information

MDPI

Article

A Novel Method for Router-to-AS Mapping Based on Graph Community Discovery

Hangyu Hu *, Weiyi Liu, Gaolei Fei, Song Yang and Guangmin Hu

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; unique_liu@163.com (W.L.); fgl@uestc.edu.cn (G.F.); syang201411@163.com (S.Y.); hgm@uestc.edu.cn (G.H.)

* Correspondence: huhangyuuestc@gmail.com; Tel.: +86-17313110233

Received: 29 January 2019; Accepted: 22 February 2019; Published: 27 February 2019

Abstract: The last decades have witnessed the progressive development of research on Internet topology at the router or autonomous systems (AS) level. Routers are essential components of ASes, which dominate their behaviors. It is important to identify the affiliation between routers and ASes because this contributes to a deeper understanding of the topology. However, the existing methods that assign a router to an AS, based on the origin AS of its IP addresses do not make full use of the information during the network interaction procedure. In this paper, we propose a novel method to assign routers to their owners' AS, based on community discovery. First, we use the initial AS information along with router-pair similarities to construct a weighted router level graph; secondly, with the large amount of graph data (more than 2M nodes and 19M edges) from the CAIDA ITDK project, we propose a fast hierarchy clustering algorithm with time and space complexity, which are both linear for graph community discovery. Finally, router-to-AS mapping is completed, based on these AS communities. Experimental results show that the effectiveness and robustness of the proposed method. Combining with AS communities, our method could have the higher accuracy rate reaching to 82.62% for Routers-to-AS mapping, while the best accuracy of prior works is plateaued at 65.44%.

Keywords: Router-to-AS mapping; community discovery; global router topology; fast hierarchy clustering

1. Introduction

As a hot topic in the network science research area, the topology of the Internet has drawn considerable attention from many researchers. They often probe and construct the topology structure at the router or autonomous systems (AS) level [1–4], and then use it to study the connectivity characteristics of the topology, design the topology of generators, develop network protocols and evaluate their performance. The definition of an AS is a set of routers under a single technical administration [5], which means that routers are the foundation of an AS and all of the routers in an AS observe the same routing policy decided by the AS. Thus, AS can affect the behaviors of their routers, and the routers can also reflect some characteristics of their owner ASs. In addition, router-level topology can provide a more detailed intra- and inter-connection of AS. Therefore, it is necessary to merge the router and the AS level topology, i.e., mapping each router to the AS that owns it, which can help deepen our understanding of the network topology. For example, through mapping we can obtain an accurate AS-level trace-route tool [6], and study the correlation between the degree of the AS and the number of their routers [7], which has become a critical task in the field of network monitoring.

The essential component of router-to-AS mapping, is extracting AS-level information from routers' IP addresses. To the best of our knowledge, there are few works focusing on router-to-AS

mapping. Tangmunarunkit et al. [7,8] were the first to try to solve this problem. They directly mapped the router to the AS, which appears most frequently on the origin AS of the router. In addition, Pansiot et al. [9] proposed five simple rules that include two probabilistic rules and three empirical rules to identify the owner AS of the AS border routers (ASBR), such as global election, which is the same as the method adopted by Tangmunarunkit. Researchers working for the Center for Applied Internet Data Analysis (CAIDA) [10] have measured the Internet topology from Ark infrastructure [11]. In order to understand the incompleteness of the Internet AS-level data, they mapped the observed IP addresses to their DNS names in real time, and analyzed the Internet as a critical infrastructure in router-to-AS assignments. Accurately determining the routers' IP addresses that are used for AS links from trace-route traces can be hard because these interfaces of routers are often assigned addresses from neighboring AS. To address this problem, Alexander Marder et al. [12] developed a new algorithm, Multipass Accurate Passive Inferences from Traceroute (MAP-IT), for inferring the exact interface addresses, used for point-to-point inter-AS links, as well as the specific AS involved. Enrico Gregori et al. [13] analyzed BGP data by the router collector project and found that large areas of the Internet are not properly captured, due to the geographical location of route collector feeders and BGP filters. They proposed a method based on a new metric, named p2c-distance, to identify the minimum number of AS required to obtain an Internet AS-level topology. However, both of them have not validated the accuracy of their method. As far as we know, to date, the only method that has been validated was proposed in [14]. Huffaker et al. designed five router assignment heuristics based on the origin AS of routers, and validated them on the ground truth provided by two ISPs and five research networks. They also tested all combinations of pairs of heuristics, and found that the most successful pair was election + degree. Throughout these works, the researchers only took the information from the router itself or its neighbors into account, which is clearly limited.

Since various types of real-world complex networks (including router-level topology graphs) have certain community structures [15], where a community refers to a set of vertices that behave differently from the rest of the system [16], they may have a dense connection within them [17,18]. Benjamin Fabian et al. [19] modelled the Internet structure in the context of known online financial services as a graph, at the level of AS, and assessed their connectivity using multiple graph measures, such as distance measures, centrality measures and neighborhood measures. But they only focused on calculating AS-level graph metrics to analyze the network connectivity. The findings reveal an impressive diversity between the AS, and help us to understand AS' structure more clearly. According to [5], AS is defined in the same collection of technical management of a group of routers. Such a definition means that a router within the same AS may have more connections, while others have less. It seems that AS behaviors from the Internet are identical to communities from complex networks. In order to enhance the community attributes of router-level topology graphs, correlativity is being introduced to depict router-pair relationships, because although routers may not have dense connections under an AS, their correlational relationships cannot vary differently. Inspired by this, we used the initial AS information of IP interfaces to calculate the node similarity, which can be used to quantify the router-pair correlations. Based on the router-pair correlations, we constructed a weighted graph, where vertices represent the routers and edges represent their correlation relationships. Then, we implemented community discovery methods on the weighted router level graph. We believe that community discovery methods can be used with this graph to find the underlying AS communities, and router-to-AS mapping can leverage these communities' information instead of the old fashioned methods. The experimental results in chapter 5 support our idea of utilizing AS communities under such weighted router level topology to help the router-to-AS mapping, and the best accuracy rate can increased to 82.62%.

The contributions of our paper lie in three parts:

- We construct a weighted router level graph by using the initial AS information of IP interfaces and router-pairs similarities simultaneously;

- We propose a fast hierarchy clustering with time and space complexity which are both linear, which is capable of finding AS communities;
- We demonstrate our method using community discovery upon a weighted router-level graph, which can lead to a drastic increase of the accuracy rate.

In Section 2, we give a brief overview of our method, and then we discuss some key issues in weighted router graph construction and router community discovery in Section 3. Section 4 firstly introduces datasets, used for constructing the global router topology and validation, and then presents baseline methods for comparison. In Section 5, we compare our router-to-AS mapping method with other base methods, while Section 6 concludes the paper.

2. Framework of Weighted Router Graph Construction and Router-to-AS Mapping

In a weighted graph, the edges' weights can contain more information than the graph structure itself, and community discovery can leverage these weights to obtain better performance than methods which only use the graph structure information. To construct a weighted router graph, we combine IP port information with the routers' ego-network to weight the edges between router pairs. As shown in Figure 1, the weighted router graph construction mainly contains three parts. In this section, we give a brief introduction on how weighed router graph be constructed, and the key issues among the construction procedure are discussed in Section 3.

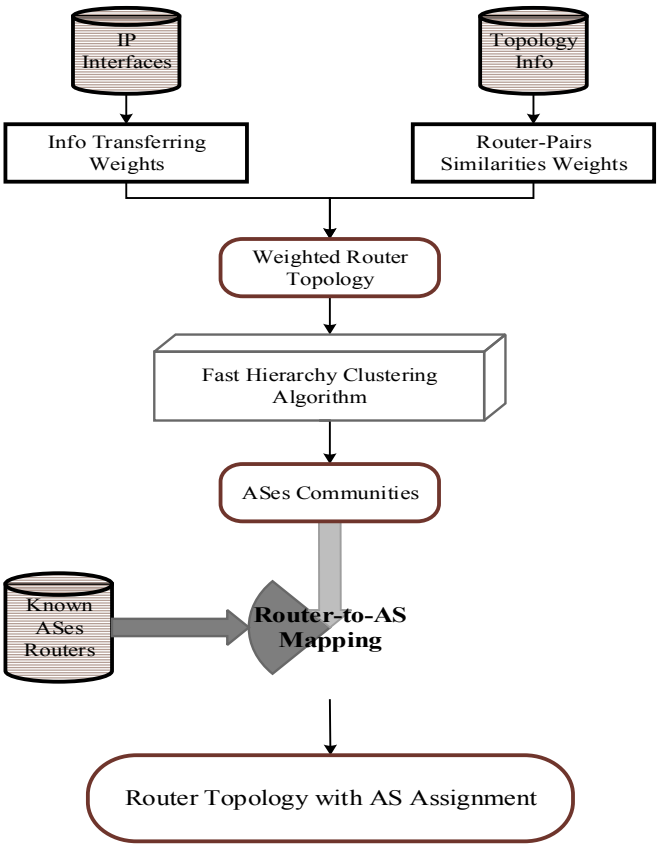


Figure 1. Router-to-AS mapping framework from community discovery.

2.1. Obtaining Weights

Firstly, we use IP-AS mapping to find out which AS a router-pair may communicate with, since there may be more than one IP port opened during two routers interaction and AS frequency must be taken into consideration. Here, we use the "Info Transferring Weight" to represent the AS frequency:

The higher of the current AS frequency, the bigger of the "Info Transferring Weight". Secondly, we use the "Router-Pairs Similarity Weight" to describe router-pair similarities based on node similarity measurement from the network science area.

2.2. Weight Fusion

Since the "Info Transferring Weight" and "Router-Pairs Similarity Weight" are irrelevant, we also need to utilize a basic operator to combine these two weights together for simplification as shown in Section 3. Moreover, we could have used more complicated methods, such as optimization method, to combine two weights together, which is not included due to the length of the paper.

2.3. Weighted Community Discovery

As the AS number is given in advance, we could have used a set of community discovery methods, such as the hierarchy clustering, which can control output community numbers to avoid the inequality between AS numbers and community numbers. As shown in Section 5, the hierarchy clustering algorithm works perfectly in generating the community numbers, and for massive datasets, our proposed method also has good robustness.

3. Key Issues of Our Method

As is well-known, there are two distinguishing features for router level graph. First and foremost, a router should be claimed by one AS, which indicates that AS communities in the router-level network should be non-overlapping; second, since the router level is the most basic level of the graph, the node size is generally tremendous. Focusing on these two features, we intend to use hierarchy clustering where time and space complexities are both linear, to perform non-overlapping community discovery on our weighted router level graph. In conclusion, the steps of our method are presented as follows: (1) combining router-pair IP port information with router-pair similarity together to generate a weighted graph; (2) using hierarchy clustering algorithm for community discovery based on this weighted graph; (3) perform router-to-AS mapping according to these communities. In this section, we have discussed the key issues appeared among the steps above.

A. Obtaining Info Transferring Weights

As router pairs use their opened ports for network information transferring, using IP-to-AS mapping we can determine which AS router-pairs may be useful for communication. Here, we use AS frequency to quantify the weight between router-pairs. As illustrated in Figure 2, blocks and lines represent routers and their opened ports, the AS information above the lines represents the AS number which current opened port belongs to, and l_{AB} represents the Info Transferring Weight between *Router A* and *Router B*. From Figure 2, *Router A* has four ports opened for interaction, while three of them belong to AS1, another belongs to AS2, and the information transferring procedure between *Router A* and *Router B* is achieved through AS1. Thus, for *Router A*, the transferring probability of information through AS1 is 75%; and for *Router B*, the probability is equal to 1/3 as there is an only one port belonging to AS1. Therefore, the value of Info Transferring Weight between *Router A* and *Router B* is 75%*1/3 = 25%. Generally speaking, we can use Equation (1) to calculate the router-pair information transferring weight W_p.

$$W_p(A,B) = \begin{cases} \sum\limits_{i,j \in \{AS(A),AS(B)\}} w_i(A) \times w_i(B), \; if \; adj(A,B) = 1 \\ 0, \; if \; adj(A,B) = 0 \; or \; i \neq j \end{cases} \tag{1}$$

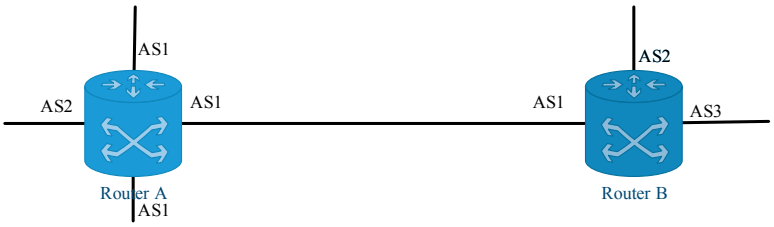


Figure 2. Toy example of Router-Pairs AS Info.

B. Obtaining Router-Pairs Similarity Weights

In the network science area, there are two metrics that could be used to quantify node-pair similarity: (1) The Local Similarity Metric: while only uses node-pair neighbor topologies to depict similarity, such as the Jaccard metric; (2) the Global Similarity Metric: Utilizing not only neighbor topologies, but also the whole network structure to calculate similarity. In general, the Global Similarity Metric performs better than the local metrics because it contains more information. But for router networks, the communication between a router-pair only involves these two routers, and is not related to other pairs. In such case, the local similarity metric is suitable for describing router-pair relationships. However, as shown in Figure 3, there is a weakness in using Local Similarity Metric. For example, when using the Jaccard Metric to quantify node-pairs' weight, the original Jaccard Metric can be defined by Equation (2) where $\Gamma(A)$ stands for *Router A*'s neighbors, and then we have that $\Gamma(A) \cap \Gamma(B) = \Gamma(B) \cap \Gamma(C) = \varnothing$, leading to $\text{Jaccard}(A,B) = \text{Jaccard}(B,C) = 0$. The "zero similarity"—represents that there is no edge between routers A and C, which goes against Figure 3. To address this problem, we introduce a novel conception of Generalized Neighbors $\Gamma^+(A)$ which is defined by Equation (3) against with the 10 Local Similarity Metric proposed by Zhou T [20], and all these updated metrics are presented in Table 1. In addition, meanings and comparison details are not included in this paper, as [20] (pp. 5–8, 10–12) has a complete introduction and discussion.

$$\text{Jaccard}(A,B) = \frac{|\Gamma(A) \cap \Gamma(B)|}{|\Gamma(A) \cup \Gamma(B)|} \tag{2}$$

$$\Gamma^+(A) = \Gamma(A) \cup \{A\} \tag{3}$$

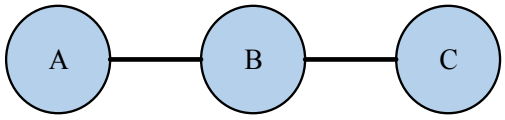


Figure 3. Toy Example of Jaccard Similarity Failure, with $\text{Jaccard}(A,B) = \text{Jaccard}(B,C) = 0$; $\text{Jaccard}^+(A,B) = \text{Jaccard}^+(B,C) = 2/3$.

Table 1. 10 kinds of Local Based Node-Pair Similarity Metrics.

Metrics	Similarity Metric Definition
CN^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert$
Salton^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / \sqrt{K(A) \times K(B)}$
Jaccard^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / \lvert\Gamma^+(A) \cup \Gamma^+(A)\rvert$
Sorenson^+	$2\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / K(A) + K(B)$
HPI^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / \min(K(A) \times K(B))$
HDI^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / \max(K(A) \times K(B))$
LHN-I^+	$\lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert / K(A) \times K(B)$
PA	$K(A) \times K(B)$
AA^+	$\sum_{Z \in \lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert} \frac{1}{\log K(Z)}$
RA^+	$\sum_{Z \in \lvert\Gamma^+(A) \cap \Gamma^+(B)\rvert} \frac{1}{K(Z)}$

C. Weighted Topology Generation

After obtaining the Info Transferring Weights and Router-Pairs Similarity Weights, we generate a weighted graph based on these two independent values. Since there is no theory or previous research results to take as examples, we choose four basic operators to perform the combination experiments; as described as below, we present the key idea of using such operators to generate a weighted topology.

- Operator **"Plus"**: this operator means that the Info Transferring Weights and Router-Pairs Similarity Weights take the same importance when we generate the weighted topology.
- Operator **"Times"**: because Times means performing an independent observation on the target system, here we use this operator to represent the independence.
- Operator **"Max & Min"**: this operator is to eliminate average or relative error during the "Plus" or "Times" two process.

D. Fast Hierarchy Clustering

There are two steps in the traditional hierarchy clustering algorithm. The first step is to perform node fusion to form a dendrogram, based on the node-pair's weight; the second step is to use the graph-partitioning method to cut this dendrogram to obtain various communities.

As illustrated in Figure 4, $W_{adj} = (w(i,j))_{N*N}$ represents the weighted Adjacent Matrix of the original network by using Jaccard Similarity Metric to calculate node-pair similarity, then node-pair fusion is used to form dendrogram. After that, Modularity Maximum model was used to discover different communities. From the best cut shown in Figure 4, it is easily to divide the original network structure into two non-overlapping communities: {1, 2, 3} and {4, 5, 6}. Taking this example in router-to-AS mapping, if node 1 and node 2 in one community belong to AS1, then node 3 may also belongs to AS1. However, this simple method may fail when processing massive datasets, because there are N nodes in a dendrogram, and the fusion process can only merge two nodes together which leads to the dendrogram height also being N. Thus the traditional hierarchy clustering algorithm suffers a high space complexity $O(N^2)$. Generally speaking, the reason for a high space complexity in traditional hierarchy clustering is that the algorithm needs to consider the entire topology at the same time; that means it has to utilize global information to generate the dendrogram, and this "GLOBAL" leads to the consumption of a large amount of memory space. For example, the global router topology described below has more than 2M nodes; thus using traditional hierarchy clustering method would spend roughly 2M $\times$ 2M = 4 $\times$ 10^3 G memory to store this massive dendrogram. This is impossible for a standard PC whose memory space is often 4~8G. Thus, we introduce a novel hierarchy clustering method—"Fast Hierarchy Clustering" to address this problem.

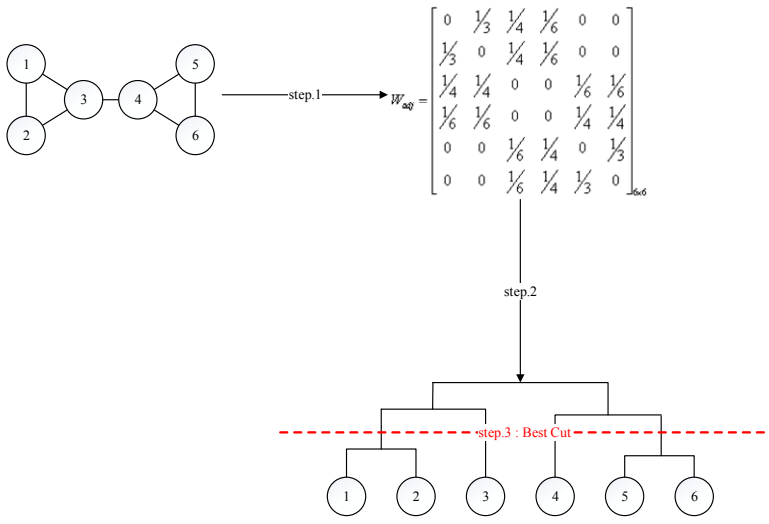


Figure 4. Toy Example of Ordinary Hierarchy Clustering.

It is clear that the general mechanism of hierarchy clustering is to aggregate two nodes together that share the same weight, which will never be separated once aggregated. Leveraging these two characteristics, a novel algorithm that only relates to the local information of the router's ego network was conceived. In traditional hierarchy clustering, we can only perform community discovery if we obtain the entire dendrogram prior, but if we merely discover which community a target node belongs to, there is no need to acquire the entire dendrogram; instead, we can just utilize the target node's ego network. As shown in Figure 5, the Arabic numerals upon edges means the node-pairs' weight, and then if we want to discover which community that node A belongs to, firstly we can extract node A's ego network, and then perform aggregation based on node A's neighbors' weights. Instead of generating the entire dendrogram, if we take "NODE A" as the aggregating target, then the aggregation of this node become a local problem; that is, to find another node (node B) in node A's ego network who shares the highest weight with node A. If we implement this mechanism iteratively, we can also build a dendrogram equal to that using the traditional clustering method.

After constructing the local dendrogram, how to identify the community information of every unknown node is still a difficult problem which needs to be solved. First of all, we use routers for which all ports have the same AS number as the ground truth; secondly, we combine the local dendrogram and Markov mechanism to assign the unknown routers AS number. In this case, if we have an unknown router A as a target router, then the most probable AS number of router A is the nearest neighbor (maybe router B) of router A in the dendrogram; if router B's AS number is still unknown, then we make router B as the new target node, using the same method to find this new target's (router B) most probable AS number. Using this mechanism recursively, we can finally acquire router A's most probable AS number.

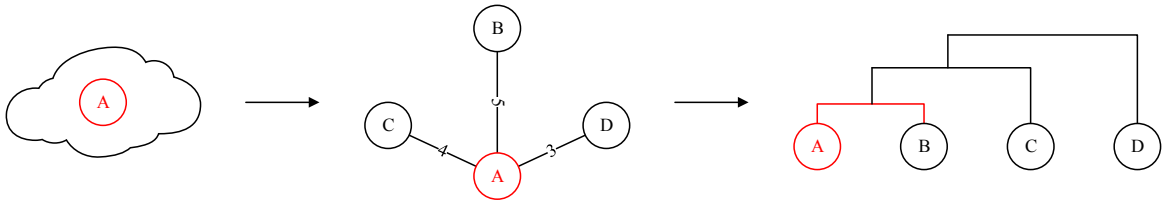


Figure 5. Toy Example of Local Node Clustering.

One concern with our method is whether the sub-node community may grow and lead to a non-convergence problem. Here, we let AS1 be the current mapping target AS, T_1 be the sub tree where all nodes belongs to AS*, $N(T_1)$ be the number of nodes in T_1, and the height of T_1 is H_1. For all nodes in T_1, if there are another AS2 beneath T_1, the height and number of nodes is H_2 and $N(T_2)$. Two cases may occur: (1) $H_1 \geq H_2$; and (2) $H_1 < H_2$. For case 1, as in the dendrogram, more height means more nodes will lead to $N(T_1) \geq N(T_2)$ naturally this means that when $T_1 = T_2$, the node number in the subtrees will decrease; in this case, our method shows a good convergence. For case 2, $H_1 < H_2$ can infer to $N(T_1) < N(T_2)$, which means all nodes in T_1 may belong to AS1, which also leads to a good convergence to our method. Hence, our method, Fast Hierarchy Clustering algorithm, will not suffer the non-convergence problem.

4. Experiment Study

In this section, we describe the network topology data we collected for experiments and the baseline methods for comparison.

4.1. Data Set Introduction

We examine how our algorithm behaves on global router-level topology. Two datasets are collected: One from the CAIDA [21] ITDK project for generating the global router-level graph; and the other from PeeringDB [22] for verifying the Router AS assignment.

4.1.1. CAIDA Router-Level Data for Generating Global Router Topology

The CAIDA's Macroscopic Internet Topology Data Kit (ITDK) is a project that aims to find all global router-level topology. The ITDK contains data about connectivity and routing gathered from a large cross-section of the global Internet [23]. At present (we are using the router-level topology data that was collected in July 2012), this ITDK release consists of four parts: (1) Two related router-level topologies; (2) router-to-AS assignments; (3) the geographic location of each router; and (4) DNS look ups of all observed IP addresses. In our study, we use part 1 for generating the global router-level topology, and part 4 for assigning every router's ports' IP addresses. Because ITDK contained 34,935,241 routers in July 2012, with a large amount of them having only one port opened, for simplicity and accuracy, we only maintain routers which have more than one port (port number ≥ 2); deleting one port router processing leads to only 2,520,154 routers and 19,291,581 edges remaining. Based on these routes and their edges, we construct the global router-level topology and map all routers' port and IP address to relate AS numbers using IP alias analysis.

4.1.2. PeeringDB IP-to-AS Ground Truth Data for Validation

PeeringDB is a database of networks that are interested in facilitating peering between networks and peering coordinators [24]. In this database, we could find the true AS that an IP may belong to. After crawling on PeeringDB, we collected 5210 pairs of IP-to-AS correspondences on 2012-02-05. Since our method is directly mapping routers to an AS, we needed to firstly map these IP addresses to Routers manually and then assign a related AS to them. We mapped the IP address to Router and Router to AS in 3 steps as follows:

- Step 1: We found routers in ITDK with port and IP addresses, containing an IP host in ground truth. After this process, there are only 617 routers that satisfy the condition;
- Step 2: We looked up the router-level topology generated before to examine whether these routers were isolated nodes or not, and we find that there are 3 of them which are solo nodes, and then we take them out;
- Step 3: We took out the current router's port and IP address, which have unknown AS information; for example, if router-A has 3 IP hosts where AS refers to AS1, AS2, and AS3 according to the IP alias analysis, but router-A's AS of ground truth is AS4 according to step 1, thus we cannot use AS1~AS3 to infer AS4, so that we have to take router-A out since lacking of other information.

During this process, we found that there are 125 of 617 routers suffering from this situation, and we have to take them out as well. After all of these three steps, there are 489 routers remaining for validation. Therefore, we take these 489 routers as ground truth.

4.2. Baseline Methods

We applied our method to perform router-to-AS mapping based on the router-level topology data, and then we use the ground truth to evaluate them. We also compared our method to some representative router AS clustering methods.

4.2.1. CAIDA Method: Election Tech + Degree Tech

Huffaker proposed a method that can find a router's current AS only based on its ports [14]. This method is mainly based on Election tech, and Degree tech can be only used when Election has failed to assign a current AS to the router. First, we present Figure 6 to see how Election Tech works: for Router A has three ports opened whose ASes are assigned as AS1, AS1 and AS2. Then Router A's AS may be assigned as AS1 as two of three ports are AS1. Secondly, a toy example of how Degree Tech works has been shown in Figure 7: Router B has three ports opened with AS number as AS1, AS2 and AS2, and in order to obtain which AS number Router B belongs to, the Degree Tech first generates an AS-level topology graph by assuming full-mesh connectivity among ASes from each

router's AS frequency matrix (Step 1), then we use this topology graph to calculate each AS's degree (Step 2). According to the basic knowledge of AS that "the AS is most likely to be the customer AS, based on similar intuition as the Customer heuristic [14] (p. 4)", the Degree Tech assigns Router B to AS1 because of it having the smallest degree value in AS topology.

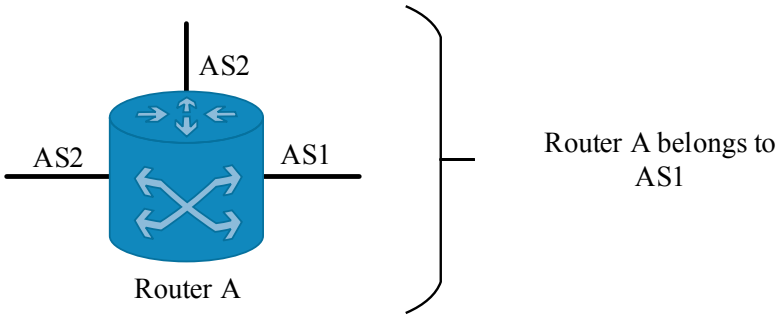


Figure 6. Toy Example of Election Tech.

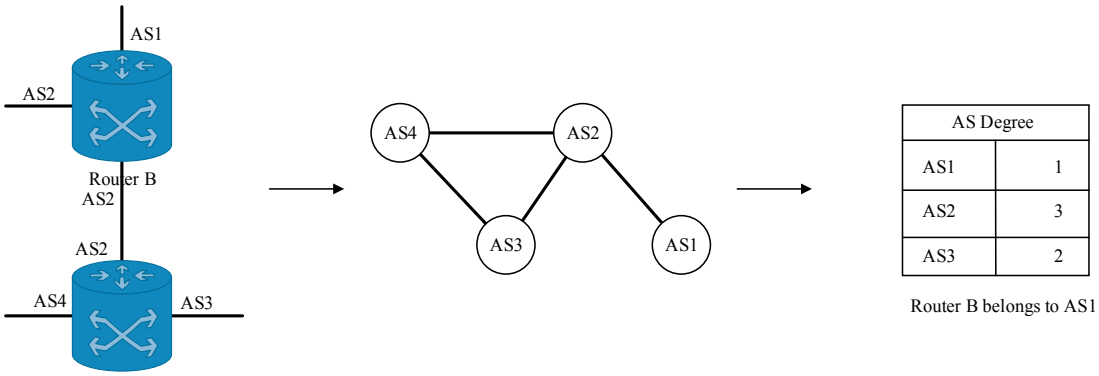


Figure 7. Toy Example of Degree Tech.

4.2.2. Hierarchy Clustering Based on Topologyonly & Hierarchy Clustering Based on Portonly

In our baseline methods, not only we use the classical and simplest router AS assignment method from CAIDA as our base line method; but also for integrality, we test our hierarchy clustering based on topology and port information only. This means we only construct our weighted router adjacent matrix with topology information or port number information. In the next, we can see that in some metric which we have discussed before, even using partial information from topology structure, the clustering results still performed better than the classical method from CAIDA.

5. Experiment Results and Discussions

In this section, we evaluate the performance of our Router-to-AS Mapping method with other baseline methods behave on global router-level topology, as discussed above, we use PeeringDB data as ground truth, and employ the clustering algorithm on the global topology based on the CAIDA ITDK project. It is easily to find out that more routers from ITDK can be assigned as real ASes, the more accurate the algorithm will be. With comparing the performance of different Router-to-AS Mapping methods, we explain the reason why our Router-to-AS Mapping method can be out-performed than other methods under such framework.

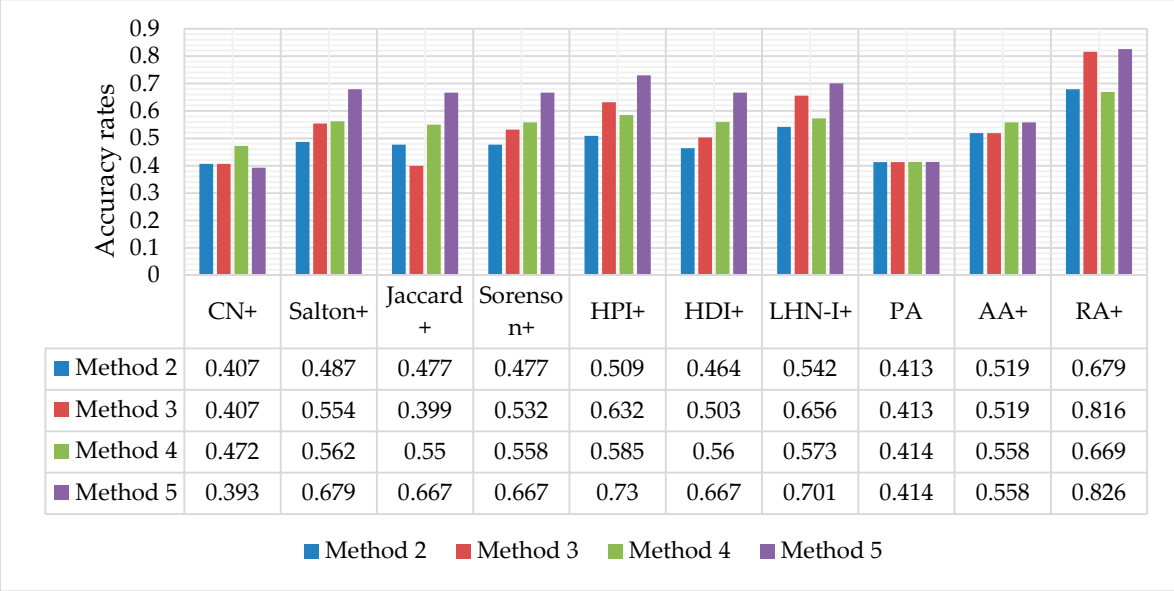
5.1. Comparison of Accuracy Rates Evaluation and Efficiency of Our Router-to-AS Mapping Method

First of all, we implement the CAIDA method (Election + Degree) to the ground truth, and find that only 320 routers are correct. It means the CAIDA method has only a 65.44% accuracy rate. Secondly, we employ the Fast Hierarchy Clustering algorithm to perform router AS assignment.

There are three kinds of information we can leverage to construct the weighted router topology network, thus we use these different information separately to form 5 validating methods as follows:

- Method 1: use the port information to generate the weight between two nodes;
- Method 2: use the traditional definition of node similarity to generate the weight;
- Method 3: use the modified node similarity metric to generate the weight;
- Method 4: combine the traditional definition of node similarity and router's port information;
- Method 5: combine the modified node similarity metric and router's port information.

For method 1, since we consider the router's port information alone, we find that only 336 routers can be assigned as real AS, so the accuracy rate is 68.71%, which is a little better than the CAIDA method. For methods 2~5, as there are 10 metrics to calculate node pair similarity, then we calculate the accuracy rate for each metric in each method as shown in Figure 8. In Figure 8, horizontal axis stands for 10 metrics, vertical axis stands for accuracy rate ranging from 0% to 90%.



	CN+	Salton+	Jaccard+	Sorenson+	HPI+	HDI+	LHN-I+	PA	AA+	RA+
Method 2	0.407	0.487	0.477	0.477	0.509	0.464	0.542	0.413	0.519	0.679
Method 3	0.407	0.554	0.399	0.532	0.632	0.503	0.656	0.413	0.519	0.816
Method 4	0.472	0.562	0.55	0.558	0.585	0.56	0.573	0.414	0.558	0.669
Method 5	0.393	0.679	0.667	0.667	0.73	0.667	0.701	0.414	0.558	0.826

Figure 8. Comparison of accuracy rates evaluation under four different validating methods.

From method 2, if we only use the traditional definition of node similarity metric, both of the accuracy rates are lower than the CAIDA method 65.44% except for the RA+ metric of which the accuracy rate increases to 67.89%, which is also lower than only using Port Information, at 68.71%. This experiment gives another proof of the incompleteness of using topology information alone without port information from initial router network.

In method 3, we have demonstrated that only using "the Modified Node Similarity" results in a tiny increment of accuracy rates, but it also suffers from the same incompleteness problem which has already been described above. It is worth noting the decreased accuracy rate while using the Jaccard+ metric. The main reason for the decrease is because there are a large number of router connections, such as in Figure 3 in the Router Global Network Topology, as there is only one connection between *routers A, B and C*. Using the traditional definition of node similarity metric, results in zero similarity between these three routers, which leads to non-combination in Fast Hierarchy Clustering, but the modified node similarity metric will never assign zero similarity between two routers. This distinguishing feature brings in some unexpected noise during the Fast Hierarchy Clustering process, and that may lead to the decreased accuracy rate using the Jaccard+ metric.

Otherwise, in method 4 we combine the traditional definition of Node Similarity and Port Information together. From the Figure 8, the performance of method 4 is only better than method 2, and it proves that only use simple information may not be good enough for Router-to-AS Mapping.

Combining the Modified Node Similarity with Port information can obtain the best accuracy rates as shown in Method 5. We can see that except for CN+, PA+, AA+ metrics, all metrics perform better than 65.44%, while using RA+ metric, the accuracy rate can reach 82.62%.

Furthermore, in order to validate the efficiency of our proposed method, we compared our method with other traditional IP-to-AS mapping approaches by the ground truth data, as shown in Figure 9. There are 404 routers have been identified correctly in 489 routers of the ground truth data by our method, which is perform better than other IP-to-AS mapping methods. Among the traditional IP-to-AS mapping methods, the CAIDA method (Election + Degree) has the best performance which only could correctly identify 320 routers.

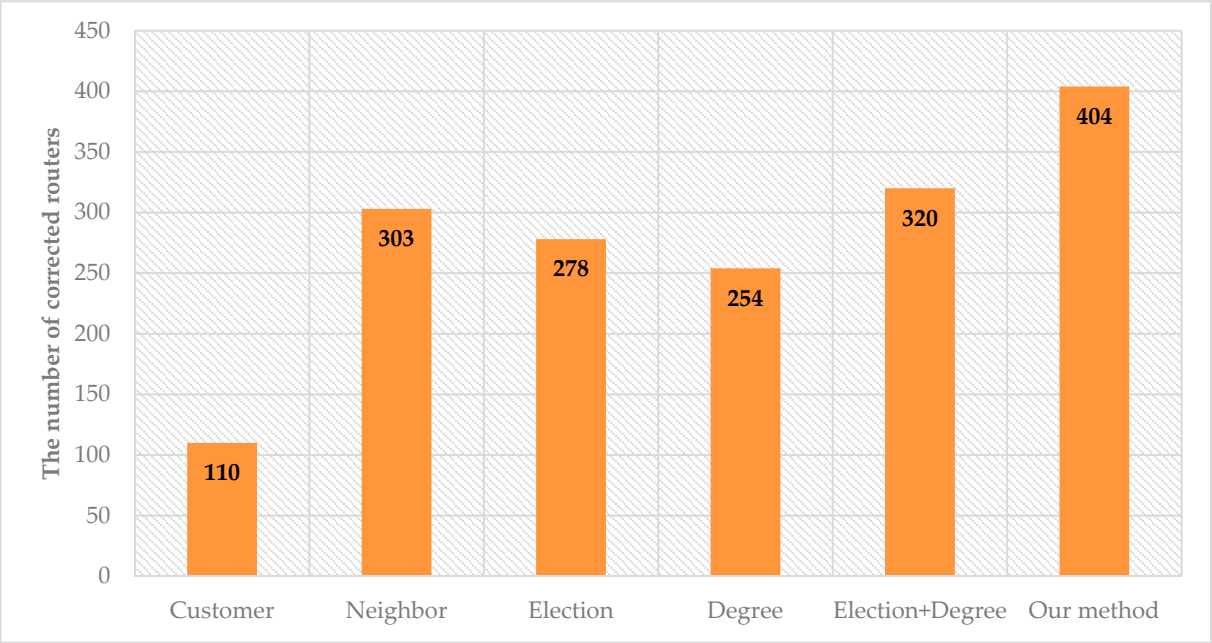


Figure 9. Comparison of the number of routers correctly identified by different router-to-AS mapping methods. There are 404 routers that were identified correctly in 489 routers of the ground truth data so that we can have the best accuracy of 82.62%.

5.2. Result Explanation for Different Similarity Metrics

After performing an analysis for each method, we attempt to explain why these different similarity metrics would acquire different accuracy rates' results. First, these 10 similarity metrics can be divided into three types according to the emphasis of features: (1) Common Neighbor as main Feature (CN+, Salton+, Jaccard+, Sorenson+ and LHN-I+); (2) Hub Node as main Feature (HPI+ and HDI+); and (3) Resource Allocation Theory (AA+ and RA+). Here we discuss these three kinds of features separately.

First of all, from the CN+ metric and metrics 2~7 according to different normalization methods, LHN-I+ obtains the highest accuracy rate 70.14%, while CN+ has the lowest accuracy rate 39.26%. The reason why CN+ exhibits the lowest accuracy rate lies in the lack of a normalization process, which leads to CN+ metric's range being beyond [0, 1]; however, the Port Information range is limited to [0, 1], and thus the low accuracy could be normal.

Secondly, the high accuracy rate of the LHN-I+ metric shows that for the global router network topology, the similarity metric of router-pairs has a reverse ratio with the router's degree, but a direct ratio with router's neighbors numbers. And for HPI+ and HDI+ metric, the accuracy rate of HPI+ performs better than HDI+, indicating that routers are more willing to connect to other routers who have higher degrees. This phenomenon can be also found in complex network analysis.

Finally, according to resource allocation theory, the mainly difference between the RA+ metric and AA+ metric is the weight increment pattern $1/K(Z)$ or $1/(\log K(Z))$. In Figure 8, the RA+ metric performs better than AA+, which is another powerful proof of Occam's razor principle: Entities should not be multiplied unnecessarily.

The above analysis of the experiments shows that *it is best to use combination of the Modified Node Similarity and Port Information to depict router-pairs similarity features*. This means that when employing community discovery methods in some particular fields, we must combine the traditional community discovery with some specific features. In this paper, we adopted these guidelines to come up with a new framework to deal with the router-to-AS Mapping problem, which combines the Port Information feature from the traditional router-to-AS Mapping technique, with the Modified Node Similarity metric from community discovery. Through the discussion of the above, it can be seen that these two features are indispensable.

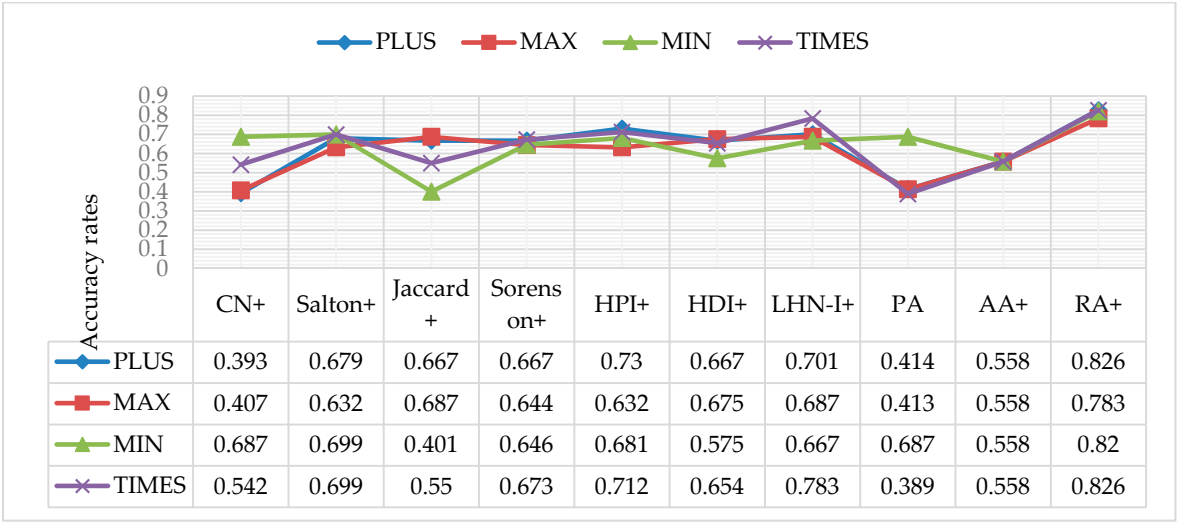
In addition, it was clearly observed that when we apply fast hierarchy clustering on routers, we not only used routers that have only one port AS number as ground truth, but also used routers that had been assigned before, which may lead to a problem: Do these routers' sequences influence the results? To answer this question, we randomly resort these routers' sequences 10 times, and apply our clustering algorithm on each of them. Table 2 shows the average accuracy rate and standard error based on random sorted routers' sequences. In this table, our Fast Hierarchy Clustering algorithm has a strong robustness and stability as almost every standard error is less than 0.01.

Table 2. Accuracy rate and Standard Error based on random sorted routers' sequences.

Metrics	Average Accuracy Rate	Standard Error
CN+	38.79%	0.0053
Salton+	67.95%	0.0010
Jaccard+	66.69%	0.0006
Sorenson+	66.83%	0.0012
HPI+	73.05%	0.0008
HDI+	66.73%	0.0009
LHN-I+	69.33%	0.0000
PA	39.53%	0.0103
AA+	55.83%	0.0000
RA+	82.62%	0.0000

5.3. Comparison of Accuracy Rate under Four Different Operators

Furthermore, there remains one question about how to combine the Node Similarity and Port information together, in order to acquire more accurate result. Here we compared the clustering accuracy rate of router-to-AS, by generating the weighted topology using 4 different operators mentioned in Section 3: We employed "PLUS", "MAX", "MIN", and "Times", respectively to generate a weighted topology, combined with the two characteristics. The experimental result of the accuracy rate comparison of router-to-AS mapping, under 4 operators, is shown in Figure 10. We can conclude that the "Times+RA+" and "Plus+RA+" share the same highest accuracy rate reaching 82.62%, which again proves that the Port Information and Modified Node Similarity metric are indispensable.



	CN+	Salton+	Jaccard +	Sorens on+	HPI+	HDI+	LHN-I+	PA	AA+	RA+
PLUS	0.393	0.679	0.667	0.667	0.73	0.667	0.701	0.414	0.558	0.826
MAX	0.407	0.632	0.687	0.644	0.632	0.675	0.687	0.413	0.558	0.783
MIN	0.687	0.699	0.401	0.646	0.681	0.575	0.667	0.687	0.558	0.82
TIMES	0.542	0.699	0.55	0.673	0.712	0.654	0.783	0.389	0.558	0.826

Figure 10. Accuracy rates comparison under different operators.

6. Conclusions

In this paper, we proposed a novel method for router-to-AS mapping. By drawing on the study of network science, our method mainly combines router-to-AS mapping techniques with community discovery, in order to obtain higher accuracy. The biggest advantage of this work is in showing a way to combine irrelevant information from different research domains to achieve better performance. In our work, we first updated the node similarity definition from $\Gamma(\cdot)$ to $\Gamma^{+}(\cdot)$, in order to avoid unnecessary noise, and then developed a novel fast hierarchy clustering method to perform Router-to-AS mapping under a massive dataset (global router topology). The results show that, by using the RA+ metric and AS frequency to quantify node-pair similarity weights, and IP host port information weights to generate the weighted topology, through our Fast Hierarchy Clustering algorithm to assign AS to routers, the accuracy rate can reach to 82.62% with 0 stand error, which is better than "Election + Degree" method, which can only reach 65.44% introduced by CAIDA. Furthermore, we compare four basic operators "Plus", "Times", "Max", and "Min" by performing weight fusion to uncover how these two irrelevant weights affect each other, the result shows that both "Plus" and "Times" can obtain the best accuracy rate of 82.62%, which means these two irrelevant information sources are weighted equally. Finally, we also use the experimental results to prove that this novel framework of Router-to-AS Mapping method has strong robustness.

Author Contributions: Conceptualization, H.H. and G.H.; methodology, H.H., W.L. and G.H.; experiment validation, W.L., G.F. and S.Y.; algorithm design, H.H. and W.L.; data collection and processing, G.F. and S.Y.; paper writing, H.H. and W.L.

Funding: This research is supported by the National Natural Science Foundation of China (No. 61471101, No. 61301274).

Acknowledgments: We are grateful to Mrs. Qing Jiang for her knowledge on IP-AS mapping and providing us useful advice in improving this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAIDA	Center for Applied Internet Data Analysis
ITDK	Internet Topology Data Kit
BGP	Border Gateway Protocol
AS	Autonomous System

References

1. He, Y.; Siganos, G.; Faloutsos, M.; Krishnamurthy, S. Lord of the Links: A Framework for Discovering Missing Links in the Internet Topology. *IEEE/ACM Trans. Netw.* **2009**, *17*, 391–404. [CrossRef]
2. Oliveira, R.; Pei, D.; Willinger, W.; Zhang, B.; Zhang, L. The (In)Completeness of the Observed Internet AS-level Structure. *IEEE/ACM Trans. Netw.* **2010**, *18*, 109–122. [CrossRef]
3. Keys, K.; Hyun, Y.; Luckie, M.; Claffy, K. Internet-scale IPv4 alias resolution with MIDAR. *IEEE/ACM Trans. Netw.* **2013**, *21*, 383–399. [CrossRef]
4. Gunes, M.H.; Sarac, K. Resolving Anonymous Routers in Internet Topology Measurement Studies. In Proceedings of the 27th Conference on Computer Communications (INFOCOM 2008), Phoenix, AZ, USA, 13–18 April 2008; pp. 1076–1084.
5. Rekhter, Y.; Katz, D.; Mathis, M.; Yu, J.Y.; Honig, J.C. Application of the Border Gateway Protocol in the Internet. 1990. Available online: https://buildbot.tools.ietf.org/html/rfc1164 (accessed on 22 February 2019).
6. Mao, Z.M.; Rexford, J.; Wang, J.; Katz, R.H. Towards an Accurate AS-Level Traceroute Tool. In Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Karlsruhe, Germany, 25–29 August 2003; pp. 365–378.
7. Tangmunarunkit, H.; Doyle, J.; Govindan, R.; Willinger, W.; Jamin, S.; Shenker, S. Does AS Size Determine Degree in AS Topology? *ACM SIGCOMM Comput. Commun. Rev.* **2001**, *31*, 7–10. [CrossRef]
8. Tangmunarunkit, H.; Govindan, R.; Shenker, S.; Estrin, D. The Impact of Routing Policy on Internet Paths. In Proceedings of the IEEE INFOCOM 2001, Conference on Computer Communications, Twentieth Annual Joint Conference of the IEEE Computer and Communications Society, Anchorage, AK, USA, 22–26 April 2001.
9. Pansiot, J.J.; Mérindol, P.; Donnet, B.; Bonaventure, O. Extracting Intra-domain Topology from mrinfo Probing. In Proceedings of the 11th International Conference on Passive and Active Measurement, Zurich, Switzerland, 7–9 April 2010; pp. 81–90.
10. Claffy, K.; Fomenkov, M. *Supporting Research and Development of Security Technologies through Network and Security Data Collection*; University of California San Diego: La Jolla, CA, USA, 2018.
11. Archipelago Measurement Infrastructure (Ark). Available online: http://www.caida.org/projects/ark/ (accessed on 22 February 2019).
12. Marder, A.; Smith, J.M. MAP-IT: Multipass accurate passive inferences from traceroute. In Proceedings of the 2016 Internet Measurement Conference, Santa Monica, CA, USA, 14–16 November 2016; pp. 397–411.
13. Gregori, E.; Improta, A.; Lenzini, L.; Rossi, L.; Sani, L. A novel methodology to address the internet as-level data incompleteness. *IEEE/ACM Trans. Netw.* **2015**, *23*, 1314–1327. [CrossRef]
14. Huffaker, B.; Dhamdhere, A.; Fomenkov, M. Toward Topology Dualism: Improving the Accuracy of AS Annotations for Routers. In *Passive and Active Measurement (PAM)*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 101–110.
15. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47. [CrossRef]
16. Loe, C.W.; Jensen, H.J. Comparison of communities detection algorithms for multiplex. *Phys. A Stat. Mech. Its Appl.* **2015**, *431*, 29–45. [CrossRef]
17. Newman, M.E.J. The structure and function of complex networks. *SIAM Rev.* **2003**, *45*, 167–256. [CrossRef]
18. Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818. [CrossRef] [PubMed]
19. Fabian, B.; Ghazaryan, Z.; Ermakova, T. Internet Connectivity of Financial Services—A Graph-Based Analysis. 2018. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213204 (accessed on 22 February 2019).
20. Zhou, T.; Lü, L.; Zhang, Y.C. Predicting missing links via local information. *Eur. Phys. J. B Condens. Matter Complex Syst.* **2009**, *71*, 623–630. [CrossRef]
21. Center for Applied Internet Data Analysis. Available online: http://www.caida.org/home/ (accessed on 22 February 2019).
22. PeeringDB Database. Available online: https://www.peeringdb.com/ (accessed on 22 February 2019).

23. ITDK Data. Available online: http://data.caida.org/datasets/topology/ark/ipv4/itdk/2012-07/ (accessed on 22 February 2019).
24. PeeringDB Resources. Available online: https://docs.peeringdb.com/faq/ (accessed on 22 February 2019).