Visual Exploration of Collaboration Networks based on Graph Degeneracy

Christos Giatsidis
LIX, École Polytechnique
Palaiseau Cedex, France
xristosakamad@gmail.com

Klaus Berberich
LIX, École Polytechnique
Palaiseau Cedex, France
klaus@berberi.ch

Dimitrios M. Thilikos
Department of Mathematics,
National & Kapodistrian
University of Athens
Athens, Greece
sedthilk@math.uoa.gr

Michalis Vazirgiannis
Department of Informatics,
Athens University of
Economics and Business
Athens, Greece
mvazirg@aueb.gr

ABSTRACT

We demonstrate a system that supports the visual exploration of collaboration networks. The system leverages the notion of fractional cores introduced in earlier work to rank vertices in a collaboration network and filter vertices' neighborhoods. Fractional cores build on the idea of graph degeneracy as captured by the notion of k-cores in graph theory and extend it to undirected edge-weighted graphs. In a co-authorship network, for instance, the fractional core index of an author intuitively reflects the degree of collaboration with equally or higher-ranked authors. Our system has been deployed on a real-world co-authorship network derived from DBLP, demonstrating that the idea of fractional cores can be applied even to large-scale networks. The system provides an easy-to-use interface to query for the fractional core index of an author, to see who the closest equally or higher-ranked co-authors are, and explore the entire co-authorship network in an incremental manner.

Categories and Subject Descriptors

H.3.3 [**Information Search & Retrieval**]

General Terms

Algorithms, Experimentation.

Keywords

Collaboration networks, graph degeneracy, fractional cores.

KDD'12, August 12–16, 2012, Beijing, China.

1. INTRODUCTION

Our focus in this work is on providing tools to visually explore collaboration networks. Such networks arise in different contexts including science, entertainment and society. Co-authorship networks, as a concrete example from science, encode researchers' collaborations on joint publications. Given the large scale of such networks (e.g., the co-authorship network derived from DBLP that we use for our demo consists of more than 0.8M vertices and 4.4M edges), tools to explore them must provide a clutter-free view on vertices and their neighborhoods. This mandates meaningful ways to rank and filter vertices.

To these ends, we build on the idea of *graph degeneracy* that is tightly connected with the concept of k-cores. In a nutshell, the k-core of an undirected graph is its largest subgraph in which every vertex has a degree of at least k. The concept of k-cores is fundamental in graph theory and has been investigated since the 1960's [4,8,10]. The existence of k-cores of large size in sufficiently dense graphs, for instance, has been theoretically studied by [2,9,11] for random graphs generated by the Erdős-Rényi model [5].

In earlier work [6], we proposed the notion of fractional cores as an extension of the k-core concept to undirected edge-weighted graphs. This extension takes into account not only the number but also the strength of connections that a vertex in the graph has. When looking at a co-authorship network (e.g., the one derived from DBLP as our showcase application), we thus take into account both the number of collaborations that authors have as well as their contribution to each of them. Our demonstrated system leverages the notion of fractional cores to rank and filter vertices in the network. It is available at the following URL:

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

We stress that the visualization of the hop-1 coauthoring community we propose conveys much more meaning and information than the simple co-author graphs presented in other approaches such as the *Co-author Graph* at `http://academic.research.microsoft.com/` or the respective in Arnet Miner at `http://arnetminer.org/`). More

specifically, in our case the visualization depicts the neighbors of authors in the DBLP co-authorship network that are *inside* the highest index core they belong to. This implies that the appearing co-authors is the subset of all coauthors that also belong to the current or better k-core. Thus these authors are those that are at least as collaborative as the one we are concerned. Therefore this graph is the best co-authorship community of the author bearing as well weights on the edges. These edges indicate the amount of the co-authorship effort shared among the authors at the ends of the edge.

Organization. The rest of this paper unfolds as follows: In Section 2, we describe the concept of fractional cores and explain how it can be used to rank and filter vertices in collaboration networks. Section 3 describes the architecture of our system. Our demo is outlined in Section 4. Finally, Section 5 lays out other potential application scenarios.

2. FRACTIONAL K-CORE RANKING

We now describe the notion of fractional cores and explain how it can be used to rank and filter vertices in a collaboration network.

Collaboration Networks are formally represented as follows: Consider a bipartite graph

$$G = (A, P, E)$$

where A is the set of authors, P is the set of papers, and E is a set of edges. Each edge $\{x, y\}$ (where $x \in A$ and $y \in P$) expresses the fact that x is one of the authors of paper y. We also assume that all papers are written by at least two authors. This is a safe assumption, given that single-author papers can be ignored for our application. Given G, we construct an edge-weighted graph $(H_G, \mathbf{w})$ whose vertex set is A and where an edge connects two authors if and only if they appear as co-authors in at least one paper. We call $(H_G, \mathbf{w})$ the *co-authorship graph* of G. The mere existence of some edge H_G does not convey how intensively the two authors have collaborated in the past. Therefore, we set up a weighing function $\mathbf{w}$ that assigns a weight $w(e)$ on each edge $e = \{x, x'\}$ of H_G according to the following scheme:

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

where $N_G(x)$ is the set of neighbors in G of the vertex x, i.e., the set of papers that were authored by x. Notice that in the above formula, we implicitly assume that all authors of a paper have put equal effort in its preparation. Given that we have assigned a weight to each of the edges of G, we can extend the notion of the degree of a vertex x by defining:

$$\mathbf{deg}_{G,\mathbf{w}}(x) = \sum_{e \in E(x)} \mathbf{w}(e). \quad (1)$$

Fractional k-Cores. Let $\mathbb{Q}^+$ be the set of all positive rational numbers. Given a $k \in \mathbb{Q}$, we define the *fractional k-core* of $(H_G, \mathbf{w})$ as the maximum size subgraph of H such that for each of its vertices, the sum of the weights of its incident edges is no smaller than k. It is easy to prove that this graph is unique. For this reason it can be computed easily by discarding from H_G vertices of fractional degree less than k until no such vertices exist anymore. We refer to this procedure as the *Trim* procedure. For more details and pseudo-code of this procedure, the reader is referred to [6]. The procedure runs in $O(n+m)$ time where n is the number of vertices and m is the number of edges in H_G [1].

Ranking. To determine a ranking of vertices in the graph (i.e., authors in our case), we consider the infinite sequence $\mathcal{G} = G_0, G_1, \ldots,$ recursively defined as follows: $G_0 = G, h_0 = 0$, and for $i > 0$, $G_i = Trim(G_{i-1}, h_{i-1})$ where $h_i = \delta(G_i, \mathbf{w}_{G_i})$ refers to the minimum weighted degree according to (1) in the graph G_i. Then, the *fractional core sequence* of an edge-weighted graph $(G, \mathbf{w})$ is the prefix of $\mathcal{G}$ that contains all the non-empty graphs of $\mathcal{G}$. Notice that this sequence divides the original graph into "levels of cohesion". That is, as we increase h_i the corresponding G_i contains fewer vertices – corresponding to those authors who have collaborated more often and/or more intensively. We define the *fractional core index* of a vertex x to be the maximum h_i for which x belongs to the fractional h_i-core G_i.

3. SYSTEM ARCHITECTURE

Figure 1 depicts the overall architecture of our system, which is the subject of this section. For our demonstration, we use a co-authorship network derived from the DBLP bibliographic dataset[1]

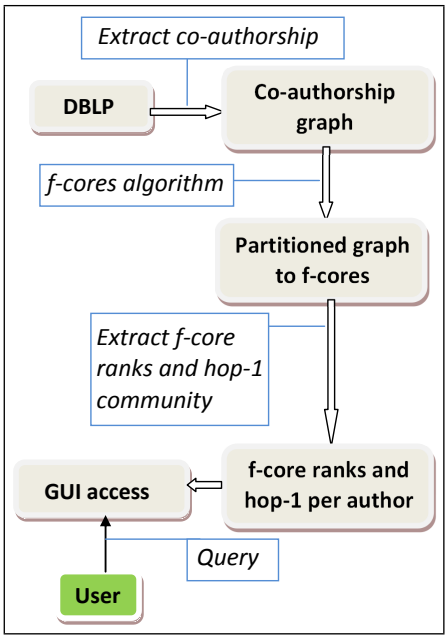

Figure 1: System Architecture

Going from top to bottom in Figure 1, we first convert the bibliographic dataset into an undirected edge-weighted co-authorship graph as described in Section 2. If two authors have co-authored a paper, this graph contains an undirected edge with a weight according to (1).

Next, as a one-time pre-computation, we iteratively compute the fractional core sequence of this graph. That is, we repeatedly invoke the *Trim* procedure, removing more and more vertices from the graph and thus implicitly partition it. When we remove a vertex from the graph, we keep track

[1]Freely available at `http://dblp.uni-trier.de`.

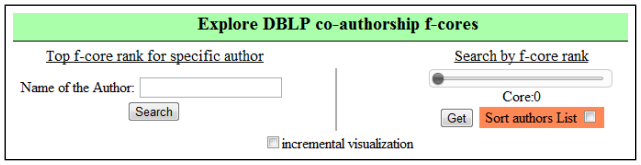

Figure 2: Main input interface

of its fractional core index and its hop-1 neighborhood consisting of immediate neighbors in the remaining graph. This information is stored in a relational database, so that it can be retrieved efficiently by our system at runtime. Note that, despite of the large scale of the DBLP co-authorship network, we were able to run the entire computation on a commodity notebook equipped with a 2-core CPU and 2 GB of main memory – an indication that fractional cores are computationally lightweight and can be applied to large-scale collaboration networks.

Users interact with our system through a web-based GUI that was implemented building on the Dracula Graph javascript library and can be accessed at the aforementioned URL. Its functionality, described in more detail in the following Section 4, includes showing rankings of authors by their fractional core index but also an interactive visualization of an author's neighborhood, displaying only those co-authors with an equal or higher fractional core index to provide a clutter-free view on the collaboration network.

4. DEMONSTRATION

The main user interface of our system is shown in Figure 2, offering users two options to interact. On the right, a slider allows users to select a threshold on the fractional core index and browse through qualifying authors. On the left, an input box allows users to search for a specific author by name – both exact match and fuzzy match are supported here. Either way, once an author has been selected, the system shows the fractional core index of the author together with a visualization of the surrounding hop-1 neighborhood, i.e., the author's closest co-authors who have at least an equally high fractional core index. Figure 3 shows an example hop-1 neighborhood (in this case for Michalis Vazirgiannis). From the visualization, users can see author's fractional core index (here 10.6) and his "tightest" collaborators each linked with a weighted edge indicating the "strength" of their partnership.

On this initial star-like graph the user can explore the surrounding authors by clicking on the "Find" function that appears when the mouse hovers over the author's vertex. If the "incremental visualization" option is not activated the result of the "Find" function is a fresh star-like graph centered around the newly selected author. Otherwise, if the option is activated, the user gets to explore the intersection of the two authors' (the original and the newly selected one) strongest collaborators. This "incremental" visualization can continue for multiple steps to reveal an increasingly broader and possibly highly interconnected community around an author. Figure 5 demonstrates this functionality building on our earlier example. Here, the star-like graph from Figure 3 was expanded by selecting Timos K. Sellis from the surrounding hop-1 neighborhood of collaborators. Interestingly, as can be observed from the figure, there is an overlap between the "tightest" collaborators of the two authors,

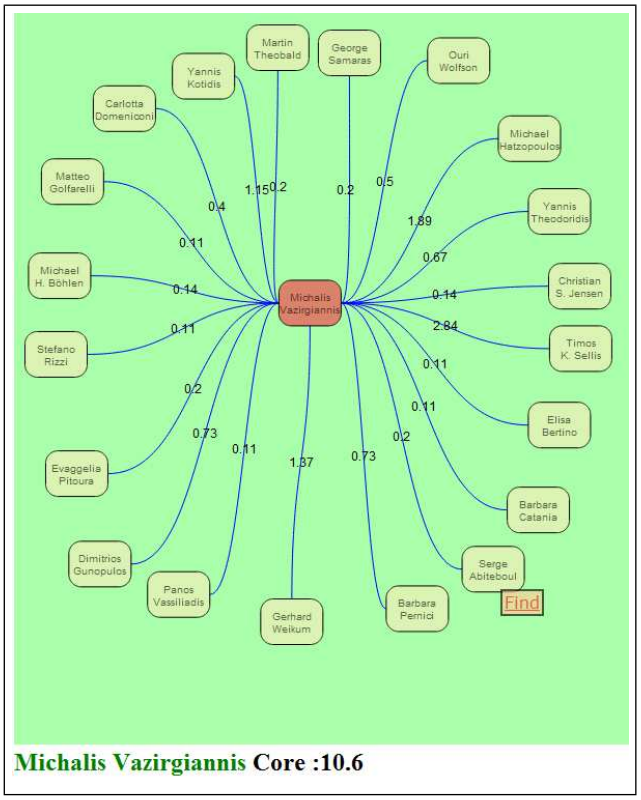

Figure 3: Output for the query "Vazirgiannis"

revealing a much richer view on the community that they belong to.

When the user selects a threshold for the fractional core index with the slider shown in Figure 2, using the "incremental" function more than one author can be selected and the user can see how different communities form within the same fractional core from the interconnection of the authors within that core. Figure 4 shows a small example based on authors from the 19.6-core – it is interesting to see the selected authors belong to two disconnected communities.

We will encourage conference attendants to interact with the system and use it to explore the(ir) DBLP co-authorship network. Until then, our system can be tried out online at the URL mentioned above. A screencast of the above test cases on our demo can be found at

`http://www.youtube.com/watch?v=XnzZqvzd0yo`

5. APPLICATION SCENARIOS

What are other application scenarios where a visual exploration of collaboration networks, as provided by our system, can provide useful insights?

Bibliographic data and measures derived therefrom (e.g., the H-Index [7] and G-Index [3]) nowadays play a big role in *academic hiring*. Measures like the aforementioned one have focused on citations and sought to capture a candidate's scientific impact. As networking skills and the ability to work in teams become more important, even in academia, our system provides the means to inspect a candidate in these regards. By relying on the notion of fractional cores, our system automatically zooms in on the connections to peers

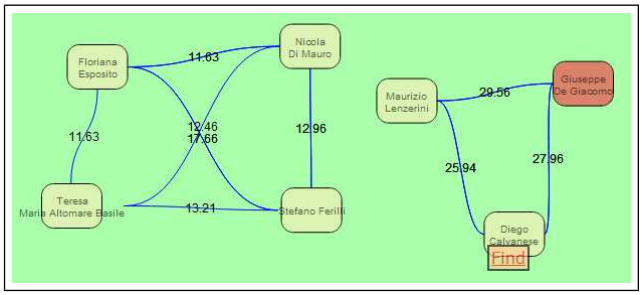

Figure 4: Example of browsing the 19.6-core with the "incremental" function activated

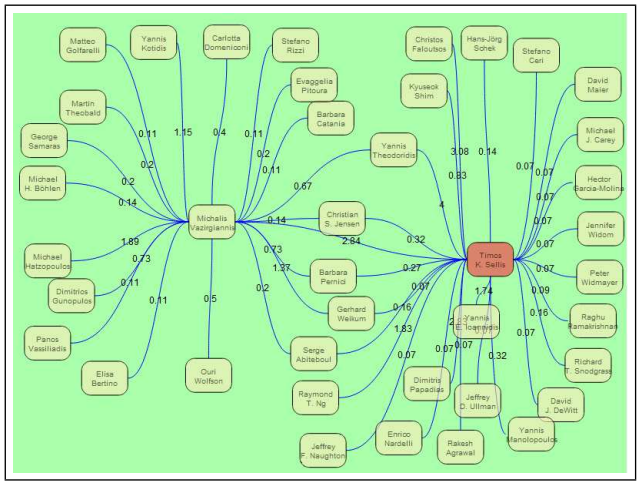

Figure 5: Example of browsing the hop-1 neighborhood with the "incremental" function activated

and more senior scientists and, as a consequence, provides a clutter-free view on the candidate's collaborations. Thus, collaborations that are ephemeral or with less prolific individuals (e.g., students who left academia after graduating) are automatically filtered out thanks to the use of fractional cores. We foresee similar applications for *human resources* and *team building* in industry, where collaboration networks can be derived from other datasets, e.g., patents (based on the listed inventor names), e-mails (based on recipient lists) or records about who worked together on which projects in the past. We believe that a visual exploration of collaboration networks can also turn out insightful for other collaboration networks, for example, those of actors who played in the same movie (e.g., derived from a movie database such as IMDB[2]) or politicians who co-signed a petition or participated in other joint activities.

The ideas behind our system are not limited to collaboration networks but can be applied to any dataset from which undirected weighted graphs can be derived in a sensible manner. This includes various facets of social networks with ties that signify, for instance, joint interests (based on group memberships) or reciprocal activities. Also in other contexts, the ability to visually explore large-scale datasets becomes more important in face of the ongoing data deluge.

[2]`http://www.imdb.com`

Some of the now available data has natural interpretations as graphs, for example, RDF datasets like those connected in the linked data cloud[3]. We expect that a visual exploration of such datasets can greatly profit from the use of fractional cores that, as in the applications above, help to get a clutter-free view on the data that focuses on essential highly-connected data items. Deploying our system on such datasets and evaluating its usefulness here is part of our ongoing work.

6. ACKNOWLEDGMENTS

All the authors of this paper are generously supported by the DIGITEO Chair grant LEVETONE in France

7. REFERENCES

[1] V. Batagelj and M. Zaversnik. An $o(m)$ algorithm for cores decomposition of networks. *CoRR*, cs.DS/0310049, 2003.

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

[3] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

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

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

[6] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *ASONAM*, pages 87–93, 2011.

[7] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569, 2005.

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

[9] 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.

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

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

[3]`http://linkeddata.org`