D-cores: measuring collaboration of directed graphs based on degeneracy

Christos Giatsidis, Dimitrios M. Thilikos, Michalis Vazirgiannis

Knowl Inf Syst (2013) 35:311–343
DOI 10.1007/s10115-012-0539-0

REGULAR PAPER

D-cores: measuring collaboration of directed graphs based on degeneracy

Christos Giatsidis · Dimitrios M. Thilikos · Michalis Vazirgiannis

Received: 12 March 2012 / Revised: 5 June 2012 / Accepted: 11 August 2012 /
Published online: 27 September 2012

Abstract Community detection and evaluation is an important task in graph mining. In many cases, a community is defined as a subgraph characterized by dense connections or interactions between its nodes. A variety of measures are proposed to evaluate different quality aspects of such communities—in most cases ignoring the directed nature of edges. In this paper, we introduce novel metrics for evaluating the collaborative nature of directed graphs—a property not captured by the single node metrics or by other established community evaluation metrics. In order to accomplish this objective, we capitalize on the concept of graph degeneracy and define a novel D-core framework, extending the classic graph-theoretic notion of k-cores for undirected graphs to directed ones. Based on the D-core, which essentially can be seen as a measure of the robustness of a community under degeneracy, we devise a wealth of novel metrics used to evaluate graph collaboration features of directed graphs. We applied the D-core approach on large synthetic and real-world graphs such as Wikipedia, DBLP, and ArXiv and report interesting results at the graph as well at the node level.

Keywords Graph mining · Community evaluation metrics · Degeneracy · Directed cores

1 Introduction

The Web, social network, and citation graphs form a context where the detection and evaluation of communities constitutes an important and challenging task. The research methods

C. Giatsidis · M. Vazirgiannis
LIX, École Polytechnique, Palaiseau, France

D. M. Thilikos
Department of Mathematics, National and Kapodistrian University of Athens,
Athens, Greece

M. Vazirgiannis (✉)
Department of Informatics, Athens University of Economics, Athens, Greece
e-mail: mvazirg@aueb.gr

M. Vazirgiannis
Télécom ParisTech, LTCI, Paris, France

in this area have mainly capitalized on the Hub/Authority concepts (see [39,43]) evaluating communities based on the centrality of nodes in terms of incoming/outgoing links. We claim that inherent mechanisms of community creation and evolution are not solely based on the Hub/Authority concepts. An important constituent of such a mechanism, generally neglected, is the community cohesion in terms of a dense distribution of in/outlinks within the community—as opposed to sparse connections across them. We are interested in quantifying the degree of cohesion of a community subgraph as a measure of collaboration among its members. Here, we have to stress the distinguishing feature of the graphs under concern in this paper: the directed nature of the edges – representing endorsement, recommendation, citation, and, in general, non-symmetric relationship among entities.

In order to study this collaboration aspect, we capitalize on the k-core concept—an established, computationally efficient technique for identifying dense graph areas with dense edge connectivity. A core is broadly defined as a maximum size subgraph of a graph that is coherent and dense in the sense that for every node in this subgraph, there are least k incident edges that are adjacent to vertices of the same subgraph (formal definitions follow in Sect. 2).

The objective of our study is to deal with and evaluate the "collaborative" behavior of communities (represented as D-cores of a directed graph) rather than dealing with authorities or hubs. This work follows up the work in [29] on evaluating collaboration in undirected graphs.

The paper contributions are the following:

- We vastly extend the existing theoretical framework on cores (mainly defined in [2,9,10,15,30]) to the case od directed graphs. Such graphs emerge naturally from social/citation networks and the Web. D-cores constitute dense directed subgraphs of the original one involving intensive and mutual collaboration in terms of directed links. Interestingly, all these notions induce a two-dimensional setting indicating qualitative differences from the directed case and are later employed and visualized in our experimentation.
- We define new structures and metrics for evaluating the collaborative nature of directed graphs. Such are the D-core matrix for a graph, its frontier, and a series of novel metrics to evaluate: (a) the robustness of the directed graph under degeneracy, as a metric of cohesiveness and hence the collaboration among the members of the graph under study and (b) the dominant patterns of the graph with respect to the inlink/outlink trade-off indicating macroscopic graph patterns related to whether the graph is extrovert or "selfish". A salient feature of our work is the low (in fact optimal) complexity for computing the D-core structures and the related structures and metrics.
- Extensive experimental evaluation: We conducted large-scale experiments in scale-free/preferential attachment synthetic graphs as well as real-world large-scale directed graphs: the (English) Wikipedia—2004 edition, the ArXiv, and DBLP citation graphs. We computed and explored the respective D-cores matrices, frontiers, and metrics, and we derived interesting results and observations both at the macroscopic (graph) and at the microscopic (node) level.

We claim that the D-core concept and the relevant structures and metrics that we define in this paper constitute a framework of tools for efficient and valid evaluation of cohesiveness and collaboration in directed networks.

We should also stress that the current paper is an extension of the paper appeared in [28] in the following aspects:

- We articulate the theoretical foundation of the proposed framework in a more principled manner, adopting valid terminology from related work, more specifically in [2,9,10,15,30]. We extend the set of metrics and structures proposed in [28] with directed

analogues of the notions of k-cores, k-core sequence, k-cells, and k-cell sequence. Moreover, we introduce the concept of a *Core Decomposition Forest* for the case of directed graphs, extending the similar concept for undirected graphs was introduced in [30] and used in [29].
- We investigate the behavior of the new concepts and metrics in the case of synthetic preferential attachment graphs—dominant in real-world cases. The study is extended to various parameters values in an attempt to fit the features of the real-worlds graphs. In order to achieve this, we developed a multiparametric graph generator.
- We extend the experimental evaluation with a. an alternative visualization of the differential behavior of the graphs under concern with regard to the pace of graph size reduction for both the inlinks and the outlinks aspect b. an exhaustive study and interpretation of the in-/out-degree distributions in the synthetic preferential attachment datasets and c. study of another citation dataset, ArXiv HEP-TH (high energy Physics Theory), featuring an alternative important representative scientific discipline.

2 Related work

A thorough review on community detection in graphs is offered by Fortunato [25].

In that work techniques, methods and datasets are presented for detecting communities in sociology, biology, and computer science, disciplines where systems are often 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.

In the recent literature, various metrics are proposed to evaluate the graph structure of a social network. Such are "Betweenness" [43], "Centrality" [39], Clustering coefficient (a measure of the likelihood that two associates of a node are associates themselves).

A higher clustering coefficient indicates a greater "cliquishness", that is, 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 the capacity of an individual to reach out the whole network.

Other interesting measures include "Structural cohesion". While cohesion metrics have been studied a lot in sociology, there does not seem to be a general agreement. Cohesion in its essence is the ability of any network not to split up when changes are made, and from this point of view, ideas like the density of interactions in the network [1,24,26] and the relational distance between nodes [35] are used as basic features for cohesion. The issues with these ideas are that—as it is also noted in [38]—the cohesion of a group could depend on only one node; additionally, these ideas are conceived for a non-directed network where each interaction is in both directions, thus making these metrics not directly applicable in a directed network. In [38], the cohesion, in a connected group of nodes, is defined by the number of nodes that, if removed, would disconnect the group. The measurement of this feature is connected with the number of paths a node has to another one, which would make the calculation of the cohesion in a large graph computationally difficult.

An popular alternative for evaluating cohesiveness is the k-core structure. A k-core in a graph G is a maximum subgraph H where each nice in H has at least k neighbors inside H; in this case, k is the *index* of the k-core H. A high-index core in a graph indicates a "strongly interrelated community" where everybody has some minimum connection with the rest. As graph theoretic concepts cores are being studied from the 1960s [22,42,36,41]. However, their use, in an experimental level, for evaluating and detecting strongly cohesive communities in real-word graphs has been used extensively in several topics such as the study of internet topology [2,15], large-scale network visualization [2,3,8], networks of protein interaction [5,44], and complex network modeling and organization [10,19]. A more general notion of k-cores was introduced in [9] where, instead of vertex degrees, more general functions were considered. In the more recent work of [29], a fractional analogue of the k-core concept is defined in order to detect and evaluate communities in citation datasets where the corresponding graphs are bipartite, that is, relate members of distinct entities such as authors with papers.

Part of our work focuses on applying our evaluation techniques on citations graphs (DBLP, ArXiv). Recent work on citation graphs can be found in [4] where a study is carried out on the citation graph of Computer Science Literature and [31]. In [4], an attempt is made to extract a descriptive summary of the graph through a study of fundamental and well-established properties (degree distribution, giant component size etc.). In contrast, our work focuses on novel techniques for evaluating community graphs and expands on a wider scope of study. In [31], the focus is on community detection and the evolution through time. The community detection is performed on the authors through the papers they have co-cited, and the evaluation of the citation graph is based on the detected clusters.

3 Our results

So far, no analogous notion or study has been done in datasets where the relations are ordered, that is, they are represented by directed graphs. A typical such graph is a citation graph where nodes are papers and where the existence of a directed edge (p_1, p_2) reflects the fact that paper p_1 is citing paper p_2. In this case, to demand that vertices of a subgraph have many neighbors is not enough as we have to take into account the tendency of vertices to be hubs (i.e. to have many out-going edges) or authorities (i.e. to have many in-going edges). Therefore, we need a more refined notion of a core in order to capture an analogue of a coherency measure in such graphs.

In this paper, we define, for the first time, an analogue of the k-core structure for directed graphs called D-*core*. It appears that directed graphs require a *two-dimensional analogue* of k-cores as each vertex v of a directed graph as two type of neighbors: those where v is pointing to (the *out-neighbors*) and those that are pointing to v (the *in-neighbors*). Then, the (k, l)-core of a directed graph D (in short: *digraph*) is a maximum subdigraph F os D where all vertices have at least k out-neighbors in F and at least l in-neighbors in F. This generates a criterion of cohesiveness where, by taking distinct values of k and l, we may tune the relation between hubs and authorities in the related D-core.

In [34], an idea similar to the D-cores is used to filter out less significant nodes, by pruning them out. The main difference to our approach is that it removes only a sufficient portion of the nodes. The cores are then fed to a generalized HITS algorithm used to expand the communities within them.

In [16], 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 in 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.

4 D-cores and relevant structures

In this section, we introduce the D-core concept along with the structures that enable finding the optimal subgraphs (with regard to cohesion) and identifying highly collaborative parts in directed graphs.

4.1 Preliminaries

Let $G = (V, E)$ be a graph. A *subgraph* H of G is a graph obtained by G after removing vertices or edges and we denote this by $H \subseteq G$. Given a vertex $x \in V$ we define its *degree* as the number of vertices that are adjacent with x in G and we denote it by $\mathbf{deg}_G(x)$. The *min-degree* of a graph G is defined as

$$\delta(G) = \min\{x \mid \mathbf{deg}_G(x) \mid x \in V(G)\}.$$

A *k-core* in a graph G is a subgraph H of G where $\delta(H) \geq k$. The *degeneracy* of a graph G denoted by $\delta^*(G)$ is the maximum k for which G contains a non-empty k-core. Assume that for $i = 0, \ldots, \delta^*$, G_i is the i-core of G. Then, we call the sequence $V(G_0), V(G_1), \ldots, V(G_{\delta^*(G)})$ *core sequence* of G and observe that $V(G_i) \subseteq V(G_{i+1})$ for $i \in \{0, \ldots, \delta^*(G) - 1\}$. We also call the sequence $V(G_1) - V(G_1), \ldots, V(G_{\delta^*(G)}) - V(G_{\delta^*-1})$ *cell sequence* of G and we observe that its elements form a partition of $V(G)$.

k-cores are fundamental structures in graph theory and their study dates back to the 1960s [22,36,42]. The parameter of *degeneracy* appeared with several names such as width [37], linkage [27], and the coloring-number [18]. The existence of a k-core in a graph indicates the existence of a highly interconnected community where every node is linked with at least k other nodes. The existence of k-cores of large size in sufficiently dense graphs has been theoretically studied by [40] for random graphs generated by the Erdős-Rényi model [23]. As shown in [40], a k-core whose size is proportionate to the size of G (i.e. a "giant" k-core) 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.

Here, we extend the notion of a k-core to directed graphs so that they can represent well interconnected communities on networks whose links are of directional nature, i.e. are represented by directed edges. To our knowledge, this is the first time such an extension is proposed.

4.2 D-cores

Let $D = (V, E)$ be a digraph that is a set V of vertices and a set E of directed edges between them. Each edge $e \in E$ can be seen as a pair $e = (v, u)$ and we say that v is the *tail* of e, while u is the *head* of e. We denote the set of vertices of a digraph D by $V(D)$. Given a vertex $x \in V$, its *in-degree*, we denote it by $\mathbf{deg}^{\text{in}}_D(x)$, is the number of *in-links* of x, that is, the edges in D with x as a head. Similarly, the *out-degree* of x, we denote it by $\mathbf{deg}^{\text{out}}_D(x)$, is the number of *out-links* of x, that is, edges in D with x as a tail. The *min-in-degree* and the *min-out-degree* of a digraph D are defined as

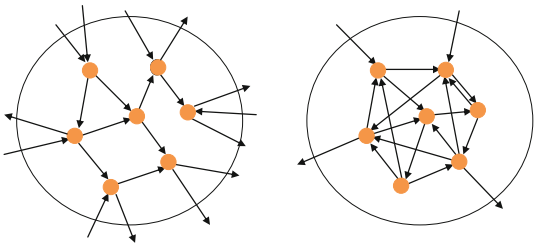

Fig. 1 Two portions of a digraph. The one in the *left* does not contain any non-trivial (k, l)-core and the one in the *right* is a (2,2)-core

$$\delta^{\text{in}}(D) = \min\{x \mid \mathbf{deg}^{\text{in}}_D(x) \mid x \in V(D)\} \quad \text{and}$$
$$\delta^{\text{out}}(D) = \min\{x \mid \mathbf{deg}^{\text{out}}_D(x) \mid x \in V(D)\},$$

respectively. Given two positive integers k, l and a digraph $D = (V, E)$, a (k, l)-D-*core* of D is a maximal size subdigraph F of D where $\delta^{\text{out}}(F) \geq k$ and $\delta^{\text{in}}(F) \geq l$; if no such digraph exists, then the (k, l)-D-core of D is the empty digraph. It is easy to see that when such a subdigraph exists, it is unique.

Given a digraph D, we denote by $\mathbf{DC}_{k,l}(D)$ the (k, l)-D-core of D. We also denote by $\mathbf{dc}_{k,l}(D)$ the size of $\mathbf{DC}_{k,l}(D)$, that is, the number of its vertices. As D will always be the network under study, we may just use the simpler notations $\mathbf{DC}_{k,l}$ and $\mathbf{dc}_{k,l}$ instead.

The intuition behind (k, l)-D-cores is to find a subdigraph where all nodes have enough out-links and in-links to the rest of it. Clearly, it is not enough for a node to have big in-degree and/or out-degree in order to be a member of such a core. What counts, on the top of this, is that the node forms part of a community where each of its members satisfies the same in-degree and/or out-degree requirements with respect to all the other community members (see Fig. 1 for an example). This indicates that nodes in a D-core exhibit a strong collaboration behavior among them.

The detection of $\mathbf{DC}_{k,l}$ is computationally easy and can be done by the following procedure:

Procedure $Trim_{k,l}(D)$
Input: A digraph D and positive integers k, l
Output: $\mathbf{DC}_{k,l}(D)$
1. let $F \leftarrow D$.
2. while *there is a node x in F such that*
$\mathbf{deg}^{\text{out}}_F(x) < k$ *or* $\mathbf{deg}^{\text{in}}_F(x) < l$,
delete *node x from F*.
3. return F.

Let $L = (v_1, \ldots, v_m)$ be a layout of the vertices of D. For every $i = 1, \ldots, n$, we denote by D_i the digraph induced by the vertices in $\{v_1, \ldots, v_i\}$. We say that L is (k, l)-*eliminable* if for every $i \in \{0, \ldots, n\}$, either $\mathbf{deg}^{\text{out}}_{D_i}(v_i) < k$ or $\mathbf{deg}^{\text{in}}_{D_i}(v_i) < l$.

The following Lemma on (k, l)-D-cores generalizes the classic min-max result of [37] (see also [27,32]).

Lemma 1 *Given a digraph D and two positive integers k and l, the (k, l)-D-core is empty if and only if there exists a (k, l)-eliminable layout of $V(D)$.*

Lemma 1 essentially indicates that the elimination procedure of the algorithm $Trim_{k,l}(D)$ works correctly and (optimally) runs in $O(m)$ steps where $m = |E(G)|$. The proof is easy and follows the arguments of [27] for the non-directed case (see also [9]).

For an optimal implementations of the $Trim_{k,l}(D)$ procedure, see the general algorithm of [9] that is based on the same ideas for the undirected case. In our implementation of this procedure, $\mathbf{DC}_{k,l}(D)$ is incrementally computed for all pairs of k and l.

4.3 Degeneracy of digraphs

The degeneracy of a directed digraph differs radically from its undirected counterpart. Actually, it has a two-dimensional nature since different choices of the lower bounds to the number of incoming/outcoming edges result to different D-cores.

The *degeneracy* of a digraph D is defined as follows.

$$\delta^*(D) = \frac{1}{2}\max\{\delta^{\text{out}}(H) + \delta^{\text{in}}(H) \mid H \subseteq D\}. \tag{1}$$

The intuition behind the definition of $\delta^*(D)$ is to return the maximum r (for some pair k, l where $k + l \geq 2r$) such that D contains a non-empty (k, l)-D-core (δ^* takes semi-integer values). Also the value of $\delta^*(D)$ may correspond to multiple (k, l)-D-cores for different choices of k and l (those where $k + l = 2 \cdot \delta^*(D)$).

Notice that if we replace each edge of a graph by two opposite direction edges, the degeneracy of the resulting digraph is equal to the degeneracy of G. Thus δ^* is indeed a valid generalization of undirected degeneracy to directed graphs. We stress that δ^* is the first density parameter on digraphs that takes into account Hub/Authority trade-offs as it differs radically (and is not comparable) with previous digraph density measures such as the ones defined in [16] and [34]. A powerful extension of the classic notion of a k-core was given in [9] where the k-core is defined as a set of vertices where some general vertex property function is bounded. While the results in [9] can also provide a natural concept of k-core for directed graphs, they are not able to capture the "two-dimensional" nature of our (k, l)-core concept where degree bounds are applied *simultaneously* on both the in-degrees and the out-degrees.

Let τ be a real number in the interval $[0, \pi/2]$ representing an angle. The τ-*degeneracy* of a digraph D is defined as follows.

$$\delta^*_\tau(D) = \max\left\{ \frac{\lceil k \rceil + \lceil l \rceil}{2} \mid G \text{ contains a non-empty } (k,l) - \text{D-core where } k = r \cdot \cos(\tau) \text{ and } l = r \cdot \sin(\tau) \text{ for some } r \text{ where } r^2 = l^2 + k^2 \right\}$$

In the above definition, one may see each pair (k, l) as a point of a Cartesian system of coordinates, corresponding to the D-core $\mathbf{DC}_{k,l}(D)$. To compute $\delta^*_\tau(D)$, we essentially follow the τ-slope segment starting from (0,0) until $\mathbf{DC}_{k,l}(D)$ becomes empty along this line. The last such non-empty D-core is the one determining the degeneracy of D with respect to the angle τ. The value of τ reflects the Hub/Authority trade-off in the considered D-cores and we refer to it as H/A-*angle*.

Again it is easy to observe that $\delta^*_{\pi/4}$ deteriorates to classic degeneracy when we replace each edge of an undirected graphs by two (opposite) directed edges. Observe that δ_τ can also provide an another definition of δ^*, equivalent to the one in (1), as $\delta^*(D) = \max\{\delta^*_\tau(D) \mid \tau \in [0, \pi/2]\}$.

D-core matrix. Our objective is to define a series of digraph-based metrics, based on directed degeneracy, in order to evaluate the dense collaboration of nodes in networks whose links

have directional nature. The whole network is represented by a digraph D and there is a unique $\mathbf{DC}_{k,l}$ for each $k, l \geq 0$. The sizes $\mathbf{dc}_{k,l}$, (for $k, l \geq 0$) define an (infinite) matrix $A_D = (\mathbf{dc}_{k,l})_{k,l \in \mathbb{N}}$ that we call D-*core matrix of* D. The notion of $A_D(k, l)$ is the two-dimensional digraph analogue of the notion of core sequence defined in Sect. 4.1 for the undirected case. For each $k, l \geq 0$, we define

$$\mathbf{DCL}^{\text{out}}_{k,l} = V(\mathbf{DC}_{k,l}) - V(\mathbf{DC}_{k+1,l}) \quad \text{and}$$
$$\mathbf{DCL}^{\text{in}}_{k,l} = V(\mathbf{DC}_{k,l}) - V(\mathbf{DC}_{1,l+1})$$

Also, we set $\mathbf{dcl}^{\text{out}}_{k,l} = |\mathbf{DCL}^{\text{out}}_{k,l}|$ and $\mathbf{dcl}^{\text{in}}_{k,l} = |\mathbf{DCL}^{\text{in}}_{k,l}|$. In other words, the values of $\mathbf{DCL}^{\text{out}}_{k,l}$ and $\mathbf{DCL}^{\text{in}}_{k,l}$ represent the "differential" of the matrix A_D taken in both horizontal and vertical direction. For this reason, we define the matrices $\partial^{\text{out}} A_D = (\mathbf{dcl}^{\text{out}}_{k,l})_{k,l \in \mathbb{N}}$ and $\partial^{\text{in}} A_D = (\mathbf{dcl}^{\text{in}}_{k,l})_{k,l \in \mathbb{N}}$. To visualize them, one may see the values of A_D as being assigned to the squares of an infinite two-dimensional grid centered to the esquire (0,0) and the values of $\partial^{\text{out}} A_D$ and $\partial^{\text{in}} A_D$ as are assigned to the vertical and horizontal edges of this grid. We identified the matrix A_D and its differentials $\partial^{\text{out}} A_D$ and $\partial^{\text{in}} A_D$ for the digraph formed by the Wikipedia (2004, English edition). The nodes correspond to Wikipedia pages and each directed edge $e = (x, y)$ is a link from page x to page y. Cell (k, l) in the matrix A_D stores the size $(\mathbf{dc}_{k,l})_{k,l \in \mathbb{N}}$ of the respective d-core $\mathbf{DC}_{k,l}$. As agreed before, we see the coordinates (k, l) as squares of an infinite two-dimensional grid and we assign the values $\mathbf{dcl}^{\text{out}}_{k,l}$ and $\mathbf{dcl}^{\text{in}}_{k,l}$ to its edges.

The result for the case of A_D is depicted in Fig. 11. As there is no Wikipedia entry with more than 51 out-links or more than 43 in-links, we restrict this matrix to its lower 51×43 portion. For each digraph D that we examine, we call this matrix D-*core matrix* of D, we visualize its cells as squares of an infinite two-dimensional grid Γ_D, and we depict the size of its (k, l)-cores by coloring the corresponding squares with different colors. According to Fig. 11, the value of $\delta^*(D_{\text{Wiki}})$ for the Wikipedia digraph D_{Wiki} is obtained in cell (38, 41) and is equal to $\frac{38+41}{2} = 39.5$. In other worlds, 39.5 is the half of the Manhattan distance between a cell of the D-core matrix of D_{Wiki} and the cell (0,0); in our case, this cell is (38, 41) and this justifies the value of $\delta^*(D_{\text{Wiki}})$.

For the cases of $\partial^{\text{out}} A_D$ and $\partial^{\text{in}} A_D$, we adopt the visualization of Fig. 2 that makes it possible to depict together differential values in both directions: Consider the grid Γ_D depicting A_D in Fig. 11. For each square in this grid, we add a new vertex in its center, we draw an edge connecting it to its 4 corners, and then we remove the square. Notice that the resulting graph is a new infinite grid, we denote it by $\partial\Gamma_D$, whose squares are corresponding either to horizontal or to vertices edges of Γ_D. That way we can assign the values of $\partial^{\text{out}} A_D$ to "vertical" squares of $\partial\Gamma_D$ and the values of $\partial^{\text{in}} A_D$ to "horizontal" squares of $\partial\Gamma_D$. The colors of the squares of $\partial\Gamma_D$ correspond to the different sizes of $\mathbf{DCL}^{\text{out}}_{k,l}$ and $\mathbf{DCL}^{\text{in}}_{k,l}$. That way the visualization of Fig. 2 can be seen as a visualization of the discrete differential values of the matrix A_D depicted in Fig. 11.

We call a sequence of squares in Γ_D *incremental* if for each two consecutive squares (x, y), (x', y'), it holds either $x' = x + 1$ and $y' = y$ or $x' = x$ and $y' = y + 1$. Each incremental sequence that starts from (0,0) corresponds to a possible scenario of considering consecutive D-cores of D by gradually incrementing either the demand on the minimum out-degree or the demand on the minimum in-degree.

We notice in Fig. 2 that the size of Wikipedia drops fast (at pace that can reach up to 10.000 nodes per step) as the minimum values of in/outlinks increase, especially in the range [1–10]. The drop in graph size is more significant for the outlinks case, showing that the outlinks graph is less robust to degeneracy. Then, the graph size reduces less aggressively—thus these

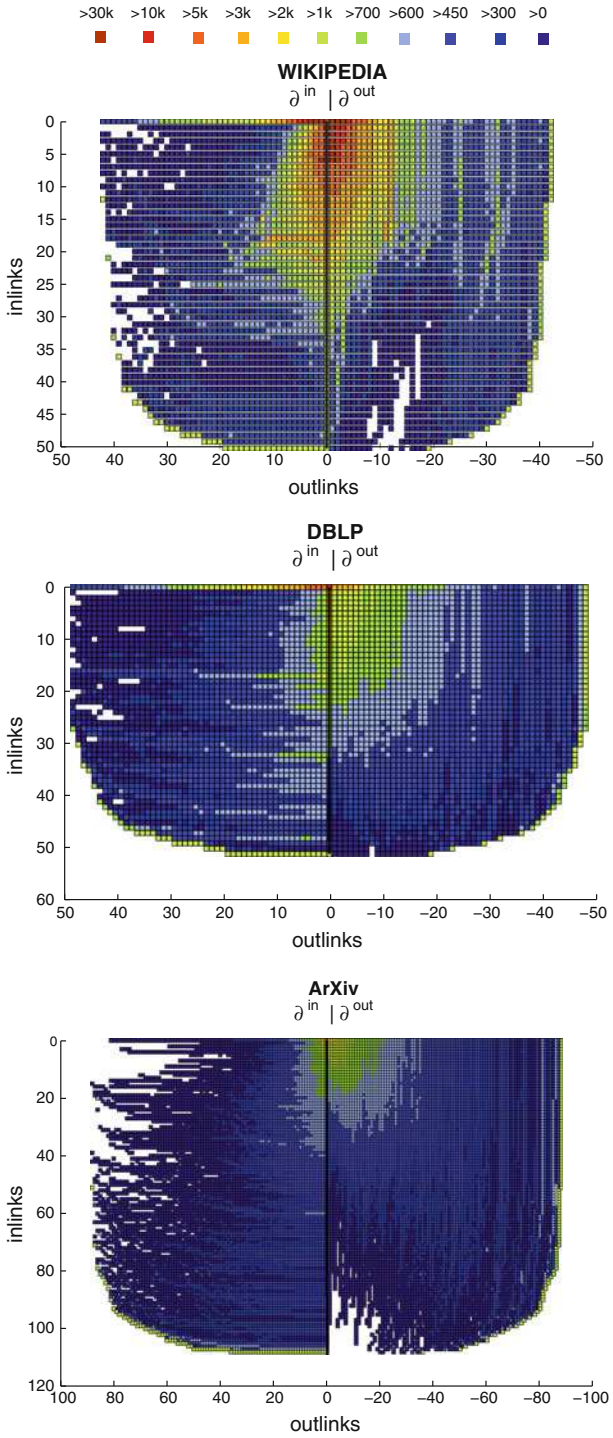

Fig. 2 The differentials $\partial^{\text{out}} A_D$ and $\partial^{\text{in}} A_D$ for the digraphs formed by the Wikipedia, DBLP, and ArXiv shown from *top* to *bottom* in that order. *White squares* indicate a value of zero

graphs are robust. It is clear that for the range of values inlinks/outlinks [20,35]/[30,40], the Wikipedia graph is very robust with regard to the inlink reduction rate. Similar behavior is observed for the DBLP and ArXiv graph, but the pace of size drop is much smaller in DBLP and even smaller for ArXiv. Overall, the most robust under degeneracy graph seems to be ArXiv—indicating a dense citation network. This is also due to the high link density observed.

4.4 Digraph degeneracy frontiers

The following observation follows directly from the definitions:

Observation 1 *For every k, k', l, l' where $k \geq k'$ and $l \geq l'$ it holds that $\mathbf{DC}_{k,l}$ is a subdigraph of $\mathbf{DC}_{k',l'}$ and therefore, $\mathbf{dc}_{k,l} \leq \mathbf{dc}_{k',l'}$.*

We call a cell (k, l) *frontier cell* for a digraph D if $\mathbf{dc}_{k,l} > 0$ and $\mathbf{dc}_{k+1,l+1} = 0$—thus the frontier consists of the cells corresponding to the last non-empty D-cores as k or l increase. The set of frontier cells of a digraph D is denoted as $F(D)$. Formally:

$$F(D) = \{(k, l) : \mathbf{dc}_{k,l} > 0 \,\&\, \mathbf{dc}_{k+1,l+1} = 0\}$$

See Fig. 11 where the frontier appears as the squares that have some common point with 0-valued squares (i.e. the white area).

The (k, l)-D-cores corresponding to the frontier cells are the *frontier* D*-cores* of D and all of them together constitute the D*-core frontier of* D. Intuitively, these D-cores exhibit the highest collaboration behavior in the network for different Hub/Authority trade-offs (i.e. H/A-angles).

Let $k_{\max}$ be the maximum k for which $(k, 0) \in F(D)$ and $l_{\max}$ be the maximum l for which $(0, l) \in F(D)$. We call $(k_{\max}, 0)$, $(0, l_{\max})$ *extreme cells* of $F(D)$. Observe that number of frontier cells is always equal to $k_{\max} + l_{\max} - 1$. Thus, the extreme $\mathbf{DC}_{0,l_{\max}}$ represents the D-core with no in-links and a maximum number of out-links. In the Wikipedia graph, the $\mathbf{DC}_{0,50}$ represents the subdigraph bearing to a maximum the Hub-property (i.e. many out-links thus a very "extrovert" D-core). On the contrary, the extreme $\mathbf{DC}_{k_{\max},0}$ represents the D-core with no out-links and a maximum number of in-links. In case of the Wikipedia digraph, this graph is $\mathbf{DC}_{42,0}$.

Consider a core sequence $\mathcal{L}$ in Γ_D that starts from square (k, l) and finishes in square (k', l'). let $e_1, \ldots, e_r$ be the sequence of edges that belong to consecutive squares of $\mathcal{L}$. Notice that each e_i corresponds to some square of $\partial\Gamma_D$ that, in turn, corresponds to some vertex set that is either $\mathbf{DCL}^{\text{out}}_{x,y}$ (in case e_i is a vertical edge) or $\mathbf{DCL}^{\text{in}}_{x,y}$ (in case e_i is an horizontal edge) for some value of x and y. We conclude that each monotone sequence $\mathcal{L}$ corresponds to a sequence of vertex sets that form a partition $\mathcal{P}$ of the vertex set $V(\mathbf{DC}_{k,l}) - V(\mathbf{DC}_{k',l'})$. That way, the size of $V(\mathbf{DC}_{k,l}) - V(\mathbf{DC}_{k',l'})$ (or, equivalently, the value $\mathbf{dc}_{k,l} - \mathbf{dc}_{k',l'}$) is the number of vertices that are discarded in order to transform $\mathbf{DC}_{k,l}$ to $\mathbf{DC}_{k',l'}$, following the core sequence $\mathcal{L}$. Notice that this number always the same no matter the choice of the elimination sequence $\mathcal{L}$ (while certainly the partition $\mathcal{P}$ may vary a lot). Therefore, we can say that the edge weighting of Γ_D defined by $\partial^{\text{out}} A_D$ and $\partial^{\text{in}} A_D$ is *adiabatic* in the sense that all paths between two vertices have the same total weight.

We are now in position to define the mono-dimensional analogue of core sequence and cell sequence in directed graphs. A *core sequence* of a directed graph D is an incremental sequence of squares in Γ_D that starts from (0,0) and finishes in some square of the D-core frontier of D.

We conclude that each core sequence $\mathcal{L}$ corresponds to a sequence of vertex sets that form a partition of the vertex set of D. This sequence is called *cell sequence* of D and we denote it by $\mathcal{P}(\mathcal{L})$. As there exist an exponential number of core sequences, the same holds also to the number of different partitions we may consider. This sharply contrasts the mono-dimensional undirected case where the corresponding cell partition is uniquely defined.

5 Digraph collaboration indices

In this section, we treat the issue of choosing the optimal D-core on the frontier, as the most representative of the specific graph D-cores, with regard to the collaborative features as implemented via dense in/outlinks connectivity. To this end, we take into account different properties of digraph degeneracy, especially with regard to the frontier. Intuitively, we are interested in the dominant trend in the frontier D-cores, that is, whether they contain more in-links or out-links.

Following this line, we define a series of metrics quantifying distinct measures of robustness.

Balanced collaboration index (BCI)

One possibility is to choose a D-core with a balanced rate of in/outlinks. Thus, we define the *balanced collaboration index* of D as the unique integer r for which $\mathbf{DC}_{r,r}$ is a frontier (r, r)-D-core. In other words, we find the coordinates of the cell where the diagonal intersects the D-core frontier of D. Formally, the *balanced collaboration index* of D, $\mathsf{BCI}(D)$ is equal to $\delta^*_{\pi/4}(D)$ (i.e. the H/A-angle is of 45°). The choice of the diagonal focuses on the D-cores with a balanced Hub/Authority trade-off - thus containing vertices that are connected to others, on average, with equal lower bounds their in- and outlinks.

Optimal collaboration index (OCI)

In this case, we choose the frontier D-cores $\mathbf{DC}_{k,l}$ for which $(k + l)/2$ is maximized. In terms of the D-core diagram, the position of such D-core has the maximum (among other frontier D-cores) Manhattan distance from the origin (0,0) and corresponds. Formally, the *optimal collaboration index*, $\mathsf{OCI}(D)$, is equal to $\delta^*(G)$. Notice that the frontier (k, l)-D-cores where $\frac{k+l}{2}$ is maximized can be multiple and may correspond to *several* H/A-angles.

Inherent collaboration index (ICI)

This index aims to represent the inherent hubs/authority trade-off in the graph and is based on the average ratio of out-links to in-links of the vertices in the digraph. Based on this, we define the average H/A-angle of a digraph D as follows.

$$\rho_{\mathsf{av}} = \tan^{-1}\left(\frac{1}{|V(\mathbf{DC}_{1,1}(D))|} \cdot \sum_{v \in V(\mathbf{DC}_{1,1}(D))} \frac{\mathbf{deg}^{\text{in}}_D(v)}{\mathbf{deg}^{\text{out}}_D(v)}\right).$$

To make the above formula feasible, we excluded vertices with zero in- or outlinks, that is, we applied the averaging inside the D-core $\mathbf{DC}_{1,1}(D)$. The *inherent collaboration index*, $\mathsf{ICI}(D)$, of the digraph D is equal to be $\delta^*_{\rho_{\mathsf{av}}}(D)$ where ρ_{av} is defined as above.

Thus, we use the terms: BCI/OCI/ICI—optimal D-core(s), respectively, for the D-cores corresponding to each particular optimization. See Fig. 11 for a depiction of the above indices on the Wikipedia D-cores matrix frontier.

Average collaboration index (ACI). This index is the average of the τ-degeneracies over all possible H/A-angles corresponding to the cells of the D-core frontier of D. Thus, the *average collaboration index*, $\mathsf{ACI}(D)$, of the digraph D is defined as

$$\frac{1}{|F(G)|}\sum_{(k,l)\in F(D)}\delta^*_{\tan^{-1}(\frac{l}{k})}(D).$$

In other words, $\mathsf{ACI}(D)$ is the half of the average Manhattan distance of the frontier cells of D. Alternatively, we may define $\mathsf{ACI}(D) = \frac{\sum_{(k,l)\in F(D)}(k+l)}{2\cdot|F(D)|}$.

Robustness. Notice that the maximum value of the average collaboration index of a digraph D with extreme positions $(k_{\max}, 0)$ and $(0, l_{\max})$ is obtained in the case where

$$F(D) = \{(k_{\max}, 0), (k_{\max}, 1), \ldots, (k_{\max}, l_{\max}), \\ (k_{\max}-1, l_{\max}), \ldots, (0, l_{\max}))\}.$$

In this extreme and, in a sense, ideal case, the digraph D has the maximum possible robustness under degeneracy with respect to its extreme positions and the Average Collaboration Index of such a graph is equal to

$$\frac{2k_{\max}l_{\max} - k_{\max} - l_{\max} + \binom{k_{\max}+1}{2} + \binom{l_{\max}+1}{2}}{2\cdot|F(D)|}.$$

We denote the above quantity by $\mu(k_{\max}, l_{\max})$. That way, we define the *robustness* of a digraph D with extreme positions ($k_{\max}$ and $l_{\max}$) as the ratio:

$$\frac{\sum_{(k,l)\in F(D)}(k+l)}{\mu(k_{\max}, l_{\max})}$$

and it always results in a real value in [0, 1] (Table 1).

The above definition implies that the robustness is essentially the surface enclosed between the $F(D)$ frontier and the $(0, 0), \ldots, (k_{\max}, 0), (0, 0), \ldots (0, l_{\max})$ coordinates divided by $\mu(k_{\max}, l_{\max})$. This represents the endurance of the D-core graph to degeneracy, that is, the degree of cohesion among the graph nodes—in terms of globally distributed in/out links.

5.1 Set frontiers and indices

Let X be a subset of nodes in a digraph D. In a similar manner as above, we define the D-core matrix of X, $\mathbf{DC}^X_{k,l}(D)$, as the cells (k, l) where X is a subset of $\mathbf{DC}_{k,l}$ and $\mathbf{dc}_{k,l} > 0$. Similarly, we define the **D**-core frontier of X, as the set of the extreme non-empty D-cores corresponding to the cells (k, l) where $\mathbf{dc}_{k,l} > 0$ and $\mathbf{dc}_{k+1,l+1} = 0$. Thus:

$$F_D(X) = \{(k, l) : X \subseteq D \,\&\, \mathbf{dc}_{k,l} > 0 \,\&\, \mathbf{dc}_{k+1,l+1} = 0\}$$

The D-core matrix of a nodes set $X \subseteq V(D)$ is defined in an analogous way as in Sect. 5.1, which represents the capacity of the nodes of X to be part, *all-together*, in subgraphs with strong mutual linking and thus presenting a noteworthy collaboration behavior.

The five collaboration indices for a set $X \subseteq V(D)$ as well as its robustness are defined analogously as in previous sections.

The *balanced collaboration index of* X, $\mathsf{BCI}_D(X)$ is the maximum r for which $X \subseteq V(\mathbf{DC}_{r,r})$.

The *optimal collaboration index of* X, $\mathsf{OCI}_D(X)$ is the maximum value of $\frac{k+l}{2}$ for which $X \subseteq V(\mathbf{DC}_{k,l})$.

Table 1 Collaboration indices values for the Wikipedia graph

	Wikipedia	Continental Congress	United States Congress
BCI(k,l)/Size of optimal DC	38 (38,38)/237	9	19
ICI(k,l)/angle/size of optimal DC	36.5/(40,33) 51.34/190	10.8	22,18
OCI$(k,l)/\langle (k,l)$/angle/ size of optimal DC$\rangle$	39.5/⟨ (43,36)/47,66/⟩ 228 ⟨(41,38)45,42/233⟩	19.7	42.8
Robustness.Local	x	0.78	0.389
Robustness. Global	0.96	0.1	0.791
ACI	32.46	9.5	20.31
AC H/A-angle (degrees)	41.8	54.57	56.957
AC H/A-angle (rads)	0.73	0.95	0.994
	Progressive Conservative Party of Canada	Congress of Vienna	Gregorian Calendar
BCI(k,l)/Size of optimal DC	8	12	27
ICI(k,l)/angle/size of optimal DC	8.7	13.11	28.24
OCI$(k,l)/\langle (k,l)$/angle/ size of optimal DC$\rangle$	1.50	8.23	42.12
Robustness.Local	0.166	0.153	0.54
Robustness. Global	0.762	0.861	0.85
ACI	16.042	12.474	23.904
AC H/A-angle (degrees)	13.316	34.76	51.458
AC H/A-angle (rads)	0.232	0.606	0.898

The *Inherent collaboration index of* X, $\mathsf{ICI}_D(X)$ is the maximum $(\lceil k\rceil + \lceil l\rceil)/2$ for which $X \subseteq V(\mathbf{DC}_{k,l})$ where $k = r \cdot \cos\rho_{\mathsf{av}}$ and $l = r \cdot \sin\rho_{\mathsf{av}}$, for some r where $r^2 = k^2 + l^2$ (ρ_{av} is the average H/A-angle, defined as in the previous subsection).

The *robustness of a set* X with extreme positions ($k_{\max}$ and $l_{\max}$) is defined as the ratio: $\frac{\sum_{(k,l)\in F_D(X)}(k+l)}{\mu(k_{\max},l_{\max})}$ where the function μ is defined as in the previous section.

The *Average collaboration H/A-angle of a set* with extreme positions ($k_{\max}$ and $l_{\max}$) is defined as:

$$\sigma_D(X) = \frac{\sum_{(k,l)\in F_D(X)}(k+l)\cdot \tan^{-1}(\frac{l}{k})}{\Sigma_{(k,l)\in F_D(X)}(k+l)}$$

As before, this angle conveys the Hub/Authority trade-off for the D-cores in which X is a subgraph.

These indices can be applied also to every individual node $x \in V(D)$ by setting $X = \{x\}$. In this case, all above notations and concepts can also be used for nodes instead of sets of nodes. Notice that all indices defined in this subsection are anti-monotone. In particular:

Observation 2 *Let X_1 and X_2 are subsets of the vertex set of some digraph D. If $X_1 \subseteq X_2$, then the balanced/optimal/inherent/diagonal collaboration index of X_1 will be at least the balanced/optimal/inherent/diagonal collaboration index of X_2.*

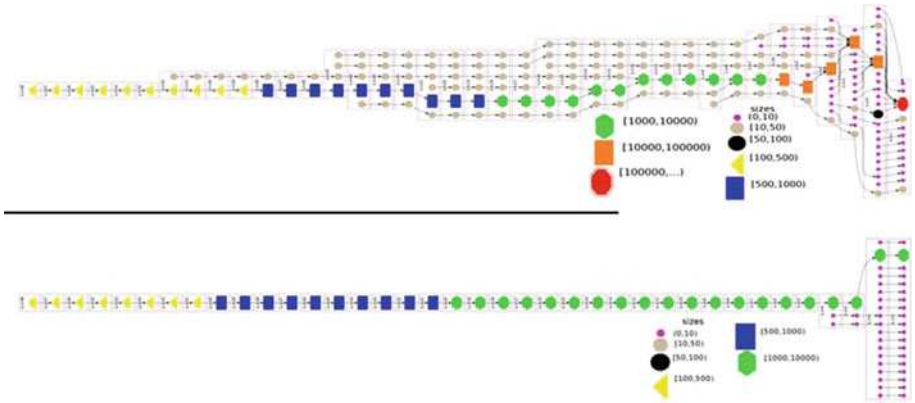

Fig. 3 The CDF corresponding to the diagonal D-cores (i, i) for Wikipedia 2004 (*upper*), DBLP (*bottom*). SCCs are depicted with *different colors* depending on their sizes

5.2 Core decomposition forests

In this section, we define the concept of a core decomposition forest in order to examine the structure of a digraph and the connected components of its cores. We need first some definitions. We say that a digraph D is *strongly connected* when every pair x and y of vertices in D has been met by some directed cycle, that is, there is a directed path from x to y and a directed path from y to x. A *strongly connected* component (in short: SCC) of a digraph D is any maximum subdigraph of D that is strongly connected. Finding the strongly connected components of a digraph, graph can be done in time linear to the sum of its edges and vertices.

Let $\mathcal{D} = D_0, D_1, \ldots, D_d$ be a sequence of digraphs such that for each i, j where $i \leq j$, D_i is a subgraph of D_j (we call such a sequence *monotone*). The *Decomposition Forest* of a monotone digraph sequence $\mathcal{D}$ is the digraph $\mathbf{DF}(\mathcal{D})$ defined in the following way: For each $i = 0, \ldots, d$, we denote the strongly connected components of D_i by $D_i^1, \ldots, D_i^{m_i}$ and each such strongly connected component is a vertex of $\mathbf{DF}(\mathcal{G})$. An edge $(G_i^j, G_{i'}^{j'})$ is added in $\mathbf{DF}(\mathcal{G})$ if $j' = j + 1$ and G_i^j contains $G_{i'}^{j'}$ as a subdigraph.

The above definition implies directly that $\mathbf{DF}(\mathcal{D})$ is a union of trees, each rooted to some of the strongly connected components of D_0. Given now a directed graph D and one say $\mathcal{L}$ of its cell sequences,

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 D_0. Given that, by its definition, each core sequence $\mathcal{L}$ of D is monotone, we define the *Core Decomposition Forest (CDF) of D with respect to* $\mathcal{L}$ as the decomposition forest corresponding to $\mathcal{L}$. The undirected analogue of the core decomposition forest appeared for the first time in [30] under the name *hierarchical degree core tree* and was used in order to visualize the connected components of several real-word graphs including the graph extracted by the common author relation of the papers of the DBLP citation graph. The same notion was also used in [29] for studying citation graphs from DBLP and ARXIV with the difference that in [29] now the common paper relation of the authors is being studied. In our case, this notion is extended in digraphs, and in our experimental study, we visualize the core decomposition forests for both Wikipedia 2004 and DBLP, where the sequence $\mathcal{L}$ corresponds to the cells in the diagonal of each D-core matrix (see Fig. 3).

6 Experimental evaluation

In this section, we present the experiments we performed applying the above algorithms and definitions on real-world datasets.

6.1 Directed graph degeneracy for scale-free graphs

Real-world web graphs have been found to display scale-free characteristics [6,7,33] evident by the power law degree distribution. Here, we are also exploring author citation graphs, which share the same properties (as it can been seen in their degree distributions). Scale-free graphs are frequently modeled by the combination of growth with preferential attachment. There have been many variations in this modeling for both directed and undirected cases, but the main idea is that the graph grows one vertex at the time and edges are added (between vertices that may be new or old). The key idea in the preferential attachment scheme is that the probability of taking an edge is proportional to the respective degrees of its endpoints. This intuitively matches with the mechanism of the evolution of both web graphs and citations graphs of authors (i.e. a "popular" page is more likely to get in-links and a "famous" author is more likely to get a citation from a new page/paper following the "rich get richer" of the preferential attachment process).

As the scale-free model seems to approximate the graphs we examine, we have chosen it for evaluation with our D-core computation procedure to see whether the results are similar for both various parameters and parameters that produce graphs with approximately similar degree distributions with the real-world graphs.

6.1.1 Preliminaries for preferential attachments

Barabasi and Albert in [6] were the first to introduce a scale-free model for undirected graphs. In that model, the graph is generated with a small number of initial vertices m_0 and grows by adding each time a new vertex with $m(\leq m_0)$ edges from the new vertex to the old ones. Preferential attachment is introduced in the selection of the old nodes; the probability a vertex i depends on the degree of that vertex, so that $\Pi(k_i) = k_i / \sum_j k_j$ where k_i the degree of the vertex. The Barabasi–Albert model was examined in more detail by Bollobás et al. in [12] and in [13] where a detailed model called *Linearized Chord Diagram* (LCD) was designed. This applies to directed and undirected graphs as well; a parameter m is used and if $m = 1$ then at each step t a new vertex v_t is added to a given graph $G_1^{(t-1)}$ with a single edge between v_t and v_i where i is chosen randomly with

$$\mathbb{P}(i = s) = \begin{cases} \frac{\deg_{G_1^{(t-1)}}(v_s)}{2t-1} & 1 \leq s \leq t-1, \\ \frac{1}{2t-1} & s = t \end{cases}$$

For $m > 1$, m edges are added from v_t to v_i one at a time, each time counting the previous edges in the total degree of each v_i.

In [21] and [20], a variation in the Barabasi–Albert model is introduced, where a constant parameter α represents the "initial attractiveness" of a node. Here, the old vertices are chosen based on a probability proportional to their degree plus the "initial attractiveness". Thus, the selection probability, defined in detail in [14], is:

$$\mathbb{P}(i = s) = \begin{cases} \frac{\deg_{G_1^{(t-1)}}(v_s)+\alpha}{2t-1} & 1 \leq s \leq t-1, \\ \frac{\alpha}{(\alpha+1)t-1} & s = t \end{cases}$$

The constant parameter here is important as it introduces a mixture of uniform and preferential attachment behavior (where if $\alpha = 1$ we have only preferential attachment). This model is also important as it resembles the directed one we utilized for our experiments. Another model that also introduced a mixture of uniform and preferential attachment was in the work of Cooper and Frieze [17]. Here instead, the uniformity was defined explicitly by defining additional parameters that would determine the probability of selecting a uniform or preferential attachment model. Furthermore, they define two different steps : a) one of growth and b) one that chooses to connect two old vertices together with a new edge. This model is also important as it gives the opportunity to control the density of a graph by controlling the probability between the two steps.

As these models seemed to be better suited for models of undirected graphs, we finally chose the model introduced by Bollobás, Borgs, Chayes, and Riordan in [11]. This model, as seen below in the description, has a initial preference parameter for both the in- and out-degrees, while also following the general idea between different steps as in the Cooper-Frieze model. Following we offer a description of that model:

> We consider a graph that grows by adding single edges at discrete time steps. At each such step, a vertex may or may not also be added. For simplicity, we allow multiple edges and loops. More precisely, let α, β, γ, δ_{in}, and δ_{out} be non-negative real numbers, with $\alpha + \beta + \gamma = 1$. Let G_0 be any fixed initial graph, for example a single vertex without edges, and let t_0 be the number of edges of G_0. (Depending on the parameters, we may have to assume $t_0 \geq 1$ for the first few steps of our process to make sense.) We set $G(t_0) = G_0$, so that at time t the graph $G(t)$ has exactly t edges, and a random number $n(t)$ of vertices. In what follows, to choose a vertex v of $G(t)$ according to $d_{\text{out}} + \delta_{\text{out}}$, means to choose v so that $Pr(v = v_i)$ is proportional to $d_{\text{out}}(v_i) + \delta_{\text{out}}$, that is, so that $Pr(v = v_i) = (d_{\text{out}}(v_i) + \delta_{\text{out}})/(t + \delta_{\text{out}} n(t))$. To choose v according to $d_{\text{in}} + \delta_{\text{in}}$, means to choose v so that $Pr(v = v_i) = (d_{\text{in}}(v_i) + \delta_{\text{in}})/(t + \delta_{\text{in}} n(t))$, where all degrees are measured in $G(t)$.
> For $t \geq t_0$ we form $G(t + 1)$ from $G(t)$ according to the following rules:
>
> (A) With probability α, add a new vertex v together with an edge from v to an existing vertex w, where w is chosen according to $d_{\text{in}} + \delta_{\text{in}}$.
> (B) With probability β, add an edge from an existing vertex v to an existing vertex w, where v and w are chosen independently, v according to $d_{\text{out}} + \delta_{\text{out}}$ and w according to $d_{\text{in}} + \delta_{\text{in}}$.
> (C) With probability γ, add a new vertex w and an edge from an existing vertex v to w, where v is chosen according to $d_{\text{out}} + \delta_{\text{out}}$.
>
> The probabilities α, β, and γ clearly should add up to one. To avoid trivialities, we will also assume that $\alpha + \gamma > 0$. When considering the web graph, we take $\delta_{\text{out}} = 0$; the motivation is that vertices added under step (C) correspond to web pages which purely provide content—such pages never change, are born without out-links and remain without out-links. Vertices added under step (A) correspond to usual pages, to which links may be later added. While mathematically it seems natural to take $\delta_{\text{in}} = 0$ in addition to $\delta_{\text{out}} = 0$, this gives a model in which every page not in G_0 has either no in-links or no out-links, which is rather unrealistic and uninteresting! A non-zero value of δ_{in} corresponds to insisting that a page is not considered part of the web until something points to it, typically one of the big search engines. It is natural to consider these edges from search engines separately from the rest of the graph, as they are of a rather different nature; for the same reason, it is natural not to insist that δ_{in} is an

integer. We include the parameter δ_{out} to make the model symmetric with respect to reversing the directions of edges (swapping α with γ and δ_{in} with δ_{out}), and because we expect the model to be applicable in contexts other than that of the web graph.

Our choice for this model was based both on the sophistication it displayed and the ability to produce graphs with behavior, in the degree distribution, very similar to our real datasets (see bellow).

6.1.2 Generating preferential attachment graphs

We created a set of graphs adopting the preferential attachment model according to [11] (see the previous section) for various parameters. In this section, we present findings on this model for a set of 4 different parameters:

1. $\alpha = 0.018,\ \beta = 0.102,\ \gamma = 0.88,\ \delta_{\text{in}} = 1,\ \delta_{\text{out}} = 2$
2. $\alpha = 0.018,\ \beta = 0.102,\ \gamma = 0.88,\ \delta_{\text{in}} = 5,\ \delta_{\text{out}} = 1$
3. $\alpha = 0.102,\ \beta = 0.238,\ \gamma = 0.66,\ \delta_{\text{in}} = 1,\ \delta_{\text{out}} = 3$
4. $\alpha = 0.001,\ \beta = 0.009,\ \gamma = 0.99,\ \delta_{\text{in}} = 1,\ \delta_{\text{out}} = 1$

The size of the graph is 16,500 nodes so that it will approximate the number of nodes that have in/out-degree of at least 1. The reader can see the distributions of resulting graphs in Fig. 5 in the same order from left to right and top to bottom. It is clear that all these graphs are scale-free. We ran the defined algorithms and metrics and, in what follows, we report on their expressive power and features.

6.1.3 D-core matrices for the synthetic data

Following the same sequence of parameters as before, here we describe the findings on the datasets we created. Firstly, we explain the meaning of the parameters starting with the γ, the parameter that controls the density of the network. Parameters α and β control the out- and in-degree behavior, respectively, while δ_{in} and δ_{out} represent the aforementioned"initial preference" for the respective in and out degrees.

For the first two datasets, we chose the same values for α, β, and γ so that we compare how the other two affect the results. The value of γ was chosen, experimentally, to produce an "average" density. Given the fact that the α parameter is lower than β, we expect to have a more extrovert behavior but we expect that to change for the second dataset as the δ_{in} parameter is a lot larger than the δ_{out}. These expectations are confirmed by the D-core matrix behavior as seen in Fig. 4. It is clearly visible that the ICI angle changes when the δ_{out} increases and the ICI line (in green) moves closer to the diagonal (in dark gray).

The next two datasets demonstrate how the γ parameter affects the "extend" of the D-cores. Since it is closely correlated to the density of a graph, we expect that the degeneracy would be affected accordingly. This would mean that, for a low value of γ, we would get graphs that would produce only low-degeneracy D-cores and for a high value the opposite. This is also confirmed by the results. As the reader can see in the two D-core matrices in the bottom part of Fig. 4, we get a graph that degrades really fast for a γ value of 0.66. On the other hand, when we chose a value of 0.99, we can easily see that the resulting graph is much more robust. This is evident by the high numbers for in- and out-degrees that the graph survives in the D-core matrix.

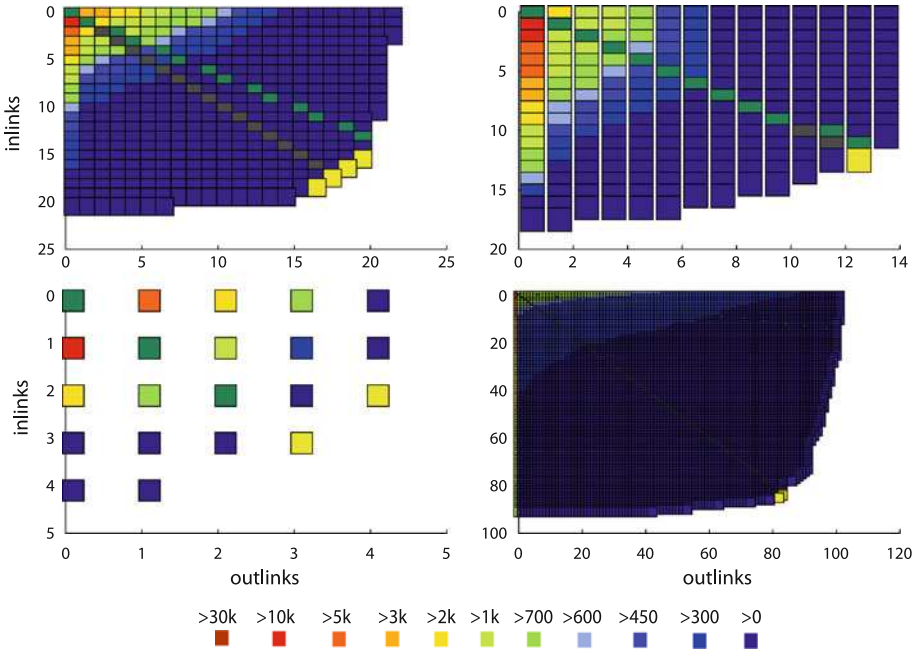

Fig. 4 D-core matrices for 4 different parameter sets on the adopted model. **a** *Top left* $\alpha = 0.018$, $\beta = 0.102$, $\gamma = 0.88$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 2$. **b** *Top right* $\alpha = 0.018$, $\beta = 0.102$, $\gamma = 0.88$, $\delta_{\text{in}} = 5$, $\delta_{\text{out}} = 1$. **c** *Bottom left* $\alpha = 0.102$, $\beta = 0.238$, $\gamma = 0.66$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 3$. **d** *Bottom right* $\alpha = 0.001$, $\beta = 0.009$, $\gamma = 0.99$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 1$

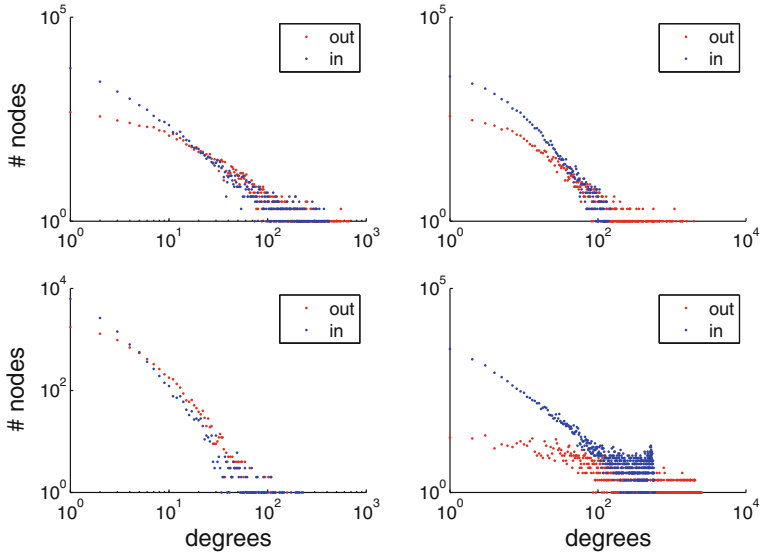

Fig. 5 Distributions for 4 different parameter sets on the adopted model. **a** *Top left* $\alpha = 0.018$, $\beta = 0.102$, $\gamma = 0.88$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 2$. **b** *Top right* $\alpha = 0.018$, $\beta = 0.102$, $\gamma = 0.88$, $\delta_{\text{in}} = 5$, $\delta_{\text{out}} = 1$. **c** *Bottom left* $\alpha = 0.102$, $\beta = 0.238$, $\gamma = 0.66$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 3$. **d** *Bottom right* $\alpha = 0.001$, $\beta = 0.009$, $\gamma = 0.99$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 1$

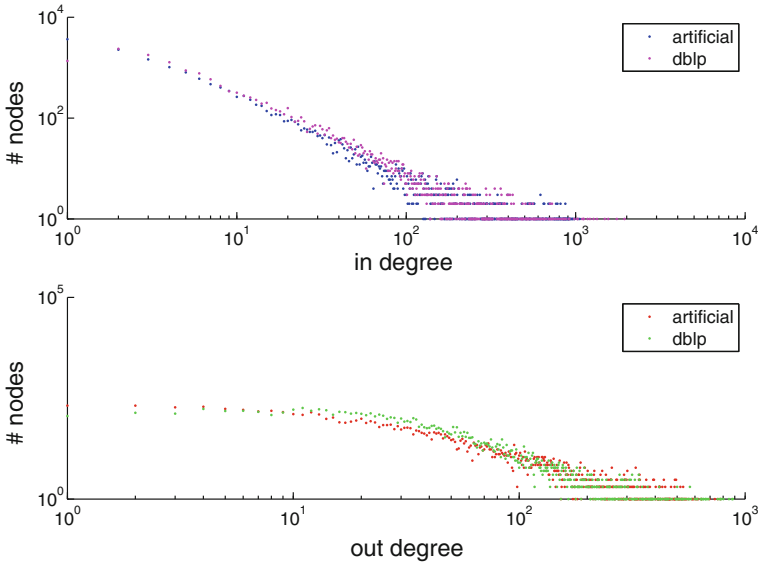

Fig. 6 Comparison of the distributions for the in/out degrees between the chosen parameters ($\alpha = 0.011$, $\beta = 0.031$, $\gamma = 0.958$, $\delta_{\text{in}} = 2$, $\delta_{\text{out}} = 5$) and the DBLP graph

6.1.4 Comparison to real-world Data

In this section, we chose parameters producing a graph with degree distributions similar to a real-world dataset and verify this via a comparison to the DBLP data. For this reason, we chose, experimentally, the following parameters for approximating the DBLP digraph: $\alpha = 0.011$, $\beta = 0.031$, $\gamma = 0.958$, $\delta_{\text{in}} = 2$, and $\delta_{\text{out}} = 5$. We can see evidence of the approximation from comparison of the in-/out-degree distributions in Fig. 6.

In Fig. 7, we can see that the behavior is quite similar to the previous one. The single interesting difference is how the size of the D-cores drops. On the synthetic graph case, we see a dramatic drop indicating that the inner structure is less connected.

When we look at the CDF forest comparison in Fig. 8—if we exclude the small SCCs in the initial cores in the DBLP digraph—the two figures look similar as in both cases, there is a giant component that survives robust until the end. Again there are some insignificant differences mostly on the rate at which the size of giant SCC drops.

In conclusion, the synthetic digraph seems to approximate quite well the DBLP graph with regard to the D-core behavior. This is important as it could be possible to predict the D-core metrics of a real-world graph of immense scale simply by producing a down-scaled 'miniature' of it by its parameters.

6.2 Data sets description

The Wikipedia dataset is a snapshot of the English version of Wikipedia, the digraph consists of about 1.2M nodes and 3.662M links. The snapshot depicts Wikipedia as it was in the January 2004 and was extracted from a database dump containing the entire history of the encyclopedia; available at http://download.wikipedia.org/.

In our experiments, we also used a popular bibliographic dataset derived from the available snapshot of DBLP, which is freely available in XML format at: http://dblp.uni-trier.de/xml.

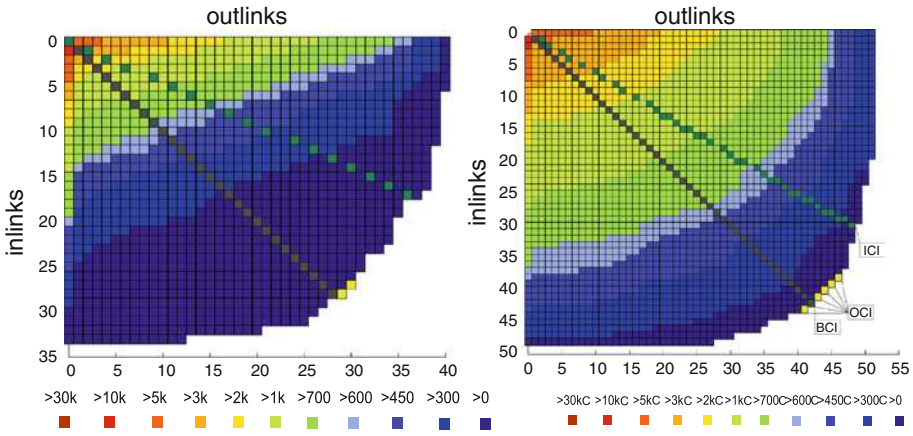

Fig. 7 The D-core matrices of the synthetic digraph (*left*) and the DBLP digraph (*right*)

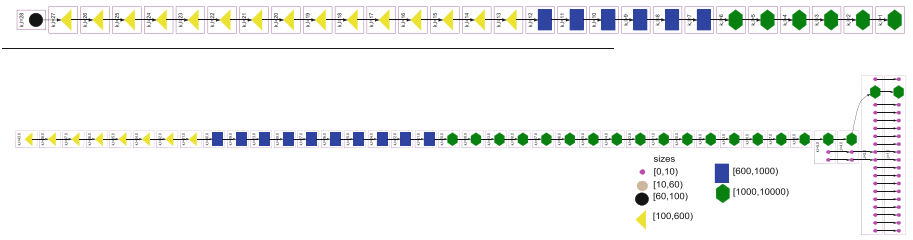

Fig. 8 The CDF corresponding the diagonal D-cores (i, i) for the synthetic/artificial (*upper*), DBLP (*bottom*). SCCs are depicted with *different colors* depending on their sizes

We obtained a digraph structure from the dataset as follows: authors correspond to the nodes of the digraph and each directed edge $e = (x, y)$ express the fact that author x cited in his/her papers a paper of author y. That way, obtain a digraph containing about 825K author nodes and 351K edges. The vast majority of them have no in-/out-links (about 800K) thus we remain with the rest 25K authors that are minimally connected.

Additionally, we have run experiments on the Arxiv HEP-TH (high energy physics theory) citation graph. This is a paper citation graph is originally from the e-print arXiv with 27.700 papers and is freely available at http://snap.stanford.edu/data/cit-HepTh.html.

From the paper citation graph, we extracted the author citation digraph similar to the DBLP one, containing 8,821 authors and 391K edges/citations.

If Fig. 9, the reader can see the degree distribution of both in- and out-degree for the three datasets. We see that all of them display a scale-free behavior governed by a power law; we carried out a parameter fitting to identify approximately that behavior. In more detail, we see that all three of them display a clear preferential attachment behavior with regard to the in-degree, probably with no "initial attractiveness" (see the described models above). Instead, in the out-degree, we see that even though there is a general scale-free behavior, there is also evidence of the "initial attractiveness" parameter being larger than the absolute minimum. This is evident by the somewhat uniform behavior for the "smaller" degrees (not including the degree of 1). Intuitively, papers with more than zero citations to other papers will cite a few papers, meaning more than one. On the other end, a paper cannot have too many citations, that is, outlinks. The previous applies naturally to authors as well. This in

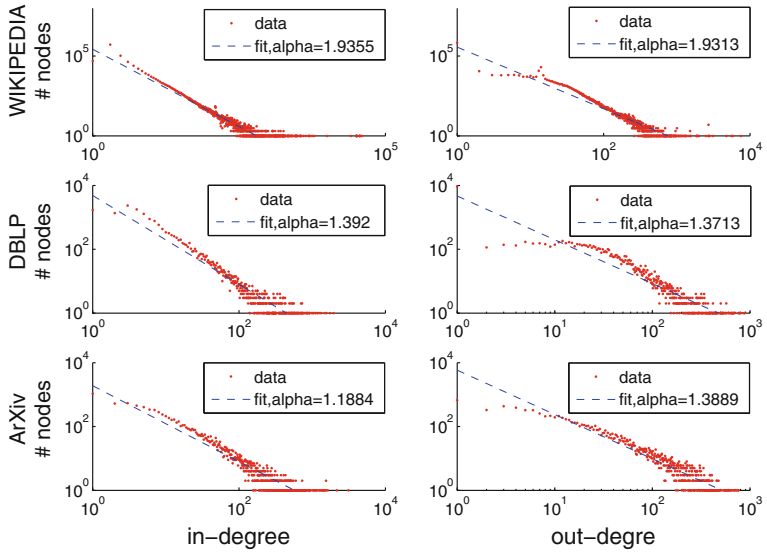

Fig. 9 Distributions of the in- and out-degree for the real-world datasets as noted above in log–log scale with power-law fitting. The exponent of the power law is also displayed

a way resembles the δ_{out} parameter of our adopted model. Thus, the δ_{out} (for the model we adopt) has to be larger than 1 for the citations networks. As we can see later, the parameters that fitted the closest to the DBLP dataset adapt to this intuition.

6.3 Algorithms complexity

The proposed D-core algorithm is of low complexity; thus, D-core computations are feasible even in large-scale digraphs. As shown in procedure $Trim_{k,l}(D)$ in Sect. 4.2, the computation of each D-core is linear to the number of its edges and thus optimal. Moreover as the digraphs we examine are sparse, the identification of the D-cores is very fast.

The D-core matrix computation starts from the original digraph and reduces it until the degeneracy leads to an empty one. This procedure involves about $(40 \times 50) \sim 2000$ repeated executions, in the case of the Wikipedia digraph, of the basic $Trim_{k,l}(D)$ procedure. Depending on the implementation, each execution can be done on commodity desktops in the scale of minutes even in million scale-sized graphs, as it is also noted in [9] for the case of non-directed graphs.

6.4 Experimental methodology

The experimental method for processing the previously mentioned digraphs involved the following phases:

1. D-*core matrix computation:* this involves computing the D-core $\mathbf{DC}_{k,l}$ subgraph, where $(k,l) \in \{0,\ldots,k_{\max}\} \times \{0,\ldots,l_{\max}\}$ where $(k_{max},0)$, $(0,l_{max})$ are the extreme cells of $F(D)$. According to Observation 2, a D-core $\mathbf{DC}_{i,j}$ is a subgraph of every D-core $\mathbf{DC}_{i',j'}$ where $i' \leq i$ and $j' \leq j$. Based on this property, we can efficiently compute, for example, the D-core $\mathbf{DC}_{0,2}$ having computed and stored in memory the D-core $\mathbf{DC}_{0,1}$. Therefore, in order to compute the entire D-core diagram, we started by computing only

the D-cores in row 0 and column 0 and used those two sets of D-cores to "fill in" the rest of the matrix (note that the D-cores $\mathbf{DC}_{0,1}$ and $\mathbf{DC}_{1,0}$ are not correlated so we need to compute both but we only need *one of them* to fill the rest of the matrix). Each D-core occupies moderate storage space, such that the whole D-cores matrix occupies less than 4GB of disk space, so storing them for subsequent use was an obvious choice.

2. *Collaboration indices computation*: We compute the values that optimize the criteria set along with the sizes of the corresponding D-cores. Namely, we compute the corresponding BCI/ICI/OCI/ACI, indices and the Robustness.
3. *Strongly Connected Components (SCCs) and Core Decomposition Forests (CDF's)*: Let D be the digraph corresponding to Wikipedia 2004 or DBLP. For each D-core $\mathbf{DC}_{i,i}$ – that is, on the D-core matrix diagonal—we computed the strong connected components. We also considered the core elimination sequence $\mathcal{L} = \mathbf{DC}_{0,0}, \ldots, \mathbf{DC}_{r,r}$ where r is the BCI of D and we computed the corresponding Core Decomposition Forests for both graphs.
   SCCs indicate groups of strong cohesiveness in the D-core. See Fig. 3 for a detailed view on the SCCs size evolution and subdigraph relationships as i, running along the D-core matrix diagonal, increases for both datasets considered.
4. *Frontiers for sets of entries*: We also computed the frontiers for single terms/authors for theWikipedia and DBLP digraphs, respectively. This is also extended, as defined above, to sets of terms/authors. These indicate the robustness (represented by the values of the indices) for the D-cores containing them.

6.5 Experimental results on Wikipedia

The D-core matrix and indices values. We processed the Wikipedia digraph and computed for each (k, l) cell of the D-core matrix the sizes of the resulting D-cores (see Fig. 11) as well as the sizes of the SCC's in each of the D-core(i, i), i.e. on the diagonal of the matrix as mentioned before.

We computed all the above-defined indices for the global Wikipedia digraph as well as for selected representative terms and sets of terms (see Fig. 10). For Wikipedia 2004, the balanced collaboration index(BCI) value is 38, while the respective D-core $\mathbf{DC}_{38,38}$ contains 237 nodes. For the same digraph, the inherent collaboration index ICI is 36 and is obtained for the D-cores $\mathbf{DC}_{39,33}$ that contains 206 nodes. For the OCI index, we obtain two OCI-optimal frontier cells corresponding to the $\mathbf{DC}_{38,41}$ and $\mathbf{DC}_{36,43}$ D-cores containing 228 and 233 nodes, respectively. The *robustness* of the global Wikipedia digraph is remarkably high at 0.963, while the maximum value is 1, indicating a very robust digraph.

D-cores frontiers for terms and sets of terms. Then, we investigate the cohesion and in/outlinks trade-off of D-cores containing specific term pages. These metrics are perceived as indication of the collaborativeness and authority/hubness of the digraphs containing these term pages. Further, we present representative terms-pages D-core matrices evaluating them.

As defined in Sect. 5.1, the D-core diagram of a vertice containing term X corresponds to the D-cores of the D-core diagram of D whose vertices sets contain X. In Fig. 10, we see the D-cores matrix frontiers for the digraphs containing the terms: Congress of Vienna, Continental Congress, Gregorial Calendar, Progressive Conservative party of Canada, and United States Congress. In each subfigure, we see the frontier of the respective digraphs degeneracy, each presenting different features and trends. The frontier for the term Continental Congress for example is presenting a low BCI index with regard to the global digraph (the BCI index is 38), as the page is participating in D-cores with low degeneracy.

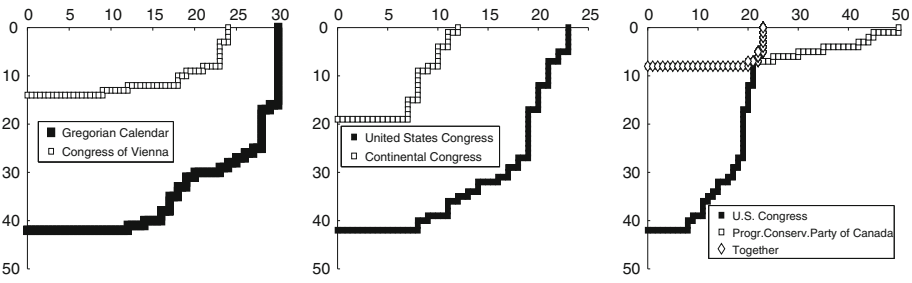

Fig. 10 Selected term pages and sets of term-pages frontiers from Wikipedia

Its respective ICI index is (19.7) much lower than the global ICI value 36. This is a rather "selfish" page as it participates in D-cores dominated by in-links.

Contrary to the previous, the Gregorian Calendar page participates in much more robust D-cores as its BCI index reaches a high 26, while its OCI is a very high – occurring at cell (42,12)—indicating a very "selfish behavior" dominated by inlinks and thus having an authority digraph behavior. On the other hand, the Congress of Vienna page is presenting a rather extrovert behavior as its OCI index occurring at cell (8,23), an indication of outlinks domination in the optimal subgraphs. The robustness of the digraph is rather low with a BCI index at 11, a low value as compared to the global BCI 38.

In Fig. 10 (right), we present the joint D-core matrix and frontier of two term pages (Progressive Conservative Party of Canada and United States Congress). The "together" frontier represents the frontier of the D-core digraphs containing both terms. The joint D-core frontier can exhibit much worse robustness under degeneracy (i.e. removing in/outlinks) than the individual ones. This can be the case when the D-core frontiers of term pages with contradictory trends are put together; as it is in our example, where the joint frontier is at $\mathbf{DC}_{8,22}$. Thus, we obtain a much weaker digraph than the ones of the individual terms.

Thematic focus of Wikipedia SCCs

We computed the SCCs of the Wikipedia D-cores $\mathbf{DC}_{i,i}$ on the balanced diagonal direction (BCI direction). The intuition is that the SCCs are considered as digraph areas with high cohesion. In Fig. 3, the reader can see the cardinality of the SCCs in each Wikipedia D-core $\mathbf{DC}_{i,i}$, the size of the SCCs and their hierarchical containment relation as i increases along the BCI axis. As we notice, starting in D-core $\mathbf{DC}_{1,1}$, there are several SCCs moderately sized (<100 nodes)—excluding one significantly larger-sized SCC (>100K nodes in D-core $\mathbf{DC}_{1,1}$. Many of the SCCs survive until the D-core $\mathbf{DC}_{32,32}$, after this only the initial giant component survives until the extreme BCI D-core $\mathbf{DC}_{38,38}$.

Further, we investigate the thematic focus of the SCCs as we study the D-cores along the BCI optimal axis, see Table 2. We observe a giant component that dominates and almost all the pages contain the terms "time". We pruned the digraph, removing those pages and we noticed a similar behavior, this time with the term Grammy awards dominating the single giant SCC remaining. It is interesting to stress that in D-core $\mathbf{DC}_{1,1}$ there are 1,034 SCCs (apart from the giant one). The size of the top-5 SCCs ranges between 5 and 24 nodes, while for each one, there is a remarkably narrow focus in their thematic area. For instance, see Table 2, the top sized SCC is about Wisconsin. The rest of the SCCs are thematically focused in: Cynodonts species, Iowa, Eurovision, History of the British penny, Submarines, Wyoming. In D-core $\mathbf{DC}_{2,2}$, we have only 23 SCCs (apart from the giant one). The size of the top-5 SCCs ranges between 3 and 30 nodes, while the thematic focus of the top sized SCCs is to a large degree identical to the top SCCs in D-core $\mathbf{DC}_{1,1}$. A similar trend continues as i increases along the diagonal $\mathbf{DC}_{i,i}$.

Table 2 The thematic focus of the Wikipedia SCCs for increasing degeneracy along the BCI axis

(k, k)	# SCCs	Top-k SCCs size	Thematic focus
1	1024	24	Wisconsin
		10	Cynodonts Species
		10	Iowa
		10	Eurovision
		5	History of the British penny
		5	Submarines
		10	Wyoming
2	23	30	Music albums
		10	Eurovision
		6	Cynodonts Species
		6	Metal Deficiencies
		5	History of the British penny
		3	Helladic
3	13	23	Extinct species
		10	Eurovision Young Dancers
		6	Metal Deficiencies
		6	Books
		5	Cynodonts Species
		5	History of the British penny
4	12	26	poker jargon
		10	Eurovision
		6	Metal Deficiencies
		5	History of the British penny
		5	films by decade
		4	Fayette
5	8	26	poker jargon
		17	Sibley-Monroe checklist
		10	Eurovision
		7	North Carolina
...			...
38	1		Dates

6.6 Experimental results on DBLP

We processed the DBLP digraph and found for each cell (k, l) of the D-core matrix the size of the resulting D-cores (see Fig. 11 bottom) as well as the number of strongly connected components (SCC's) in each of the D-cores $\mathbf{DC}_{i,i}$—that is, on the diagonal (see Fig. 3 bottom). We computed all the above-defined indices for the global DBLP digraph as well as for selected representative authors and sets of authors.

For the case of the DBLP digraph, the value of BCI is 42 (see Table 3 a summary of all indices values) while the respective D-core $\mathbf{DC}_{42,42}$ contains 188 nodes (see the lower part of Fig. 11). For the same digraph, the inherent collaboration index ICI is 39 and is obtained for the D-core $\mathbf{DC}_{30,48}$ that contains 220 nodes. For the OCI index, we get a value 42, which

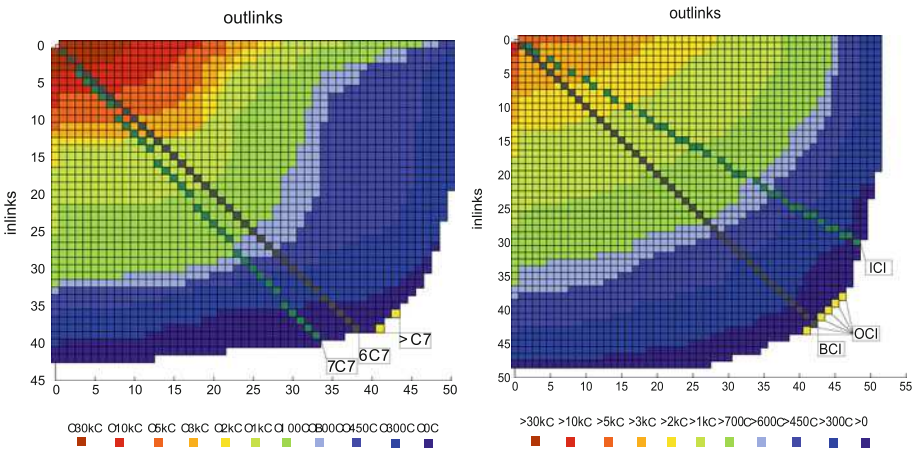

Fig. 11 The D-core matrices of the Wikipedia 2004 digraph (*left*) and the DBLP digraph (*right*)

occurs in six D-cores located at the positions: (38, 46),(39, 45),(40, 44),(41, 43),(42, 42),(43, 41) on the D-core matrix frontier. The *robustness* of the global DBLP digraph is remarkably high at 0.966, indicating a very robust to degeneracy digraph. It is evident that the DPLP digraph has significant extrovert features (i.e. more out than in citations, an expected result).

We also computed the SCCs of the DBLP D-cores $\mathbf{DC}_{i,i}$ on the balanced diagonal direction (BCI direction). In Fig. 3, bottom, one can see the cardinality of the SCCs in each DBLP D-core $\mathbf{DC}_{i,i}$, the size of the SCCs, and their containment relation as i increases. As we notice, starting in D-core $\mathbf{DC}_{1,1}$, there are few SCCs poor sized (<10 nodes)—excluding one significantly larger-sized SCC (>1,000 nodes in $\mathbf{DC}_{1,1}$—that survive until $\mathbf{DC}_{4,4}$. After this only the initial giant component continues until the extreme BCI D-core $\mathbf{DC}_{42,42}$. This SCC apparently contains the nodes/authors with a large number of mutual citations.

The giant SCC contains 188 authors (Table 4) presenting both top publication activity, thus many outgoing citations, as well as high rate of incoming citations. This group of authors indeed contains well known and reputable scientists' names and looks pretty reasonable. Of course, we have to stress the partial coverage of the DBLP dataset as its citation bulk is before 2004. Also, in the first years of its function, the emphasis is on database-related papers.

We further studied the D-cores corresponding to specific authors and computed the respective D-core matrices and frontiers. We selected two characteristic cases of seminal authors. In Fig. 12 (left), we see the D-core matrix and frontier for "E.F Codd", founder of the relational database area. His BCI extreme is $\mathbf{DC}_{42,23}$ indicating an intensive inlinks (incoming citations) trend. This is natural as he was authoring in the early years of computer science with few previous works to cite. On the contrary his works enjoy a very high number of citations, thus a high number of inlinks in the citations digraph.

On the other hand, a more modern seminal author G. Weikum presents a very robust to degeneracy D-core structure for both in/outlinks tendency. This is explained by the facts. i. his works are highly cited during many years and ii. he is intensively authoring and thus citing other authors. In Fig. 12, (right) we present the joint D-core matrix and frontier for the two aforementioned authors. The "together frontier" represents the frontier of the D-cores that contain both E.F. Codd and G. Weikum author (nodes), thus representing the D-cores (i.e. citation subgraphs) in which the two aforementioned cite in common and they are commonly cited (Table 5).

Table 3 Collaboration indices values for the DBLP digraph

	DBLP	E.F. Codd	G. Weikum
BCI(k,l)/ Size of optimal DC	42/188	22/913	41/221
ICI/(k,l)/angle/ size of optimal DC	39/(30,48)/32.01/220	19/(15,23)	38/(29,47)
OCI/⟨(k,l)/angle/ size of optimal DC ⟩	42/⟨(43,41). . .(38, 46)/ 43.63,. . .,50.44/165,188,217, 187,185,188 ⟩	31.5/(42,21)	41.5/(38,45)
Robustness,Local	–	0.457	0.966
Robustness, Global	0.966	0.952	0.928
ACI	35.17	23.083	33.66
AC H/A-angle (deg)	43.90	55.66	41.91

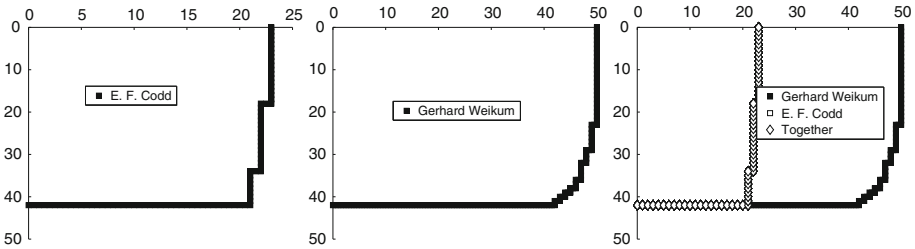

Fig. 12 Representative authors D-core frontier from the DBLP digraph

6.7 Experimental results on ARXIV

Following the same procedure, with the other two datasets, we produced the D-core matrix and the CDF for the ArxiV citation graph. The reader can see the results in Fig. 13. It is interesting to stress that this graph has a much denser core resulting in much larger metric values as it can be easily seen in the respective D-core matrix. Additionally, we see that the CDF is dominated by one SCC in its largest part. Furthermore, we noticed that the initial giant component survives until the extreme BCI D-core $\mathbf{DC}_{83,83}$. Thus, this graph is much more robust to degeneracy than all the others we tested, indicating thus a very dense collaboration among the members of the theoretical Physics community. The authors of this core can be seen in Table 6. It is evident that all the senior names in this scientific area appear here justifying their close collaboration to the community in terms of in/out citations. We stress here that we used the abbreviated version for most of the author names as these were more frequent in the dataset.

As for the other characteristics, the inherent collaboration index ICI is characterized by an angle of 25 degrees at the $\mathbf{DC}_{50,107}$ D-core with size of 306. For the OCI index, we obtain three cells $\mathbf{DC}_{78,95}$, $\mathbf{DC}_{79,94}$ and $\mathbf{DC}_{80,93}$ with respective sizes of 237, 241, and 244 nodes, respectively.

The *robustness* of the ArXiv graph is high as well at 0.9704, indicating, much like the DBLP one, very high robustness to degeneracy digraph. Again we observe overall some very extrovert features meaning that the graph is featured mostly by outgoing citations. In this case, we could say that the ArXiv digraph displays higher extroversion than the DBLP.

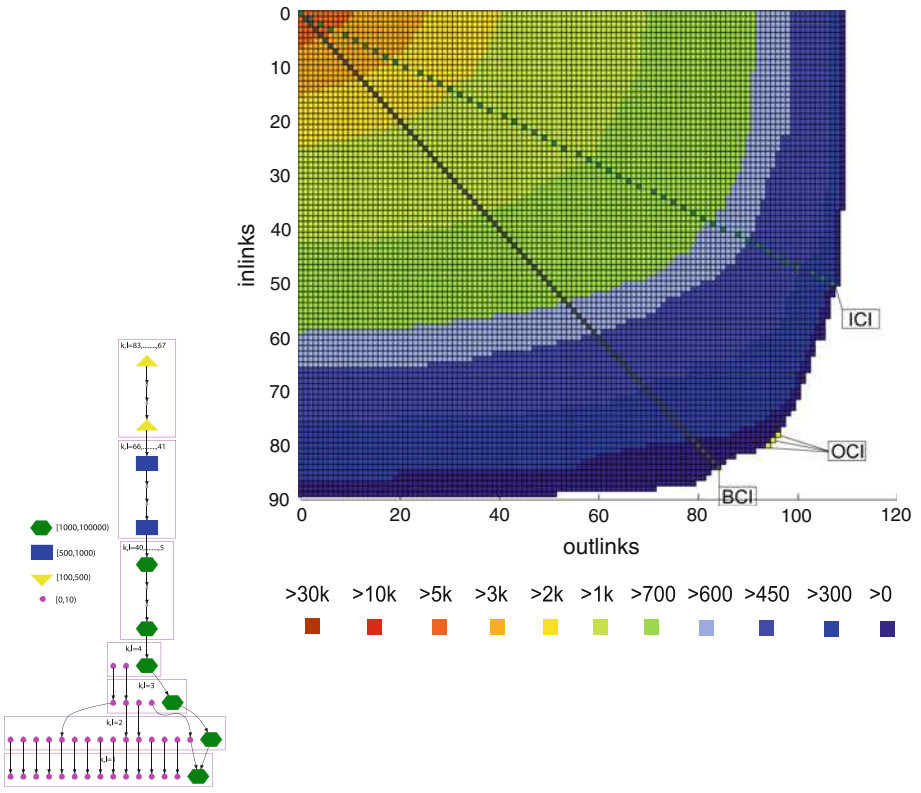

Fig. 13 *Left* the CDF corresponding to the diagonal D-cores (i, i) for ArXiv. SCCs are depicted with *different colors* depending on their sizes. *Right* the D-core matrix of the same data

On the other hand, this could be attributed to the fact that the DBLP dataset is not very well maintained thus lots of citations missing.

7 Conclusions

Cohesion and collaboration in graphs are cornerstone features for their evaluation, especially with the advent of large-scale applications such as the Web, social networks, citations graphs. The traditional way to look at graphs is through the authority/hub notion based on *per node* in/outlinks patterns. Other group evaluation measures do not take into account the directed nature of the aforementioned graphs. On the contrary, in this paper, we stress the importance of cohesion and collaboration among groups of nodes in the case of directed graphs (digraphs). The intuition is that subgraphs with many in/outlinks among their nodes convey a high degree of collaboration (adapted to the local application semantics). Thus, we defined D-core, a novel extension of the k-core concept to cover the directed graph case, as means of representing their collaborative features based on their robustness under degeneracy.

Capitalizing on the D-core structure, we define interesting and novel evaluation metrics and structures. Specifically, the D-core matrix for a graph, its frontier, and a wealth of metrics to evaluate (a) the robustness of the directed graph under degeneracy and (b) the dominant patterns of the graph with regard to inlinks/outlinks trade-offs.

Table 4 Authors in the D-core $\mathbf{DC}_{42,42}$ of the DBLP digraph

José A. Blakeley	Hector Garcia-Molina	Abraham Silberschatz	Umeshwar Dayal
Eric N. Hanson	Jennifer Widom	Klaus R. Dittrich	Nathan Goodman
Won Kim	Alfons Kemper	Guido Moerkotte	Clement T. Yu
M. Tamer Ö?zsu	Amit P. Sheth	Ming-Chien Shan	Richard T. Snodgrass
David Maier	Michael J. Carey	David J. DeWitt	Joel E. Richardson
Eugene J. Shekita	Waqar Hasan	Marie-Anne Neimat	Darrell Woelk
Roger King	Stanley B. Zdonik	Lawrence A. Rowe	Michael Stonebraker
Serge Abiteboul	Richard Hull	Victor Vianu	Jeffrey D. Ullman
Michael Kifer	Philip A. Bernstein	Vassos Hadzilacos	Elisa Bertino
Stefano Ceri	Georges Gardarin	Patrick Valduriez	Ramez Elmasri
Richard R. Muntz	David B. Lomet	Betty Salzberg	Shamkant B. Navathe
Arie Segev	Gio Wiederhold	Witold Litwin	Theo Härder
François Bancilhon	Raghu Ramakrishnan	Michael J. Franklin	Yannis E. Ioannidis
Henry F. Korth	S. Sudarshan	Patrick E. O'Neil	Dennis Shasha
Shamim A. Naqvi	Shalom Tsur	Christos H. Papadimitriou	Georg Lausen
Gerhard Weikum	Kotagiri Ramamohanarao	Maurizio Lenzerini	Domenico Saccà?
Giuseppe Pelagatti	Paris C. Kanellakis	Jeffrey Scott Vitter	Letizia Tanca
Sophie Cluet	Timos K. Sellis	Alberto O. Mendelzon	Dennis McLeod
Calton Pu	C. Mohan	Malcolm P. Atkinson	Doron Rotem
Michel E. Adiba	Kyuseok Shim	Goetz Graefe	Jiawei Han
Edward Sciore	Rakesh Agrawal	Carlo Zaniolo	V. S. Subrahmanian
Claude Delobel	Christophe Lécluse	Michel Scholl	Peter C. Lockemann
Peter M. Schwarz	Laura M. Haas	Arnon Rosenthal	Erich J. Neuhold
Hans-Jörg Schek	Dirk Van Gucht	Hamid Pirahesh	Marc H. Scholl
Peter M. G. Apers	Allen Van Gelder	Tomasz Imielinski	Yehoshua Sagiv
Narain H. Gehani	H. V. Jagadish	Eric Simon	Peter Buneman
Dan Suciu	Christos Faloutsos	Donald D. Chamberlin	Setrag Khoshafian
Toby J. Teorey	Randy H. Katz	Miron Livny	Philip S. Yu
Stanley Y. W. Su	Henk M. Blanken	Peter Pistor	Matthias Jarke
Moshe Y. Vardi	Daniel Barbará	Uwe Deppisch	H.-Bernhard Paul
Don S. Batory	Marco A. Casanova	Jürgen Koch	Joachim W. Schmidt
Guy M. Lohman	Bruce G. Lindsay	Paul F. Wilms	Z. Meral Özsoyoglu
Gultekin Özsoyoglu	Kyu-Young Whang	Shahram Ghandeharizadeh	Tova Milo
Alon Y. Levy	Georg Gottlob	Johann Christoph Freytag	Klaus Küspert
Louiqa Raschid	John Mylopoulos	Alexander Borgida	Anand Rajaraman
Joseph M. Hellerstein	Masaru Kitsuregawa	Sumit Ganguly	Rudolf Bayer
Raymond T. Ng	Daniela Florescu	Per-Åke Larson	Hongjun Lu
Ravi Krishnamurthy	Arthur M. Keller	Catriel Beeri	Inderpal Singh Mumick
Oded Shmueli	George P. Copeland	Peter Dadam	Susan B. Davidson
Donald Kossmann	Christophe de Maindreville	Yannis Papakonstantinou	Kenneth C. Sevcik
Gabriel M. Kuper	Peter J. Haas	Jeffrey F. Naughton	Nick Roussopoulos
Bernhard Seeger	Georg Walch	R. Erbe	Balakrishna R. Iyer
Ashish Gupta	Praveen Seshadri	Walter Chang	Surajit Chaudhuri
Divesh Srivastava	Kenneth A. Ross	Arun N. Swami	Donovan A. Schneider
S. Seshadri	Edward L. Wimmers	Kenneth Salem	Scott L. Vandenberg
Dallan Quass	Michael V. Mannino	John McPherson	Shaul Dar
Sheldon J. Finkelstein	Leonard D. Shapiro	Anant Jhingran	George Lapis

Table 5 Authors in the D-core $\mathbf{DC}_{83,83}$ of the of the ArXiv digraph

A. Klemm	S. Theisen	P.S. Aspinwall	B.R. Greene
B.R. Greene	D.R. Morrison	A. Giveon	M. Porrati
M. Porrati	E. Rabinovici	N. Seiberg	E. Witten
E. Witten	M. J. Duff	Andrew Strominger	Shyamoli Chaudhuri
Shyamoli Chaudhuri	Shamit Kachru	Cumrun Vafa	S. Ferrara
S. Ferrara	J. A. Harvey	J. Polchinski	A. Ceresole
A. Ceresole	R. D'Auria	Katrin Becker	Melanie Becker
Melanie Becker	James T. Liu	J. Rahmfeld	W. Lerche
W. Lerche	P. Mayr	M. Bershadsky	Jan Louis
Jan Louis	Sheldon Katz	M. Ronen Plesser	Michael R. Douglas
Michael R. Douglas	Gregory Moore	Micha Berkooz	Robert G. Leigh
Robert G. Leigh	John H. Schwarz	J. Distler	K. Intriligator
K. Intriligator	B. Craps	A. Van Proeyen	Julie D. Blum
Julie D. Blum	Kentaro Hori	Hirosi Ooguri	Ashoke Sen
Ashoke Sen	Ruben Minasian	Moshe Rozali	Mirjam Cvetic
Mirjam Cvetic	Burt A. Ovrut	K.S. Stelle	Daniel Waldram
Daniel Waldram	H. Lu	C.N. Pope	Klaus Behrndt
Klaus Behrndt	A. Zaffaroni	N.P. Warner	A. Kehagias
A. Kehagias	K. Sfetsos	Steven S. Gubser	D.Z. Freedman
D.Z. Freedman	A. Brandhuber	A. Karch	Per Kraus
Per Kraus	J. de Boer	E. Verlinde	H. Verlinde
H. Verlinde	A. Sagnotti	N. Dorey	Matthias Blau
Matthias Blau	T. Banks	W. Fischler	L. Susskind
L. Susskind	A. Fayyazuddin	Juan M. Maldacena	B. Pioline
B. Pioline	Edi Halyo	G.W. Gibbons	I.R. Klebanov
I.R. Klebanov	Mans Henningson	Kostas Skenderis	Cesar Gomez
Cesar Gomez	L. Girardello	Vijay Balasubramanian	G. Papadopoulos
G. Papadopoulos	P.K. Townsend	A.A. Tseytlin	Jerome P. Gauntlett
Jerome P. Gauntlett	E. Kiritsis	T.R. Taylor	Gary T. Horowitz
Gary T. Horowitz	Robert C. Myers	Donam Youm	E. Sezgin
E. Sezgin	Chris M. Hull	Anamaria Font	Yaron Oz
Yaron Oz	Zheng Yin	Ilka Brunner	Albion Lawrence
Albion Lawrence	John McGreevy	Joaquim Gomis	N. Nekrasov
N. Nekrasov	T. Tada	D. Minic	M.B. Green

We articulate the theoretical foundation of the proposed framework in a more principled manner, adopting valid terminology from related work. We extend the set of metrics and structures proposed in [28] with directed analogues of the notions of k-cores, k-core sequence, k-cells, and k-cell sequence. Moreover, we introduce the concept of a *Core Decomposition Forest* for the case of directed graphs, extending the similar concept for undirected graphs was introduced in [30] and used in [29]. We investigate the behavior of the new concepts and metrics in the case of synthetic preferential attachment graphs—dominant in real-world cases. The study is extended to various parameters values in an attempt to fit the features of the real-worlds graphs. In order to achieve this, we developed a multiparametric graph

Table 6 Authors in the D-core $\mathbf{DC}_{83,83}$ of the of the ArXiv digraph (continued)

M.B. Green	M. Gutperle	Petr Horava	Clifford V. Johnson
Clifford V. Johnson	Gabriel Lopes Cardoso	Dieter Lust	O. Bergman
O. Bergman	G. Lifschytz	Atish Dabholkar	Barton Zwiebach
Barton Zwiebach	Nathan Berkovits	R.R. Metsaev	S. Yankielowicz
S. Yankielowicz	Philip C. Argyres	Amihay Hanany	M. Bianchi
M. Bianchi	Duiliu-Emanuel Diaconescu	Ofer Aharony	Hong Liu
Hong Liu	P.S. Howe	P.C. West	Nakwoo Kim
P.C. West	Richard Corrado	Horatiu Nastase	Amanda W. Peet
Amanda W. Peet	M.R. Gaberdiel	Piljin Yi	Rajesh Gopakumar
Rajesh Gopakumar	C. Bachas	Akikazu Hashimoto	Marco Billo
Marco Billo	I. Pesando	Keshav Dasgupta	Sunil Mukhi
Sunil Mukhi	Joseph A. Minahan	D. Kutasov	Juan Maldacena
Juan Maldacena	Jeremy Michelson	C. Kounnas	R. Dijkgraaf
R. Dijkgraaf	Nissan Itzhaki	Jacob Sonnenschein	S. Gukov
S. Gukov	David Berenstein	Pei-Ming Ho	Angel M. Uranga
Angel M. Uranga	Sangmin Lee	J.G. Russo	E. Bergshoeff
E. Bergshoeff	M. de Roo	Soo-Jong Rey	Jung-Tay Yee
Jung-Tay Yee	Finn Larsen	Sandip P. Trivedi	I. Antoniadis
I. Antoniadis	E. Gava	K. S. Narain	R. Kallosh
R. Kallosh	J. Kumar	H.J. Boonstra	Kyungho Oh
Kyungho Oh	Radu Tatar	Mina Aganagic	Jaemo Park
Jaemo Park	David A. Lowe	Andrei Linde	Eric G. Gimon
Eric G. Gimon	L. E. Ibanez	Zurab Kakushadze	F. Quevedo
F. Quevedo	Ramzi R. Khuri	J. X. Lu	S. Sethi
S. Sethi	Sanjaye Ramgoolam	Sumit R. Das	Miao Li
Miao LI	Chris Hull	Washington Taylor	Curtis G. Callan
Curtis G. Callan	Samir D. Mathur	E. Martinec	Daniel Kabat
Daniel Kabat	BS Acharya	JM Figueroa-O'Farrill	Bernard de Wit
Bernard de Wit	Chong-Sun Chu	T. Ortin	Michael Dine
Michael Dine	Eva Silverstein	Laura Andrianopoli	Leonardo Rastelli
Leonardo Rastelli	Ulf H. Danielsson	Ori J. Ganor	Anastasia Volovich
Anastasia Volovich	H. Partouche	Barak Kol	Shmuel Elitzur
Shmuel Elitzur	A. Rajaraman	J.L.F. Barbon	Gabriele Ferretti
Gabriele Ferretti	Adel Bilal	S. P. de Alwis	Steven B. Giddings
Steven B. Giddings			

generator. Moreover, we offer an alternative visualization of the differential behavior of the graphs under concern with regard to the pace of graph size reduction for both the inlinks and the outlinks aspect and conduct an exhaustive study and interpretation of the in-/out-degree distributions in the synthetic preferential attachment datasets. Finally, we conduct an extensive experimental evaluation for scale-free/preferential attachment synthetic graphs as well as real-world large-scale directed graphs: the (English) Wikipedia—2004 edition, the ArXiv, and DBLP citation graphs. We computed and explored the respective D-cores matrices, frontiers, and metrics, and we derived interesting results and observations both

at the macroscopic (graph) and at the microscopic (node) level. We claim that the D-core concept and the relevant structures and metrics that we define in this paper constitute a framework of tools for efficient and valid evaluation of cohesiveness and collaboration in directed networks.

Future research will be focused on the following: (1) dealing with the temporal evolution of D-cores to capture collaboration evolution and (2) using D-cores as a preprocessing step in directed graph clustering. As D-cores are structures of high cohesion, we seek to research if it can be a beneficial pre-processing step for graph clustering, resulting in lower overall complexity with good quality results.

References

1. Alba RD (1973) A graph-theoretic definition of a sociometric clique. J Math Sociol 3:113–126
2. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2005) *k*-core decomposition: a tool for the visualization of large scale networks. CoRR, cs.NI/0504107
3. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2006) Large scale networks fingerprinting and visualization using the *k*-core decomposition. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in neural information processing systems, vol 18. MIT Press, Cambridge, pp 41–50
4. An Y, Janssen J, Milios EE (2004) Characterizing and mining the citation graph of the computer science literature. Knowl Inf Syst 6:664–678. doi:10.1007/s10115-003-0128-3
5. Bader GD, Hogue CWV (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformat 4:1–1
6. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
7. Barabási A-L, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology of the world-wide web. Phys A Stat Mech Appl 281:69–77
8. Batagelj V, Mrvar A (2002) Pajek—analysis and visualization of large networks. In: Mutzel P, Jünger M, Leipert S (eds) Graph Drawing, volume 2265 of Lecture Notes in Computer Science. Springer, Berlin, pp 8–11
9. Batagelj V, Zaversnik M (2002) Generalized cores. CoRR, cs.DS/0202039
10. Baur M, Gaertler M, Görke R, Krug M, Wagner D (2007) Generating graphs with predefined *k*-core structure. In: Proceedings of the European conference of complex systems (ECCS'07), Oct. 2007
11. Bollobas B, Borgs C, Chayes J, Riordan O (2003) Directed scale-free graph. In: Proceedings of 14th ACM-SIAM symposium on discrete algorithms, pp 132–139
12. Bollobás B, Riordan O (2004) The diameter of a scale-free random graph. Combinatorica 24:5–34
13. Bollobs B, Riordan O, Spencer J, Tusnády G (2001) The degree sequence of a scale-free random graph process. Random Struct Algorithms 18(3):279–290
14. Buckley PG, Osthus D (2001) Popularity based random graph models leading to a scale-free degree sequence. Discrete Math 282:53–68
15. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2006) MEDUSA—new model of internet topology using k-shell decomposition, arXiv:cond-mat/0601240
16. Charikar M, (2000) Greedy approximation algorithms for finding dense components in a graph. In: Approximation algorithms for combinatorial optimization (Saarbrücken), (2000) volume 1913 of Lecture Notes in Computer Science. Springer, Berlin, pp 84–95
17. Cooper C, Frieze A (2003) A general model of web graphs. Random Struct Algorithms 22:311–335
18. Diestel R (2005) Graph theory, volume 173 of Graduate texts in mathematics. Springer, Berlin
19. Dorogovtsev SN, Goltsev AV, Mendes JFF (2006) *k*-core organization of complex networks. Phys Rev Lett 96:040601
20. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of growing networks with preferential linking. Phys Rev Lett 85(21):4633–4636
21. Drinea E, Enachescu M, Mitzenmacher M (2001) Variations on random graph models for the web. Computer Science Group Harvard University, Cambridge
22. Erdős P (1963) On the structure of linear graphs. Israel J Math 1:156–160
23. Erdős P, Rényi A (1960) On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5:17–61
24. Fershtman M (1997) Cohesive group detection in a social network by the segregation matrix index. Social Netw 19:193–207

25. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
26. Frank KA (1995) Identifying cohesive subgroups. Social Netw 17:27–56
27. Freuder EC (1982) A sufficient condition for backtrack-free search. J Assoc Comput Mach 29(1):24–32
28. Giatsidis C, Thilikos DM, Vazirgiannis M (2011) D-cores: measuring collaboration of directed graphs based on degeneracy. In: ICDM, pp 201–210
29. Giatsidis C, Thilikos DM, Vazirgiannis M (2011) Evaluating cooperation in communities with the *k*-core structure. In: ASONAM. IEEE Computer Society, pp 87–93
30. Healy J, Janssen J, Milios E, Aiello W (2008) 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, Banff, Nov. 30–Dec. 1, 2008
31. Kandylas V, Upham S, Ungar L (2008) Finding cohesive clusters for analyzing knowledge communities. Knowl Inf Syst 17:335–354. doi:10.1007/s10115-008-0135-5
32. Kirousis LM, Thilikos DM (1996) The linkage of a graph. SIAM J Comput 25(3):626–647
33. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E (2000) Stochastic models for the web graph. In: Proceedings of the 41st annual symposium on foundations of computer science. IEEE Computer Society . Washington, DC, USA, p 57
34. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Extracting large-scale knowledge bases from the web. In: VLDB '99: proceedings of the 25th international conference on very large data bases. Morgan Kaufmann, San Francisco, pp 639–650
35. Luce D (1950) Connectivity and generalized cliques in sociometric group structure. Psychometrika 15:169–190
36. Matula DW (1968) A min-max theorem for graphs with application to graph coloring. SIAM Rev 10:481–482
37. Matula DW, Marble G, Isaacson JD (1972) Graph coloring algorithms. In: Graph theory and computing. Academic Press, New York, pp 109–122
38. Moody J, White DR (2007) Structural cohesion and embeddedness: a hierarchical concept of social groups. Am Sociol Rev 68(1):103–127
39. Papadimitriou S, Sun J, Faloutsos C, Yu PS (2008) Hierarchical, parameter-free community discovery. In: ECML/PKDD (2), pp 170–187
40. Pittel B, Spencer J, Wormald N (1996) Sudden emergence of a giant *k*-core in a random graph. J Combinatorial Theory Ser B 67(1):111–151
41. Seidman SB (1983) Network structure and minimum degree. Social Netw 5(3):269–287
42. Szekeres G, Wilf HS (1968) An inequality for the chromatic number of a graph. J Combinatorial Theory 4:1–3
43. Wasserman S, Faust K (1994) Social networks analysis: methods and applications. Cambridge University Press, Cambridge
44. Wuchty S, Almaas E (2005) Peeling the yeast protein network. Proteomics 5(2):444–449

Author Biographies

Christos Giatsidis is a graduate of the Department of Informatics of the Athens University of Economics and Business (A.U.E.B) and is currently a Ph.D. Student at LIX, École Polytechnique in the are of data mining with specialized interest in social networks and community graphs.

Dimitrios M. Thilikos studied Mathematics at the University of Patras, Greece, and received his Ph.D. from the Department of Computer Engineering and Informatics of the same university. He worked as a researcher at the Department of Information and Computing Sciences of Utrecht University, Netherlands, and at the School of Computer Science of the University of Waterloo, Canada. He has been an Associate Professor at the Computer Science Department of the Polytechnic University of Catalonia, Spain. Currently, he is an Associate Professor of the Department of Mathematics of the Athens University, Greece. His research interests are in Graph theory, Parameterized Complexity, and their algorithmic applications.

Michalis Vazirgiannis has conducted research in Frauhofer (Germany), INRIA/Paris, IBM, and Max Planck Institut für Informatik (Germany). He is recipient of the Marie Curie Intra-European and the ERCIM fellowships and of the French DIGITEO Chair grant. His research interests include Web Graph analysis & evolution monitoring, distributed machine learning algorithms, dimensionality reduction, and aspects of text retrieval and mining . Currently, he holds a Professor position at AUEB, Greece. He has participated in international research projects in the area of web mining. He has chaired and participated in many programme committees of international conferences in the areas of data bases, data mining/machine learning and the web.