International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, 2011.

Evaluating cooperation in citation datasets using core structures

C Giatsidis[1], D M Thilikos[2], M Vazirgiannis[1,3], K Berberich[4]

[1]LIX, École Polytechnique, Palaiseau Cedex, France
[2]Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece
[3]Department of Informatics, Athens University of Economics, Athens, Greece
[4]Max-Planck-Institut für Informatik, Saarbrücken, Germany

E-mail: xristosakamad@gmail.com, sedthilk@thilikos.info, mvazirg@yahoo.gr, kberberi@mpi-inf.mpg.de

Abstract. Community subgraphs are characterized by dense connections or interactions among their nodes. Community detection and evaluation is an important task in graph mining. A variety of measures have been proposed to evaluate the quality of such communities. In this paper, we evaluate communities capitalizing on the k-core structure, as means of evaluating their collaborative nature – a property not captured by the single node metrics or by the established community evaluation metrics. Based on the k-core concept, which essentially measures the robustness of a community under degeneracy, we extend it to graphs with weighted edges, devising the novel concept of fractional core for undirected graphs with edge-weighted edges. We applied these approaches to large real-world graphs investigating the co-authorship case for citation datasets from Computer Science (DBLP) and High Energy Physics (ARXIV.hep-th). Our findings are intuitive and we report interesting results and observations with regards to collaboration among authors.

1. Introduction

Large and evolving graphs constitute an important element in current large-scale information systems. Common cases of such graphs are the Web, social networks, citation graphs, CDRs (Call Data Records) where nodes – featured with, in some cases, many attributes – are connected to each other with directed edges, representing a relation such as endorsement, recommendation or friendship. In all cases, due to the economic significance of these networks, the ranking of individual nodes is a cornerstone necessity.

Graphs of real-word data with community structure have vertex degree with a wide range. As pointed out in [1], nodes of low degree coexist with nodes of high degree making the graph inhomogeneous both globally and locally which usually indicates particularities in its structure, for instance, communities.

Community sub-graphs are characterized by dense connections or interactions among its nodes. Community detection and evaluation is an important task in graph mining. A variety of measures has been proposed to evaluate the quality of such communities. In this paper, we evaluate communities based on the k- core concept, as a means of evaluating their collaborative nature. A *k-core* (or a *core of index k*) in a graph G is a subgraph H with all vertices having at least k neighbors in H. If the vertices of a graph G represent a set of entities and its edges represent collaboration links between them, then a core of high index in a graph G can be seen and treated as a community of entities that demonstrates a strong collaboration between them. The *degeneracy* of a graph G is the maximum index k of a non-empty core in G and can be seen as a measure of the overall density of a graph. The graph theoretic study of degeneracy and k-cores dates back to the 60's [2–5]. Moreover, both notions have been extensively used, in an experimental level, for evaluating and detecting strongly cohesive communities in real-word graphs [6–12]. In this paper, we use and extend these concepts in order to evaluate cohesiveness in bipartite graphs where edges represent relations between two different types of entities. Our experiments concern bipartite graphs corresponding to author/paper relations in several bibliographical databases.

Let $G = (A, P, E)$ be a bipartite graph where edges represent relations between two disjoint sets A and P. For instance, such a graph G may represent a bibliographic dataset where A are the authors, B are the papers, and an edge $\{a, p\}$ belongs in E if a is an author of the paper p. A natural way to extract the collaboration graph relating the authors of A is the following: we define the *co-authorship graph* H_G where the vertices are the authors and where we add an edge between two authors if and only if they are co-authors of some paper. As a first step of our work, we construct this graph for the considered bibliographical datasets and we use the k-core structure in order to detect in them groups of authors with a strong collaborative relation. We stress that a different graph is defined by the "dual" construction where the vertices are the papers of the dataset and an edge is added between two papers if and only of they have some common author (this alternative approach has been adopted in [13]).

As our experiments indicate, the high rank cores of the co-authorship graph H_G are strongly biased towards including papers written by many authors. Moreover, the same criterion of building a co-authorship graph does not take into account neither the number of papers two authors may have in common nor the number of authors contributing to each paper. For these reasons and towards defining a more objective way of quantifying collaboration, we introduce an edge-weighted enhancement of H_G where weights of edges reflect the "essential collaboration effort" between the authors represented by their endpoints (see Section 3 for details). As this weighting assigns fractional weights to the edges, the degrees of the vertices are now fractional. For this reason, we introduce a fractional analogue of the k-core concept in order to study and detect coherent communities of authors with a high level of collaboration. Also we define the fractional analogues of the main combinatorial concepts related to the cores and the degeneracy of a graph. To our knowledge, this is the first time that the concept of k-cores is adapted edge-weighted graphs and we believe that this new concept has independent interest both in theory and in practice, especially when dealing with relations represented by hypergraphs (represented by their incident bipartite graph) where relations may have arbitrary arity.

Our experiments concern two datasets: the DBLP and the ARXIV on High Energy Physics - Theory (ARXIV.hep-th). We perform an extended experimental evaluation studying in depth the core subgraphs, both integer and fractional, of their edge-weighted co-authorship graphs and we visualize a series of cores, selected under certain filtering criteria. Also, for each dataset, we visualize the connected component forests (a concept defined in [13]) depicting the way connected components of fractional cores evolve as their index is increasing. Our experiments resul in several interesting and intuitive findings. A fully functional demonstrator for the DBLP co-authorship graph available at:

`http://www.graphdegeneracy.org/`

A preliminary version of this work appeared in [14]. In this paper we significantly extended that preliminary work as follows:

- We articulate the method and the metrics in a more principled and rigorous manner.
- We define the additional concepts of *fractional core sequence* and *fractional index sequence*. This concepts are critical for obtaining a deeper intuition on the concept of degenerate graphs, especially in what concerns our proposed extension from the integer to the fractional setting.
- We extended our experimental evaluation by introducing an additional bibliographic dataset from theoretical Physics ARXIV.hep-th and we applied an extensive comparison and evaluation of the derived experiments.
- We additionally used the *Core Decomposition Forest* of a graph in order to visualize and discuss the interesting properties of the connected components in the k-cores –integer or fractional– for both considered datasets.

The paper is organized as follows. In Section 2, we present related work on cores and their use to evaluate the behavior and structure of complex network topologies while in Section 3 we introduce the theoretical aspects of our methodology. In Section 4 we present our experiments on the integer and fractional cores of our datasets and we make several visualizations of the extracted information. The notion of Core Decomposition Forest is introduced in Section 5, where we use it in order to visualize the evolution of the cores (both integer or fractional) of our datasets. We conclude in Section 6, by presenting several issues on further work and experimentation based on the concepts defined in this paper.

2. Related work

A thorough review on community detection in graphs is offered by Fortunato in [1]. In that work techniques, methods, and datasets are presented for detecting communities in sociology, biology and computer science, disciplines where systems are typically represented by graphs. Most existing relevant methods are presented, with a special focus on statistical physics, including discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other.

Cohesion measures on graphs. Studying the general behavior and properties of real graphs, both edge-weighted and unweighted, is the subject of [15] where a pattern on the behavior of connected components over time is observed and, upon that, a generative model is build.

In recent literature, various metrics are proposed relevant to the graph structure of a social network. Such are "Betweenness" [16], "Centrality" [17], and "Clustering coefficient (a measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater "cliquishness", i.e. cohesion degree or density. Of special interest here is the eigenvector centrality – a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to nodes having a high score contribute more to the score of the node in question. Other measures include "path length" (i.e. distances between pairs of nodes in the network), "prestige/authority", a measure in directed graphs to describe a node's centrality, and "radiality", a notion representing an individual's capacity to reach out the whole network (i.e., its influence). Other interesting measures include "Structural cohesion" - the minimum number of members who, if removed from a group, would disconnect the group [18]. In [19] an alternative "core notion" is considered for the case of directed graphs where a core is seen as a complete bipartite graph where all edges are directed from the one part to the other. In [19], such cores are detected and are then fed to a generalized HITS algorithm used to expand the communities within them. In [20], greedy approximation algorithms are proposed for finding the dense components of a graph. Both undirected and directed graphs are examined. In the case of directed graphs the vertices are divided into hubs

(S) and authorities (T). Then, based on a value of $|S|/|T|$, a greedy algorithm removes the vertex of minimum degree from either S or T until both sets are empty. Also, in [21], the subject of finding dense subgraphs, based on query nodes, is studied, where the issue is to find a community that contains certain given nodes.

Cores. The k-cores are fundamental structures in graph theory and their study dates back to the 60's [2–5]. A *k-core* of a graph G is the maximum subgraph of G where each vertex in H has at least k neighbors in H. The *degeneracy* of a graph is defined as the biggest k for which a graph contains a non-empty k-core [22]. The same notion has appeared with several names such as *width* [23], *linkage* [24,25], or *coloring number* [26] and has been proven to be equal to the smallest k for which we can find a linear ordering of the vertices of the graph such that for each vertex v, the number of its neighbors that appear before v in the ordering is at most k (see [4,22,24]).

The existence of k-cores of large size in sufficiently dense graphs has been theoretically studied by [27] for random graphs generated by the Erdős-Rényi model [28]. As shown in [27], a k-core whose size is proportional to the size of G (i.e. a "giant" k-core) "suddenly" appears in a random graph with n vertices and m edges when m reaches a threshold $c_k \cdot n$, for some constant c_k that depends exclusively on k. Also, it was proved in [29,30] that, in the Erdős-Rényi model, almost all k-cores are k-connected (see [31] for more recent results on this topic).

An efficient algorithm for the computation of the k-core of a graph was given in [32] and its running time is proportional to the number of edges of the input graph. Actually, the algorithm in [32] can compute the *core decomposition* of a graph consisting of the sequence of all the non-empty the i-cells of G where each i-cell is defined as the vertices contained in the i-core but not in the $(i+1)$-core. Core decompositions give useful information on the way subgraphs of a graph are clustered according to their degrees and has been used extensively in several topics such as the study of internet topology [6,7], large scale network visualization [8–10], networks of protein interaction [11,12], and complex network modeling and organization [33,34]. A more general notion of k-cores was introduced in [35] where, instead of vertex degrees, more general functions where considered.

3. The fractional core method

Our motivation is the detection of the cohesive parts of community graphs with special emphasis to the DBLP and the ARXIV.hep-th co-authorship datasets. Our main theoretical tolls are the notions of k-cores and fractional k-cores that will be defined in the next two subsections.

3.1. Preliminaries

All graphs we consider are undirected and simple (i.e. they do not have multiple edges or loops). We denote the vertex and the edge set of a graph by $V(G)$ and $E(G)$ respectively. We also refer to the cardinality of $V(G)$ as the *size* of G. We also consider *edge-weighted* graphs (or, simply *weighted* graphs) and we denote them by pairs $(G, \mathbf{w})$ where $\mathbf{w}$ is a weighting function assigning rational numbers to the edges of G. We say that a graph H is a subgraph of a graph G if H occurs from H after removing vertices or edges (the removal of an edge implies the removal of all edges that are incident to it). A graph is *connected* if for every pair of its vertices there is a path connecting them. A *connected component* of a graph is a maximal connected subgraph of it. Given a graph G, we denote by $g(G)$ the size of the biggest connected component of G and we call it *giant* component.

Given a vertex $v \in V(G)$, the *degree* of v in G is the number of edges that are incident to it. We also denote by $\delta(G)$ the minimum degree of a vertex in G. The *degeneracy* of a graph G is defined as follows.

$$\delta^*(G) = \max\{\delta(H) \mid H \text{ is a non-empty subgraph of } G\}.$$

Definition 3.1 *Given a graph G and a non-negative integer k, the* k-core *of G is defined as the maximum size subgraph H of G where $\delta(H) \geq k$. It is easily to see that such a subgraph is unique. Given a k-core, we refer to k as its* core index *or simply* index.

Notice that, for each $i \leq j$, the j-core of a graph is a subgraph of its i-core. Notice that the degeneracy of a graph is the maximum k for which G contains a non-empty k-core. Given a graph G where $\delta^*(G) = d$ and an integer i where $0 \leq i \leq d$, we denote by G_i the i-core of G and we define $\mathcal{G}(G) = G_0, G_1, \ldots, G_d$ as the *core sequence* of G, where $G_0 = G$ and G_d is the *densest* core of G. For every $i \geq 0$, the graph G_{i+1} can be computed by the following simple procedure.

Procedure *Trim*(G, k)
Input: An undirected graph G and positive integer k.
Output: the $(k+1)$-core of G
1. let $F := G$.
2. **while** there is a node x in F such that $\mathbf{deg}_F(x) \leq k$
   delete node x from F.
3. **return** F.

The *Trim*(G, k) procedure runs in $O(kn)$ steps, thus computations are feasible even in large scale graphs [32]. Applying successively *Trim*(G, i), for $i = 0, \ldots, \delta^*(G) - 1$, gives a fast way to compute the core sequence of G. In fact an optimal implementation of the above pruning procedure that is able to produce the core sequence of a graph in $O(\delta^*(G) \cdot n)$ steps has been given in [32]. In fact, the procedure in [32] works for much more general variants of the core notion, including the fractional core notion that will be defined later in this section.

Definition 3.2 *The* core index *of a vertex v of G is the maximum k for which v belongs in the k-core of G.*

Notice that one may also define the *core index of a set S* of vertices in G as the maximum k for which all vertices of S belong in the k-core of G [35]. It is easy to see that this number is the minimum core index of all the vertices in S.

3.2. Cores for bipartite graphs

The datasets that we study are represented by bipartite graphs where edges denote relations between papers and authors. We denote such a graph by $G = (A, P, E)$ where A is the set of authors, P is the set of papers, and E is a set of edges. Each edge $\{x, y\}$ (where $x \in A$ and $y \in P$) expresses the fact that x is one of the authors of paper y. As what we aim is to evaluate the collaboration between authors, we restrict our study to the papers that are written by at least two authors, i.e., we assume that all the vertices in P have degree at least two.

The *co-authorship graph* corresponding to G is defined as follows:

$$H_G = (A, \{\{x, x'\} \mid \exists y \in P : \{x, y\}, \{x', y\} \in E), \tag{1}$$

i.e., two authors are adjacent if they appear as co-authors in at least one paper. Notice that the above definition of H_G is radically different from the one used in [13], where they study graphs whose vertices correspond to authors and edges indicate joint publications between two authors. In fact, the construction in [13] can be seen as being the dual of the one we used for creating H_G in the sense of vertex-edge duality of hyper-graphs.

For each dataset (represented by a bipartite graph G), we compute $\delta^*(H_G)$ and the core index of each vertex/set of vertices in H_G in order to evaluate the collaboration behavior in the bipartite graph G and the dataset that it represents. The idea of our criterion is to locate communities of authors with a high collaboration between them in the sense that we do not just demand that they have authored many papers but also that they have all authored them with authors in the same community.

However, this is not an entirely satisfactory evaluation, since the number of authors on a paper has no impact in this measure. For this reason, we introduce below a more refined way to define cores based on the notion of a fractional core.

3.3. Fractional k-cores for edge-weighted graphs

Let $G = (A, P, E)$ be a bipartite graph where all vertices in P have minimum degree 2. Given an author vertex $X \in A$, we define the neighborhood $N_G(x)$ of x as the set containing each paper $y \in P$ for which $\{x, y\} \in E$, i.e., $N_G(x)$ is the set of papers co-authored by x. Symmetrically, we define the neighborhood $N_G(y)$ of a paper $y \in P$, i.e., the set of the authors of paper y. Also, given an author x we denote by $E_G(x)$, the set of all edges that are incident to x in G. In what follows, we denote by $\mathbb{Q}^+$ the set of all non-negative rational numbers.

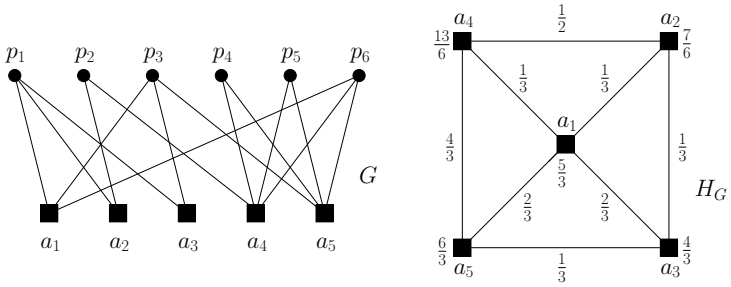

Figure 1. An example of a bipartite graph G and its edge-weighted co-authorship graph, $(H_G, \mathbf{w})$.

Definition 3.3 *Given a bipartite graph $G = (A, P, E)$, we define the* edge-weighted co-authorship graph*, denoted by $(H_G, \mathbf{w})$, by taking H_G, as defined in* (1)*, and setting up a rational weight function $\mathbf{w} : E \to \mathbb{Q}^+$ on the edges of H_G as follows: For every edge $e = \{x, x'\}$ we set*

$$\mathbf{w}(e) = \sum_{y \in N_G(x) \cap N_G(x')} \frac{1}{N_G(y)}.$$

Notice that, $\sum_{e \in H_G} \mathbf{w}(e) = |V(P)|$, i.e. the sum of all the weights on the edges is the size of the graph, i.e., the number of its vertices. For example, in Figure 1, in order to compute the weight of the edge $e = \{a_1, a_3\}$, one should observe that the authors a_1 and a_3 are co-authors of the papers p_1 and p_3. As p_1 and p_3 have 3 authors each, they contribute $1/3$ to the weight of e, that is $\mathbf{w}(e) = 2/3$. This weighting of e expresses the fact that the collective effort of author a_1 to the papers he/she co-authored with p_3 is of $2/3$ papers, and vice versa.

As we agreed before, we use notation $(G, \mathbf{w})$ for the graph G to denote that it is edge-weighted by $\mathbf{w}$.

Definition 3.4 *Given an edge-weighted graph $(G, \mathbf{w})$ and a vertex $x \in V(G)$, we define the* fractional-degree *of x in (G,w) as*

$$\mathbf{deg}_{G,\mathbf{w}}(x) = \sum_{e \in E_G(x)} \mathbf{w}(e)$$

In our co-authorship context, the degree $\mathbf{deg}_{G,\mathbf{w}}(x)$ of an author x is the collective effort of author x for all the papers she/he wrote. For instance, in Figure 1, the degree author a_4 is the sum of all the weights of the edges that are incident to it, i.e., $1/3 + 2/3 + 4/3 = 13/6$.

We say that $(H, \mathbf{w}_H)$ is an edge-weighted subgraph *of $(G, \mathbf{w})$ if H is a subgraph of G and $\mathbf{w}_H$ is the restriction of $\mathbf{w}$ on $E(H)$. Given any such subgraph $(H, \mathbf{w}_H)$ of $(G, \mathbf{w})$, we define*

$$\delta(H, \mathbf{w}_H) = \min\{\mathbf{deg}_{H,\mathbf{w}_H}(x) \mid x \in V(H)\}.$$

For example, if $(G, \mathbf{w})$ is the edge-weighted graph in Figure 1, then $\delta(G, \mathbf{w}) = \mathbf{deg}_{G,\mathbf{w}}(a_2) = 7/6$. If H is the subgraph of G containing all edges that are incident to the vertices a_1, a_2, and a_4, then $\delta(H, \mathbf{w}_H) = \mathbf{deg}_{H,\mathbf{w}_H}(a_1) = 2/3$.

Definition 3.5 *Let $(G, \mathbf{w})$ be an edge-weighted graph. The* fractional-degeneracy *of $(G, \mathbf{w})$ is defined as follows:*

$$\delta^*(G, \mathbf{w}) = \max\{\delta(H, \mathbf{w}_H) \mid (H, \mathbf{w}_H) \text{ is a non-empty edge-weighted subgraph of } (G, \mathbf{w})\}.$$

Let $k \in \mathbb{Q}^+$. Then the k-core of $(G, \mathbf{w})$ is the maximun-size edge-weighted subgraph $(H, \mathbf{w}_H)$ of $(G, \mathbf{w})$ where $\delta(H, \mathbf{w}_H) \geq k$.

The *Trim* procedure can also compute k-cores where k is a rational number. The only modification is that $\mathbf{deg}_F(x) \leq k$ should be replaced by $\mathbf{deg}_{F,\mathbf{w}_F}(x) \leq k$, i.e., we check the fractional degree of x in the edge-weighted graph $(F, \mathbf{w}_F)$, where $\mathbf{w}_F$ is the restriction of $\mathbf{w}$ to the edges of F. In fact, we have to be more careful in the definition of the fractional analogue of the core sequence, as it now should be indexed by rational numbers. For his, consider the infinite sequence $\mathcal{G} = G_{h_0}, G_{h_1}, \ldots$, recursively defined as follows: $G_{h_0} = G, h_0 = 0$, and for $i > 0$, $G_{h_i} = Trim(G_{h_{i-1}}, h_{i-1})$ where $h_i = \delta(G_i, \mathbf{w}_{G_i})$. Then, the *fractional core sequence* of an edge-weighted graph $(G, \mathbf{w})$ is the prefix of $\mathcal{G}$ that contains all non-empty graphs of $\mathcal{G}$ and is denoted by $\mathcal{G}(G, \mathbf{w})$. The *size* of a *fractional core sequence* is the number of its terms minus one. Notice that the size l of the fractional core sequence of an edge-weighted graph $(G, \mathbf{w})$ can never exceed the size of G. We finally call the sequence $h_1, \ldots, h_l$ *fractional index sequence* of $(G, \mathbf{w})$.

The *fractional core index* of a vertex of an edge-weighted graph $(G, \mathbf{w})$ is the maximum rational number k for which v belongs in the k-core of G. As in the unweighted case, the *fractional core index* definition can be naturally extended to sets instead of vertices. Again the fractional core index of a set of vertices is the minimum fractional core index of its members.

As an example of the above definitions, the edge-weighted graph $(H_G, \mathbf{w})$ depicted in Figure 1, has fractional degeneracy $\frac{7}{6}$, i.e. $\delta(H_G, \mathbf{w}) = \delta^*(H_G, \mathbf{w})$. Indeed if we apply $Trim(H_G, \frac{7}{3})$ then the first vertex to be removed is a_2. This removal drops the fractional degrees of a_1, a_3, and a_4 below $\frac{7}{3}$. Therefore, they are also removed and, for the same reason, the remaining vertex a_5 is removed as well. Therefore, G_1 is the empty graph, the fractional core sequence contains only graph $G_0 = G$, and the length of the fractional index sequence of $(H_G, \mathbf{w})$ is 0. We should mention that a less trivial example would be too complicated to present in a figure and even more complicated to be processed by the reader.

At this point we stress that, as a graph-theoretic notion, fractional cores are defined on bipartite graphs, encoding relations between two sets representing different entities (in our case, papers and authors). Equivalently, we can define fractional cores in hypergraphs by considering the fractional cores of their (bipartite) incident graphs. In this case, the hypergraph corresponding to the graph G would contain the authors

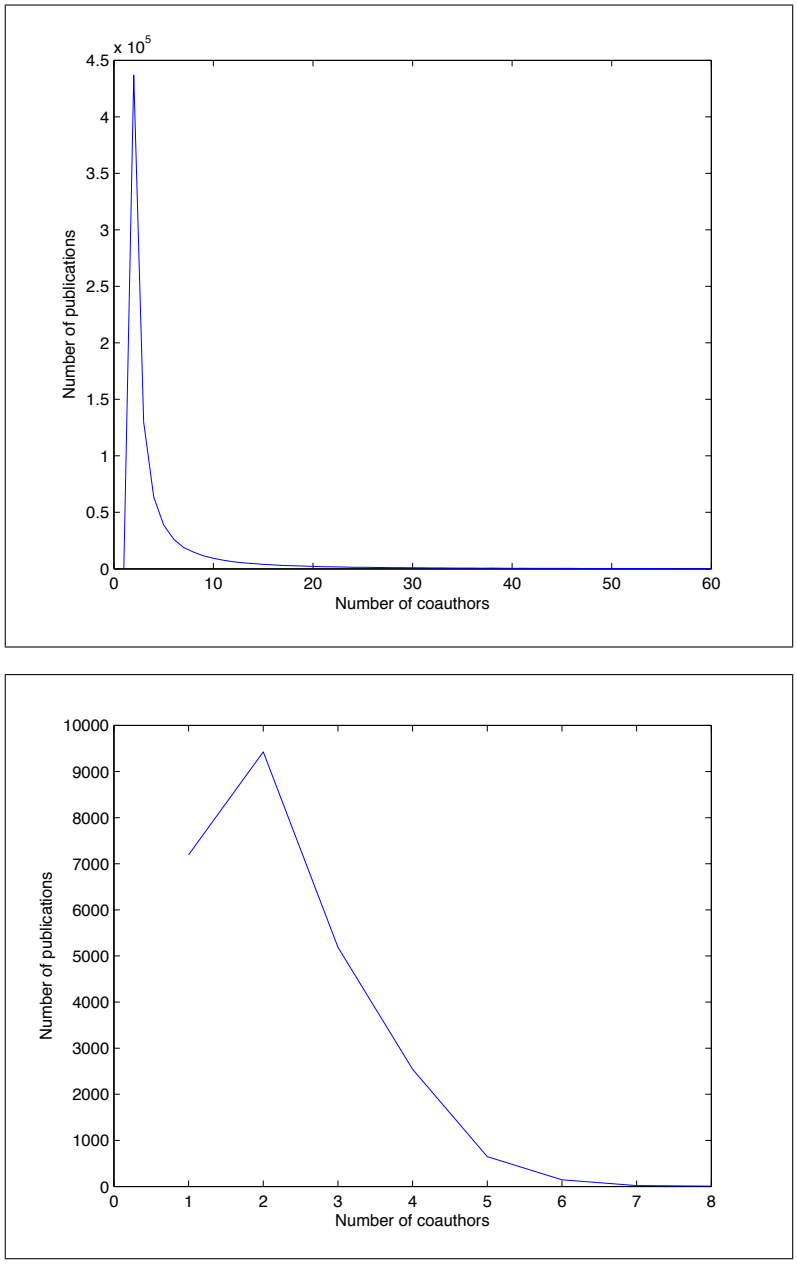

Figure 2. Distribution of number of publications versus cardinality of co-author set for DBLP and ARXIV.

as vertices and the papers as hyperedges. In this paper we chose to avoid hypergraph notation and, for simplicity, we adopted the definition that uses bipartite graphs.

4. Experimental evaluation of the DBLP and ARXIV.hep-th datasets

In this section we present the application of the above defined framework on the bipartite graphs corresponding to the DBLP dataset, concerning publications in computer science, and the ARXIV on High Energy Physics - Theory (ARXIV.hep-th) dataset. From now

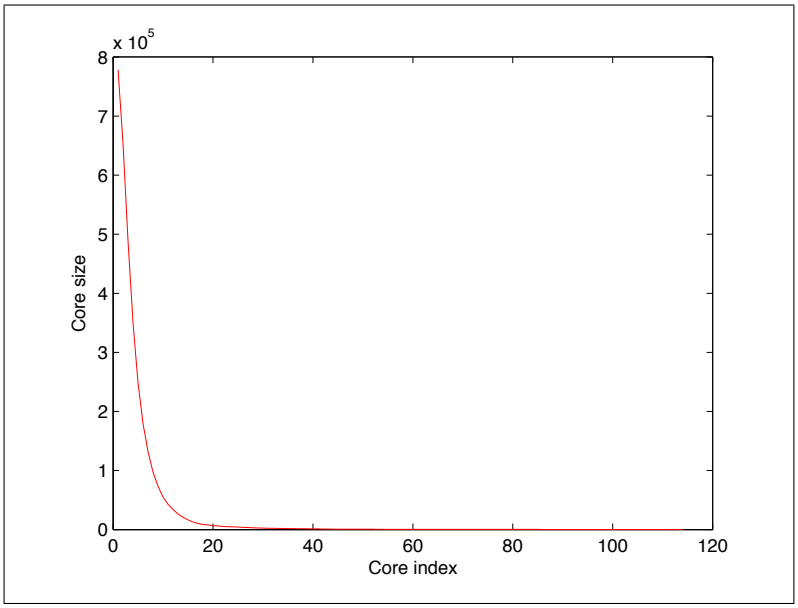

Figure 3. Distribution of the core sizes vs core indices in H_{DBLP}.

on, for notational convenience, we use ARXIV as an abbreviation of ARXIV.hep-th. Our aim is to detect, in each dataset, the sets of authors that correspond to the most coherent community in terms of co-authorship collaboration.

4.1. Data set description and preprocessing

The DBLP dataset is freely available in XML format at

$$\texttt{http} : //\texttt{dblp.uni-trier.de/xml/}$$

and the ARXIV dataset on High Energy Physics Theory is available in simple text format at:

$$\texttt{http} : //\texttt{snap.stanford.edu/data/ca-HepTh.html}$$

We extracted, from these datasets, the bipartite graphs DBLP and ARXIV. In the current snapshot, DBLP has 2208512 papers while ARXIV has 25170 papers. Among them, 817 of the papers in DBLP have only one author, while the same holds for 7196 of the papers of ARXIV. Also, DBLP has 825761 authors and ARXIV has approximately 8862 authors. In total, DBLP has 4446765 edges and ARXIV has 56065 edges.

In Figure 2, one can see the distribution of the number of co-authors per publication in the DBLP graph and the ARXIV graph. It is clear the vast majority of the papers are authored by few authors. However, there are some extremities where one specific paper in DBLP has 114 authors! On the other side all papers in ARXIV have at most 8 co-authors.

We computed, as described in Subsections 3.2 and 3.3, the unweighted graphs H_{DBLP} and H_{ARXIV} and their edge-weighted versions $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$. Clearly,

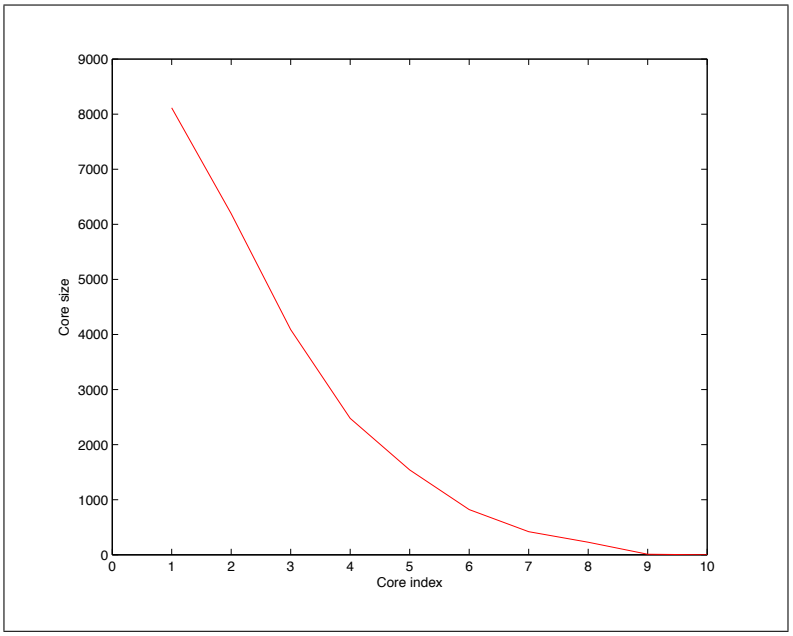

Figure 4. Distribution of the core sizes vs core indices in H_{ARXIV}.

H_{DBLP}	**Name of author**	**Index**
	Serge Abiteboul	28
	Christos Faloutsos	28
	Gerhard Weikum	22
	Christos H. Papadimitriou	17
	Paul Erdős	16
	Andrew Tanenbaum	48
H_{ARXIV}		
	Mirjam Cvetic	9
	Riccardo D'Auria	8
	Christoph Schweigert	7
	John Ellis	6
	Jürgen Fuchs	6
	Dimitris Nanopoulos	6

Table 1. Ranks of selected authors in H_{DBLP} and H_{ARXIV}.

single-author papers will not create any edge between authors and all isolated vertices in H_{DBLP} and H_{ARXIV} correspond to authors that have written only single-author papers.

4.2. k-cores in co-authorship graphs

We applied the *Trim* procedure to find the core sequences of the graphs H_{DBLP} and H_{ARXIV}. In this computation, we took into account all the papers regardless of the number of the authors each may have. In Figure 4, we can see the distribution of cores sizes for each graph.

In Table 1, we present a ranking of a few selected authors for both datasets. As

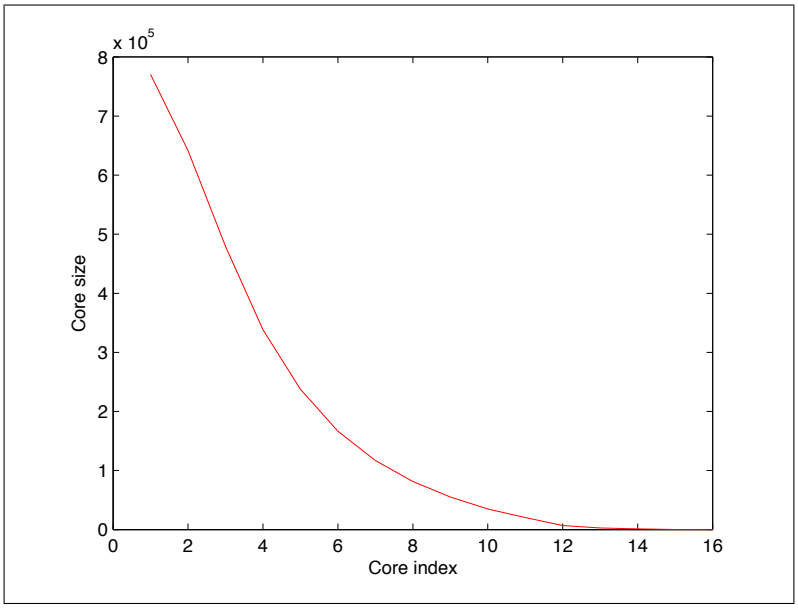

Figure 5. Distribution of the core sizes vs core indices in H^*_{DBLP}.

mentioned before, one paper with a large number of co-authors can "push" authors with otherwise low co-authorship to the densest k-core. For example, in DBLP, at $k = 113$ we have 114 authors all of which have participated in the same publication and some of them do not appear anywhere else on the dataset. Actually, the results on the H_{DBLP} graphs are apparently quite biased, i.e. a maximum-index 113-core exists in H_{DBLP} because of the existence of a single paper regardless of their other publication activity. In graph theoretic terms H_{DBLP} this core is a clique of 114 vertices that is created because of the existence of a vertex in DBLP of degree 113. However, this does not hold for the case of the – smaller in size – graph H_{ARXIV} where the maximum number of authors in a paper is 8. The densest core in H_{ARXIV} is the 9-core and is a clique on 10 vertices. The members of this core are presented in the lower part of Table 2. It is interesting to note that the edges of this clique are formed by many different papers. In fact there are at least 118 papers in ARXIV that have been co-authored by at least two of the members of the 9-core of H_{ARXIV}.

The biased situation that we detected in H_{DBLP} motivated us to consider filtering out papers with excessively high number of co-authors. In this case, we computed a filtered version of H_{DBLP}, by taking into account only the the papers whose number of co-authors is within the 99% of the corresponding distribution shown in Figure 2. This excludes from DBLP papers with more than 15 co-authors. We call this version of the graph H_{DBLP} *filtered* and we denote it by H^*_{DBLP}.

We applied the *Trim* procedure to find the core sequence of the graph H^*_{DBLP}. The distribution of the resulting core sizes appears in Figure 5. In the filtered case, the densest core of H^*_{DBLP} has index 15 and has a size of 76 authors. These authors appear in the upper part of Table 2.

H^*_{DBLP}

Pankaj K. Agarwal	Hee-Kap Ahn	Oswin Aichholzer	Greg Aloupis
Helmut Alt	Esther M. Arkin	Boris Aronov	Tetsuo Asano
Mark de Berg	Therese C. Biedl	Prosenjit Bose	David Bremner
Hervé Brönnimann	Sergio Cabello	Timothy M. Chan	Bernard Chazelle
Otfried Cheong	Sébastien Collette	Mirela Damian	Erik D. Demaine
Martin L. Demaine	Olivier Devillers	Vida Dujmovic	Herbert Edelsbrunner
Alon Efrat	David Eppstein	Jeff Erickson	Hazel Everett
Sándor P. Fekete	Joachim Gudmundsson	Leonidas J. Guibas	Dan Halperin
Sariel Har-Peled	John Hershberger	Ferran Hurtado	John Iacono
Christian Knauer	Danny Krizanc	Stefan Langerman	Sylvain Lazard
Giuseppe Liotta	Anna Lubiw	Rolf Klein Mark	Jirí Matousek
Kurt Mehlhorn	Henk Meijer	Joseph S. B. Mitchell	Pat Morin
Joseph O'Rourke	Mark H. Overmars	Belén Palop	Richard Pollack
Suneeta Ramaswami	David Rappaport	Günter Rote	Vera Sacristan
Otfried Schwarzkopf	Raimund Seidel	Micha Sharir	Thomas C. Shermer
Michiel H. M. Smid	Jack Snoeyink	Michael A. Soss	Diane L. Souvaine
Bettina Speckmann	Ileana Streinu	Subhash Suri	Perouz Taslakian
Godfried T. Toussaint	Marc J. van Kreveld	Jorge Urrutia	Sue Whitesides
David R. Wood	Stefanie Wuhrer	Chee-Keng Yap	Emo Welzl

H_{ARXIV}

Mirjam Cvetič	Michael J. Duf	P Hoxha	R Martinez-Acosta
James T. Liu	Hong Lu	Jian-Xin Lu	Christopher N. Pope
Hisham Sati	Tuan A. Tran		

Table 2. Authors of the 15-core of H^*_{DBLP} (up) and the 9-core of H_{ARXIV} (down).

Name of author	**Index**
Serge Abiteboul	14
Paul Erdős	14
Christos Faloutsos	14
Christos H. Papadimitriou	14
Gerhard Weikum	14
Andrew Tanenbaum	12

Table 3. Ranks of selected authors in H^*_{DBLP}.

As expected, in the filtered graph H_{DBLP}, several of the authors "move down" in cores of smaller index. The new indices for the selected sets of authors of Table 1 for DBLP are now depicted in Table 3. As we can see, in the case of H_{DBLP}, the authors of Table 3 get now accumulated in the second densest core, i.e the 14-core.

It is interesting that for some authors of DBLP, such as Andrew Tanenbaum, the core index in the filtered case is much lower (12) that in the unfiltered one (48). Apparently, this happens due to his participation in multi-author papers that were

DBLP															
34.0	**34.3**	35.2	36.4	37	37.3	38.8	42.7	44.2	47.8	48.4	53.8	55.3	64.6	77.8	149.2
42	**39**	35	31	29	25	22	20	18	16	13	11	8	6	4	2
8	7	7	4	4	3	3	3	3	3	3	3	2	2	2	2
ARXIV															
10.4	10.5	10.6	10.7	11.0	11.4	11.5	**12.0**	13.1	**13.4**	13.7	14.9	16.0	21.7	24.5	34.9
51	50	36	35	33	26	23	**21**	16	**14**	11	9	6	5	4	2
37	36	20	19	19	16	16	**14**	6	6	6	6	6	5	2	2

Table 4. Data of the last 16 graphs of the fractional core sequence of $(H_{\mathsf{DBLP}}, \mathbf{w})$ (up) and $(H_{\mathsf{ARXIV}}, \mathbf{w})$ (down). For each dataset, the first line depicts h_i, the second line contains the size of the h_i-core and the third one contains the size of the biggest connected component of the h_i-core.

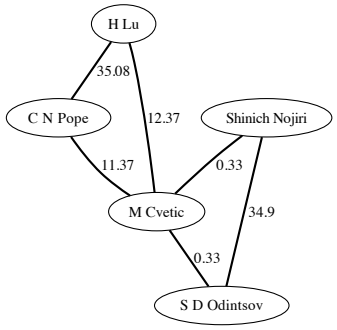

Figure 6. The 27.1-core of $(H_{\mathsf{ARXIV}}, \mathbf{w})$.

filtered out.

4.3. Fractional cores on the weights graph

Here we articulate the need for assigning weights to the edges of the previously defined co-authorship graphs. Assume that two authors x, y have co-authored several papers and therefore they are connected by an edge $e = \{x, y\}$. This co-authorship relation represents a strong collaboration among the two that escapes the unweighted setting of the previous section. This collaborative effort is apparently larger as the number of co-authored papers increases. On the other hand, the effort to author a paper is naturally divided among all the co-authors (we assume in equal parts). This justifies the definition in Section 3.3 of an edge-weighted co-authorship graph where the contribution of each author is now fractional.

In the fractional case, we do not need to apply any filtering of papers with a huge number of authors, as they are now filtered indirectly because the weight they contribute to their authors is tiny. Recall that the weight $\mathbf{w}(e)$ assigned to each edge is proportional to the number of papers they have co-authored and inversely proportional to the number of co-authors per co-authored paper. Thus $\mathbf{w}(e)$ represents the "essential amount" of collaboration among authors x, y in terms of the effort committed for common publications (which is normalized in each case by the number of contributing co-authors). This implies that the best fractional k-core communities contain authors that are intensively co-authoring with others and, while the number of co-authors is not

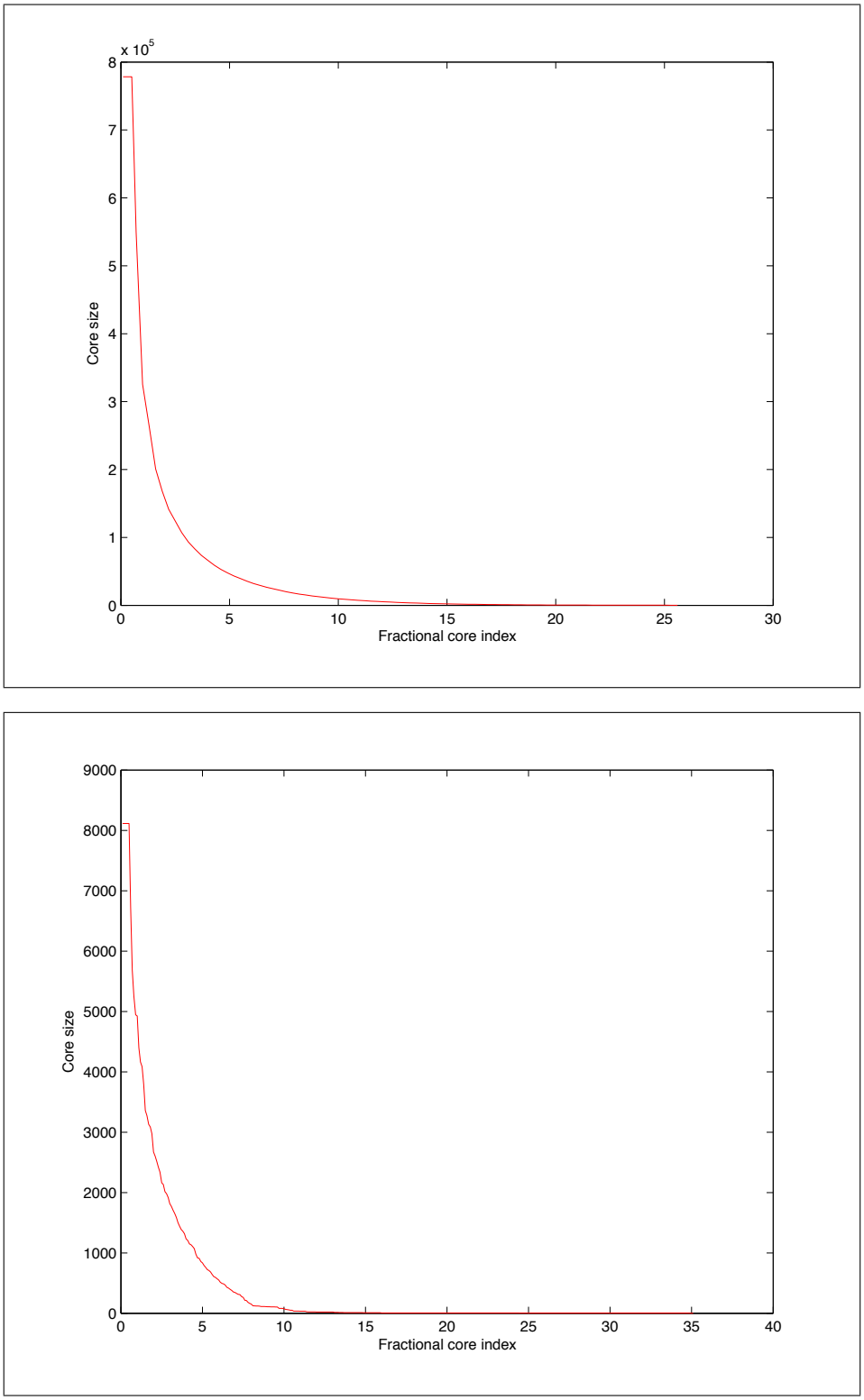

Figure 7. Distribution of the fractional core sizes vs core indices in the edge-weighted co-authorship graph of DBLP (up) and ARXIV (down).

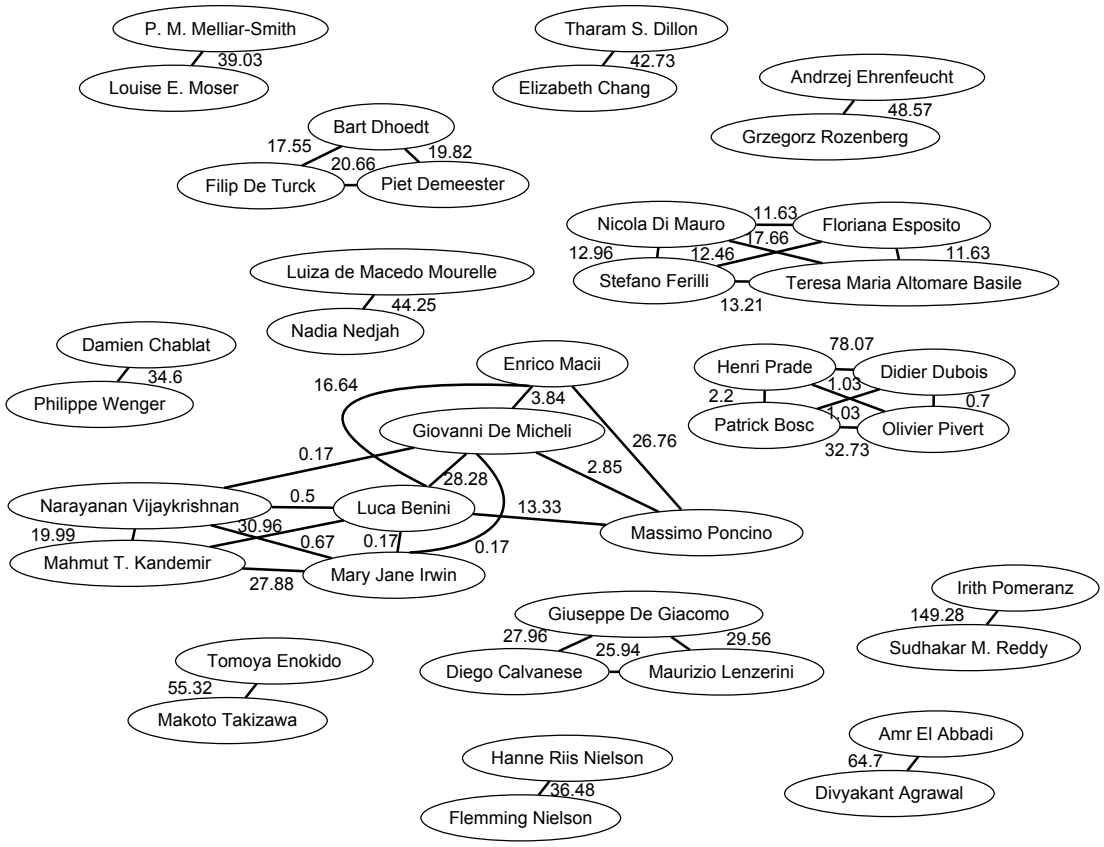

Figure 8. The 34.30-core of $(H_{\mathsf{DBLP}}, \mathbf{w})$

high, it follows that the share of collaborative effort is high.

In Figure 7, we can see the size distribution of the graphs in the fractional core sequence of $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$, i.e. the edge-weighted co-authorship graph of DBLP and ARXIV respectively. For both $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$, the behavior is of similar flavor in terms of the relation of the h_i-core size and h_i. The fractional index sequence of $(H_{\mathsf{DBLP}}, \mathbf{w})$ contains a big number of rational numbers that becomes "sparsest" as it increases, i.e., the differences between two consecutive elements is increasing, especially in the end. The 16 last terms of the fractional index sequence of $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$ are depicted in Table 4.

4.4. Rank vs size

For $(H_{\mathsf{DBLP}}, \mathbf{w})$, the densest fractional core has index 149.2 and contains only two authors (Sudhakar M. Reddy, Irith Pomeranz) whose publication record indeed verifies the claims as they have co-authored 373 papers, 256 of which as the only authors! The second densest core of $(H_{\mathsf{DBLP}}, \mathbf{w})$ is the 77.8-core that includes the additional authors: Henri Prade, Didier Dubois whose intense collaboration is verified by the number of co-authored papers (223 according to DBLP). In other words, the 77.8-core of H_{DBLP} consists of just two isolated edges. This trend continues for some of the next members

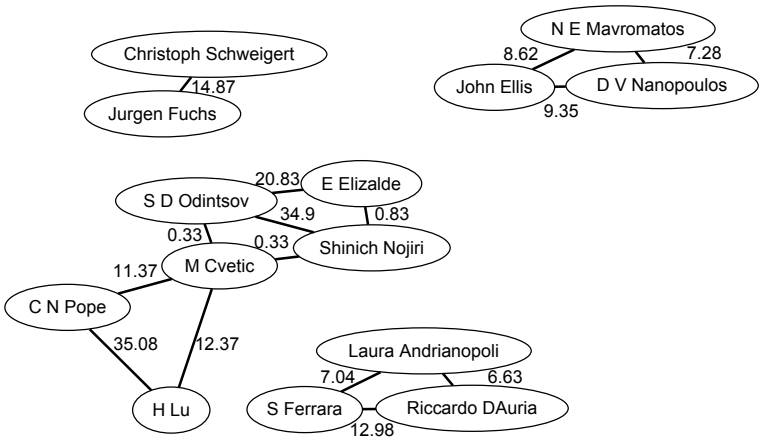

Figure 9. The 13.40-core of $(H_{\mathsf{ARXIV}}, \mathbf{w})$

of the fractional core sequence until the cores become greater and thus more complex.

In the case of $(H_{\mathsf{ARXIV}}, \mathbf{w})$, similar behavior is observed for the densest cores. However now the cores swiftly develop large connected components. The densest $(H_{\mathsf{ARXIV}}, \mathbf{w})$ core, the 34.9-core, contains only two authors: H. Lu and C. N. Pope that have co-authored 114 papers. The second densest 24.5-core contains two more authors: Shinich Nojiri and Sergey D. Odintsov who co-authored 76 papers. Interestingly, this set of authors becomes connected in the next 27.1-core because of the insertion of Mirjam Cvetic in it who has published papers with all aforementioned authors. The 21.7-core of $(H_{\mathsf{ARXIV}}, \mathbf{w})$ is depicted in Figure 6.

To amortize the effect of having tiny dense cores or dense cores of small connected components, we introduce two criteriatwo criteria to focus on dense cores:

- SVR (Size Versus Rank) Criterion: we discard from the core sequence of H_G all G_{h_i} for which $h_i > |V(G_{h_i})|$, i.e., we do not consider the cores whose size is less than their index.
- GCVR (Giant Component Versus Rank) Criterion: we discard from the core sequence of H_G all G_{h_i} for which $h_i > g(G_{h_i})$, i.e., we do not consider the cores for which the size of their giant component is less than their index.

Both above criteria are balancing the high index with some quantity criterion on the number of authors that generate it. SVR asks that the essential degree of effort of each author (i.e. the fractional degree of each vertex) is bigger than the total number of authors in the core with whom this effort has been shared. Clearly, GCVR is at least as strict as SVR and reflects the fact that, as cores grow in size, most of their authors are accumulated in on the giant component (see Section 5). The application of GCVR on $(H_{\mathsf{DBLP}}, \mathbf{w})$ considers the 34.3-core (depicted in Figure 8): it has 39 authors while the next 35.2-core has 35 authors. The same criterion applied to $(H_{\mathsf{ARXIV}}, \mathbf{w})$ considers the 13.4-core that has 14 authors (depicted in Figure 9). Notice that in both Figures 8

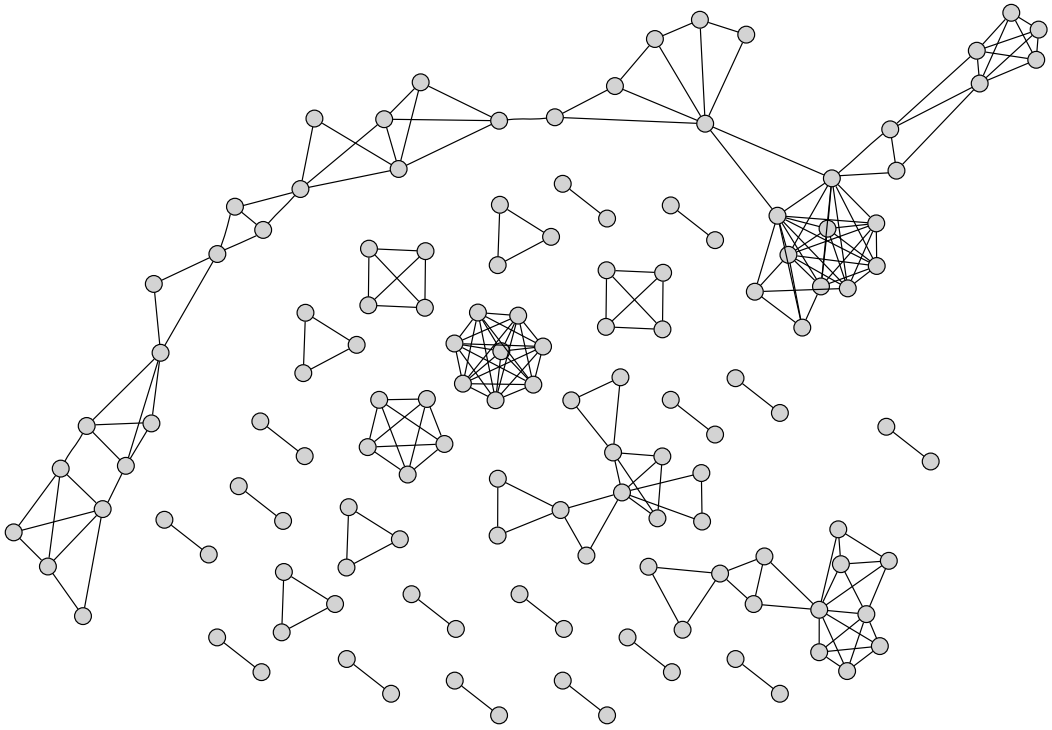

Figure 10. The 27.70-core of $(H_{\mathsf{DBLP}}, \mathbf{w})$

and 9, the graphs are still quite fragmented and, at the same moment, already big enough to reveal several collaboration communities.

Our next step is to apply the GCVR criterion on H_{DBLP}. In this case, the biggest k in the fractional index sequence of H_{DBLP} for which the giant component of the k-core is bigger than k is 27.7. Indeed, the 27.7-core has 132 authors and its giant component has 42 authors, while the next index is 28.0 and the 28.0-core has size 122 and its giant component has 23 authors. The 27.7-core is depicted in Figure 10 (as it has 122 vertices, we do not include the names of the authors).

The application of the GCVR criterion on H_{ARXIV} implies that the the 12-core, that has 21 vertices, is the last one whose giant component has more vertices, that is 12 than its index. Indeed, the next index is 13.1 and the 13.1-core has 16 authors and, among them, 6 are in its giant component. The 12-core of H_{ARXIV} is depicted in Figure 11.

4.5. Hop-1 lists

In Table 5 we depict the index of the previous sample of selected authors of both DBLP and ARXIV, based on the fractional cores computation. It is interesting that indices are different in this case due to the weighting scheme that favors not just a big number of publications but also repetitive co-authorship with limited number of co-authors. In this case, intensive collaboration with certain co-authors over a long series of publications

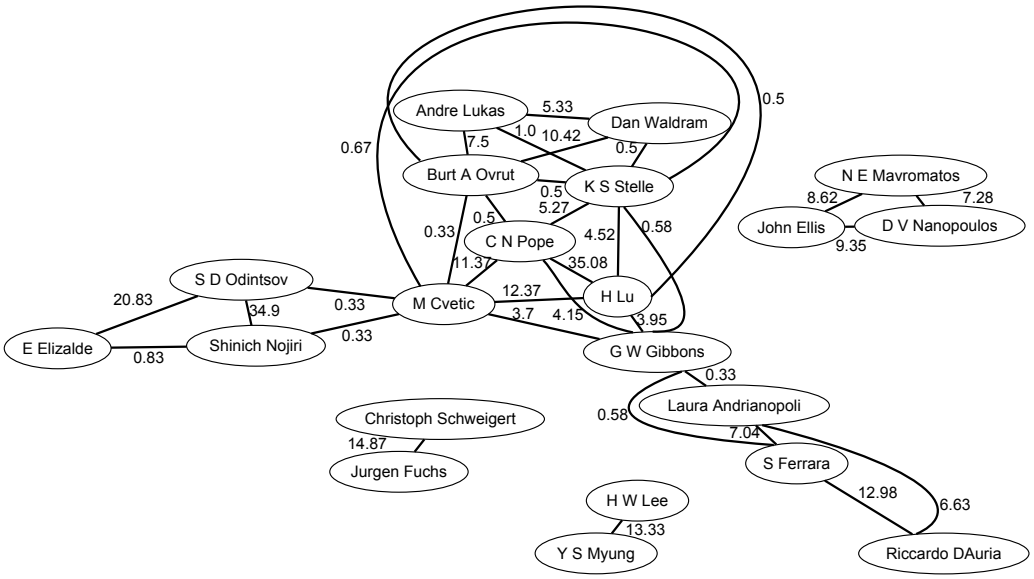

Figure 11. The 12-core of $(H_{\mathsf{ARXIV}}, \mathbf{w})$

	Name of author	**Index**
$(H_{\mathsf{DBLP}}, \mathbf{w})$		
	Christos H. Papadimitriou	20.8
	Serge Abiteboul	20.5
	Christos Faloutsos	18.7
	Gerhard Weikum	16.3
	Paul Erdős	13.9
	Andrew Tanenbaum	13.0
$(H_{\mathsf{ARXIV}}, \mathbf{w})$		
	Mirjam Cvetic	21.7
	John Ellis	14.9
	Dimitris Nanopoulos	14.9
	Christoph Schweigert	13.7
	Riccardo D'Auria	13.1

Table 5. Ranks of selected authors in $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$.

increases the mutual edge weights and thus the indices in the fractional k-cores.

Assuming an author x in H_{DBLP} it should be stressed that his/her best hop-1 co-authorship k-core (i.e. immediate co-authors) are those that have at least k co-authors in the same core.

In Table 6, we see the relevant data for fractional cores for a selection of well known and seminal authors from DBLP representing their degree of collaboration with their co-authors. C. H. Papadimitriou has a top score in this measure (20.8) while having a very small but cohesive community of co-authors, with the prominent example of Michalis Yannakakis contributing an awesome weight (19.62) to the vertex fractional degree of

Author	Fractional Rank	Size
C.H. Papadimirtiou	20.80	417
Michalis Yannakakis (19.62)	Erik D. Demaine (0.14)	Georg Gottlob (1.00)
Gerhard Weikum	16.30	1506
Hans-Jörg Schek (7.43)	Surajit Chaudhuri (5.05)	Yuri Breitbart (1.49)
Gautam Das (0.70)	Jeffrey F. Naughton (0.57)	Divesh Srivastava (0.53)
DanSuciu (0.50)	Rakesh Agrawal (0.48)	Gustavo Alonso (0.43)
Raghu Ramakrishnan (0.41)	Catriel Beeri (0.33)	Michael Backes (0.33)
Serge Abiteboul (0.33)	Divyakant Agrawal (0.29)	Amr El Abbadi (0.29)
Stefano Ceri (0.275)	Yannis E. Ioannidis (0.23)	Henry F. Korth (0.23)
S. Sudarshan (0.20)	Jennifer Widom (0.19)	David J. DeWitt (0.19)
Abraham Silberschatz (0.17)	David Maier (0.16)	Krithi Ramamritham (0.15)
Hector Garcia-Molina (0.14)	Christos Faloutsos (0.13)	Victor Vianu (0.13)
Edward A. Fox (0.09)	Beng Chin Ooi (0.08)	Richard Snodgrass (0.07)
Jeffrey D. Ullman (0.07)	Timos K. Sellis (0.07)	Umeshwar Dayal (0.17)
Michael J. Carey (0.14)		
Andrew Tanenbaum	13.0	4016
M. Frans Kaashoek (7.00)	Robbert van Renesse (5.40)	Maarten van Steen (4.68)
Frances M. T. Brazier (0.98)	Anne-Marie Kermarre (0.25)	Howard Jay Siegel (0.13)
Michael S. Lew (0.02)		
Paul Erdős	13.9	2678
János Pach (2.53)	Boris Aronov (0.28)	Leonard J. Schulman (0.28)
Ronald L. Graham (1.83)	Fan R. K. Chung (1.74)	Zoltán Füredi (1.58)
Noga Alon (0.50)	Endre Szemerédi (1.40)	Vojtech Rödl (1.33)
Nathan Linial (1.0)	Miklós Ajtai (0.25)	János Komlós (0.25)
László Lovász (0.33)	Shlomo Moran (0.53)	Andreas Blass (0.33)
Michael E. Saks (0.33)	Richard Pollack (0.25)	Shmuel Zaks (0.20)

Table 6. Fractional indices and hop-1 list for selected authors from DBLP.

Papadimitriou. This implies that they have co-authored many papers together (46) out of which more than 30 are co-authored by the two of them only! On the other hand, G. Weikum has a much more distributed collaboration circle in terms of co-authors that almost uniformly (except the case of Scheck, that is 7.43) contribute to his vertex fractional degree. Finally, Andrew Tanenbaum with a vertex fractional degree 13.0 has a rather small collaboration community with main collaborators Maarten van Steen (contributing a weight 4.68) and Robbert van Renesse (5.4) while the rest is uniformly distributed to the others.

In Table 7, we see the respective data for selected authors from ARXIV. There we also see very well known names in the scientific area together with their closet collaborators. Actually in this case all authors indicated in Figure 7 are present both the 13.40-core an the 12-core of $(H_{\mathsf{ARXIV}}, \mathbf{w})$. Especially in the 13.40 they appear in different connected components. Observe that, in the 12-core, Mirjam Cvetic and Riccardo D'Auria appear in the same component while this is not the case in the higher rank 13.40 core. However, the "clusters" of John Ellis, and Christoph Schweigert are

Author	Fractional Rank	Size	
Mirjam Cvetic H Lu (12.36) S D Odintsov (0.33)	21.7 C N Pope (11.36)	5 Shinich Nojiri (0.33)	
John Ellis N E Mavromatos (8.61)	14.9 D V Nanopoulos 9.35	9	
Christoph Schweigert Jurgen Fuchs (14.86)	13.7	11	
Riccardo D'Auria S Ferrara (12.98)	13.1 Laura Andrianopoli (6.62)	16	

Table 7. Fractional indices and hop-1 list for selected authors from ARXIV.

already becoming disconnected in the 12-core. Observe also that S D Odintsov enters the hop-1 list of Mirjam Cvetic because of a link of relatively low weight, i.e., 0.33. However, S D Odintsov enters in the hop-1 list of Mirjam Cvetic because of his strong collaboration with Shinich Nojiri and E Elizaide (that, however is not in the hop-1 list of Mirjam Cvetic).

4.6. Community-focused rankings

In our final experiment, we focus on authors belonging to specific scientific communities and compare their rankings according to our fractional cores method against rankings determined using simpler measures of collaborativeness. More precisely, we extracted the names of programme committee members of SIGMOD, SIGIR, and SIGKDD for the years 2009, 2010, and 2011 to obtain subsets of the database, information retrieval, and data mining community, respectively. Most of the authors could be mapped automatically to their entries in DBLP using string matching; for some we had to perform a best-effort manual mapping (e.g., because of missing middle initials or nicknames in the programme committee lists); about a handful of authors could not be mapped with confidence and are thus missing from our rankings. For each community, we rank authors therein according to the following measures:

(a) *fractional index*

(b) *number of co-authors*

(c) *number of publications*

(d) *average number of co-authors per publication*

(e) *years active*

The resulting top-10 rankings are given in Table 8, Table 9, and Table 10. Note that for our fractional cores method, as before, an author's fractional index is determined on the entire DBLP co-authorship graph and not only based on collaborations with authors within the same scientific community. When looking at the top-10 rankings presented, we observe that across all communities rankings according to **(b)**, **(c)**, and

(a)	(b)	(c)
Amr El Abbadi	Wei Wang	Wei Wang
Divyakant Agrawal	Hans-Peter Kriegel	Christos Faloutsos
Christian S. Jensen	Christos Faloutsos	Michael Stonebraker
Richard T. Snodgrass	Divyakant Agrawal	Michael J. Carey
Sourav S. Bhowmick	Elke A. Rundensteiner	Wolfgang Nejdl
Beng Chin Ooi	Kian-Lee Tan	Stefano Ceri
Kian-Lee Tan	Amr El Abbadi	Christian S. Jensen
Pierangela Samarati	Christian S. Jensen	Raghu Ramakrishnan
Sabrina De Capitani di Vimercati	Ming-Syan Chen	Jian Pei
Mong-Li Lee	Richard T. Snodgrass	Beng Chin Ooi

(d)	(e)
Michael Stonebraker	Nesime Tatbul
David B. Lomet	Anastasia Ailamaki
Theo Härder	Laura M. Haas
Philip A. Bernstein	Mitch Cherniack
Hans-Peter Kriegel	John McPherson
Michael Hatzopoulos	Brian Cooper
Carlo Zaniolo	Daniel J. Abadi
Umeshwar Dayal	Jayavel Shanmugasundaram
Stefano Ceri	Tim Kraska
Meral Ozsoyoglu	Fatma Ozcan

Table 8. Database community ranking.

(e) are biased in favor of senior authors (e.g., Michael Stonebraker, W. Bruce Croft, and Jiawei Han) and overlap sometimes significantly. This is natural, given that authors who have been active longer, tend to have more publications, co-authored with different people at different points in time. The rankings according to **(e)**, the average number of co-authors per publication, contain for all three communities relatively junior alongside senior authors. However, it can also be seen that this is not a robust measure, bringing up authors who have published and collaborated modestly, but happen to have one publication with a large number of publications. Finally, the rankings according to **(a)**, our fractional cores method, seem less biased toward senior authors, bringing up a mix of prolific authors with long-lasting intensive collaborations between them (e.g., Amr El Abbadi and Divyakant Agrawal, Ophir Frieder and Abdur Chowdhury, Annalisa Appice and Donato Malerba).

5. Core Decomposition Forest

In this section we examine the relation between core structure of a graph and the connected components of its cores. We need first some definitions.

Definition 5.1 *Let $\mathcal{G} = G_0, G_1, \ldots, G_d$ be a sequence of graphs such that for each*

(a)	(b)	(c)
Ee-Peng Lim	Lei Zhang	Lei Zhang
Paolo Boldi	Jun Wang	Jun Wang
Jie Lu	Gerhard Weikum	Yi Zhang
Steven M. Beitzel	Hsinchun Chen	Tao Li
Abdur Chowdhury	Tao Li	Qiang Yang
Ophir Frieder	Wei-Ying Ma	Wei-Ying Ma
Juan M. Fernández-Luna	Qiang Yang	Jun Xu
Juan F. Huete	C. Lee Giles	Gerhard Weikum
Wei-Ying Ma	Lee Giles	Hsinchun Chen
Yong Yu	Ricardo A. Baeza-Yates	Yong Yu

(d)	(e)
Michael Lesk	Michael Taylor
Erich J. Neuhold	Gerald Benoit
Jun-ichi Tsujii	Yifen Huang
W. Bruce Croft	Claus-Peter Klas
Fredric C. Gey	Mark Greenwood
Donald H. Kraft	Yantao Zheng
Jaime G. Carbonell	Maria M. Nikolaidou
David Lewis	Jinhui Tang
William R. Hersh	Jayavel Shanmugasundaram
Nicholas J. Belkin	David Smith

Table 9. Information retrieval community ranking.

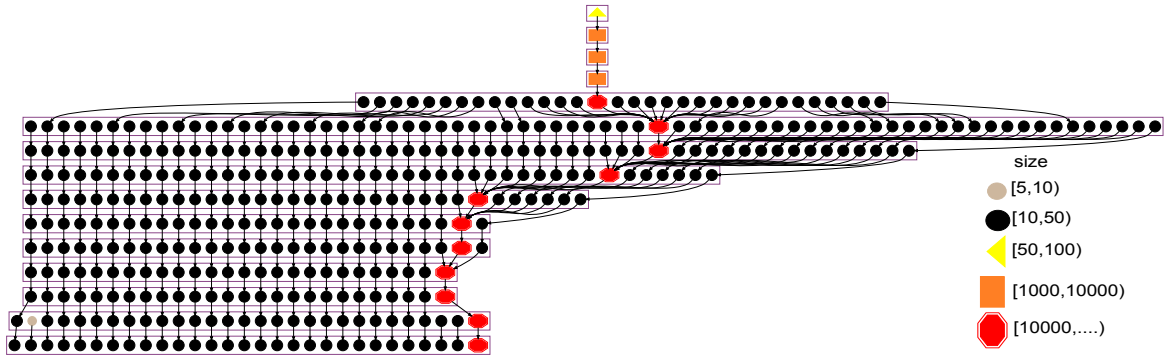

Figure 12. The Core Decomposition Forest of the core sequence of H^*_{DBLP}

i, j where $i \leq j$, G_i is a subgraph of G_j (we call such a sequence monotone*). The* Decomposition Forest *of a monotone graph sequence $\mathcal{G}$ is the graph $\mathbf{DF}(\mathcal{G})$ that is defined as follows. For each $i = 0, \ldots, d$ we denote the connected components of G_i by $G_i^1, \ldots, G_i^{m_i}$ and each such connected component is a vertex of $\mathbf{DF}(\mathcal{G})$ (we treat isomorphic graphs as different graphs). The pair $(G_i^j, G_{i'}^{j'})$ is a directed edge of $\mathbf{DF}(\mathcal{G})$ if $j' = j + 1$ and G_i^j contains $G_{i'}^{j'}$ as a subgraph.*

(a)	(b)	(c)
Floriana Esposito	Jiawei Han	Jiawei Han
Ee-Peng Lim	Christos Faloutsos	Christos Faloutsos
Annalisa Appice	Alok N. Choudhary	Gang Wang
Donato Malerba	Alberto Del Bimbo	Alok N. Choudhary
Charu C. Aggarwal	C. Lee Giles	C. Lee Giles
Alok N. Choudhary	Gonzalo Navarro	Jian Pei
Diane J. Cook	Ee-Peng Lim	Bing Liu
Alberto Del Bimbo	Jeffrey Xu Yu	Jeffrey Xu Yu
Jeffrey Xu Yu	Floriana Esposito	Aoying Zhou
Carlo Zaniolo	Carlo Zaniolo	Ee-Peng Lim

(d)	(e)
Andrzej Skowron	Jonathan Chang
Carlo Zaniolo	Jeffrey Yu
Christos Faloutsos	Byron J. Gao
Heikki Mannila	Jennifer Dy
Daniel Barbara	Edwin V. Bonilla
Dennis Shasha	Gui-Rong Xue
Alberto Del Bimbo	Ashok Savasere
Foto N. Afrati	Benoit Huet
David Poole	Jiangtao Ren
C. Lee Giles	Dou Shen

Table 10. Data mining community ranking.

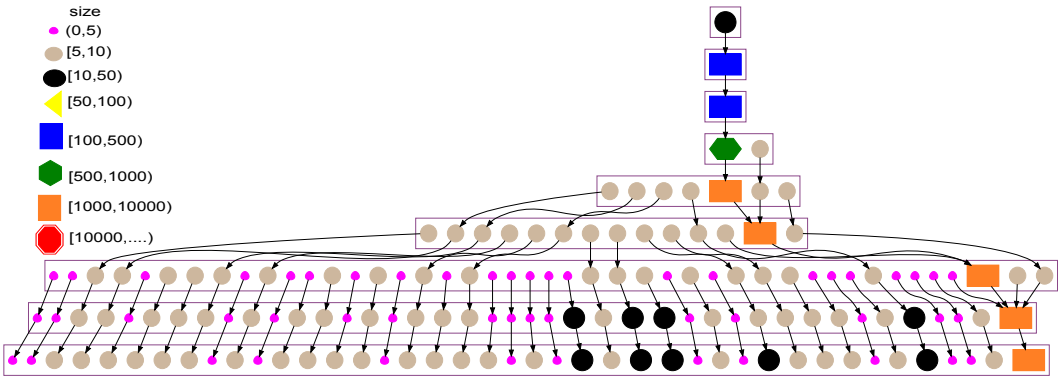

Figure 13. The Core Decomposition Forest of the core sequence of H_{ARXIV}

It is easy to verify that the directed graph defined above is a rooted forest. In fact, each of its components is a rooted tree where all its edges are directed away from the root and each root is a connected component of G_0. Given that the core sequence of G is monotone, we define the Core Decomposition Forest of a graph (edge-weighted or not) as the decomposition forest corresponding to its core sequence. The notion of the core decomposition forest appeared for the first time in [13] under the name *hierarchical degree core tree* and was used in order to visualize the connected components of several

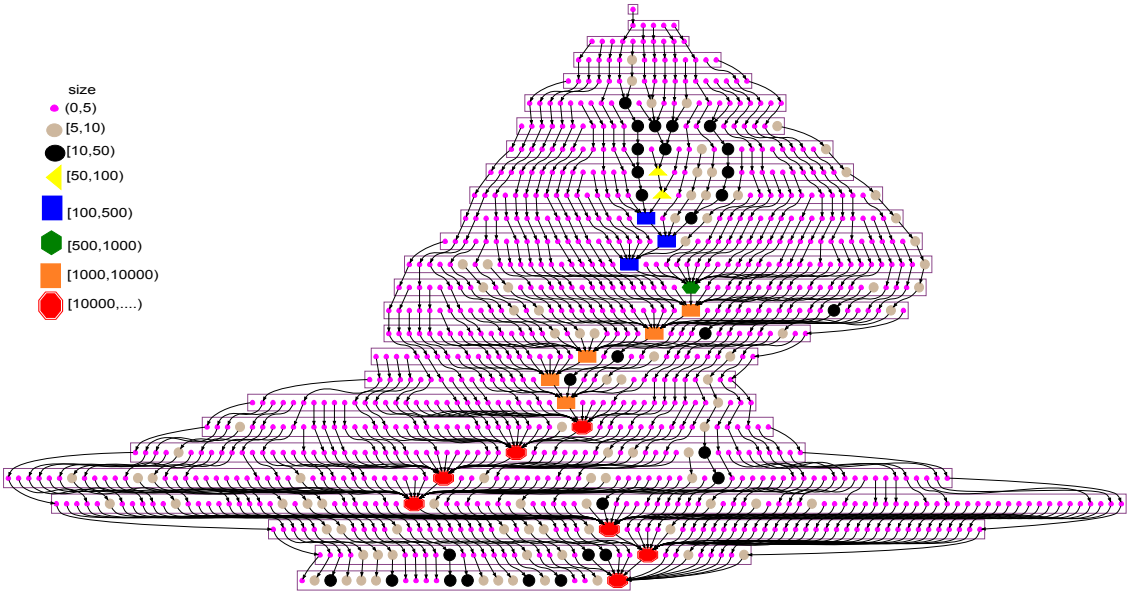

Figure 14. The Core Decomposition Forest of the core sequence of $(H_{\mathsf{DBLP}}, \mathbf{w})$

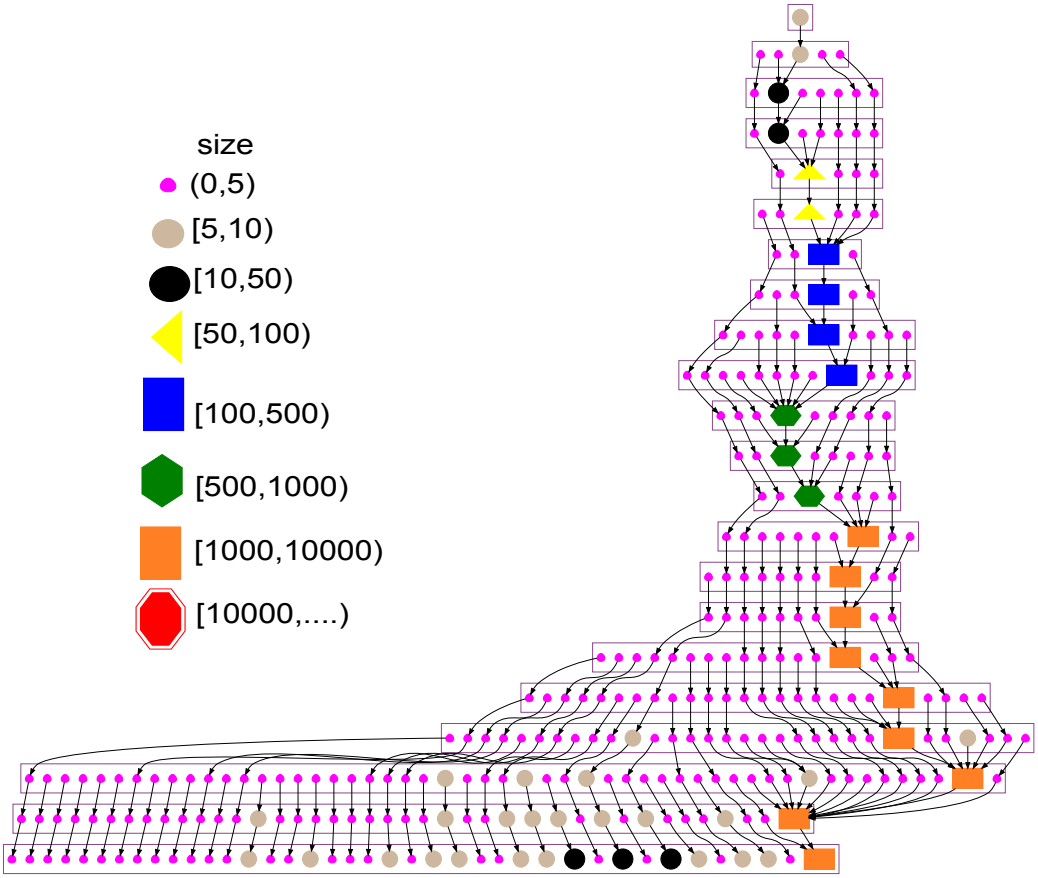

Figure 15. The Core Decomposition Forest of the core sequence of $(H_{\mathsf{ARXIV}}, \mathbf{w})$.

real-word graphs including the graph extracted by the common-author relation of the papers of the DBLP citation graph. As the graphs that we extract from DBLP and ARXIV are expressing relations between authors, the cores decomposition forests that we describe below are of radically different nature than the one extracted in [13].

In our study, we computed the Core Decomposition Forests $\mathbf{DF}(\mathcal{G}(H^*_{\mathsf{DBLP}}))$, $\mathbf{DF}(\mathcal{G}(H_{\mathsf{ARXIV}}))$, $\mathbf{DF}(\mathcal{G}(H_{\mathsf{DBLP}}), \mathbf{w})$, and $\mathbf{DF}(\mathcal{G}(H_{\mathsf{ARXIV}}), \mathbf{w})$. The results for the case of H^*_{DBLP} and H_{ARXIV} are depicted in Figures 12 and 13 respectively, while the results for $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$ are depicted in Figures 14 and 15 respectively. We stress that these figures depict only an approximation of these trees as their sizes are too big to fit in a visible way. To facilitate the visualization of the core decomposition forests we applied the following relaxations parameterized by α and n:

(1) suppress in $\mathcal{G}$ consecutive terms that are the same,

(2) consider only the members of the resulting sequence that are indexed by multiples of α, and

(3) in the core decomposition forest of the (fractional) core sequence remaining after relaxations (1) and (2), exclude all subtrees that do not have ancestors after the n-th core, of this sequence.

For the visualization of the core decomposition forest for H^*_{DBLP} and H_{ARXIV} we applied steps (1)–(3) for $\alpha = 1, n = 8$ and $\alpha = 1, n = 1$ respectively. For the visualization of the core decomposition forests for $(H_{\mathsf{DBLP}}, \mathbf{w})$ and $(H_{\mathsf{ARXIV}}, \mathbf{w})$, we only applied the relaxation steps (2) and (3) (relaxation step (1) is unnecessary on fractional sequences) for $\alpha = 5$, $n = 10$ and $\alpha = 5$ and $n = 8$ respectively. In each case the values of the parameter n and α have been chosen as to optimize the visualization of the corresponding datasets.

As we see in Figure 12, the H^*_{DBLP} dataset presents the following behavior in terms of connected components: There is clearly a giant component that evolves as k increases and survives until the last 15-core. It is interesting that many connected components survive until core index 11, thus the H^*_{DBLP} dataset is rather robust under degeneracy.

In Figure 13 we see the robustness of the cores of the ARXIV co-authorship graph under degeneracy. Again there is a giant component that evolves as k increases and survives until the last 9-core. It is interesting that many connected components survive until core index 5.

As for the edge-weighted graphs there is a remarkable behavior. In Figure 14 we see the evolution of the connected components of the $(H_{\mathsf{DBLP}}, \mathbf{w})$ graph. In this case the graph is much more robust as the steps of degeneracy are fractional while we see again a giant component that splits into other components that merge before they shrink again.

In Figure 15 the evolution of the connected components of the $(H_{\mathsf{ARXIV}}, \mathbf{w})$ graph is depicted. In this case the graph is much less robust as the number of connected components is swiftly shrinking and only a few - together with the giant component survive until the highest index fractional core.

6. Conclusions

Large graphs constitute an important element in current large scale information systems. Common cases of such graphs are the Web graph, social networks, citation graphs, CDRs (call data records) where nodes (featured with attributes - in some cases with a large cardinality) are connected to each other with edges representing a relation such as endorsement, recommendation, and/or friendship. Community detection and evaluation is an important task in graph mining. A variety of measures have been proposed to evaluate the quality of such communities. In this paper, we evaluated communities, using the k-core concept, as a means of evaluating their collective collaborative nature - a property not captured by individual node metrics or by other community evaluation metrics. Based on the k-core concept, which essentially measures the collaboration robustness of a community, we extended it to edge-weighted graphs, devising a novel concept of fractional k-cores on weighted graphs. We applied the (fractional) k-core approach on large real-world graphs – such as DBLP and report interesting results. Notice that further research in this direction could study different ways to weight edges and vertices of a collaboration graph. For example, as a continuation of our research, one might also weight vertices according to, for example, the H-Index of the corresponding author. This would reveal communities where the collaborative strength of an author would be would be mixed with his/her standing in research. We believe that the semantics of such type of weightings under the k-core methodology are worth to investigate.

The findings of the experiments of this paper, as well as an applet visualizing the hop-1 lists of each individual author, can be accessed online at

`http://www.graphdegeneracy.org/#kfcores`

Acknowledgements

All the authors are supported by the DIGITEO Chair grant LEVETONE in France.

Acknowledgements

[1] Santo Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.
[2] Paul Erdős. On the structure of linear graphs. *Israel J. Math.*, 1:156–160, 1963.
[3] George Szekeres and Herbert S. Wilf. An inequality for the chromatic number of a graph. *J. Combinatorial Theory*, 4:1–3, 1968.
[4] David W. Matula. A min–max theorem for graphs with application to graph coloring. *SIAM Reviews*, 10:481–482, 1968.
[5] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.
[6] José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the analysis of large scale Internet graphs. *CoRR*, cs/0511007, 2005.
[7] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. Medusa - new model of internet topology using k-shell decomposition, 2006.

[8] José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the visualization of large scale networks. *CoRR*, cs.NI/0504107, 2005.

[9] J. Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 41–50, Cambridge, MA, 2006. MIT Press.

[10] Vladimir Batagelj and Andrej Mrvar. Pajek— analysis and visualization of large networks. In Petra Mutzel, Michael Jünger, and Sebastian Leipert, editors, *Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, pages 8–11. Springer Berlin / Heidelberg, 2002.

[11] Stefan Wuchty and Eivind Almaas. Peeling the yeast protein network. *PROTEOMICS*, 5(2):444–449, 2005.

[12] Gary D. Bader and Christopher W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, pages –1–1, 2003.

[13] John Healy, Jeannette Janssen, Evangelos Milios, and William Aiello. Characterization of graphs using degree cores. In *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006*, volume LNCS-4936 of *Lecture Notes in Computer Science*. Springer Verlag, Banff, Canada, Nov. 30 - Dec. 1 2008.

[14] Christos Giatsidis, Dimitrios M. Thilikos, and Michalis Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *ASONAM*, pages 87–93. IEEE Computer Society, 2011.

[15] Mary McGlohon, Leman Akoglu, and Christos Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 524–532, New York, NY, USA, 2008. ACM.

[16] Stanley Wasserman and Katherine Faust. *Social Networks Analysis: Methods and Applications*. Cambridge: Cambridge University Press., 1994.

[17] Spiros Papadimitriou, Jimeng Sun, Christos Faloutsos, and Philip S. Yu. Hierarchical, parameter-free community discovery. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 170–187, Berlin, Heidelberg, 2008. Springer-Verlag.

[18] James Moody and Douglas R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):pp. 103–127, 2003.

[19] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 639–650, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[20] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization*, APPROX '00, pages 84–95, London, UK, 2000. Springer-Verlag.

[21] Mauro Sozio and Aristides Gionis. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 939–948, New York, NY, USA, 2010. ACM.

[22] Don R. Lick and Arthur T. White. k-degenerate graphs. *Canad. J. Math.*, 22:1082–1096, 1970.

[23] David W. Matula, George Marble, and Joel D. Isaacson. Graph coloring algorithms. In *Graph theory and computing*, pages 109–122. Academic Press, New York, 1972.

[24] Eugene C. Freuder. A sufficient condition for backtrack-free search. *J. Assoc. Comput. Mach.*, 29(1):24–32, 1982.

[25] Lefteris M. Kirousis and Dimitrios M. Thilikos. The linkage of a graph. *SIAM J. Comput.*, 25(3):626–647, 1996.

[26] Reinhard Diestel. *Graph theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Berlin, third edition, 2005.

[27] Boris Pittel, Joel Spencer, and Nicholas Wormald. Sudden emergence of a giant k-core in a random graph. *J. Combin. Theory Ser. B*, 67(1):111–151, 1996.

[28] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[29] Tomasz Łuczak. Size and connectivity of the k-core of a random graph. *Discrete Mathematics*, pages 61–68, 1991.

[30] Béla Bollobás. The evolution of sparse graphs. In *Graph theory and combinatorics (Cambridge, 1983)*, pages 35–57. Academic Press, London, 1984.

[31] Svante Janson and Malwina J. Łuczak. Asymptotic normality of the k-core in random graphs. *Ann. Appl. Probab.*, 18(3):1085–1137, 2008.

[32] Vladimir Batagelj and Matjaz Zaversnik. An o(m) algorithm for cores decomposition of networks. *CoRR*, cs.DS/0310049, 2003.

[33] Michael Baur, Marco Gaertler, Robert Görke, Marcus Krug, and Dorothea Wagner. Generating Graphs with Predefined k-Core Structure. In *Proceedings of the European Conference of Complex Systems (ECCS'07)*, October 2007.

[34] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. k-core organization of complex networks. *PHYS.REV.LETT.*, 96:040601, 2006.

[35] Vladimir Batagelj and Matjaz Zaversnik. Generalized cores. *CoRR*, cs.DS/0202039, 2002.