K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases.

J. I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, A. Vespignani

2006
K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases.

J. I. Alvarez-Hamelin,1, 2 Luca Dall'Asta,1 Alain Barrat,1 and Alessandro Vespignani3

1Laboratoire de Physique Théorique (UMR du CNRS 8627), Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex (France)
2Facultad de Ingeniería, Univ. de Buenos Aires, Paseo Colón 850, C1063ACV, Argentina
3School of Informatics, Indiana University, Bloomington, IN 47408, USA

We consider the k-core decomposition of large scale Internet graphs at the autonomous system level. We characterize more and more central cores of the graphs, conveniently uncovering the hierarchical and structural properties of the Internet network. We compare the obtained results with the analysis of several complex network models and discuss the differences and similarities with the real Internet graph. The Internet exhibits a k-core structure with invariant statistical properties recovered only in models with heavy tailed distributions and we discuss the implications of this finding on routing and hierarchical properties. We address the issue of the incompleteness of Internet maps by analyzing the k-core structure of graphs obtained by simulated trace-route explorations of different network models. We show that the k-core structure of sampled graphs provides a clear fingerprinting of the actual underlying networks, clearly discriminating its topological properties even at low level of sampling. Finally, the temporal evolution of the k-core structure of actual Internet maps provides the quantitative characterization of the fluctuations and incompleteness inherent to the sampling process.

I. INTRODUCTION

In recent times, mapping projects of the World Wide Web (WWW) and the physical Internet have 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, 17, 34]. The study of large scale networks, however, faces us with an array of new challenges. The definitions of centrality, hierarchies and structural organizations are in particular hindered by the large size of these networks and the complex interplay of connectivity patterns, traffic flows, geographical, and economical attributes characterizing their elements.

In this paper, we investigate the k-core decomposition of large scale networks, focusing on well-known network models and on Internet maps. The k-core decomposition [5, 8, 37] 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. The k-core decomposition therefore provides a probe to study the hierarchical properties of large scale networks, focusing on the network’s regions of increasing centrality and connectedness properties. More central cores are indeed more strongly connected, with larger number of possible distinct paths between vertices: this allows to obtain not only more robust routing properties but also better Quality of Service (QoS). Here, we analyze models of complex networks as well as Internet maps at the Autonomous Systems (AS) level obtained by various Internet mapping projects. 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 in the AS maps. This feature is shared by the network models with heterogeneous topology. The k-core decomposition therefore exploits the self-similar properties of networks, uncovering the same structural ordering at different hierarchical levels. It also appears as a valuable tool for model validation. We show moreover how the k-core decomposition can be used as a sensitive probe of temporal evolution of networks by comparing the k-core decompositions of the AS maps at various times between 2001 and 2005. Since such maps are a result of a sampling (typically path-based, i.e. using a merging of many routes between various points of the networks), we also discuss how sampling biases can alter the structure of the k-core decomposition of various complex networks. Our findings indicate that the k-core decomposition’s fingerprints allow the discrimination between heterogeneous and homogeneous topologies even after an incomplete sampling: this shows that the signatures observed in the AS Internet maps are qualitatively reliable, even if some biases are unavoidable at a detailed quantitative level.

II. RELATED WORK

The discussion of the topological properties of Internet maps has highlighted a very complex and heterogeneous topology with fluctuations extending over many scale lengths. Starting with the seminal paper by Faloutsos, Faloutsos
and Faloutsos[19] an impressive number of papers has dealt with the characterization of the large scale properties and hierarchies of the Internet [25, 26, 34, 36, 39, 40]. Here, we consider the use of the k-core decomposition as a probe for the hierarchical and self-similar structure of large-scale networks, and in particular of Internet maps.

The k-core decomposition has mostly been used in biologically related contexts, where it was applied to the analysis of protein interaction networks or in the prediction of protein functions [3, 43]. An interesting application in the area of networking has been provided by Gaertler et al. [21], where the k-core decomposition is used for filtering out peripheral Autonomous Systems (ASes) in the case of Internet maps. The k-core decomposition has also recently been used as a basis for the visualization of large networks, in particular for AS maps [2, 7, 24]. Finally, recent works using the k-core analysis have focused on the analysis of the Internet maps obtained by the DIMEs project [27]. In Refs.[11, 12], an approach based on the k-core decomposition has been used to provide a conceptual and structural model of the Internet, the so-called Medusa model for the Internet. Up to now, no study has however considered the k-core decomposition of the various commonly used models for complex networks, nor compared it to the one of real-world networks.

Subramanian et al. [39] have proposed to classify ASes in five different levels or “tiers”, and given a method to extract this classification from the AS directed graph. This method can however lead to some biases when the knowledge of the all peer-to-peer relationships is not complete. The k-core decomposition studied in this paper considers on the other hand undirected networks, and yields a finer hierarchy, not based on the commercial relations between vertices, and in which the number of levels is not fixed a priori but depends on the characteristics of the network. It is moreover not restricted to AS maps but can be applied as well for example to Internet router maps or more generally to any real or computer generated graph.

Finally, recent works have been devoted to a better understanding of the possible sources of errors and biases presented by the experimental data [13, 15, 16, 22, 23, 35]. Since Internet maps are typically based on a sampling of routes between sources and destinations (obtained by tools such as traceroute), these studies have dealt with a simplified model of traceroute-like sampling, applied to graphs with various topological properties. They have shown that, except in some peculiar cases [13], the sampling process allows to distinguish qualitatively between networks with strongly different properties (homogeneous vs. heterogeneous), indicating that the heterogeneous properties of Internet maps are genuine [15, 16]. However, the precise quantitative form of the degree distribution can suffer important biases, and a huge mapping effort may be necessary to distinguish between close topologies. On the other hand, the effect of sampling biases on the k-core structure of networks has not been studied yet.

III. K-CORE DECOMPOSITION

Let us consider a graph $G = (V, E)$ of $|V| = n$ vertices and $|E| = e$ edges, the definition from [5] of k-cores is the following

**Definition 1:** A subgraph $H = (C, E|C)$ induced by the set $C \subseteq V$ is a k-core or a core of order $k$ if and only if $\forall v \in C : \text{degree}_H(v) \geq k$, 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 worth remarking that this process is not equivalent to prune vertices of a certain degree. Indeed, a star-like subgraph formed by a vertex with a high degree that connects many vertices with degree one, and connected only with a single edge to the rest of the graph, is going to belong to the first shell no matter how high is the degree of the vertex. We will also use the following definitions

**Definition 2:** A vertex $i$ has shell index $k$ if it belongs to the k-core but not to $(k + 1)$-core.

**Definition 3:** A k-shell $S_k$ is composed by all the vertices whose shell index is $k$. The maximum value $k$ such that $S_k$ is not empty is denoted $k_{\text{max}}$. The k-core is thus the union of all shells $S_c$ with $c \geq k$.

**Definition 4:** Each connected set of vertices having the same shell index $c$ is a cluster $Q^c$, where the corresponding set of edges are those connecting vertices of the cluster. Each shell $S_c$ is thus composed by clusters $Q^c_m$, such that $S_c = \cup_{1 \leq m \leq q_{\text{max}}^c} Q^c_m$, where $q_{\text{max}}^c$ is the number of clusters in $S_c$.

The k-core decomposition therefore identifies progressively internal cores and decomposes the networks layer by layer, revealing the structure of the different k-shells from the outmost one to the most internal one, as sketched in Fig. 1.

It is interesting to note that the k-core decomposition can be easily implemented: the algorithm by Batagelj and Zaversnik [6] presents a time complexity of order $O(n + e)$ for a general graph. This makes the algorithm very efficient for sparse graphs, where $e$ is of order $n$.

A very interesting feature of the k-cores concerns their connectivity properties. It has been for example shown experimentally in [11] that the k-cores of the AS map obtained by the DIMEs project [27] are $k$-connected, which means that $k$ disjoint paths are available between any two vertices belonging to the k-core. In fact, for any two vertices $u$ and $v$ of the network, with shell indices respectively $c_u$ and $c_v$, there are (with some exceptions for small
values of $c_u$ and $c_v$) at least $\min(c_u, c_v)$ disjoint paths between $u$ and $v$ [11]. Such property has important practical consequences since it implies larger and larger robustness and routing capacities for more and more central cores. The knowledge of such capacities identifies a very important hierarchy of ASes that could be taken advantage of by newly created ASes in order to choose to which other ASes to establish connections. We will come back to this point in section IV B 3.

IV. $K$-CORE STRUCTURE OF INTERNET MAPS AND MODELS

A. Structure of some models

In order to better understand the properties of the $k$-core decomposition of networks, we first apply it to a set of well known and commonly used models of complex networks, whose main characteristics are summarized in Table I. Various topological properties can lead to various decompositions so we consider both homogeneous and heterogeneous networks.

![FIG. 1: Sketch of the $k$-core decomposition for a small graph. Each closed line contains the set of vertices belonging to a given $k$-core, while different types of vertices correspond to different $k$-shells.](image)

<table>
<thead>
<tr>
<th>source</th>
<th>$n$</th>
<th>$e$</th>
<th>$\langle d \rangle$</th>
<th>$d_{max}$</th>
<th>$k_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER</td>
<td>$10^5$</td>
<td>$10^6$</td>
<td>20</td>
<td>41</td>
<td>14</td>
</tr>
<tr>
<td>Weibull</td>
<td>$10^5$</td>
<td>307500</td>
<td>6.15</td>
<td>377</td>
<td>9</td>
</tr>
<tr>
<td>RSF $\gamma = 2.3$</td>
<td>97315</td>
<td>293891</td>
<td>6.04</td>
<td>938</td>
<td>22</td>
</tr>
<tr>
<td>BRITE</td>
<td>$10^4$</td>
<td>156145</td>
<td>3.63</td>
<td>433</td>
<td>54</td>
</tr>
<tr>
<td>INET 3.0</td>
<td>10000</td>
<td>19676</td>
<td>3.936</td>
<td>984</td>
<td>8</td>
</tr>
</tbody>
</table>

TABLE I: Main properties of the models considered in the present study: number of vertices $n$ and of edges $e$, average degree $\langle d \rangle$, maximum degree $d_{max}$ and maximum shell index $k_{max}$.

1. Size of shells

We first consider for reference the random graph model of Erdős and Rényi (ER) [18], which is the most standard example of graphs with a characteristic value for the degree (the average value $\langle d \rangle$). In this case, the maximum index is clearly related to the average degree $\langle d \rangle$. The vertex degrees have only small fluctuations, thus most vertices belong to the same $k$-core that is also the highest. Noticeably, the size of the shells is increasing with the index, showing that only few vertices can be considered as peripherical (see Fig. 2), and that the network contains no clear hierarchy between nodes.

Since many real-world networks have been shown to display a very heterogeneous topology as measured by broad degree distributions, many models and mechanisms have been proposed to construct heterogeneous networks. Because of its renown, it could be tempting to consider the Barabási-Albert (BA) model [4] as a paradigm for Internet modeling, and we therefore consider here its $k$-core decomposition. In this model, a growing network is constructed according to the preferential attachment mechanism: each new vertex is connected to $m$ already existing vertices chosen with a probability proportional to their starting degree. This model produces graphs with power-law degree distributions, thus characterized by a very large variety of degree values. Strikingly, the corresponding $k$-core decompositions are however trivial, with only few shells at very small index. The construction mechanism provides a simple explanation.
Each new vertex enters the system with degree \( m \), but at the following time steps new vertices may connect to it, increasing its degree. Inverting the procedure, we obtain exactly the \( k \)-core decomposition. The minimum degree is \( m \), therefore all shells \( C_i \) with \( c < m \) are empty. Recursively pruning all vertices of degree \( m \), one first removes the last vertex, then the one added at the preceding step, whose degree is now reduced to its initial value \( m \), and so on, up to the initial vertices which may have larger degree. Hence, all vertices except the initial ones belong to the shell of index \( m \). This somewhat pathological property holds for all growing networks with fixed initial number of links for new vertices, even if tunable exponents for the degree distributions can be obtained (by e.g. generalizing to linear preferential attachment).

Variations of the basic growing procedure result in more complicate structures. The simple introduction of stochasticity, obtained with a randomly varying \( m \), allows to obtain a slightly larger number of shells, with however a maximal index \( k_{\text{max}} \) still strongly linked to the average \( m \) (and thus to the average degree). Strikingly, one also obtains that the shell size is increasing with its index (not shown), except for the most central shells that are very small.

On the other hand, the BRITE generator [28] proposes a growth mechanism combining the addition of vertices with \( m \) new links according to the preferential attachment with the addition of new links between already existing vertices, also through a preferential attachment mechanism. In this case, a non-trivial structure of shells is obtained, with a largest shell index \( k_{\text{max}} \) much larger than the average degree, and a shell size decreasing as a power-law function of the index. This implies a similar power-law relation between the size of each \( k \)-core and its index, meaning that each \( k \)-core is composed of a constant fraction of the \((k-1)\)-core, building a hierarchy of more and more central vertices. At large \( k \), large fluctuations are observed, with a relatively large central core (see Fig. 2). Such a difference between BRITE and BA networks highlights the structural relevance of the addition of new links between already existing vertices in a growing heterogeneous network model.

The various afore-mentioned heterogeneous graphs can be considered as particular since they are growing networks, constructed according to the preferential attachment mechanism. We therefore consider as well “static” graphs obtained by various algorithms.

The INET3.0 topology generator [42] has been specifically designed to represent the Internet at the AS level by obtaining a closely similar topology. As shown in Fig. 2, such network presents a small number of \( k \)-cores, with a shell size behavior that is exponentially decreasing for increasing shell index.

Other algorithms are widely used to obtain random graphs with prescribed broad degree distributions. In the literature, different definitions of heavy-tailed like distributions exist. While we do not want to enter the detailed definition, we have considered two classes of such distributions: (i) scale-free or Pareto distributions of the form \( P(k) \sim k^{-\gamma} \) (RSF), and (ii) Weibull distributions (WEI) \( P(k) = (a/c)(k/c)^{a-1} \exp(-(k/c)^a) \). The scale-free distribution has a diverging second moment and therefore virtually unbounded fluctuations, limited only by eventual size-cut-off. The Weibull distribution is akin to power-law distributions truncated by an exponential cut-off which are often encountered in the analysis of scale-free systems in the real world. Indeed, a truncation of the power-law behavior is generally due to finite-size effects and other physical constraints. Both forms have been proposed as representing the topological properties of the Internet [10]. We have generated the corresponding random graphs by using the algorithm proposed by Molloy and Reed [29, 30]: the vertices of the graph are assigned a fixed sequence of degrees \( \{k_i\} \), \( i = 1, \ldots, N \), chosen at random from the desired degree distribution \( P(k) \), and with the additional constraint that the sum \( \sum k_i \) must be even; then, the vertices are connected by \( \sum k_i/2 \) edges, respecting the assigned degrees and avoiding self- and multiple-connections. The parameters used are \( a = 0.4 \) and \( c = 0.6 \) for the Weibull distribution, and \( \gamma = 2.3 \) for the RSF case.

As shown for an example in Fig. 2, random networks with heavy-tailed degree distributions present systematically a large number of shells (we have also checked that \( k_{\text{max}} \) increases if \( \gamma \) decreases), and much larger than the average degree \( \langle d \rangle \). The shell size is decreasing as a power-law of the index, with a quite large central core of index \( k_{\text{max}} \), as for BRITE. On the contrary, Weibull distributed networks have relatively few shells with a much smaller \( k_{\text{max}} \). It is interesting that networks with relatively similar degree distributions can present in fact strongly different \( k \)-core decompositions. This points to the \( k \)-core decomposition as a supplementary valuable tool for network investigation.

2. Self-similarity

In this paragraph, we compare the statistical characteristics of the different cores, i.e. of more and more central parts of the network, focusing in particular on heterogeneous networks.

Figure 3 shows the cumulative degree distribution for some \( k \)-cores, for some of the studied models; namely, the probability \( P_{\chi}(d) \) that any vertex in the networks has a degree larger than \( d \). Strikingly, the shape of the distribution (power-laws or Weibull) is not affected by the decomposition. This feature points to a striking property of statistical self-similarity of the generated \( k \)-cores, which resemble one with each other under the opportune rescaling of the average degree.
In order to better characterize and check this self-similarity, we have computed also the two and three points correlations functions of the various \( k \)-cores. A useful measure to quantify correlations between the degrees of neighboring vertices is the average degree of nearest neighbors \( d_{nn}(d) \) of vertices of degree \( d \) [33]:

\[
d_{nn}(d) = \frac{1}{n_d} \sum_{j/d_j=d} \frac{1}{d_j} \sum_{i \in V(j)} d_i,
\]

where \( V(j) \) is the set of the \( d_j \) neighbors of vertex \( j \) and \( n_d \) the number of vertices of degree \( d \). This last quantity is related to the correlations between the degree of connected vertices since on the average it can be expressed as

\[
d_{nn}(d) = \sum_{d'} d' P(d'|d),
\]

where \( xP(d'|d) \) is the conditional probability that a vertex with degree \( d \) is connected to a vertex with degree \( d' \). If degrees of neighboring vertices are uncorrelated, \( P(d'|d) \) is only a function of \( d' \) and thus \( d_{nn}(d) \) is a constant. When correlations are present, two main classes of possible correlations have been identified: assortative behavior if \( d_{nn}(d) \) increases with \( d \), which indicates that large degree vertices are preferentially connected with other large degree vertices, and disassortative if \( d_{nn}(d) \) decreases with \( d \) [32]. From a routing point of view, a disassortative behavior corresponds to a network structure where vertices with small degree are preferentially connected to the hubs (i.e., large degree vertices). A second, and often studied, relevant quantity is the clustering coefficient [41] that measures the local group cohesiveness and is defined for any vertex \( j \) as the fraction of connected neighbors of \( j \)

\[
cc_j = 2 \cdot n_{1nj}/(d_j(d_j - 1)) ,
\]
where \( n_{\text{link}} \) is the number of links between the \( d_j \) neighbors of \( j \). The study of the clustering spectrum \( cc(d) \) of vertices of degree \( d \), defined as

\[
cc(d) = \frac{1}{n_d} \sum_{j:d_j=d} cc_j,
\]

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). Large clustering has a clear relevance for routing purposes since it indicates the presence of alternative paths thanks to the presence of many triangles: if a link from a vertex \( u \) to a neighbor \( v \) goes down, the message can be sent from \( u \) to \( v \) through a common neighbor.

**FIG. 4:** Nearest neighbors degree distribution of some \( k \)-cores, rescaled by the corresponding average values, for some model networks. The degree of each node is normalized by the average degree of each \( k \)-core.

In Figs. 4 and 5, we report the \( d_{nn}(d) \) and \( cc(d) \) computed for the various \( k \)-cores. Strikingly, the behavior of the two quantities is preserved in all cases as the network is recursively pruned of its low-degree vertices. In other words, the overall network topology is invariant for \( k \)-cores of increasing centrality.

**FIG. 5:** Clustering coefficient spectrum of some \( k \)-cores for some model networks. The degree of each node is normalized by the average degree of each \( k \)-core, and the clustering coefficient is rescaled by the average clustering of each \( k \)-core.

### 3. Summary

In summary, the \( k \)-core decomposition allows to uncover very different behaviors for different models which may otherwise share e.g. very similar degree distributions. For example, a growing network obtained with the linear preferential attachment rule may have a scale-free distribution of degrees \( P(k) \sim k^{-\gamma} \) but will have a trivial shell
structure because of its construction mechanism. On the other hand, randomly constructed scale-free networks, which may have trivial correlation properties and small clustering, can present a rich hierarchical decomposition with a large central core of high shell index.

Interestingly, the more and more central cores present a high level of self-similarity, preserving the shape of the degree distribution as well as the correlation properties as more and more external vertices are pruned.

The $k$-core decomposition appears therefore as a very valuable additional tool to characterize the topology of complex networks, discriminating among models with apparently similar topologies. On the other hand, the fact that even random scale-free networks with trivial correlation properties present a rich $k$-core decomposition highlights the difficulty of constructing a model based on the sole characteristics of such decomposition.

B. Internet AS maps

In this section, we inspect Internet maps at the AS level and compare their $k$-core structure with the insights obtained from models. In order to obtain Internet connectivity information at the AS level it is possible to inspect routing tables and paths stored in each router (passive measurements) or directly ask the network with a software probe (active measurements). In the following we consider data from two recent large scale Internet mapping projects using an active measurement approach. The skitter project at CAIDA [14] has deployed several strategically placed probing monitors using a path probing software. All the data are then centrally collected and merged in order to obtain Internet maps that maximizes the estimate of cross-connectivity. The second set we consider is provided by the Distributed Internet Measurements and Simulations (DIMES) project [27, 38]. At the moment the project consists of more than 5,000 measuring agents that perform Internet measurements such as traceroute and ping.

Table I displays a summary of the basic properties of the considered Internet maps. In the following we show how the application of the $k$-core decomposition can shed light on important hierarchical properties of Internet graphs, focusing on the AS maps obtained by each project in 2005.

The first observation about the structure of the $k$-cores is that, as also happened in the models, they remain connected. This is not a priori an obvious fact since one can easily imagine networks whose $k$-core decomposition yields several connected components corresponding, e.g. to various communities. Instead, each decomposition step is just peeling the network leaving connected the inner part of the network, showing a high hierarchical structure, i.e. the most connected part of the network is also the most central. Figure 6 displays the size in terms of vertices of each $k$-shell as a function of its index. As for RSF or BRITE networks, power-law like shapes are obtained: each $k$-core is composed of a constant fraction of the $(k - 1)$-core. Important fluctuations appear at large $k$, which is not very surprising since such shells of large index are relatively small, except for the most central core which contains respectively 50 vertices at $k_{\text{max}} = 26$ and 82 vertices at $k_{\text{max}} = 39$ for CAIDA and DIMES. Such a structure has also been observed in the independent study of [11].

<table>
<thead>
<tr>
<th>source</th>
<th>n</th>
<th>e</th>
<th>$\langle d \rangle$</th>
<th>$d_{\text{max}}$</th>
<th>$k_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAIDA, 2005/04</td>
<td>8542</td>
<td>25492</td>
<td>5.97</td>
<td>1171</td>
<td>26</td>
</tr>
<tr>
<td>DIMES, 2005/05</td>
<td>20455</td>
<td>61760</td>
<td>6.04</td>
<td>2800</td>
<td>39</td>
</tr>
</tbody>
</table>

TABLE II: Main properties of the Internet maps considered in the present study: number of vertices $n$ and of edges $e$, average degree $\langle d \rangle$, maximum degree $d_{\text{max}}$ and maximum shell index $k_{\text{max}}$.

Interestingly, a much larger $k_{\text{max}}$ is obtained for the DIMES AS map than for the CAIDA one. It is likely that such discrepancy is linked to the diversity of the exploration methods. The maximum core depends indeed largely on the amount of discovered edges and lateral connectivity. The set of “observers” is 22 for CAIDA but more than 5,000 for DIMES. It is therefore reasonable that the latter has more probability to discover edges, and therefore a larger value of $k_{\text{max}}$.

1. Self-similarity

As in the case of the models, we study the statistical properties of the successive $k$-cores of Internet maps by considering their degree distributions and correlation properties.

Figure 7 shows the cumulative degree distribution for the first $k$-cores, for the various AS maps. Similarly to the case of the network models with heterogeneous degree distributions, the shape of the distribution, i.e. an approximate power-law, 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. In other words, each core conserves a broad degree distribution: AS with
FIG. 6: Shell size as a function of their index for the AS maps. The dashed line is a power-law $\propto k^{-2.7}$.

significantly different number of neighbors are present in each core or hierarchy level. Figure 8 shows moreover that the clustering and correlations structures of the Internet maps are as well essentially preserved as the more and more external parts of the network are pruned. We note however that, as also shown in [11], the largest $k$-cores are no more scale-free: since they are very densely connected, their degree distribution is rather peaked around an average value and their topology is closer to that of a random graph with large average degree.

FIG. 7: Rescaled cumulative degree distributions of some $k$-cores of the AS Internet maps. The degree is normalized by the corresponding average degree in each $k$-core. The shapes of the distributions are preserved by the successive pruning.

In summary, the AS networks exhibit a statistical scale invariance with respect to the pruning obtained with the $k$-cores decomposition for a wide range of $k$. Indeed, while this decomposition identifies subgraphs that progressively correspond to the most central regions of the network, the statistical properties of these subgraphs are preserved at many levels of pruning. This hints to a sort of global self-similarity for regions of increasing centrality of the network, and to a structure in which each region of the Internet as defined in terms of network centrality has the same properties than the whole network. This is particularly interesting since the properties of Internet (heterogeneous degree distributions, correlations, clustering...) have been up to now studied at the level of the whole map, while one can be interested to restrict the analysis to some particular regions of the map, focusing for example on parts of the network with certain routing capabilities (QoS, failure support). At a general level, the $k$-core decomposition appears therefore as a suitable way to define a pruning procedure equivalent to a scale-change preserving the statistical properties of graphs while focusing on their more and more connected parts.

2. Shell index and centrality

The identification of the most central vertices is a major issue in networks characterization [20]. While a first intuitive and immediate measure of the centrality of vertices is given by their degree, more refined investigations are needed in order to characterize the real importance of various vertices: for example, some low-degree vertices may be essential because they provide connections between otherwise separated parts of the network. In order to uncover such important vertices, the concept of betweenness centrality (BC) is now commonly used [20, 31]. The betweenness centrality of a vertex $v$ is defined as $g(v) = \sum_{s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where $\sigma_{st}$ is the number of shortest paths going from $s$ to $t$...
and \( \sigma_{st}(v) \) is the number of shortest paths from \( s \) to \( t \) going through \( v \). This definition means that central vertices are part of more shortest paths within the network than peripheral vertices. Moreover, the betweenness centrality gives in transport networks an estimate of the traffic handled by the vertices, assuming that the number of shortest paths is a zero-th order approximation to the frequency of use of a given vertex (e.g. the load of an AS), in the case of an all-to-all communication.

The \( k \)-core decomposition intuitively provides a hierarchy of the vertices based on their shell index that is a combination of local and global properties. (e.g.: [43] shows that the shell index is a better criterium for centrality than the degree in protein interaction networks). In this perspective, it becomes very interesting to study the correlation between the degree, the betweenness centrality and the shell index of a vertex in order to quantify the statistical level of consistency of the various measures. We show in Fig. 9 the average betweenness centrality (computed on the original graph) of vertices as a function of their shell index, and the shell index as a function of the degree \( d \). A strong correlation is expected, but the fluctuations observed should not be a surprise: while a low-degree vertex has clearly low shell index, large or medium degree vertices do not have necessarily a large shell index. In the AS maps, we observe in fact that all large degree vertices belong to the most central core, while large fluctuations are observed for intermediate degree values. Moreover, the betweenness centrality is a highly non-local quantity which can be large even for small-degree vertices. These quantities are thus pinpointing different kinds of centrality. The shell index appears therefore as a very interesting quantity to uncover central vertices and it has the advantage of a much faster computation time than those required for the betweenness centrality (of order \( n^2 \log n \) [9]).
### Potential practical implications

The $k$-core decomposition has interesting immediate applications. First of all, as already mentioned in section III, it has been shown in ref [11] that each $k$-core of the DIMES AS map is $k$-connected, and that the number of disjoint paths between two vertices $u$ and $v$ of this map is bounded from below by the minimum of the shell indices of $u$ and $v$.

Moreover, it is quite easy to show and understand that similar properties hold for a network under certain assumptions. In particular, if the central core (of shell index $k_{\text{max}}$) of a given network is $k_{\text{max}}$-edge-connected, and if there exists enough edges between the various shells (in particular if any cluster - see Def. 4- of each $k$-shell is connected to the $k+1$-shell by at least $k$ edges), then each $k$-core of the network turns out to be $k$-edge connected. We have in fact checked that these conditions are verified for the CAIDA and DIMES maps as well as for the network models under study. Note that $k$-edge connectivity (i.e. the existence of $k$ distinct paths which do not share any common edge) is less restrictive than $k$-connectivity. In the context of Autonomous Systems and evaluation of routing capacities or of failure possibilities however, it is particularly relevant since a vertex of the AS map represents in fact many different routers, so that different paths may cross at a given AS while being effectively physically disjoints.

Such connectivity properties highlight the fact that the $k$-core decomposition provides a natural definition for a hierarchy in the network, in which the more central vertices (with larger shell index) have better routing capabilities, and each $k$-core constitutes an ensemble of ASes able to provide a certain QoS, with global larger robustness for larger $k$.

It is therefore interesting to compare the $k$-core decomposition with the tiers hierarchy proposed by Subramanian et al. [39]. These two hierarchies have different origins and motivations: on the one hand, the tiers classification is based on the inference of AS commercial relationships; on the other hand, and in a somehow opposite point of view, the $k$-core decomposition gives a classification of the network’s vertices which does not have an a priori fixed number of classes or levels, but which adapts itself to the situation of the network. Moreover, the shell index of a vertex is not fixed once and for all but may fluctuate in time due to possible connectivity changes (as investigated in the next section). In this aspect, such a hierarchy provides very relevant information about the state of the network at a given time. While the actual routing protocols do not take advantage of such information, one could imagine that future routing protocols may be able to exploit it.

We finally note that the use of the $k$-core decomposition in order to find a certain hierarchy of connectedness properties is not limited to the analysis of AS maps: it can as well be applied to other kinds of Internet maps, for example at the router level, or to any communication or transportation network.

### $K$-CORES, DYNAMICS AND SAMPLING BIASES

#### A. Temporal variations of the $k$-core structure

The availability of data obtained by the various projects makes it possible to study the temporal evolution of the Internet maps. We have considered the maps obtained by the CAIDA project at various times between 2001 and 2005. Table III shows the main characteristics of the analyzed maps, each of which was obtained from the archives of one complete month.

While statistical signatures such as degree distribution, disassortative behavior and clustering spectrum are typically very stable over time, the $k$-core structure analysis reveals some finer variations. For example, the number of vertices and edges and the maximal shell index fluctuate in the CAIDA maps. This can be tracked down to the fact that the number of sources used by CAIDA changes (14 for the 2001/05 map, 21 for 2002/03, 24 for 2003/05 and 2004/04, and 22 for 2005/04), and that the locations of some of these sources also change.

Interesting informations also arise from the study of the change in the composition of the various $k$-shells: we show

<table>
<thead>
<tr>
<th>Year/</th>
<th># Vertices</th>
<th># Edges</th>
<th>Average Degree</th>
<th>Max Degree</th>
<th>Max Shell</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001/05</td>
<td>7400</td>
<td>24791</td>
<td>6.700</td>
<td>1820</td>
<td>28</td>
</tr>
<tr>
<td>2002/03</td>
<td>8489</td>
<td>28871</td>
<td>6.802</td>
<td>2007</td>
<td>32</td>
</tr>
<tr>
<td>2003/05</td>
<td>8755</td>
<td>27300</td>
<td>6.236</td>
<td>1560</td>
<td>26</td>
</tr>
<tr>
<td>2004/04</td>
<td>9238</td>
<td>28016</td>
<td>6.065</td>
<td>1406</td>
<td>26</td>
</tr>
<tr>
<td>2005/04</td>
<td>8542</td>
<td>25492</td>
<td>5.969</td>
<td>1171</td>
<td>26</td>
</tr>
</tbody>
</table>

**TABLE III: Characteristics of the CAIDA AS maps considered for the time analysis.**
FIG. 10: The grayscale code gives the probability of a change in shell index, from the CAIDA map of 2004/04 (x axis) to the one of 2005/04 (y axis). The points in line 0 correspond to ASes that are present in 2004 but not in 2005, and the column 0 corresponds to the reverse situation.

as an example in Fig. 10 the probability for a given AS to change from a shell of index $x$ in a map obtained at a given time to a shell of index $y$ in the successive map. While most vertices do not change their shell index, as shown by the dark area around the diagonal, some suffer an important change of status, from a highly central shell to a peripherical one or vice-versa. This highlights the presence of strong structural fluctuations in the evolution of CAIDA AS maps.

A further fingerprint of such structural changes is provided by the analysis of the shell index of vertices that appear in or disappear from the maps between one snapshot and the other, as shown in Fig.11: vertices in all shells, even central ones, disappear from the CAIDA maps even in the most recent maps, between 2004 and 2005. The fluctuations observed in the shell index of ASes may be related to two factors. A first one is the natural evolution of the Internet structure. A second factor is the uncertainty and bias in the data collection. In this respect, CAIDA maps seem to exhibit a high level of instability, indicative of a mapping process less stable in time. In this context, the $k$-core analysis appears as an interesting tool to highlight the temporal changes of the Internet structure as well as the measurement reliability in each particular experimental set-up, at an intermediate level between global quantities and local ones such as the degree. It will certainly be of interest in the future to study similar data for evolving DIMES maps, which are obtained with a much larger set of sources.

B. Sampling biases

In this paragraph, we perform a sensitivity analysis of the $k$-core decomposition with respect to eventual sampling biases. Indeed, Internet maps are currently obtained through sampling methods of the real Internet, which are
based on a merging of paths between sources and destinations, obtained either through Border Gateway Protocol routing tables or through active traceroute measurements. Such sampling processes present possible sources of errors and biases whose effect has been up to now studied essentially for the degree distributions [13, 15, 16, 22, 23, 35]. The analysis of idealized sampling processes on networks with various topologies has in particular revealed that the broadness of the degree distributions observed in Internet maps is a genuine feature, although important biases can remain on the exact form of the distribution, due to an undersampling of vertices with small degree. Moreover, although a path-based sampling process can produce a heterogeneous graph out of an homogeneous initial network (such as an ER graph), as rigorously shown in [13], this is restricted to the case of a single source probing. It is therefore interesting to note that a single source traceroute-like probing of any network yields essentially a tree, whose $k$-core decomposition is by definition trivial (with $k_{\text{max}} = 1$). Another obvious but important constatation regards the largest shell index: by definition, a sampling cannot discover paths or edges that do not exist, so that the maximal shell index of a network, $k_{\text{max}}$, can not be increased by partial sampling (nor can the maximal degree observed). In fact the actual $k_{\text{max}}$ is thus at least equal to the one found by a sampling of the true network.

Since more central cores are more connected, and more paths go through them, path-based sampling should intuitively discover and sample better more central cores, while the peripheral shells could suffer from stronger biases. In order to check such ideas, we perform a traceroute-like probing of the various model networks considered in section IV A, and compare their $k$-core decomposition before and after sampling. We use the same model for traceroute as in [15, 16, 22]: a set of $N_S$ sources sends probes to $N_T$ destinations randomly placed on the network, and the shortest paths between the source-destination pairs are merged to compose the sampled network. We use $N_S = 50$ sources, and various probing efforts measured by $\epsilon = N_S N_T / N$ (where $N$ is the size of the initial network), from a small value $\epsilon = 0.1$ (corresponding to a small density of targets $N_T / N = 2.10^{-3}$) to a much larger $\epsilon = 5$ (relatively large density of targets $N_T / N = 10^{-1}$).

Figure 12 presents the curves of the $k$-shell size as a function of the index for various network models and various sampling efforts. For ER networks, the populated shells change from being at index values only slightly under $k = (d)$ to much smaller values, with an almost uniform population of shells. The observed behavior is therefore completely different from the one observed in AS maps. On the contrary, the power-law shape obtained for RSF or BRITE networks, and comparable to the one of the AS maps, is very robust, even if the slope is affected. Indeed, shells of smaller indices are less well sampled. In particular, the size of the first shell is most strongly decreased by the sampling procedure: in some cases in fact, the first shell is larger than the second in the original network, but becomes smaller in the sampled network. We note that in the available AS maps, the first shell is indeed typically smaller than the second, and that the true AS network thus very probably exhibits a much larger shell of index $k = 1$. Similarly, one can expect that the exponent close to 2.7 of the power-law behavior of the shell size vs. its index (see [11] and Fig. 6) is a lower bound and that such value might be reconsidered in the future thanks to more and more extensive sampling efforts. On the other hand, the fact that the shell of largest index is substantially larger than the ones with immediately lower indices is well preserved, even if its index is substantially decreased by the fact that many edges are ignored during the sampling process.

Figures 13 and 14 moreover show that the self-similar properties of the $k$-core decomposition are preserved by the sampling process. Although the precise form of the degree distribution of the whole network is slightly altered, the basic correlation properties are conserved by the sampling. Moreover, the self-similar structure of the $k$-core...
decomposition is also preserved, as a comparison of Fig.s 13 and 14 with Fig.s 4 and 5 clearly shows.

While the main statistical properties of the $k$-core decomposition are therefore largely conserved by the sampling process, allowing to distinguish between networks with different topological structures, important quantitative biases can appear and compromise the accuracy of the measurements, as we now investigate. In order to understand such effects in more details, we indeed show in Fig.s 15 and 16 the probability for a vertex of given shell index in the original network to have another shell index in the sampled network, in the case of an original network obtained by the BRITE generator. At low sampling effort, many vertices are simply left undiscovered, and their index properties can be strongly affected in a seemingly erratic way, as shown by the important scattering of data in Fig. 15. As soon however as the sampling effort is increased to a more reasonable level, a strong correlation appears between the true shell index and its value in the sampled graph, even if a systematic downwards trend is observed (Fig. 16).

In summary, our results indicate that the sampling biases do in fact affect only slightly the measure of the statistical properties of heterogeneous graphs and of their $k$-core decomposition, even at relatively low level of sampling. In fact, the routing properties as “measured” by the shell indices will be in fact rather underevaluated due to the incomplete sampling of edges, which can be taken as a rather good news showing that the AS network probably offers better performance (QoS, robustness) than what can be measured by the present maps.

VI. CONCLUSIONS

We have presented the application of the $k$-core decomposition to the analysis of large scale networks models and of large scale Internet maps. The $k$-core decomposition allows the progressive pruning of the networks and the identification of subgraphs of increasing centrality. These subgraphs have the property of being more and more densely
FIG. 15: The grayscale code gives the probability of a change in shell index due to the traceroute-like sampling, from a certain index before sampling (x axis) to another one after sampling (y axis). The line at y = 0 represents the probability of vertices of shell index x to be absent from the sampled graph. The initial network is obtained by the BRITE generator. Here $N_S = 50$ sources and a fraction $N_T/N = 2 \times 10^{-3}$ of targets are used.

FIG. 16: Same as Fig. 15 for $N_S = 50$ sources and $N_T/N = 2 \times 10^{-2}$ (top) and $N_T/N = 10^{-1}$ (bottom).

connected, and therefore of presenting more and more robust routing capabilities. The study of the obtained subgraphs uncovers the main hierarchical layers of the network and allows for their statistical characterization. Strikingly, we observe for random heterogeneous networks and for the Internet at the Autonomous System a statistical self-similarity of the topological properties for cores of increasing centrality.

The $k$-core decomposition proves useful to uncover not only the hierarchical decomposition of real maps, but also for model validations. For example, many models, although having, e.g. degree distribution and clustering properties similar to those of real maps, do not present shell index values as large as the real data, nor a similar structure in which each $k$-core is composed by a constant fraction of the $k-1$-core. On the other hand, such a structure can appear even for uncorrelated random scale-free networks, which warns about the limitations of a modeling strategy based uniquely on the $k$-core structure. The $k$-core decomposition should therefore be considered as a supplementary
valuable tool for network characterization and model validation.

It is also worth mentioning that the router level $k$-core structure of the Internet appears to have different properties than those appearing at the AS level [24, 27]. This calls for repeating the present analysis for different router level maps available at the moment in order to better emphasize the structural difference exhibited by the two different mapping granularities.

Moreover, the $k$-core analysis allows to compare maps obtained by different mapping processes, follow their temporal evolution and assess the stability of these maps. It also appears as an interesting way of discriminating between various topologies, even after sampling biases have been introduced: for example, a sampled ER network may display a power-law like degree distribution in case of a very limited sampling effort, but its $k$-core decomposition will in any case remain very different from the one of sampled heterogeneous networks.

Finally, the $k$-core decomposition may be used also to define a computational feasible centrality measure and a hierarchy between the nodes of a network. It combines the degree ranking with more global structural properties, connectedness and routing capabilities, providing a centrality measure that is highly correlated with the various standard definitions such as degree and betweenness centrality. Moreover, the $k$-core decomposition presents a useful hierarchical classification of the ASes, in an adaptive and complementary manner to the tiers hierarchy, and may be used in other communication networks.

In conclusion, the $k$-core decomposition appears at a general level as a very interesting and useful additional tool for analysis of complex networks, with particular relevance in the context of technological and communication networks.

VII. ACKNOWLEDGMENTS

This work has been partially supported by the EU within the 6th Framework Programme under contract 001907 (DELIS). We acknowledge also interesting discussions with Pierre Fraigniaud.


