert, T. Evans, V. Blondel, R. Lambiotte (2011) 'Uncovering space-independent communities in spatial networks', *Proceedings of the National Academy of Sciences USA* **108** (19), 7663

[19] C. Bothorel, J. D. Cruz, M. Magnani, B. Micenkova (2015) 'Clustering attributed graphs: models, measures and methods', *Network Science* **3** (3), 408

[20] M. E. J. Newman, A. Clauset (2016) 'Structure and inference in annotated networks', *Nature Communications* **7**, 11863

[21] M. Girvan, M. E. J. Newman (2002) 'Community structure in social and biological networks', *Proceedings of the National Academy of Sciences USA* **99** (12), 7821

[22] L. Waltman, N. J. Van Eck (2012) 'A new methodology for constructing a publication-level classification system of science', *Journal of the American Society for Information Science and Technology* **63** (12), 2378

[23] K. W. Boyack, R. Klavans (2010) 'Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately?', *Journal of the American Society for Information Science and Technology* **61** (12), 2389

[24] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner (2011) 'Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches', *PloS ONE* **6** (3), e18029

[25] P. Glenisson, W. Glänzel, F. Janssens, B. De Moor (2005) 'Combining full text and bibliometric information in mapping scientific disciplines', *Information Processing & Management* **41** (6), 1548

[26] An electronic archive and distribution server for research articles, `http://arxiv.org`

[27] R. Prokofyev, G. Demartini, A. Boyarsky, O. Ruchayskiy, P. Cudré-Mauroux (2013) 'Ontology-based word sense disambiguation for scientific literature', *Advances in Information Retrieval*, 594

[28] K. S. Jones (1973) 'Index term weighting', *Information Storage and Retrieval* **9** (11), 619

[29] M. E. J. Newman, M. Girvan (2004) 'Finding and evaluating community structure in networks', *Physical Review E* **69** (2), 026113

[30] D. M. Blei, A. Y. Ng, M. I. Jordan (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* **3**, 993

[31] M. J. Barber (2007) 'Modularity and community detection in bipartite networks', *Physical Review E* **76** (6), 066102

[32] R. Guimerá, M. Sales-Pardo, L. A. N. Amaral (2007) 'Module identification in bipartite and directed networks', *Physical Review E* **76** (3), 036102

[33] D. B. Larremore, A. Clauset, A. Z. Jacobs (2014) 'Efficiently inferring community structure in bipartite networks', *Physical Review E* **90** (1), 012805

[34] M. Meilă (2007) 'Comparing clusters – an information based distance', *Journal of Multivariate Analysis* **98** (5), 873

[35] E. K. Jacob (2004) 'Classification and categorization: a difference that makes a difference', Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign