Abstract

A k-core of a graph is a maximal subgraph whose nodes have degree at least k in that subgraph. A node can appear in multiple k-cores. The core number of a node is the largest k among its k-cores. The degeneracy of a graph is the maximum core number among its nodes. Applications of k-cores and core numbers are numerous, including community detection and modeling spread on a network. We present a study of k-cores on the DBLP co-authorship graphs over 30 years starting at 1980. Our key observations are as follows. (1) Over time, collaborations, in terms of co-authorships, have increased dramatically. From 1980 to 2010, the number of papers has increased 38 times and the number of authors has increased 64.6 times, but the number of co-authorship relations has increased 153.7 times and the degeneracy of the co-authorship graph has increased 7.1 times. Specifically, the 1980 graph has a degeneracy of 11 (which means that it has one maximal subgraph where the authors have at least 11 co-authors), while the 2010 graph has a degeneracy of 78 (which means that it has one maximal subgraph where the authors have at least 78 co-authors). (2) The k-cores with the largest k values in the DBLP co-authorship graphs are often cliques (representing a specific publication). (3) We observe two types of authors. The first type consists of authors with large core numbers. They have relatively few papers but many co-authors per paper. The second type consists of high-degree authors who have more papers but few co-authors per paper.

Introduction

A graph’s k-core [1] is its maximal subgraph such that all the nodes in the subgraph have degree at least k in that subgraph. A node may be part of several k-cores. The core number of a node is the largest value k such that the node is in a k-core. For example, a graph in which all nodes have degree at least one is a 1-core. A k-core can be a clique. For example, a 5-core containing six nodes is a clique. In a graph, k-cores with large k values (a.k.a. high k-cores) generally represent the more densely connected parts of that graph. The degeneracy of a graph is the maximum k for which that graph has a k-core. Equivalently, the degeneracy of a graph is the maximum core number among its nodes. K-cores have wide applications in problems such as community detection, selecting nodes for network experiments, and modeling spread of information. For example, in community detection it is helpful to know the set of nodes that are part of a k-core with a large value of k [2]. Another example is the use of k-core decomposition and core numbers to understand the structure of protein interaction networks [3]. In this work, we investigate the evolution of k-cores in co-authorship graphs. Our data spans 30 years and is from DBLP, a computer science bibliography.

Experimental Setting and Initial Observations

We consider the DBLP co-authorship graphs, one per year from 1980 to 2010. In each (unweighted) graph, the nodes represent authors and the edges represent co-authorship relations between authors in the given year. Figure 1(a) depicts the number of authors (i.e., nodes), the number of co-author relations (i.e., edges) and the number of papers published from 1980 to 2010. Over time, all measures have increased. For example, the 1980 DBLP graph has 4,452 authors and 4,778 co-authorship relations based on 5,346 papers. In contrast, the 2010 DBLP graph has 287,440 authors and 734,357 co-authorship relations based on 203,386 papers.

For each DBLP graph, we compute the k-cores and the core numbers of the nodes. Figure 1(b) shows the trend in the degeneracy values of DBLP graphs over time. The increasing degeneracy observed corresponds to an increase in collaborations among authors. k-cores with large k values in the DBLP graphs are usually from papers with many co-authors. In the subsequent sections, we will discuss this point in further detail. Figures 1(c) and 1(d) show the distributions of core numbers in the 1980 and 2010 DBLP graphs, respectively. These plots show that the majority of

---

1 http://dblp.uni-trier.de/

2 A co-authorship relation in a specific year is an edge between two authors who have co-authored at least one paper in that year.
the nodes belong to $k$-cores with small $k$ values. Although the fraction of nodes in the $k$-cores with large $k$ values is small, they represent densely connected network of authors (e.g., many authors on a single paper). For instance, in the 2010 DBLP graph (with 287,440 authors), 84.8% of the authors are in the 2-core. The number of authors in a $k$-core drops quickly as $k$ increases. The 10-core in the 2010 DBLP graph only has 3.9% of the authors; and the 41-core (which is the $k$-core with the largest $k$ value that is not a clique) only has 0.11% of the authors. The $k$-core with the maximum $k$ value is a clique of 79 nodes (or a mere 0.03% of the nodes). We observed similar patterns in the other DBLP graphs.

![Graph](image)

(a) DBLP graphs (per year): As the numbers of papers have increased over time, so have the sizes of the co-authorship graphs.

(b) Degeneracy (i.e., the largest $k$ for which that graph has a $k$-core). 2006 has the largest degeneracy of 118, resulting from a single paper with 119 authors.

(c) Distribution of core numbers in the 1980 DBLP graph. The axes are in log scale. The maximum core number (a.k.a. degeneracy) is 11.

(d) Distribution of core numbers in the 2010 DBLP graph. The axes are in log scale. The maximum core number (a.k.a. degeneracy) is 78

**Figure 1:** (a) Sizes of the DBLP graphs over time. (b) Degeneracy values for the DBLP graphs over time. (c), (d) Distributions of core numbers in the 1980 and 2010 DBLP graphs. Observations: (1) We observe an increase in collaborations (depicted by the growth in the sizes of the DBLP graphs and in their degeneracy values over time), which is likely due to the increase in digital communication. (2) The distributions of core numbers among nodes in DBLP graphs are skewed. The bulk of the nodes belong to $k$-cores with small $k$ values.

Observations from Figure 1 lead us to ask the following questions:

- How are the nodes in $k$-cores with large $k$ values different from the set of high-degree nodes?
- Are $k$-cores with large $k$ values similar in structure?
- What kind of collaborations result in $k$-cores with large $k$ values?
In the following section, we present our findings and discuss how they relate to collaboration networks.

**More Observations and Discussion**

**Degeneracy, clique and non-clique k-cores with large k values.** Recall that degeneracy is defined as the largest $k$ among all the $k$-cores in the graph. The *degeneracy-core* is the $k$-core with the largest $k$ value. A large degeneracy value is the result of papers with many common authors.

For the DBLP graphs, Figure 2(a) depicts the degeneracy values and the largest $k$ values for non-clique $k$-cores over time. Except for the 1982 DBLP graph (where the blue and green lines in Figure 2(a) meet), the degeneracy-core is a clique, resulting from a single paper with many authors. For example, the 2006 DBLP graph has a degeneracy-core that is a clique of 119 nodes [4]. Also in Figure 2(a), we observe that many non-clique $k$-cores have large $k$ values. For example, the 2002 DBLP graph has a non-clique 64-core with 115 nodes. Thus, $k$-cores with large $k$ values (a.k.a. high $k$-cores) represent either a single paper with many authors (clique $k$-core) or a set of authors who co-author multiple papers with each other in small groups (non-clique $k$-core).

Figure 2(b) depicts the fraction of nodes in the degeneracy-core and in the highest non-clique $k$-core. Over time, although the sizes of the graphs increase, as shown in Figure 1(a), and the degeneracy values increase, as shown in Figure 1(b), the fraction of authors in the $k$-cores with the largest $k$ values decreases; see Figure 2(b). Thus, for the (DLBP) collaboration networks, our observation implies that over time the increase in the number of authors outpaces the fraction of authors that are co-authoring multiple papers in small groups.

![Figure 2](image_url)

(a) The value of $k$ in $k$-cores with the largest $k$ values over time. The blue line depicts the clique $k$-cores. The green line depicts non-clique $k$-cores.

(b) The fraction of nodes in $k$-cores with the largest $k$ values over time. The blue line depicts the clique $k$-cores. The green line depicts non-clique $k$-cores.

**Figure 2:** (a) The values of $k$ and (b) the fractions of nodes in the degeneracy-cores and in the highest non-clique $k$-cores of the DBLP graphs over time. Observations: (1) Except for 1982, the degeneracy-core in each graph is a clique, which implies that the degeneracy-core is a result of a single paper with many authors (typically from the same research lab). The highest non-clique $k$-cores are the result of several papers among overlapping sets of authors, typically across different research labs. (2) Over time, the increase in the number of authors, shown in Figure 1(a), outpaces the fraction of authors that are co-authoring multiple papers in small groups.

**Distributions of core numbers.** Figure 3 depicts heatmaps of degree by core numbers for the 2000 and 2010 DBLP graphs. We observe the following:

- There are many low-degree nodes (i.e., authors with few collaborators). These authors also have low core numbers since the core number of a node cannot be larger than its degree.
- In Figures 3(a) and 3(c), we observe cone-line patterns that are the result of two distinct forms of collaborations, which can be explained with Figures 3(b) and 3(d). Consider Figure 3(c). The upper end of the cone represents nodes with high core numbers. These nodes are typically authors in papers with many co-authors. Thus, as shown in Figure 3(d), their degree is a result of a few papers with many authors. However,
the lower end of the cone represents nodes with high degrees and lower core numbers. These nodes are authors with many collaborators and many papers, but only a few co-authors on each paper. Thus, the highest degree nodes do not necessarily have high core numbers.

- Figures 3(b) and 3(d) have been annotated with the paper titles that led to $k$-cores with the largest $k$ values. We observe the degrees and core numbers of some well-known data miners. Note that the points towards the far right-bottom corner of Figures 3(b) and 3(d) are authors who, according to the DBLP data, have a high number of published papers on a diverse set of topics. Examples of such authors in the 2010 DBLP graph include Wei Wang (whose number of co-authors is 204), Yan Zhang (whose number of co-authors is 228), and Wei Liu (whose number of co-authors is 228).

**Recurring nodes in the degeneracy-core.** We observe that the Jaccard index of the authors in the degeneracy-cores of consecutive years is less than 0.05. Recall that the degeneracy core is the $k$-core with the largest $k$ value. Thus, we observe that large but different groups of authors publish papers with many authors.

**Conclusion**

$K$-cores are known to be important structures with many applications. In this work, we present a case study of $k$-cores in affiliation graphs over time, by considering the DBLP co-authorship graphs from 1980 to 2010. We observe the following:

- $K$-cores have skewed distributions in terms of their sizes (i.e., number of nodes in the $k$-core).
- The degeneracy-core (i.e., the $k$-core with the largest $k$) in DBLP graphs is very often a large clique resulting from a single paper with many co-authors (e.g., a 2006 paper with 119 authors [4]).
- Non-clique $k$-cores with large $k$ values are the result of several papers among overlapping sets of authors, typically from different research labs.
- Regardless of whether a $k$-core is a clique or not, the value of the largest $k$ increases over time.
- High-degree nodes are authors with many papers and few co-authors per paper; the reverse is true for nodes in $k$-cores with large $k$ values.
- The overlap between authors in the degeneracy-cores of consecutive years is very small (less than 0.05).

**References**


3 http://dblp.uni-trier.de/pers/hd/a/Avino:S=
The 2000 DBLP graph with a degeneracy of 48. The cell color in (a) is the log scale of the number of authors. The cell color in (b) is the log scale of the average number of papers published.

The 2010 DBLP graph with a degeneracy of 78. The cell color in (c) is the log scale of the number of authors. The cell color in (d) is the log scale of the average number of papers published.

Figure 3: Heatmaps of node degree by core number. Observations: (1) The distribution of core numbers in DBLP follows a similar pattern across time. (2) A large fraction of the nodes have low degree and hence have low core numbers. (3) As degree increases, so does the range of core numbers. (4) The cone-like patterns in (a) and (c) are due to two types of collaborations, which can be explained by (b) and (d). First, the upper end of the cone represents authors who publish fewer papers with a large number of co-authors in each paper. Second, the lower end of the cone represents authors who publish many papers with few co-authors per paper. These authors have high degree because their total number of distinct co-authors is large. (d) is annotated to point out influential authors with high degree and papers that lead to the highest k-core (here 78-core). Note: The scale of the axes in each plot varies due to different ranges of degree and core numbers. Due to rounding of cell values close to zero, some cells in these figures were plotted in white color.