UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria “Leonardo da Vinci”

Corso di Dottorato di Ricerca in
INGEGNERIA DELL’INFORMAZIONE

Tesi di Dottorato di Ricerca

A Structural Analysis of the Internet AS-level topology

Chiara Orsini

Anno 2013

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

A Structural Analysis of the Internet AS-level topology

Autore:
Chiara Orsini ____________________

Relatori:

Prof. Luciano Lenzini ____________________

Prof.ssa Gigliola Vaglini ____________________

Ing. Enrico Gregori ____________________

Anno 2013
SSD ING-INF/05

To Gabriella, Lorenzo, and Piero,
for their unconditional support.

Sommario

Lo studio delle proprietà strutturali della topologia di Internet a livello Autonomous System (AS) è un importante tema di ricerca che ha attratto un significativo interesse negli ultimi anni. Una conoscenza dettagliata della struttura della topologia consente, infatti, la definizione di modelli sempre più accurati della *Rete*. Inoltre, poichè considerare la struttura sottostante facilita lo sviluppo di algoritmi efficienti, tali modelli sono, a loro volta, utilizzati per lo sviluppo e il test (*syntethic graphs*) di nuove applicazioni e protocolli. Due tematiche tipiche di quest'area di ricerca sono, rispettivamente, l'analisi e l'interpretazione dell'organizzazione complessiva del grafo. Spesso, l'approccio utilizzato è quello delle communities, ovvero la decomposizione del grafo in sottocomponenti. Tuttavia, mentre il tema della *community detection* risulta ampiamente trattato in letteratura, l'interpretazione delle community risulta un argomento poco affrontato.

Il contributo di questa tesi si compone di due parti: uno studio dell'evoluzione della rete Internet negli anni, dal 2004 al 2012, mediante tecniche di community detection e dK-analysis; un'analisi delle classi di AS e delle varie tipologie di connessione che creano le communities individuate. Sebbene, col passare del tempo la topologia di Internet sia cresciuta visibilmente (ad esempio, il numero di nodi è duplicato nell'arco di soli 9 anni), alcune proprietà strutturali sono, invece, rimaste invariate. Un importante risultato derivante dall'analisi strutturale è il fatto che, dopo opportune normalizzazioni, le statistiche ottenibili mediante la decomposizione della rete in k-dense communities rimangono stabili. Per quanto riguarda l'interpretazione delle communities, osserviamo che la continua crescita del traffico e del numero di connessioni instaurate presso gli Internet eXchange Point risulta essere la principale causa della presenza di strutture sempre più dense nel grafo Internet. Infatti, tutti gli AS che formano queste zone estremamente connesse sono membri di almeno un IXP. Un altro aspetto interessante che emerge è il fatto che gli AS che causano la creazione di strutture *ben connesse* sono principalmente di una di queste categorie: Network Service Provider, Content Provider o Content Delivery Network.

Abstract

The study of the structural characteristics of the Internet topology at the Autonomous System (AS) level of abstraction is an important and interesting subject that has attracted significant interest over the last few years. Above all, a deep knowledge of the Internet underlying structure helps researchers in designing a more accurate model of the network; as a result, engineers can design applications and protocols that can take into account the underlying structure and test their projects on synthetic graphs, thereby developing more efficient algorithms. A significant challenge for researchers analyzing the Internet is how to interpret the global organization of the graph as the coexistence of its structural blocks associated with more highly interconnected parts, namely communities. While a huge number of papers have already been published on the issue of community detection, very little attention has so far been devoted to the discovery and interpretation of Internet communities.

The contribution of this work is twofold. First, we study the evolution of the Internet AS-level topology over the last 9 years by means of two innovative approaches: the k-dense method and the dK-analysis. Second, we focus on substructures that play a key role in the Internet connectivity, and we investigate the classes of the ASes and the nature of the connections that create such communities. We find that as the Internet grows over time, some of its structural properties remain unchanged. Although the size of the network, as well as the k_{MAX}-dense index (an index of the maximum level of density reached in a network), has doubled over the last 9 years, we show that after proper normalizations the k-dense decomposition has remained stable. Besides, we provided a clear evidence that the formation of denser and denser sub-graphs over time has been triggered by the proliferation of Internet eXchange Points (IXP) and public peering connections. We found that ASes within most densely-connected substructures are usually Network Service Providers, Content Providers, or Content Delivery Networks; in addition, all of them participate to at least one IXP.

Acknowledgements

This thesis would not have been possible without the support of many people over the years. I take this opportunity to thank some of them. First and foremost, I would like to thank my advisors Ing. Enrico Gregori and Prof. Luciano Lenzini for their guidance and constant encouragement. My sincere thanks also go to Dr. Dmitri Krioukov for offering me two internship opportunities in the Cooperative Association for Internet Data Analysis and leading me working on the dK-graph project. I would like to show my gratitude to my colleagues for making the office a pleasant place to work and for their valuable discussions. Last but not the least, I would like to say thank you to my family for always supporting me.

Contents

List of Figures

List of Tables

1

Introduction and Related Work

The Internet is a constantly changing network that evolves according to independent decisions made by each Autonomous System (AS). Since each AS is a separately managed network, the overall Internet appears as an extremely heterogeneous system. First, the business objectives of the network operators can be very different, thereby determining different network sizes and policies. Second, even if two ASes have the same function, there is no best-practice that could be defined as a *dogma*: for instance, one AS could decide to connect to an Internet Exchange Point and thus have the incentive to create multiple BGP connections to the other members of the facility; whereas the other AS could decide to set up just a single connection to its transit provider. In addition, due to its distributed nature, there is no centralized entity having an exact and global understanding of how the Internet network is evolving. For these reasons, characterizing the Internet topology organization represents a valuable tool to have a clearer view of the interplay between ASes; also, correlating the economic forces behind such interactions with the underlying structures can help predict the future *shape* of the network.

Analyzing and modeling real-world phenomena is a research challenge common to many fields. Social Networks, Communication Systems, Economy, Computer Science, Transportation, Medicine, Biology, and many other disciplines benefit from understanding the structure of their networks [16]. The motivation behind the specific analysis of the Internet topology at the AS level of abstraction is mostly driven by the following purposes:

- the design of more efficient routing protocols [29, 91, 94, 96, 99],
- the development of customized algorithms for searching and for flow optimization [2, 52], and
- the evaluation of the consequences of node failures or virus spreading (what-if scenarios in general) [99].

In order to achieve such objectives, it is necessary to test new protocols and applications against the current Internet topology and its future predictions (e.g. synthetic graph generators) [61, 45, 81, 62, 88, 80, 24, 87, 75, 18, 44]. In other words, it is necessary to have a valid model of the Internet topology, which in turn requires a detailed understanding of the structural characteristics of the network and how these properties change over time [22, 9, 23, 15, 79].

In this thesis we address the specific problem of analyzing the structure of the Internet topology at the AS level of abstraction and its evolution. Such research challenge has a complex nature and it answers to the following questions:

- how do we describe such an heterogeneous network?
- how do we define the *important substructures*?
- how do structures change over time?
- how can we correlate real world phenomena driving the Internet evolution and structural properties?

The main components of this thesis address each research question as follows.

The first contribution of this thesis is to provide a description of the Internet AS-level topology inspired from different data sources. In the current literature Internet topology is commonly described as an undirected and unweighted graph [80]; such approach is actually the simplest and more natural way to represent the inferred Internet AS-level topologies [47, 93, 33] as *it captures both entities and relationships between those entities* [4]. For instance, the discovery of power-laws in the degree distribution [24] and the consequent preferential-attachment models [9] use that descriptive solution. On one hand, maintaining a simple data structure, such as an edge-list, simplifies the process of analyzing the structure of the graph. On the other hand, considering such an approach provides a biased view of the real network. In fact, [44] showed in 2009 that different economic relationships between ASes (provider-customer and peering connections) contributed in a different way to the final shape of the degree distribution, thereby drawing the attention on the influence of Internet eXchange Points (IXP) on Internet topology dynamics. A more recent example is [18] which separates the study of peering and customer-provider connections and consider four distinct classes of ASes to discuss the Internet evolution. Although this trend of enriching the Internet topology has been recently spreading in the specialized literature, the description of IXPs has mostly remained a stand-alone topic. [8] and [3] provide a very detailed description of the IXP *panorama*, yet giving surprising statistics related to the proliferation of public peering connections, however no work correlates Internet structural properties and IXPs so far. In our work we combine three different information, i.e.: topology, inferred AS relationships and IXP data. In detail, we first consider the network as an undirected and unweighted graph in order to be able to apply all kind of structural analyses that derive from graph theory. Then, we focus on specific substructures and we investigate their inferred relationships and the properties that can be derived from their participation at an IXP.

A second contribution of this thesis is to thoroughly discuss what are the most suitable community detection methods on the Internet AS-level topology. The study of the graph as a single entity can hide the underlying structure of the graph. For instance, observing the average values or distributions of some common graph metric (e.g. degree, clustering coefficient) can provide some interesting description of the network, however it does not describes its structural organization or its functional blocks. For these reasons, we decide to focus on a description of a network by means of communities. In most of the approaches published in the specialized literature, communities have been characterized and discovered by exploiting some global property of the graph [26], and the optimization of the modularity is so far the most used approach [68, 12]. However due to Internet organization, such definition may lead to *wrong* results. [68] defines the community as a partition which is densely connected, but it has relatively few connections directed outside. However, if we consider a group of well connected Service Providers, since their business objective is to sell transit, we will find that most of them will have many provider customer connections directed to their customers. We believe that the presence of many outgoing connections should not be a valid reason to deny the presence of a community of Service Providers, thereby providing evidence that modularity is not a universal approach to find communities. In this thesis, we first describe the characteristics of an ideal community within the Internet context: precisely, we sustain that the presence of well-clustered group of ASes resembles a community regardless the number of connections directed outside the community. Then, we show and discuss the differences between three selected community detection algorithms, i.e. the k-core decomposition [84], the k-dense method [83], and the Clique Percolation Method [78]. In order to perform such comparison, we propose a new method to visualize the nesting process that characterizes such community detection methods, i.e. the k-tree; also we investigate the statistical significance of the communities by using the dK-analysis. k-core decomposition is the only method that has been already applied to the Internet AS-level topology: [6], for instance, points out the hierarchies of the graph emerging from k-cores; [15] provides a descriptive model of the Internet made up of three components, i.e. a nucleus, a peer-connected component and a group of dendrites. Although, k-core decomposition has been proved to provide valuable insights into the structure of the Internet, the results of our comparison show that k-dense decomposition better individuate those densely connected components of the network. To the best of our knowledge, neither the k-dense method nor the dK-analysis have ever been applied to the Internet topology and to k-dense communities before our works [38, 76].

The k-dense method and dK-analysis characterize the way this thesis answer to the third research question too. Although many works have covered the description of the Internet evolution, our work is the first that combines classical graph theory properties with more complex and insightful characteristics obtained with the k-dense method. For example, [58] analyzes the growth and the densification of the network showing that the distance parameters surprisingly decrease as time goes by, while [18] observes the evolution of the graph in terms of growth and rewiring. Both works provide a valuable contribution to the description of the Internet evolution, but they do not provide a description of how the structure is changing. The same criticism can be applied to [86] and [75] that describe a new preferential attachment-based model of the Internet evolution and a discussion of the impact of data incompleteness on Internet observed topology over time respectively. A major advance in this thesis with respect to the current literature is the idea of decomposing the graph and then analyzing how the different substructures are interconnected. Such study reveals that the k-dense organization of the Internet graph is stable over time, yet the fact that densest community has a key role for the Internet overall connectivity is time-invariant. Also we focus on these cohesive communities, namely k_{MAX}-denses, and we apply the dK-analysis to understand if the size of the building blocks of these specific substructures change over time.

Finally, we interpret the patterns emerging from our analysis of the structure with the support of additional information, such as the inferred AS relationships or the participation at IXPs. The development of IXPs and the proliferation of public peering connections is an interesting phenomenon that has been driving the evolution of the Internet AS-level topology, yet it is a primary cause behind the formation of denser and denser sub-graphs. The various drivers behind the evolution of the Internet are usually considered as separated topics in the current literature, e.g.: [55] is an interesting dissertation on Internet inter-domain traffic, [54] discusses the economics behind the settlement of peering connections, the aforementioned [3] provides a detailed description of a large European IXP, [18] shows the different evolution of peering and provider customer connections. An appreciable advance with respect to the current state of the art is a thorough interpretation of how each k-dense component contributes to the overall structure of the graph through the analysis of the general business drivers. For instance, we find that the loosely connected components that are the main contributors to the Internet growth correspond to enterprise customer ASes adopting a single- or a multi-homed connection to their providers. On the other hand, we find that the most well-connected sub-graph of the Internet is due to the increasing traffic requirements of Content Providers and Content Delivery Networks, yet a growing number of Network Service Providers adopting open peering policies at IXPs. Surprisingly, the most densely connected community does not include the so-called Tier-1s, i.e. those ASes that are on the top of the *routing hierarchy*.

Thesis organization

The rest of this thesis is organized as follows. In Chapter 2, we present the topologies that we use to analyze the Internet structure evolution and the additional datasets that are required to provide an interpretation of the structural organization. In Chapter 3, we discuss the tools that we adopt to capture the structural properties of the Internet topology. We start with the definition of dK-analysis which is a systematic approach for describing the network and unveiling its building block. Then, we describe and thoroughly compare three community detection algorithms (k-core decomposition, k-dense method, and clique percolation method) that well suit the idea of Internet community that we propose. In Chapter 4, we study the evolution of the Internet AS-level topology from 2004 to 2012, outlining both the growing and the shrinking trends and the time-invariant organization resulting from the k-dense analysis. Then, we provide an innovative analysis on the internal structure of the most densely-connected subgraphs using the dK-series. In addition, we study in detail the outcome of the k-dense decomposition on the most recent Internet topology. In Chapter 5, we analyze the correlation between the observed structural organization of the Internet and the business drivers behind the evolution of the network components. We conclude with outlining the contribution of this thesis in Chapter 6.

2

Internet AS-level datasets

2.1 Background

Internet at the AS level of abstraction can be described as an heterogeneous system made up of different kinds of players, i.e ASes, connecting one to each other according to business and technical (and other) drivers, and exploiting different technologies. Autonomous System definition in [43] states:

> `An AS is a connected group of one or more IP prefixes run by one or more network operators which has a SINGLE and CLEARLY DEFINED routing policy.`

This definition does not give any details related to the business run by the network operator, it does not provide neither a limit on the size of the network, nor the traffic exchanged. As a result, ASes populating the current topology play different roles. BGP (Border Gateway Protocol) is the inter-domain protocol adopted by ASes and it is flexible enough to accommodate all the different policies required by such different players. BGP connections can be roughly classified into two main categories: *provider-customer* and *peering*. In a provider-customer relationship one AS (transit provider) gives access to all destinations in its routing table to the customer AS. Providers often charge their customers using the 95th percentile measurement schema [71], i.e. the cost of the service depends on the amount of traffic exchanged. On the other hand, a peering relationsip is usually free of charge. When such kind of connection is established, both ASes (peers) have access to the other AS' customer cone, i.e. to all the customers of the other AS. Thus, traffic is a key factor in order to understand the dynamics behind the Internet AS-level topology and predict the settlement of new connections. Traffic information are confidential and are largely not available (unless a direct access to routers is usable [55]). Other information, such as the topology, can be retrieved using active or passive measurements (Section2.2), also it is possible to infer the category of the BGP connection (Section 2.4) or collect data associated to a single AS (Section 2.3). This Chapter describes the data available for research and the frameworks used to manipulate those information.

2.2 Internet Topologies

The analysis the structure of the Internet at the AS level of abstraction requires the presence of one or more (if the evolution of the network is studied) topologies representing the status of the network in terms of connections. The collection of the Internet AS-level topology is an on-going research topic, indeed, there is no public tool or registry specifically designed to provide the complete list of all the BGP connections between ASes. In addition, since BGP connections are the results of strategic business decisions, companies are not encouraged to make such information easily available.

Internet topologies are mainly inferred using traceroute-based measurements or collecting BGP dumps. In the first case, an automated process sends active probes (traceroutes) to a set of IP address from multiple vantage points and manipulates the ICMP packets, received as a response, obtaining a list of adjacent IP addresses and then a list of adjacent AS addresses (after de-aliasing). In the latter case, multiple processes collect the BGP messages received by their respective peers, then AS-PATHs are transformed in a list of adjacent ASes.

Both BGP- and traceroute-based data provides an incomplete view of the current Internet AS-level topology. According to [35] BGP-based data has two main drawbacks: many connections are not discovered since the topology inferred using feeder information is biased (e.g. peering connections between leaves are not visible from a large ISP feeder), AS paths gathered are the results of multiple decision processes thus, since only the best path is announced, the information provided might be incomplete. Also a non-fixed number of monitors could cause differences in the resulting topology observed. On the other hand, there are still unresolved issues with the mapping of IP addresses to AS numbers when we are dealing with traceroute-based data: no mapping is available; there are IP addresses that appear to originate from more than one AS; there are AS-sets in the AS-PATH that is used to map IP addresses; however, the most important problem is that there are not-responding AS (especially leaf nodes that represent the vast majority of Internet ASes). For a much more detailed discussion of these and other issues, see for instance [49], [98], [73], [74], [35], [30].

The Cooperative Association for the Internet Data Analysis (CAIDA) [46, 47] (*traceroute*) and the Internet Research Lab (IRL) [73, 93] (*BGP*) are two leaders in the Internet mapping research. We collect 9 Internet topologies, one snapshot for each year from January 2004 to January 2012, for each project using the following procedure:

- CAIDA - *Jan. 2004 - Jan. 2007* - for each year we merge into a single file all the links seen by the Skitter tool from January, 1st to January, 31st [46];
- CAIDA - *Jan. 2008 - Jan. 2012* - for each year we merge into a single file all the links seen by the Ark tool from January, 1st to January, 31st [47];

- IRL - *Jan. 2004 - Jan. 2012* - for each year we download the data related to the last day of January and we keep all the links listed in the file having the last seen attribute more recent than the first day of the month (i.e. January, 1st).

In Figure 2.1 we compare the growth of the number of ASes of the two projects, CAIDA and IRL, and the number of ASes provided by the CIDR report [10]. Although data collected by the ARK tool (2008-2012) are clearly richer than the data collected by the Skitter tool (2004-2007), only IRL has a number of discovered ASes compliant with the CIDR report.

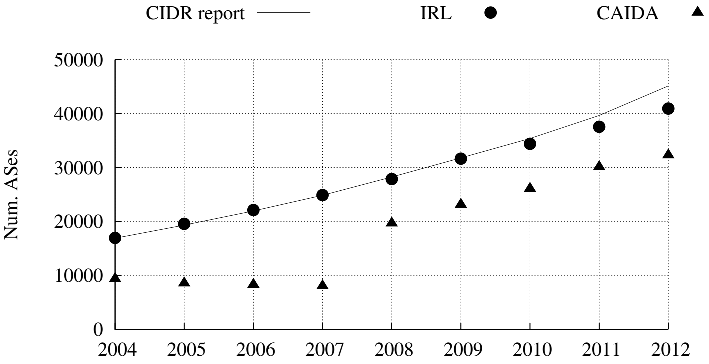


Figure 2.1: Number of unique AS numbers identified by CIDR report, IRL, and CAIDA from 2004 to 2012.

Hereinafter, we refer to the IRL topologies as Internet topologies. We decide to avoid using CAIDA topologies as: a) data are collected using two different tools which have different performances (Figure 2.1), b) the number of ASes discovered is much lower than the number declared by the CIDR report.

2.3 Internet Exchange Points

According to the European Internet Exchange Association[1], an Internet eXchange Point (IXP) can be defined as:

> `A physical network infrastructure operated by a single entity with the purpose to facilitate the exchange of Internet traffic between Autonomous Systems. The number of Autonomous Systems connected should at least be three and there must be a clear and open policy for others to join.`

[1] EURO-IX, `https://www.euro-ix.net/`

IXPs are layer 1 or layer 2 network structures[2] and are not visible within the Internet AS-level topology. However, their presence has been noticeably affecting the dynamics of the Internet evolution over the last decade [44],[8],[3]; in addition, we prove that they are a main driver behind the formation of dense structures - Chapter 5, [32], [38], [40], [39],[76]. When an AS decides to be a member of an IXP, it connects to the exchange point facility and attaches an own router to the LAN of the IXP, i.e. the physical location where members peer. Then, the IXP member adopts one of these *peering policies*:

- *open* - a member with an open peering policy is disposed to peer with any other AS.
- *selective* - a member with a selective peering policy is disposed to peer with those ASes that satisfy certain conditions (e.g. traffic levels, peer is not a customer).
- *restrictive* - it reflects a general willingness not to peer[71].

In order to correlate the presence of IXPs and public peering connections (peering connections using the IXP) to the formation of specific structures in the Internet AS-level topology (Chapter 5), we collect information from PeeringDB [82]. PeeringDB is a freely available database containing information related to public peering [92] and filled by AS participating at IXPs that want to share/show their peering data.

We have a single snapshot of PeeringDB related to January 2012. It contains information related to 2,367 (or 2,345) networks connecting at 317 IXPs. Data retrieved are summarized in Table 2.1.

2.4 Inferred AS relationships

Generally speaking, each BGP connection represents a *provider-customer* or *peering* relationships between two ASes - Section 2.1. Most importantly, identifying the category of a BGP connection enables us to better understand how Internet traffic is routed. Indeed, Internet traffic usually follows the no-valley-and-prefer-customer policy described in [28], i.e.:

- an AS does not provide transit between any of its providers or peers;
- an AS prefers the free of charge customer route over the peer or provider route.

AS relationships are not publicly available, however they can be inferred analyzing the AS-PATH within BGP packets and considering the rules of the no-valley-and-prefer-customer routing policy. Several examples of inference algorithms are present in literature, [27, 89, 19, 42, 17, 13, 33]. However, we use two projects that provide publicly available datasets:

[2] APNIC, `http://www.apnic.net/services/services-apnic-provides/helpdesk/faqs/ixp-address-assignment---faqs`

Table 2.1: Summary of properties extracted from PeeringDB.

(a) Business type.

Business type	AS count
Cable/DSL/ISP	777
Content	582
Educational/Research	93
Enterprise	75
NSP	767
Non-Profit	73

(b) Geographic type.

Geographic scope	AS count
Asia Pacific	205
Europe	754
Global	365
North America	218
Regional	825

(c) Traffic volume.

Traffic volume	AS count
Not disclosed	597
0-1000 Mbps	773
1 - 100 Gbps	761
100 - 1000 Gbps	117
1 Tbps+	19

(d) Traffic ratio.

Traffic ratio	AS count
Balanced	943
Heavy Inbound	77
Heavy Outbound	190
Mostly Inbound	559
Mostly Outbound	598

(e) Peering policy.

Peering policy	AS count
Open	1832
Restrictive	42
Selective	493

- Isolario - data downloaded[3] provides an economic tag for each link using the tagging algorithm described in [33, 34]. This tagging algorithm relies on the information provided by RouteViews/RIS/PCH feeders, then since these feeders provide a view of the network as seen from large ISPs, then many peering connections are unlikely to be seen.
- AS-rank - we collected information related to the size of the customer cone of each AS in the Internet AS-level topology. Customer cone size can be expressed in terms of number of ASes, number of IPv4 prefixes or, IPv4 addresses, that can be reached from a given AS following only customer links [13, 14].

[3] Economic Topologies - February 2012, `http://www.isolario.it/`

3

Tools for Structure Analysis

In this Chapter we present the two approaches that we use to analyze the structure of the Internet AS-level topology graph, i.e.: dk-analysis and community detection methods. We start with providing the definition of dK-series, which are the basis of the dK-analysis, and we continue with describing a method that enable us to understand if a given network can be simply described by the correlation of nodes at distance $d-1$ - Section 3.1. Then we provide the definition and a thorough comparison of three different community detection methods: k-core decomposition, k-dense method, and clique percolation method - Section 3.2. In addition, we show the relationships between these three approaches and we point out the main differences between them analyzing in detail their outcome on simple example, showing their nested structure by means of an innovative visualization method - k-tree, and investigating their statistical significance using the dK-analysis.

3.1 dK-analysis

A systematic approach for topology analysis is represented by dK-analysis [63]. The main idea behind such technique is that the structure of a graph can be described by identifying the statistics related to the degree correlations of nodes at distance $d-1$, namely dK-series. For instance, a $0K$-series defines the constraints required to obtain a graph with the same $0K$-distribution, i.e. the same average degree of nodes; a $1K$-series defines the constraints required to obtain a graph with the same $1K$-distribution, i.e. the same distribution of the degree of nodes; a $2K$-series defines the constraints required to obtain a graph with the same $2K$-distribution, i.e. the same joint degree distribution of pair of nodes. dK-series of probability distributions (or dK-properties) are able to capture progressively (as d is increased) more structural properties of the graph, in addition, when d is equal to the number of nodes of the analyzed graph, the structure of that graph is fully defined (i.e. all the graphs with the same dK-series probability distributions are isomorphic). Such properties are inclusive as each d^*K-property subsumes all dK-properties when $d < d^*$.

Since dK-analysis provides a description of the structure of the graph at different levels of granularity, it is interesting to investigate if there is a minimum d value such that, all the synthetic random graphs having the same dK-properties provide a good approximation of the local and global scale properties of the analyzed graph (e.g. the Internet AS-level topology graph) [48]. The size of the building blocks of a given graph, i.e. d, can be computed by combining the methodologies described in [63] and [48]. The procedure takes as input a graph topology that will be referred to as target graph (as the main goal is to approximately reproduce properties of this graph). Then, it generates from scratch a set of graphs having the same dK-properties of the target graph, namely dK-random graphs. dK-properties are a collection of distributions describing the correlations of degrees of d connected nodes, thus:

- a $0K$-random graph has the same number of nodes and links of the target graph;
- a $1K$-random graph has the same degree distribution (or degree correlation of nodes at distance 0);
- a $2K$-random graph maintains the same correlation of degrees of nodes at distance 1;
- a $3K$-random graph preserves the correlation of degrees of nodes at distance 2 and so on.

Specifically, we start with generating random graphs with a $d = 0$ and we increment d if the current dK-random graphs do not provide a good approximation of the local and global scale properties of the target graph. Each dK-random graph is extracted uniformly at random from the set of all the graphs having the same dK-properties. For $d = 0$ and $d = 1$ we built dK-random graphs using the Erdös-Rényi model [22] and the generalized Havel-Hakimi algorithm [50] respectively. The process behind the creation of dK-random graphs for higher d values is based on the dK-targeting $d(K-1)$-preserving rewiring [63].

When $d \geq 3$, such framework requires a noticeable computational load, indeed the degrees each motif of size d have to be annotated and taken into account in order to match the required statistics, i.e. dK-properties. we use this approach to investigate the statistical significance of different community detection methods - Section 3.3 - and to fully understand the structure of some specific communities of the network emerging from our structural analysis - Chapter 4.

3.2 Community detection

The identification of communities within complex networks is an interesting methodology which provides an insight into the structural characteristics of the overall network. Community structures can reveal the functional organization in networks [57]; in addition, the interactions of many components and the topological properties fundamentally affect the dynamics of the network [77]. 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 [67].
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 [67]. 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 [67].

Due to the great importance of identifying community structure in graphs, a huge variety of community detection algorithms have been developed in computer science, physics, economics, and sociology [26, 67, 95, 12, 25, 31, 66, 20, 77, 57, 68, 56, 59, 97, 84, 83, 78].

A commonly used approach is to evaluate the quality of a community decomposition by its modularity [68], 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. 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. 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.

At the AS-level of abstraction we are interested in finding communities made up of ASes 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 regional transit providers which 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 ASes is likely to form a community although, it is highly probable that the vast majority of their connections will be directed to customer ASes, i.e. outside the community. If the number of connections directed outside the community is very high the product $k_i k_j$ 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, regardless the value of their average out degree fraction. The link density of a subgraph is defined as the fraction of existing connections to possible connections [56]:

$$\rho = \frac{2 \cdot e}{n \cdot (n-1)} \tag{3.1}$$

where e is the number of internal connections and n is the number of nodes within the community. If the community is made up of a single connected component 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). The ODF, Out Degree Fraction of a node i ([77]), is defined as the ratio between the external degree and the total degree:

$$ODF_i = \frac{external\ degree_i}{totaldegree_i} \tag{3.2}$$

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. In other words, we are interested in a definition of community as a form of local organization of the graph, i.e. a community could be defined from some property of the groups of vertices themselves, regardless of the rest of the graph. For all the previous reasons, we investigate the structure of the Internet at the AS level of abstraction considering the following concept of community:

`an unusually densely connected set of ASes`

Selecting dense zones of the Internet AS-level topology graph helps researchers to understand classes of ASes interested in interconnecting with each other, also it helps to shed light on the organization of the graph or the underlying properties of the graph nodes. We focus our attention on three community detection techniques: k-core decomposition [84], k-dense method [83], and Clique Percolation Method (CPM) [78]. These three approaches share the following properties:

- their definition is deterministic;
- each community identifies a set of cohesive nodes;
- they detect a set of “nested" communities with an increasing internal density.

Also, they can be formally correlated one to each other. Briefly, k-core decomposition detects communities by recursively removing nodes with a degree lower than k - Section 3.2.1: k-dense method is based on a recursive removal of those links connecting nodes with less than $k - 2$ common neighbors - Section 3.2.2; lastly, CPM definition is based on specific sets of maximal cliques - Section 3.2.3.

3.2.1 k-core Decomposition

k-core decomposition has been widely applied to a variety of networks [6, 51, 5, 102, 11, 15, 7] since its introduction in 1983 [84]. Such technique can be used to locate the most efficient spreaders in static [51] or in dynamic [64] networks. It can also reveal structural properties: in 2007 [15] showed that Internet topologies obtained with the DIMES project[1] [1] could be modeled as a three-components structure; in 2008 [6]

[1] A traceroute-based tool used to infer Internet AS-level topologies.

proved that cores extracted from traceroute-based Internet topologies where statistical self-similar; in the same year [100] showed the stability of k-core properties over time considering BGP-based Internet topologies; in 2011 we showed the correlation of highly dense k-cores and IXPs in Internet[32].

A k-core is the largest sub-graph made up of nodes having a degree greater or equal to k within the sub-graph. Thus, it can be computed by removing all the nodes with degree lower than k recursively. A more rigorous definition follows. Each graph G is defined as a set of nodes and a set of links. $V_G = \{1, ..., N\}$ is the set of nodes, $E_G = \{e_1, ..., e_M\}$ is the set of links, where $e_m = \{i, j\} \subset V_G$ and $i \neq j$. We indicate with N and M the number of nodes and links respectively. Also, we indicate the degree of a node i in graph G with $k_i(G)$. Then, a k-core of graph G is a sub-graph H_k defined as follows:

$$H_k = \begin{cases} V_{H_k} & = \{i : k_i(H_k) \geq k\} \\ E_{H_k} & = \{e_m : e_m = i, j \subset V_{H_k}\} \end{cases} \tag{3.3}$$

By definition, the k-core decomposition extracts a set of nested sub-graphs, indeed each $k+1$-core is included into a k-core. Also, we refer to the maximum k index providing a non-empty k-core as k_{MAX}.

Each connected component of a k-core is referred to as k-core community. 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, i.e.:

$$i \in H_{k^*} \wedge i \notin H_{k^*+1} \tag{3.4}$$

We define a k-core-shell as the set of nodes having a k-core-index equal to k.

$$k\text{-core-shell} = \{i : i \in H_{k^*} \wedge i \notin H_{k^*+1}\} \tag{3.5}$$

3.2.2 k-dense Method

The k-dense community concept is based on the following intuition. If two nodes are connected together by an edge, it does 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 [83]. In other words, if two ASes share several neighbors they are likely to be part of a same community. The method has been defined in 2009 by [83] and it has been originally applied to a Blog Trackback Network, to a Word Association Network, and to the Wikipedia Reference Network. To the best of our knowledge, we have been the firsts to apply the k-dense method to the Internet AS-level topology graph [38, 41, 76].

k-dense communities can be formally defined using the concept of edge multiplicity [85, 103]. The multiplicity $m_G(i,j)$ of edge (i,j) in graph G is the number of triangles in G containing the edge, or equivalently, the number of common neighbors of connected nodes i and j. By definition, the k-dense H_k of graph G is the sub-graph induced by all the links with multiplicity larger or equal to $k-2$ *in the sub-graph*:

$$m_{H_k}(i,j) \geq k-2. \tag{3.6}$$

This sub-graph can be obtained from G by iterative pruning of all the links with multiplicity smaller than $k-2$. Since all the nodes in the sub-graph H_k share at least $k-2$ neighbors with each of their neighbors, it turns out that all of them have a degree larger or equal to $k-1$. Then, each k-dense is part of a $k-1$-core. Similarly to the k-core, the k-dense method provides a set of nested sub-graphs, indeed each $k+1$-dense is included into a k-dense. Also, we refer to the maximum k index providing a non-empty k-dense as k_{MAX}.

Each connected component of a k-dense 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.:

$$i \in H_{k^*} \wedge i \notin H_{k^*+1} \tag{3.7}$$

We define a k-dense-shell as the set of nodes having a k-dense-index equal to k.

$$k\text{-dense-shell} = \{i : i \in H_{k^*} \wedge i \notin H_{k^*+1}\} \tag{3.8}$$

3.2.3 Clique Percolation Method

A *k*-clique community [78] is defined as the union of all *k*-cliques (complete sub-graphs of size *k*) that can be reached from one or the other through a series of adjacent *k*-cliques (where adjacency means sharing $k-1$ nodes). On the basis of the *k*-clique community definition we can prove that, for each *k*-clique community of order *k*, $community_i(k)$, there exists one and only one *k*-clique community of order $k-1$ (or $(k-1)$-clique community), $community_j(k-1)$, such that:

$$community_i(k) \subseteq community_j(k-1) \tag{3.9}$$

i.e. $community_i(k)$ is a sub-graph of $community_j(k-1)$ (a proof of this can be found in [37]). Hence, all those k-clique communities that are unique (i.e. there is a single community for that *k*) include all the relative *k+1*-clique communities. In addition, by applying (3.9) recursively, we can assert that given a k-clique community of order k^*, there is a k-clique community that completely contains it for each $k < k^*$. A node i is said to have a k-clique-index k^* if it belongs to at least one k^*-clique but is not part of any (k^*+1)-cliques. Also, we refer to the maximum k index providing a non-empty k-clique as k_{MAX}.

Since each node within a k-clique is part of at least one maximal clique of size k, then it turns out to have at least $k-1$ neighbors, each one sharing with it at least $k-2$ neighbors. This implies that each k-clique is part of a k-dense. In summary:

$$k - clique \subseteq k - dense \subseteq (k-1) - core \tag{3.10}$$

The computational load required by CPM is much more demanding than the k-core decomposition or the k-dense method. The extraction of such communities from the Internet AS-level topology has been made doable for the first time in 2011, when we developed a new implementation which store the required data structures in an efficient way and it exploits parallel architectures - FLIP-CPM [36].

3.3 Discussion

k-core decomposition, k-dense method and CPM are similar approaches to detect nested cohesive sub-graphs of increasing density. However, 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. CPM detects tightly connected set of nodes, however if we keep in mind the incompleteness issues related to the collected topologies (Chapter 2), its definition might be too restrictive. In order to point out the differences between these three community detection techniques, we perform the following analyses:

1. we report the outcome of the three community detection algorithms applied to an example topology made up of 14 nodes and 24 links - Figure 3.1.
2. we provide a k-tree representation of the communities extracted from the Internet AS-level topology graph related to January 2012 - Figure 3.2.
3. we discuss the statistical significance of the properties obtained using these community definitions.

3.3.1 Example topology

The example topology in Figure 3.1 highlights some important differences between the three community detection methods. k-cores detected in Figure 3.1 have a lower link density than k-denses or k-cliques, and the k_{MAX}-core, which is supposed to be the most tightly connected community, contains nodes that are loosely related. For example, the distance between node F and node D (both of them belong to the k_{MAX}-core) is 3 hops and the whole graph is made of 8 nodes. CPM detects the same dense zones emerging from the k-dense analysis and, in addition, it selects the maximal clique of size 4 - Figure 3.1. On one hand, the 4-clique community detected by CPM has a higher link density than the k_{MAX}-dense community, On the other hand, the exclusion of node D from the k_{MAX}-clique community highlights how tight are the requirements of CPM. Node D, for example, appears to be strongly related to the “A-B-C-E" community, in addition, if D had a link to E, it would have generated a 5-clique community.

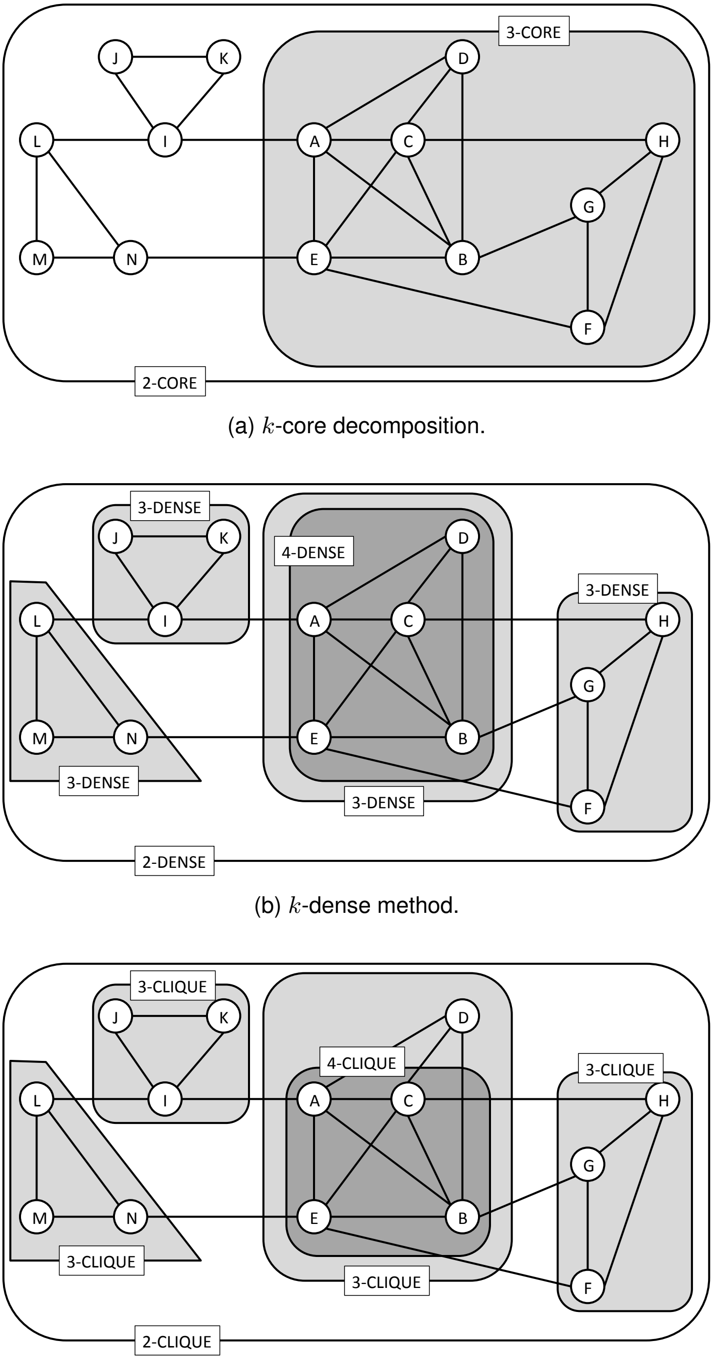
3.3.2 k-trees

In addition to Figure 3.1, we compare the three community detection methods using an innovative representation named k-tree. A k-tree is a graphical representation that we developed in [40] in order to have a better understanding of the nesting process that similarly characterizes k-core decomposition, k-dense method, and CPM. The construction of a k-tree consists of three phases. First, we define *main communities* all those communities that include the k_{MAX} community[2]. Then, we refer to the remaining communities as *parallel communities*. Starting from this definition we can represent communities by means of a tree. Each k-community is a node and we can plot an edge connecting a k-community with its relative $(k-1)$-community (i.e. the $(k-1)$-community which fully contains it). For each k, there is a main community and, very often, more than one parallel community. In Figure 3.2 we report the k-trees related to the k-core, the k-dense and the k-clique communities extracted from the Internet AS-level topology graph in January 2012. CPM provides a huge number of communities if compared to the other two methods. Specifically, it provides more than a single community for each k, however, even if two communities with the same k-clique-index appear separate in the k-tree, they are likely to be highly overlapping [40]. The k-tree structure related to k-cores does not have any parallel community, while the k-tree structure related to k-denses has some parallel communities only for low values of the k-dense index. Such k-tree differences could be interpreted as follows: a) when the k value is low, k-dense and k-clique are able to extract separate communities, while k-core tends to unify all of them into a broader and more loosely connected single community; b) when the k index is high there is only a single cohesive sub-graph of nodes (even if there are many k-clique communities, they are overlapping [40]). Again, the k-dense communities seem to provide the best solution.

3.3.3 Statistical significance

Finally, since understanding the community structure of a graph is a step toward the development of new topology generators, it is fundamental to investigate the statistical significance of the community properties detected. For instance, if random graphs having the same degree distribution of the Internet fully reproduce all its k-core properties, then such properties are a statistical consequence of the observed degree

[2] If there are multiples k_{MAX} communities, we pick one at random. This does not affect the results of the visualization.

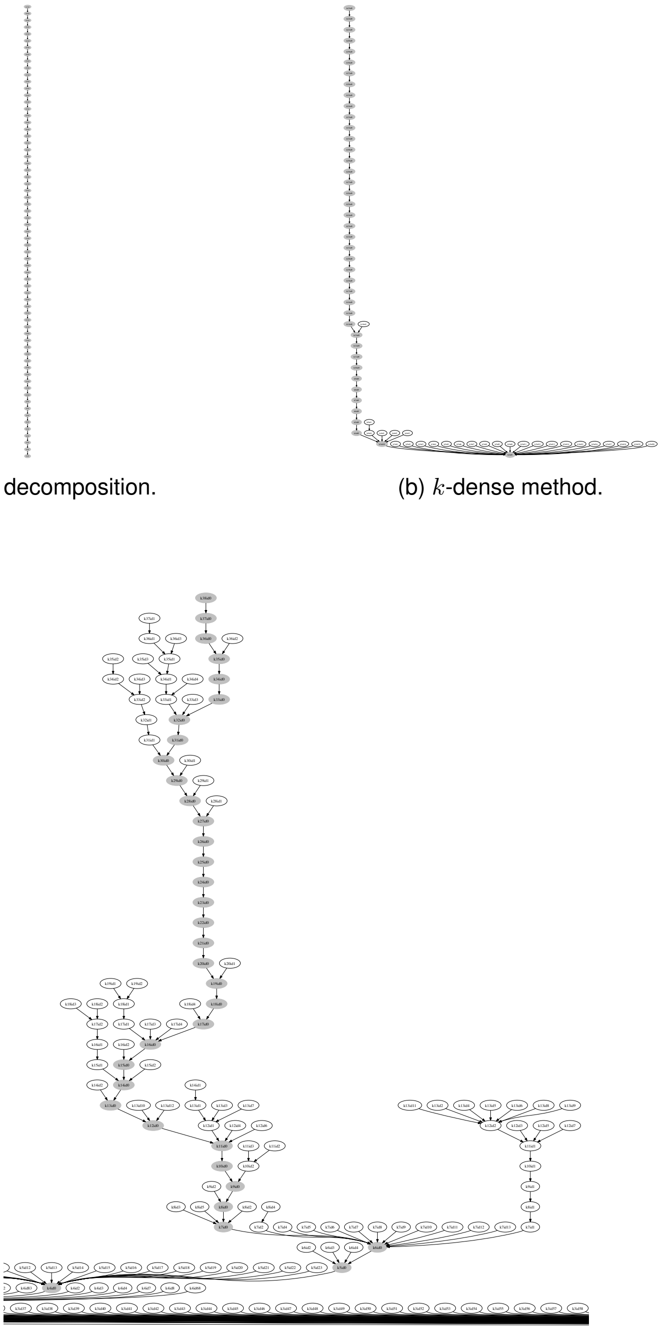


(a) k-core decomposition.

(b) k-dense method.

(c) CPM.

Figure 3.1: Example topology.



(a) k-core decomposition.

(b) k-dense method.

(c) CPM (k-clique communities with k-index lower than 6 have been trimmed).

Figure 3.2: k-trees resulting from the analysis of Internet 2012 snapshot.

Table 3.1: k_{MAX} values in dK-random graphs.

	k_{MAX}-core		k_{MAX}-dense		k_{MAX}-clique	
	mean	*stDev*	*mean*	*stDev*	*mean*	*stDev*
$0K$-random	5	0	3	0	3	0
$1K$-random	102.8	3.05941	68	4.04969	66.9	4.27668
$2K$-random	64.7	0.458258	44.3	0.458258	37.4	0.489898
Internet	73		48		41	

degree distribution, then all $1K$-random graphs would provide the same k-core decomposition. First of all, a degree correlation of nodes at distance 1 is embedded in the definitions of all the three methods. Due to the pruning process that characterizes the k-core decomposition, each node within a k-core has a degree $k^* \geq k$ and it is connected to at least k other nodes having degrees $k^* \geq k$ too. Similarly, each link within a k-dense or a k-clique connects nodes having degrees $k^* \geq k - 1$. Nevertheless, these correlations are not enough to prove any community property dependence of dK-series.

To address this issue we consider a recent Internet topology (May 2012) and we construct dK-random graphs [63] for $d = 0, 1, 2$ as described in Section 3.1. These graphs are random graphs with the same average degree, degree distribution, or joint degree distribution as in the Internet snapshot. For each d we generate 10 realizations, then we extract the k-core, the k-dense communities and the maximal cliques[3] from each of them. For each randomization we compute the k_{MAX}-core, the k_{MAX}-dense, and the k_{MAX}-clique indexes and we compare them against the Internet's value - Table 3.1. When $d \leq 2$, dK-random graphs do not match the k_{MAX} index detected on the Internet graph.

To further prove the independence of these community detection approaches from dK-series with $d \leq 2$, we compute the number of nodes with a given k-core index, the number of links with a given *k*-dense-index, and the distribution of maximal cliques of a given size k and juxtapose them against the Internet's - Figure 3.3. Since the dK-random graphs have different k_{MAX} indexes, to be able to properly compare the different dK-graphs we next perform the following normalization, mapping k-indexes and corresponding numbers of nodes and links to fractions with values between 0 and 1:

- *x-axis normalization*: map each index *k* to what we call the *k-dense-index fraction*:

$$x = \frac{k - k_{MIN}}{k_{MAX} - k_{MIN}}, \tag{3.11}$$

  where k_{MIN} and k_{MAX} are the minimum and maximum values of the k-index in the graph;

[3] Since the definition of k-clique community relies on maximal cliques, we use this information as a proxy.

- *y-axis normalization*: divide the corresponding number of nodes or links by the total number of nodes or links in the graph.

In Figure 3.3 we show, for each d the average value of the property computed on the 10 realizations and the confidence interval with probability 0.8. We observe that neither degree distribution nor joint degree distribution fully reproduce the k-core, k-dense, and k-clique properties of the Internet, meaning that these properties have their own statistical significance.

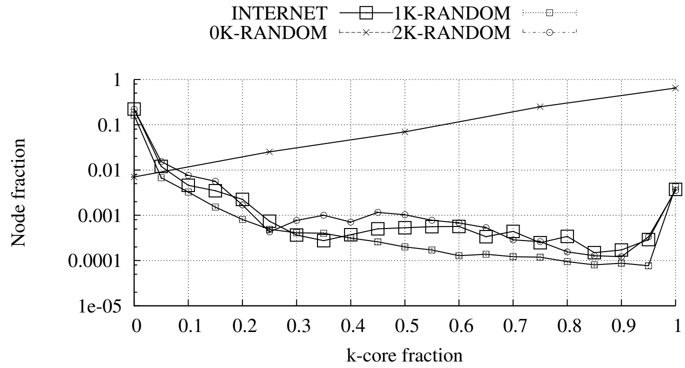
3.3.4 Conclusions

A thoroughly analysis of the properties related to k-core decomposition, k-dense method and CPM - Section 3.3 - has shown the different characteristics of these three approaches, yet it has proved that they provide a peculiar representation of the Internet topology that neither degree distribution nor joint degree distribution are able to reproduce. k-dense method emerges as the best method (compared to k-core decomposition and CPM) to analyze the Internet topology.

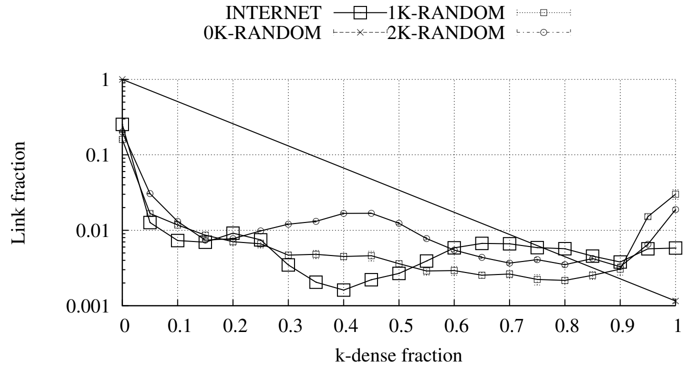
Indeed, although CPM is the only algorithm that accounts both for the locality of the community definition and the possibility of having overlapping communities, it has two main drawbacks. First, since maximal cliques are very fragile structures, we cannot use CPM communities (k-cliques) to analyze some evolutionary trend. We experimentally proved that few edge-swaps can significantly change the number of maximal cliques of a given size - Figure 3.2c. Second, due to the high computational complexity, the detection of k-clique of the Internet AS-level topology graph requires the use of highly parallel machines in order to converge in small amounts of time.

On the other hand, k-core decomposition has the lowest computational complexity (compared to k-dense and CPM), i.e. $O(N + M)$. However it provides more coarse-grained and loosely-connected communities - Figure 3.2a. Finally, k-dense can be thought of as an interpolation between the *k*-core decomposition and CPM - Figure 3.2b. The definition of k-dense suggests a stronger relationship between nodes within the community than the definition of k-core, at the same time, it is more robust than k-cliques. In addition, the amount of computational load is much lower than the one required to compute k-cliques, indeed the computational complexity of the k-dense algorithm is closely related to that of clustering coefficients calculation [65].

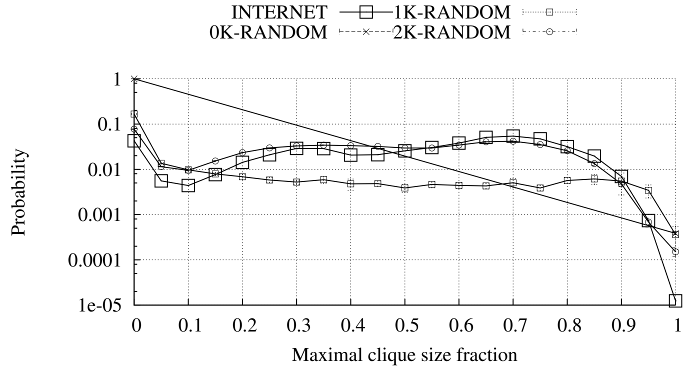
For these reasons, we use the k-dense method both to analyze the evolutionary trends of the Internet topology over the last 9 years and to unveil the details the structural properties of the most recent topology - Chapter 4.



(a) Fraction of nodes with a given k-core index (i.e. nodes in k-core shell divided by the total number of nodes of the graph) related to Internet 2012 and its randomizations.



(b) Fraction of links with a given k-dense index (i.e. links in k-dense shell divided by the total number of links of the graph) related to Internet 2012 and its randomizations.



(c) Distribution of maximal cliques size related to Internet 2012 and its randomizations.

Figure 3.3: Statistical significance of k-core, k-dense, and k-clique properties.

4

Evolutionary Trends of the Internet Structure

We study the evolution of the Internet AS-level structure describing this network as an undirected graph. All the data we collected in Chapter 2 are described as set of undirected graphs sorted in chronological order. In this Chapter we exclusively use the terms nodes and links as we focus on the structural aspects only. An interpretation of the emerging phenomena resulting from this analysis is provided in Chapter 5.

4.1 Related Work

The study of the evolution of the Internet structure provides insight into the creation (and the validation) of new synthetic graph generators, also it helps in evaluating the performance of new protocols as the topology changes [75, 18]. Many works have covered the description of the Internet evolution, a brief summary of the current state of the art follows.

First of all, the Internet evolution can be studied from different perspectives; [101], for example, discusses the different growth of IPv4 and IPv6 topologies (from 1997 to 2009). The main result shown is that IPv4 had a phase transition in 2001, while IPv6 had a phase transition in 2006. Also, both these transitions should be taken into account when developing new models. Although our goal is to study the Internet AS-level topology, such information can be crucial when dealing with Internet AS-level topologies obtained by traceroute measurements. A work that deals directly with Internet AS-level topologies is [21] which evaluate how eight measures (related to node centrality, path length, community structure and scale free structures) of the graph change over time. In detail, it analyzes the Internet topology from January 2002 to January 2010 exploiting Cramér-von Mises Criterion to identify changes between distributions. Authors find that the distributions of most of the measures remain unchanged, except for average path length and clustering coefficient. It is interesting how they discuss this shift as a consequence of peering policies change. A different point of view on Internet evolution is given by [75] which studies the change of the graph from January 2004 to December 2006. In this paper authors focus on topology liveness and completeness problems comparing different data sources. Two evolution trends are highlighted in this paper: a) customer networks are the major cause of the network size growth, b) transit providers tend to form denser and denser structures. In order to provide a more accurate analysis of how different connection strategies influence the Internet evolution, [18] analyzes the topology changes from January 1998 to January 2010 exploiting tagged links. In other words, a business tag is applied to each connection, thus customer-provider and peering relationships are studied separately. Also in this work authors assert that enterprise networks and content/access providers at the periphery are the main contributors to the growth of the Internet. They also study the rewiring activity and they find that content/access providers seem to be the most active.

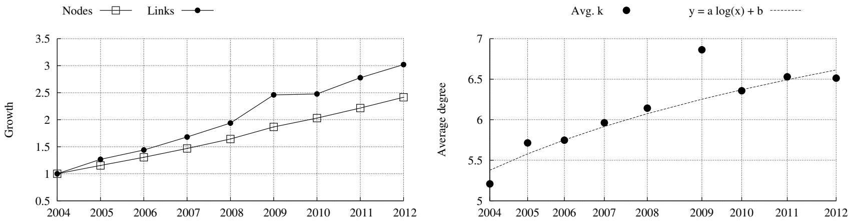
Very few works analyze the evolution exploiting the idea of communities. A singular example is [100], a work that studies Internet evolution from December 2001 to December 2006 by applying the *k*-core decomposition to each topology, and monitoring the properties of the nucleus over time. However, it does not provide details on how the different substructure are connected, also it does not deepen the analysis of the main drivers behind the communities evolution.

In this Chapter we present an innovative framework for the Internet structure evolution analysis. We start with observing how aggregated statistics changes over time in Section 4.2. Then, we present the result obtained by applying the k-dense method to each of the Internet topologies gathered in Chapter 2 and we point out the presence of time-invariant properties in Section 4.3. Furthermore, we focus on the densest sub-graphs of the network, namely k_{MAX}-denses, and we perform a dK-analysis of them in order to show their building blocks in Section 4.4. We conclude our structural analysis with presenting a detailed k-dense description of the most recent Internet topology.

4.2 Basic Trends

Analyzing the Internet graph structure evolution through classic graph theory indexes gives a high-level description of how the graph changes, but it does not provide any insights into the sub-structures that cause such transformation. In order to demonstrate such thesis, we discuss the results we obtain by investigating the Internet topologies from 2004 to 2012. First, we comment the growth of the number of nodes, the number of links, and the average degree. Then, we point out the time-invariance of the average clustering coefficient and the average shortest path. Finally, we plot the increase of the k_{MAX} indexes over time and we show how different is the information embedded in these *innovative* metrics.

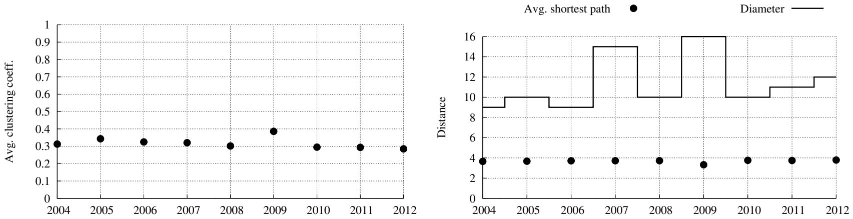
First of all, we observe in Figure 4.1a how the number of nodes and and the number of links grow over time at different rates. In order to emphasize such difference, instead of plotting the absolute values of Table 4.1, we consider the following normalized values:



(a) Nodes and links.

(b) $\bar{k}$ growth and its logarithmic approximation.

Figure 4.1: Growth of the graph over time.



(a) $\bar{c}$.

(b) $\bar{\ell}$ and d.

Figure 4.2: Average clustering coefficient, $\bar{c}$, average shortest path length, $\bar{\ell}$, and diameter, d, over time.

- *nodes growth*, $N(t)/N(t_0)$, where $N(t)$ is the number of nodes at time t and $N(t_0)$ is the number of nodes in 2004, i.e. $16,943$.
- *links growth*, $M(t)/M(t_0)$, where $M(t)$ is the number of links at time t and $M(t_0)$ is the number of links in 2004, i.e. $44,129$.

Figure 4.1a shows that the number of links grows faster than the number of nodes (while the number of nodes more than doubled, the number of links almost tripled). As a results, the average degree, $\bar{k} = 2M/N$, has been increasing too. This increase appears to be a logarithmic function of the number of nodes, i.e.:

$$\bar{k} \approx a \cdot ln(N) - b \tag{4.1}$$

with $a = 1.402$ and $b = -8.2723$, similarly to [79, 53]. Figure 4.1b shows the growth of the average degree, $\bar{k}$, over time and proves the quality of the approximation in Expression 4.1. Although such information may be useful to tune a synthetic graph generator, it does not give any information related to internal structural changes.

Another commonly used property in graph theory is the average clustering coefficient [80], i.e.:

$$\bar{c} = \frac{1}{N} \sum_{i=1}^{N} \frac{ntri_i}{k_i \cdot (k_i - 1)} \tag{4.2}$$

where $ntri_i$ is the number of triangles involving node i, and k_i is the degree of node i. Values in Figure 4.2a show that $\bar{c}$ does not change over time, but it remains stable at 0.3 over the whole observed period. The same considerations apply to the analysis of the average shortest path length, $\bar{\ell}$ over time - Figure 4.2b. Although the graph significantly grows over time, the average shortest path remains stable at 4 for the entire observation period. Such information reveals that the graph evolution embeds some mechanism that preserves the *small-world* property, however it does not provide insight into the sub-structures that make this happen. Figure 4.2b also shows that the diameter, d, has an oscillating trend. Again, since the diameter represents a worst-case by definition, it might not represent a real change of the graph organization.

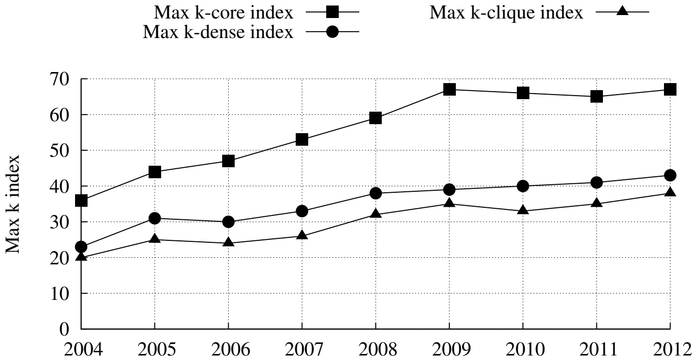


Figure 4.3: Growth of k_{MAX} indexes.

Finally, we show in Figure 4.3 the growth of the k_{MAX} indexes related to k-cores, k-dense and k-clique communities. All the three k_{MAX} indexes have an increasing trend over time, then while the graph was growing more densely connected parts of the graph were forming too. Although a variation of the k_{MAX} index does not necessarily imply a deep change of the structure of the graph, neither is k_{MAX} a property generated by a single node. For instance, a k_{MAX}-dense equal to k^* reveals the presence of a group of at least $k^* - 1$ nodes, each one with at least $k^* - 1$ links directed to nodes of the same group.

In summary, monitoring the change of classic graph theory indexes like the average degree, the average clustering coefficient, or the average shortest path, is useful to both understand some general trends and to tune synthetic graph generators; however, the study of the graph through more innovative techniques, such as community detection methods, enhances our understanding of the evolution at the mesoscopic scale. For these reasons, in Section 4.3 we further investigate the k-dense properties of the Internet over the last 9 years.

Table 4.1: Summary of Internet properties over time.

	2004	2005	2006	2007	2008	2009	2010	2011	2012
N	16943	19559	22106	24876	27854	31630	34393	37546	40936
M	44129	55884	63533	74167	85545	108537	109324	122564	133301
$< k >$	5.20911	5.7144	5.74803	5.96294	6.14239	6.86291	6.35734	6.52874	6.51265
$< k_{nn} >$	521.789	524.684	504.479	526.254	537.261	1024.92	538.418	563.979	586.868
$< c >$	0.312826	0.343384	0.324822	0.32046	0.30197	0.385876	0.294936	0.29359	0.285019
$< b >$	22506.7	26169.6	29997	33869.6	38007.4	36812.9	47454.4	51520.3	57084.1
$< ell >$	3.65692	3.6761	3.71405	3.72319	3.729	3.32666	3.75962	3.74446	3.78901
d	9	10	9	15	10	16	10	11	12
$core_{MAX}$	36	44	47	53	59	67	66	65	67
$dense_{MAX}$	23	31	30	33	38	39	40	41	43
$clique_{MAX}$	20	25	24	26	32	35	33	35	38

4.3 k-dense analysis

In order to gain insight into the structural changes affecting the Internet AS-level topology, we analyze the results of the k-dense decomposition of each snapshot. The rationale behind this choice has been described in Section 3.3.4. In addition, k-dense decomposition characteristics have been presented in Section 3.2 and thoroughly compared to the other community detection methods in Section 3.3, yet its statistical significance have been proved in Section 3.3.3.

We start with computing the size of each k-dense, in terms of nodes and link, for each year. Since the graph size and the k_{MAX}-dense index (hereinafter k_{MAX}) change over time, we applied a normalization similar to the one presented in Section 3.3.3. In detail:

- *x-axis normalization* - we substituted each k by the quantity:

$$k^* = \frac{k - k_{MIN}}{k_{MAX} - k_{MIN}} \tag{4.3}$$

Values obtained with this procedure are referred to as *k*-dense index fractions.

- *y-axis normalization* - we divided each value by the total number of nodes (or links) in the graph, thus each *y* value is a fraction within the range $[0 : 1]$. Values obtained with this procedure are referred to as nodes or links fractions.

Once these normalizations have been applied we cannot use the resulting graphics to deduce an average number of nodes (or links) or refer to a specific *k*-dense index, because both axes provide relative values. Nevertheless, since all the graphs analyzed do not contain isolated nodes, 2 is always the minimum k-dense index, then for each snapshot $x = 0$ corresponds to $k = 2$. Also, 2 is the minimum density level that a node can achieve: nodes with a *2*-dense index are connected to the network, but there is no evidence that they form communities. On the other hand x = 1 means $k = k_{MAX}$. Nodes belonging to this set form the most well-connected sub-graph of the network, however, since each graph has a different k_{MAX}, we cannot say how dense these communities are until we look at a single snapshot.

In Figure 4.4a, we aggregate data related to the 9 snapshots in a single chart, i.e.: we report the average fraction of nodes (and links) and the confidence interval with probability 0.8. We observe that the sizes of confidence intervals are moderate, then average values are highly representative, in the sense that they can approximate pretty well each snapshot (from 2004 to 2012) with the average trend. As a result, we can safely assert that the functions representing the fraction of nodes and links within k-denses are time-invariant.

Due to the nesting process that characterizes the k-dense definition, both functions have a decreasing trend. The rapid decrease of the nodes and links fraction is a peculiarity of the Internet topology. The vast majority of nodes and links belong to k-denses with low k-indexes: 90% of nodes has k-dense fraction lower than 0.1;

k_{MAX}-dense, on the other hand, is usually made up of the 0.2% of the graph nodes and the 2% of the graph links.

In order to gain insight into the overall organization of the graph from a *k*-dense perspective, we analyze how nodes with different k-dense indexes are connected. First, we report the fraction of nodes within each k-dense shell - Figure 4.4b. Then, we count for each k the number of links involving at least one node in the corresponding k-dense shell, and we report the ratio between the counted links and the total number of links of the graph - Figure 4.4c. Both Figures represent aggregated data, i.e. we do not show a function for each year, but we draw the average values and the confidence intervals with probability 0.8. The small size of confidence intervals in Figure 4.4b and in Figure 4.4c confirms the quality of the approximation, thereby demonstrating that such trends are time-invariant.

Figure 4.4c shows that there are two classes of k-dense shells involved in a huge number of links: those with a low k-dense index (k-dense shell fraction is close to 0) and those with a high k-dense index (k-dense shell fraction is close to 1). If we look into the single snapshots and we find that:

- k-dense indexes 2 and 3 are responsible for the leftmost peak;
- k_{MAX} is the only *k*-dense shell which provides the rightmost peak.

In Table 4.2 we report the global properties of the nodes with k-dense index equal to 2, 3, and k_{MAX} averaged over the 9 snapshots. Precisely, we report the average fraction of nodes in the shell divided by the total number of nodes in the snapshot, the average fraction of links involving the nodes in the shell divided by the total number of links in the snapshot, the average Internet degree, the average Internet neighbor degree, the average Internet clustering coefficient and the average Internet betweenness.

A noticeable number of links involving k equal to 2 and 3 is not surprising given that the vast majority of nodes belongs to low k-dense shells - Figure 4.4b. Furthermore, 2 and 3 k-dense shells are the most populated in each snapshot. These nodes have a low average degree and a low average betweenness- Table 4.2.

In order to visualize how these nodes are connected to the rest of the graph we generate Figures 4.5a and 4.5b. We apply the normalization and we report the average values and the confidence intervals with probability 0.8. Also in this case the properties shown are time-invariant, i.e. each snapshot has the same trend. Figures related to links involving 2 and 3 dense shells are pretty similar: a noticeable percentage of their links is directed to low k-dense shells, the rest is directed to medium-high k-dense shells (we discuss Internet business patterns behind these structures in Chapter 5).

k_{MAX}-dense shells, on the other hand, are made up of a small number of nodes characterized by a pretty high average degree and a high average betweenness - Table 4.2. The latter index is a consequence of the high number of links involving the k_{MAX}-dense shell shown in Figure 4.4c. Since betweenness is defined as the number of shortest paths from all nodes to all others that pass through a considered

Table 4.2: Summary of global properties of 2- 3- and k_{MAX}- dense shells nodes.

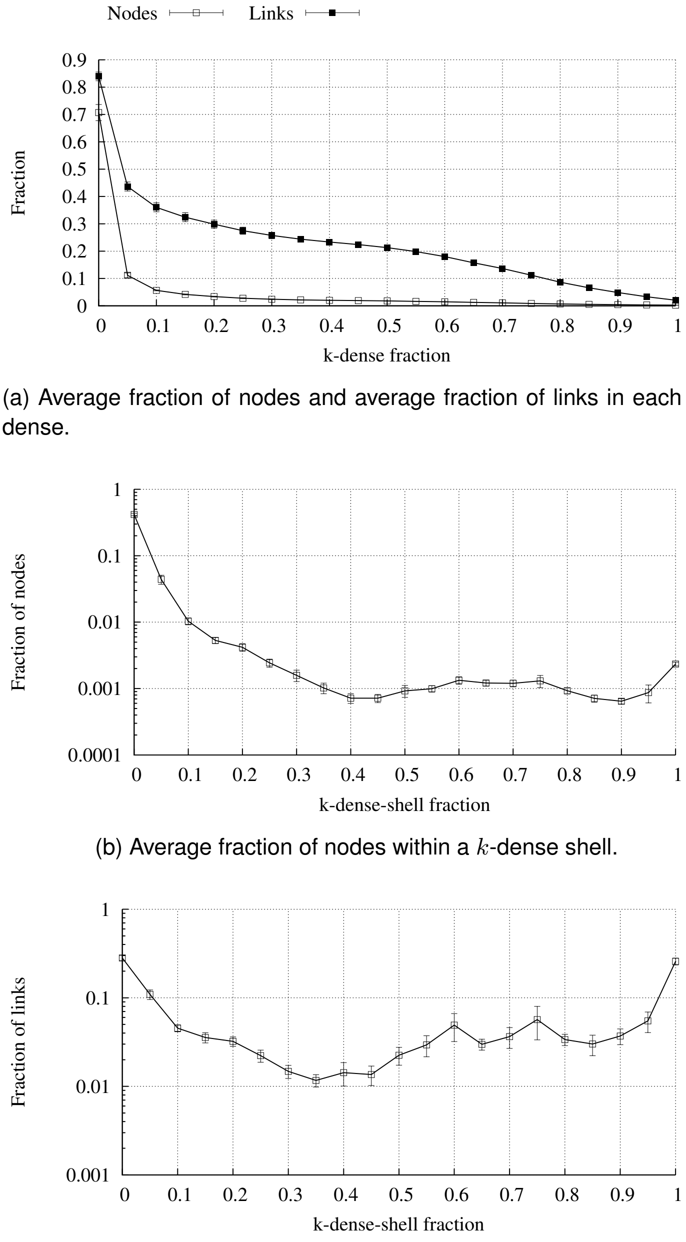
	nodes %	links %	$< k >$	$< k_{nn} >$	$< c >$	$< b >$
2	0.543	0.270	1.635	413.244	0	2047.94
3	0.320	0.304	3.12	846.418	0.751	8979.8
k_{MAX}	0.002	0.257	370.915	160.743	0.205	2754640

node, the higher is the number of links involving a node, the higher is the chance to have some shortest paths traversing it. Nodes in the k_{MAX}-dense shell have a central position (in terms of betweenness centrality) in the graph, and are a key element for the overall connectivity, yet these properties hold for all the considered snapshots.

In Figure 4.5c we report the average fraction of links involving nodes in the k_{MAX}-dense shell (or k_{MAX}-dense[1]). The confidence intervals confirm that also this property is time-invariant. Figure 4.5c is characterized by the presence of multiple *peaks*: k_{MAX}-dense nodes direct a considerable percentage of their connections to nodes within 2- and 3- dense shells, but also to nodes which are part of more densely-connected parts of the graph (we discuss the rationale behind this property in Chapter 5).

The importance of the k_{MAX}-dense in the graph connectivity has been a constant outcome of the k-dense analyses over the 9 considered snapshots. Thus, we further investigate in Section 4.4 the internal structure of these cohesive *nuclei*.

[1] When $k^* = k_{MAX}$ the set of nodes within the k_{MAX}-dense and the set of nodes within the k_{MAX}-dense shell is the same.

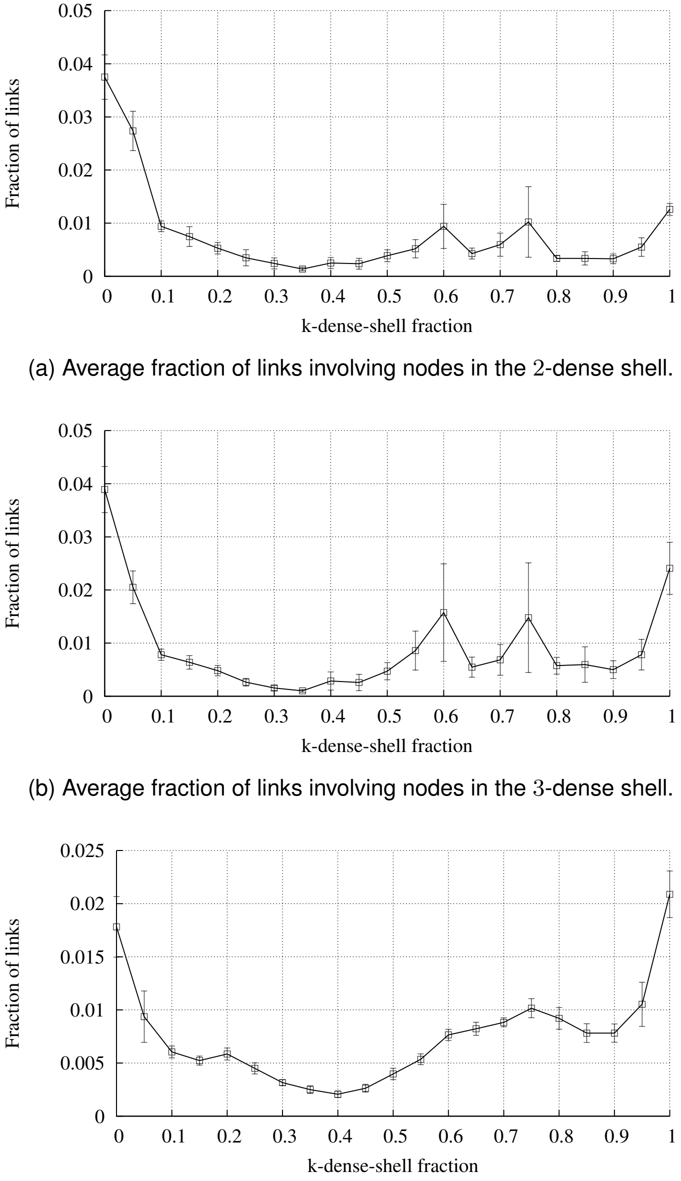


(a) Average fraction of nodes and average fraction of links in each k-dense.

(b) Average fraction of nodes within a k-dense shell.

(c) Average fraction of links involving nodes in a k-dense shell.

Figure 4.4: Organization of k-denses and k-dense shells. Average values and confidence intervals with probability 0.8 are provided for each function.



(a) Average fraction of links involving nodes in the 2-dense shell.

(b) Average fraction of links involving nodes in the 3-dense shell.

(c) Average fraction of links involving nodes in the k_{MAX}-dense shell.

Figure 4.5: Average fraction of links involving nodes in a k-dense shell and originating in the 2-, 3- or k_{MAX}- dense shells. Average values and confidence intervals with probability 0.8 are provided for each function.

4.4 Structure of the k_{MAX}-denses

In this Section we discuss the structure of the k_{MAX}-denses using a novel approach based on *dk*-series [63] - Chapter 3. Specifically, we investigate the problem of building blocks in order to gain insight into the complexity of the k_{MAX}-dense structures. Indeed, uncovering the structural properties of k_{MAX}-denses is useful for those who are interested in developing new models of the Internet topology. For instance, according to our findings in Section 4.3 an accurate model should include a very dense community which is *largely* connected to the rest of the graph.

We start the analysis of k_{MAX}-denses with listing in Table 4.3 the main properties of the k_{MAX}-denses. Each sub-graph is made up of a small number of nodes and a pretty high number of links. In order to understand how dense are these graphs, we computed the link density. Link density values reported indicate that, on average, each selected graph is made up of 85%[2] of links of a correspondent full-mesh topology. Due to the high link density, one may think that a $0K$-random graph would reproduce the main properties of the k_{MAX}-dense graph. In other words, a natural question that arises: does any graph with the same average degree, i.e. the same number of nodes and links (and then the same link density), approximate the k_{MAX}-dense?

Table 4.3: Summary of k_{MAX}-dense properties: k_{MAX} it the k_{MAX}-dense index, N is the number of nodes; M is the number of links, LD is the link density, $< ODF >$ is the average out degree fraction.

	k_{MAX}	N	M	LD	$O\bar{D}F$
2004	23	36	567	0.90	0.81
2005	31	53	1198	0.87	0.75
2006	30	59	1419	0.83	0.76
2007	33	52	1197	0.90	0.81
2008	38	60	1586	0.89	0.73
2009	39	97	3422	0.73	0.72
2010	40	74	2289	0.85	0.78
2011	41	82	2742	0.83	0.77
2012	43	80	2670	0.84	0.80

We tackle this problem using the standard dK-statistical analysis described in Chapter 3 [63, 48]. We consider each k_{MAX}-dense graph, one at a time and we refer to this selected topology as *target* graph. Then, we generate random graphs with a $d = 0$ and we increment d if the current dK-random graphs do not provide a good approximation of the local and global scale properties of the target graph. Specifically, starting from $d = 0$ we perform the following procedure:

1. generate 20 independent dK-random graphs (also referred to as *realizations*);

[2] The average link density is equal to 0.850055

2. compare the properties[3] of the target graph with the properties of the dK-random graphs;
3. stop the procedure if the current dK-properties are sufficient to describe the target graph, otherwise increment d and restart from 1).

Each dK-random graph is extracted uniformly at random from the set of all the graphs having the same *dk*-properties. Also, as explained in Chapter 3.1, for $d = 0$ and $d = 1$ we build dK-random graphs using the Erdős-Rényi model [22] and the generalized Havel-Hakimi algorithm[50] respectively; the process behind the creation of dK-random graphs for higher d values is based on the dK-targeting $d(K-1)$-preserving rewiring [63].

The standard dK-statistical analysis on all the k_{MAX}-denses detected points out all the k_{MAX}-denses are $1K$-random graphs, thereby highlighting another time-invariant feature of the graph related to k-dense features. In Figures 4.6, 4.7 and 4.8 we report the results of such analysis for the k_{MAX}-denses related to the 2004, 2008, and 2012 snapshots respectively.

Degree distribution

The degree distribution is a property fully defined by $1K$-series. By definition, $0K$-random graphs have the same average degree of the target graph, however Figures 4.6a, 4.7a, and 4.8a show that they are not sufficient to approximate the degree distribution. Indeed, degree distributions of $0K$-random graphs are characterized by a different trends and large confidence intervals. $1K$-random graphs have the same degree distribution of the target graph by definition.

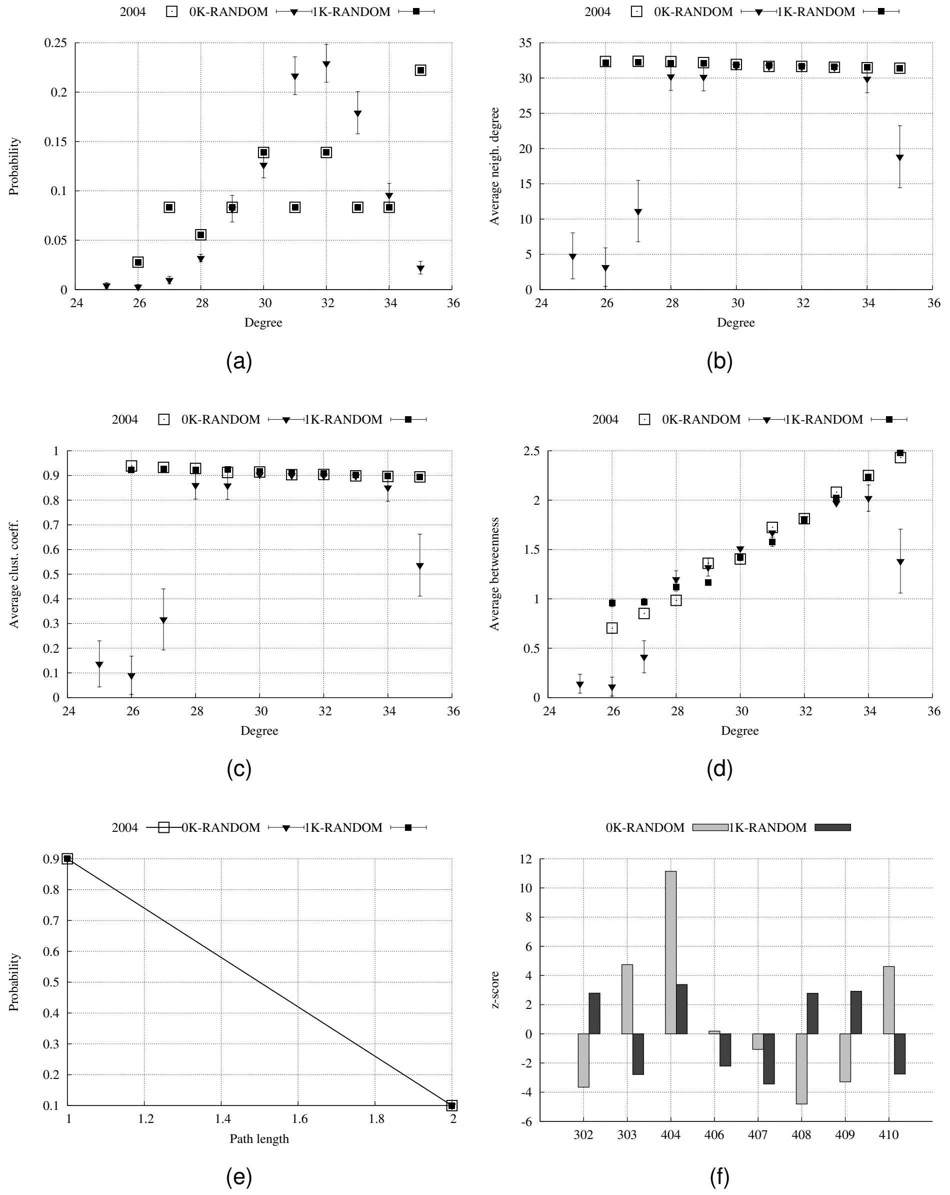
Average neighbor degree over degree

The average neighbor degree over degree function is a property fully defined by the $2K$-series. Similarly to the degree distribution case, the average neighbor degree functions representing $0K$-random graphs visibly differ from the target distribution; in addition, they are characterized by large confidence intervals. In contrast, $1K$-random graphs well approximates average neighbor degree over degree function, and the small sizes of the confidence intervals prove the quality of such approximation. Results are shown in Figures 4.6b, 4.7b, and 4.8b.

Average clustering coefficient over degree

The average clustering coefficient over degree function is a property fully defined by the $3K$-series. Figures 4.6c, 4.7c, and 4.8c provide the same result obtained for the average neighbor degree over degree function, i.e.: while $0K$-random graphs average values are distant from the target, $1K$-random graphs almost match the target function and are characterized by small confidence intervals.

[3] We consider properties that do not depend on the current d, e.g. average neighbor degree or average clustering when $d < 2$, or diameter.



(a)

(b)

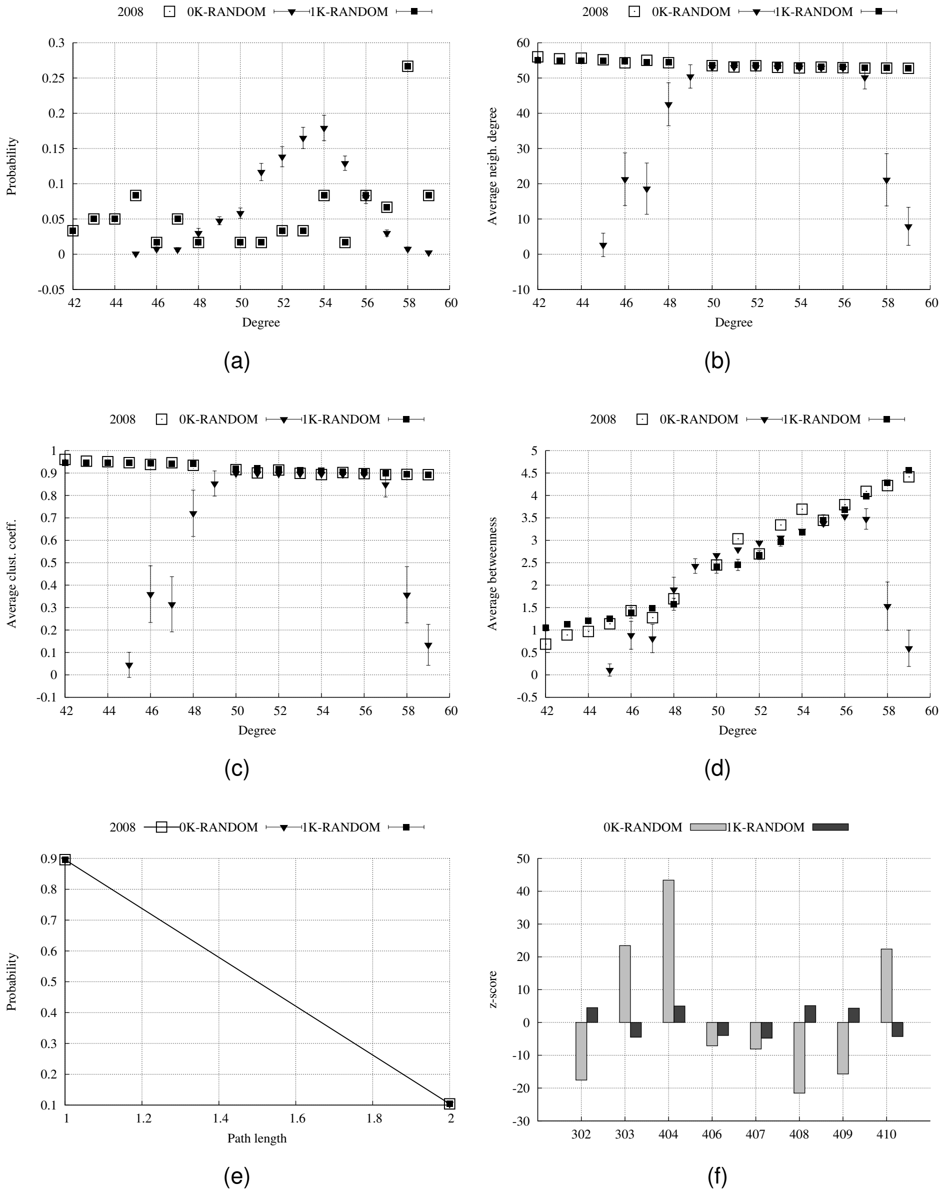
(c)

(d)

(e)

(f)

Figure 4.6: **2004 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: 4.6a degree distribution, 4.6b average neighbor degree over degree, 4.6c average clustering coefficient over degree, 4.6d average betweenness over degree, 4.6e average shortest path distribution, 4.6f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

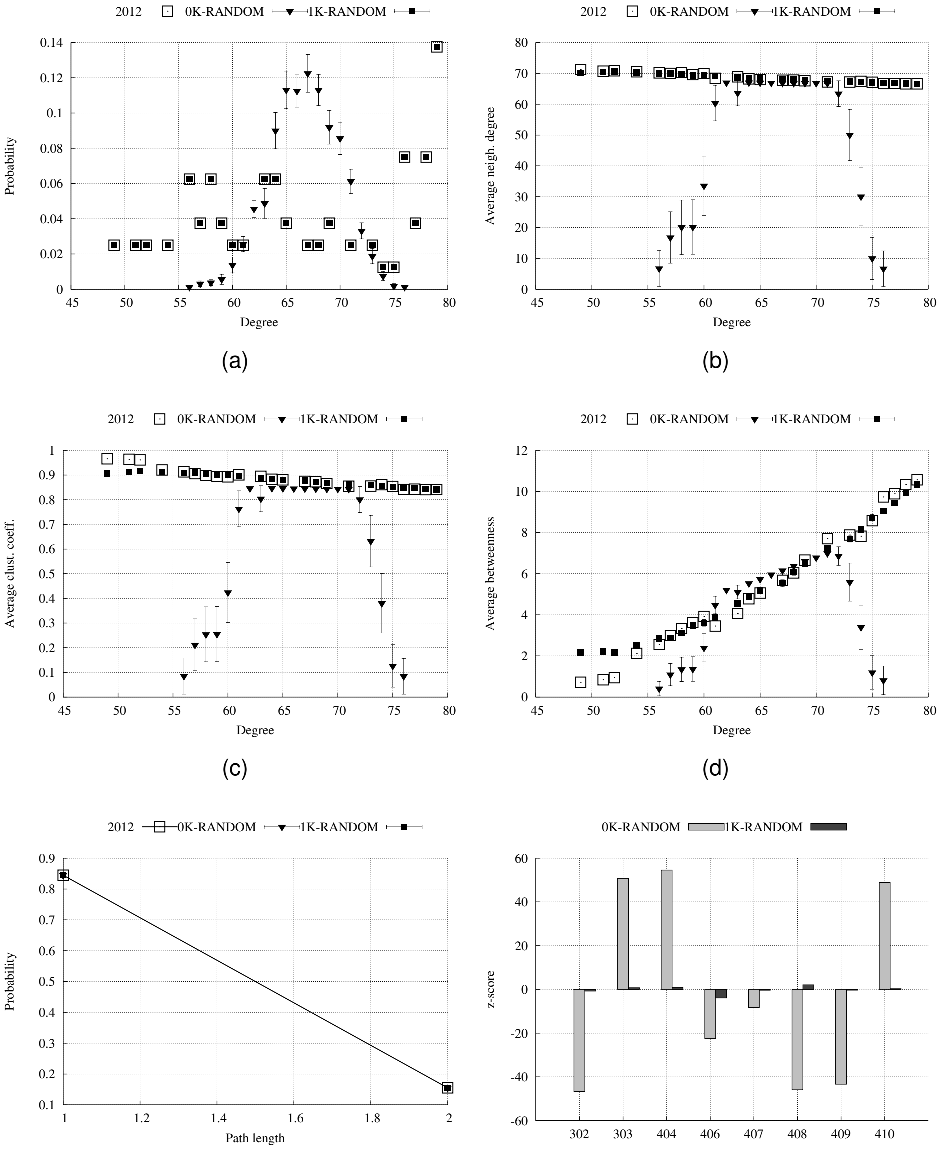
(c)

(d)

(e)

(f)

Figure 4.7: **2008 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: 4.7a degree distribution, 4.7b average neighbor degree over degree, 4.7c average clustering coefficient over degree, 4.7d average betweenness over degree, 4.7e average shortest path distribution, 4.7f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a) (b) (c) (d) (e) (f)

Figure 4.8: **2012 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: 4.8a degree distribution, 4.8b average neighbor degree over degree, 4.8c average clustering coefficient over degree, 4.8d average betweenness over degree, 4.8e average shortest path distribution, 4.8f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.

Average betweenness over degree

Average betweenness over degree is not fully defined by a specific d in theory. Figures 4.6d, 4.7d, and 4.8d show the results related to this property and support the conclusions made for the previous metrics. Although the differences between the target function and the $1K$-random approximation are more evident than in the above properties, the target and the $1K$-random functions are strongly correlated, have close values; also, $1K$-random functions do not present visible confidence intervals. On the other hand, functions related to $0K$-random graphs do not approximate the target average betweenness over degree functions.

z-score of motifs of size 3 and 4

- The number of motifs of size 3 and 4 in a graph is an information embedded within $3K$-series and $4K$-series respectively. We do not compare the number of motifs of a given type directly but we use the *z-score*, a statistical tool that expresses how much an experimental result is statistically *far* from a known distribution. z-score is a dimensionless index that count the distance between a raw score and the mean of the distribution in units of standard deviation, i.e.:

$$z = \frac{x - \mu}{\sigma} \tag{4.4}$$

The z-score shown in Figures 4.6f, 4.7f, and 4.8f is the difference between the number of occurrences of a motif in the target graph (x) and the average number of its occurrences in the considered dK-random graph (μ), divided by the standard deviation of its occurrences in the considered dK-random graph (σ). Precisely, mean and standard deviation are computed over the different realizations of a specific dK-random graph. z-score gives a measure of how statistically significant is each motif in the target graph when compared to a dK-random graph. Results confirm that there is no statistically significative motif in the target graph when compared to $1K$-random realizations, that is the distribution of motifs in the target graph is statistically *similar* to the distribution of motifs within $1K$-random graphs. On the other hand, z-scores related $0K$-random graphs prove that they have a different distribution of motifs.

Properties shown in Figures 4.6, 4.7, and 4.8 demonstrate the $1K$-randomness of the k_{MAX}-denses related to 2004, 2008, and 2012 snapshots. In order to provide a complete description of the structural properties of all the k_{MAX}-denses we apply the above framework to 2004, 2005, 2006, 2007, 2009, 2010, and 2011 snapshots. Results in Appendix A are compatible with findings related to 2004, 2008, and 2012 snapshots.

In summary, the standard dK-statistical analysis of the k_{MAX}-denses proves that:

- although these graphs have a high link density (Table 4.3), a graph having the same number of nodes and links is not sufficient to reproduce most of their structural properties.

- graphs created constraining the degree correlation of nodes at distance 0, i.e. imposing the number of nodes with a given degree, well approximate the target graphs.
- $1K$-randomness of k_{MAX}-denses is another time-invariant feature of the Internet graph.

4.5 Current Internet Structure

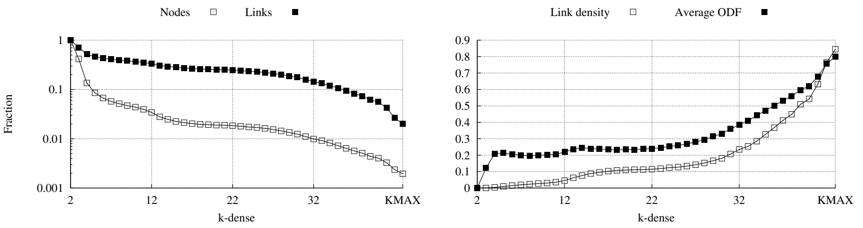
In this Section we investigate the structural properties of the current Internet AS-level topology graph. We provide the absolute values related to the k-dense properties that have been shown in an aggregated way in Section 4.3, thereby characterizing the Internet 2012 snapshots.

Internet 2012 AS-level topology graph, hereinafter Internet graph, is made up of 40,936 nodes and 133,301 links. The application of the *k*-dense algorithm to the Internet AS-level topology graph provides a k_{MAX} equal to 43. The vast majority of k-denses are made up of a single connected component, thus terms k-dense and k-dense community often refer to the same sub-graph. Precisely, k-denses are made up of more than a single connected component when k is equal to 3, 4, 5, and 14 only - Figure 3.2b. When we analyze these k-denses we observe a common feature: there is a single large connected component which represents the vast majority of all the k-dense nodes, while the remaining community (or the remaining communities) is made up of a negligible number of nodes. For instance, the *main*[4] 3-dense community is made up of 16,955 nodes and 95,186, while the remaining 20 parallel communities have an average size of 3.5 nodes and 4.1 links. In this Chapter we focus on the analysis of the k-denses and k-dense shells and we postpone the discussion of these specific sub-structures in Chapter 5.

We start the analysis of the Internet k-dense structure by reporting the size in terms of fraction of nodes and links in each k-dense - Figure 4.9a. By definition each k-dense is a sub-graph of a $(k-1)$-dense, hence these 42 k-denses are nested ($k \in [2:43]$). As a result, the number of nodes and links in a k-dense decreases as k increases. The vast majority of nodes belong to low k-denses, only 10% have a k-dense index larger than 5. Also, the fraction of links within each k-dense decreases more slowly than the percentage of nodes, for the higher the k-dense index the better-connected the relative sub-graph.

In Figure 4.9b we plot the average out degree fraction (ODF) and the link density of each k-dense. 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 nodes within a k-dense are. The increasing trend of the link density function indicates that the higher the k-dense index is, the more clustered the corresponding k-dense is. All the k-denses with a $k \geq 40$ have a link

[4] See k-tree definition in Chapter 3



(a) Nodes and links in each k-dense.

(b) Link density and average out degree fraction of k-denses.

Figure 4.9: k-dense size, link density and average out degree fraction.

density larger than 0.5. This kind of nodes also presents high average ODF values which means that, although they are really well connected to the other k-dense nodes, they mostly direct their connections outside such sub-graph. On the other hand, low k-denses are characterized by a small link density and a small ODF, for only a small percentage of nodes are outside these k-denses.

From Figures 4.9 we cannot infer how k-denses interact. To better understand such feature, we plot the volume of connections originated by each k-dense-shell and the number of nodes in the corresponding shell - Figure 4.10. Using k-dense-shells instead of k-denses helps separating the contribute of each group of nodes to the graph connectivity. Figure 4.10 highlights the presence of two groups of nodes involved in a very high number of links: the first group is made up of nodes with a k-dense-index equal to 2 and 3, the second group of is nodes in the k_{MAX}-dense shell. In detail, nodes in the 2- and 3- dense shells are involved in 34,697 and in 33,845 links respectively, thereby being involved in more than 50% of the entire Internet links. Such high percentage is not surprising if we consider that the sum of nodes in these two k-dense shells represent the 86% of the total number of nodes in the graph. On the other hand, there are only 80 nodes in the 43-dense shell (or 43-dense), yet they are involved in 39,677 links, i.e. almost 30% of the total links of the Internet graph. These nodes have a central position within the graph, and they play a primary role in Internet connectivity. Since the k_{MAX}-dense and the k_{MAX}-dense shell correspond by definition, we can further discuss the k_{MAX}-dense shell properties observing the link density and the average out degree fraction values in Figure 4.9b. Due to the *small* size of the k_{MAX}-dense, even if the internal structure is close to a complete graph (link density is larger than 0.8), each node direct the majority of its links outside the k_{MAX}-dense.

In order to visualize how nodes in the 2-, 3-, and k_{MAX}- dense shells are connected to the rest of the graph we generate Figures 4.11a, 4.11b, and 4.11c. Figures related to links involving 2- and 3- dense shells are pretty similar: a noticeable percentage of their links is directed to low k-dense shells; the rest is directed to medium-high k-dense shells. The function of links involving k_{MAX}-dense shell nodes is characterized by the presence of multiple *peaks*: k_{MAX}-dense nodes direct a considerable

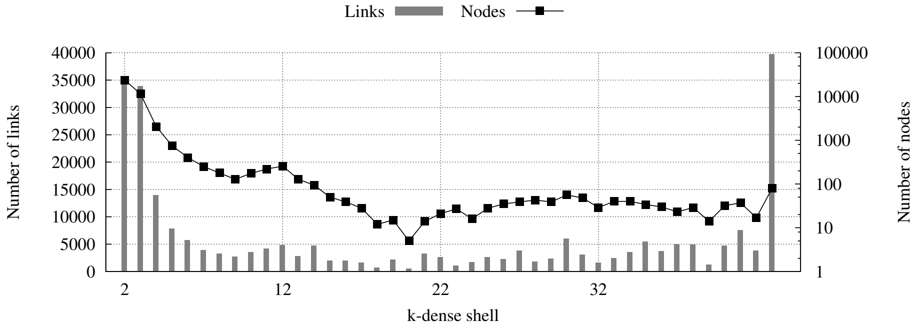


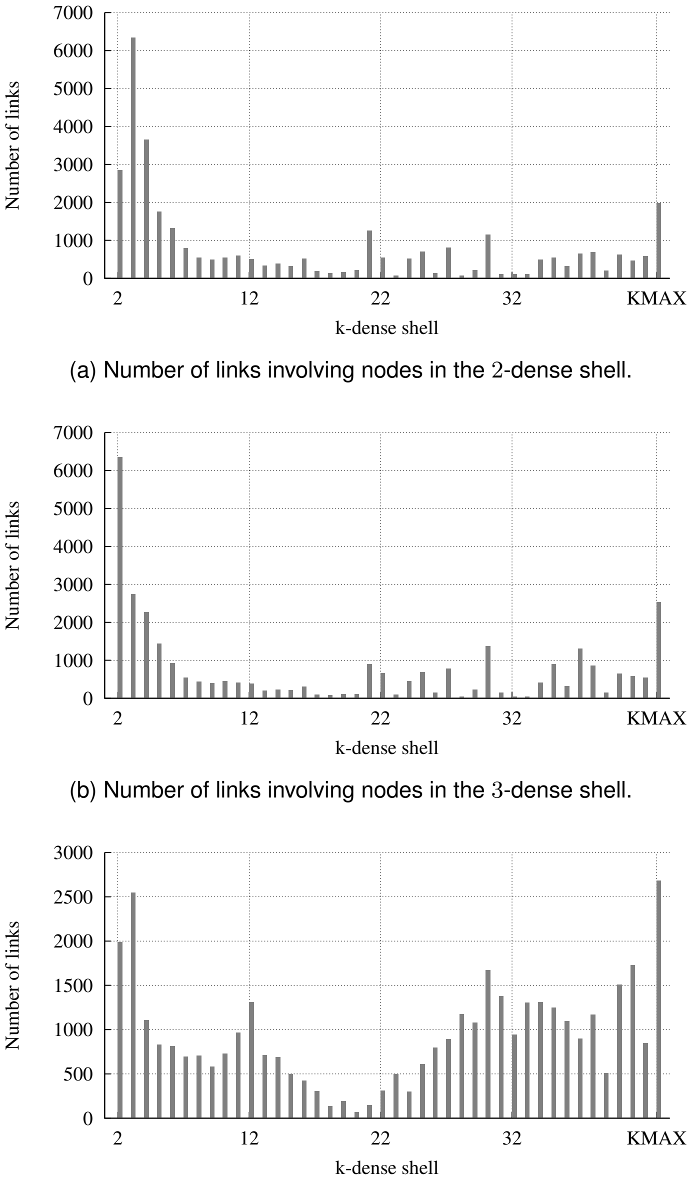
Figure 4.10: Number of nodes in each k-dense shell and number of links involving each k-dense shell.

percentage of their connections to nodes in 2- and 3- dense shells, but also to nodes which are part of more densely-connected parts of the network.

In summary, the analysis of the current Internet graph topology indicates that:

- only a small subset of nodes belong to very well-connected denses;
- low k-dense shells are highly populated, also they are involved in more tha 50% of the overall links. In addition, such links are mostly directed to low k-dense shells or the the k_{MAX}-dense
- higher k-denses are characterized by a high level of clusterization, nevertheless they tend to direct most of their link to nodes outside the dense;
- k_{MAX}-dense nodes have a primary role in graph connectivity since they are involved in a huge number of Internet links (almost 30% of the entire number of links). Links involving k_{MAX}-dense nodes are directed both to nodes with a medium-high k-dense index and to nodes with a low k-dense index.

Results shown above describe an Internet structure which is compliant with the Jellyfish model proposed in [90]. 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 links. [90] provides 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 and systematic description of the maximum level of density which characterize each node.



(a) Number of links involving nodes in the 2-dense shell.

(b) Number of links involving nodes in the 3-dense shell.

(c) Number of links involving nodes in the k_{MAX}-dense shell.

Figure 4.11: Number of links involving nodes in a k-dense shell and originating in the 2-, 3- or k_{MAX}- dense shells.

5

Internet Evolution drivers

In Chapter 4 we presented the results of our analysis on 9 snapshots (2004 - 2012) of the Internet AS-level topology. Each snapshot has been treated as an undirected graph made up nodes and links, however no *context* information have been added so far. In this Chapter we interpret previous results keeping in mind that nodes represent ASes and that links represent connections at the AS-level, or in other words, they are business agreements between two entities that enables their networks to exchange traffic. As reported in Chapter 2, connections at the AS-level can be of two kinds: *provider-customer* or *peering*. From a technical point of view the difference between these two kinds of connection resides in the routing information exchanged: transit providers announce all the destinations to their customers, and thus forward all the traffic that their customers send uplink; peers mutually announce a limited set of destinations, i.e. their own network prefixes and their customer networks. Typically the transit provider charges its customers using the 95th percentile measurement method[1], i.e. the cost of the service depends on the amount of traffic exchanged; on the other hand, a peering connection is usually free of charge (if maintenance costs are not considered). In addition, we provide two further definitions that are typical of the Internet AS-level ecosystem: public peering connections and Tier-1 ASes .

First, a peering agreement is referred to as public peering if the two ASes use an IXP to settle such connection. As described in Chapter 2, IXPs are facilities where each AS member can easily settle peering connections; in detail, each member can create either single peering connections directed to other selected members (*selective/restrictive peering* policy) or peering connections to an undetermined number of members (*open peering* policy).

Second, a provider which does not need to pay a transit provider in order to reach any destination, is known to be a Tier-1. As a result, all the Tier-1 ASes connect mutually with peering connections forming a clique topology. These ASes are characterized by a high number of provider-customer connections (directed to their customer),

[1] http://drpeering.net/white-papers/Internet-Service-Providers-And-Peering.html

whereas the number of peering connections is less remarkable. Also, if the are member of an IXP, they usually adopt a restrictive peering policy.

Although we cannot infer exactly neither Tier-1 ASes nor connections on IXPs, we use the information provided by PeeringDB [82], Isolario [33] and AS-rank [13] to show the correlation of these two practices to the Internet structure. In detail, if two ASes are connected and, at the same time, they are members of the same IXP, we assume that such connection is likely to be settled using the IXP facility. Such hypothesis is even stronger if the two ASes declare an open peering policy.

In the following Sections we comment the structural characteristics of the Internet AS-level topology obtained using the k-dense method and we correlate such information to IXP related data and AS relationships - Chapter 2. We start with discussing the growth of the k_{MAX}-dense index over time and we support its relation with the proliferation of public peering connections. Then, we describe the most recurrent patterns that characterize 2- and 3- denses. We continue our discussion with focusing on the presence of Tier-1 ASes in the medium k-dense shells. Subsequently, we thoroughly investigate the nature of ASes forming the k_{MAX}-dense and we discuss its $1K$-randomness. Finally we take into account the incompleteness problem raised in Chapter 2 and we provide some insight into the differences that an analysis of a more complete topology would introduce.

5.1 k_{MAX}-dense index growth

k_{MAX}-dense-index growth is mostly due to the development of IXPs and, the proliferation of settlement of public peering connections. A first piece of evidence is the structural change of the Internet topology that we witnessed in the last decade due to a different proliferation of peering and provider-customer relationships. Initially peering relationships born to drop the cost of customer provider relationships, however, as the price of transit noticeably decreased, such solution became less attractive (unless a large amount of traffic had to be exchanged). As described in [54] peering continues to grow and is still contributing to *a much larger middle tier compared to the backbone*. Large Content Providers (CP) and Content Delivery Networks (CDN), which nowadays represent a noticeable percentage of the overall Internet traffic [54, 55]), are a primary driver in this process, in fact:

1. a shorter path between these networks and subscribers provides better performances;
2. although the traffic exchanged is highly asymmetric, for most ISPs the connection to the content providers is vital.

For instance, [55] asserts that the majority of AS-level traffic flows directly between large CPs, CDNs, data centers and consumer networks, also it shows that 150 ASes originate more than 50% of all Internet inter-domain traffic.

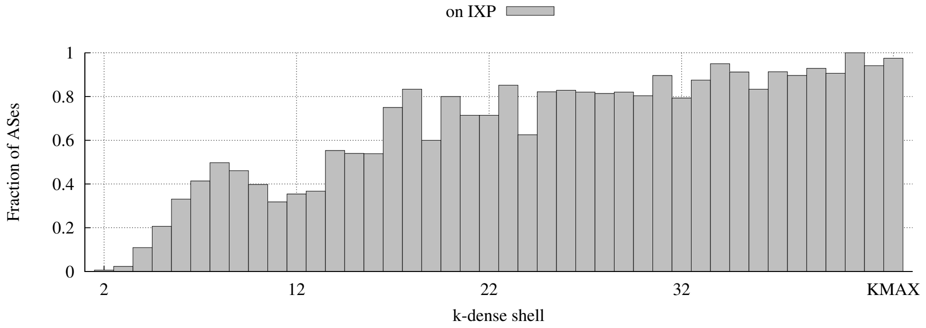


Figure 5.1: Fraction of ASes on IXP in each k-dense shell.

In order to show the relation between k-dense shells and IXPs, we plot the fraction of ASes tagged as IXP members in each k-dense shell, i.e. the number of ASes in a given shell that are members of at least one IXP divided by the total number of ASes in the k-dense shell - Figure 5.1. Low k-dense shells are mainly made up of not-on-IXP ASes, whereas IXP members have a strong presence in high k-dense shells. This indicates that the presence of well-connected zones within our Internet AS-level topology is mainly triggered by IXP member ASes.

Current literature also supports the correlation between the k_{MAX}-dense index growth and the evolution of IXPs facilities. For instance, we show in [38] and in [76] that the percentage of IXP members within each k-dense shell rapidly increases as the k-dense index increases, yet all the ASes in the k_{MAX}-dense are members of at least one IXP. Such results refer to the Internet AS-level topologies related to April 2010 and May 2012 respectively. Also [3] supports our claim describing the ground truth of a large European IXP (data refers to a measurement campaign made in 2011): authors shows that the amount of peering links established exploiting this facility is pretty huge, more in detail, 350 members are connected using 50,000 public peering links.

We find that 78 out of 80 ASes in the 2012 Internet k_{MAX}-dense have a peering record in the PeeringDB database, and declare to be members of at least one IXP. We investigate the profile of these 2 missing networks and we find that both of them are actually members of a german IXP, DE-CIX[2]. Moreover, 60 ASes of the the k_{MAX}-dense are member of the same IXP, DE-CIX.

Finally, we count the number of peering connections within the k_{MAX}-dense using the Isolario dataset. We find that 1,651 out of 2,670 connections are tagged as peering connections. The presence of peering connections is not as large as expected (about 60%). We discussed these results with [33] authors and they confirmed that this result was actually expected. Their tagging algorithm relies on the information provided by

[2] DE-CIX (Deutscher Commercial Internet Exchange) is one of the largest Internet Exchange Points worldwide. It is located in Frankfurt (DE) and it currently counts more than 450 providers connected. Please consider `http://www.de-cix.net/` for more details.

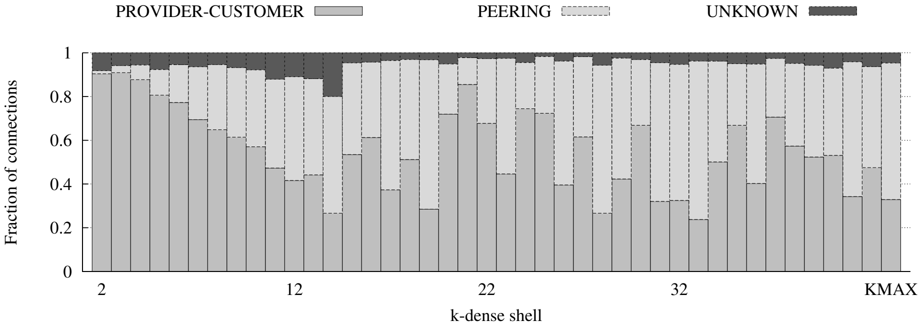


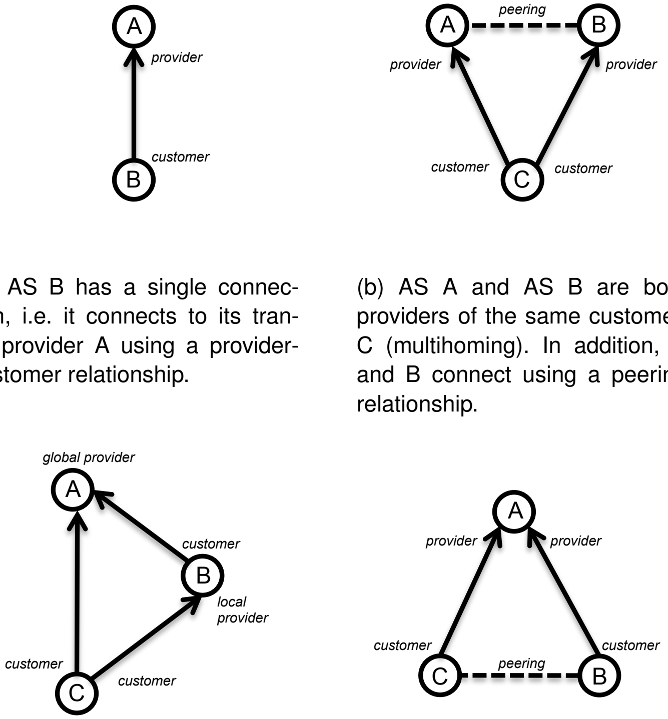
Figure 5.2: Fraction of provider-customer and peering connections involving k-dense shells.

RouteViews/RIS/PCH feeders, then since these feeders provide a view of the network as seen from large ISPs, then peering connections are unlikely to be seen. Some connections are tagged as *unknown* because the set of connections extracted from IRL, i.e. the Internet topology related to January 2012, differs from the topology used by [33, 34] to infer AS relationships.

5.2 k-dense-structure: 2- 3- dense shells

Nodes with k-dense-index equal to 2 or 3 are the vast majority of Internet ASes. They are periphery ASes that contributed the most to the network growth: they are characterized by a small degree, however they are involved in a huge number of connections [75, 18]. All the nodes with a degree equal to 1 (and all the nodes with clustering coefficient equal to 0) are part of the 2-dense shell. Figure 5.1 shows that these k-dense shells have a low percentage of IXP members.In addition, we compute the fraction of provider-customer and peering connections involving nodes in each k-dense shell - Figure 5.2, and we demonstrate that an high fraction of provider-customer connections involving the 2- and 3- denses. Finally, [38] indicates that most of these ASes have a national scope.

This structural description and this AS characterization are compliant with those new ASes that enter in the network and whose business is not Internet-driven: they set up connections in order to obtain connectivity, and all they need is a transit provider. Sometimes they set up multiple agreements, i.e. they purchase transit from more than a single provider (multihoming). If the two providers of a multihomed AS are connected then a triangle is formed and thus all three nodes are part of the 3-dense. These stub ASes do not transit traffic for other ASes and are likely to be customers in provider-customer relationships. These types of ASes are national ASes unless a continental or a worldwide presence is required by their own business. In Figure 5.3 we show the most common patterns behind the formation of 2- and 3- denses.



(a) AS B has a single connection, i.e. it connects to its transit provider A using a provider-customer relationship.

(b) AS A and AS B are both providers of the same customer, C (multihoming). In addition, A and B connect using a peering relationship.

(c) AS A and AS B are both providers of the same customer, C (multihoming). In addition, A and B connects using a provider-customer relationship. A and B are likely to be a global and a local provider respectively.

(d) AS B and AS C are both customers of the same provider, A. In addition, B and C connect using a peering relationship in order to avoid the transit cost when they exchange traffic.

Figure 5.3: Common patterns behind 2- 3- denses formation.

5.3 k-dense-structure:Tier-1s

Currently there is no broadly accepted list of Tier-1 ASes. however we can list a number of properties that characterize them. First of all, the primary goal of Tier-1 ASes is to sell transit to their customers. As a result, Tier-1 ASes are expected to have a high number of providers-customers connections directed to enterprise customers. Another consequence of their business profile is the adoption of a restrictive peering policy at the IXPs. In fact, the adoption of an open peering policy would not increase their revenues, rather it would result in peering with potential customers and thus in loosing some potential revenue. On the contrary, Tier-1 ASes are likely to peer large

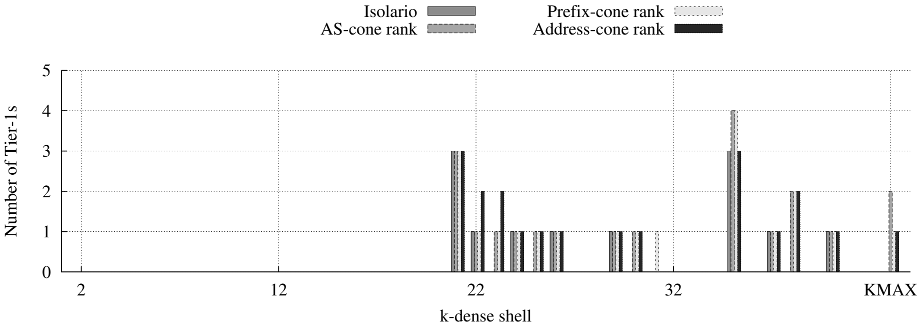


Figure 5.4: Presence of Tier-1s within k-dense shells.

Content Providers and Content Delivery Networks as they are important sources of traffic and that makes them ideal peers. Finally, Tier-1 ASes are supposed to have a very large customer cone. In fact, the ensemble of all the Tier-1 ASes' customer cones is made up of all the Internet ASes by definition.

In order to understand the relationship between Tier-1 ASes and k-denses, we considered two possible sets of Tier-1:

- the list of Tier-1 ASes adopted by Isolario [33];
- the list of 20 ASes with the largest customer cones according to [13] (AS customer cone, prefix customer cone, and address customer cone).

We plot in Figure 5.4 the number of Tier-1 ASes in each k-dense shell. Tier-1 ASes populate medium-high k-dense shells, however only few of them are part of the k_{MAX}-dense. Due to their definition, all Tier-1 ASes are required to be mutually connected, thereby forming a clique topology. Such imposed structure guarantees a non-trivial k-dense index to each Tier-1 AS, i.e. if N_T is the number of Tier-1 ASes, then a k-dense index equal to N_T is guaranteed. On the other hand, their large number of provider-customer connections create hierarchical structures that do not contribute to the creation of dense structures.

5.4 k-dense-structure: k_{MAX}-denses

The k_{MAX}-dense ASes form the densest-connected community by definition. The easiest way to establish such dense connectivity in practice is by connecting to a large IXP and declaring an open peering policy. In this Section we further investigate how IXPs contribute to the creation of these dense zones of the Internet, and which kind of ASes build these substructures.

We start with analyzing the connections involving k_{MAX}-dense ASes. Connections involving k_{MAX}-dense are directed to other k_{MAX}-dense nodes, to low k-dense shells, but also to other medium/high k_{MAX}-dense shells - Figure 4.11c. In Figure 5.5

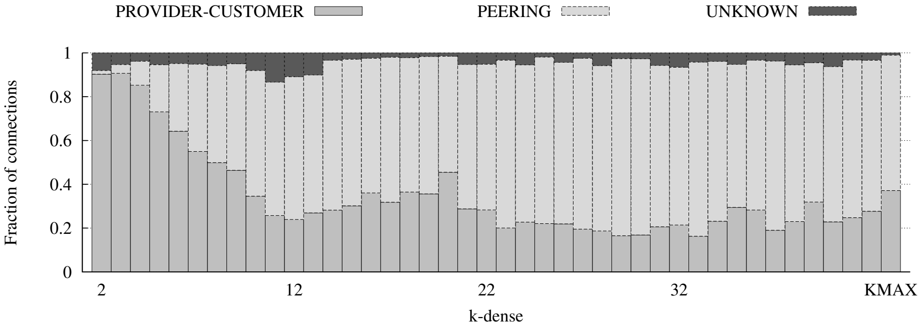


Figure 5.5: Fraction of provider-customer and peering connections in each k-dense.

we show the fraction of provider-customer and peering connections in each k-dense, and we observe that connections within the k_{MAX}-dense are mostly tagged as peering connections. The same observation applies to the fraction of peering connections involving ASes in the k_{MAX}-dense shell - Figure 5.2. Connections directed to low k-dense indexes are likely to be customer-provider relationships; connections directed to Tier-1 ASes could be customer-provider relationships as well. On the other hand, peering connections are likely to be caused by IXPs.

In order to support this last claim, we use the PeeringDB [82] dataset to extract and discuss the information related to k_{MAX}-dense ASes and IXPs. First, all the ASes within the k_{MAX}-dense are connected to at least one IXP - Section 5.1. Then, we plot in Figure 5.6 the number of IXPs that are joined by each k_{MAX}-dense AS, we find that each k_{MAX}-dense AS is connected on average to 8.35 IXPs. Also, we plot in Figure 5.7 the number of k_{MAX}-dense ASes in each IXP (we show the top-20 IXPs that have the highest number of k_{MAX}-dense members) and we observe the tendency of k_{MAX}-dense ASes are likely to connect to large IXPs. In fact, AMS-IX, DE-CIX, MSK-IX and LINX are the largest IXPs in terms of members.

Since participation at IXPs ensues from the AS business strategy, we can extend the analysis of these ASes by investigating their business profile. It is interesting to observe that ASes participating in more than a single IXPs are typically Network Service Providers (NSP). Also, according to [38] Content Delivery Networks and Content Providers are likely to connect to multiple IXPs. In Table 5.1a we show the business profile of those ASes that form the k_{MAX}-dense, and we observe that these three categories are well represented. In order to understand the rationale behind the formation of a densely-connected sub-graph by these categories, we provide a brief description of their business.

A NSP or, more in general, a network operator can be of two types: 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. Internet Backbone Providers are organizations that supply the ISPs with access to the lines that connect ISPs to each other, thereby allowing ISPs to

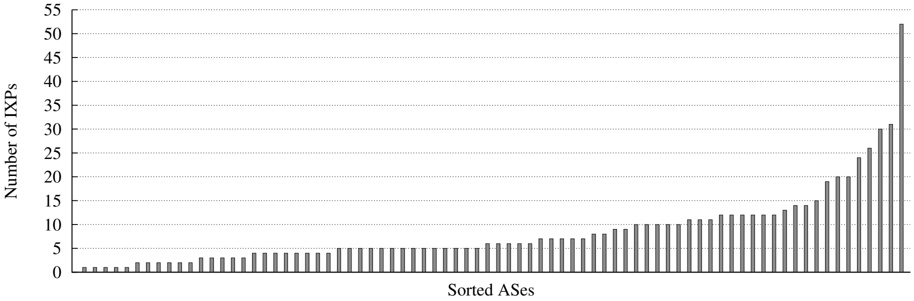


Figure 5.6: Number of joined IXP for each k_{MAX}-dense AS (sorted).

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.

On the other 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.*"[3]

Given that CPs and CDNs benefit from peering with any willing-to-peer ASes [69], it is quite plausible that CPs and CDNs are main players behind the formation of this densely interconnected substructure [38]. Surprisingly, the adoption of an open peering policy is an emerging phenomenon among Network Service Providers (NSPs) [60], i.e. tier-2 ASes provided with an own backbone network that purchase transit from an upstream provider and resell it to other ASes. Although these ASes usually adopt a selective peering policy, as they do not want to peer with potential customers, such peering connections help tier-2 ASes to provide a better end-user experience to their customers [70].

The percentage of ASes with selective peering policies (almost all NSPs) is also significant, but all these ASes are good candidates for selective peering as well, explaining their high k-density. Indeed, one commonly considered aspect in the peer selection process is the symmetry of the exchanged traffic [72]. We do not have access to traffic statistics, but we can use the number of IP addresses in the customer

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

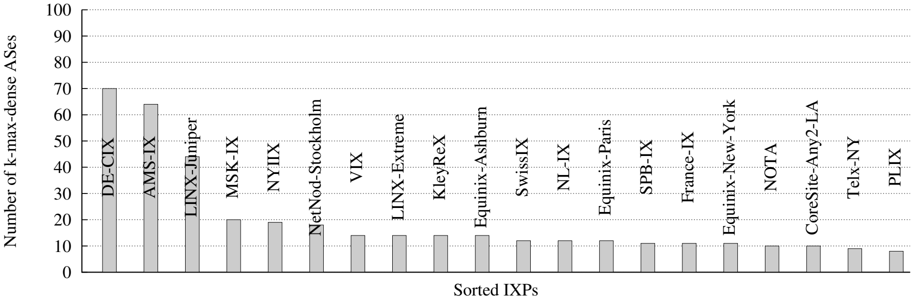


Figure 5.7: Number of k_{MAX}-dense AS in each IXP (top-20 IXPs).

cone [13] as a proxy. The customer cone of an AS is the set of ASes that can be reached from the AS following only provider-to-customer links. In other words, it is the set of destinations that can be reached *for free* upon peering with the AS. [76] shows that the size of the customer cone of ASes within the k_{MAX} is much higher than the Internet average, and it does not show a high variance. In other words, ASes within the k_{MAX} dense have similar customer cone sizes and thus setting up peering connections is a plausible strategy as it does not involve potential customers.

Finally, we use data in Table 5.1 to summarize the properties of k_{MAX}-dense ASes described so far. k_{MAX}-dense is mostly composed of network operators (NSP or Cable/DSL providers) and Content and Educational networks - Table 5.1a. The presence of such kind of ASes within the k_{MAX}-dense is likely to be caused by their membership at one or multiple IXPs. In detail, the so-called tier-2 ASes benefit from settling peering connection at IXPs because they can offer a shortest path between content providers and their customers, thereby providing a better service. Educational and research networks, on the other hand, are more interested in avoiding the transit costs. The traffic information shown Tables 5.1b and 5.1c reveals that the vast majority of ASes have either a balanced or a mostly outbound peering ratio, and it indicates that most of the ASes have a similar traffic volumes. The presence of mostly outbound traffic ASes underlies presence of content services[4]. In addition, the fact that many ASes have the same traffic volume supports the creation of many peering connections even if a selective policy is adopted. Table 5.1e shows that almost all of the k_{MAX}-dense ASes declare either an open or a selective peering policy, and that is compatible with the absence of Tier-1s. We conclude this summary with pointing out the high percentage of European ASes, whose presence is correlated with the noticeable number of ASes which are members of AMS-IX, DE-CIX and LINX - Figure 5.7.

[4] ASes in Internet can have multiple businesses, thus even if there are only 5 content providers, many NSPs could offer content services too.

5.5 $1K$-randomness of k_{MAX}-denses

A network that preserves the same number of nodes and links of the target graph does not approximate accurately the structure of the k_{MAX}-dense graphs that we considered in Chapter 4. On the other hand, the degree distribution is sufficient to describe the structure of these dense sub-graphs. Degree correlation of nodes can be interpreted studying the *process* behind the set up of peering connections. When $d = 1$ is sufficient it means that an AS is likely to connect using an *open peering policy*: it simply sets up a certain number of connections (which is described by its own degree) but it does not *choose* its neighbors (a degree correlation between nodes at greater distance is not evident). On the other hand, a $d = 2$ could be the result of a peering selection process which is typical of a *selective peering policy*: in this case the observed AS chooses its neighbors and thus chooses their degrees too. Supposing that all the k_{MAX}-dense connections take place at a single IXP[5](DE-CIX would be the main candidate) and supposing that the peering policies declared in PeeringDB represent the peering policy that each AS adopts on each IXP, then our explanation of $1K$-randomness using IXP policies seems to be fully supported by data shown in Table 5.1e.

5.6 Topology incompleteness

According to [74] results the vast majority of missing links are peering connections which involve ASes that do not host any monitor (or ASes whose downstream customers do not host any monitor). Thus we are likely to miss peering connections involving small ASes 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 the dataset we considered since it is likely to find a monitor in one of them or at least within their downstream customers. In our analysis we find that the most well-connected zone of the Internet involves ASes participating at IXPs with a pretty high Internet degree, while ASes 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 the Internet topology would provide the following changes:

- 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;
- there could be many ASes with a low k-dense index shifted to k-dense shells with a higher k-dense index, moreover this behavior 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] We do not have access to the peering matrix of IXPs, thus we cannot confirm this hypothesis. However, results presented in [3] guarantee the feasibility of our guess.

Table 5.1: PeeringDB properties related to k_{MAX}-denseASes.

(a) Business type.

Business type	AS count
Cable/DSL/ISP	12
Content	5
Educational/Research	5
NSP	56
Non-Profit	1
Unknown	1

(b) Traffic volume.

Traffic volume	AS count
Unknown	1
Not disclosed	7
0-1000 Mbps	2
1 - 100 Gbps	43
100 - 1000 Gbps	25
1 Tbps+	2

(c) Traffic ratio.

Traffic ratio	AS count
Heavy Inbound	1
Mostly Inbound	9
Balanced	44
Mostly Outbound	23
Heavy Outbound	2
Unknown	1

(d) Geographic type.

Geographic scope	AS count
Europe	38
Global	19
Regional	22
Unknown	1

(e) Peering policy.

Peering policy	AS count
Open	40
Restrictive	1
Selective	38
Unknown	1

6

Conclusions

A deep understanding of the underlying structure of the Internet AS-level topology, as well as its evolution, helps the development of new models, which in turn, are vital for testing new protocols and applications. Also, a better knowledge of the Internet structure can provide insights into the design of new routing protocols, it can ease the evaluation of the consequences of node failures, and it can facilitate the development of more efficient algorithms for searching and flow optimization.

In this work we carried out a detailed analysis of the structure of the Internet topology at the AS-level of abstraction and its evolution over the last 9 years. In order to understand how this evolution process works, we divided the problem into four sub-problems:

- first, due to the high heterogeneity of the network components, it is important to identify a method to represent the network such that it enables both to efficiently perform a structural analysis and to understand the phenomena that drive its evolution;
- second, once the Internet representation is formalized, the sub-structures of interest have to be defined, or in other words it is required to have a structural description that provides insights into the underlying organization of the network;
- third, the description of the topology evolution requires the computation of the properties of interest on different snapshots (of the network) collected in different dates, then these information have to be aggregated and compared in order to reveal the underlying trends;
- finally, since the structural changes affecting the topology are driven by the independent decisions of each AS, it is interesting to understand the strategies that ASes use to optimize their objective functions and how these are correlated with the topology structure.

A summary of the main contributions of this thesis follows.

Network description

Internet AS-level topology has been often described as an ecosystem due to the variety of its building blocks, i.e. ASes, as well as the complex dynamics that drive its evolution. Although the information available to research is limited and incomplete because of its confidentiality, it is possible to collect data describing the Internet topology over time, the business relationship between ASes, the size of ASes customer cones and the presence of an AS at an IXP. In this thesis we tackled the problem of network description using the following approach: first we focused on the sole topology and we computed the structural properties without adding any context information; then, we added the details related to inferred AS-relationships and regarding IXPs. This strategy enabled to easily compute the structural properties of the network, also it provided a way to further study the sub-structures emerging from the structural analysis, thereby helping us in understanding the drivers behind Internet structure evolution.

Meaningful substructures

In this thesis we analyzed the Internet AS-level structure using the k-dense method. Since there is no broadly accepted definition of community, we first described the main criteria that should guide the detection of a community within the Internet environment, i.e.: basically we were looking for sub-graphs with a high internal density without any concern about external links. Furthermore, we selected three deterministic community detection algorithms that satisfied such criteria, i.e. k-core decomposition, k-dense method, and clique percolation method, and we thoroughly compared them. In detail, we studied the result of the community detection algorithms on a sample topology and we observed the different community organizations using an innovative visualization tool, i.e. k-tree. In addition we determined their statistical significance using the dK-analysis. To the best of our knowledge, this is the first time the dK-series are computed to analyze the *complexity* of a community detection method. We found that neither $2K$-series nor dK-series with $d < 2$ are able to reproduce the output of these community detection algorithms meaning that they are statistical interesting. k-dense method resulted the best tool to use since the obtained structures were not as fragile as k-clique communities (i.e. communities obtained using the clique percolation method), whereas their communities suggested a stronger positive relationship between ASes if compared to the k-core decomposition.

Network evolution

Our analysis of the Internet AS-level topology by means of the k-dense method was inspired by the fact that, although the common properties, like number of nodes or the average degree, provide a general vision of how the network evolves, they do not give insight into the structure change. We extracted the k-dense communities from each snapshot of the network from 2004 to 2012 and we found that k_{MAX}-dense index, i.e. a proxy measure of the maximum level of density achieved within a graph, is constantly growing, thereby revealing the presence of denser sub-graphs as the graph evolved. After proper normalizations we found that the distribution of ASes in each k-dense shell and the distribution of connections involving each k-dense shell are time-invariant. We also found that all the snapshots share the following property: 2-, 3-, and k_{MAX}- dense shells are the classes of ASes involved in the maximum number of connections. 2-, 3- dense shells represent a noticeable portion of the entire set of ASes in Internet and that is a sufficient to justify their involvement in such a high number of connections. Surprisingly, k_{MAX}-dense ASes represent (on average) just 0.2% of the total number of Internet ASes and are involved in more than 25% of the Internet connections. Such information revealed that k_{MAX}-dense ASes have a main role in the Internet connectivity. Due to its importance we investigated the internal structure of the k_{MAX}-dense using the dK-analysis to infer the statistical properties of such sub-graph. Although these networks are highly dense (on average they have a link density equal to 0.84, meaning that they resemble a complete topology) we proved that these graphs are $1K$-random, yet this property is shared between all the snapshots. In other words, a synthetically generated graph with the same number of nodes and links does not accurately approximate the structure of the k_{MAX}-dense. On the other hand, a graph with the same distribution approximates correctly the k_{MAX}-dense. This kind of information might be very useful to test if a synthetic graph generator is able to generate a k_{MAX}-dense with the same characteristics. Finally, we analyzed the organization of the current Internet structure, i.e. the snapshot related to 2012, providing further details such as comparing the number of links within the communities and the number of connections directed outside the communities, providing the number of nodes and connections in each k-dense, showing the the number of connections involving each k-dense shell and finally providing a thorough description of the how ASes in the 2-, 3-, and k_{MAX}- dense shells direct their connections.

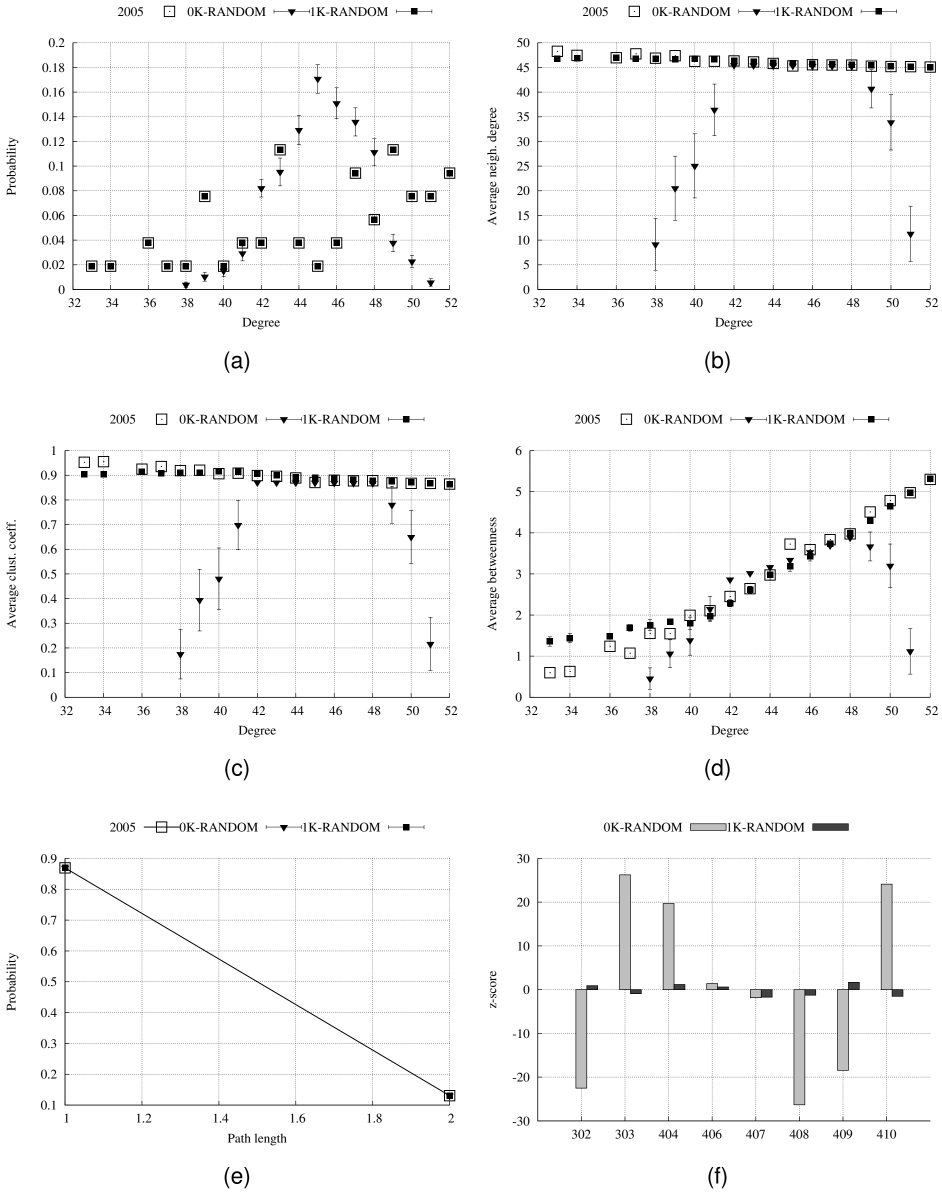
Internet drivers

By combining the results of the structural analysis and the information related to the inferred relationships between ASes and the set of ASes participating at IXPs, we were able to unveil some of the driving forces behind Internet structure changes. We found that the growth of the k_{MAX}- dense index, or simply the creation of denser and denser sub-graphs, is mostly due to the proliferation of public peering connections, i.e. peering connections exploiting IXPs facilities. On the other hand, we observed that ASes within the low k-dense shells, or periphery ASes, are the main responsible for the network growth. In details, small enterprise customers whose business is not Internet-driven with one or more transit providers are the reason behind the presence of so many 2- and 3- dense shells ASes. Another property that we investigated was the presence of Tier-1 ASes within the k-dense shells. This kind of ASes have a primary role in Internet routing as they are defined as the set of ASes that can reach all the Internet prefixes through their customers and peers. By definition such group has to form a complete topology, thus a minimum k-dense index is automatically guaranteed; however, they hardly are part of the k_{MAX}-dense. An interpretation of this phenomenon is that even if they are IXP members, they use to have a restrictive peering policy because all ASes are potential customers of a Tier-1, thus Tier-1 ASes are not likely to form extremely meshed topologies. Finally we investigated the set of ASes within the k_{MAX}-dense and we provided an explanation of its $1K$-randomness. k_{MAX}-dense are mostly Network Service Providers, Content Providers, or Content Delivery Networks, also all of them participate at least to an IXP. CPs and CDPs usually adopt an open peering policy , indeed the more connections they open, the better service they can provide. Also, they do not sell transit, thus they do not have connections potential customers to avoid, whereas a peering connection can let them save the cost of transit. On the other hand, NSPs usually adopt a selective peering policy. A peering connection to a high source of traffic can be vital for an NSP in order to provide a better service to its customers, however it cannot peer indistinctly, as its main business objective is to sell transit. As a result, an NSP usually avoids peering with potential customers, thereby preferring a selective peering policy over an open peering policy. Finally, we found that this combination of open and selective peering policies characterizing the k_{MAX}-dense provides an explanation of its $1K$-randomness.

A

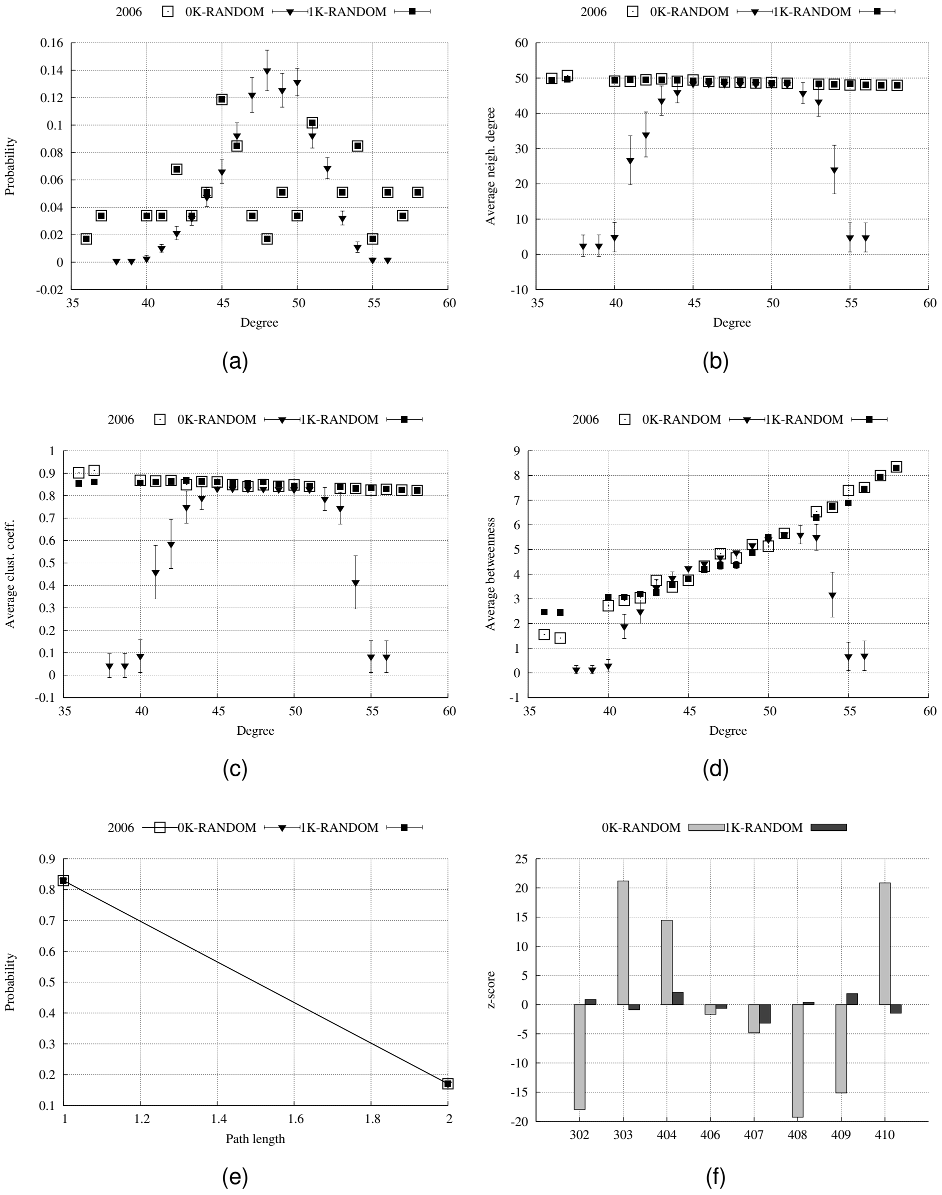
dK-statistical analysis of k_{MAX}-denses

In this Appendix we show the results of the dK-analysis related to 2005, 2006, 2007, 2009, 2010, and 2011 snapshots. Figures A.1, A.2, A.3, A.4,A.5, and A.6 provide the same outcome found for 2004, 2008, and 2012 snapshots, i.e. the k_{MAX}-dense sub-graph is a $1K$-random network.



(a) (b) (c) (d) (e) (f)

Figure A.1: **2005 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.1a degree distribution, A.1b average neighbor degree over degree, c average clustering coefficient over degree, A.1d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

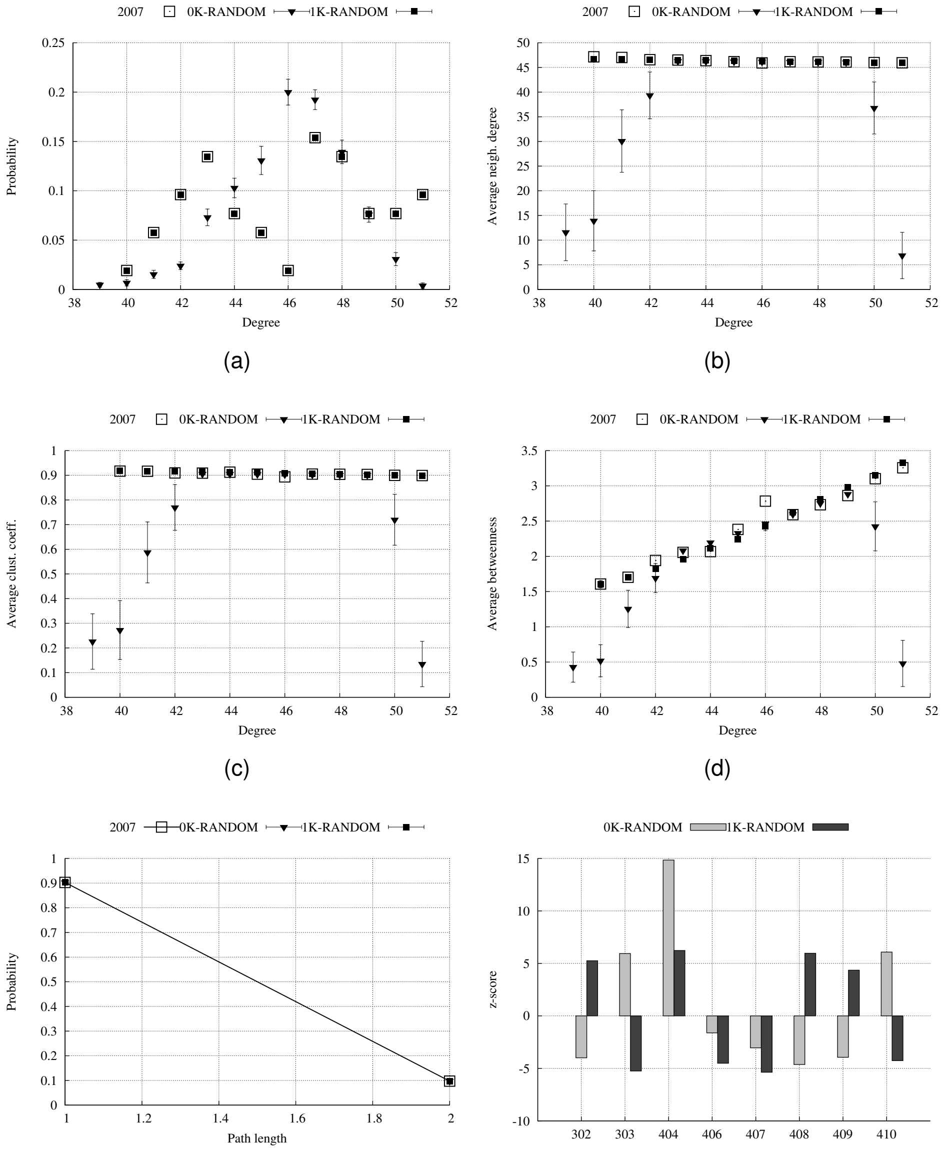
(c)

(d)

(e)

(f)

Figure A.2: **2006 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.2a degree distribution, A.2b average neighbor degree over degree, c average clustering coefficient over degree, A.2d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

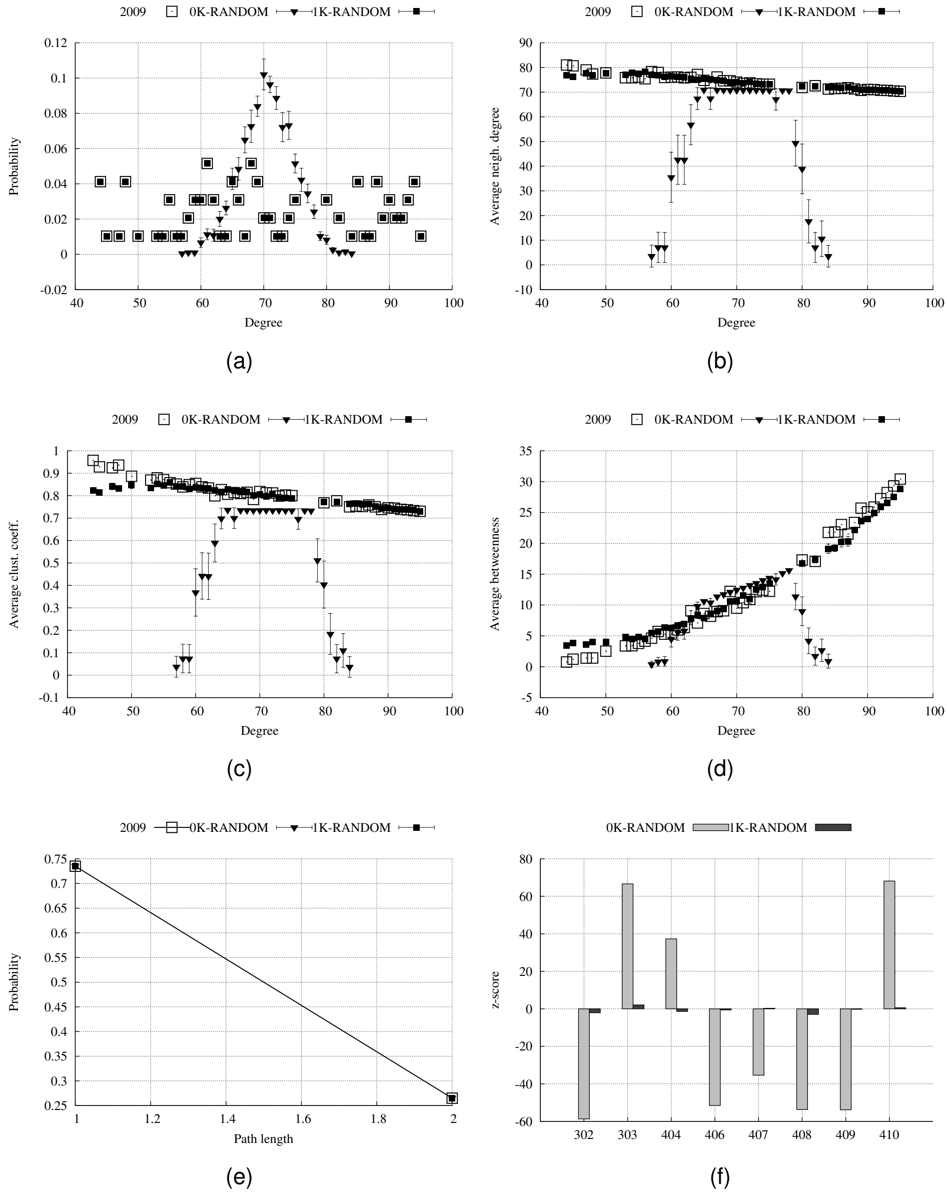
(c)

(d)

(e)

(f)

Figure A.3: **2007 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.3a degree distribution, A.3b average neighbor degree over degree, c average clustering coefficient over degree, A.3d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

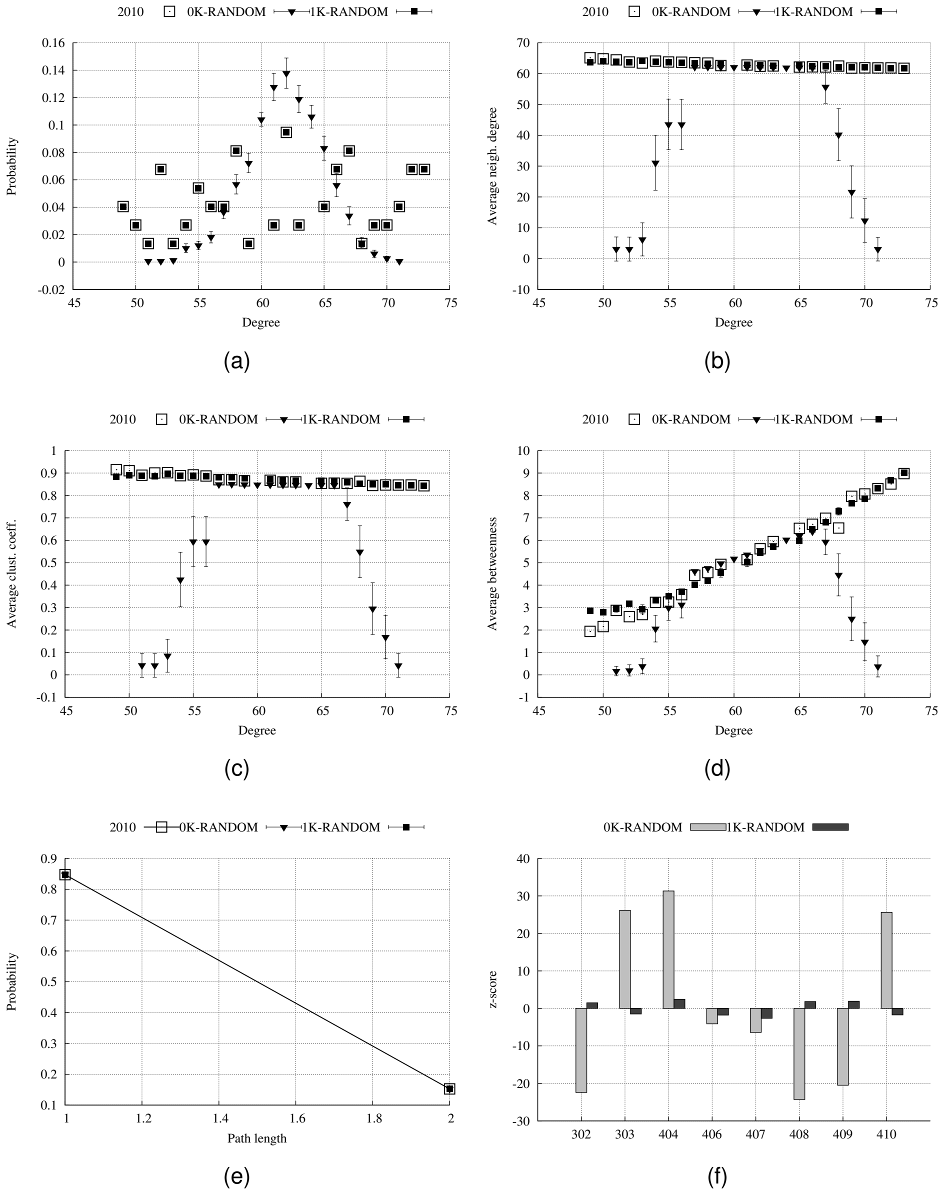
(c)

(d)

(e)

(f)

Figure A.4: **2009 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.4a degree distribution, A.4b average neighbor degree over degree, c average clustering coefficient over degree, A.4d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

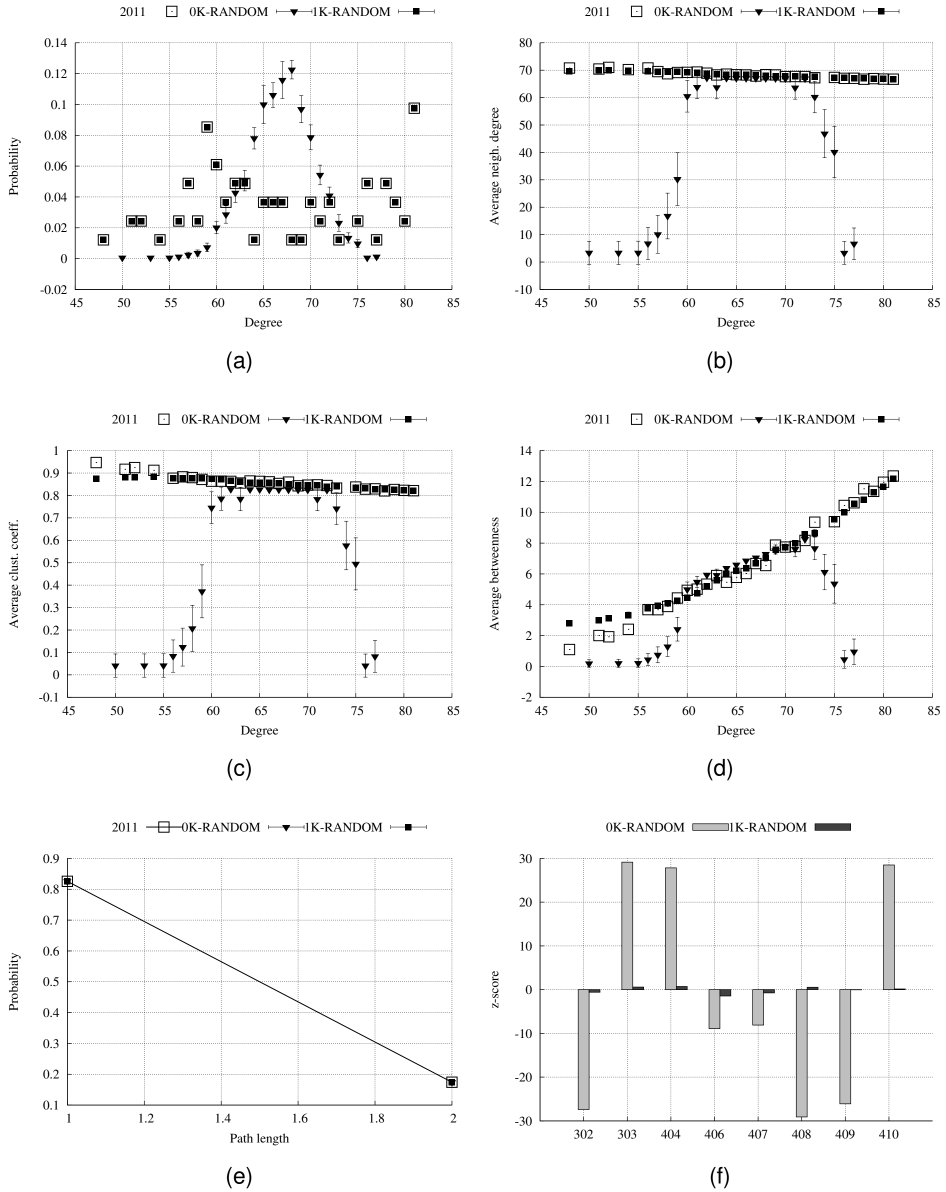
(c)

(d)

(e)

(f)

Figure A.5: **2010 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.5a degree distribution, A.5b average neighbor degree over degree, c average clustering coefficient over degree, A.5d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.



(a)

(b)

(c)

(d)

(e)

(f)

Figure A.6: **2011 k_{MAX}-dense** properties vs. 0k-random and 1k-random graphs properties: A.6a degree distribution, A.6b average neighbor degree over degree, c average clustering coefficient over degree, A.6d average betweenness over degree, e average shortest path distribution, f z-score of motifs of size 3 and 4. Figures report, for each property, the average value and the confidence interval with probability 0.8.

References

1. Distributed Internet MEasurements and Simulations dataset. http://www.netdimes.org/, 2010.
2. Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, Sep 2001.
3. Bernhard Ager, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. Anatomy of a large european IXP. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '12, pages 163–174, New York, NY, USA, 2012. ACM.
4. Alexander Tsiatas. *Diffusion and Clustering on Large Graphs*. PhD thesis, University of California, San Diego, 2012.
5. J. Ignacio Alvarez-Hamelin, Mariano G. Beiro, and Jorge R. Busch. Understanding Edge Connectivity in the Internet through Core Decomposition. *Internet Mathematics*, 7(1):45–66, 2011.
6. J. Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases. *Networks and Heterogeneous Media*, 3(2):371–293, 2008.
7. Sara Amr, Mohammed El-Betagy, and Mohammed Helmi. Analyzing Internet Connectivity Data Using Modified k-shell Analysis. In *INFOS 2008*, 2008.
8. Brice Augustin, Balachander Krishnamurthy, and Walter Willinger. IXPs: mapped? In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 336–349, New York, NY, USA, 2009. ACM.
9. Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.
10. Tony Bates, Philip Smith, and Geoff Huston. CIDR report. http://www.cidr-report.org/as2.0/, 2013.
11. G.J. Baxter, S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes. k-core organization in complex networks. In My T. Thai and Panos M. Pardalos, editors, *Handbook of Optimization in Complex Networks*, Springer Optimization and Its Applications, pages 229–252. Springer US, 2012.
12. Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, July 2008.
13. CAIDA, the Cooperative Association for the Internet Data Analysis. AS Rank. http://as-rank.caida.org/, 2012.
14. CAIDA, the Cooperative Association for the Internet Data Analysis. The CAIDA AS Relationships Dataset, <2012-06-01>. http://www.caida.org/data/active/as-relationships/, 2012.

15. Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of Internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, July 2007.
16. L. da F. Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. C. Rocha. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *Advances in Physics*, 60(3):329–412, 2011.
17. Wenping Deng, Wolfgang Mühlbauer, Yuexiang Yang, Peidong Zhu, Xicheng Lu, and Bernhard Plattner. Shedding light on the use of AS relationships for path inference. *Journal of Communications and Networks*, 14(3):336–345, 2012.
18. A. Dhamdhere and C. Dovrolis. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Transactions on Networking*, 19(5):1420–1433, Sep. 2011.
19. X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, k. claffy, and G. Riley. AS Relationships: Inference and Validation. *ACM/SIGCOMM Computer Communication Review*, 37(1):29–40, Jan 2007.
20. Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):027104+, Aug 2005.
21. Benjamin Edwards, Steven A. Hofmeyr, George Stelle, and Stephanie Forrest. Internet Topology over Time. In *Proceedings of the 8th CS UNM Student Conference*, Apr. 2012.
22. P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.
23. Alex Fabrikant, Elias Koutsoupias, and Christos H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ICALP '02, pages 110–122, London, UK, UK, 2002. Springer-Verlag.
24. Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. *ACM/SIGCOMM Computer Communication Review*, 29(4):251–262, Aug. 1999.
25. Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-Organization and Identification of Web Communities. *Computer*, 35(3):66–71, 2002.
26. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.
27. Lixin Gao. On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on Networking*, 9:733–745, 2000.
28. Lixin Gao and Feng Wang. The extent of AS path inflation by routing policies. In *IEEE Global Internet Symposium*, 2002.
29. Cyril Gavoille. Routing in Distributed Networks: Overview and Open Problems. *ACM SIGACT News - Distributed Computing Column*, 32:36–52, 2001.
30. Phillipa Gill, Michael Schapira, and Sharon Goldberg. Modeling on quicksand: dealing with the scarcity of ground truth in interdomain routing data. *ACM/SIGCOMM Computer Communication Review*, 42(1):40–46, 2012.
31. M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.
32. Enrico Gregori, Alessandro Improta, Luciano Lenzini, and Chiara Orsini. The impact of IXPs on the AS-level topology structure of the Internet. *Computer Communications*, 34(1):68 – 82, 2011.
33. Enrico Gregori, Alessandro Improta, Luciano Lenzini, Lorenzo Rossi, and Luca Sani. BGP and inter-AS economic relationships. In *Proceedings of the 10th international IFIP TC 6 conference on Networking - Volume Part II*, NETWORKING'11, pages 54–67, Berlin, Heidelberg, 2011. Springer-Verlag.

34. Enrico Gregori, Alessandro Improta, Luciano Lenzini, Lorenzo Rossi, and Luca Sani. An Enhancement for Networking '11 Tagging Algorithm. Technical report, IIT/CNR, 2012.
35. Enrico Gregori, Alessandro Improta, Luciano Lenzini, Lorenzo Rossi, and Luca Sani. On the Incompleteness of the AS-level Graph: a Novel Methodology for BGP Route Collector Placement. In *Proceedings of IMC 2012*, 2012.
36. Enrico Gregori, Luciano Lenzini, Simone Mainardi, and Chiara Orsini. FLIP-CPM: A Parallel Community Detection Method. In *ISCIS 2011*, 2011.
37. Enrico Gregori, Luciano Lenzini, and Chiara Orsini. k-clique Communities in the Internet AS-level Topology Graph. Technical report, 2010.
38. Enrico Gregori, Luciano Lenzini, and Chiara 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.
39. Enrico Gregori, Luciano Lenzini, and Chiara Orsini. k-clique communities and economic relationships in the internet as-level topology graph. In *Inf-Q: Informatica Quantitiva*, 2011.
40. Enrico Gregori, Luciano Lenzini, and Chiara Orsini. k-clique Communities in the Internet AS-Level Topology Graph. In *SIMPLEX 2011*, 2011.
41. Enrico Gregori, Luciano Lenzini, and Chiara Orsini. k-dense communities in the internet as-level topology graph. *Computer Networks*, 2012. (In press).
42. S. Hasan and S. Gorinsky. Obscure Giants: Detecting the Provider-Free ASes. In *Proceedings of IFIP Networking 2012*, 2012.
43. J. Hawkinson and T. Bates. Guidelines for creation, selection, and registration of an Autonomous System (AS) - RFC 1930. `http://tools.ietf.org/html/rfc1930`, 2013.
44. Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy. Lord of the Links: a Framework for Discovering Missing Links in the Internet Topology. *IEEE/ACM Transactions on Networking*, 17(2):391–404, 2009.
45. J. Herzen, C. Westphal, and P. Thiran. Scalable routing easy as PIE: A practical isometric embedding protocol. In *Network Protocols (ICNP), 2011 19th IEEE International Conference on*, pages 49 –58, Oct. 2011.
46. Bradley Huffaker, Dan Andersen, Emile Aben, Matthew Luckie, kc claffy, and Colleen Shannon. The Skitter AS Links Dataset, <2004-01-01 - 2008-01-01>. `http://www.caida.org/data/active/skitter_aslinks_dataset.xml`, 2013.
47. Young Hyun, Bradley Huffaker, Dan Andersen, Emile Aben, Matthew Luckie, kc claffy, and Colleen Shannon. The IPv4 Routed /24 AS Links Dataset, <2008-01-01 - 2012-01-01>. `http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml`, 2013.
48. Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Marián Boguñá, and Dmitri Krioukov. How small are building blocks of complex networks. *CoRR*, abs/0908.1143, 2009.
49. K. Keys, Y. Hyun, M. Luckie, and k. claffy. Internet-Scale IPv4 Alias Resolution with MIDAR. *IEEE/ACM Transactions on Networking*, 2012.
50. H. Kim, C. I. Del Genio, K. E. Bassler, and Z. Toroczkai. Constructing and sampling directed graphs with given degree sequences. *New Journal of Physics*, 14(2):023012, 2012.
51. M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identification of influential spreaders in complex networks . *Nature Physics*, 6(11):888–893, Aug 2010.
52. G. Korniss, R. Huang, S. Sreenivasan, and B. K. Szymanski. Optimizing Synchronization, Flow, and Robustness in Weighted Complex Networks. In *Handbook of Optimization in Complex Networks*, volume 58 of *Springer Optimization and Its Applications*, pages 61–96. Springer New York, 2012.
53. D. Krioukov, M. Kitsak, R. Sinkovits, D. Rideout, D. Meyer, and Marián Boguñá. Network Cosmology. *Nature Scientific Reports*, 2(793), Nov. 2012.

54. Bill Krogfoss, Marcus Weldon, and Lev Sofman. Internet Architecture Evolution and the Complex Economies of Content Peering. *Bell Lab. Tech. J.*, 17(1):163–184, Jun. 2012.
55. Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. Internet inter-domain traffic. *ACM/SIGCOMM Computer Communication Review*, 41(4):–, Aug. 2010.
56. Andrea Lancichinetti, Mikko Kivela, Jari Saramaki, and Santo Fortunato. Characterizing the community structure of complex networks. *PloS One 5(8)*, 2010.
57. Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. In *Workshop on Social Network Mining and Analysis*, 2010.
58. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.
59. Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *CoRR*, abs/0810.1355, 2008.
60. Aemen Lodhi, Amogh Dhamdhere, and Constantine Dovrolis. Peering strategy adoption by transit providers in the internet: a game theoretic approach? *ACM/SIGMETRICS Performance Evaluation Review*, 40(2):38–41, Oct. 2012.
61. Huaiyuan Ma, B.E. Helvik, and O.J. Wittner. An impact of addressing schemes on routing scalability. *Communications and Networks, Journal of*, 13(6):602 –611, Dec. 2011.
62. Huaiyuan Ma, Bjarne E. Helvik, and Otto J. Wittner. The Stability of Compact Routing in Dynamic Inter-Domain Networks. In *Communication Theory, Reliability, and Quality of Service (CTRQ), 2010 Third International Conference on*, pages 61 –66, Jun. 2010.
63. Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. *ACM/SIGCOMM Computer Communication Review*, 36(4):135–146, Aug. 2006.
64. Daniele Miorandi and Francesco De Pellegrini. K-Shell Decomposition for Dynamic Complex Networks. In *WiOpt'10: Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pages 499–507, Avignon, France, 2010.
65. M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.
66. M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330, 2004.
67. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104+, 2006.
68. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, February 2004.
69. William B. Norton. Content Providers (aka Content Publishers). `http://drpeering.net/white-papers/Ecosystems/Content-Providers.html`, 2011.
70. William B. Norton. *The Internet Peering Playbook: Connecting to the Core of the Internet*, chapter The Top Five Motivations to Peer. DrPeering Press, 2011.
71. WIlliam B. Norton. Internet Service Providers and Peering v3.0. `http://drpeering.net/white-papers/Internet-Service-Providers-And-Peering.html`, 2012.
72. William B. Norton. The Folly of Peering Ratios (as a Peering Candidate Discriminator). `http://drpeering.net/white-papers/The-Folly-Of-Peering-Ratios.html`, 2012.
73. Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. Quantifying the Completeness of the Observed Internet AS-level Structure. Technical Report 080026, UCLA, September 2008.
74. Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. The (in)completeness of the Observed Internet AS-level Structure. *IEEE/ACM Transactions on Networking*, 18(1):109–122, Feb. 2010.

75. Ricardo V. Oliveira, Beichuan Zhang, and Lixia Zhang. Observing the Evolution of Internet AS Topology. In *Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '07, pages 313–324, New York, NY, USA, 2007. ACM.
76. Chiara Orsini, Enrico Gregori, Luciano Lenzini, and Dmitri Krioukov. Evolution of the internet k-dense structure. *IEEE/ACM Transactions on Networking*, 2013. (submitted).
77. Jure ovec, Kevin J. Lang, and Michael W. Mahoney. Empirical Comparison of Algorithms for Network Community Detection. In *WWW2010: ACM WWW International Conference on World Wide Web*, 2010.
78. Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
79. F. Papadopoulos, M. Kitsak, M. Serrano, Marián Boguñá, and D. Krioukov. Popularity versus Similarity in Growing Networks. *Nature*, 489:537–540, Sep. 2012.
80. Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA, 2004.
81. P. Pedroso, D. Papadimitriou, and D. Careglio. Dynamic Compact Multicast Routing on Power-Law Graphs. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1 –6, Dec. 2011.
82. PeeringDB. PeeringDB. `http://www.peeringdb.com/`, 2012.
83. Kazumi Saito, Takeshi Yamada, and Kazuhiro Kazama. Extracting Communities from Complex Networks by the k-Dense Method. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E91-A(11):3304–3311, 2008.
84. Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.
85. M. Angeles Serrano and Marián Boguñá. Clustering in complex networks. I. General formalism. *Physical Review E*, 74, 2006.
86. S. Shakkottai, M. Fomenkov, R. Koga, D. Krioukov, and k. claffy. Evolution of the Internet AS-Level Ecosystem. In *International Conference on Complex Sciences (Complex)*, Shanghai, China, Feb. 2009. International Conference on Complex Sciences (Complex).
87. Georgos Siganos, Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. Power Laws and the AS-level Internet Topology. *IEEE/ACM Transactions on Networking*, 11(4):514–524, Aug. 2003.
88. S.D. Strowes, G. Mooney, and C. Perkins. Compact routing on the Internet AS-graph. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 852 –857, Apr. 2011.
89. L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. volume 2, pages 618–627 vol.2, 2002.
90. S. L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the Internet topology. In *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, volume 3, pages 1667–1671 vol.3, 2001.
91. Mikkel Thorup and Uri Zwick. Compact routing schemes. In *SPAA '01: Proceedings of the thirteenth annual ACM symposium on Parallel algorithms and architectures*, pages 1–10, 2001.
92. Mark Tinka. PeeringDB 'I&' The Role of Peering Coordinators. In *The 3rd African Peering and Interconnection Forum (AfPIF)*, Aug. 2012.
93. UCLA Computer Science Department's Internet Research Lab. Internet Topology Collection. `http://irl.cs.ucla.edu/topology/`, 2013.
94. Jiajing Wu, Chi K. Tse, Francis C.M. Lau, and Ivan W.H. Ho. Complex network approach to communication network performance analysis. In *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, pages 1632 –1635, may 2012.

95. Jierui Xie, Stephen Kelley, and B. K. Szymanski. Overlapping Community Detection in Networks: the State of the Art and Comparative Study. *ACM Computing Surveys*, 45(4), 2013.
96. Gang Yan, Tao Zhou, Bo Hu, Zhong-Qian Fu, and Bing-Hong Wang. Efficient routing on complex networks. *Phys. Rev. E*, 73:046108, Apr 2006.
97. Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 3:1–3:8, New York, NY, USA, 2012. ACM.
98. Beichuan Zhang, Raymond Liu, Daniel Massey, and Lixia Zhang. Collecting the Internet AS-level Topology. *ACM SIGCOMM Computer Communication Review (CCR) special issue on Internet Vital Statistics*, Jan. 2005.
99. Guo-Qing Zhang, Di Wang, and Guo-Jie Li. Enhancing the transmission efficiency by edge deletion in scale-free networks. *Phys. Rev. E*, 76:017101, Jul 2007.
100. Guo-Qing Zhang, Guo-Qiang Zhang, Qing-Feng Yang, Su-Qi Cheng, and Tao Zhou. Evolution of the Internet and its cores. *New Journal of Physics*, 10(12):123027, 2008.
101. Guoqiang Zhang, Bruno Quoitin, and Shi Zhou. Phase changes in the evolution of the IPv4 and IPv6 AS-Level Internet topologies. *Computer Communications*, 34(5):649–657, Apr. 2011.
102. Haohua Zhang, Hai Zhao, Wei Cai, Jie Liu, and Wanlei Zhou. Using the k-core decomposition to analyze the static structure of large-scale software systems. *The Journal of Supercomputing*, 53:352–369, 2010.
103. V. Zlatić, D. Garlaschelli, and G. Caldarelli. Networks with arbitrary edge multiplicities. *EPL (Europhysics Letters)*, 97(2):28005–p1–p5, 2012.