Computer Networks 57 (2013) 213–227

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

k-Dense communities in the Internet AS-level topology graph

Enrico Gregori [a], Luciano Lenzini [b,*], Chiara Orsini [a,b]

[a] Institute of Informatics and Telematics, Italian National Research Council, Pisa, Italy
[b] Department of Information Engineering, University of Pisa, Pisa, Italy

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form 10 September 2012
Accepted 13 September 2012
Available online 20 September 2012

Keywords:
Internet AS-level topology
Community detection
k-Dense
IXP

ABSTRACT

In this paper we investigate the structure of the Internet by exploiting an efficient algorithm for extracting *k*-dense communities from the Internet AS-level topology graph. The analyses showed that the most well-connected communities consist of a small number of ASs characterized by a high level of clusterization, although they tend to direct a lot of their connections to ASs outside the community. In addition these communities are mainly composed of ASs that participate at the Internet Exchange Points (IXPs) and have a worldwide geographical scope. Regarding *k-max*-dense ASs we found that they play a primary role in the Internet connectivity since they are involved in a huge number of Internet connections (42% of Internet connections). We also investigated the properties of three classes of *k-max*-dense ASs: Content Delivery Networks, Internet Backbone Providers and Tier-1s. Specifically, we showed that CDNs and IBPs heavily exploit IXPs by participating in many of them and connecting to many IXP participant ASs. On the other hand, we found that a high percentage of connections originated by Tier-1 ASs are likely to involve national ASs which do not participate at IXPs.

1. Introduction

Extracting a set of well connected sub-graphs as communities from the Internet AS-level topology graph enables researchers to gain more insight into the structure of the Internet. In this paper we investigate the structure of the Internet at the AS level of abstraction exploiting the concept of a community which is informally defined as "an unusually densely connected set of ASs" [1]. Such communities quite often shed light on the structure of graphs or the underlying properties of the graph nodes. For example, these dense zones of the Internet AS-level topology graph enable to find which classes of ASs are interested in interconnecting with each other. The rationale behind our decision to exploit communities for the Internet analysis was driven by the need to select a portion of the graph with similar properties.

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, it could be a common topic or language. Thus, by analyzing communities, it is possible to infer semantic attributes.
2. By identifying communities, it is possible to carry out a focused analysis for communities on an individual basis. Different communities often exhibit significantly different properties, which may get blurred in a global analysis. On the other hand, a more focused analysis of single communities may lead deeper or more meaningful insights, for instance into the roles of individuals.
3. Conversely, each community can be "collapsed" into a single "meta-node", thus enabling a graph to be designed 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 on this topic, see for instance [2].

[*] Corresponding author. Tel.: +39 050 568511.
E-mail addresses: enrico.gregori@iit.cnr.it (E. Gregori), l.lenzini@iet.unipi.it (L. Lenzini), chiara.orsini@iet.unipi.it (C. Orsini).

1389-1286/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2012.09.011

Structural properties provided by the *k*-dense analysis could be useful to test the validity of future Internet AS-level topology models. In addition, having a better understanding of the Internet structure allows engineers to design applications and protocols which can take into account the underlying network and thus can be more efficient than those which do not consider these information [3]. Consider, for example, routing on random graphs compared to routing on trees or grids, the best possible performance of routing on these structures are really different [4,5]. Internet-like topologies are particularly well-structured for routing efficiency [6,7], but current Internet routing architecture does not exploit this fact. In addition, a deep knowledge of Internet underlying structure could also help researchers in designing a more accurate model of the Internet topology at the AS-level of abstraction.

Interpreting the obtained communities with the support of additional information, such as the participation to Internet Exchange Points (IXPs) or the geographical location of the community members, helps in terms of a more conscious view of the Internet AS-level topology graph structure. In our work we used two datasets: the IXP dataset lists, for each IXP its set of participant ASs; and the geographical dataset which lists, for each AS, the set of countries where the AS is geographically located. The fundamental role of IXPs in Internet connectivity was recently highlighted by He et al. [8] and by Augustin et al. [9]. A recent work on the geographical location of Internet Points of Presence (PoPs) is presented in [10].

A huge number of community extraction algorithms have been proposed in the literature, including the *k*-core decomposition and the *k*-dense community detection algorithms. Both algorithms are computationally efficient, however the *k*-dense is able to discover more well-connected communities. When we use the term *well-connected* we refer to portion of the network whose link density is orders of magnitude greater than the overall link density (Internet link density is $\sim 2.4 \times 10^4$, please see Section 3 for link density definition).

In this work we extend our analysis of the Internet AS-level topology graph proposed in [11]. The main aims of this paper are to:

- Investigate the completeness of previously known AS-level Internet topologies, using the method shown in [12] in order to merge them together to capture a more "representative" snapshot of the actual Internet AS-level topology graph.
- Thoroughly analyze the structural characteristics of the Internet AS-level topology graph using the *k*-dense community detection algorithm by also exploiting geographical information and statistics related to the IXPs.
- Highlight the presence of IXPs within the most well-connected communities of the Internet AS-level topology graph.

The novelty of this paper is a detailed analysis of the characteristics of the ASs (i.e. companies) that form the most well-connected zone of the Internet AS-level topology according to the *k*-dense definition. We picked up some of the networks within the *k-max*-dense and we discussed the relation between their business profile and our additional datasets. Another feature which characterizes this paper is a comparison between different community detection algorithms when applied to the Internet AS-level topology graph. Finally, we also discuss how *k*-dense results could be affected by biases caused by the incompleteness of the Internet AS-level topology graph.

We discovered the existence of a small percentage of Internet ASs (about 0.3%) that form a very dense community and are involved in a very large number of connections, i.e. 42% of Internet connections. We then concentrated on studying this community by integrating the topological graph with our additional datasets, i.e. IXP and geographical datasets. We then discovered that this dense community is not homogeneously spread worldwide and heavily exploits the IXP facilities (as each AS of the community connects to, on average, 10.32 IXPs). Since business strategies can influence the geographical scope of an ASs and the decision to participate in IXPs, the exploitation of our additional datasets enabled us to unveil characteristics which are shared among nodes within a community. Specifically, we discussed the participation in many IXPs of several Content Delivery Networks (CDNs) and Internet Backbone Providers (IBPs) which make up the *k-max*-dense community. These ASs typically connect to other ASs participating in IXPs. On the other hand, we found that Tier-1 ASs tend to direct a high percentage of their connections to national ASs which are not likely to participate in IXPs.

The paper is organized as follows. In Section 2 we describe in detail how the *k*-dense community detection algorithm works. In Section 3 we describe the metrics we used to analyse the Internet AS-level topology graph. In Section 4 we describe how we derived the Internet AS-level topology graph, the IXPs dataset and the geographical dataset. We also introduced some new parameters which enabled us to create a taxonomy of Internet ASs exploiting the IXP and the geographical datasets. In Section 5 we evaluate the results that were obtained by applying the *k*-dense community detection algorithm to the Internet AS-level topology graph. We analyse *k*-dense communities using the metrics described in Section 3 and by exploiting the IXP and the geographical datasets (see Section 4). We then study the characteristics of the *k-max*-dense community since its ASs have been shown to play a very important role in the Internet connectivity. In Section 6 we present some related works concerning community detection algorithms and we provide a comparison between different community detection algorithms. In Section 7 we present our conclusions.

2. *k*-Dense method

In this Section we review concepts and introduce notations which are relevant for an analysis of the Internet AS-level topology graph carried out in Section 5. More specifically we describe the *k*-dense community detection algorithm and we provide a formal comparison to other algorithms. For a comprehensive description of the algorithms specified therein, see [13,1,14].

The *k*-dense community concept is based on the following intuition. Two nodes connected together by an edge do not necessarily imply that they belong to the same community unless there is 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 a stronger positive relation [1]. In other words, if two ASs share several neighbours they are likely to be part of a same community. In the following paragraphs we give a formal description of the community definition that implements this idea.

Firstly, Internet AS-level topology can be thought of as an undirected graph where nodes represent ASs and edges (or links) represent connections between ASs. Thus we can start by outlining some basic definitions. For a graph $G = (V_G, E_G)$, let $V_G = \{1, \dots, N\}$ be a set of nodes and $E_G = \{e_1, \dots, e_M\}$ a set of edges, where $e_m = \{i,j\} \subset V_G$ and $i \neq j$, which means that we focus on an undirected graph without self-links. The set of adjacent nodes of node i (in the graph G) is defined as follows:

$$F_G(i) = \{j : \{i,j\} \subset E_G\} \tag{1}$$

Thus $F_G(i)$ represents the set of ASs connected to AS i, i.e. the set of i neighbours. Expression (1) can be extended to a set of nodes (ASs) as follows:

$$F_G(V) = \bigcap_{i \in V} F_G(i) \tag{2}$$

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

Saito et al. [1] define the k-dense as the subgraph $D(k) = \{V_{D(k)}, E_{D(k)}\}$ of G with the following characteristics:

$$\begin{aligned} V_{D(k)} &= \bigcup_{\{i,j\} = e_m \in E_{D(k)}} \{i,j\} \\ E_{D(k)} &= \{e_m = \{i,j\} : |F_{D(k)}(\{i,j\})| \geqslant k-2\} \end{aligned} \tag{3}$$

where e_m denotes the edge connecting nodes $\{i,j\}$. $V_{D(k)}$ is the set of nodes having at least one connection within a given k-dense. $E_{D(k)}$ is the set of connections whose endpoints (i and j) have at least $k-2$ common neighbours. From a topological point of view, a k-dense connection is a part of the graph where at least k-2 triangles share a common base (see Fig. 1).

The symbol $S_{D(k)}$ indicates the number of connected components of the k-dense graph $D(k)$. Each connected component of k-dense $D(k)$, i.e. $D^s(k) (1 \leqslant s \leqslant S_{D(k)})$, 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, i.e.:

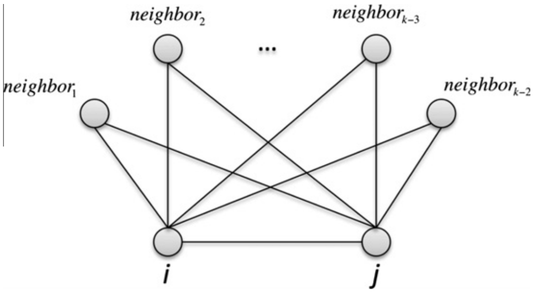

Fig. 1. A k-dense connection.

$$i \in V_{D(k)} \wedge i \notin V_{D(k+1)} \tag{4}$$

Moreover, we define a *k-dense-shell* as the set of nodes having a k-dense-index equal to k.

$$k\text{-}dense\text{-}shell = \{i : i \in V_{D(k)} \wedge i \notin V_{D(k+1)}\} \tag{5}$$

The maximum k-dense-index will be referred to as *k-max*. In Fig. 2 we report a sample network (firstly appeared in [11]) in order to better visualize how the k-dense works. Observing Fig. 2 we can conclude that: (a) the *k-max* is equal to 5; (b) there are 10 nodes with a k-dense-index equal to 5, there are 4 nodes with a k-dense-index equal to 3 and 4 nodes with a k-dense-index equal to 2. A detailed description of the k-dense algorithm can be found in [1].

k-Core and k-dense concepts can be formally correlated [1]. Since, k-core is a sub-graph $C(k) = \{V_{C(k)}, E_{C(k)}\}$ of G satisfying the following requirements:

$$\begin{aligned} V_{C(k)} &= \{i : |F_{C(k)}(i)| \geqslant k\} \\ E_{C(k)} &= \{e_m : e_m \subset V_{C(k)}\} \end{aligned} \tag{6}$$

it follows that k-dense implies *(k-1)*-core, i.e. $D(k) \subset C(k-1)$. Since the sub-graph $D(k)$ is obtained considering all the edges whose endpoints share, at least $k-2$ neighbours, it follows that each node in $D(k)$ has, at least, $k-1$ neighbours and hence is part of the *(k-1)*-core. Formally:

$$D(k) \subset C(k-1) \tag{7}$$

Although k-core and k-dense pruning algorithms[1] are able to identify well-connected zones (see Fig. 12), 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 neighbours and hence, suggests a stronger relationship between nodes of the same community. To better appreciate this we compared k-dense and in k-core in Fig. 2.

Fig. 2 shows that communities obtained by the k-dense method provide much more insight into the structural properties of the sample network than the single community provided by the k-core decomposition. It is significant that k-dense allows to separate the two complete subgraphs of order 5 (i.e. 5-clique), while k-core consider the whole graph as a single community.

Computational complexity. k-core algorithm 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 algorithm are coarse-grained and loosely-connected. The k-dense algorithm, whose computation time is comparable with the time obtained by the k-core algorithm, is an interesting trade off between the k-clique and the k-core algorithms[2]. The k-clique algorithm, which is cited in Section 6, is able to

[1] k-Core and k-dense communities can be found through a graph pruning process that involves nodes and edges respectively.

[2] See higher level k-dense community dissertation in [1] for more details.

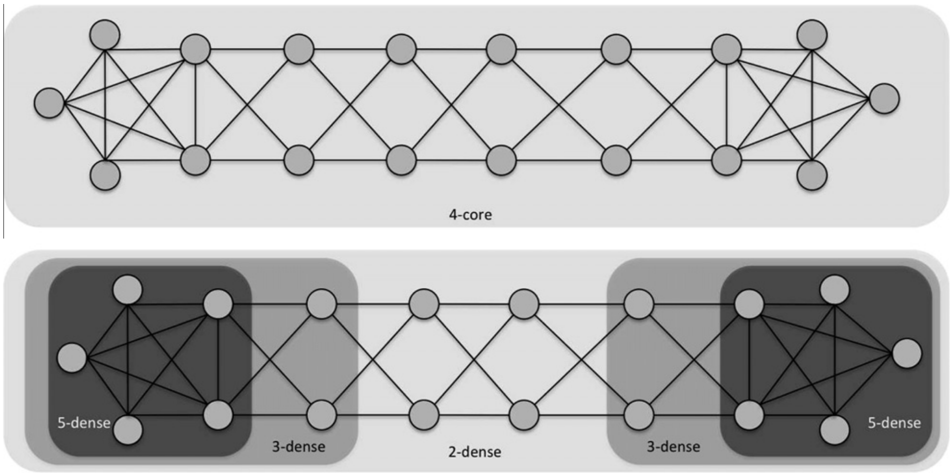

Fig. 2. Sample network.

identify sub-graphs which are more tightly connected than those found by the *k*-dense algorithm but it requires a huge amount of computational resources. This algorithm first finds all the maximal cliques within a graph, then it extracts communities from the set of maximal cliques found. We were able to compute all the maximal cliques of our derived Internet AS-level topology graph in a few minutes (2 GHz processor, 1 GB RAM) using the efficient algorithm proposed by Bron and Kerbosch [15], however the extraction of *k*-clique communities took an unreasonable amount of time (even when we used a higher performing machine, i.e. 3 GHz processor, 8 GB RAM).

In terms of connectivity point of view, the *k*-dense algorithm finds communities that are more densely-connected than the corresponding *(k-1)*-core communities. Moreover, since it requires the presence of *k-2* common neighbours between each pair of connected nodes, it helps to isolate nodes which are part of separated densely-connected zones.

3. Community metrics

The *k*-dense method applied to the Internet AS-level topology graph provides nested sub-graphs (i.e. *k*-dense communities) that can be studied using classical graph theory indices. Hereafter *k*-dense communities will be referred to as communities for the sake of clarity. In addition, we define: (i) the *internal degree* as the number of connections an AS has within the community; (ii) the *external degree* as the number of connections an AS has outside the community; (iii) the *Internet degree* as the number of connections an AS has on the Internet AS-level topology. In this Section we define two metrics which will be useful for the analysis of the communities extracted: the link density, that gives a measure of how densely connected is a sub-graph, and the Out Degree Fraction (ODF) that provides a measure of the number of connections directed outside the community.

The *link density* of a subgraph is defined as the fraction of existing connections to possible connections [16]:

$$\rho = \frac{2 \cdot e}{n \cdot (n-1)} \tag{8}$$

where e is the number of internal connections and n is the number of ASs within the community. If the community is made up of a single connected community the link density has values in the range $[\frac{2}{n} : 1]$ (the lower bound is the link density of a tree topology, the upper bound is the link density of a clique topology). If the community is made up or 2 (or more) connected components, the link density can also have values in the range $[0 : \frac{2}{n}]$, in these cases there is a number of internal connections that is lower than the minimum number of connections with a single connected component (i.e. the number of internal connections is lower than $n-1$).

The ODF, Out Degree Fraction [17], is defined as the ratio between the external degree and the Internet degree:

$$\text{ODF} = \frac{\text{external degree}}{\text{Internet degree}} \tag{9}$$

ODF takes values in the range $[0 : 1]$: ODF is 0 if there are no connections on the boundary, ODF is 1 if there are no internal connections. ODF indicates, for each AS, the percentage of connections on the boundary which contribute to its Internet degree.

4. Data sources

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

4.1. 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, on the Internet there is no tool specifically designed to derive topology information, hence researchers have had to derive it using various indirect measurements. Topology data

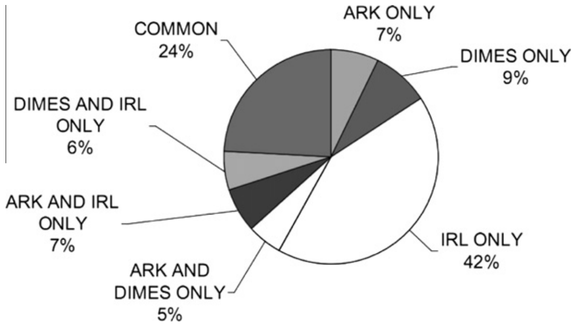

Fig. 3. Source of the connections that makes up the Internet AS-level topology.

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 [18–20]. Although we cannot have a complete map of the Internet AS-level graph, we can still try to reduce its incompleteness by merging topologies obtained from different projects (e.g. CAIDA and IRL, see below for more details) which have been gathered within the same time slot. We thus decided to adopt the methodology described in [12]. Specifically, we built the Internet AS-level topology graph using the following procedure:

1. We downloaded three public available datasets considering the measurement campaigns that they performed in April 2010:
   - the *IPv4 Routed/24 AS Links* dataset [21] (hereafter CAIDA) and the *Distributed Internet MEasurements and Simulations* dataset [22] (hereafter DIMES), which are based on traceroute-like methods;
   - the *Internet Topology Collection at the Internet Research Lab* dataset [23] (hereafter IRL), which gathers topology information based on static snapshots of the BGP routing tables and dynamic BGP data.
2. We then 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[3];
   - AS 23456 which, according to RFC 4893 is reserved and assigned for AS_TRANS[4];
   - AS 3130 which, according to the Cyclops website[5], shows false AS adjacencies due to an experiment by Randy Bush.[6]

At the end of this procedure we obtained a dataset made up of **35,390** ASs and **152,233** connections. The pie chart in Fig. 3 shows the data source of each connection that makes up the Internet AS-level topology graph and demonstrates how each single data source contributes to obtaining a more detailed view of the graph (see [12]).

Despite the efforts to have a detailed map, Internet AS-level topology graph is still incomplete [3,18,24]. Available measurements do provide valuable information [3], however biases arisen from the incompleteness of data should be taken into account when studying the structural properties. [18] states that the coverage of customer-provider connections is much more complete than the coverage of peering connections. No-valley prefer-customer routing policies [25], a pretty good approximation of inter-AS routing rules, lead to the following statements: customer-provider connections are not invisible, thus measurements should be able to unveil all of them; on the other hand, it is not possible to discover a peering connection between two generic ASs, AS *A* and AS *B*, unless there is a monitor installed in either A or B, or in a downstream customer of A or B [18]. In Section 5 we will discuss the influence of these biases on the analysis of the *k*-dense communities.

4.2. IXP dataset

An IXP is a physical hub-and-spoke infrastructure, which enables ASs (participants[7]) to exchange traffic with each other as if they were connected directly via a physical link.

IXPs are a cost effective solution for those ASs which need to be well-connected with each other within a specific geographical location. By exploiting IXP infrastructure these ASs can avoid multiple ad hoc point-to-point connection costs among participants, which are otherwise needed when BGP operates among (all or a subset of) them. IXPs also help Internet traffic to remain localized in the geographical region it belongs to. In fact, much Internet traffic is directed within national borders, since it is made up 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 Africa or submarine fibre connections in Australia), which in addition to improving network performances, saves costs. Hereafter we will use the term *IXP size* to indicate the number of participants in an IXP.

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 where the IXP is located) and the AS numbers of its participants. The dataset was built by applying a procedure similar to that presented in [12]:

1. We collected a potential list of all IXPs by exploiting the information gathered in the Packet Clearing House [26], peeringDB [27], Euro-IX [28] and bgp4.as [29] websites.

[3] A full list of private ASN can be found on IANA website at http://www.iana.org/assignments/as-numbers/as-numbers.xml.

[4] It is used to permit peering between a BGP speaker using 2-octet ASN and a BGP speaker using a 4-octet ASN.

[5] http://cyclops.cs.ucla.edu/blog/?m=200904.

[6] http://psg.com/173-174.

[7] ASs that are connected to at least one IXP will be termed as participants. We avoided using other terms, such as members or customers, because these names depend on the IXP policies these ASs belong to.

Table 1
Geographical list of IXPs found.

Continents	# IXPs	Average IXP size
Africa	12	9.83
Asia	32	28.34
Europe	108	45.63
Latin America	20	19.90
North America	44	40.75
Oceania	16	32.81
World	232	37.37

2. From the previous list we selected only those IXPs which were active. Specifically, we verified each IXP's activity by observing its traffic statistics or by observing the results obtained by 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 by browsing the IXP website or by 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 obtained a collection of **232** active IXPs from all over the world. Table 1 summarizes the results of our collection campaign (the third column indicates the average size of IXPs belonging to a given continent).

In Section 5 we use the IXP dataset intensively in order to identify two classes of ASs: those who participate in IXPs and those that do not. We propose two tags:

- *on-IXP AS* is associated with ASs that belong to an IXP participant list;
- *not-on-IXP AS* is associated with ASs that do not belong to any IXP participant list.

Knowledge of an IXP participants list does not provide any information on the peering matrix, which represents the BGP connections (set up by AS administrators) among their participants. Since ASs treat peering relationships between other participants as proprietary information, the current peering matrices of IXPs are in most cases unknown. For these reasons, we were unable to discover which connections cross the IXPs. To the best of our knowledge, there are two main papers that present strategies to discover these connections: [8,9]. Since their measurement campaigns were performed in May 2005 and April 2009 respectively, we could not use their information to tag connections in our dataset. BGP connections are in fact highly volatile, and hence we could not use datasets that referred to old measurement campaigns.

4.3. Geographical dataset

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

1. We downloaded the GeoLite Country and the GeoLite ASN free databases from MaxMind website[8]. Both of them were uploaded on 1 May 2010. The GeoLite Country database associates IPv4 addresses with 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 ⟨`AS number`, `Country code`⟩ tuples. Note that, for each AS number multiple country codes could exist, hence the geographical database key is the entire tuple.

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

In Section 5 we provide a geographic attribute to each AS, according to the following taxonomy:

- An AS is called a *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 a *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 are placed outside Europe.
- An AS is called a *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 the Netherlands and one geographical location in the United States is referred to as worldwide AS.

5. Structural properties of the Internet AS-level topology graph

In this Section we use the k-dense approach to investigate the main structural properties of the Internet AS-level topology graph. Our topology (see Section 4.1) is made up of 35,390 ASs and 152,233 connections. If we apply tags derived from the IXP dataset (see Section 4.2) we find that the 13% of Internet ASs participate in at least on IXP while the remaining 87% can be tagged as not-on-IXP. This means that IXPs facilities are used by a small percentage of Internet ASs.

To further analyze the properties of the Internet AS-level topology graph, Table 2 shows the geographical distri-

[8] In this work we use GeoLite data created by MaxMind, available from http://www.maxmind.com/.

Table 2
Internet features related to the geographical dataset.

	ASs	ASs (%)
Worldwide	1568	4.43
Continental	1115	3.15
National	31,228	88.24
Unknown	1479	4.18

bution of ASs according to the taxonomy presented in Section 4.3. The results presented in Table 2 indicate that the vast majority of Internet ASs are national. Only the 7.58% are worldwide or continental. This distribution can be explained by the existence of many stub ASs, i.e. ASs that are interested in connecting to other ASs to obtain connectivity to the Internet. These ASs 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 (”unknown”) identifies the percentage of ASs whose geographical location was not inferred by MaxMind. This depends on MaxMind information retrieval methods. In future research, we plan to use other datasets (e.g. MaxMind GeoIP Country, IPligence, hostip.info) in order to have complete ASs coverage.

5.1. k-Dense results

The application of the *k*-dense algorithm to the Internet AS-level topology graph produces a *k-max* equal to **47**. The vast majority of *k*-denses are made up of a single connected component, thus terms *k*-dense and *k*-dense community refer to the same sub-graph. Nevertheless, when *k* is equal to 3, 4 and 6, *k*-denses are made up of more than a single connected component (3-dense has 2 connected components, 4-dense has 3 connected components, 6-dense has 2 connected components). When we analyze these *k*-denses we observe a common feature: there is a single large connected component which represents more than the 99% of all the *k*-dense ASs, while the remaining community (or the remaining communities) is made up of a negligible number of ASs. To make the notation easier, hereafter we will use the terms 3-dense community, 4-dense community and 6-dense community to indicate the largest community extracted from these *k*-denses.

Briefly, we will analyse the behaviour of 46 communities, i.e. each *k*-dense community with a *k* in the range [2 : 47]. By definition each *k*-dense is a sub-graph of a *k-1*-dense, hence these 46 communities are nested. The number of ASs and connections which belong to a *k*-dense community thus decreases as *k* increases.

In Fig. 4 we show the percentage of Internet ASs and connections which are part of each *k*-dense community. As expected, these two indices decrease as *k* increases. Nevertheless the vast majority of Internet ASs belong to low *k*-dense communities while 10% have a *k*-dense-index larger than 7. Also, the percentage of Internet connections within each *k*-dense community decreases more slowly than the percentage of Internet ASs, since the higher the *k*-dense the better-connected the relative sub-graph.

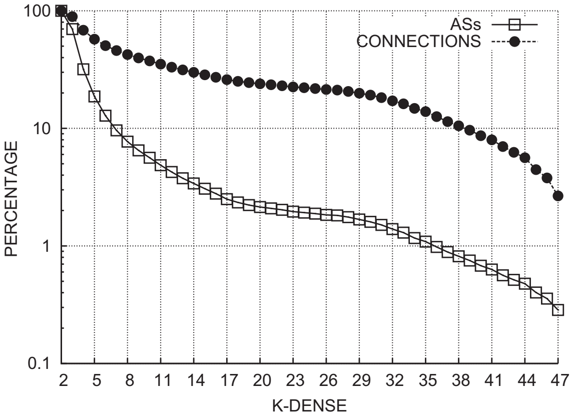

Fig. 4. Percentage of ASs and connections in each *k*-dense.

However from Fig. 4 we cannot infer how a *k*-dense community contributes to the overall Internet AS-level topology. To better understand this, in Fig. 5 we plot the volume of connections originated by each *k*-dense-shell (see Section 2). Note that in this Figure we use *k*-dense-shell instead of *k*-dense since this approach enables us to clearly find which ASs are involved in a considerable number of Internet connections.

Fig. 5 highlights the presence of two groups of ASs involved in a very high number of connections: the first group is made up of ASs with a low *k*-dense-index, the second group of the *k-max*-dense ASs[9]. Note that the *3*-dense-shell ASs are involved in 33,495 connections. The *3*-dense-shell is made up of a huge number of ASs, i.e. 13,464, hence there is a considerable number of connections which originate from this *k*-dense-shell. The presence of so many ASs belonging to the *3*-dense-shell could be the result of multi-homing. A multi-homed AS is defined as a network which is connected to multiple providers at the same time. With this approach if one of the two (or more) links fails, the AS is still able to reach Internet. Very often the two providers of a multi-homed ASs are connected, thus, these three ASs form a 3-clique and are thus part of a *3*-dense community. ASs belonging to the *47*-dense-shell, are involved in 63,973 connections, i.e. 42% of Internet connections, although they only represent 0.28% of the Internet ASs. These percentages indicate that these ASs have a central position within the graph and that they play a primary role in Internet connectivity. As shown in Fig. 4, *47*-dense internal connections represent 2.66% of the overall Internet connections, hence *47*-dense ASs direct the vast majority of their connections to ASs belonging to other *k*-dense-shells. The tendency of *k*-dense ASs to direct their degree inside or outside the community can be measured using the ODF metric (see Section 3).

In Fig. 6 we plot the average ODF for each *k*-dense community and its link density. The ODF strictly depends on the number of external connections, while the link density relies on the number of internal connections and enables us to evaluate how well clustered the ASs within a community are (see Section 3 for more details).

[9] Please note that, if *k* = *k-max* then *k-max*-dense ASs and *k-max*-dense-shell ASs identify the same set of ASs.

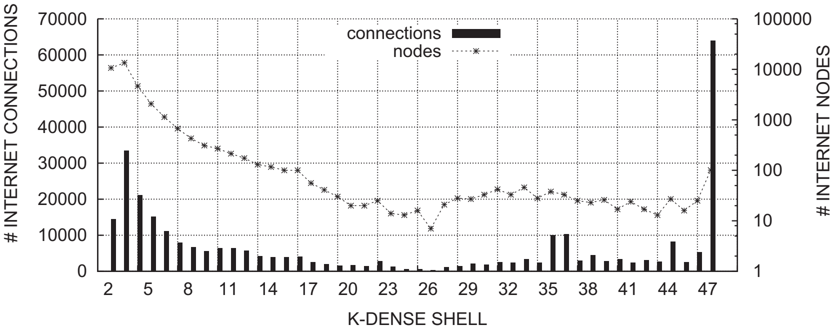

Fig. 5. Number of connections (and number of nodes) vs. *k*-dense-shell.

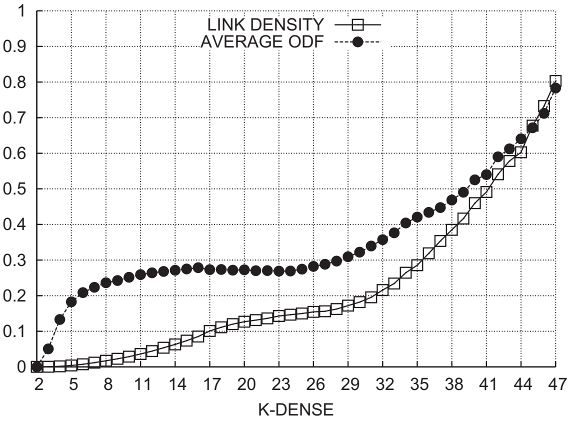

Fig. 6. *k*-Dense link density and ODF.

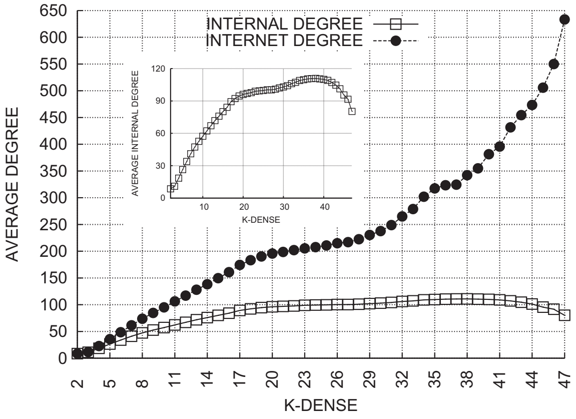

Fig. 7. *k*-Dense internal degree and total degree.

The link density in Fig. 6 shows that the higher *k* is, the more clustered the community is. All the *k*-dense communities with a *k* higher than 41 have a link density larger than 0.5. This kind of AS also presents high average ODF values which means that, although they are really well connected to the other *k*-dense community ASs, they mostly direct their connections outside the community. On the other hand, low *k*-dense communities are characterized by a small link density and a small ODF. This is because only a small percentage of ASs are outside these communities (e.g. all the communities with a *k* lower than 6 are made up of more than the 50% of Internet ASs).

ODF values, shown in Fig. 6, indicate that on average higher *k*-dense community ASs tend to have an external degree larger than their internal degree. To better understand this behaviour, in Fig. 7 we plot the average internal degree and the average Internet degree for each *k*-dense community. Given that Internet and external degrees are correlated (Internet degree is equal to the sum of internal degree and external degree), we can derive the average external degree of each *k*-dense community if we know the average internal degree and the average Internet degree.

Fig. 7 shows that the higher *k* is, the higher the average Internet degree is. On the other hand, the average internal degree does not have a monotonic trend (see the enlarged sub-figure for a clearer view of the trend). This behaviour indicates that a high average internal degree is not a good criterion to select dense zones, indeed link density can increase even if the internal degree decreases.

Results shown above describe an Internet structure which is compliant with the Jellyfish model proposed in [30]. Both descriptions highlight the presence of a dense sub-graph of nodes strongly connected to the rest of the graph by means of a huge number of connections. Siganos et al. [30] provide several interesting structural observation and present a model of the network which has been proved to accurately describe many snapshots of the Internet AS-level topology. *k*-dense analysis does not provide a visualization of the network as appealing as the Jellyfish model (*k*-denses represent the network as a series of concentric circular shells), nevertheless it provides a formal description of the maximum level of density which characterize each node. The main difference between this work and [30] is the interpretation of the *structure*: in the following paragraphs we give a more detailed description of the nodes which populate *k*-dense layers. Specifically we apply tags obtained from the IXP and the geographical datasets (see Sections 4.2 and 4.3 respectively) to unveil the characteristics of ASs which are part of the densest Internet subgraphs. We start this analysis with plotting the percentage of each ASs category in each *k*-dense.

IXP dataset tags. In Fig. 8 we plot the percentage of ASs tagged as on-IXP for each *k*-dense and the percentage of ASs tagged as not-on-IXP. Fig. 8 highlights that low *k*-dense communities are mainly made up of not-on-IXP ASs (see [2 : 4]-dense communities). On-IXP ASs have a strong presence in high *k*-dense communities, moreover they completely populate [36 : 47]-dense communities. This indicates that the presence of well-connected zones within our Internet AS-level topology is mainly (or totally) triggered by on-IXP ASs. In our analysis we also found that on-IXP ASs have a higher probability of participating in high *k*-dense communities with respect to not-on-IXP ASs. In fact, IXP facilities ease the creation of connections between their participants (see Section 4.2), hence ASs on the same IXP are likely to set up more connections with each other.

Geographical dataset tags. Fig. 9 shows for each *k*-dense the percentage of ASs tagged as worldwide, the percentage of ASs tagged as continental and the percentage of tagged as national. The percentage of national ASs within a *k*-dense community considerably decreases as *k* increases. In contrast, continental and worldwide ASs have the opposite behaviour. National ASs represent the majority of ASs within the [2 : 12]-dense communities, worldwide ASs represent the majority of ASs within the [33 : 47]-dense communities. There is no prevailing AS geographical scope in the [13 : 32] range.

The results presented in Figs. 4–6 indicate that: (a) only a small subset of Internet ASs belong to very well-connected communities; (b) higher *k*-dense communities are characterized by a high level of clusterization, nevertheless they tend to direct most of their connections to ASs outside the community; (c) 47-dense ASs have a primary role in Internet connectivity since they are involved in a huge number of Internet connections (42%). For this reason we perform a more detailed analysis of the *k-max*-dense community in Section 5.3.

The results presented in Figs. 7–9 indicate that: (a) the most well-connected communities are mainly made up of ASs connected with at least one IXP and with a worldwide geographical scope. These communities are characterized by ASs with a high average Internet degree. (b) the least well-connected communities are characterized by a strong presence of not-on-IXP ASs with a national geographical scope. In addition, by observing the average Internet degree of these ASs, we can deduce the presence of a considerable number of low degree ASs.

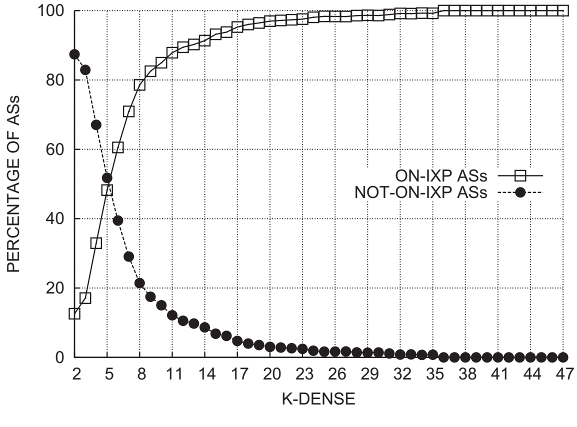

Fig. 8. Percentage of on-IXP/not-on-IXP ASs within each *k*-dense.

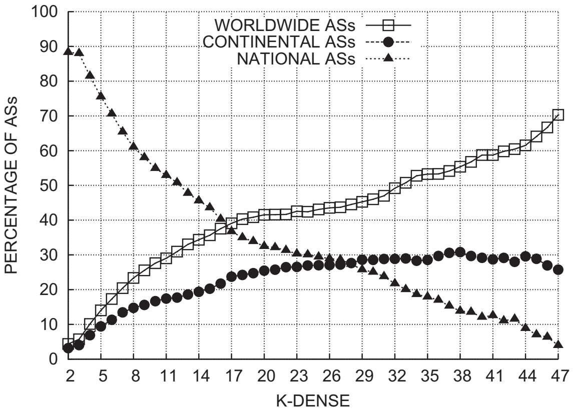

Fig. 9. Percentage of worldwide/continental/national ASs within each *k*-dense.

5.2. Incompleteness

According to [18] results the vast majority of missing links are peering connections which involve ASs that do not host any monitor (or ASs whose downstream customers do not host any monitor). Thus we are likely to miss peering connections involving small ASs that do not host monitor or which do not have customers. On the other hand, we believe peering connections involving large providers are fully captured by our datasets since it is likely to find a monitor in one of them or at least within their downstream customers. In our analysis we found that the most well-connected zone of the Internet involves ASs participating at IXPs with a pretty high Internet degree, while ASs belonging to the lower *k*-denses are likely to have a lower degree and, usually, do not participate at IXPs. Based on these statements, we believe the addition of currently hidden peering links to our Internet topology would provide the following changes: (a) the *k-max* would be probably increased as the peering connections are less hierarchical than the customer-provider connections and hence they are likely to form dense zones; (b) there could be many ASs with a low *k*-dense shifted to *k*-dense communities with a higher *k*, moreover this behaviour could also provide the formation of communities separated from the *giant* component (i.e. the single and large *k*-dense community) that could be interpreted as local communities.

5.3. *k*-Max-dense analysis

The *k-max*-dense (or 47-dense) community is a single connected component with 101 ASs and 4056 connections. Hereafter we will refer to the *k-max*-dense community as *k-max*-dense or simply *47*-dense.

We decided to deepen our analysis of the *47*-dense community since it is involved in the largest number of Internet connections (63,973) and thus it has a main role in the overall Internet connectivity. In addition, providing a deep analysis also for *k*-dense communities with a pretty lower *k* would not change significantly the results provided by the *47-dense* community analysis.

As expected, *47*-dense is made up of ASs with a high Internet degree and a rather high internal degree (see Fig. 7). The minimum Internet degree is 123, and more than 50% of *47*-dense ASs have a larger Internet degree than 407. The internal degree has values in the interval [60:100], which means that each *k-max*-dense AS is connected to at least another 60 ASs of the *k-max*-dense. This confirms the presence of a high level of connectivity within the community. There are two ASs (WV Fibre and RETN) which are connected to all the *k-max*-dense ASs. The diameter is thus equal to 2 (i.e. each AS in the community can reach any other *47*-dense AS with at most two hops).

k-Max-dense ASs usually have an Internet degree which is much larger than the internal degree. This characteristic, which can be seen by observing the average ODF in Fig. 6, holds for the vast majority of the *k-max*-dense ASs. Although the ODF of *k-max*-dense ASs has values in the range [0.392:0.783], 97% of ASs have an ODF greater than 0.5 (i.e. the number of connections on the boundary is larger than the number of internal connections). Since the degree does not provide much insight into the level of clusterization within the community, we evaluate the compactness of the *47*-dense using the link density and the clustering coefficient. The link density of the *47*-dense is equal to 0.803 and indicates that the community has a number of connections that is equal to the 80% of connections we could find in a *101*-clique topology. Hence, it points to a very high level of internal connectivity. In addition, we found that *47*-dense ASs have a clustering coefficient in the range [0.794:0.833]. This tight interval indicates that each AS has a very well connected set of neighbours within the community.

To better characterize ASs which populate the *k-max*-dense community, we apply the tags derived from the IXP and the geographical datasets.

IXP dataset tags. All the ASs within the *k-max*-dense are connected to at least one IXP. Thus by exploiting our IXP dataset we found that each *k-max*-dense AS is connected on average to 10.3 IXPs whose average number of participants is 178.43. This average IXP size is really high compared to the Internet average IXP size shown in Table 1 (see the last row), also considering that only 18 out of 232 IXPs have an IXP size larger than 100. This means that *47*-dense ASs are likely to connect to large IXPs. Using the information in our dataset we discovered that all the 101 *k-max*-dense ASs are connected to at least one of the largest IXPs (AMS-IX, DE-CIX, LINX), while 63% of them are connected to all these three large European IXPs.

In order to characterize ASs belonging to *47*-dense, Fig. 10 shows the number of IXPs that each *k*-dense AS is connected to (ASs were sorted by decreasing number of IXPs). There are four ASs which participate in a very high number of IXPs (more than 35), on the other hand there are many *k-max*-dense ASs which are not interested in participating in many IXPs. Since participation at IXPs ensues from the AS business strategy, we can extend the analysis of these ASs by investigating their business profile. It is interesting to observe that ASs participating in more than 20 IXPs are typically Content Delivery Networks (CDN) or Internet Backbone Providers (IBP). The main goal of CDNs is to deliver content for their clients by reducing latency and packet loss. In order to avoid bottlenecks near central servers (which host data) they usually place their edge servers (a sort of mirror of the central server) close to their client networks. Then participation in many IXPs allows CDNs to be closer to many of their customers using a single connection (connection to IXP eases the set up of connections to other IXP participants). "*In addition, a Content Provider has to pay transit fees to reach some destinations within the region, therefore tends to seek peering with others with whom there is a large amount of traffic to exchange. This leads to a generally open-peering inclination as articulated by an open peering policy.*"[10]

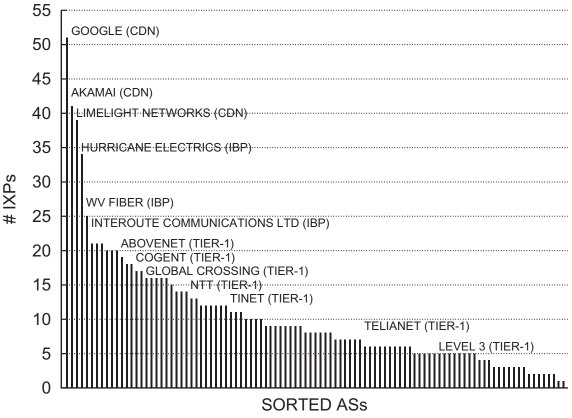

Fig. 10. Number of IXPs for each AS of the *k-max*-dense.

A network operator can be of two categories: Internet Service Provider (ISP) or Internet Backbone Provider (IBP). ISPs offer retail network access for individuals and institutions, while IBPs provide high-speed, long haul communication links for ISPs [31]. These categories can overlap, in this paper we will use the term Internet Backbone Providers to refer to organizations that supply the ISPs with access to the lines that connect ISPs to each other, thus allowing ISPs to offer their customers Internet access at high speeds. These backbone providers usually provide connection facilities in many cities for their clients, and they themselves connect with other backbone providers at IXPs. Google (CDN), Akamai (CDN), Limelight Networks (CDN), Hurricane Electrics (IBP), WV Fibre (IBP) and Interoute Communications Ltd (IBP) are *k-max*-dense ASs have a higher number of participations in IXPs (see Fig. 10). Google's (CDN) and Limelight Networks' (CDN) participation in a high number of IXPs was also highlighted in [9].

For some kinds of ASs, participation in a large or a small number of IXPs is not a common strategy. Within the

[10] William B. Norton, http://drpeering.net/white-papers/Ecosystems/Content-Providers.html.

47-dense community there are Tier-1 ASs which participate in many IXPs (e.g. AboveNet, Cogent and Global Crossing participate in more than 15 IXPs) and Tier-1 ASs which participate in only 5 IXPs (e.g. Level 3).

ASes with large customer cones have an especially important role in the Internets capital and governance structure. At the top of this hierarchy are ISPs commonly known as Tier-1 ISPs[11]. A Tier-1 AS is a network that can reach each other AS in the Internet via settlement-free peering and without purchasing IP transit (by definition ASs which belong to this category form a clique). There is no authority that defines Tiers of networks participating in the Internet, moreover it is not possible to find out business relationships between ASs since this is confidential information. Although we cannot guarantee that cited ASs (e.g. Cogent, Global Crossing, Level 3) are fully transit-free ASs, they nevertheless present many Tier-1 properties, e.g. they form a clique, they have a worldwide presence, they have a huge number of connections.

Since all the *k-max*-dense ASs are tagged as on-IXP ASs, it follows that all the internal connections are directed to on-IXP ASs. Exploiting our IXP dataset, we found that 99.9% of cases were connections whose endpoint ASs participate in a common IXP[12]. This behaviour indicates that ASs participating in a common IXP are really involved in the creation of this well-connected zone of the Internet AS-level topology identified by *47*-dense.

The strong presence of IXPs also holds for Internet connections involving *k-max*-dense ASs (i.e. internal connections + connections on the boundary). Consider all the connections that start from a generic *47*-dense AS: on average, 82% are directed to on-IXP ASs, moreover, 74% involve ASs participating in a common IXP. Fig. 11 shows for each *k-max*-dense AS the percentage of connections directed to on-IXP ASs (continuous line) and the percentage of connections directed to ASs participating in a common IXP (dashed line). The *k-max*-dense ASs are sorted by increasing percentage of connections directed to on-IXP ASs.

As seen in Fig. 11, 92 out of 101 *k-max*-dense ASs, direct more connections to on-IXP ASs than to not-on-IXP ASs. In this set are the CDNs and IBPs highlighted in Fig. 10. On the other hand, the nine ASs which direct the majority of their connections to not-on-IXP are mostly Tier-1 ASs: in this set are Level 3, Cogent, Global Crossing, TeliaNet, NTT, Abovenet and TiNet.

Geographical dataset tags. We end our analysis of *k-max*-dense by showing the distribution of geographical scopes (Table 3). For some ASs worldwide coverage and availability is required for their business, as with Tier-1 ASs, CDNs and IBPs. These ASs are part of the 71 ASs with worldwide geographical scope. Table 3 highlights that *47*-dense is mainly made up of worldwide ASs. Using the geographical dataset, we found that the set of 30 ASswith a national or a continental geographical scope (4 + 26, see Table 3) is made exclusively of European ASs.

Each *k-max*-dense AS directs its Internet connections at national, continental and worldwide ASs. These connections are distributed on average as follows: 37.047% are directed to worldwide ASs, 21.396% are directed to continental ASs, 40.851% are directed to national ASs. Those ASs, which were previously indicated as Tier-1, do not follow the average behaviour but are characterized by a very high percentage of connections directed at national ASs (these ASs are likely to be their customers).

6. Related work and comparison

This Section provides a comprehensive review of the state of the art of the community detection algorithms which in our opinion, could well highlight structural features of the Internet AS-level topology graph. In addition, we provide a comparison of different community detection algorithm.

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. Detecting communities in a graph is very hard given that there is no formal definition of a community and second because most of the algorithms are computationally very demanding. Other problems may also arise both from the possible occurrence of hierarchies, i.e. communities which are nested inside larger communities, and from the existence of overlaps between communities, due to the presence of nodes belonging to several groups. 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 [32,2,33–36]). For an exhaustive description of the state of the art related to community detection algorithms see [32].

The quality of a community decomposition is often measured by the modularity [37], denoted by Q. This metric is defined to be the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. According to this definition, a good partition of the network is that in which there are dense internal connections between the nodes within the community but only sparse connections between different communities. More in detail, the modularity of a partition is defined as follows:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where A_{ij} is an element of the adjacency matrix of the graph, k_i is the degree of node i, m is the total number of connections of the graph, c_i is the community to which node i is assigned and δ is the Kronecker delta. At the AS-level of abstraction we are interested in finding communities made up of ASs which form very dense sub-graphs, but we do not require they have few connections directed outside the community. Consider, for instance, a group of

[11] CAIDA, http://www.caida.org/research/topology/rank_as/.

[12] Please note that, if the two endpoints participate at more than one IXP we count a connection if there is IXP which lists both of them in its participant list. In order to be more clear, consider the following example. AS_A has two connections, one is directed to AS_B, one is directed to AS_C. Moreover, AS_A participate at IXP_1 and IXP_2, AS_B participate at IXP_2 and IXP_3, AS_C participate at IXP_4. Hence, AS_A has one connection directed to an on-IXP AS (AS_B) which participate in a common IXP (IXP_2).

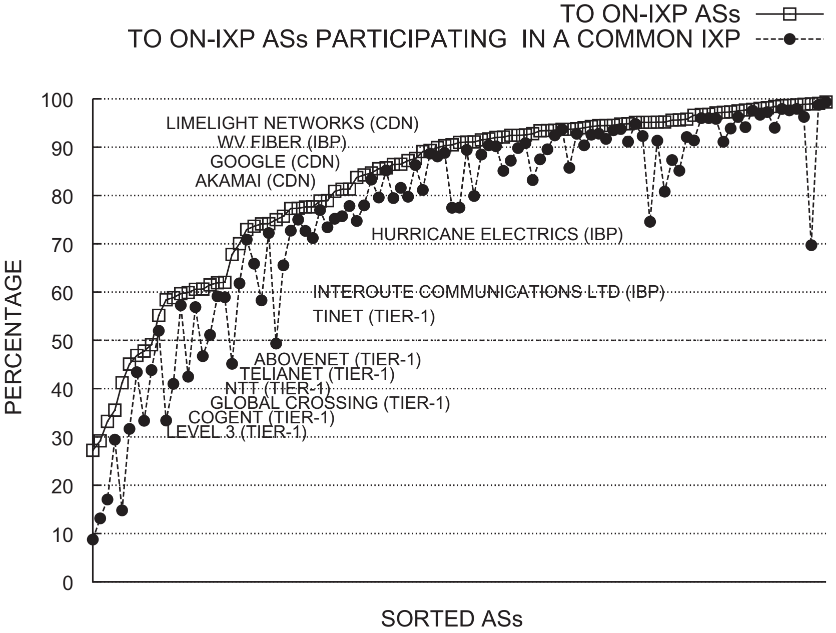

Fig. 11. Percentage of connections directed to on-IXP ASs for each AS of the *k-max*-dense.

Table 3
k-Max-dense features related to the geographical dataset.

	ASs	ASs (%)
Worldwide	71	70.30
Continental	26	25.74
National	4	3.96
Unknown	0	0

regional transit providers who are really interested in connecting to each other in order for the traffic to remain localized and to prevent traffic from unnecessarily traversing other transit networks. This set of ASs is likely to form a community although, it is highly probable that the vast majority of their connections will be directed to customer ASs, i.e. outside the community. If the number of connections directed outside the community is very high the product k_ik_j yields in a negative modularity, thus a community detection method based on modularity would not provide this kind of communities. Communities extracted from the Internet AS-level topology graph should be characterized by a pretty high link density (see Section 3), it does not matter the value of their average Out Degree Fraction (see Section 3). In other words, we are interested in finding dense sub-graphs regardless of the connections directed outside the community.

In most of the approaches published in the specialized literature, communities have been characterized and discovered by exploiting some global property of the graph, such as betweenness, modularity, etc. However, communities can also be 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. One method that accounts both for the locality of the community definition and the possibility of having overlapping communities is the Clique Percolation Method (CPM) by Palla et al. [14]. A number of concepts were introduced in [14] as a support for specifying a *k*-clique community. Specifically, two *k*-cliques are adjacent if they share *k-1* vertices. The union of adjacent *k*-cliques is called a *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 algorithm by Palla et al. [14] requires a huge amount of computation for the Internet AS-level topology graph. Some public tools are available, such as CFinder[13] which implement the *k*-clique algorithm, however they are not able to extract *k*-clique communities within our Internet topology dataset in a reasonable amount of time.

There are two other community detection algorithms whose aim is to detect well-connected zones of the graph: the *k*-core decomposition [38] and the *k*-dense algorithm [1]. The study of the Internet AS-level topology graph structure through the *k*-core decomposition [38] has been conducted in [13,39,40]. On the other hand, the algorithm in [1] 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 analyse the results of this community detection algorithm since it can be thought of 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 algorithm.

[13] http://cfinder.org/.

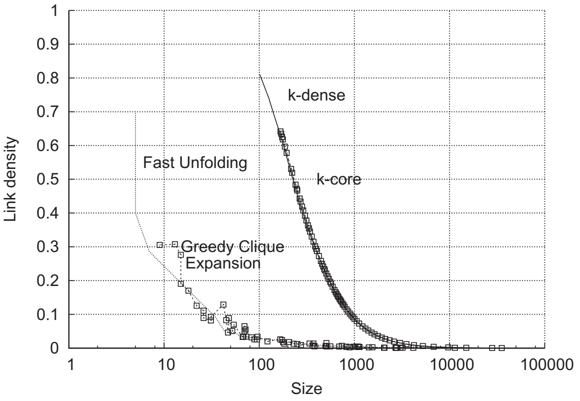

Fig. 12. Link density vs. size.

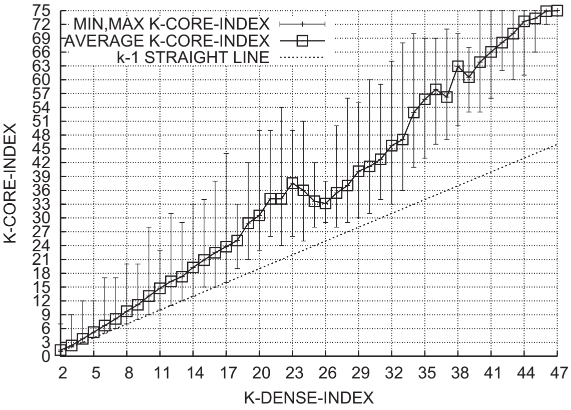

Fig. 13. *k*-Core-index in each *k*-dense-shell.

The rationale behind the choice of the *k*-dense algorithm also arises from the comparison of the link density of communities provided by different methods when applied to the Internet AS-level topology graph (see section 4). More in detail, we show in Fig. 12 the link density vs. the size of each community extracted by: (a) the *k*-dense algorithm, (b) the *k*-core decomposition, (c) the Greedy Clique Expansion algorithm (GCE) [41] and (d) the Fast Unfolding method [42]. GCE algorithm allows to find overlapping communities starting from maximal cliques and by greedily optimizing a local fitness function (defined in [16]). On the other hand, Fast Unfolding method is a heuristic method that is based on modularity optimization. Both of these algorithms allow to extract communities from the Internet AS-level topology very quickly in terms of computation time. Fig. 12 shows that communities provided by the *k*-dense algorithm and the *k*-core decomposition have larger values of link density if compared those provided by the Greedy Clique Expansion and the Fast Unfolding. More in detail, GCE takes link density values greater than 0.2 only for very small communities (their size is lower than 20). The same behaviour applies to the Fast Unfolding communities. *k*-core and *k*-dense allow to obtain high link density values also for large communities (e.g. size greater than 100). For this reason we will continue our comparison focusing on *k*-core and *k*-dense only.

The application of the *k*-core decomposition algorithm to the Internet AS-level topology graph generates 76 nested communities (*k* has values in the range $[1:75]$), one for each *k*-core. This decomposition, although its communities are less well-connected, yields quite similar results with respect to the *k*-dense decomposition. *k*-core communities have a size distribution that is very similar to the one in Fig. 4, i.e. only a small percentage of Internet ASs participate in the most well-connected *k*-core communities. Moreover, the average Internet degree in each *k*-core community increases as *k* increases. Another common feature is the presence of two groups of ASs (those with a low *k*-core-index[14] and those with a *k-max* *k*-core-index) which are involved in a very high percentage of Internet connections (as in Fig. 5).

The main difference between *k*-dense and *k*-core community extraction algorithms is in the link density. All the *k*-dense communities have a larger link density if compared to their corresponding *k-1*-cores. This behaviour cannot be foreseen a priori: each *k*-dense is a subset extracted from a *k-1*-core (see Expression (7)) thus it has a smaller (or equal) number of ASs and connections. However we cannot say anything about the ASs' connections ratio. In fact the link density is computed as a ratio between the number of connections and the maximum number of connections that we can evaluate starting from the number of ASs. Fig. 13 shows the average *k*-core-index in each *k*-dense-shell. Obviously, the *k*-core-index grows as the *k*-dense-index increases (this follows from the definition of *k*-dense). The distance between the $k-1$ line and the average *k*-core-index line, highlights that ASs in a given *k*-dense-shell typically have a *k*-core-index greater than $k-1$. Moreover, since all of *47*-dense community ASs are part of the *75*-core, at least with our topology dataset, the *k-max*-dense community is a tightly connected zone of the *k-max*-core community.

7. Discussion and conclusions

In this work we have analyzed the structural characteristics of the Internet AS-level topology graph using the *k*-dense community detection algorithm and by also exploiting geographical information and statistics related to IXPs.

We started our dissertation by comparing different community detection methods. We evaluated their goodness in identifying communities within the Internet topology graph at the AS-level of abstraction by supporting our motivations with real examples.

The Internet AS-level topology derived from our datasets is made up of 47 nested *k*-dense communities. Those communities are more compact than the corresponding *k-1*-core communities (they have a greater link density). Also, *k*-dense communities ASs typically have a *k*-core-index higher than *k-1*, i.e. the *k*-dense community extraction

[14] A node i is said to have a *k*-core-index *k* if it belongs to the *k*-core but is not part of the *(k + 1)*-core. Moreover, we define *k*-core–shell the set of nodes having a *k*-core-index equal to k.The maximum *k*-core-index will be referred to as k-max.

algorithm selects ASs with a high *k*-core-index within the *k-1*-core.

We found that low *k*-dense communities (i.e. those whose *k* is close to 2) have a very large number of ASs and connections. In addition, ASs belonging to these communities are usually low degree ASs with a national geographical scope and do not participate in IXPs. On the other hand, the high *k*-dense communities (i.e. whose *k* is close to 47) have a small number of ASs and a high level of clusterization (i.e. high link density). These ASs typically connect to IXP infrastructures and have a worldwide or continental geographical scope.

In addition, since current measurements provide an incomplete map of the Internet AS-level topology we discussed the possible changes that would apply to our analysis supposing to have a complete view of the Internet graph.

The 101 ASs that make up the *k-max*-dense (or *47*-dense) are involved in a huge number of Internet connections (42% of overall Internet connections). This highlights the primary role of these ASs in the Internet AS-level topology. The *47*-dense community is made up of a cohesive set of ASs since it has a link density equal to 0.8. By exploiting the IXP dataset we found that *47*-dense ASs participate in at least one IXP and, surprisingly, we discovered that the vast majority (82%) of their Internet connections (i.e. internal connections + connections on the boundary) are directed at on-IXP ASs. Using the geographical dataset we found that the vast majority of *k-max*-dense ASs have a worldwide geographical scope. They all have at least one location in Europe, and each *47*-dense AS participates in at least one of these three large European IXPs: AMS-IX, DE-CIX and LINX.

We have shown that CDNs and IBPs seem to be interested in participating in many IXPs and they are likely to direct a large percentage of their connections at on-IXP ASs. On the other hand, Tier-1 ASs typically connect to a smaller number of IXPs with respect to CDNs and IBPs. They tend to direct a large part of their connections at not-on-IXP ASs, especially those with a national geographical scope.

The *47*-dense community has a very high number of connections on the boundary (its average ODF is close to 0.8) and this seems to contradict the legacy community definition, i.e.: a community is typically thought of as a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network [43]. This definition does not seem to well represent Internet AS-level communities. For example Tier-1 ASs can be thought as an Internet community since they are well connected with each other (they form a complete graph by definition), and they are likely to have a similar business strategy. These ASs (about 20) have a very high Internet degree, each has more than one thousand connections, hence they have a huge amount of connections directed at ASs outside their community.

We are planning to derive a new definition of community that would fit the Internet AS-level topology graphs. To this end, we intend to improve our additional datasets in order to have a better view of the Internet AS-level topology. It would be useful to tag Internet connections with business relationships labels (e.g. customer-provider, peer-to-peer) or to highlight those connections that are set up through IXP facilities. We could also improve the geographical dataset by associating set of Points of Presence (PoPs) with each AS.

References

[1] K. Saito, T. Yamada, K. Kazama, Extracting communities from complex networks by the *k*-dense method, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E91-A (11) (2008) 3304–3311.

[2] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104+.

[3] D. Krioukov, K. Claffy, M. Fomenkov, F. Chung, A. Vespignani, W. Will-Inger, The workshop on Internet topology (wit) report, SIGCOMMComput. Commun. Rev. 37 (2007) 69–73. ISSN:0146-4833.

[4] C. Gavoille, Routing in distributed networks: overview and open problems, SIGACT News 32 (2001) 36–52. ISSN:0163-5700.

[5] M. Thorup, U. Zwick, Compact routing schemes, in: SPAA '01: Proceedings of the Thirteenth Annual ACM Symposium on Parallel Algorithms and Architectures, 2001, pp. 1–10.

[6] D. Krioukov, K. Fall, Compact routing on Internet-like graphs, in: Proc. IEEE INFOCOM, 2004, pp. 209–219.

[7] A. Brady, L. Cowen, Compact routing on power-law graphs with additive stretch, in: ALENEX, 2006.

[8] Y. He, G. Siganos, M. Faloutsos, S.V. Krishnamurthy, Lord of the links: a framework for discovering missing links in the Internet topology, IEEE/ACM Trans. Netw. 17 (2) (2009) 391–404.

[9] B. Augustin, B. Krishnamurthy, W. Willinger, IXPs: mapped?, in: IMC '09: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, ACM, New York, NY, USA, 2009, pp 336–349.

[10] Y. Shavitt, N. Zilberman, A structural approach for PoP geo-location, in: NetSciCom 2010: Second International Workshop on Network Science for Communication Networks, 2010.

[11] E. Gregori, L. Lenzini, C. Orsini, k-Dense communities in the Internet AS-level topology, in: COMSNETS 2011: Proceeding of the Third International Conference on COMmunication Systems and NETworkS, 2010.

[12] E. Gregori, A. Improta, L. Lenzini, C. Orsini, The impact of IXPs on the AS-level topology structure of the Internet, Comput. Commun. 34 (1) (2011) 68–82.

[13] J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, K-core decomposition of Internet graphs: hierarchies self-similarity and measurement biases, Netw. Heterogen. Media 3 (2) (2008) 293–371.

[14] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.

[15] C. Bron, J. Kerbosch, Algorithm 457: finding all cliques of an undirected graph, Commun. ACM 16 (9) (1973) 575–577. ISSN:0001-0782.

[16] A. Lancichinetti, M. Kivelä, J. Saramäki, S. Fortunato, Characterizing the Community Structure of Complex Networks, PLoS ONE 5 (8) (2010) e11976.

[17] J. Leskovec, K.J. Lang, M.W. Mahoney, Empirical comparison of algorithms for network community detection, in: WWW2010: ACM WWW International Conference on World Wide Web, 2010.

[18] R. Oliveira, D. Pei, W. Willinger, B. Zhang, L. Zhang, The (in)completeness of the observed Internet AS-level structure, IEEE/ACM Trans. Netw. 18 (1) (2010) 109–122.

[19] A. Lakhina, J.W. Byers, M. Crovella, P. Xie, Sampling biases in IP topology measurements, in: IEEE INFOCOM, 2003, pp. 332–341.

[20] D. Achlioptas, A. Clauset, D. Kempe, C. Moore, On the bias of traceroute sampling: or, power-law degree distributions in regular graphs, in: STOC '05: Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing, ACM, New York, NY, USA, 2005, pp. 694–703.

[21] Y. Hyun, B. Huffaker, D. Andersen, E. Aben, M. Luckie, K.C. Claffy, C. Shannon, The IPv4 Routed/24 AS Links Dataset, 2010. <http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml>.

[22] Distributed Internet MEasurements and Simulations dataset, 2010. <http://www.netdimes.org/>.

[23] Internet Topology Collection at the Internet Research Lab dataset, 2010. <http://irl.cs.ucla.edu/topology/>.

[24] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, A. Vespignani, Exploring networks with traceroute-like probes: theory and simulations, Theor. Comput. Sci. 355 (2006) 6–24. ISSN:0304-3975.

[25] L. Gao, On inferring autonomous system relationships in the Internet, IEEE/ACM Trans. Netw. 9 (2001) 733–745.

[26] Packet Clearing House, 2010. <http://www.pch.net/>.

[27] Peering DB, 2010. <http://www.peeringdb.com/>.

[28] Euro-IX, 2010. <http://www.euro-ix.net/>.

[29] BGP4 Website, 2010. <http://bgp4.as/>.

[30] G. Siganos, S. Tauro, M. Faloutsos, Jellyfish: a conceptual model for the AS Internet topology, J. Commun. Netw. 8 (3) (2006) 339–350.

[31] Y. Tan, I.R. Chiang, V.S. Mookerjee, An economic analysis of interconnection arrangements between Internet backbone providers, Oper. Res. 54 (4) (2006) 776–788.

[32] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.

[33] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization and identification of web communities, Computer 35 (3) (2002) 66–71.

[34] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (12) (2002) 7821–7826.

[35] M.E.J. Newman, Detecting community structure in networks, Eur. Phys. J. B – Condens. Matter. Complex Syst. 38 (2) (2004) 321–330.

[36] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2) (2005) 027104+.

[37] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113+.

[38] S.B. Seidman, Network structure and minimum degree, Soc. Netw. 5 (1983) 269–287.

[39] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of Internet topology using k-shell decomposition, Proc. Natl. Acad. Sci. 104 (27) (2007) 11150–11154.

[40] S. Amr, M. El-Betagy, M. Helmi, Analyzing Internet connectivity data using modified k-shell analysis, in: INFOS 2008, 2008.

[41] C. Lee, F. Reid, A. McDaid, N. Hurley, Detecting highly overlapping community structure by greedy clique expansion, in: Workshop on Social Network Mining and Analysis, 2010.

[42] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech.: Theory Exp. (10) (2008) P10008+. ISSN:1742-5468.

[43] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Statistical properties of community structure in large social and information networks, in: WWW '08: Proceeding of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 695–704.

Enrico Gregori received the Laurea in electronic engineering from the University of Pisa in 1980. In 1981 he joined the Italian National Research Council (CNR) where he is currently a CNR research director. In 1986 he held a visiting position in the IBM research center in Zurich working on network software engineering and on heterogeneous networking. He has contributed to several national and international projects on computer networking. He has authored more than 100 papers in the area of computer networks and has published in international journals and conference proceedings and is co-author of the book "Metropolitan Area Networks" (Springer, London 1997). He was the General Chair of the IFIP TC6 conferences: Networking2002 and PWC2003 (Personal Wireless Communications) and IEEE Pervasive Computing and Communication (PERCOM) 2006. He served as guest editor for the Networking2002 journal special issues on: Performance Evaluation, Cluster Computing and ACM/Kluwer Wireless Networks Journals. He is on the editorial board of the Cluster Computing, of the Computer Networks and of the Wireless Networks Journals. His current research interests include: Ad hoc networks, Sensor networks, Wireless LANs, Quality of service in packetswitching networks, Evolution of TCP/IP protocols

Luciano Lenzini is a full professor in computer networking at the Faculty of Engineering of the University of Pisa. His current research interests include the design and performance evaluation of MAC protocols for wireless networks and the Quality of Service provision in integrated and differentiated services networks. He is the author and co-author of a high number of papers published in journals and conferece proceedings. He is the author of numerous industrial patents. He is currently on the Editorial Boards of Computer Networks and the Journal of Communications and Networks. He served as chairman for the 1992 IEEE Workshop on Metropolitan Area Networks and for the 2002 European Wireless (EW2002) conference. He served as guest editor of the IEEE Journal on Selected Areas in Communications, special issue entitled Analysis and Synthesis of MAC Protocols and guest editor of the ACM/Kluwer Wireless Networks, special issue devoted to the best papers of EW2002. He has directed several national an international projects in the area of computer networking.

Chiara Orsini is a Ph.D. student in Information Engineering under the supervision of Eng. Enrico Gregori (Institute of Informatics and Telematics of CNR, Pisa) and Prof. Luciano Lenzini (Department of Information Engineering, University of Pisa). Her research interests focus primarily on studying the structural characteristics of the Internet topology at the Autonomous System level of abstraction. In 2011/2012 she spent six months in CAIDA working on dk-graphs.