D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy

Christos Giatsidis*, Dimitrios M. Thilikos‡, Michalis Vazirgiannis†*
*LIX – École Polytechnique, Palaiseau Cedex, France
Email: xristosakamad@gmail.com
†Department of Informatics, Athens Univ. of Economics & Business, Athens, Greece
Email: mvazirg@aueb.gr
‡Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece
Email: sedthilk@math.uoa.gr

***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 among its nodes. A large variety of measures have been proposed to evaluate the quality 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 real-world graphs such as Wikipedia and DBLP and report interesting results at the graph as well at node level.**

***Keywords*-Graph Mining; Community evaluation metrics; Degeneracy; Cores;**

I. Introduction

The Web, social network, and citation graphs form a context where the detection and evaluation of communities is a challenging task. The research methods in this area have mainly capitalized on the Hub/Authority concepts [21], [18] 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 sub-graph 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 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 are least k incident edges that are adjacent to vertices of the same subgraph (formal definitions follow in Section 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 [11] on evaluating collaboration in undirected graphs The paper contributions are the following:

- we extent the notion of k-core by introducing the D-core concept on graphs where the edges are directed. 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.
- 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 inlink/outlink trade offs 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 two real-world, large scale graphs: the (English) Wikipedia - 2004 edition, and the DBLP graph. We computed and explored the respective D-cores matrix, 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 form a set of

tools for efficient and valid evaluation of cohesiveness and collaboration in directed networks.

II. Related work

A thorough review on community detection in graphs is offered by Fortunato [8]. In that work techniques, methods, and data sets 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 recent literature, various metrics are proposed to evaluate the graph structure of a social network. Such are "Betweenness" [21], "Centrality" [18], Clustering coefficient (a measure of the likelihood that two associates of a node are associates themselves). A higher clustering coefficient indicates a greater "cliquishness", i.e. cohesion degree or density. Of special interest here is the eigenvector centrality – a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to nodes having a high score contribute more to the score of the node in question. Other measures include "path length" (i.e. distances between pairs of nodes in the network), "prestige/authority", a measure in directed graphs to describe a node's centrality and "radiality", a notion representing the capacity of an individual to rech 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], [7], [9] and the relational distance between nodes [14] are used as basic features for cohesion. The issues with these ideas are that -as it is also noted in [17]- 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 [17] 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.

In [13] 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 [3], 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. Finally, a fractional version of the k-core structure was introduced in [11] towards evaluating and detecting collaboration communities in bipartite graphs where the edges represent relations between different entities such as papers and authors.

III. 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.

A. 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. k-cores are fundamental structures in graph theory and their study dates back to the 60's [5], [15], [20]. The parameter of *degeneracy* appeared with several names such as width [16], linkage [10], and the coloring-number [4]. 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 [19] for random graphs generated by the Erdős-Rényi model [6]. As shown in [19], 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. For this, we introduce below some definitions.

B. 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, i.e. 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, i.e. edges in D with x as a tail. The *min-in-degree* and the *min-out-degree* of a digraph D are defined as

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

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 $\delta^{\text{in}}(F) \geq k$ and $\delta^{\text{out}}(F) \geq l$; if no such digraph exists then the (k,l)-D-core of D is the empty digraph.

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 $\mathbf{DC}_{k,l}(D)$, i.e. 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.

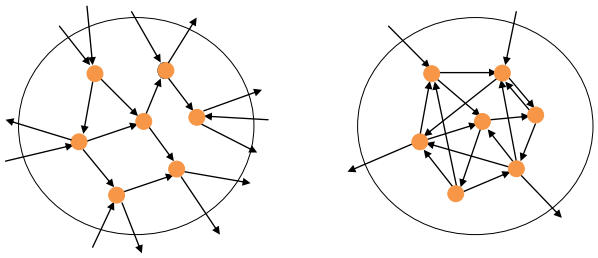

Figure 1. Two portions of a digraph. The one in the left does not contain any non-trivial (k,l)-core and the one in the right is a $(2,2)$-core.

The intuition behind (k,l)-D-cores is to find a subgraph 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 satisfy the same in-degree and/or out-degree requirements with respect to all the other community members (see Figure III-B 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*
$\mathbf{deg}^{\text{in}}_F(x) < k$ *or* $\mathbf{deg}^{\text{out}}_F(x) < l$,
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 $\mathbf{deg}^{\text{in}}_{D_i}(v_i) < k$ or $\mathbf{deg}^{\text{out}}_{D_i}(v_i) < l$.

The following Lemma on (k,l)-D-cores generalizes the classic min-max result of [16] (see also [10], [12]).

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 [10] for the non-directed case (see also [2]). In our implementation of this procedure, $\mathbf{DC}_{k,l}(D)$ is incrementally computed for all pairs of k and l.

C. 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.

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

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 [3] and [13]. A powerful extension of the classic notion of a k-core was given in [2] where the k-core is defined as a set of vertices where some general vertex property function is bounded. While the results in [2] 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 an 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.

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

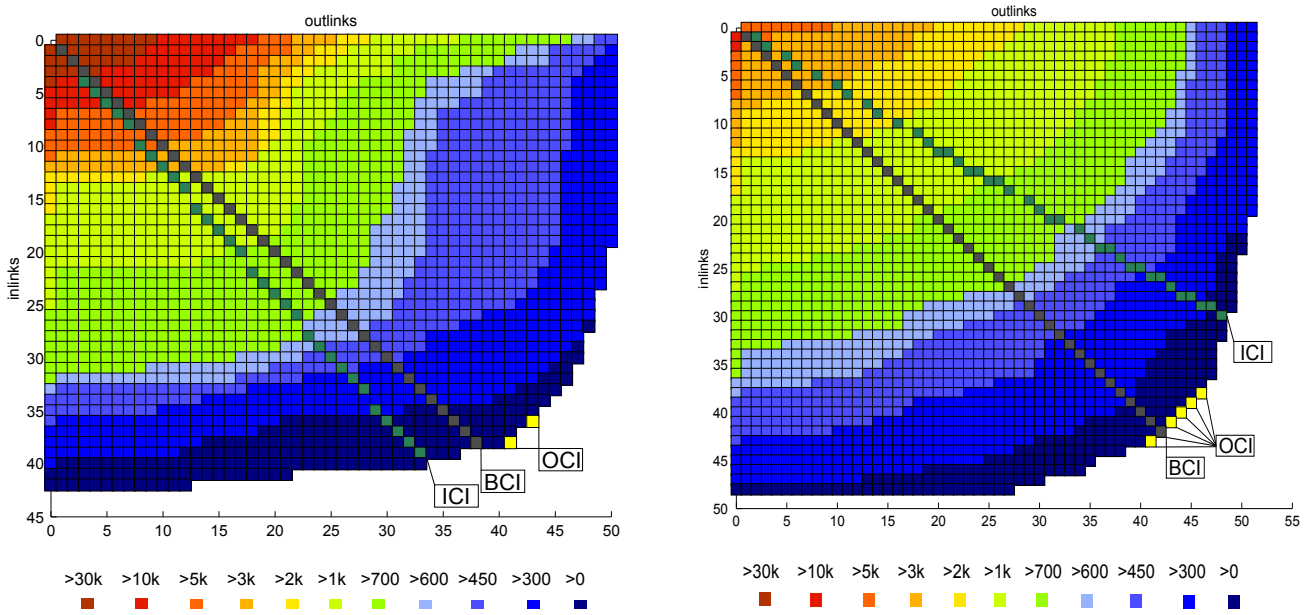

Figure 2. The D-core matrices of the Wikipedia 2004 digraph (upper) and the DBLP digraph (bottom)

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 $A_D(k, l) = (\mathbf{dc}_{k,l})_{k,l\in\mathbb{N}}$ that we call D-*core matrix of* D.

We identified this matrix 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. Each cell in this matrix $A_D(k, l)$ stores the size $(\mathbf{dc}_{k,l})_{k,l\in\mathbb{N}}$ of the respective $\mathbf{DC}_{k,l}$. The result is depicted in Figure 2. 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. According to Figure 2, 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 $\delta^*(D_{\text{Wiki}})$.

D. 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 sub-digraph 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 Figure 2 where the frontier appears as the dark color contour surrounded by 0 values.

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 subgraph 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 $\mathbf{DC}_{42,0}$.

IV. 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/out links 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 i.e. whether they contain more in-links or out-links. The other important issue is the robustness of the D-cores in terms of degeneracy. According to the previous definitions, this is proportional to the Manhattan distance of the extreme frontier points from the matrix origin, i.e. cell $(0,0)$. 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/out links. 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 out links.

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 in-links to out-links of the vertices in the digraph. Based on this we define the average H/A-angle of a digraph D as follows.

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

To make the above formula feasible, we excluded vertices with zero in or out links, i.e. 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 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 Figure 2 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

$$\begin{aligned} F(D) \quad = \quad & \{(k_{\max},0),(k_{\max},1),\ldots,(k_{\max},l_{\max}), \\ & (k_{\max}-1,l_{\max}),\ldots,(0,l_{\max}))\}. \end{aligned}$$

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 $[0,1]$.

The above definition implies that the robustness is essentially the 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, i.e. the degree of cohesion among the graph nodes – in terms of globally distributed in/out links.

A. 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,

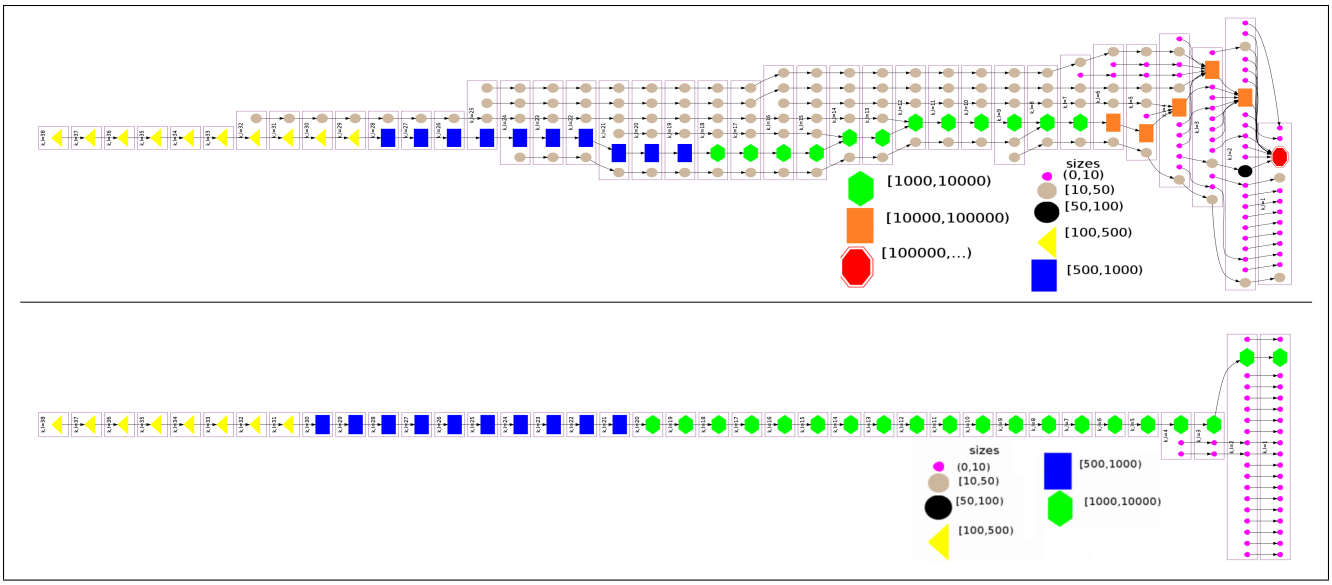

Figure 3. The SCCs sizes on the diagonal D-cores(i,i) and their hierarchical containment, Wikipedia 2004 (upper), DBLP (bottom).

	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/$\langle$(43,36) /47,66/$\rangle$228 $\langle$(41,38) 45,42/233$\rangle$	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

Table I
COLLABORATION INDICES VALUES FOR THE WIKIPEDIA GRAPH.

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 subsection IV-A, 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 four collaboration indices for a set $X \subseteq V(D)$ as del as its robustness are defined analogously as in previous sections. We omit the definitions due to lack of space.

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 collaboration index of X_1 will be at least the balanced/optimal/inherent collaboration index of X_2.

V. EXPERIMENTAL EVALUATION

In this section we present the experiments we performed applying the above algorithms and definitions on real-world data sets, obtaining valuable and convincing results.

A. 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 of 2004 and was extracted from a database dump containing the entire history of the encyclopedia; which can be found 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/. 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.

B. 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 subsection III-B, 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 (40x50) $\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 [2] for the case of non directed graphs.

C. 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 e.g. 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 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/ OCIACI, indices and the Robustness.

3. *Strongly Connected Components (SCCs)*: for each D-core $\mathbf{DC}_{i,i}$ – i.e. on the D-core matrix diagonal we computed the strong connected components. A *strong connected component* of a digraph D is a maximal sub-digraph where every two vertices are in a directed cycle. SCCs indicate groups of strong cohesiveness in the D-core. See Figure 3 for detailed view on the SCCs size evolution and hierarchical relationships as i, running along the D-core matrix diagonal, increases for both datasets considered. This hierarchy is depicted by a collection of rooted trees where the roots correspond to the strong connected components of the whole digraph (level zero components) and each level contains the strong connected components of the diagonal D-cores. Moreover each directed edge points from a strong connected component of some level to a super-digraph of it to the previous level.

4. *Frontiers for sets of entries*: We also computed the frontiers for single terms/authors for Wikipedia/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.

(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

Table II
THE THEMATIC FOCUS OF THE WIKIPEDIA SCCS FOR INCREASING DEGENERACY ALONG THE BCI AXIS.

D. 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 Figure 2) 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 Figure 4). For Wikipedia 2004

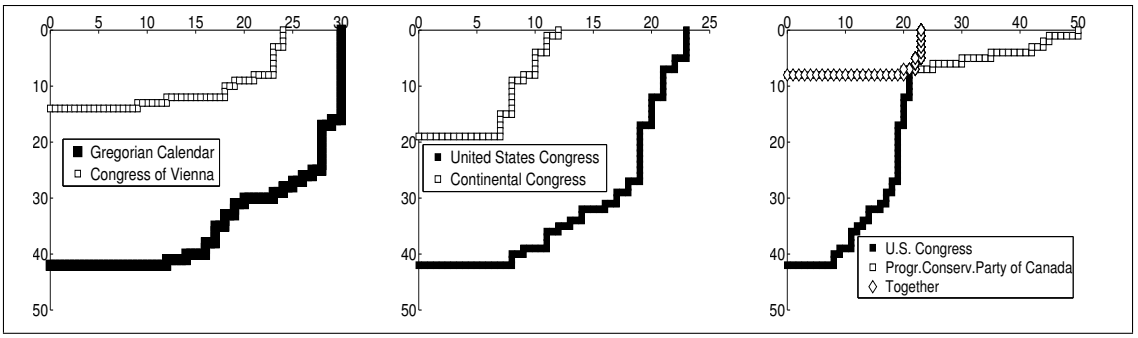

Figure 4. Selected term-pages and sets of term-pages frontiers from Wikipedia.

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 IV-A, 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 Figure 4 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 sub-figure, 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 the page is participating in D-cores with low degeneracy. 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 Figure 4 (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/out links) that 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 Figure 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 II. 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 1034 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 II, 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 $\mathbf{DC}_{i,i}$.

E. 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 Figure 2 bottom) as well as the number of strongly connected components (SCC's) in each of the D-cores $\mathbf{DC}_{i,i}$ – i.e. on the diagonal (see Figure 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.

	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

Table III
COLLABORATION INDICES VALUES FOR THE DBLP DIGRAPH

For the case of the DBLP digraph, the value of BCI is 42 (see Table III a summary of all indices values) while the respective D-core $\mathbf{DC}_{42,42}$ contains 188 nodes (see the lower part of Figure 2). 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 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 Figure 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 ($<$10 nodes) – excluding one significantly larger sized SCC ($>$1000 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[1] 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 data set 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 Figure 5 (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/out links 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 Figure 5 (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.

VI. 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 etc. The traditional way to look at graphs is though the authority/hub notion based on *per node* in/out links 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/out links 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.

[1]The names of these authors can be accessed at: http://www.db-net.aueb.gr/michalis/DCORE_authors.html

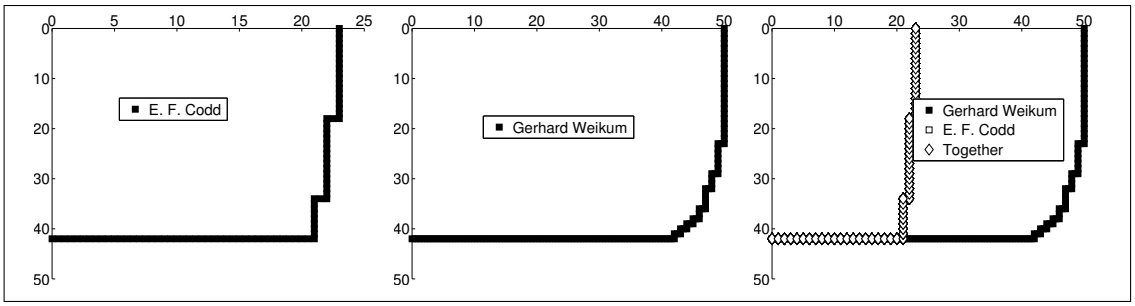

Figure 5. Representative authors D-core frontier from the DBLP digraph

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 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.

We evaluate these structures and metrics in large scale real world graphs (i.e. a Web and a citation graph). The results are interesting and justify the D-core structure and related metrics as a new framework for evaluating collaboration and cohesion in applications where directed graphs are the dominant structures.

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] R. D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology 3:113-26*, 1973.

[2] V. Batagelj and M. Zaversnik. Generalized cores. *CoRR*, cs.DS/0202039, 2002.

[3] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Approximation algorithms for combinatorial optimization (Saarbrücken, 2000)*, volume 1913 of *Lecture Notes in Comput. Sci.*, pages 84–95. Springer, Berlin, 2000.

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

[5] P. Erdős. On the structure of linear graphs. *Israel J. Math.*, 1:156–160, 1963.

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

[7] M. Fershtman. Cohesive group detection in a social network by the segregation matrix index. *Social Networks 19:193-207*, 1997.

[8] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.

[9] K. A. Frank. Identifying cohesive subgroups. *Social Networks 17:27-56*, 1995.

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

[11] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, pages 87–93, 2011.

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

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

[14] D. Luce. Connectivity and generalized cliques in sociometric group structure. *Psychometrika 15:169-90*, 1950.

[15] D. W. Matula. A min–max theorem for graphs with application to graph coloring. *SIAM Reviews*, 10:481–482, 1968.

[16] D. W. Matula, G. Marble, and J. D. Isaacson. Graph coloring algorithms. In *Graph theory and computing*, pages 109–122. Academic Press, New York, 1972.

[17] J. Moody and D. R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review 68(1):103-127.*, 2007.

[18] S. Papadimitriou, J. Sun, C. Faloutsos, and P. S. Yu. Hierarchical, parameter-free community discovery. In *ECML/PKDD (2)*, pages 170–187, 2008.

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

[20] G. Szekeres and H. S. Wilf. An inequality for the chromatic number of a graph. *J. Combinatorial Theory*, 4:1–3, 1968.

[21] S. Wasserman and K. Faust. *Social Networks Analysis: Methods and Applications*. Cambridge: Cambridge University Press., 1994.