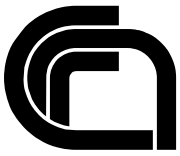


Consiglio Nazionale delle Ricerche

k-clique Communities in the Internet AS-level Topology Graph

E. Gregori, L. Lenzini, C. Orsini

IIT TR-32/2010

Technical report

Dicembre 2010

Istituto di Informatica e Telematica

k-clique Communities in the Internet AS-level Topology Graph

Enrico Gregori[1] Luciano Lenzini[2] Chiara Orsini[3]

December 9, 2010

[1]Institute of Informatics and Telematics, Italian National Research Council, Pisa, Italy. E-mail: enrico.gregori@iit.cnr.it

[2]Department of Information Engineering,University of Pisa, Pisa, Italy. E-mail: l.lenzini@iet.unipi.it

[3]Institute of Informatics and Telematics, Italian National Research Council, Pisa, Italy and Department of Information Engineering,University of Pisa, Pisa, Italy. E-mail: chiara.orsini@iet.unipi.it

Abstract

A significant challenge for researchers analysing the Internet AS-level topology graph is how to interpret the global organization of the graph as the coexistence of its structural blocks (communities) associated with more highly interconnected parts. 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 at the various levels of abstractions.

We believe that by discovering and interpreting a priori these unknown building blocks (i.e. communities), this will then pave the way for new types of analysis which are crucial in understanding of the structural and functional properties of the Internet at least at the AS level of abstraction.

We thus propose a novel type of analysis of the Internet AS-level topology graph by exploiting the k-clique community definition. First, we show that detected communities can be described by a tree representation. Then we show the presence of two classes of k-clique communities: those that are strictly affected by the nesting process which is embedded in the k-clique community definition, and, on the other hand, those that appear as branches in the tree. We conclude our analysis by highlighting the properties that characterize k-clique communities with different k values by exploiting both geographical data and information related to IXPs.

Chapter 1

Introduction and Related Work

The identification of communities within complex networks is an interesting methodology which provides an insight into the structural characteristics of the overall network. Knowledge of community structures can help to reveal the functional organization in networks [18]. In addition, the interactions of many components and the topological properties fundamentally affect the dynamics of the network [27]. Thus, the study of the structural and functional properties of complex networks through the identification of their community structure is a hot research topic. Although there is no broadly accepted definition of a community, many methods have been proposed to reveal the community structure of complex networks. According to [27] there are two categories of community detection methods: those that provide a partition of the network and those that provide a cover of the network. The main difference between these two techniques is that the former category does not allow communities to overlap, while the latter does. Regarding the Internet topology graph at the Autonomous System (AS) level of abstraction, we are interested in exploiting the second category of community detection methods since we believe that Internet AS-level communities should satisfy the following properties: communities should identify dense subgraphs of the graph indicating that each community AS is really interested in connecting to other community ASes, in addition overlapping communities should be allowed (e.g. consider for instance worldwide ASes or ASes that take part in many IXPs). There are many studies on the structural properties of the Internet AS-level topology graph that partition the network into communities. For instance, k-core decomposition [26] and k-dense methods [25] have been used in [6], [3],[10] and [12]. In contrast, there are several works that present the structural properties of the network by adopting covers, for instance, [23], [18], [27], [21] and [16]. See [9] and [17] for a detailed survey on community detection algorithms.

The Clique Percolation Method (CPM) [23], the Greedy Clique Expansion (GCE) algorithm [18] and the EAGLE (agglomerativE hierarchicAl clusterinG based on maximaL cliquE) algorithm [27] detect communities within complex networks starting from maximal-cliques, hence they are likely to identify very dense zones of the overall graph. However, we decided to consider the k-clique community definition and to avoid using GCE and EAGLE algorithms for the following reasons: the community fitness function used in GCE is not compliant with an Internet AS-level environment since it searches for sub-graphs where nodes have more internal connections than external connections, EAGLE neglects all maximal cliques with a k smaller than a threshold, hence it discards small cliques which could otherwise represent *local* communities within the Internet AS-level graph. In addition, the authors themselves admit that EAGLE is more time-consuming than the *k-clique* algorithm.

In this paper we study the structural properties of the Internet AS-level topology by using the k-clique communities defined in [23]. This approach enables us to detect overlapping communities with a high internal density. Although communities are often thought of as a set of nodes that has more connections between its members than with the rest of the network [19], we believe that this feature is not required in the Internet AS-level environment. Consider, for instance, a group of regional transit providers who are really interested in connecting to each other in order for the traffic to remain localized and to prevent traffic from unnecessarily traversing other transit networks. This set of 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. In order to illustrate this concept further, let us look at another example. Consider the set of Tier-1 ASes. By definition this set is made up of ASes that can reach every other network on the Internet without purchasing IP transit, hence they have to form a full-mesh topology and can be regarded as an Internet AS-level community. On the other hand, Tier-1 ASes are characterized by a huge number of connections (e.g. thousands of connections) that are directed to their customers. Thus, like the previous example, a community detection method based on the-more-connections-between-members property [19] would not provide for this kind of community. k-clique communities are not affected by this issue. In terms of an analysis of the community structure of the Internet AS-level topology we considered the following papers: [21] and [16]. [16] uses inconsistent community detection methods as its starting point, such as [5], [7] and [28], and provides a solution for improving consistency without compromising the modularity. On the other hand [21] analyses 12 real networks by adopting the methodology described in [16]. [16] proposes an interesting method to interpret communities extracted from the Internet AS-level topology. Specifically, by using the RIPE Data Search

tool[1] it shows that there are communities made up of ASes which are geographically close, and it also uses the AS type dataset of [29]. [21] provides a more detailed analysis of Internet AS-level communities by studying communities extracted from [8] and [22] using the z-P analysis [13]. In our analysis of Internet k-clique communities we avoided using methods such as [13], since they rely on threshold based on heuristics. Nevertheless, we interpret the detected communities by exploiting both the geographical and the IXP datasets (see Chapter 2 for more details) as done in [10] and [12].

To the best of our knowledge k-clique communities have never been extracted from the Internet AS-level topology graph due to the computational requirements of the Clique Percolation Method (CPM is the algorithm proposed by [23], which enables k-clique communities to be extracted from a graph). However, by using the algorithm described in [11] we were able to extract the k-clique communities from our Internet AS-level topology dataset in about 93 hours by running the Lightweight Parallel Clique Percolation Method on a 48-core computer. The main contributions of this paper are as follows:

- An analysis of the structural properties of the Internet AS-level topology graph by exploiting the k-clique communities definition [23];
- An analysis of the relationships between overlapping communities of a given order k;
- An interpretation of the driving factors behind such structural properties by means of two additional datasets, i.e. the IXP and the geographical datasets.

The remainder of this paper is organized as follows: in Chapter 2 we describe the topological dataset we used to compute the k-clique communities and the additional datasets in order to interpret the results of community detection. In Chapter 3 we describe the main characteristics of k-clique communities. In Chapter 4 we present our analysis of the k-clique communities detected within the Internet AS-level topology graph, and we summarize our results in Chapter 5.

[1] `http://www.db.ripe.net/whois`

Chapter 2

Data sources

In this section we describe the datasets we used to study the structural properties of the Internet AS-level topology graph. All the datasets, i.e. the topology, the IXP and the geographical datasets, were obtained at the end of April 2010. Thus, data belonging to different datasets are compliant and can be correlated by means of tags.

2.1 Topology dataset

The Internet AS-level topology dataset, hereinafter *Topology dataset*, is a collection of connections between ASes that describes the Internet AS-level topology as an undirected unweighted graph. This dataset was built according to the methodology described in [10], briefly: a) we downloaded three public available datasets (the IPv4 Routed /24 AS Links dataset [15], the Distributed Internet MEasurements and Simulations dataset [1] and the Internet Topology Collection at the Internet Research Lab dataset [2]) considering the measurement campaigns performed in April 2010; b) we merged the three datasets; c) we then removed spurious data from the topology. The resulting Topology dataset is made up of **35,390** ASes and **152,233** connections.

2.2 IXP dataset

Internet Exchange Points (hereinafter IXPs) make up a physical infrastructure that allows its participants, i.e. ASes, to easily establish connections with each other. ASes can reduce their costs by participating in these facilities (i.e. a part of their network is in the colocation center) since they can connect directly with other participants and settle BGP connections on the IXP rather than setting up multiple ad-hoc point-to-point connections or exploiting one or more third party networks. Over the last few years, researchers have proved the fundamental role of IXPs in Internet connectivity

([14], [4]). In addition, [10] and [12] highlighted that ASes participating at IXPs are responsible for the creation of well-connected zones of the Internet AS-level topology. In this paper we use the IXP dataset that was exploited in [10] and [12]. This dataset is a collection of information related to **232** IXPs, from all over the world, that were active in April 2010. Each IXP is associated with a geographical location and a list of ASes which participate in it. See [10] for an exhaustive description of IXP data gathering process.

2.3 Geographical dataset

ASes are made up of networks that can be placed in more than one location. Currently, ASes can be present in more than one city, country or, sometimes, in more than one continent. This information, which it is often correlated to the business profile of the considered AS, can be useful in order to understand how geography can affect the creation of communities. We thus used a Geographical dataset which enabled us to associate to each AS with a list of countries in which it has at least one point of presence. This collection was created in April 2010 by exploiting the MaxMind Geolite data[1] following to the method described in [12]. The resulting geographical database associates **34,190** ASes with at least one country code.

2.4 Tags

The Topology dataset and the IXP and Geographical datasets were correlated by defining two kinds of tags. The first category of tags is related to the IXP dataset: an AS is called an **on-IXP AS** if it belongs to at least one IXP participant list; otherwise, an AS is referred to as a **not-on-IXP AS**. The second category of tags refers to the Geographical dataset: an AS is called a **national AS** if all of its geographical locations belong to the same country, i.e. its networks are placed within a single country. An AS is called a **continental AS** if all of its geographical locations are placed within the same continent. An AS is called a **worldwide AS** if it owns at least two geographical locations that are located in two different continents.

on-IXP	*not-on-IXP*
4,462	30,928

Table 2.1: Summary of tagging results.

In Table 2.1 and in Table 2.2 we show the number of ASes of the Topology dataset that belong to each category. Please note that we refer to those ASes

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

National	*Continental*	*Worldwide*	*Unknown*
31,228	1,115	1,568	1,479

Table 2.2: Summary of tagging results.

whose geographical locations have not been discovered (i.e. they are not part of the Geographical dataset) as **unknown ASes**. These ASes are mostly stub ASes with a low degree [12].

In the following we will also exploit the concept of the **tag-induced subgraph**. According to [24], a subgraph of G (i.e. a generic graph) induced by the tag α is made up of all the edges of G whose endpoints are both tagged with the tag α. If we apply this definition to the previously described datasets, we can build, for instance, an IXP-induced subgraph or a country-induced subgraph by considering all the participant ASes of a single IXP or all the ASes with a geographical location in a given country respectively.

Chapter 3

k-clique Communities Detection

As introduced in Chapter 1, in this paper we analyse the structural properties of the Internet AS-level topology graph by using k-clique communities. The definition of a k-clique community reported in [23] is the following: a k-clique community is a union of all k-cliques (complete subgraphs 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). These k-clique communities have the following properties: a) their definition is deterministic; b) overlapping is allowed; c) each community identifies a set of cohesive nodes, since it can be described as a *chain* of fully-connected subgraphs (i.e. k-cliques). This is mainly why we adopted k-clique communities to analyze of the Internet AS-level topology. 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:

$$community_i(k) \subseteq community_j(k-1), \tag{3.1}$$

i.e. $community_i(k)$ is a subgraph of $community_j(k-1)$ (A proof of this is shown in Section 3.1). Thus, if we have a single community for each k, we can model the graph as a set of nested communities. On the other hand, when we have more than a single community for each k, there can exist *k-1*-clique communities that do not include any of the k-clique communities.

The k-clique communities studied in this paper were extracted from our Topology dataset by applying the Lightweight Parallel Clique Percolation Method [11]. Although this process of community detection took about four days to complete on a 48-core machine (see [11] for more details), the Lightweight Parallel Clique Percolation method, to the best of our knowledge, was the only algorithm that would enable us to obtain these k-clique communities, at least with our Topology dataset. This large execution time

depends on the number of maximal k-clique found in the Topology dataset and their distribution. In the Topology dataset there are 2,730,916 maximal k-cliques, 88% of which have k values in the range $[18:28]$.

3.1 *k*-clique nesting proof

Graph G is an undirected graph without self-links, i.e.:

$$G = \begin{cases} V_G &= \{1, \ldots, N\} \\ E_G &= \{e_1, \ldots, e_M\} \end{cases} \tag{3.2}$$

where V_G is a set of nodes and E_G is a set of edges where $e_m = \{i, j\} \subset V_G$ and $i \neq j$.

***k*-clique** We define k-clique as the subgraph $clique(k) \subset G$ such that:

$$clique(k) = \begin{cases} V_{clique(k)} &= \{1, \ldots, N_{clique(k)}\} \\ E_{clique(k)} &= \{e_1, \ldots, e_{M_{clique(k)}}\} \end{cases} \tag{3.3}$$

where:

- $\left|V_{clique(k)}\right| = N_{clique(k)} = k$;
- $\left|E_{clique(k)}\right| = M_{clique(k)} = k \cdot (k-1)$;
- $e_m = \{i, j\} \subset V_{clique(k)}$.

Since $clique(k) \subset G$, it follows that $i \neq j$, hence: each node $i \in V_{clique(k)}$ connects to each other node $j \in V_{clique(k)}$, i.e. it is involved in $k-1$ connections.

***k*-clique community** We define k-clique community a connected component of the graph $C(k)$:

$$C(k) = \begin{cases} V_{C(k)} &= \{1, \ldots, N_{C(k)}\} \\ E_{C(k)} &= \{e_1, \ldots, e_{M_{C(k)}}\} \end{cases} \tag{3.4}$$

where:

- $i \in V_{C(k)}$ is a k-clique, i.e. $i = clique_i(k) \subset G$;
- $e_m = \{i, j\} \subset V_{C(k)}$ such that $\left|V_{clique_i(k)} \cap V_{clique_j(k)}\right| = k - 1$.

Hereafter, we will refer to a k-clique community as $community_{nc}(k)$, where nc has values in the range $[1 : N_C(k)]$ ($N_C(k)$ is the number of distinct k-clique communities of order k within a graph G).

Theorem 1. *For each k-clique community, $community_i(k)$, there exists a unique k-1-clique community, $community_j(k-1)$ such that:*

$$community_i(k) \subseteq community_j(k-1)$$

Proof. $community_i(k)$ is a connected component of the graph $C(k)$ (see Expression 3.4). $community_i(k)$ can be represented as a graph whose nodes are k-cliques and edges connect k-cliques that share $k-1$ members. Each connection e_m of a k-clique-community connects two distinct k-cliques, suppose they are $clique_I(k)$ and $clique_{II}(k)$. Hence, each e_m, or e_{I-II} can be thought of as a set of $k+1$ distinct nodes:

- $node_{I*}$ is part of $clique_I(k)$ but it does not belong to $clique_{II}(k)$;
- $node_{II*}$ is part of $clique_{II}(k)$ but it does not belong to $clique_I(k)$;
- there are $k-1$ nodes that make up $clique_I(k) \cap clique_{II}(k)$. Hereinafter, we will refer to this set of nodes as $V_{I \cap II}$.

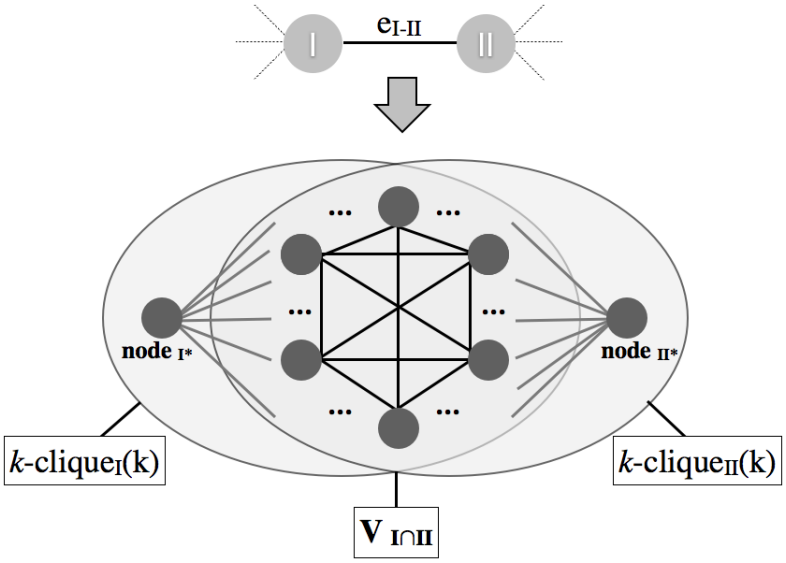


Figure 3.1: Connection between $clique_I(k)$ and $clique_{II}(k)$.

Each subset of a $clique(k)$ is a complete subgraph by definition. Hence nodes belonging to $V_{I \cap II}$ form a $clique(k-1)$. Hereinafter we will refer to this *k-1*-clique as $clique_{I \cap II}(k-1)$. Now select a node within the set $V_{I \cap II}$ and refer to it as $node_{(I \cap II)*}$. There can be individuated 3 distinct *k-1*-cliques:

- $clique_I(k-1)$ is a subset of $clique_I(k)$ composed of this set of nodes: $V_{clique_I(k-1)} = \{V_{clique_I(k)} \setminus node_{(I \cap II)*}\} = \{node_{I*} \cup V_{I \cap II}\}$.
- $clique_{II}(k-1)$ is a subset of $clique_{II}(k)$ composed of this set of nodes: $V_{clique_{II}(k-1)} = \{V_{clique_{II}(k)} \setminus node_{(I \cap II)*}\} = \{node_{II*} \cup V_{I \cap II}\}$.
- $clique_{I \cap II}(k-1) = clique_I(k) \cap clique_{II}(k)$.

$clique_I(k-1)$ and $clique_{I \cap II}(k-1)$ share $k-2$ nodes. $clique_{II}(k-1)$ and $clique_{I \cap II}(k-1)$ share $k-2$ nodes too. Then these three *k-1*-cliques are part of a common connected component within the graph $C(k-1)$. Hence they are part of a common *k-1*-clique community. We can extend this reasoning to all the connections that make up the $community_i(k)$ and find that $community_i(k)$ is a subset of a $community_j(k-1)$. $community_j(k-1)$ is unique since there cannot exist two distinct connected components within the $C(k-1)$ graph sharing edges by definition of distinct connected components.

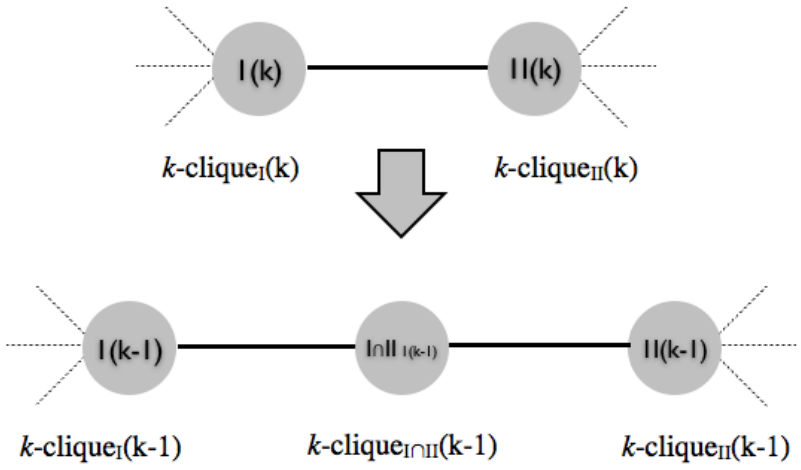


Figure 3.2: Translation of a connection belonging to a $C(k)$ graph into two connections within the $C(k-1)$ graph.

□

Chapter 4

Analysis of the k-clique Communities within the Internet AS-level Topology Graph

When the Lightweight Parallel Clique Percolation Method is applied to the Topology dataset the result is 627 k-clique communities. In Figure 4.1 we show the number of k-clique communities for each k. While low k values are characterized by the presence of several communities, high k values present a small number of communities. Note that, since the Topology dataset corresponds to a single connected component, it follows that there is a single 2-clique community.

If we analyse Figure 4.1 bearing in mind Expression 3.1, we can state that 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. Consider, for instance, k equal to 25, since Expression 3.1 holds, all the three 26-clique communities are subgraphs of the 25-clique community. In addition, since for each 27-clique community, there exists one and only one 26-clique community that includes it, it follows that the 25-clique community considered also contains all the 27-clique communities. If we extend this to the higher k values, then the 25-clique community can be said to include all the k-clique communities with a k value higher than 25. This property holds for all unique k-clique communities (i.e. 2, 21, 22, 25, 36).

On the other hand, by applying Expression 3.1 recursively, we can assert that given a k-clique community of order $\tilde{k}$, there is a k-clique community that completely contains it for each $k < \tilde{k}$. Thus, there are 34 k-clique communities that contain the 36-clique community. Hereinafter, we refer to these communities as *main* communities (the 36-clique community is also part of this set), and we refer to the remaining communities as *parallel* com-

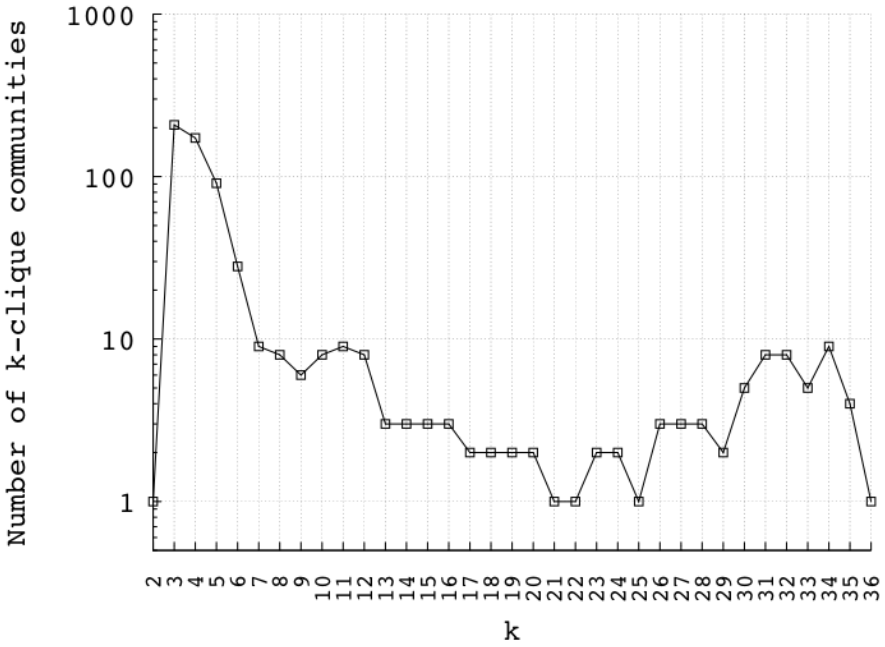


Figure 4.1: Number of k-clique communities vs. k.

munities. The relationships discussed above can be summarized with a tree representation, namely a k-clique community tree. We can represent each k-clique community with a node and we can plot an edge connecting a k-clique community with its relative *k-1*-clique community (i.e. the *k-1*-clique community which fully contains it). For each k there is a main community (the node filled with black) and, very often, more than one parallel community (unfilled nodes). The k-clique community tree is shown in Figure 4.2. We found that the properties that characterize main communities and parallel communities can be very different, unless we observe k-clique communities with a high k value close to 36.

We can start our analysis of k-clique communities by considering the size of the community, i.e. the number of ASes which belong to a community. In Figure 4.3 we plot the size of each community using two different point styles in order to distinguish the values related to main communities from the values related to parallel communities. The main community is made up of all the Topology dataset for $k = 2$, i.e. its size is equal to 35,390, then its size decreases rapidly as k increases. The size of the main community is comparable to that of parallel communities only for k close to 36. The vast majority of parallel communities have a size value which is close to the k value, i.e. they are made up of a small number of maximal cliques. Their size trend obviously increases as the minimum size of a k-clique community is k by definition. Those decreasing trends for the parallel communities with

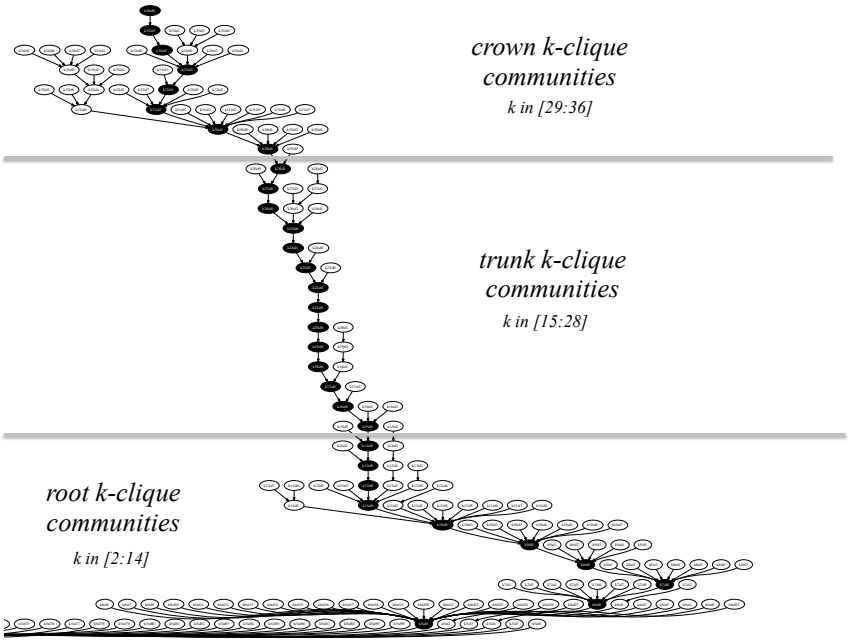


Figure 4.2: k-clique community tree (for readability no k-clique communities with a $k \leq 5$ are shown).

k in one of these ranges, $[11:17]$ or $[18:20]$ or $[26:29]$ or $[31:35]$, appear as branches of the tree.

In order to better understand the role of k-clique communities, both main and parallel, within the Internet AS-level topology graph, we plot in Figure 4.4(a) their link density [17] and in Figure 4.4(b) their average Out Degree Fraction [20], hereinafter ODF. Link density is defined as the fraction of existing connections (within the community) to possible connections. This metric has values in the range $[0:1]$ and indicates how densely-connected a subgraph (community) is by comparing its number of edges with the number of edges of a relative full mesh topology. ODF also has values in the range $[0:1]$. Given a node i, its ODF is the ratio between its degree within the subgraph (community) and its overall degree, i.e. its degree within the Internet AS-level topology graph. Thus, the average ODF of a subgraph (community) expresses the tendency of community nodes to direct their degree inside or outside the community.

By analysing Figure 4.3, Figure 4.4(a) and Figure 4.4(b), we can identify three behaviors. The first case refers to those main communities with a k value in the range $[2:30]$. These communities are likely to be made up of long k-clique chains, thus, although ASes are locally well-connected, we are very unlikely to find full-mesh-like topologies. This structural property

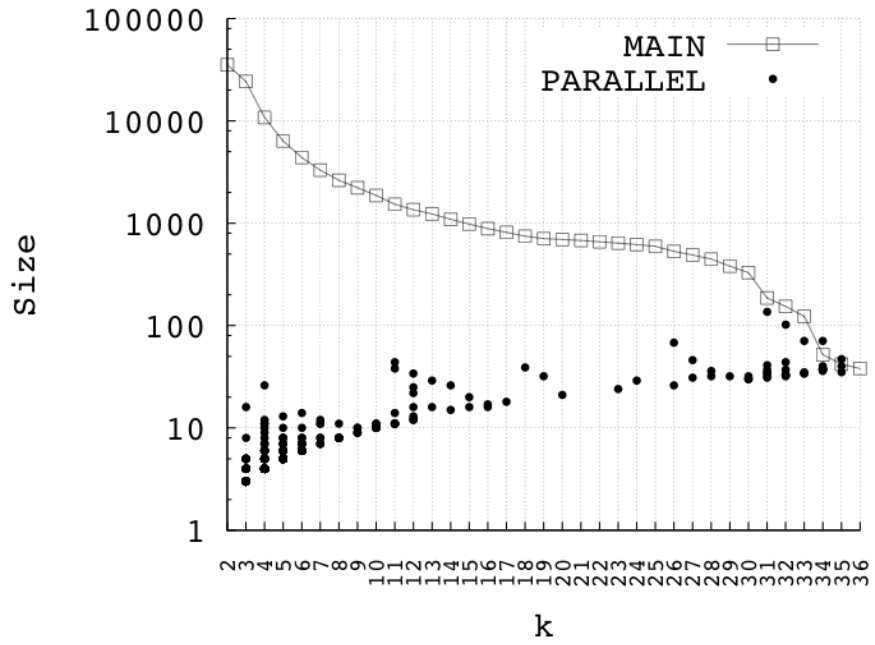
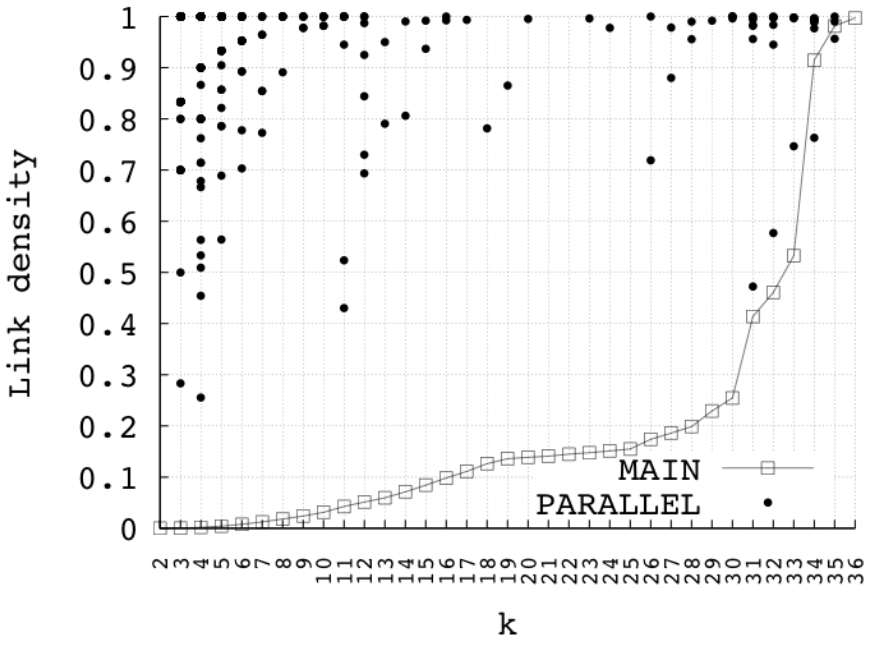


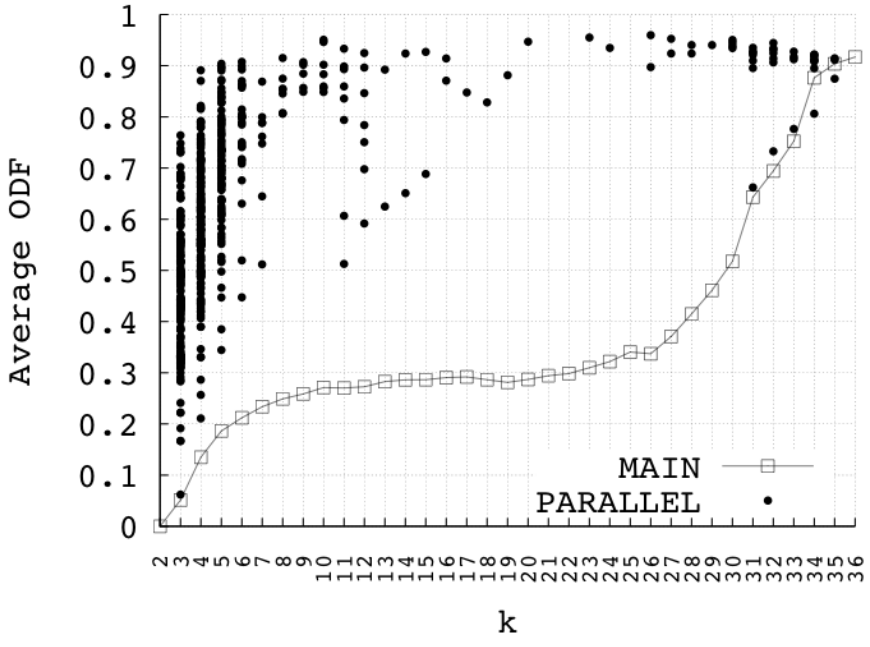
Figure 4.3: Size of k-clique communities vs. k.

implies a low link density. On the other hand, many members of these communities are likely to have more connections within their community than connections directed outside. Note that for k equal to 3, the number of ASes within the main community is equal to 69% of the Topology dataset ASes while the remaining 31% of ASes belong to the main 2-clique community or to the parallel 3-clique communities. Thus ASes of the main 3-clique community are very unlikely to have a high average ODF. The second case refers to the main k-clique communities with a size comparable to the k value, such as those with a k in the range $[31 : 36]$, and several parallel k-clique communities that present a high link density value. From a structural point of view, these k-clique communities are very similar to clique topologies. These k-clique communities also present a high average ODF, hence, even if they appear as cohesive sets of ASes, they have a very large number of connections directed outside the community. The third case refers to those parallel k-clique communities with a low k value and a very variable link density and ODF. This can be justified by observing that, since these communities are made up of a very small number of ASes (see for instance the size of the parallel k-clique communities with a k in the range $[3 : 7]$) the presence or the absence of few connections can strongly influence both the link density and the average ODF values.

The properties described above provide some insight into the differences between main and parallel k-clique communities and their connections in



(a)



(b)

Figure 4.4: Link density, Figure 4.4(a) and Average ODF, Figure 4.4(b) vs. k.

the Internet AS-level graph. We will now extend our analysis by investigating the relationships between these communities. To do this we exploit the overlapping between communities sharing the same k. At the same time we avoid computing the overlap between communities with a different k, since the nesting process provides results that are very difficult to interpret. Overlap is defined as the number of members that are shared by two communities. This metric has values in the range $[0 : max_{overlap}]$ ($max_{overlap}$ is the size of the smaller community. In fact, the maximum overlap occurs when all the members of a community are also members of the other community). In order to compare overlap values related to different pairs of communities, we define the *overlap fraction* as the ratio between the overlap and the $max_{overlap}$ value. The computation of the *overlap fraction* between communities with the same k value provides the following results: a) every parallel community shares at least one AS with its relative main community[1]; b) there are parallel communities that do not overlap with other parallel communities; c) we can identify small sets of parallel communities with a pretty high overlap fraction. We found that the average overlap fraction between the main k-clique community is always larger than 0.432 for each k. In addition by averaging this metric over k we obtain 0.704 and a variance equal to 0.023. Since, for the vast majority of k-clique communities the size of parallel communities is always lower than the relative main community, we can interpret the previous overlap fraction as follows: on average, the 70.4% of ASes belonging to a parallel community also participate in the main community. On the other hand, the average overlap fraction between the parallel k-clique communities can vary a lot. There are parallel communities that do not share any member as well as communities sharing the vast majority of their ASes. Due to a high variance value (i.e. 0.136) we avoid reporting the overlap fraction value averaged over k.

In order to understand which factors might lead to these structural properties we study the k-clique community tree using tags (see Chapter 2). Firstly, we computed the percentage of on-IXP ASes in each k-clique community. As in [10] and [12], we found that the most well-connected communities are made up of a large number of ASes participating in IXPs. We found that all the k-clique communities with a k greater than or equal to 16 have more than 90% on-IXP ASes. For those communities with a k lower than 16, the percentage of on-IXP ASes is highly variable. We then refined our analysis of IXP tags by building an IXP-induced subgraph for each IXP and then computing the overlap between these subgraphs and k-clique communities. We discovered that **35** k-clique communities were subgraphs of an IXP-induced subgraph. This means that there are communities made up

[1] As a matter of fact, there are 6 exceptions: there is a 6-clique community, a 5-clique community, three 4-clique communities and a 3-clique community that do not share any AS with their respective main community.

of ASes belonging to a common IXP. There are essentially three behaviors: a) if $k > 28$ we can find communities that are fully included in DE-CIX- or LINX- induced subgraphs only; b) if $k < 14$ there are communities that are fully included in small IXPs; c) if $k \in [14 : 28]$ none of the communities are fully included in an IXP-induced subgraph. On the basis of these three ranges, we extend our analysis of the k-clique communities by studying *crown* communities (i.e. $k > 28$), *trunk* communities (i.e. $k \in [14 : 28]$) and *root* communities (i.e. $k < 14$) separately. These three categories can be associated with three separate parts of the k-clique community tree as shown in Figure 4.2. Hereinafter, we will refer to the IXP with the maximum number of participants in common with a community as its *max-share-IXP*. In addition, if the community is fully included in the IXP-induced-subgraph, this IXP is called *full-share-IXP* for that community. A community with a full-share-IXP can be interpreted as follows: the community considered is made up of a subset of the full-share-IXP participants.

4.1 Crown k-clique communities

This category is made up of 42 k-clique communities with a k value in the range $[29 : 36]$. We pay special attention to the 36-clique community, since it is the most dense subgraph in our Topology dataset according to the k-clique community definition. Although this community is made up of 38 on-IXP-ASes participating in several IXPs worldwide, it does not have a full-share-IXP. However, it shares the 89% of its members (ASes) with AMS-IX, which is also its max-share-IXP. Since all the main communities include, by definition, the 36-clique community, none of the main communities will have a full-share-IXP. All the ASes belonging to crown k-clique communities are in Europe and participate in at least one IXP. The only exceptions are the following four ASes: 2905 (TICSA), 3.236 (MIT-GATEWAYS), 3.392 (MIT-GATEWAYS) and 37179 (AFRICAINX), at least with our datasets, are not present in Europe. In addition, there are three that do not participate in any IXPs. Another common feature of the crown k-clique communities is their max-share-IXP which is always one of these three: AMS-IX, DECIX or LINX. In order to gain insight into the presence of these three IXPs within the most well-connnected part of the Internet AS-level topology, let us now analyse the 34-clique communities in detail. There are nine 34-clique communities which can be described as follows: the main community shares 92% of its members with AMS-IX; four communities have LINX as their full-share IXP, three communities have DE-CIX as their full-share IXP and one community shares 98% of its members with DE-CIX. Although these nine communities have different max-share-IXPs, they overlap with each other. This can be explained by considering that AMS-IX, DE-CIX and LINX share several participants (119). On the other hand, if we observe the

overlap fraction values of the 34-clique communities, it is clear that those couples of communities with the same max-share-IXP have a higher overlap fraction than the other couples.

4.2 Trunk *k*-clique communities

This category is made up of 30 k-clique communities with a k value in the range $[15 : 28]$. This part of the tree is characterized by a small number of communities sharing the same k value (compared to the other parts of the tree). Although the percentage of on-IXP ASes in each community is very high (higher than 90%) there are no full-share-IXPs. If we consider the percentage of ASes shared between communities and their max-share-IXP, we find that parallel communities have high percentages, while main communities do not. Consider for instances the branch made up of the three parallel nested communities with k equal to 20, 19 and 18. These three nested communities have a size equal to 21, 32, 39 respectively, moreover all have the same max-share-IXP, MSK-IX and share more than 95% of their ASes with it. Since the size of the main community is now much larger than the size of its relative parallel communities, it is unlikely to find main communities fully included in an IXP-induced-subgraph. From a structural point of view, these main communities are large and dense k-clique chains. In addition, their ASes present a pretty high average Internet degree, i.e. 500.2. By exploiting our Geographical dataset we also found that a high percentage of trunk k-clique community ASes are worldwide or continental. These characteristics, suggest that there are Internet providers within this category of communities.

4.3 Root *k*-clique communities

This category is made up of 554 k-clique communities with a k value in the range $[2 : 14]$. With the exception of the main communities, their average size is very low, i.e. 5.09 ASes per community. As mentioned before, there are 14 parallel communities with a full-share IXP. In contrast to crown communities, there are several IXPs acting as a full-share IXP, and some of them are not European. We identified parallel communities that were fully included in one of the following IXPs: WIX (New Zealand), KhIX (Russia), SIX (United States of America), SIX.SK (Slovakia), PIPE-NSW (Australia), NIXI-Delhi (India), SPB-IX (Russia), PTTMETRO Sao Paulo (Brasil), NIX (Czech Republic), SWISS-IX (Switzerland), MIX-IT (Italy) and VIX (Austria). Very often, parallel communities are fully included in country-induced-subgraphs. If this happens, it means that all the members of the considered community have a geographical location in a common country. We discovered 382 root communities with this property. Thus,

most of the root k-clique communities are likely to be originated by regional environments.

Chapter 5

Discussion and Conclusion

In this work we have analysed the community structure of the Internet AS-level topology graph by adopting the k-clique community definition and by exploiting both the IXP and the Geographical datasets.

We propose a novel approach to represent Internet k-clique communities and how the nesting process affects them: the k-clique community tree. In addition, we define two classes of k-clique communities, main and parallel communities. The Internet AS-level topology community structure can be broadly described as a system made up of these two classes of communities: as k decreases we can always find a growing main community including all the previous main communities. In addition, there are parallel branches of the tree (i.e. parallel communities) characterized by a limited size which are rapidly incorporated into a main community with a lower k. Main communities and parallel communities present very different characteristics unless we consider k-clique communities with a k value close to 36. Main communities are typically large and have a low ODF and link density values. On the other hand, parallel communities are usually small and with a large ODF and link density values.

We extended our analysis of k-clique communities by using our additional datasets (i.e. those containing geographical information and data related to the IXPs) and by studying three categories of communities separately: crown, trunk and root communities. Crown communities, which represent the most dense Internet subgraphs, are made up almost exclusively of ASes participating in AMS-IX, DE-CIX and LINX IXPs. Trunk communities are also made up of several on-IXP ASes.Unlike the other categories, none of the trunk communities can be thought of as a subgraph of an IXP-induced graph. ASes populating these communities present a high average degree and typically have several geographical locations in more than one country. Thus these ASes are likely to be service providers (e.g. CDNs, IBPs, transit providers). Root parallel communities can be thought of as a small set of ASes that form regional dense subgraphs. Using our datasets we found

that there are 382 k-clique communities made up of ASes with a common geographical location in a country. These communities are likely to be made up of small groups of customers and providers forming a clique because of multi-homing.

Bibliography

[1] Distributed Internet MEasurements and Simulations dataset, 2010.

[2] Internet Topology Collection at the Internet Research Lab dataset, 2010.

[3] 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.

[4] Brice Augustin, Balachander Krishnamurthy, and Walter Willinger. IXPs: mapped? In *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 336–349, New York, NY, USA, 2009. ACM.

[5] 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+, 2008.

[6] 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, 2007.

[7] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, 2004.

[8] Amogh Dhamdhere and Constantine Dovrolis. Ten years in the evolution of the internet ecosystem. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, IMC '08, pages 183–196, New York, NY, USA, 2008. ACM.

[9] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[10] 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.

[11] Enrico Gregori, Luciano Lenzini, Simone Mainardi, and Chiara Orsini. Lightweight Parallel Clique Percolation Method. In *Submitted to IFIP Networking 2011*, 2011.

[12] 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.

[13] Roger Guimera and Luis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[14] Yihua He, Georgos Siganos, Michalis Faloutsos, and Srikanth V. Krishnamurthy. Lord of the links: a framework for discovering missing links in the internet topology. *IEEE/ACM Trans. Netw.*, 17(2):391–404, 2009.

[15] Young Hyun, Bradley Huffaker, Dan Andersen, Emile Aben, Matthew Luckie, k c Claffy, and Colleen Shannon. The IPv4 Routed /24 AS Links Dataset, 2010.

[16] Haewoon Kwak, Yoonchan Choi, Young-Ho Eom, Hawoong Jeong, and Sue Moon. Mining Communities in Networks: A Solution for Consistency and its Evaluation. In *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 301–314, New York, NY, USA, 2009. ACM.

[17] Andrea Lancichinetti, Mikko Kivela, Jari Saramaki, and Santo Fortunato. Characterizing the community structure of complex networks. *PloS One 5(8)*, 2010.

[18] 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.

[19] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[20] Jure Leskovec, 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.

[21] Sue Moon, Jinyoung You, Haewoon Kwak, Daniel Kim, and Hawoong Jeong. Understanding Topological Mesoscale Features in Community Mining. In *COMSNETS 2010: Proceeding of the Second International Conference on COMmunication Systems and NETworkS (invited paper)*, pages 1 –10, 2010.

[22] 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.

[23] 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.

[24] Gergely Palla, Ills J Farkas, Pter Pollner, Imre Dernyi, and Tams Vicsek. Fundamental statistical features and self-similar properties of tagged networks. *New Journal of Physics*, 10(12):123026, 2008.

[25] Kazumi Saito, Takeshi Yamada, and Kazuhiro Kazama. Extracting Communities from Complex Networks by the k-Dense Method. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E91-A(11):3304–3311, 2008.

[26] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

[27] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu. Detect overlapping and hierarchical community structure in networks. *CoRR*, abs/0810.3093, 2008.

[28] Ken Wakita and Toshiyuki Tsurumi. Finding Community Structure in Mega-scale Social Networks. *CoRR*, abs/cs/0702048:9, 2007.

[29] Walter Willinger, Ricardo Oliveira, and Beichuan Zhang. Quantifying the Completeness of the Observed Internet AS-level Structure. *Technical Report, UCLA CS Department*, 2008.