Analysis and visualization of large scale networks using the k-core decomposition

Ignacio Alvarez-Hamelin[1], Luca Dall'Asta[1], Alain Barrat[1] and Alessandro Vespignani[2]
[1] Laboratoire de Physique Théorique CNRS, Bât 210, Université de Paris-Sud, 91405 ORSAY Cedex France.
[2]School of Informatics and Department of Physics, University of Indiana, Bloomington, IN 47408, USA.

1 Introduction

In recent times, the possibility of accessing, handling and mining large-scale networks datasets has revamped the interest in their investigation and theoretical characterization along with the definition of new modeling frameworks. In particular, mapping projects of the World Wide Web (WWW) and the physical Internet offered the first chance to study topology and traffic of large-scale networks. Gradually other studies followed describing population networks of practical interest in social science, critical infrastructures and epidemiology [1, 3, 7, 10]. The study of large scale networks, however, faces us with an array of new challenges. The definitions of centrality, hierarchies and structural organizations are hindered by the large size of these networks and the complex interplay of connectivity patterns, traffic flows and geographical, social and economical attributes characterizing their basic elements.

In this paper, we propose the use of the k-core decomposition to study the hierarchical properties of large scale Internet maps. The k-core decomposition [4] consists in identifying particular subsets of the network, called k-cores, each one obtained by recursively removing all the vertices of degree smaller than k, until the degree of all remaining vertices is larger than or equal to k. Larger values of coreness clearly correspond to vertices with larger degree and more central position in the network's structure. The k-core decomposition therefore provides a probe to study the properties of the network's regions of increasing centrality. Here we analyze Internet networks at both the Autonomous Systems and router level. In both cases we find that k-cores are always made by a single connected component, indicating the presence of a hierarchy of well defined regions of which it is possible to investigate the statistical properties. Strikingly, the various distributions and quantities analyzed appear to be invariant in the various k-cores. This characteristic appears extremely important in providing an operative definition of a topological self-similarity of scale-free graphs and prompts to the k-core decomposition as a suitable transformation equivalent to a scale-change in the topological space of networks non-embedded in the geometrical space.

Motivated by the previous finding we developed a visualization algorithm based on the k-core decomposition that allows the identification of networks' fingerprints, according to properties such as hierarchical arrangement, degree correlations and centrality, etc. The distinction between networks with seemingly similar properties is achieved by inspecting the different layouts generated by the visualization algorithm. The running time of the algorithm grows only linearly with the size of the network, granting the scalability needed for the visualization of very large scale networks. We apply the proposed visualization algorithm to real Internet maps and several computer generated graphs aimed at their modeling. The visualization algorithm appears to be a convenient tool able to clearly pinpoint the differences of Internet maps obtained at different granularities and with different experimental techniques. In addition, the inspection of computer generated networks provides a first approach to models validation. The presented visualization algorithm is publicly available [8].

Keywords: network analysis, k-cores, Internet, visualization

2 k-core decomposition

In this section, first we introduce the definition of k-core decomposition, then we show how the application of this decomposition can shed light on important hierarchical properties of graphs, and finally we present a visualization that aids to identify networks.

Let us consider a graph $G = (V, E)$ of $|V| = n$ vertices and $|E| = e$ edges, the definition from [4] of k-cores is the following

Definition: *A subgraph $H = (C, E|C)$ induced by the set $C \subseteq V$ is a k-core or a core of order k iff $\forall v \in C : \mathtt{degree}_H(v) \geq k$, and H is the maximum subgraph with this property.*

A k-core of G can therefore be obtained by recursively removing all the vertices of degree less than k, until all vertices in the remaining graph have degree at least k. It is shown by V. Batagelj and M. Zaversnik [4] that the complexity of this decomposition is e, the number of edges.

2.1 Analyzing AS and IR graphs

We apply the k-core decomposition to Internet's maps. The autonomous system level is represented by collected routes of *Oregon route-views* [9] project, called AS from May 26, 2001. For the router level, we use the graph obtained by an exploration of CAIDA project [6] between April 21st and May 8th, 2003.

Figure 1 displays the cumulative degree distribution for the first k-cores, for two maps of Internet (upper plots); namely, the probability $P_>(d)$ that any vertex in the networks has a degree larger than d. Strikingly, the shape of the distribution, i.e. a power-law with an exponential cut-off, is not affected by the decomposition. In particular the exponent of the power-law is robust although the range of variation of the degree decreases. This feature defines a striking property of statistical self-similarity of the network and the generated k-cores, which resemble one with each other under the opportune rescaling of the average degree.

Another relevant quantity is the *clustering coefficient* that measures the local group cohesiveness and is defined for any vertex j as the fraction of connected neighbors of j [11]: $cc_j = 2 \cdot n_{\mathtt{link}}/(d_j(d_j - 1))$, where $n_{\mathtt{link}}$ is the number of links between the d_j neighbors of j. The study of the clustering spectrum $cc(d) = \frac{1}{n_d}\sum_{j/d_j=d} cc_j$, allows e.g. to uncover hierarchies in which low degree vertices belong generally to well interconnected communities (high clustering coefficient), while hubs connect many vertices that are not directly connected (small clustering coefficient). Figure 1 presents $cc(d)$ in the lower plots. Also in this case the shape of the spectrums is preserved as the network is recursively pruned of its low-degree vertices.

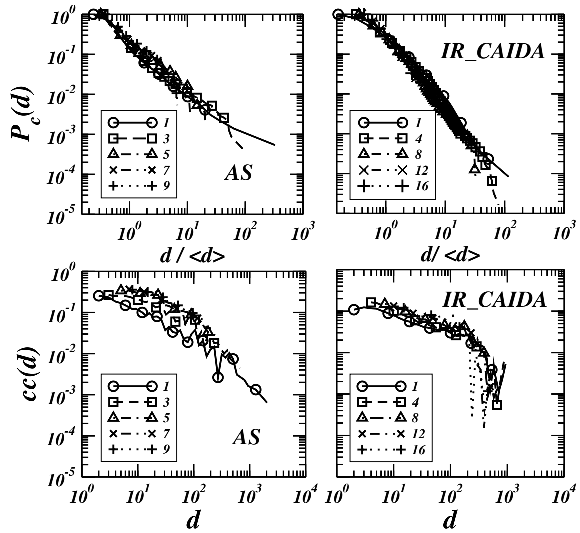

Figure 1: Top: cumulative degree distribution of the various k-cores. The degree is normalized by the average degree of each k-core. Bottom: clustering coefficient spectrum of the various cores.

2.2 Network fingerprints

Motivated by the previous analysis we propose a visualization algorithm based on k-core decomposition that places vertices in 2 dimensions, the position of each vertex depending on its coreness and on the coreness of its neighbors (see also [5]). A color code allows for the identification of core numbers, while the vertex's original degree is immediately provided by its size that depends logarithmically on the degree (see Figs. 2 and 3). For the sake of clarity, our algorithm represents a small percentage of the edges, chosen uniformly at random. The aim is to provide a clear visualization of the hierarchies, coreness shells and self-similarity observed in the context of statistical analysis presented in the previous section. As mentioned, in the most general situation, indeed, the recursive removal of vertices having degree less than a given k can break the original network into various connected components, each of which might even be once again broken by the subsequent decomposition. While this is not the case in the Internet maps that we have analyzed, we cannot exclude this possibility and a central role in our visualization method is played by the possibility of a multi-components representation of the k-core decomposition. The complete description of the algorithm can be found in [2]. The most remarkable fact is that the IR is populated at all levels, and it also has high degree nodes in low k-shells, while the AS has few nodes in higher k-shells and it has only high degree nodes in higher k-shells.

3 Conclusions

In this paper we have presented the application of the k-core decomposition to the analysis and visualization of large scale networks. The k-core decomposition allows the progressive pruning of large networks and the identification of subgraphs of increasing centrality. The study of these subgraphs and their statistical properties uncover the main hierarchical layers of the network and allows for their statistical characterization. Strikingly, we observe for heterogeneous graphs such the Internet at the Autonomous System (AS) and Router level a statistical self-similarity of the topological properties for cores of increasing centrality. We propose a general visualization algorithm that allows for the graphical distinction of the k-core hierarchy along with the degree of the vertices and their relation with the hierarchical position of the neighbors. The obtained results show the possibility of gaining clear insights on the architecture of many real world and synthetic networks. Networks with different topological properties and structural arrangement can be distinguished and the hierarchical arrangements of the elements rationalized. The present visualization strategy may be also used for determining if a certain model fits or not with the real data, providing a further interesting tool for models validation.

Acknowledgments: This work has been partially funded by the European Commission - Fet Open project COSIN IST-2001-33555 and contract 001907 (DELIS).

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–, 2000.

[2] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani. k-core decomposition: a tool for the visualization of large scale networks. *arxiv.org*, cs.NI/0504107, 2005.

[3] L. A. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley. Classes of small world networks. *Proc. Natl. Acad. Sci. (USA)*, 97:11149–11152, 2000.

[4] V. Batagelj and M. Zaversnik. Generalized Cores. *arxiv.org*, cs.DS/0202039, 2002.

[5] M. Baur, U. Brandes, M. Gaertler, and D. Wagner. Drawing the AS Graph in 2.5 Dimensions. In *"12th International Symposium on Graph Drawing, Springer-Verlag editor"*, pages 43–48, 2004.

[6] Cooperative Association for Internet Data Analysis, Router-Level Topology Measurements. `http://www.caida.org/tools/measurement/skitter/router_topology/`.

[7] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW.* Oxford University Press, 2003.

[8] LArge NETwork VIsualization tool. `http://xavier.informatics.indiana.edu/lanet-vi/`.

[9] University of Oregon Route Views Project. `http://www.routeviews.org/`.

[10] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A statistical physics approach.* Cambridge University Press, 2004.

[11] D. J. Watts and S. H. Strogatz. Collective Dynamics of Small-World Networks. *Nature*, 393:440–442, 1998.

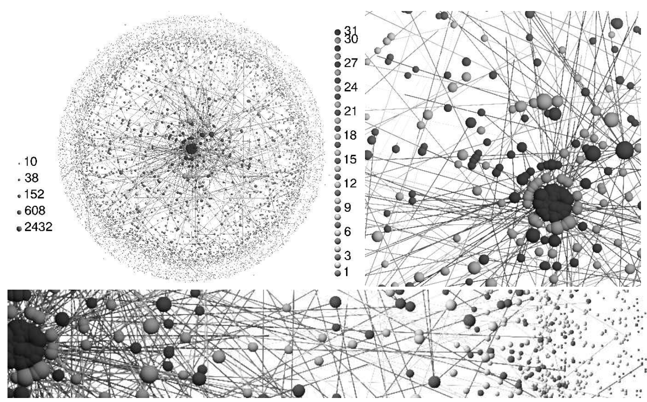

Figure 2: Graphical representation of the AS+ graph. The color code for the coreness is given on the right of the representation, while the legend for the degree of the vertices is given on the left, showing the maximum degree node. The following figure use the same legend.

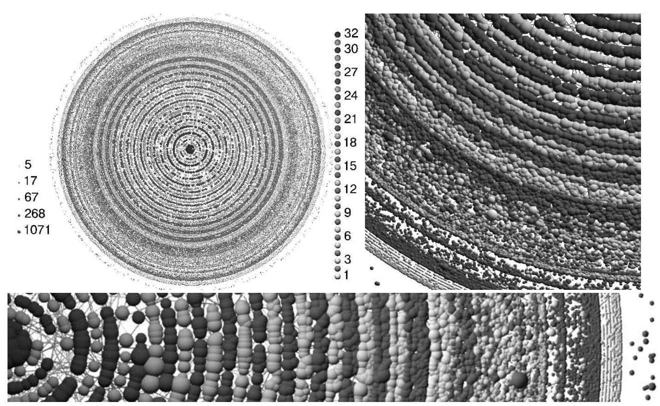

Figure 3: Graphical representation of the CAIDA IR graph.