k-dense Communities in the Internet AS-Level Topology

E. Gregori, L. Lenzini, C. Orsini

IIT TR-21/2010

Technical report

Luglio 2010

Istituto di Informatica e Telematica
k-dense Communities in the Internet AS-Level Topology

Enrico Gregori∗ Luciano Lenzini† Chiara Orsini‡

ABSTRACT
Extracting a set of well connected subgraphs as communities from the Internet AS-level topology graph is crucially important for assessing the performance of protocols and routing algorithms, for designing efficient networks, and for evaluating the impact of failures. A huge number of community extraction methods have been proposed in the literature, among which the $k$-core decomposition and the $k$-clique community extraction methods. The former method is computationally efficient, but it only discovers coarse-grained and loosely connected communities. On the other hand, $k$-clique can extract fine-grained and tightly connected communities, but is NP hard and therefore useless for analyzing the Internet AS-level topology graph. In the paper we investigate the Internet structure by exploiting an efficient algorithm for extracting $k$-dense communities, where a $k$-clique community implies a $k$-dense community, which in turn implies a $k$-core community.

The paper provides two innovative contributions. The first is the application of the $k$-dense method to the Internet AS-level topology graph - obtained from the CAIDA, DIMES and IRL datasets - to identify well-connected communities and to analyze how these are connected to the rest of the graph.

The second contribution relates to the study of the most well-connected communities with the support of two additional datasets: a geographical dataset (which lists, for each AS, the countries in which it has at least one geographical location) and the IXP dataset (which maintains, for each IXP, its geographical position and the list of its participants). We found that the $k$-max dense community holds a central position in the Internet AS-level topology graph structure since its 101 ASs (less than the 0.3% of Internet ASs) are involved in more than 39% of all Internet connections. We also found that those ASs are connected to at least one IXP and have at least one geographical location in Europe (only 70.3% of them have at least one additional geographical location outside Europe).

1. INTRODUCTION AND RELATED WORKS

In order to get insight into the Internet structure at the AS level of abstraction, we employ the concept of community which is informally defined as “an unusually densely connected set of ASs” ([16]). Such communities quite often shed light on the structure of graphs or underlying properties of the graph nodes. Detecting communities is largely used in sociology, biology and computer science where systems are often represented as graphs. To the best of our knowledge this is the first time in which the communities are exploited for discovering structural properties of the Internet AS-level topology graph. The task of detecting communities in a graph is very hard for at least two reasons. First, there is no formal definition of a community and second because most of the algorithms are NP-hard. Other problems may also arise both from the possible occurrence of hierarchies, i.e. communities which are nested inside larger communities, and by the existence of overlaps between communities, due to the presence of nodes belonging to more groups. For a comprehensive description of the state of the art related to community discovery methods see [8]. The rational behind our decision to exploit communities for the Internet analysis was based on the findings other disciplines benefit from identifying community structure in graphs they are commonly dealing with, i.e.:

1. Frequently, the nodes in a community share a specific real-world property, e.g. for social networks, this could be a common interest while for web pages, this could be a common topic or language. Thus, by analyzing communities, one can infer semantic attributes.
2. By identifying communities, one can carry out focused analysis for communities individually. Different communities often exhibit significantly different properties which however are blurred by a global analysis. On the other hand, a more focused analysis of single communities may lead to more deepen or meaningful insights, for instance into the roles of individuals.

3. Conversely, each community can be “collapsed” into a single “meta-node”, allowing the design of a graph at a higher level of abstraction or equivalently at a coarser level, and this in turn give up a focus on higher-level structure.

For a much more detailed discussion of these and other motivations, see for instance [14]. Due to the great importance of identifying community structure in graphs, there has been a large amount of work in computer science, physics, economics, and sociology (for some examples, see [8], [14], [7], [9], [13], [6]).

In most of the approaches published in the specialized literature, communities have been characterized and discovered by exploiting some global property of the graph, like betweenness, modularity, etc. However, communities can be also interpreted as a form of local organization of the graph, so they could be defined from some property of the groups of vertices themselves, regardless of the rest of the graph. Moreover, very few algorithms are able to deal with the problem of overlapping communities. A method that accounts both for the locality of the community definition and for the possibility of having overlapping communities is the Clique Percolation Method (CPM) by Palla et al. [15]. A number of concepts were introduced Palla et al. as a support for specifying a k-clique community. Specifically, two k-cliques are adjacent if they share k1 vertices. The union of adjacent k-cliques is called k-clique chain. Two k-cliques are connected if they are part of a k-clique chain. Finally, a k-clique community is the largest connected subgraph obtained by the union of a k-clique and of all k-cliques which are connected to it. Unfortunately, the k-clique method by [15] requires a huge amount of computation for the Internet AS-level topology graph, i.e. it is NP-hard and, hence, it is intractable considering the structure and dimension of the Internet AS-level topology graph. A literature survey reveals the presence of two other community detection algorithms whose aim is to detect well-connected zone of the graph: the k-core decomposition [17] and, very recently, the k-dense method [16]. The study of the Internet AS-level topology graph structure through the [17] technique has been conducted in [2], [5] and [3]. On the other hand, [16] method has been applied to a Blog Trackback Network, to a Word Association Network and to the Wikipedia Reference Network, but, to the best of our knowledge, it has never been applied to the Internet AS-level topology graph. It is interesting to analyze the result of this community detection algorithm since it can be thought as an interpolation between the k-core decomposition (whose computational load is very low, but whose detected communities are too coarse-grained to detect specific properties of the constituent ASs) and the k-clique method. Interpreting the obtained communities with the support of additional information, likewise the participation to IXPs or the geographical location of the community members, helps in having a more conscious view of the Internet AS-level topology graph structure. The fundamental role of IXPs in Internet connectivity was recently proved by [10] and by [4]. A recent work on the geographical location of Internet Points of Presence (PoPs) is presented in [18].

The main contributions of this paper are summarized as follows.

- Investigate the completeness of previously known AS-level Internet topologies, and propose a method for “fusing” them together in order to capture a more “representative” snapshot of the actual Internet AS-level topology graph.
- Analyze structural characteristics of the Internet AS-level topology graph using the k-dense community detection method also by exploiting geographical information and statistics related to the IXPs.
- Evaluate the impact of IXPs on the structure of the Internet AS-level topology graph.

We discovered the existence of a small percentage (about 0.3%) of the ASes that forms a very dense community and are involved in a very large number of connections, i.e. 39.33% of Internet connections. We then concentrate on the study of this community integrating the topological graph with other important information: the geographical location of an AS and the connection of an AS with an hub and spoke infrastructure (i.e., an Internet Exchange Point, IXP in the following). We then discovered that this dense community is not homogeneously spread worldwide, it heavily exploits the IXP facilities (as it connects to 10.32 IXPs on average).

The paper is organized as follows. In Section 2 we present the datasets we will use in our structural analysis of the Internet AS-level topology graph. Specifically, we describe how we built the Internet AS-level topology graph, the IXP's dataset and the geographical dataset. In Section 3 we describe the tools we used to analyze the Internet AS-level topology graph. We describe the community detection algorithms we used and we show some parameters which let us to associate characteristics of the IXP dataset and the geographical dataset to ASs.
and connections of the Internet AS-level topology graph. In Section 4 we evaluate the results obtained applying the $k$-core and the $k$-dense community detection methods to the Internet AS-level topology graph. Here we demonstrate how, in the Internet case, the $k$-max-dense would be more tightly connected and central with respect to the $k$-max-core community. Moreover, with the support of the IXP dataset and the geographical dataset we identify some characteristics which are shared among $k$-dense ASs, i.e. all of them connect to, at least, one European IXP and they are likely to have geographical locations in more than in one country. Section 4 ends with a more detailed analysis of the connections crossing AMS-IX, DE-CIX and LINX IXPs since they have been indicated as possible causes of the creation of well-connected zone of the Internet AS-level topology graph. In Section 5 we summarize our conclusions.

2. DATA SOURCES

In this Section we introduce three different datasets: the Internet AS-level topology graph, the IXP dataset and the geographical dataset. In the next three Subsections we describe how they have been retrieved from public projects and how they have been built.

2.1 Internet AS-level Topology graph

Collecting a complete and up-to-date map of the AS-level Internet topology is a hot research topic. Currently, in the Internet there is no tool specifically designed to derive topology information, hence researchers have had to derive it using various indirect measurements. Currently, topology data are mostly gathered using traceroute-like methods (active probing) or BGP retrieval methods (passive measurement). Both these retrieval approaches are reliable but unfortunately they are largely incomplete and affected by biases ([12], [1]). In order to have a more detailed and un-biased view of the Internet, we decided to adopt the methodology described in [11]. Specifically, we built the Internet AS-level topology graph following this procedure:

1. We downloaded three public available datasets considering the measurement campaigns they performed in April 2010:
   - the IPv4 Routed /24 AS Links dataset\(^1\) (hereafter CAIDA) and the Distributed Internet Measurements and Simulations dataset\(^2\) (hereafter DIMES), which are based on traceroute-like methods;
   - the Internet Topology Collection at the Internet Research Lab dataset\(^3\) (hereafter IRL), which gathers topology information basing on static snapshots of the BGP routing tables and dynamic BGP data.

2. Then we merged them to obtain a single dataset.

3. Finally, we performed a data hygiene process. More in detail, we removed from the topology the connections which involved:
   - AS numbers declared as private by IANA\(^4\);
   - AS 23456 which, according to RFC 4893 is reserved and assigned for AS TRANS\(^5\);
   - AS 3130 which, according to the Cyclops website\(^6\), shows false AS adjacencies due to an experiment by Randy Bush\(^7\).

At the end of this procedure we obtained a dataset composed of 35,390 ASs and 152,233 connections. The pie chart in Figure 1 shows the origin data source of each connection composing the Internet AS-level topology graph and demonstrates how each single data source contributes to obtain a more detailed view of the graph. A more detailed description of this procedure can be found in [11].

2.2 IXP dataset

An IXP is a physical hub and spoke infrastructure, which enables ASs (participants\(^8\)) to exchange traffic

\(^{1}\)The IPv4 Routed /24 AS Links Dataset, Y. Hyun, B. Huffman, D. Andersen, E. Aben, M. Luckie, K.C. Claffy, and C. Shannon, \url{http://www.caida.org/}
\(^{2}\)http://www.netdimes.org/
\(^{3}\)http://irl.cs.ucla.edu/topology/
\(^{4}\)A full list of private ASN can be found on IANA website at \url{http://www.iana.org/assignments/as-numbers/as-numbers.xml}
\(^{5}\)It is used by to permit peering between a BGP speaker using 2-octet ASN and a BGP speaker using 4-octet ASN.
\(^{6}\)http://cyclops.cs.ucla.edu/blog/?m=200904
\(^{7}\)http://psg.com/173-174
\(^{8}\)ASs that are connected to at least one IXP will be termed as participants. We avoided using other terms, like members or
with each other as if they were connected directly via a physical link. There are financial advantages of an IXP for medium-sized ASs as they can avoid multiple ad-hoc point-to-point connection costs among participants, which are otherwise needed when BGP operates between (all or a subset of) them. IXPs also help Internet traffic to remain localized in the geographical region it belongs to. In fact, a large part of Internet traffic is directed inside national borders, since it is composed of language-dependent content (e.g. national music, websites, videos) and IXPs typically host a lot of regional ASs. This prevents traffic between regional ASs from passing through expensive connections (e.g. satellite connections in the African region or submarine fiber connections in the Australian region), which other than improving network performances, saves costs.

In order to highlight the presence of IXPs within the Internet AS-level topology graph structure, we built the IXP dataset, i.e. a dataset which maintains, for each IXP, its geographical position (i.e. the city and the country in which the IXP is placed) and the AS numbers of its participants. The dataset was built applying the following procedure:

1. We collected a potential list of all IXPs exploiting the information gathered in Packet Clearing House\(^9\) (PCH), peeringDB\(^10\), Euro-IX\(^11\) and bgp4.as\(^12\) websites.

2. We selected from the previous list only those IXPs which were active. Specifically, we verify each IXP activity observing its traffic statistics or observing the results obtained querying its looking glass server or observing the freshness of its website.

3. Then, for each active IXP, we gathered its geographical position (which is always present in the IXP website) and the list of its participants. This latter information was collected browsing the IXP website or parsing the results of the `show ip bgp summary` command executed on the IXP looking glass server. Those IXPs, for which it was not possible to collect the list of participants, were removed from the dataset.

At the end of this procedure, which was carried out in April 2010, we got a collection of 232 active IXPs from all over the world. In Table 1 we resume the results of our collection campaign.

The knowledge of an IXP participants list does not provide any information on the peering matrix, which customers, because these names depend on the IXP policies these ASs belong to.

<table>
<thead>
<tr>
<th>Continents</th>
<th># IXPs</th>
<th>Average IXP participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>12</td>
<td>9.83</td>
</tr>
<tr>
<td>Asia</td>
<td>32</td>
<td>28.34</td>
</tr>
<tr>
<td>Europe</td>
<td>108</td>
<td>45.63</td>
</tr>
<tr>
<td>Latin America</td>
<td>20</td>
<td>19.90</td>
</tr>
<tr>
<td>North America</td>
<td>44</td>
<td>40.75</td>
</tr>
<tr>
<td>Oceania</td>
<td>16</td>
<td>32.81</td>
</tr>
<tr>
<td>World</td>
<td>232</td>
<td>37.37</td>
</tr>
</tbody>
</table>

Table 1: Geographical list of IXPs found.

This assumption maximizes the number of connections going through IXPs, since it may include connections that do not cross any IXP. If the topology registers a connection between two ASs, and they both coexist on an IXP, their connection does not necessarily cross that IXP. This hypothesis is optimistic but also realistic. Assuming that two ASs are interconnected and are also participants in the same IXP, these ASs thus have the opportunity to set up their connection without deploying any additional private physical connection. From a financial standpoint, they establish their connection using the shared IXP. However, the main limitation of this hypothesis is that it cannot to detect whether the connections occur exclusively on IXPs or if they are also deployed outside these facilities.

### 2.3 Geographical Dataset

The addition of geographical location information to the Internet AS-level topology graph helps in interpreting those particular Internet subgraphs structures which are strongly driven by the local economy or the geographical distribution of backbone fibers (e.g. countries which reach the global Internet connectivity through costly satellite connections, for example, tend to form full-mesh like structure to help traffic to remain localized). In this Subsection we present the framework that we developed to associate a list of geographical loca-
tion to each AS basing on the MaxMind IP geolocation service. Specifically, we built the geographical dataset following this procedure:

1. We downloaded the GeoLite Country and the GeoLite ASN free databases from MaxMind website\(^{13}\). Both of them were uploaded in May 1st, 2010. The GeoLite Country database associates IPv4 addresses to country codes. The GeoLite ASN database maps IPv4 addresses to AS numbers.

2. We joined the GeoLite Country and the GeoLite ASN databases using the IPv4 address field. Thus, we obtained a database containing \(<\text{AS number}, \text{Country code}>\) tuples. Note that, for each AS number there could exists multiple country codes, hence the geographical database key is the entire tuple.

The resulting geographical database associates 34,190 ASs to, at least, one country code.

In Section 4 we need to provide a geographic attribute to any AS, according to the following taxonomy.

- An AS is called \textbf{national} AS if all of its geographical locations belong to the same country, i.e. its networks are placed within a single country.

- An AS is called \textbf{continental} AS if all of its geographical locations are placed within the same continent. For example, an AS is called European if its geographical locations belong to European countries and none of its geographical locations is placed outside Europe.

- An AS is called \textbf{worldwide} AS if it owns at least two geographical locations which are located in two different continents. For example, an AS which has one geographical location in Netherlands and one geographical location in the United States is referred to as worldwide AS.

3. \textbf{GRAPH PROPERTIES}

In this Section we review concepts and introduce notations which are relevant for the Internet AS-level topology graph analysis carried out in the paper. More specifically we define and compare two different community detection algorithms. For a comprehensive description of the algorithms specified therein, we refer the reader to [2] and [16].

\subsection{Community Detection Algorithms}

We start with outlining some basic definitions. For a graph \(G = (V_G, E_G)\), let \(V_G = \{1, ..., N\}\) be a set of nodes (or ASs) and \(E_G = \{e_1, ..., e_M\}\) a set of edges \(^{15}\)In this work we use GeoLite data created by MaxMind, available from \url{http://www.maxmind.com/}. (or connections), where \(e_m = \{i, j\} \subset V_G\) and \(i \neq j\), meaning that we focus on undirected graph without self-links. The set of adjacent nodes of node \(i\) (in the graph \(G\)) is defined as follows:

\[ F_G(i) = \{j : \{i, j\} \subset E_G\} \quad (1) \]

Equation 1 can be extended to a set of nodes as follows:

\[ F_G(V) = \bigcap_{i \in V} F_G(i) \quad (2) \]

\(F_G(V)\), referred to as the set of common adjacent nodes, is the set of nodes which are adjacent to all nodes in \(V\).

\subsubsection{k-core Community}

The subgraph \(C(k) = \{V_C(k), E_C(k)\}\) is called \(k\)-core if it satisfies the following requirements:

\[ V_C(k) = \{i : |F_C(k)(i)| \geq k\} \]

\[ E_C(k) = \{e_m : e_m \subset V_C(k)\} \quad (3) \]

The symbol \(S_C(k)\) denotes the number of the connected components of the \(k\)-core \(C(k)\). Each connected component of \(k\)-core \(C(k)\), i.e. \(C^*(k)\) \((1 \leq s \leq S_C(k))\), is referred to as \textbf{\(k\)-core community}. A node \(i\) is said to have a \(k\)-shell-index \(k\) if it belongs to the \(k\)-core but is not part of the \((k+1)\)-core. We will refer to the maximum \(k\)-shell-index as \(k\)-max. A detailed description of the \(k\)-core decomposition algorithm can be found in [2].

The \(k\)-core is a complex measure of node connectivity. To better explain this concept we describe the results of the \(k\)-core decomposition applied to simple topologies. For example, all the nodes belonging to a hub and spoke topology graph have a \(k\)-shell index of 1. The same result holds for nodes belonging to a perfect tree topology. On the other hand, if we consider a full mesh topology, all nodes belong to the \(k\)-max-core, where \(k\)-max is equal to the number of nodes in the graph minus 1. It follows that a network with a larger \(k\)-max will present a larger well-connected set of nodes, while a hierarchical network will tend to have a smaller \(k\)-max. At the same time, a \(k\)-shell index is not a measure of the centrality of the node. A low-degree node interconnecting a few high-degree hubs has a low \(k\)-shell index value, but intuitively it is in the center of the graph.

\subsubsection{k-dense Community}

Two nodes connected together by an edge do not necessarily imply that they belong to the same community unless there is a clear evidence or witness supporting a strong positive relation between them: the fact that they are just connected by a single link may not be strong enough. The existence of more common adjacent nodes in the same community suggests stronger positive relation. The \(k\)-dense community concept is based on this intuition. The subgraph \(D(k) = \{V_D(k), E_D(k)\}\) of
$G$ is called $k$-dense ([16]) if:

$$V_{D(k)} = \bigcup_{e_m \in E_{D(k)}} e_m$$

$$E_{D(k)} = \{ e_m : |F_{D(k)}(e_m)| \geq k - 2 \}$$

where $e_m$ denotes the edge connecting nodes $\{i, j\}$. The symbol $S_{D(k)}$ indicates the number of the connected components of the $k$-dense $D(k)$. Each connected component of $k$-dense $D(k)$, i.e. $D^s(k)$ $(1 \leq s \leq S_{D(k)})$, is referred to as $k$-dense community. A node $i$ is said to have a $k$-dense-index $k$ if it belongs to the $k$-dense but is not part of the $(k+1)$-dense. We will refer to the maximum $k$-dense-index as $k$-max. A detailed description of the $k$-dense algorithm can be found in [16].

$k$-core and $k$-dense concepts can be formally correlated ([16]). $k$-dense implies $(k-1)$-core, i.e. $D(k) \subset C(k-1)$. Since the subgraph $D(k)$ is obtained considering all the edges whose endpoints share, at least $k - 2$ neighbors, it follows that each node in $D(k)$ has, at least, $k - 1$ neighbors and hence is part of the $(k-1)$-core. Formally:

$$D(k) \subset C(k-1)$$

Although $k$-core and $k$-dense pruning algorithms\(^\text{14}\) are able to individuate well-connected zones, the $k$-dense definition seems to better fit the idea of community. Generally speaking, nodes belonging to the same community should share properties: while $k$-core requires, for each node, the presence of at least $k$ connections to the other $k$-core nodes, $k$-dense imposes the presence of common neighbors and hence, suggests a stronger relationship between nodes of the same community. To better appreciate this, we report in Figure 2 the sample network proposed in [16].

\(^\text{14}\)$k$-core and $k$-dense communities can be individuated through a graph pruning process that involves nodes and edges respectively.

### 3.1.3 Computational Complexity

$k$-core method complexity is very low, i.e.$O(n + e)$ where $n$ is the number of nodes and $e$ is the number of edges composing the graph. Nevertheless, communities obtained with this method are coarse-grained and loosely-connected. The $k$-dense algorithm, whose computation time is comparable with this obtained by the $k$-core method, is an interesting trade off between the $k$-clique and the $k$-core methods\(^\text{15}\). The $k$-clique algorithm, which is cited in Section 1, is able to identify subgraphs which are more tightly connected than those found by the $k$-dense algorithm, nevertheless since it is $NP$-hard, it is intractable on this paper environment. From a connectivity point of view, $k$-dense algorithm finds communities which are more densely-connected than the correspondent $(k-1)$-core communities. Moreover, since it requires the presence of $k-2$ common neighbors between each pair of connected nodes, it helps in isolating nodes which are part of separated densely-connected zones.

### 4. STRUCTURAL PROPERTIES OF THE INTERNET AS-LEVEL TOPOLOGY GRAPH

In this Section we employ both the $k$-core and the $k$-dense approaches to investigate the main structural properties of the Internet AS-level topology graph. To assess the impact of the connections crossing IXPs and the AS geographical location on the most dense Internet communities, the IXP dataset and the geographical dataset are used.

#### 4.1 General Features

In this Subsection we present the characteristics of the Internet AS-level topology graph which can be deduced exploiting the IXP dataset and the geographical dataset.

\begin{table}[h]
\begin{tabular}{|c|c|c|}
\hline
 & ASs & Connections \\
\hline
Internet & 35,390 & 152,233 \\
On IXP & 4,462 & 55,411 \\
\hline
\end{tabular}
\end{table}

Table 2: Internet features related to the IXP dataset.

In Table 2 we present the number of ASs and connections that compose the Internet AS-level topology graph and we compare both these figures to the number of ASs on IXP and to the number of connections on IXP (according to the hypothesis in Subsection 2.2). It is interesting to observe that the percentage of connections

\(^\text{15}\)See Higher level $k$-dense community dissertation in [16] for more details.
crossing IXPs (i.e. 36.40%) is much higher than the percentage of ASs on IXP (i.e. 12.6%). This result suggests that IXPs play an important role in the Internet structure (at least at the AS-level) as they seem to contribute to create a lot of connections. To further analyze the properties of the Internet AS-level topology graph, we now present, in Table 3, the geographical distribution of ASs according to the taxonomy presented in Subsection 2.3. Results presented in Table 3 indicate that the vast majority of Internet ASs has a national scope. Only the 7.58% has a worldwide scope or continental scope.

<table>
<thead>
<tr>
<th>ASs</th>
<th>ASs percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worldwide</td>
<td>1,568</td>
</tr>
<tr>
<td>Continental</td>
<td>1,115</td>
</tr>
<tr>
<td>National</td>
<td>31,228</td>
</tr>
<tr>
<td>Unknown</td>
<td>1,479</td>
</tr>
</tbody>
</table>

Table 3: Internet features related to the geographical dataset.

This distribution can be explained by the existence of many ASs that are not service providers but are interested in connecting to other ASs to obtain a connectivity to the Internet, i.e. ASs that do not transit traffic for other ASs and hence are likely to be customers in provider-customer relationships. These types of ASs are national ASs unless a continental or a worldwide presence is required by their own business. The last row, the one tagged with unknown, identifies the percentage of ASs whose geographical location has not been inferred by MaxMind. This depends on MaxMind information retrieval methods. In our future works, we plan to use other datasets (e.g. MaxMind GeoIP Country, IPligence, hostip.info) to have a complete ASs coverage.

### 4.2 k-core Results

The application of the k-core decomposition algorithm to the Internet AS-level topology graph indicates that this graph has a k-max equal to 75. Since each k-core is composed of a single connected component, the k-core and the k-core community terms identify the same subgraph. Briefly, k-core decomposition identifies 75 nested communities. In order to better understand the role of these communities within the whole Internet structure, we plot in Figure 3 the number of distinct connections which involve at least one AS with a k-shell-index equal to k.

In Figure 3 we can easily identify the presence of two groups of ASs which are involved in a very high number of connections: the first group is composed of ASs with a shell index equal to 2, 3, or 4, the second group is composed of the k-max-shell ASs. High values related to the first group can be explained considering that the ASs that are part of the 2,3,4 shells represent the 59.98% of the Internet AS-level topology graph.

Hence, it is obvious their involvement in so many Internet connections, i.e. this part of the graph is composed of a huge number of ASs with few connections. On the other hand, the second group has completely different characteristics, i.e. there are few nodes with a high number of connections. In this region there are only 167 ASs (i.e. the 0.47% of Internet ASs), but these nodes contribute with a very high number of connections, i.e. 72,698 (47.75% of Internet connections). This second characteristic indicates that these nodes are part of a very much central zone of the Internet. Due to its fundamental role in the Internet connectivity, we focus our attention exclusively on the k-max-core.

<table>
<thead>
<tr>
<th>ASs</th>
<th>Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>167</td>
</tr>
<tr>
<td>On IXP</td>
<td>167</td>
</tr>
</tbody>
</table>

Table 4: k-core features related to the IXP dataset.

As can be see in Table 4, the k-max-core is composed of 167 ASs and 8,896 connections. All these ASs have, at least one connection to an IXP\(^{16}\), moreover, the 97.80% of connections are on IXP connections. These two considerations highlight that IXPs have a fundamental role in the creation of this well-connected zone of the graph identified by the 75-core.

<table>
<thead>
<tr>
<th>ASs</th>
<th>ASs percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worldwide</td>
<td>108</td>
</tr>
<tr>
<td>Continental</td>
<td>45</td>
</tr>
<tr>
<td>National</td>
<td>14</td>
</tr>
<tr>
<td>Unknown</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5: k-core features related to the geographical dataset.

In Table 5 we present the geographical scopes related to the k-max-core ASs. This Table outlines that the\(^{16}\)More in detail, each k-max-core AS is connected, on average, to 9.12 IXPs whose average number of participants is 187.73.

Figure 3: Number of connections vs. k-shell-index.
vast majority of the 75-core ASs have a worldwide or a continental scope, only the 8.38% of ASs has a national scope. Please note the differences between the geographical properties of the Internet AS-level topology graph (see Table 3) and those related to the \textit{k-max}-core ASs (see Table 5).

To summarize, the \textit{k-max}-core community is composed of: a) ASs connected to, at least, one IXP; b) connections that, with a high probability, are deployed through the IXP facilities; c) ASs that, with high probability, have geographical location in different countries (which often belong to different continents). A more detailed analysis of the \textit{k-core} properties of the Internet AS-level topology graph can be found in [11].

### 4.3 \textit{k-dense} Results

The application of the \textit{k-dense} method to the AS-level Internet topology graph produces a \textit{k-max} equal to 47. Each \textit{k-dense} subgraph is composed of a single connected component\(^{17}\), hence the \textit{k-dense} subgraph and the \textit{k-dense} community identify the same subgraph. Briefly, \textit{k-dense} method identifies 47 nested communities. To better understand the role of these communities within the whole Internet AS-level topology structure, we plot in Figure 4 the number of distinct connections that involve at least one AS with a \textit{k-dense}-index equal to \(k\).

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig4}
\caption{Number of connections vs. \textit{k-dense}-index.}
\end{figure}

Similarly to the \textit{k-core} case, we can appreciate the presence of two groups of ASs involved in a very high number of connections: the first group is composed of ASs with a low \textit{k-dense}-index, the second group is composed of the \textit{k-dense}-max ASs. It is worth noting the weight of the 3-dense ASs within the first group (3-dense ASs are involved in 32,155 connections). This can be explained considering that very often the two

\textit{ASs} | \textit{Connections} \\
\hline
Internet | 101 | 4,056 \\
On IXP | 101 | 4,011 \\
\end{table}

Table 6: \textit{k-dense} features related to the IXP dataset.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\textit{ASs percentage} & \textit{ASs} \\
\hline
Worldwide & 71 & 70.30% \\
Continental & 26 & 25.74% \\
National & 4 & 3.96% \\
Unknown & 0 & 0% \\
\hline
\end{tabular}
\caption{\textit{k-dense} features related to the geographical dataset.}
\end{table}

Moreover, the 99.90% of \textit{k-max}-dense connections are deployed through the IXP facilities. Hence, also in this case, IXPs have a fundamental role in the creation of this well-connected zone of the Internet AS-level graph.

### 4.4 \textit{k-max-core} vs. \textit{k-max-dense}

The analysis of the \textit{k-max-core} and the \textit{k-max-dense} communities yields quite similar results. Specifically, their central position within the Internet AS-level topology graph and their reliance on connections crossing IXPs are evident. Relationships described in Equation 5 indicate that the 47-dense community is a subgraph of the 46-core. Surprisingly, we found that the entire 47-dense community is a subgraph of the 75-core community (see Figure 5). This indicates that the \textit{k-max-dense} community is a tightly connected zone of the \textit{k-max-core}. On the other hand, 75-core ASs take \textit{k-dense} indices in the interval [39:47].

It is also interesting to observe that the 47-dense community ASs are more central than the 75-core community

\footnote{More in detail, each \textit{k-max-dense} AS is connected, on average, to 10.32 IXPs whose average number of participants is 178.43.}
ASs. This result is obtained observing the complementary cumulative distribution curves plotted in Figure 6. For this reason, in the following, we will focus on the 47-dense subgraph only.

4.5 k-max-dense analysis

As already stated, the $k$-max-dense subgraph is composed of 101 ASs. Here we deepen the analysis by focusing on the connections belonging to the 47-dense subgraph. In Figure 7 we plot the 47-dense degree, i.e. the number of connections in the 47-dense subgraph of each 47-dense subgraph node.

$k$-dense degree values shown in Figure 7 confirm the presence of a high level of connectivity within the 47-dense community. All the ASs have $k$-dense degree values in the interval $[60:100]$, meaning that each $k$-max-dense AS is connected to, at least, other 60 ASs of the $k$-max-dense. Moreover, there are 2 ASs which are connected to all the $k$-max-dense ASs. The set of 30 ASs which have a national or a continental geographical scope (see Table 7) is composed exclusively of European ASs (and are connected to, at least one, European IXP).

On the other hand, the remaining 71 ASs can be partitioned in two groups: 12 are connected to European IXPs only, and the remaining 59 are connected to, at least, one European IXP, and to another IXP outside Europe. These considerations lead us to focus our attention on European IXPs, since the connection to them is a shared property among all the $k$-max-dense community ASs. To this end we consider the top-10 European IXPs (i.e. the 10 IXPs with the highest number of participants) and we apply the $k$-dense method to their subgraphs. Specifically, we define the subgraph of an IXP as the network composed of all those connections that have been tagged as crossing the considered IXP by the hypothesis in Subsection 2.2. We summarize the results of this process in Table 8.

<table>
<thead>
<tr>
<th>IXP subgraph</th>
<th>Size</th>
<th>$k$-max</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINX (UK)</td>
<td>347</td>
<td>41</td>
</tr>
<tr>
<td>DE-CIX (DE)</td>
<td>332</td>
<td>42</td>
</tr>
<tr>
<td>AMS-IX (NL)</td>
<td>324</td>
<td>42</td>
</tr>
<tr>
<td>MSK-IX (RU)</td>
<td>293</td>
<td>20</td>
</tr>
<tr>
<td>NL-IX (NL)</td>
<td>232</td>
<td>18</td>
</tr>
<tr>
<td>PaNAP (FR)</td>
<td>150</td>
<td>17</td>
</tr>
<tr>
<td>PL-IX (PL)</td>
<td>141</td>
<td>8</td>
</tr>
<tr>
<td>KleyReX (DE)</td>
<td>129</td>
<td>13</td>
</tr>
<tr>
<td>Packet Exchange (UK)</td>
<td>119</td>
<td>7</td>
</tr>
<tr>
<td>SwissIX (CH)</td>
<td>117</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 8: European IXPs rank. For each IXP: the size column indicates the number of participants; the $k$-max column indicates the maximum $k$-dense-index.

Table 8 clearly highlights the presence of three IXPs whose $k$-max-dense-index is close to the overall Internet $k$-max-dense-index: LINX, DE-CIX and AMS-IX. MSK-IX, despite of the comparable number of participants,
has a $k$-max-dense value that is, approximately, half that of AMS-IX. Before focusing on the three IXPs with a very high $k$-max-dense value, we need to underline that the $k$-max-dense values obtained strictly depend upon the considered Internet AS-level topology graph and the hypothesis in Subsection 2.2 which led us to generate IXP topologies (or subgraphs). To be more precise, we focus on the MSK-IX case. As stated in its website\textsuperscript{19}, MSK-IX participants can have the use of the Route Server in order to reduce the number of individually administered peering sessions. Since this network service (i.e. the Route Server) retransmits BGP announcements between the connected participants, all the connected participants have a peering relationship with each others. From the Internet AS-level topology graph point of view, this reflects in a star topology in which the AS of the Route Server represents the center. Thus, since many connections among MSK-IX participants could be hidden by this network service, this could be the reason why MSK-IX subgraph has a low $k$-max-dense index. Generally speaking, the presence of the route server AS numbers within the path which describe two participants connection could significantly reduce the number of connections crossing the IXPs that are identified by our hypothesis\textsuperscript{20}. This service (the route server) is a common characteristic of the IXPs, indeed, 7 out of 10 IXPs listed in Table 8 offer the route server service to their participants. In our future works we plan to extend the AS-level topology dataset to better match the full mesh connectivity that can be obtained through the Route Server.

Back to the analysis, we show in Table 9 and in Table 10 the number of ASs and connections that are shared between the Internet $k$-max-dense and the IXPs $k$-max-denses. More concretely, the cell individuated by AMS-IX row and the Internet column (in Table 9) contains the number of ASs (or connections, if we consider Table 10) which are present both in the AMS-IX $k$-max-dense (i.e. the 42-dense individuated applying the $k$-dense method to the IXP subgraph) and in the Internet $k$-max-dense (i.e. the 47-dense individuated applying the $k$-dense method to the Internet topology). All the other cells values can be computed following the same procedure (both for ASs and connections).

Observing Table 9 and Table 10 values indicate that the Internet $k$-max-dense and the three IXPs $k$-max-denses have a big common part both in terms of ASs and in terms of connections. More in detail, AMS-IX, DE-CIX and LINX $k$-max-denses share with the Internet $k$-max-dense the 81.18\% , the 69.31\% and the 75.25\% of ASs and the 68.24\%, the 51.24\% and the 58.2\% connections respectively. These percentages demonstrate that the geographical and IXP information exploited were a good indicator, in other words, focusing on the European IXPs subgraphs leads us to individuate one of the possible causes of the creation of the $k$-max-dense zone\textsuperscript{21}. The $k$-max-denses of these three IXPs share 52 ASs (or participants). All these ASs are also member of the Internet $k$-max-dense. Thus, it is interesting to observe the relationship between the $k$-max-dense resulting from the merge of AMS-IX, DE-CIX and LINX subgraphs, and the Internet $k$-max-dense. We will refer to this merge topology as ADL (AMS-IX DE-CIX and LINX acronym). ADL topology has a $k$-max-dense-index equal to 46. In Table 11 and 12 we show the number of ASs and connections which are both in the ADL 46-dense and in the Internet 47-dense. Data in Table 11 and in Table 12 are computed as in Table 9 and in Table 10.

As expected, ADL $k$-max-dense share with the Internet $k$-max-dense a high percentage of ASs and connections. More in detail, it shares the 89.11\% of ASs and the 82.25\% of connections. These results indicate that AMS-IX, DE-CIX and LINX are an important aggregation point which play a fundamental role in the formation of such well-connected zone represented by

\textsuperscript{19}http://www.msk-ix.ru/eng/routeserver.html

\textsuperscript{20}The presence of $N$ connections to the route server in the topology, could hide $N \cdot (N - 1) / 2$ peering connections that are really established by the IXP participants.

\textsuperscript{21}We cannot assure IXPs are the only cause of the creation of the $k$-max-dense zone. Connections which are considered as crossing IXP can be also deployed elsewhere. We expect this could be a very frequent circumstance since most of the ASs we are considering (i.e. ASs belonging to the Internet $k$-max-dense) have a worldwide scope and, hence, are likely to be interested in short-cutting their traffic wherever possible.

<table>
<thead>
<tr>
<th></th>
<th>I (47)</th>
<th>A (42)</th>
<th>D (42)</th>
<th>L (41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (47)</td>
<td>101</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (42)</td>
<td>82</td>
<td>86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (42)</td>
<td>70</td>
<td>61</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>L (41)</td>
<td>76</td>
<td>64</td>
<td>60</td>
<td>107</td>
</tr>
</tbody>
</table>

Table 9: Number of shared ASs among $k$-max-denses. I refers to Internet, A refer to AMS-IX, D refers to DE-CIX, L refers to LINX. Numbers in parentheses are the $k$-max-dense-indices.

<table>
<thead>
<tr>
<th></th>
<th>I (47)</th>
<th>A (42)</th>
<th>D (42)</th>
<th>L (41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (47)</td>
<td>4,056</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (42)</td>
<td>2,768</td>
<td>3,013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (42)</td>
<td>2,104</td>
<td>1,574</td>
<td>3,149</td>
<td></td>
</tr>
<tr>
<td>L (41)</td>
<td>2,363</td>
<td>1,672</td>
<td>1,478</td>
<td>4,162</td>
</tr>
</tbody>
</table>

Table 10: Number of shared connections among $k$-max-denses. I refers to Internet, A refer to AMS-IX, D refers to DE-CIX, L refers to LINX. Numbers in parentheses are the $k$-max-dense-indices.
<table>
<thead>
<tr>
<th></th>
<th>Internet (47)</th>
<th>ADL (46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet (47)</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>ADL (46)</td>
<td>90</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 11: Number of shared ASs between Internet and ADL k-max-denses. Numbers in parentheses are the k-max-dense-indices.

<table>
<thead>
<tr>
<th></th>
<th>Internet (47)</th>
<th>ADL (46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet (47)</td>
<td>4,056</td>
<td></td>
</tr>
<tr>
<td>ADL (46)</td>
<td>3,336</td>
<td>3,659</td>
</tr>
</tbody>
</table>

Table 12: Number of shared connections between Internet and ADL k-max-denses. Numbers in parentheses are the k-max-dense-indices.

If we consider the Internet line (Figure 8) we can observe that the major part of the ASs belong to lower k-denses while k-denses whose k-dense-index is greater than 10 represent a percentage of Internet ASs that is lower than the 5%. From a structural point of view, this means that ASs composing Internet are mostly: a) ASs with a low degree value that are necessarily members of the low k-denses, or b) ASs with no or few common neighbors with the ASs they are connected to. The high k-dense values, i.e. those close to 47, indicate that there exist zones that are densely-connected. Nevertheless, these zones are created by a small minority of the Internet ASs (as stated before less than the 5% of nodes has a k-dense index higher than 10).

If we consider the results relative to the IXP topologies (Figure 8), we observe four similar trends which are really different from the Internet trend we have just described. If we concentrate on AMS-IX, DE-CIX and LINX results we can see that the vast majority of the participants of these IXPs has high k-dense-index values. Hence, the well-connected zones that can be individuated within these IXPs subgraphs are developed by an high percentage of these IXPs participants, in contrast to the result relative to the Internet AS-level topology graph. In other words, almost all of the participants contribute to the formation of well-connected zones.

If we consider the ADL trend is similar to the AMS-IX, DE-CIX and LINX trends. It is interesting to observe that the merge of the three IXP dataset is able to develop a 46-dense, i.e. a zone that is more densely-connected than the those that can be obtained from the single IXP subgraphs. This phenomenon can be explained noting that there are 52 participants which are common to the three IXPs.

5. DISCUSSION AND CONCLUSIONS

In this work we have analyzed structural characteristics of the Internet AS-level topology graph using the k-dense community detection method also by exploiting geographical information and statistics related to the IXPs. The decomposition of the Internet AS-level topology graph through the k-core and the k-dense methods yields high k-max-index values, 75 and 47 respectively. From a structural point of view, this confirms the results shown in [11], i.e. the Internet structure is somehow hierarchical with densely-connected zones created by horizontal connections. We were able to assess the deviation from a hierarchical structure by exploiting a novel approach for the Internet environment, i.e. the k-dense communities. Specifically, the k-dense analysis (Section 4) shows that the k-max-dense subgraph represent the 0.28% of Internet ASs and the 2.66% of the Internet connections. Despite of these visibly low percentages, this set of ASs is involved in the 39.33% of the Internet connections. Thus, the importance and the centrality...
of these ASs are evident. It is also interesting to observe that the 47-dense (i.e. k-max-dense) represents a subgraph which is more tightly connected and central (see vertex betweenness distributions in Section 4) than the 75-core (i.e. the k-max-core). The exploitation of the geographical dataset reveals that, while the typical Internet ASs has a national scope, ASs belonging to the k-max-dense are, on the contrary, likely to be present in many countries. Moreover, our analysis of the 47-dense shows that all the ASs which are part of this subgraph have a geographical location in Europe and connect to, at least, one European IXP. By further deepening the analysis, we showed (in Subsection 4.5) that the three main European IXPs, i.e. AMS-IX, DE-CIX and LINX, are possible causes of the creation of such well-connected zone that is individuated by the 47-dense. This assertion can be supported considering that the ADL k-max-dense (i.e. the k-max-dense obtained applying the k-dense method to the merge of AMS-IX, DE-CIX and LINX subgraphs) shares the 89.11% of ASs and the 82.25% of connections with the k-max-dense. The importance of AMS-IX, DE-CIX and LINX facilities within the Internet structure is not unexpected. Both AMS-IX and LINX, indeed, are not so far from oceanic backbone fiber landing points. On the other hand, DE-CIX is one of the largest exchange point in terms of peak traffic\footnote{As of May 16, 2010, the all time peak of incoming traffic was 944.839 Gbit/s while the all time peak of outgoing traffic was 942.796 Gbit/s. Other AMS-IX statistics can be found at http://www.ams-ix.net/statistics/}.

The reliability of our results depends upon the accuracy of the datasets considered and the validity of IXP hypothesis regarding connections crossing it. As stated in Subsection 4.5, information related to connections crossing IXPs should be interpreted carefully. Although it is reasonable to assume that any two ASs which are sharing a peering session and which are present on the same IXP, connect to each other using the IXP facility, we do not know if they deploy their connection also out of the IXP. Moreover, we believe it is important to cope with the Route Server issue that has been presented in Subsection 4.5. Nevertheless, we want to underline that a more specific analysis of the IXPs connections which are deployed through the Route Server could amplify the relevance of the IXPs in the formation of well-connected zones.

In our future works, we plan also to extend our k-dense analysis to all the k-dense communities other than that regarding the k-max-dense community. In this way, it could be interesting to observe the Internet connectivity at a level of abstraction that is higher than the AS-level, in particular, we are interested in studying the connectivity considering the idea of a community-level Internet topology.

6. REFERENCES


