Improving the Accuracy of the Internet Cartography

Vasileios Giotsas

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

November 21, 2014

I, Vasileios Giotsas, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

As the global Internet expands to satisfy the demands of the ever-increasing connected population, profound changes are occurring in its interconnection structure. The pervasive growth of IXPs and CDNs, two initially independent but synergistic infrastructure sectors, have contributed to the gradual flattening of the Internet's inter-domain hierarchy with primary routing paths shifting from backbone networks to peripheral peering links. At the same time the IPv6 deployment has taken off due to the depletion of unallocated IPv4 addresses. These fundamental changes in Internet dynamics has obvious implications for network engineering and operations, which can be benefited by accurate topology maps to understand the properties of this critical infrastructure.

This thesis presents a set of new measurement techniques and inference algorithms to construct a new type of semantically rich Internet map, and improve the state of the art in Internet cartography. The author first develops a methodology to extract large-scale validation data from the Communities BGP attribute, which encodes rich routing meta-data on BGP messages. Based on this better-informed dataset the author proceeds to analyse popular assumptions about inter-domain routing policies and devise a more accurate model to describe inter-AS business relationships. Accordingly, the thesis proposes a new relationship inference algorithm to accurately capture both simple and complex AS relationships across two dimensions: prefix type, and geographic location. Validation against three sources of ground-truth data reveals that the proposed algorithm achieves a near-perfect accuracy. However, any inference approach is constrained by the inability of the existing topology data sources to provide a complete view of the inter-domain topology. To limit the topology incompleteness problem the author augments traditional BGP data with routing policy data obtained directly from IXPs to discover massive peering meshes which have thus far been largely invisible.

Acknowledgements

This thesis would have not been possible without the help and influence of many people. First of all I thank my Ph.D. supervisor, Shi Zhou, for his excellent guidance and encouragement throughout my Ph.D. period. Shi Zhou has been a very patient advisor and a source of intellectual stimulation and scientific inspiration.

I thank Steve Hailes who served as an examiner of my Ph.D. upgrade viva. His insightful comments offered me different perspectives that I utilized throughout my research. I also feel very fortunate for the chance I had to work with the leading research group in Internet data analysis, CAIDA, and I especially thank Matthew Luckie, Kc Claffy and Bradley Huffaker. My collaboration with them helped me greatly to become a better researcher and develop a deep appreciation for their rigorous scientific approach on Internet measurements.

I was very lucky to share the 7th floor offices of the UCL Malet Place Engineering building with many great friends and colleagues. Lynne Salameh, Tamas Jambor, Jagadeesh Gorla, Dominik Beste, Martin Parsley, Afra Mashadi, and Clovis Chapman were all a fantastic company with whom I enjoyed my breaks from work (and many games of foosball), and provided great advice and feedback on my papers. Jie Xiong and Manos Protonotarios made sure I won't stay alone in the lab during the many nights I spent there.

I also need to express my gratitude to Yiannis Andreopoulos who helped decisively towards the completion of this thesis in many different ways.

I'm grateful to my parents for their unconditional and multidimensional love and support without which it would be impossible to reach my academic goals.

Finally, I am thankful to my beloved fiance, Evangelia Lymperopoulou, for her tireless support and for sharing my burden of doing a Ph.D. She has been the most consistent source of assurance and she never failed to cheer me up whenever pressure increased. The period when she did her Master in Cultural Policy and Management at

City University was the most enjoyable and productive period of my Ph.D. because we were studying together.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The Internet is a composed by more than 45,000 independent networks, called Autonomous Systems (ASes), that interconnect to form the most widespread global communications infrastructure. It is forecasted that by 2018 the number of Internet users will correspond to over 50% of the world's population, and more than 20 billion devices will be connected online [60]. An AS can be an Internet Service Provider (ISP), a Content Distribution Network (CDN), or a smaller organisation such as a university or corporation that autonomously administers a domain of connected IP prefixes. Packets within an AS are routed according to a set of metrics and Interior Gateway Protocols (IGPs) that are determined by each AS operator separately and can differ significantly between ASes [123]. Each AS owns only a subset of the IP address space and typically covers a limited geographical area, which means that end-to-end traffic often needs to traverse multiple AS domains before reaching its destination. In inter-domain routing the Border Gateway Protocol (BGP) is used as the de-facto protocol for the exchange of reachability information at the boundary of ASes [175]. Before starting to exchange traffic two ASes need first to establish a physical connection and agree to a contractual relationship that determines the economic and technical aspects of their connectivity. Business relationships between ASes, can be broadly classified into two types: customer-to-provider (c2p) or transit and peer-to-peer (p2p). In a c2p relationship, the customer pays the provider for traffic sent between the two ASes. In return, the customer gains access to the ASes the provider can reach, including those which the provider reaches through its own providers. In a p2p relationship, the peering ASes

gain access to each others customers, typically without either AS paying the other. Peering ASes have a financial incentive to engage in a settlement-free peering relationship if they would otherwise pay a provider to carry their traffic, and neither AS could convince the other to become a customer.

BGP provides the flexibility to express routing policies on how reachability information is propagated, allowing AS operators to enforce their contractual agreements and implement complex traffic engineering techniques for load and cost balancing. As a result of policy-based routing inter-domain traffic does not necessarily follow the shortest path between two ASes. Policy-based routing has been one of the initial design aspects of BGP aiming to enable AS operators to chose which routes will be accepted, which will be preferred and which will be propagated to their neighbours.

1.2 Motivations for Internet Cartography

Internet cartography describes the research efforts to discover, annotate and characterise the Internet topology through direct measurements and inference techniques.

The need for Internet cartography stems from the fact that today there does not exist any map that fully describes the various levels of Internet connectivity due to the highly distributed ownership and administration of the Internet's infrastructure. Each AS has full knowledge and control of their own domain but treat the other ASes as black boxes. This is another aspect of BGP which was designed to hide any information regarding the internal structure of ASes. Consequently, nobody has global visibility of the Internet's topology since 1995, when the National Science Foundation Network (NSFNET) backbone was decommissioned and gave its place to the "commercial Internet" [198].

The field of Internet cartography gained significant attention due to the necessity to study the structure and dynamics of the Internet topology for the design of future Internet architectures, inform technology investment, facilitate the development of public policies, and enable the scientific analysis of the Internet from the perspectives of statistical physics and network science.

Bellow I explore in more depth how the Internet topology can be a useful tool in tackling Internet-related challenges.

1.2.1 Design of future Internet protocols

The Internet was not originally designed to be the ubiquitous global network it is today. Instead it started as a research project that evolved spontaneously and without central coordination. As the Internet pioneer Vint Cerf noted in a recent interview:

> *"I thought that if the Internet idea actually worked we would then build a production version of it. And what happened is it got loose, into use! We have been using the experimental Internet design since 1983 when we turned it on!"*

For instance, BGP was introduced in 1989 [144] to enable the transition from the centric ARPAnet architecture to the meshed NSFnet topology. The current version of BGP, BGPv4, has been adapted in 1995 and although it went through multiple extensions its designed principles remain the same for almost 20 years. Over these years the demands placed on the Internet kept changing guided by innovations in the physical and the application layers. As a result, a number of serious shortcomings have been building over time, including slow convergence, instability, weak security, non-deterministic behaviour, scalability concerns, and proneness to misconfigurations [89, 111, 151, 110, 122, 49, 81].

To respond to the emerging challenges, Internet researchers and engineers develop new architectures and protocols that can support the ever-increasing demands for higher traffic volumes, stronger security and better performance. However, operators are reluctant to adopt protocol innovations due to the difficulties involved in the transition to new technologies while ensuring the uninterrupted operation of their network and the seamless interoperability with other networks.

The problem of adopting novel protocols and architectures has been parallelised as an effort to change the engines of an aeroplane in mid-flight. The high cost of potential failures discourage the switch to new technologies leading network engineers and administrators to choose ad-hoc solutions. These are essentially temporary fixes that mitigate the short-term problems, but have limited long-term payback, as the underlay architectural problems persist [121, 32].

Simulation, emulation and virtualization have been proposed as alternatives to the limited capabilities for try-and-error experimentation [174, 32, 159]. These approaches

can greatly facilitate research efforts but require sound understanding of the Internet's large-scale structure and dynamics. Such insights can be obtained through the study of the Internet's global topology [91].

1.2.2 Development of Regulatory Policies

As the Internet ecosystem evolves the emerging challenges do not relate only with pure technical issues but also with policies to regulate money flows, competition and network neutrality.

The need to regulate the Internet business relationships is not new but becomes more pressing due to the rising market power of CDNs and Eyeball networks[1]. Today the majority of traffic which originates from content providers is delivered over CDNs at the edge of the Internet topology. These new traffic channels constitute a significant shift from the past when backbone tier-1 ASes where the major hubs for the delivery of end-to-end traffic [140, 181], with direct implications for the balance in market power. CDNs control the source of the traffic, but eyeball networks control the incoming links over which content is delivered to end users. This traffic imbalance has led to a sharp increase in interconnection disputes [165, 177, 26, 45, 205], intentional throttling of capacity, and other business practices that threaten network neutrality [142, 162]. Consequently, there are growing concerns that regulatory action is required to ensure transparency and fairness in the Internet transit market [161, 201, 1].

Policy makers can benefit from accurate topology data and models that can provide insights on interconnection strategies, facilitate the monitoring of connectivity types and help in avoiding potential unexpected effects of new regulations [71, 147].

Similarly, the study of Internet topology can provide useful insights to policy-makers who aim at stimulating the expansion of the Internet in developing countries. Recently there have been great efforts to improve the Internet infrastructure in Africa and Latin America, but without necessarily leading to the expected levels of development [96, 183]. Connectivity is a key aspect of monitoring and assessing these efforts and therefore policy-makers and investors can make better-informed decisions trough the analysis of the inter-domain topology [115, 87].

[1] *Eyeball* networks are called the ISPs that mainly operate as access providers for the end users such as home subscribers.

1.2.3 Scientific Exploration of Complex Networks

The analysis of the Internet topology has been particularly appealing among statistical physicists and mathematicians, who are mainly motivated by a scientific interest in the exploration of fundamental process that happen in complex networks.

The focus on the Internet as a complex network has been motivated by the seminal study of Faloutsos brothers that suggested that the connectivity degree of the router-level topology has a power-law shape [86]. At about the same time power-law relationships have been found in a large array of natural and artificial networks fuelling an exploration for global properties among scale-free networks processes, such as epidemiological properties, growth mechanisms, small-world phenomena [190, 193, 30, 196, 164].

The preferential attachment model [38] as a mechanism for the evolution of complex networks inspired the development of topology generators that reproduce an array topological properties of the Internet, such as a power-law degree distribution, disassortative mixing, strong clustering, and the rich club connectivity [47, 199, 156, 211, 210, 169]. Topology generators require accurate inference and annotation of the Internet topology, and need to capture the dynamics of topology evolution, in order to provide more realistic models of the actual network [119, 120].

1.3 Challenges

Despite over a decade of progress in mapping the Internet inter-domain topology many widely cited results have been characterized as inaccurate or misleading due to several open issues in Internet Cartography [179].

1.3.1 Topology Incompleteness

One of the main difficulties in mapping the inter-domain topology is the lack of readily available connectivity data. As a workaround researchers try to construct the inter-domain topology by aggregating BGP and traceroute routing data that are collected primarily for monitoring and debugging purposes. The most widely used BGP datasets are collected by two large-scale monitoring projects, RouteViews [21] and RIPE RIS [16], which connect directly to ASes to receive and publish their BGP feeds. traceroute data are collected by distributed vantage points deployed as part of data-plane monitoring infrastructures, such as CAIDA's Ark [4], DIMES [184] and RIPE Atlas [15]. However, both BGP and traceroute monitors have only limited visibility of the global inter-domain connectivity and consequently the visible inter-domain topology is incomplete even when the views of multiple vantage points are combined.

The incompleteness problem is exacerbated due to its bias against peering links at the periphery of the AS topology, and geographical areas that contain smaller concentration of BGP and traceroute monitors. Therefore, the visible topology is not only incomplete but also considerably skewed. Improving the completeness of the visible inter-domain topology is a very important research area, but despite recent progress through crowd-sourcing traceroute measurements [58, 85] and combination of multiple data sources [124, 136] the largest part of the inter-domain topology remains invisible [28].

The number of invisible links is amplified by the expansion of the IXPs that facilitate the establishment of peering relationships at the edge of the network and lead to dense but highly localised peering meshes. These peering links are difficult to be discovered even through targeted measurements [36] and it was suggested that they can be observed only by extracting data directly from IXPs [28].

1.3.2 Questionable Relationship Inference Heuristics

Inference of AS relationships has been an active area of research for over a decade [97, 202, 191, 84, 41, 76, 167, 197]. While yielding insights into the structure and evolution of the topology, this line of research is constrained by systematic measurement and inference challenges [198, 70].

Relationship inference algorithms are limited by the fact that they rely on AS connectivity information (obtained from the BGP or traceroute paths) and therefore heuristics need to be used to translate path patterns to relationship types. Past works on relationship inference used two types of techniques, Type-of-Relationship (ToR) optimisation and top-down inference. Both techniques rely on the assumption that all valid BGP paths conform to the *valley-free* export policy[2], and accordingly they assign c2p relationships to AS links. Similarly p2p relationships are assigned based on the connectivity degree of the connected ASes or the visibility of the links.

These heuristics are based on theoretical principles that have not been thoroughly validated in practice due to the lack of ground-truth data. Assuming the universality of simple BGP policies is prone to errors even if some policies are prevalent. For instance, the valley-free export policy may be violated in cases of misconfigurations or poisoning that are difficult to be detected using simple data sanitisation techniques [145, 133].

Moreover, the so-called "flattening" hierarchy of the inter-domain connectivity inhibit the effort to develop accurate heuristics, since the parallel growth of CDNs (traffic producers) and eyeball networks (traffic consumers) have radically changed the perceptions of symmetry in traffic delivery costs [88]. The implications of these Internet phenomena on inter-domain routing policies need to be properly understood, in order to improve the quality of AS relationship inferences. Ground-truth data on routing policies are needed in order to gain new insights for the development of well-grounded heuristics and the validation of the inference results.

1.3.3 Simplistic Modelling Abstraction

The layered nature of the Internet's architecture means that its topology can be described at different levels of abstraction. Accordingly, at least four levels of topology resolutions can be defined [80, 158]. The **interface-level** or **IP-level** topology describes

[2] According to the *valley-free* policy, routes learned from peers and providers are advertised only to customers. BGP policies are discussed in more detail in section 2

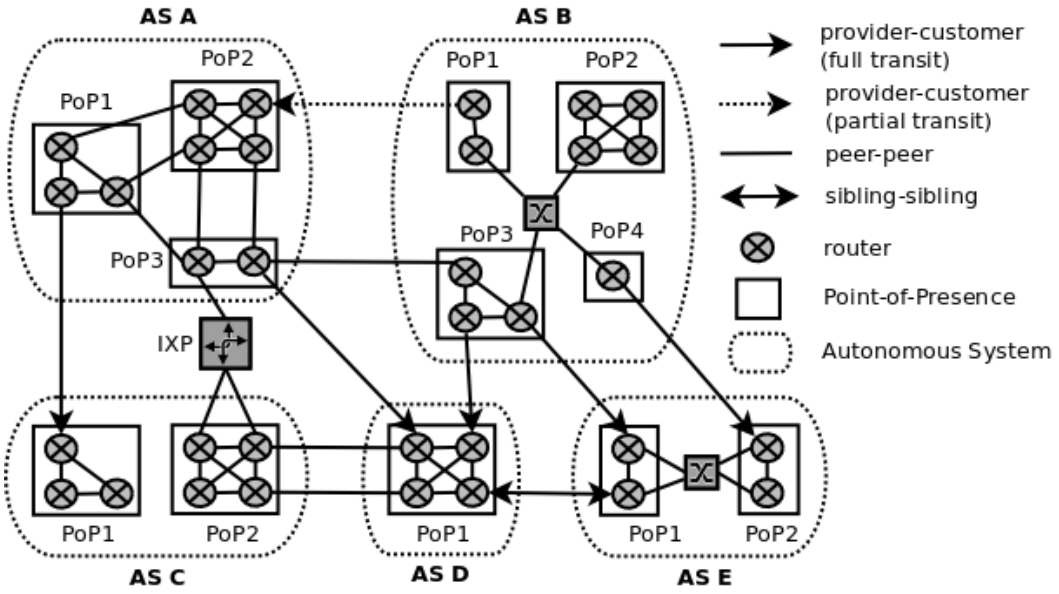


Figure 1.1: Inter-domain topology at three different levels of resolution, router-level, PoP-level and AS level.

the network layer connectivity between router interfaces. Each IP interface is a node and a physical connection between two interfaces represents a link. The **router-level** topology results by grouping the interfaces that belong to the same router in one node. Routers can be grouped in Point-of-Presence (PoPs) forming the **PoP-level topology**. PoPs are physical buildings where ISPs hosts their network devices and function as access points to their network. The PoPs that belong to the same administrative entity form an Autonomous System (AS) or a Routing Domain. In the **AS-level topology** the nodes represent Autonomous System and the links express the contractual relationships among them. Each link is a logical construct that groups together multiple router-level or PoP-level links between two ASes.

The resolution at which the topology is considered has direct implications on its applicability in the study of Internet engineering problems. Higher level of detail provides a more realistic representation of the actual network but comes at the cost of increased complexity and scarcer data sources. For this reason the majority of past topology research works have used a simplistic approach that models the inter-domain topology as a simple AS graph. This abstraction loses important details regarding the internal structure of ASes and the actual connectivity at the BGP level, making it largely irrelevant to Internet engineering problems. For example, it has been found that when used in path prediction it produces very poor results [153, 159]. Even though an AS

is supposed to exhibit a single routing policy to another AS according to the definition in RFC 1930 [123], in reality ASes often announce different prefixes at different interconnection points [48], apply different import policies [155] and can agree different relationship types [82].

To capture the diversity of routing policies the AS graph should be modelled as a multi-graph. The PoP-level topology can provide a good trade-off between fidelity and simplicity because it represents an intermediate granularity between the physical router topology and the logical AS topology [198, 163]. However, discovering the geographic footprint of an AS is non-trivial and typically it requires to be inferred through latency measurements combined with traceroute paths that can be erroneous [137].

1.4 Contributions

The main contribution of this thesis is the development of a new type of Internet map that accurately captures the inter-domain connectivity at a higher level of granularity. The data produced by the new methods will offer a fresh lens under which to study the structure and evolution of the Internet topology. The new data can facilitate better understanding of the economics and policies of inter-domain topology and improve the scientific analysis of its statistical properties. More specifically the contributions of this thesis can be summarised as follows:

1.4.1 Augmented Measurement Datasets

BGP data contain a wealth of routing information but so far related research relied almost exclusively on the AS path attribute. I augment the connectivity information obtained from AS paths with metadata extracted from the Communities BGP attribute and operators' documents that explain their routing policies. The augmented measurements provide a unique ground-truth dataset that covers over 35% of the visible AS links and allows me to develop inference and validation techniques that do not rely solely on connectivity data but also on actual policy configurations. To ensure the correctness of the collected data I have cross-validated the BGP Communities against feedback directly received by AS operators that covers about 10% of the visible AS links.

To complement the data provided by RouteViews and RIPE RIS BGP feeds I integrate BGP and traceroute data obtained through more than 500 of distributed looking glass servers. Looking glass servers provide interfaces to routers that allow the execution of non-privileged commands and therefore they can be used for targeted active querying. Different looking glass servers expose different interfaces through which they accept queries and present the output. To facilitate querying I develop an overlay API that abstracts querying and parsing, and enables the transparent execution of complex queries across diverse looking glasses. Although some past works also utilised looking glasses to obtain additional connectivity information[36, 136], I implemented my own querying overlay for two reasons: (i) The source code used in the past works for querying the looking glasses has not been shared publicly, only the list of the looking glass servers[33, 135]; (ii) Looking glass servers are highly volatile and often existing looking glasses change querying parameters, become unresponsive or unavailable,

while new looking glasses become available. Therefore past lists of looking glasses become quickly stale and of limited use. Instead of relying on hard-coded lists I automated the discovery of new looking glass servers through web crawling and web scrapping.

1.4.2 Validation of Inter-domain Modelling Heuristics

I apply the augmented datasets in the analysis of inter-domain routing data to gain new insights on import and export policies. I use the ground-truth data collected from BGP Communities to thoroughly analyse popular modelling heuristics and validate the existing relationship inference algorithms.

My results reveal that popular heuristics such as the valley-free "rule" are violated at least twice as often than previously believed. More importantly, a large fraction of these violations are not a result of transient misconfigurations but rather the outcome of complex policies that cannot be captured by coarse-grained models [104].

Additionally I find that the assumptions on the symmetry of peering and transit relationships based on the connectivity degree do not provide a realistic model of AS connectivity. Consequently, they lead to higher number of errors compared to the algorithms that do not utilise topological properties to infer AS relationships [103].

The Community values allow to study the IPv6 AS relationships, despite the considerable "noise" in IPv6 paths due to the early stage of its deployment. The results reveal a large number of links with different relationship types between the two topologies [102]. This work was one of the first to suggest that IPv6 AS relationships should be studied separately from IPv4 since the economics are completely different [61].

1.4.3 New Measurement and Inference Algorithms

The insights gained from the ground-truth data allow to rethink the problems of topology inference and annotation. Development of more realistic relationship inference heuristics and path sanitisation techniques contribute to the inference of conventional AS relationships with near-perfect accuracy (99.6% for transit links and 98.9% for peering links) [146]

To capture the complex AS relationships at the PoP-level granularity I build a new algorithm to infer hybrid and partial transit relationships with 95% accuracy. This new algorithm is a first step towards tackling a long-standing problem which has been

characterized as the "holy-grail" of relationship inference [163]. To achieve this breakthrough I develop a measurement technique that orchestrates active BGP and traceroute measurements based on hints provided by passively collected BGP data in order to obtain fine-grained connectivity without increasing exponentially the measurement cost.

Finally, I develop a new algorithm to infer with high confidence the peering links established over IXP route servers, that reveal a large fraction of the "invisible" AS topology. Route servers provide a scalable way to implement dense peering connectivity with small administrative effort and therefore they are one of the key technologies that contribute to the "flattening" of the AS topology [57]. By combining IXP presence data with route redistribution BGP filters I am able to reveal 210% more peering links compared to the combined publicly available BGP and traceroute data [105, 106].

1.4.4 New Knowledge on Inter-domain Connectivity and Policies

The new inference algorithms and measurement techniques allow to study the inter-domain connectivity and policies through a new lens. The new relationship inference algorithms informs our understanding of evolutionary trends such as the flattening of the Internet topology and the financial consolidation of the Internet transit industry.

The inference of fine-grained relationships reveals that complex relationship types are more widespread among European ASes, mainly due to the highly expanded IXP ecosystem. IXPs are driving the evolution of complex relationships as they allow content providers and eyeball ASes to interconnect at multiple countries and apply different policies for each country.

The last observation is also apparent by the inferred multilateral peering agreements. ASes that co-locate at different IXP route servers apply different sets of BGP filters, meaning that the peering policy of a network is frequently location-specific. Consequently, the self-reported peering policies should not be taken at face value. Instead my data enable a closer inspection of the actual policies and can be used to gain insights on peering policies at higher granularity.

1.4.5 New Public Datasets

I make publicly available all the datasets produced by the research presented in this thesis[3]. Overall four different datasets have been published:

- Ground-truth data on AS relationships and BGP routing policies obtained through BGP Communities for April 2012.
- AS relationship inferences for conventional relationships types between January 2003 and November 2013.
- AS relationship inferences for complex relationship types for March 2014.
- Multilateral peering links for 18 large European IXPs for February 2013.

The above datasets are accompanied by appropriate documentation and they are available either directly from myself or through CAIDA's online data repository. Overall the data have been downloaded by more than a hundred different researchers.

1.5 Thesis Outline

The rest of the thesis is organised as follows. Chapter 2 provides an overview of the preliminaries of inter-domain routing and explains the different BGP attributes and how they are used in the decision process for the selection of the active AS paths. I also present a literature review about the related works on topology discovery and relationship inference. The section concludes by discussing the open questions that this thesis aims to research.

Chapter 3 presents a new approach to augment the available connectivity datasets with data extracted from the BGP Communities and the LocPref attributes, and describes a new measurement framework to automatically interpret and sanitise the encoded values. The rest of the chapter presents how the collected ground-truth data are used to perform a preliminary study of the AS topology. The results of this study shed new light on the validity of popular heuristics used in modelling inter-domain routing. More specifically, I provide a study of the valley-free violations in IPv4 and IPv6 paths, the Gao-Rexford conditions on the ordering of Local Preference values, and the correlation between different relationship types and the connectivity degree of ASes. Chapter

[3]The only data that have not been made available are the direct feedback received from AS operators due to non-disclosure agreements with them.

4 concludes by presenting a validation of the accuracy of past relationship inference algorithms.

Chapter 4 presents how the insights gained from the previous chapter are combined to develop a new algorithm for the inference of conventional relationships. The chapter provides details on each step of the algorithm and the efforts to thoroughly validate the inferences. I then explain how the inferred relationships are used to construct the customer cones of the ASes as a metric of their market power and the flattening of the AS topology. relationships is decreasing over time, but the differences in the top transit providers remain notable.

Chapter 5 presents how the conventional inference algorithm presented in Chapter 4 is extended to infer two important types of complex AS relationships at the PoP-level topology granularity: hybrid and partial-transit. Then, it provides an analysis of the findings and validates the inferences against three sources of data, direct feedback, BGP communities and RPSL objects.

Chapter 6 introduces a novel algorithm for the inference of multilateral peering links. First I introduce the role of Route Servers in Internet peering agreements and I explain how BGP filters are used to control the multilateral peering reachability. Then I present in detail the link inference algorithm that reveals more than 180K "invisible" peering links. The chapter continues by providing a study of the observed peering policies of the Route Server participants. Based on the inference results the chapter concludes by extrapolating the total number of peering links globally to put into perspective the incompleteness problem.

Finally, Chapter 7 concludes the thesis with an overview of the research outcomes and how the contributions of my work has so far facilitated further research in inter-domain routing. The thesis concludes with a discussion on the future directions of the Internet inter-domain cartography.

1.6 Chapters Material

Most materials presented in this thesis have been published in peer-reviewed conferences and journals. This section explains how these publications are related to the chapters of this thesis. Chapter 3 is based on material published in [103, 102, 104]. Chapter 4 is based on material published in [146]. Chapter 5 is based on material pub-

lished in [101]. Chapter 6 is based on material published in [105, 106]. In all the papers that I am the first author I personally conducted all the research and technical work presented in the publications.

The only paper where I contributed as supporting author is [146], in which my contributions are limited in section 2 (related work), sections 3.5, 3.6 (collection and sanitisation of validation data extracted from BGP communities), and sections 4.6, 4.7 (validation and debugging of inference heuristics; analysis of complex relationships; comparison against past relationship inference algorithms). The algorithm itself has been designed and developed by Matthew Luckie. I did not contribute in the section 5 (inference of customer cones and topology flattening), and therefore inference and analysis of customer cones is not part of this thesis's contributions. Section 4.4 shortly presents customers cones to explain a practical application of relationships inference.

Chapter 2

Background

In this chapter I introduce the terminology and preliminaries of inter-domain routing, and I review the relevant literature in inter-domain cartography. The chapter is divided in three parts. The first part explains the different BGP attributes and how they are used to determine which AS path will be selected for routing traffic towards an IP prefix. Then I present the different sources of publicly available topology data and I explain their limitations. These two sections provide the necessary background for understanding the techniques presented in the next chapters. Finally, I review the related works on the discovery and annotation of the inter-domain links and I discuss the open research issues.

2.1 Introduction

The Internet is composed by millions of routers distributed all over the world. Routers receive traffic from end hosts (e.g. web servers) and decide how it should be forwarded to the destination hosts (e.g. Internet users) based on the unique identifiers of the sender and the receiver. These identifiers are expressed as unique numbers called IP addresses. The length of the IP addresses depend on the IP version, IPv4 uses 32-bit addresses while IPv6 uses 128-bit addresses. Usually traffic travels across many routers before reaching its destination. To decide the next hop in the routing path, each router keeps a table that maps IP addresses to neighbour routers. Due to the huge number of possible IP address (2^{32} for IPv4 addresses) such a table is not scalable. To make the routing table more compact, IP addresses are grouped into prefixes of common bits. For example all the addresses that have the first 16 bits in common are group into a prefix of the format a.b.0.0/16 (CIDR format). The prefix can be of variable length

depending on the number of common bits.

Routers are deployed and administered by large organizations such as Internet Service Providers (ISPs), enterprises or educational institutions. A network of routers under the same administrative entity is called *Autonomous System (AS)*. Each AS is identified by a unique 32-bit number (ASN) and has been assigned with one or more address blocks (IP prefixes) to distribute to the devices connected to its infrastructure. The *Internet Assigned Numbers Authority (IANA)* is responsible to assign ASNs and IP prefixes to the the *Regional Internet Registries (RIRs)* which in turn assign them to ASes. Today there are five RIRs (Afica: AFRINIC, America: ARIN and LACNIC, Asia and Australia: APNIC,Europe: RIPENCC). On June 2014 there were 65,592 allocated ASNs, 46,726 of which were actively advertised in the BGP routing table [24].

The ASes are autonomous in the sense that their operators can independently decide which *Interior Gateway Protocol (IGP)* and metrics will be used for routing inside their domain, irrespective of what protocols other ASes use. Routing inside an AS is a black box for other ASes which are only interested in which IP prefixes are reachable through it. To exchange routing information a common *External Gateway Protocol (EGP)* should be used across all the ASes. Today this protocol is *Border Gateway Protocol (BGP)* [175] and the routers that use it to link different ASes are called *border routers* or *BGP speakers*. Figure 2.1 illustrates some of the above concepts.

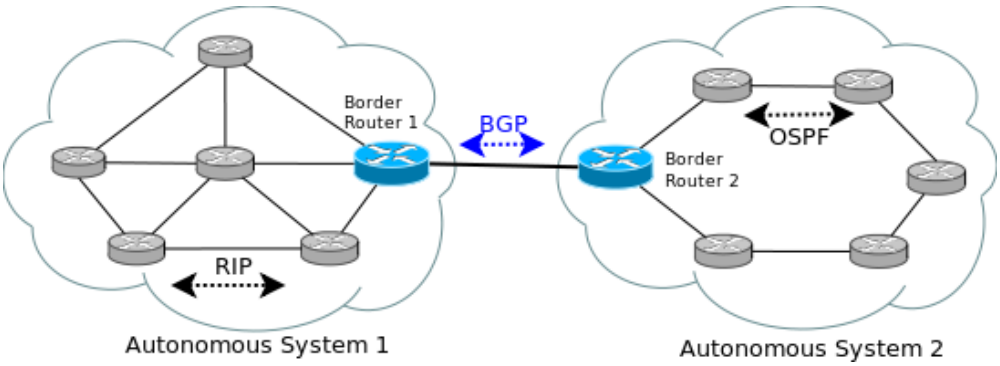


Figure 2.1: The difference between router-level and AS-level Internet topology. From the AS perspective there are only two nodes - AS1 and AS2 - and one link.

2.2 Basic Operation of BGP

Before the exchange of routing information two border routers should establish a BGP session that starts with the exchange of BGP OPEN and KEEPALIVE messages

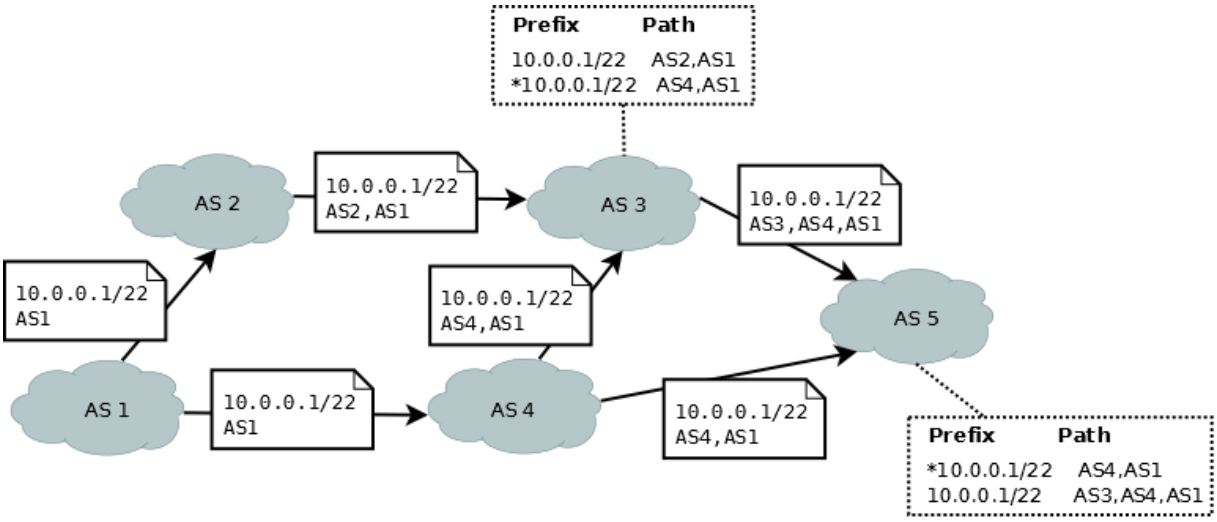


Figure 2.2: Propagation of reachability information using BGP. AS1 originates and advertises the IP prefix 10.0.0.1/22 to its neighbours AS2 and AS4 with an associated path vector $< AS1 >$. Both AS2 and AS4 receive the same prefix advertisement and forward it to their neighbours by prepending their own AS number in the path vector. AS3 will receive two advertisements from the same prefix, one from AS2 and one from AS4. It will select the most preferable based on a routing policy that orders the paths and it will advertise only the selected path to AS5. Similarly, AS5 will receive two advertisements for the prefix 10.0.0.2/22, one with path vector $< AS3, AS4, AS1 >$ and the second with path vector $< AS4, AS1 >$. Accordingly it will select its preferable path based on its own policy decision.

to identify themselves and to verify that they still exist and are engaged in the relationship. To maintain the session in established mode KEEPALIVE messages should be exchanged periodically otherwise the session changes to idle state. After the session is established reachability information can be announced using BGP UPDATE messages. There are two types of UPDATE messages, advertisements and withdrawal. When an AS wants to advertise reachability information towards an IP prefix it sends an advertisement that contains the IP prefix of the advertised network and a set of attributes (shown in table 2.1). Similarly, when an AS wants to announce that it no longer offers reachability towards an IP prefix it previously advertised, it sends a withdrawal message that includes the IP prefix of the withdrawn network.

A BGP advertisement contains two important attributes, the IP prefix for which reachability is advertised, and an associated AS Path which is a vector of ASN values that indicates the order of ASes that needed to be traversed to reach the prefix. BGP does not determine the IP-level path that traffic should follow and it does not reveal information on router-level topology, it only maintains routing information at the AS-

level. When an AS learns a path to a prefix it can propagate it to its neighbours. Each time an AS advertises a prefix it prepends its ASN in the AS Path, so that each AS that learns a prefix advertisement it receives an updated path vector with the ASes traversed thus far. This mechanism is used to avoid routing loops because an AS will reject AS paths that already include its local ASN.

Figure 2.2 illustrates how a BGP advertisement is propagated along multiple ASes. Multi-homing ASes, namely ASes that have more than one neighbours, may receive BGP advertisements for a given IP prefix from different neighbours. For example, in figure 2.2 AS 3 receives BGP advertisements for prefix 10.0.0.2/22 from both AS 2 and AS 4. In this case a single path should be selected as the default route towards that prefix and be propagated to the adjacent ASes of AS 3. The selection of the most preferable path is typically based on the local policy of each AS.

Table 2.1 shows a list of the BGP attributes included in a BGP UPDATE message.

Attribute Name	**Description**
ORIGIN	Specifies the origin of the path information. Indicates whether the path was learned from an IGP or EGP.
AS_PATH	The list of ASes that should be traversed to reach the advertised IP prefix.
NEXT_HOP	The IP address of the border router that should be used as the next hop.
MULTI_EXIT_DISC	When two ASes are connected by multiple links MED indicates which link is preferable.
LOCAL_PREF	The degree of preference for a particular route. It is used in communication between border routers within the same AS.
ATOMIC_AGGREGATE	When a border router receives IP prefixes advertisements that overlap, it can use the shorter prefix for all the overlapping prefixes. In that case this attribute is set to 1 to indicate that aggregation has been done.
AGGREGATOR	The ASN and BGP ID of the router that performed aggregation.
COMMUNITIES	Extra attributes that can influence the propagation or selection of a route. There is no standardized use of these attributes and their values are defined by the network administrations.

Table 2.1: The attributes that can be included in a BGP UPDATE message. Only the first three are mandatory. The LOCAL_PREF is required only for messages between border routers of the same AS.

The next sections describe how import and export policies can be expressed through the configuration of BGP attributes.

2.3 BGP Policy Configuration

There are two stages at which BGP policy configuration is applied, when a new route is received, and when a route is advertised. *Import* policies determine how the ingress route advertisements will be processed to decide if a route will be accepted or rejected, and to rank all the available paths towards a given prefix. *Export* policies determine which neighbours can receive an advertisement for a given prefix.

There are three types of expressions that can be used to implement the above policies:

2.3.1 Filtering

Filtering instructs a router to block ingress advertisements or suppress egress advertisements by matching certain BGP attributes against some pre-defined values.

Import filters are typically used to mitigate security threats and eliminate erroneous routes [37]. For instance, a good practice is considered to block path advertisements for reserved and unallocated prefixes [67, 44], which are often used to facilitate Denial-of-Service attacks [90].

Export filters are used to implement selective (or conditional) prefix advertisements, meaning that AS operators can select which subset of their neighbours will receive an advertisement and which will be excluded. Selective advertisements are the main apparatus through which ASes enforce their business relationships through the so-called valley free rule (explained in section 2.4). Also, load balancing and other outbound traffic engineering techniques are implemented through filtering BGP exports. Of course, export filters are also used to limit the propagation of "junk" routes. Good BGP "citizenship" requires that routes expected to be blocked by import filters are not supposed to be advertised at all in the first place.

2.3.2 Ranking

ASes with multiple neighbours may receive multiple different route advertisements for the same prefix. BGP requires that for each prefix a single path should be installed in the routing table to use for traffic forwarding. To rank all the available paths

BGP uses a numerical value that can be set for each path through the *Local Preference* (ttfamily LocPref) BGP attribute. The path with the highest ttfamily LocPref value will be selected as the most preferable path and will be used to route traffic. ttfamily LocPref is assigned locally to a router and is not propagated through BGP updates.

If two routes have equal Local Preference values, the path with the shortest AS path length is preferred. If both the ttfamily LocPref and the path length cannot determine the best path, BGP continues the path selection process by checking the value of other attributes that are used as tie-breakers. Table 2.2 shows the complete decision process of BGP. The ordering of the attributes is strict but some router vendors may allow some attributes to be deactivated, or they may add vendor-specific attributes. For example, Weight is a Cisco-defined attribute that it is considered before ttfamily LocPref.

Priority	**Attribute**	**Preference Rule**
1	Local Preference	Highest Local Preference
2	AS Path	Shortest AS Path
3	Origin	Lowest Origin type IGP is lower than EGP. EGP is lower than INCOMPLETE
4	MED	Lowest MED By default MED is considered only if neighbouring AS is the same in the compared paths.
5	eBGP/iBGP	Prefer eBGP over iBGP paths.
6	Metric	Lowest IGP metric
7	Router ID	Lowest Router ID

Table 2.2: BGP best path selection algorithm

2.3.3 Tagging

To facilitate the implementation of filtering and ranking, operators often annotate their routes with additional meta-data. Tags are applied through the BGP Communities attribute which is an optional BGP attribute that contains a series of 32-bit numerical values. The BGP Communities is a transitive attribute which means that Communities values can be propagated through different ASes which are free to edit them by adding or removing values. Communities can be applied either when a route is advertised, or when a route is received depending on the desired policy. Egress Communities are

usually used to request special treatment for a route by the neighbour to which the route is advertised. For example, an egress Community value can express a request by a customer to a provider for selective re-advertisement. Ingress Community values are usually informational applied by a border router to notify the routers that will subsequently receive the same route about its the properties. For instance, ingress Communities can encode the relationship type of the local AS with the AS from which a route was received. Communities are highly expressive because their values are not standardised, so each AS can determine its own values to implement highly complex policies.

2.4 AS Business Relationships

Internet inter-domain routing is a collaborative effort between ASes. ASes negotiate contractual agreements to define their business relations and impose technical restrictions on traffic exchange. On the Internet, connectivity does not imply traffic reachability, which is fundamentally determined by the business relationships between ASes.

The AS business relationships are coarsely divided into three categories.

1. Transit relationship, including customer-to-provider (c2p) and provider-to-customer (p2c). It is established when an AS (customer) pays a better-connected AS (provider) to transit traffic with the Internet. Essentially, providers operate as a gateway to the rest of the Internet. An AS can have multiple providers for purposes of resilience and load balancing. Such ASes are called multihomed.

2. Peering relationship (p2p), which allows two ASes to freely exchange traffic between themselves and their customers to avoid the cost of sending traffic through a provider.

3. Sibling relationship (s2s), links connect two ASes that belong to the same administrative entity without any cost or routing limitations.

According to the number and type of links an AS can be categorized as Tier-1, Tier-2, CPs, ISPs, and stubs. *Tier-1 networks* are ASes that have no provider and their main role is to offer global connectivity to other ASes. *Tier-2 networks* are also

large ASes that mainly provide IP transit to other ASes, but they are not transit-free. *CPs (Content Providers)* are global networks that mainly transit traffic between content generators and end-users. To achieve low cost and low end-to-end delay they mainly establish p2p relationships, although they have providers for redundancy and fall-over purposes. *ISPs* can have both customers and providers and usually their coverage is national or regional. ISPs provide internet connectivity either to stub ASes or to end hosts. ISPs that mainly operate as access providers for end-hosts are called *eyeballs*. *Stub ASes* have only providers and their scope is regional. Example of stub domains are university or research networks. Figure 2 shows an example of AS relationships.

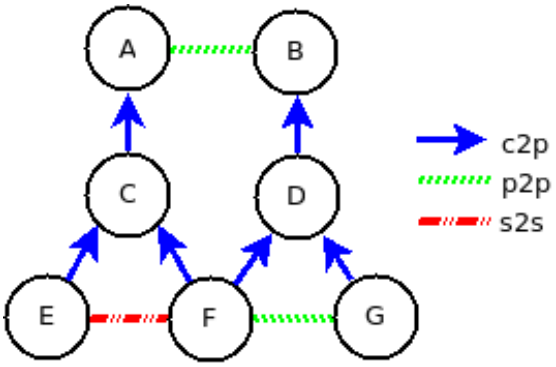


Figure 2.3: An example AS graph. Nodes A,B are transit ASes, nodes C,D are national ISPs and nodes E,F,G are stubs. Node F is multihomed since it has two providers, C and D. It should be noted that arrows show the flow of money and not flow of traffic which is bidirectional.

BGP routes are usually exported following the so-called *valley-free* rule [97], i.e. a customer route can be exported to any neighbour, but a route from a peer or a provider can only be exported to customers. Hence, a routing path (of a series of adjacent AS links) is valley-free if it follows such patterns: (1) $n\times$c2p + $m\times$p2c; or (2) $n\times$c2p + p2p + $m\times$p2c; where n and m ≥ 0. The valley-free rule aims to prevent an AS from providing free transit either to their providers or peers. It should be underlined that according to the valley-free pattern, only one p2p link is allowed in a valid path. Even though typically p2p relationships do not involve direct traffic exchange cost, ASes have indirect costs when carrying traffic over their network (consumption of resources). Therefore ASes avoid the transit traffic between their peers because they consume resources without generating profit. The sibling links can be inserted freely without changing the valley-free property of a path.

The valley-free rule describes a typical AS path. Most reachable paths which are

valid for traffic routing are valley-free, because they serves the business interest of ASes, i.e. to minimize operation cost and maximize revenue. It should be noted that the valley-free rule is not an enforcement rule. Namely, it does not mean that a routing path has to follow this rule.

2.5 Data Sources

The study of the topological properties of the AS-level graph requires the existence of reliable connectivity data. Unfortunately many ISPs are not willing to reveal the contractual relationships with other ISPs since such relationships are viewed as strategic advantages over the competition and thus are kept as business secrets. To overcome this problem there have been numerous efforts to develop methodologies and tools to collect and publish the links between the ASes. The three most important data sources are BGP data from public monitors, active traceroute measurements, and Internet Routing Registries (IRR).

2.5.1 Passive Monitoring

BGP tables can provide abundant information on AS-level connectivity, not only about AS adjacencies but also about the attribute of the links (e.g. LOCAL_PREF). Access to BGP routing tables can be obtained through BGP looking glasses, BGP route servers or BGP monitors. BGP looking glasses and route servers allow the remote execution of non-privileged BGP commands (e.g. `show ip bgp`) through a web interface or remote login to help network operators debug their configurations. An updated list of looking glasses and route servers is provided by the Traceroute Organization website[1]. Listing 1 shows the output of the `show ip bgp summary` command on a router of AS 25409 (Alsys). From the AS column we can infer the neighbours of AS 25409 as seen by its router 93.190.151.147. A router may not have the complete connectivity information. Looking glasses usually offer access to a limited number of routers so one cannot extract the full routing table of an AS through a looking glass. On the other hand, route servers offer full route tables. Currently there are 404 looking glasses and only 55 route servers.

BGP monitors offer the most complete BGP data by peering with backbone ASes

[1] `http://www.traceroute.org/`

```
BGP router identifier 93.190.151.147, local AS number 25409
RIB entries 678228, using 41 MiB of memory
Peers 8, using 20 KiB of memory

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
62.231.74.61    4  8708  671716  346056        0    0    0 01w4d12h     5867
89.37.120.193   4 39737  890437  346608        0    0    0 05w4d13h    13189
89.37.122.97    4 39737 27798745  690112        0    0    0 01w2d23h   317245
93.190.151.146  4 25409  346608 24600286        0    0    0 12w6d04h       16
193.231.3.127   4 16220  344790  345348        0    0    0 1d02h13m        0
193.231.184.233 4  8708 17827070  345932        0    0    0 01w4d12h   316590
2a00:ff0:fffe::2
                4 39737 1869480  690164        0    0    0 01w2d23h        0

Total number of neighbors 7
```

Listing 1: The routing table of AS 25409 as obtained from the looking glass at http://lg.alsysdata.net/

and by collecting full BGP tables and update messages. Routeviews [21] has deployed 15 monitors around the world (12 in the US, 1 in the UK, 1 in Japan and 1 Kenya) that continuously collect BGP tables and updates (BGP dumps) from hundreds of different backbone routers. The motivation of the ASes that offer access to their backbone routers is to understand how their prefixes are viewed by the global routing system. However this information has been widely used for discovering the adjacent ASes. RIPE has also deployed similar monitors (4 in the US, 10 in Europe, 1 in Japan, 1 in Brazil and 1 in Russia). Routeviews and RIPE data are freely available from their webpages[2, 3].

BGP data can provide a very accurate and updated state of the AS adjacenies. Since most of the BGP dumps come from backbone (default-free) routers that accumulate prefix announcements (and thus AS paths) from all over the Internet, a few BGP tables can provide a very wide view of the Internet. Also, BGP dumps contain the attributes of the advertised paths making possible an in-depth analysis of the AS links (e.g. identify backup links). However, BGP misconfigurations or malicious interventions are not uncommon [151] and can affect the quality of the collected data.

2.5.2 Active Probing

Traceroute is a tool for the discovery of router-level paths based on the Internet Control Message Protocol (ICMP) [171]. Its mechanism utilizes a field of the User Datagram Protocol (UDP) [170] header which is called Time-To-Live (TTL). When

[2] http://archive.routeviews.org/
[3] http://www.ripe.net/projects/ris/rawdata.html

a router receives a UDP datagram it decreases its TTL value by 1 and forwards the message to the next router according to the destination IP address. If a router receives a message with TTL == 1 then this router doesn't forward the datagram further and replies with a ICMP Time Exceeded Message (type 11). Traceroute exploits this mechanism by sending UDP datagrams that start with TTL equals to 1 and increases it by 1 every time a ICMP type 11 message is received. Gradually the sender receives replies with the IP addresses of all the routers in the path. The IP address can then be resolved to the corresponding AS number.

Traceroute cannot always discover complete router paths. It is common that some routers have disabled the ICMP protocol and thus they do not reply to the sender while other routers reply with the destination address of the UDP probes instead of their own address. Such routers are called *anonymous* and the inferred router path will have a '*' entry indicating these non-responding routers. The presence of anonymous routes can result to incomplete or inflated topologies. Identifying the '*' entries that belong to the same routers is a NP-complete problem [204] and heuristics have been proposed to obtain more accurate results [204, 113]. Another problem that limits the reachability of traceroute probing is the use of firewalls that may block the UDP or ICMP packets. TCP traceroutes can be used instead of UDP and ICMP traceroutes to bypass firewalls that accept TCP connections. A third problem is that many routers have multiple interfaces configured with different IP addresses which are called aliases. If these aliases are not resolved, the different interfaces will be reported as different routers. Another problem of traceroute is the inference of spurious paths when a load-balancing mechanism(per-packet or per-flow is used by some routers, as explained in figure 2.4. Per-flow mechanisms identifies different flows based on the first four bytes of the transport-layer header. Traceroute triggers per-flow load balancing by changing some of the bits of those bytes (UDP port number or ICMP Echo sequence number) to match the router's responses with the sent probes. Paris traceroute [35] is a variant of classic traceroute that keeps the flow identifiers stable to avoid segregating the successive probes to different flows.

2.5.3 IRR Databases

Internet Routing Registries (IRRs) [12] are publicly accessible databases where AS administrators voluntarily and manually register and update routing information. IRR can be used for network troubleshooting, route filtering and validation and can be queried using the WHOIS protocol. Examples of popular IRRs are the RIPE and the APNIC WHOIS database [23, 22] although different IRRs are mirrored in multiple sites. Listing 2 shows an extract from the RIPE's IRR record for AS 174 (Cogent) that describes two links to AS3 and AS14 to which it announces all the prefixes it knows and accepts all advertisements from these two ASes.

```
as-name:        COGENT
descr:          Cogent/PSI
import:         from AS3 accept ANY
export:         to AS3 announce ANY
import:         from AS14 accept ANY
export:         to AS14 announce ANY
```

Listing 2: Example of IRR record obtained from RIPE's database for AS 174.

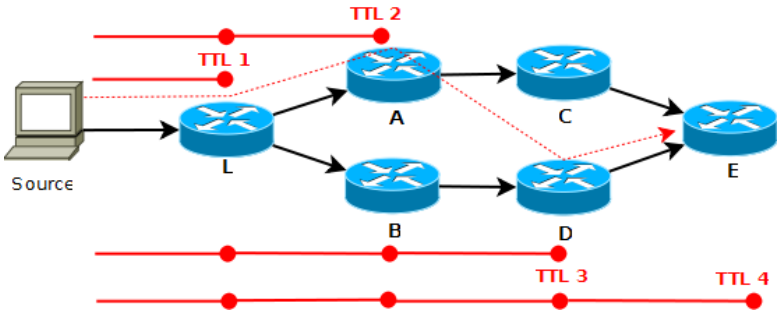


Figure 2.4: Inference of load balancing through Paris Traceroute. If router L uses a load-balancing mechanism, it may direct the first two probes (TTL =1 and TTL =2) through the link L-A and the two next probes (TTL = 3 and TTL = 4) through the link L - B. The source will receive ICMP replies from nodes L,A,D and E and infer the corresponding path L-A-D-E. The link A-D is an artifact of load balancing.

2.6 Related Research

This section reviews the related research on the discovery, modelling and annotation of the inter-domain topology.

2.6.1 Topology Incompleteness

Based on the data sources described in section 2.5 a large number of studies tried to analyse the topological properties of the AS ecosystem. Faloutsos brothers [86] were the first who observed that the AS graph degree distribution follows a power-law distribution with exponent $\gamma = 2.22$. Their results were based on three instances of a collection of BGP tables from 1997 to 1998 which contained a relatively small number of nodes and links. However latter studies confirmed the existence and persistence over time of the power-law degree distribution using different datasets - including Routeviews, RIPE, Skitter, DIMES and Mercator - and observed only small changes of the value of the power-law exponent that varies between -2.1 and -2.3 [188, 107, 149, 116, 185].

Despite the repeatedly reported power-law degree distribution, many researchers argued that it is an artifact of measurement biases. Lakhina et al. [141] were the first to empirically show that even if the real network is an Erdös-Réyni graph with Poisson degree distribution, the degree distribution of the observed graph is close to power-law. The distortion of the distribution is a result of probing a large number of destinations from a small number of sources. Effectively the real graph is approximated by a union of spanning trees where the edges near the root are sampled much more frequently than the distant edges and thus their distribution differ significantly.

Furthermore, [141] suggested that a significant source of bias is the inability of traceroute to capture the lateral edges that are not part of the shortest-path trees. Two more studies [27, 63] formally verified the results of [141] showing that for random graphs $G(n, p = c/n)$ - where n is the number of nodes, c is the average degree and p is the probability that an edge exists between two vertices - single-source traceroutes result in degree distribution $P(k) \sim k^{-1}$ up to $k \sim c$. According to [63], to mitigate the effects of the sampling bias the number of monitors should grow linearly with the average degree. Other works, demonstrated that the exponent α of the power-law distribution is underestimated since the sampling process focuses on high-betweenness vertices that usually have higher degree [69, 168]. Mahadevan et al. [150] suggested that the above studies are also relevant to BGP topology inference since it can also be approximated as a union of spanning trees whose root is the BGP collection point.

Chen et al. [59] combined a number of BGP data sources with data from the RIPE

IRR to calculate that the original dataset used by [86] to derive the power-law relationships miss 20-50% links. A notable result is that although the degree distribution of the augmented dataset is heavy tailed, it does not conform to power-law distribution but it is closer to Weibull distribution. Cohen and Raz [64] conducted a similar study in order to quantify the number of missing links in the AS topology. They concluded that 35-50% links remain hidden from the BGP tables and that the majority of these links are of p2p type. The difference in the observed number of p2p links is responsible for the different degree distribution observed by Routeviews and RIPE IRR. More specifically, they showed that both in Routeviews and RIPE graphs the c2p subgraphs follow power-law degree distribution while the p2p subgraphs follow Weibull degree distribution. Since the majority of Routeviews links (about 92%) was classified as c2p and the majority of IRR links is classified p2p (71-75%), the Routeviews graph is biased towards power-law while IRR is biased towards Weibull distribution. The big number of p2p links in RIPE database can be explained by the fact that European IXPs require from their members to register their peering links established over their fabric [4].

Oliveira et al. [167] classified missing links into two categories, *hidden* and *invisible*. Hidden links are usually backup c2p links that can be observed if the preferred path towards a prefix changes. Other (typically p2p) links are inherently invisible links due to the limited number and placement of vantage points, and the route propagation restrictions on p2p relationships associated with valley-free routing policies of most ISPs.

Invisible p2p links constitute the majority of missing links, and are mostly located in the periphery of the AS graph [178, 166]. BGP feeds are mostly provided by high-tier ASes and some geographic areas are poorly covered. Furthermore, two-thirds of all contributing ASes configure their connection with the BGP collector as a p2p link, which means they advertise only routes learned from customers.

IRR does not suffer of the systematic measurement bias of traceroute and BGP data sources. However, the registered information is frequently inaccurate, incomplete or intentionally false in order to appear more attractable to other ASes [187]. It is also reported that only 28% of the ISPs have correct data in IRRs when compared

[4] An example is the LINX Memorandum of Understanding which states in section 4.6 that "All routes to be advertised in a peering session across LINX shall be registered in the RIPE or other public routing registry" `https://www.linx.net/govern/mou.html`

to collected BGP data and that the RIPE database is the most accurate [187]. RIPE provides a tool for consistency checking of the IRR database to detect inaccurate or missing information [18]. As a result IRRs are not recommended for extracting the AS topology or they should be used in combination with other data sources. Battista et al [42] developed a tool that collects the AS adjacencies from the IRR databases after checking their consistency and classifies the extracted links based on whether the import and export policies are observed to the data of both adjacent ASes. These data are available from the website of the project [5].

Another increasingly prominent aspect of the Internet interconnection ecosystem is the proliferation of IXP infrastructure to facilitate cost-effective dense peering. This presents a challenge to peering visibility using traditional BGP data archives. Recent work [124, 36] has demonstrated a vast number of p2p links at IXPs, most of which were not visible in any public dataset. Despite the novelty of these techniques [36, 124] they have not been used to provide periodic data due to the complexities of data acquisition. Conducting large-scale targeted traceroute measurements is computationally expensive as well as time-consuming; the methodology in [36] required 16 million traceroute and LG queries (14 days to complete) to discover 44K links.

2.6.2 Topology Discovery

Missing links is one of the most significant challenges in inter-domain cartography. As a result multiple research projects focused on mitigating the topology incompleteness problem. Theoretically optimal placement of BGP monitors might mitigate this incompleteness [209, 186, 109], but in practice ASes participate voluntarily in such data collection projects so optimal placement is not possible.

A very extensive work to discover the missing links from the BGP and traceroute measurements is described in [124]. The authors try not only to quantify the percentage of missing links but also to discover which links are actually missing. They utilize data from IXPs (participating ASes) and IRR as hints for which links are likely to exist but are not yet discovered. Then they validate the existence of these links using a large-scale traceroute tool that employed public traceroute servers. Their results agree with the previous studies as they report that about 40% more links and 300% more p2p links

[5] http://tocai.dia.uniroma3.it/~irr_analysis

were discovered in comparison to the Oregon Routeviews data. In comparison to all the BGP routing tables these percentages are 15% and 65% respectively. In addition to these revealed links, it is estimated that there are about 35% p2p links that remain hidden. The results of [124] verify that the degree distribution of the c2p links can be accurately described by a power-law distribution while the degree distribution of p2p links follow a Weibull distribution. Unfortunately, their link discovery has not been repeated to provide the research community with updated data. Therefore, this work provided valuable insights but did not advance our capabilities for more accurate topology analysis.

Recently, Ager *et al.* [28] used sFlow traffic data to discover a rich peering fabric at a large European IXP. Based on actual traffic exchanged they inferred more than 50K p2p links at this IXP alone, about 10K more compared to public BGP data for the same period, and about 6K more than what the authors of [36] discovered across all IXPs using distributed traceroute measurements. Using traffic data as in [28] is trickier since such data is private and not available to the broader research community, inhibiting reproducibility. In addition, (1) the Non-Disclosure Agreement (NDA) with the IXP inhibits the derived links from being reported publicly, (2) NDA agreements have to be made with individual IXPs, and (3) not all IXPs are likely to agree to an NDA.

BGP data can be augmented by collecting AS paths through public Looking Glass (LGs) servers. Multiple measurement projects have utilized LGs to obtain additional connectivity information [36, 124]. Khan et al. attempted to quantify the amount of additional links revealed by LGs compared to BGP, traceroute (CAIDA,iPlane) and IRR data. For March 2013, they queried 245 LG servers across 110 countries to discover 11 K AS links not present in the other available topology data. Although providing an expanded topology, it appears that adding more BGP vantage points results in diminishing returns since LGs suffer from similar limitations. Given that active querying of LGs incurs significant measurement cost it may not worth it to use a large number of LGs to reveal a small fraction of additional links. However, LGs provide access to BGP attributes that may be completely invisible to RouteViews and RIPE RIS datasets, such as Local Preference values that offer a completely new dimension of policy data. Therefore, I believe that utilisation of LGs should have as primary aim the collection of policy data not found passive BGP data, instead of the collection of extra AS links

that offers only incremental improvement of the topology completeness.

Some researchers suggest highly distributed traceroute monitoring infrastructures [184, 58] are a promising approach to discover invisible AS links, yet the visibility improvement so far is limited compared with the links discovered at just a single IXP by Ager *et al.* [28]. The next section reviews some of the most important efforts in traceroute-based topology discovery.

2.6.2.1 Traceroute-based topology discovery

Mercator [107] uses hop-limited probing like traceroute to discover the router-level map of the internet. To be easily deployable it does not use any database of IP addresses to be probed, but uses a heuristic called informed random address probing. When Mercator receives a reply to a probe it assumes that the source IP address of the reply is part of an addressable prefix that contains more routed address. It also assumes that the neighbouring prefixes are also addressable. One of Mercator's aims is to map the Internet from a single arbitrary location. To extend the scope of its probes and to discover "cross-links" it uses source-routing. Source routing allows the sender to specify the route taken by the probes, so Mercator can direct probes through routers that are not observed in the routing paths to a certain destination. Essentially every source-routing capable router can act as an instance of Mercator running from the router's location. Even though only 8% of Internet routers support source routing, the authors argue that in sparse graphs 5% of source route capable routers can discover 90% of the graph's links. Mercator also employs an alias resolution technique called alias probing. A UDP packet is sent to a non-existent port on a router interface and usually routers reply with an ICMP Port Unreachable message sent through the interface of the unicast route. It is then understood that these two different interfaces belong to the same router. Figure 2.5 explains this process. One of the main drawbacks of Mercator is the use of classful prefixes which are not used anymore due to the Classless Inter Domain Routing (CIDR) [95]. Another drawback is that Mercator uses very low probing frequency to avoid generating excessive traffic. It takes about 3 weeks for Mercator to discover 200,000 links which results in maps that are time-averaged and may contain dead links or may miss links established during this period. There hasn't been any publicly available dataset generated by Mercator measurements.

Rocketfuel [189] is an active probing measurement tool that aims to reduce the number of required probes without sacrificing accuracy. To do so it uses two heuristics. The first is *directed probing* that utilizes Routeviews BGP routing tables (specifically the AS path towards an IP prefix) to identify which destination addresses will result in traceroutes that transit an ISP's infrastructure. The second heuristic is called *Path Reductions* that reduces redundant traceroutes based on the observation that often it's the next-hop AS and not the IP prefix that determines the routing path. Also it is suggested that probes from multiple monitors to a destination may be redundant as they converge to the same ingress router. Similarly probes from a single monitor to multiple destinations may converge to the same egress router. These heuristics enable Rocketfuel to execute only 0.1% of the traceroutes required by a brute-force approach while in comparison to Skitter it discovers seven time more links. Rocketfuel also integrates *Ally*, an alias resolution tool that exploits IP identifiers (IP_ID) to infer aliases. IP_ID is a counter that a router increments after sending a packet and is used for reassembly of fragmented IP datagrams. Consecutive packets will normally be tagged with consecutive IP identifiers, so Ally sends consecutive probes (e.g. x $\rightarrow$ y $\rightarrow$ z) to the suspected aliases and compares their IP_ID fields. If $ID_x < ID_y < ID_z$ and $ID_z - ID_x$ is small then the interfaces are likely to belong to the same router. To add confidence Ally also compares the TTL of the responses. Compared with the Routeviews data, Rocketfuel discovers less AS adjacencies (for AS 1239 it discovers 30% less links) although Rocketfuel's data include links not present in BGP routing tables (mainly links between large ASes). The only available dataset by Rocketfuel is a one-off measurement in 2002 and it is available from the project's website[6].

Skitter [126] is one of the most widely used topology discovery tools developed by CAIDA as part of its Macroscopic Topology Project. During the period 1998-2008 skitter has been deployed to 25 different probing locations (monitors) dispersed all over the world to collect daily data based on ICMP traceroute probes. Every monitor had a destination list of IP address that combined resulted to a set of one-half million IP addresses covering all routed /24 IP prefixes. These destinations were gathered either by collecting the source address from DNS queries or from addresses found in the web (webservers or other sources). Skitter integrates a tool called *iffinder* [10] to perform

[6] `http://www.cs.washington.edu/research/networking/rocketfuel/`

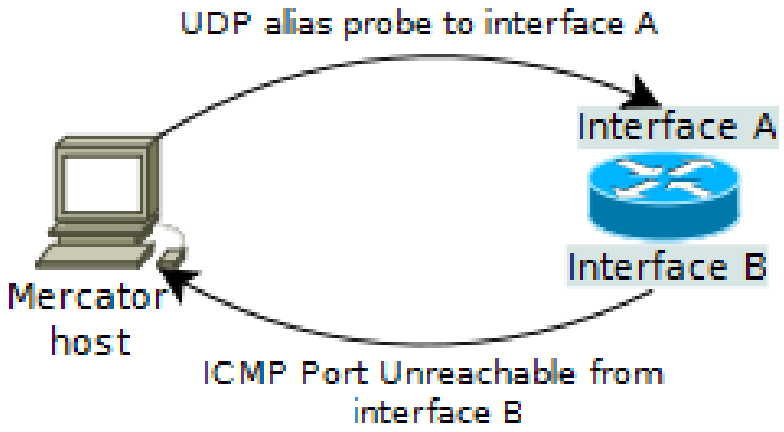


Figure 2.5: The alias probing method used by Skitter for discovering interfaces that belong to the same router. The Mercator host sends a UDP packet to a high port number with destination address A. If a ICMP Port Unreachable message is received with source address B then the addresses A and B belong to two different interfaces of the same router.

alias resolution similar to that of Mercator (see figure 2.5).

Archipelago (Ark) [62] is the successor of Skitter. It is built upon *scamper*, an active measurement tool that supports UDP, ICMP and TCP traceroutes and allows IPv6 measurements. Scamper uses Paris traceroute. A significant addition in Ark is the use of a distributed shared memory (tuple-space) that allows decentralized coordination of the distributed monitors. For alias resolution it integrates two tools, iffinder and APAR (Analytical and Probe-based Alias Resolution) [114]. APAR is based on the observation that two router interfaces connected with a point-to-point link often are consecutive and belong to the same /30 or /31 prefix. Using reverse traceroutes APAR can find the interfaces connected by point-to-point links and by aligning the reverse traces it infers the aliases as explained in figure 2.6. It has been shown that APAR combined with Ally result in the most inferred aliases [112]. Ark is deployed in 46 monitors as of May 2010. Both Ark and Skitter datasets are publicly available from the CAIDA website[7].

DIMES (Distributed Internet Measurement and Simulations) [184] is another well-known topology discovery tool. DIMES is a distributed scientific project similar to SETI@home where DIMES agents are voluntarily installed in end-user computers across the Internet and issue traceroute and ping probes at low rate. DIMES agents cover a very wide geographic area thus can discover more links than Skitter. However

[7] `http://www.caida.org/data/overview/`

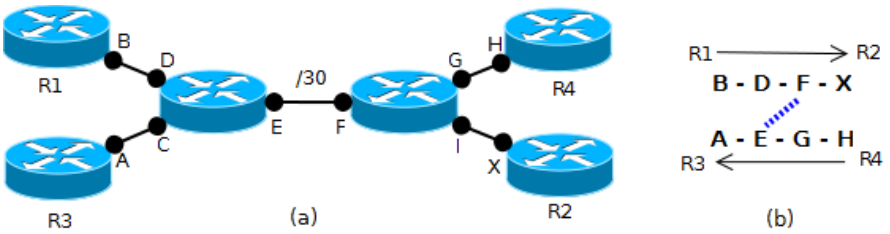


Figure 2.6: The traceroutes from R1 to R2 and from are R4 to R3 reveal that interfaces E and F are connected by a point-to-point link. By aligning the traces it can be inferred that D,E and F,G are interfaces of the same router.

many of these links overlap since DIMES agent can locate in the same geographic area. Unlike the above-mentioned tools, DIMES does not perform alias resolution. DIMES dataset is also publicly accessible from its website [8].

2.6.3 Relationship Inference Algorithms

Internet studies demand knowledge on the relationships between ASes. However most ASes try to hide their business relations. In the last decade researchers have introduced a number of algorithms to infer the AS relationships.

Gao [97] was the first to study the inference of AS relationships. Her solution relies on the assumption that BGP paths are hierarchical, or *valley-free*, i.e., each path consists of an uphill segment of zero or more c2p or sibling links, zero or one p2p links at the top of the path, followed by a downhill segment of zero or more p2c or sibling links. The valley-free assumption reflects the typical reality of commercial relationships in the Internet: if an AS were to announce routes learned from a peer or provider to a peer or provider (creating a valley in the path), it would then be offering transit for free. Gao's algorithm thus tries to derive the maximum number of valley-free paths, by selecting the largest-degree AS in a path as the top, and assuming that ASes with similar degrees are likely to be peers (p2p). Gao validated her results using information obtained from a single Tier-1 AS (AT&T). Xia and Gao [202] proposed an improvement to Gao's algorithm that uses a set of ground truth relationships to seed the inference process. Transit relationships are then inferred using the valley-free assumption. Gao's algorithm [97] is used for remaining unresolved links. They validated 2,254 (6.3%) of their inferences using 80% of their validation data and found their algorithm was accurate for 96.1% of p2c links and 89.33% of p2p links.

[8] `http://www.netdimes.org/DIMESControlCenter/MonthlyData.jsp`

Subramanian *et al.* [191] formalized Gao's heuristic into the Type of Relationship (ToR) combinatorial optimization problem: given a graph derived from a set of BGP paths, assign the edge type (c2p or p2p, ignoring sibling relationships) to every edge such that the total number of valley-free paths is maximized. They conjectured that the ToR problem is NP-complete and developed a heuristic-based solution (SARK) that ranks each AS based on how close to the graph's core it appears from multiple vantage points. Broadly, as in Gao [97], ASes of similar rank are inferred to have a p2p relationship, and the rest are inferred to have a p2c relationship. Di Battista, Erlebach *et al.* [40] proved the ToR problem formulation was NP-complete in the general case and unable to infer p2p relationships. They reported that it was possible to find a solution provided the AS paths used are valley-free. They developed solutions to infer c2p relationships, leaving p2p and sibling inference as open problems. Neither Subramanian *et al.* or Di Battista, Erlebach *et al.* validated their inferences; rather, they determined the fraction of valley-free paths formed using their inferences.

Dimitropoulos *et al.* [76] created a solution based on solving MAX-2-SAT. They inferred sibling relationships using information encoded in WHOIS databases. Their algorithm attempted to maximize two values: (1) the number of valley-free paths, and (2) the number of c2p inferences where the node degree of the provider is larger than the customer. The algorithm uses a parameter α to weight these two objectives. They validated 3,724 AS relationships (86.2% were c2p, 16.1% p2p, and 1.2% sibling) and found their algorithm correctly inferred 96.5% of c2p links, 82.8% of p2p links, and 90.3% of sibling links. Their validation covered 9.7% of the public AS-level graph and has thus far been the most validated algorithm. However, MAX-2-SAT is NP-hard and their implementation does not complete in a practical length of time for recent AS graphs.

UCLA's Internet Research Laboratory produces AS-level graphs of the Internet annotated with relationships [25]. The method is described in papers by Zhang *et al.* [206] and Oliveira *et al.* [166]. Their algorithm begins with a set of ASes inferred to be in the Tier-1 clique, then infers links seen by these ASes to be p2c; all other links are p2p. Zhang [206] describes a method to infer the clique; Oliveira [166] assumes the Tier-1 ASes are externally available, such as a list published by Wikipedia. There are a growing number of region-specific c2p relationships visible only below the provider

AS, causing this approach to assign many p2p relationships that are actually p2c. Gregori *et al.* [108] used a similar approach; for each AS path, their algorithm identifies the relationships possible and infers the actual relationship based on the lifetime of the paths. None of [206, 166, 108] describes validation.

2.6.3.1 Limitations of Relationship Inferences

It has been shown AS path inference based on AS relationships results in poor results. Mao et al [153] was the first to approach the problem of inferring the AS paths between any two ASes based on the AS relationships. One of their main findings is that the quality of the AS relationships is the most possible cause of mismatches and propose a new relationship inference algorithm based on the observed AS paths. For all the BGP gateways the achieve accuracy close to 60% in the sense that one of the inferred best paths matches with the actual AS path, and accuracy up to 62% for path length match. They further increase the accuracy of their predictions to 70% by assuming direct access to the destination hosts in order to infer the first hop. They conclude that multihoming and complicated AS relationships are the main reasons for not achieving higher correct prediction rate. Deng et al. [70] concluded that the valley-free assumption is a fundamental limitation of path prediction algorithms that depend on AS relationships.

In [172] the authors propose two algorithms for AS-level path prediction, by utilizing the AS paths that appear in BGP tables to infer AS paths not present in the available tables. Their algorithms infers AS paths on the granularity of destination prefix instead of destination ASes. The accuracy of this algorithm depends on the amount of the known BGP paths. When a substantial amount AS paths are available the prediction rate can approach 80%. This approach requires a much larger amount of input data and has the drawback that it cannot predict potential routing changes in case of link failures but only the best paths as these are observed at some specific time snapshot.

An alternative approach is followed by [159, 160] where the authors do not model the ASes as an atomic entity but each AS consists of a number of quasi-routers depending on its size. Also, recognizing that AS relationships are more complex than the ones described by the existing models their simulations are relationship-agnostic. Instead, they infer the import and export policies again by using a large number of known AS

paths as a training dataset.

2.7 Summary

The discovery of the Internet topology relies on three data sources, BGP tables [21, 16], traceroute paths [4, 148], and IRRs [12]. All of these sources of routing data have not been designed for the discovery of the Internet topology but rather for the debugging of routing policies. Therefore, their use as connectivity data sources is essentially a hack due to the absence of dedicated topology discovery tools [198]. As a result each dataset has its own limitations that introduce significant challenges in the study of the Internet topology.

Importantly, none of the available topology discovery methodologies are able to capture the complete inter-domain topology [65, 178, 124]. The incompleteness of the resulted topologies is mainly a result of policy-based routing that restricts the propagation of certain link types, and especially links with attached peering and backup relationships [167, 36]. Therefore the collected topologies are not only incomplete but also biased towards certain link types.

The incompleteness problem limits significantly the applications of the collected topologies and the research community proposed a number of different approaches to extend our current link visibility. A promising approach is the large-scale deployment of traceroute probes at edge hosts through crowd-sourcing [184, 15, 180]. Installing new traceroute vantage points is easy and inexpensive, but there are many challenges involved in translating the collected IP-level paths to router or AS paths. Anonymous and unresponsive routes, IP aliases, third-party addresses and multiple-origin addresses pose significant open problems [80, 154].

The AS topology graph alone is not enough for studying the Internet inter-domain routing. This is because the business relationships between the ASes play a crucial role in the decision process of BGP routing. Only combined knowledge of the AS relationships and the AS topology can allow researchers to run more realistic simulations and enable engineers and operators to make more informed decisions. However, for business reasons, ASes do not want to disclose their relationships. In recent years a number of algorithms have been proposed to infer AS relationships based on the AS topology data [191, 41, 76, 202, 197, 166]. They have applied a variety of heuris-

tics with increasing sophistication. The quality of their results have been questioned and it has been shown that the current inter-domain models yield poor simulation results [153, 172, 159, 160]. The lack of correctness and predictability led operators and engineers to criticize many fundamental assumptions of the existing modelling approaches [179]

The problematic lack of validation puts AS relationship inference research in a precarious scientific position. Nonetheless, research continues to build on the assumption that meaningful AS relationship inference can be achieved and applied to the study of the Internet, from deployment of security technologies [100], Internet topology mapping [36, 58, 28] and evolution [72, 73], to industry complexity [88] and market concentration [75]. Due to the diversity in inter-domain connectivity, relationship inferences do not (by themselves) consistently predict paths actually taken by packets; such predictive capabilities remain an open area of research [160].

Given its importance to the field of Internet research, the science of AS relationship inference should be revisited, with particular attention to validation.

Chapter 3

Revealing the Complexity of BGP Policies

Past studies on the inference and modelling of BPG policies and AS relationships relied almost exclusively on connectivity data that were translated to policy data through various heuristics. These heuristics typically offered a simplistic or even inaccurate abstraction of the actual policies leading to a distorted picture of inter-domain routing.

Despite the information-hiding nature of BGP, the available data include attributes beyond the AS path that encode policy data which when collected can broaden considerably our measurement horizons. Therefore, instead diving directly to the design of increasingly sophisticated but questionable heuristics for the parsing of connectivity information, it is necessary to take a step back and analyse the routing paths through the lens of policy data. The ground-truth of BGP policies is held in the BGP router configurations and is expressed through the BGP attributes, as documented in section 2.3. Two attributes, the BGP Communities and the Local Preference (LocPref) are of particular interest since they are used to implement two of the three policy functions, tagging and routing. This chapter describes a measurement framework to mine BGP Communities and LocPref values to extract ground truth policy data for around 40% of the AS links.

The collected ground truth allows me to observe and analyse a number of issues in BGP routing, such as the non-valley-free paths, the hybrid relationships, the backup links and the differences between IPv4 and IPv6 AS relationships. It also enables the evaluation the existing inference algorithms. This study offers new insights in the complexity of the actual BGP policies and creates a strong foundation for building the

```
TYPE: TABLE_DUMP_V2/IPV4_UNICAST
PREFIX: 1.22.73.0/24
FROM: 206.223.115.10 AS4589
ORIGIN: IGP
ASPATH: 4589 15412 18101 45528
NEXT_HOP: 206.223.115.10
COMMUNITY: 4589:2 4589:410 4589:612 4589:14413 15412:604
15412:614 15412:621 15412:705 15412:1431 18101:1344
18101:50120 18101:50420
```

Figure 3.1: Entry from a BGP routing table dump tagged with multiple Communities.

accurate measurement and inference techniques that will be described in chapters 4-6.

3.1 Introduction

One of the biggest challenges in developing accurate inter-domain models is the lack of ground truth data which are only held in router configurations. However, many of the BGP policies are implemented using the BGP Communities, a highly expressive transitive BGP attribute that tags BGP paths with meta-data [79]. Utilization of the Communities attribute can provide a rich source of policy data that cannot be obtained otherwise, but its utilization in previous topology studies has been minimal at the best due to the difficulties involved in interpreting the encoded data. By implementing a new framework for the systematic extraction, interpretation and evaluation of BGP Community values, I collect ground-truth policy and relationship data from more than 6,800 Community values that provide meta-data for over 40% of the visible AS links.

A second important BGP attribute that strongly expresses policy intentions is the Local Preference (LocPref) that expresses how preferable a route is to the local AS. The LocPref attribute is not transitive, namely it is not included in the BGP announcements received by RouteViews and RIPE monitors. The LocPref values can be obtained by having a direct interface to a BGP router. Remote access to such interfaces is provided through public Looking Glass servers that allow remote execution of non-privileged BGP commands over the SSH or telnet protocols. I have developed an API to query 28 Looking Glass to collect the LocPref for more than 10% of the visible AS links.

3.2 BGP Communities Attribute

The BGP Communities is an optional BGP attribute that contains a series of 32-bit numbers [54]. BGP Communities are applied to prefix advertisements to tag the relevant routes with additional information. The first 16-bits of each Community value represent an ASN while the last 16-bits encode a value with a pre-determined meaning. This meaning can correspond to relationship information, geolocation, or request for triggering traffic engineering policies. The BGP Communities is a transitive attribute which means that Communities values can be propagated through different ASes which are free to edit this field by adding or removing values. The BGP Extended Communities attribute [182] is a 64-bit value that extend the range of the BGP Communities and provide structure through a type field.

Values of the Communities attribute are not standardised. Many ASes explain the meaning of their Communities values in their Internet Routing Registry (IRR) records [12] or in the resources of their Network Operation Centres (NOC). A database of NOC websites can be found in the PeeringDB records [14].

Figure 3.1 shows an entry from a BGP table dump. From the `AS_PATH` I obtain three AS links, AS4589–AS15412, AS15412–AS18101, and AS18101–AS45528. Communities values beginning with '4589:' are determined by AS4589 to describe the AS link with AS15412. If the value 4589:612 encodes the meaning 'Route received from a LINX peer', I infer the relationship AS4589 – AS15412 as p2p. Similarly, if the value 15412:705 corresponds to 'Route received from customer', I infer the relationship AS15412 – AS18101 as p2c.

BGP Communities have been increasingly used by AS operators to implement a wide range of BGP policies, such as the valley-free rule, black-holing traffic, or complex load balancing techniques [92]. A rigorous classification of BGP Community values is provided in [79]. For June 2011 I have observed in RouteViews and RIPE RIS BGP data 26,055 distinct Community values, set by 2,964 different ASes in 15,167,572 paths, after I filter out records with (1) reserved and private AS numbers (i.e. 23456 and 56320–65535) and (2) path cycles that result from misconfigurations. Figure 3.2 show the number of distinct Community values per AS number. For the majority of ASes I observe less than 10 Community values but for a few ASes I observe some hundreds of values.

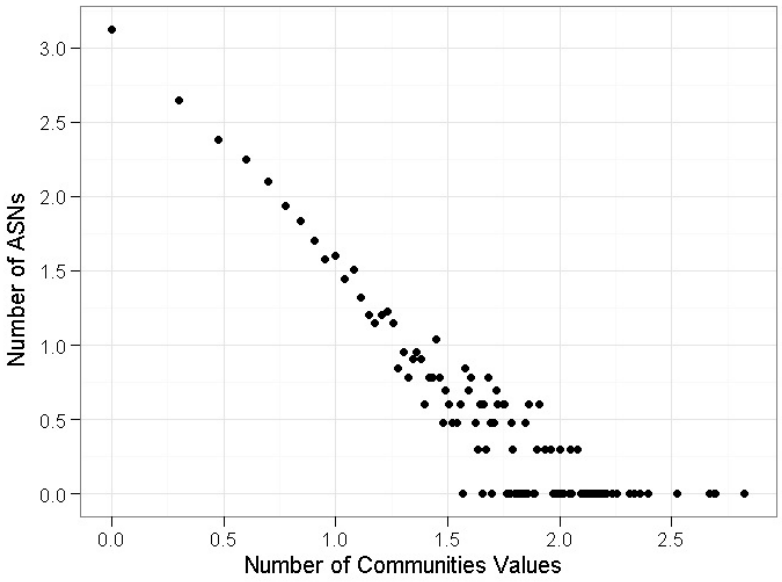


Figure 3.2: The number of ASes as a function of the number of unique Community values that an AS has (log-log scale).

3.2.1 Interpretation of BGP Community Values

Successful interpretation of the encoded information in the Community values would significantly reduce the dependence of BGP simulations on heuristics in favour of real policy data. However, the usage of the BGP Community attribute is not standardized and each AS is free to define and use arbitrary values. RFC1997 [54] defined three Community values for suppressing route advertisements, but they serve limited applications and often they are ignored by AS operators. RFC4384 [157] proposed a more complete standardization to facilitate data collection, but it has not been adopted by the operators. Hence, the interpretation of the Community values requires additional sources of documentation.

Many operators document the usage of their Community values in the *remarks* section of their Internet Routing Registry (IRR) records [12], or in the resources of their Network Operation Centres (NOC). This documentation is in free text, without specific format or terminology, and consequently significant manual work is required to process and extract the available information. As an example, Figure 3.3 lists some of the different ways observed in the IRR to document route redistribution Communities that prepend an AS path twice, and that suppress the prefix advertisement towards specific peers.

```
12713:9422 Prepend to LINX peers Twice
12578:162 Prepend AS12578 2x to Cogent AS174
13645:2 Prepend 13645 13645 to route
29113:6060 - ! Advertise to Telia (AS1299)
1836:11100 No Export to European Peers
2764:7 Announce to customers only
```

Figure 3.3: Examples of parsed BGP Communities documentation. Different ASes not only use different values to implement the same policy (e.g. prepend twice), but also may document these policies with different language. To parse such documentation I developed a Natural Language Processing tool since the lack of certain structure in documentation makes impractical the development of regular expressions.

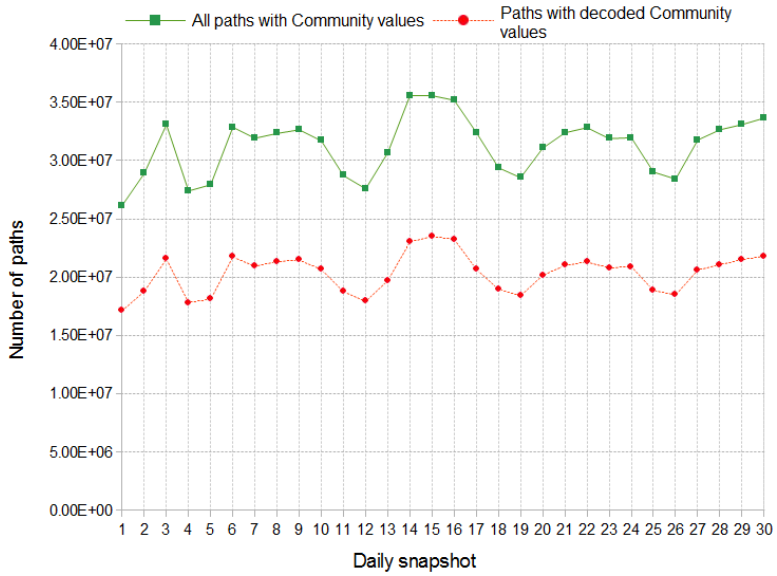


Figure 3.4: The daily number of AS paths with decoded BGP Communities versus the total number of paths with BGP Communities

Transforming these Community values to input for BGP simulation policies would require manual processing, which is not possible regarding the large number of Community values. The lack of standardisation in the usage and documentation of BGP Communities is the main reason for their under-utilization in research studies. To overcome these difficulties I developed a Natural Language Processing (NLP) tool to transform the unstructured documentation to the JSON structured data format, based on Python's Natural Language ToolKit (NLTK) [43]. Unfortunately, to the best of my knowledge there does not exist any annotated corpus of network policy documentations. Instead the million-word Brown corpus [94] was used to train a Part-of-Speech (PoS) tagger using Maximum Entropy methods as explained in [152]. For route redistribution Community values I extract the (i) type of the Community, (ii) the target

```
{"value":"29113:6060","type":"announcement","target":"AS1299","action":"deny"}
{"value":"2764:7","type":"announcement","target":"non-customers","action":"deny"}
{"value":"1836:11100","type":"announcement","target":"EU Peers","action":"deny"}
```


Figure 3.5: Examples of parsed BGP Communities documentation. The NLP parser will take as input the text in Figure 3.3 and will return a JSON file which identifies the type, target and action of each Community value.

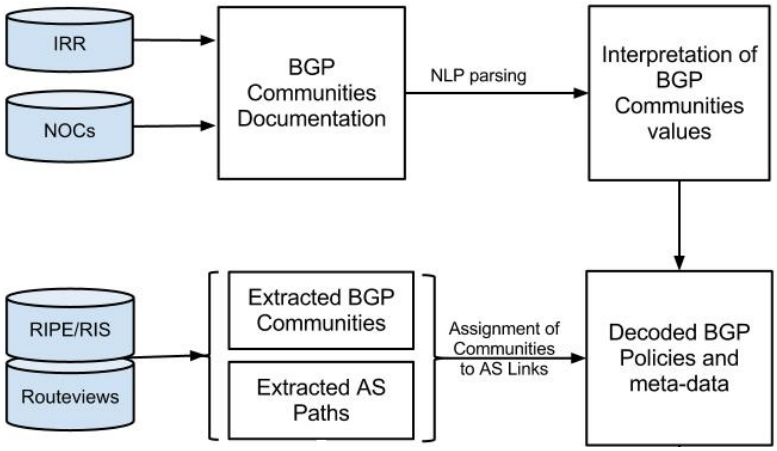


Figure 3.6: Methodology for the extraction policies from BGP Communities

of the Community action, and (iii) the actual action that determines how the policy is applied. For route tagging Communities I only extract the corresponding type and tag. Figure 3.5 illustrates how the documentation for the announcement Communities of Figure 3.3 will become.

Figure 3.6 presents the overall methodology for the extraction of policy data from BGP Communities.

3.2.2 Sanitisation of the Communities Documentation

To ensure that the correctness of the documented Community values I use three steps:

1. Sanity checks: Stub ASes are tagged only as customers, while Tier-1 ASes are never tagged as customers.

2. Cross-validation: For the AS links that I have relationship information from both ASes I test if they agree.

3. Consistency: Many ASes also provide Communities for triggering AS Path Prepending (ASPP) policies which are often conditioned in terms of relationship

Table 3.1: The number of unique Community values per Community category. The types marked with * indicate Communities set by the local AS to label a route, while the other types are set by the remote AS to request an action.

Community Type	Number of values
Announcement control	1910
*Geography	1305
*Relationship	1302
Prepending control	1204
*Point-of-Presence	446
*Origin	401
LocPref Control	208
Misc	120
Blackhole	18

types, e.g. "Prepend prefix 2x to upstream providers". I test whether ASPP control Communities affect only the paths where the next- hop AS has been tagged with the corresponding relationship Community. For example consider a path P_1: A - B - C where the link B - C is tagged by B as c2p. Consider also a non-prepended path P_2: C - B - D and a Community c which corresponds to 2x path prepending towards providers. If c is applied on P_2 I expect P_2 to become P_2^p : C - B - B - A. If that change does not happen I conclude that the Communities of the specific AS are not consistent.

Furthermore, to avoid stale information in IRR I only parse records which have been updated within one year of my measurement data (i.e. after June 2010). I am able to interpret 6,852 distinct Community values from 375 different ASes. Although the decoded Communities cover only 26% of all the observable Communities, they mostly belong to Tier-1 or Tier-2 ASes or large IXPs and cover 65% of the paths that are tagged with Community values (Figure 3.4). Table 3.1 lists the types of the decoded Communities. Note that some of the decoded Communities belong to more than one Community types. For example the Community 3491:200 corresponds to customers in North America which provides both relationship and geo-location information. Route redistribution Communities account for 50% of the decoded Communities, but - as expected - the informational Communities appear far more often in the AS paths (Figure 3.7).

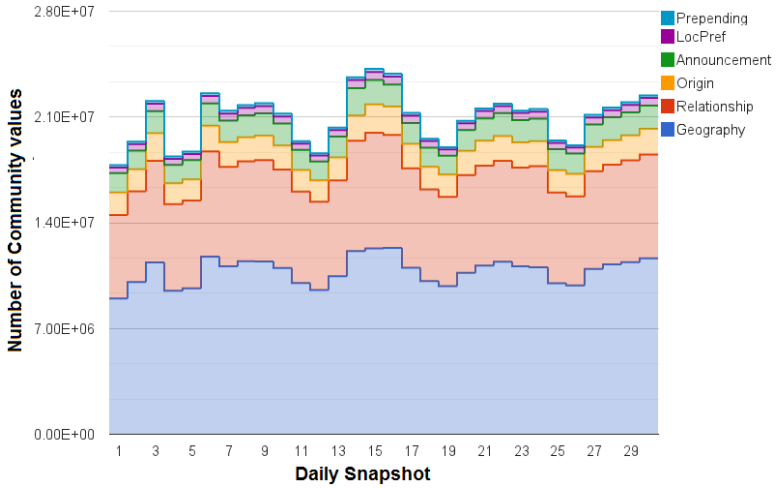


Figure 3.7: Daily number of BGP Communities per usage type

```
TYPE: TABLE_DUMP_V2/IPV4_UNICAST
PREFIX: 12.69.189.0/24
SEQUENCE: 1735
FROM: 206.223.115.61 AS4436
ORIGINATED: 06/01/11 06:00:00
ORIGIN: IGP
ASPATH: 4436 209 19581
NEXT_HOP: 206.223.115.61
MULTI_EXIT_DISC: 3
COMMUNITY: 209:209 209:13570 4436:903 4436:31611
```

Figure 3.8: Mapping of Community values to AS links. The ASPATH attribute gives a series of AS numbers (ASN) that form a routing path. The COMMUNITY attribute lists a series of 16-bit integer pairs in the format of $x : y$ where x is an ASN and y is an Community value.

3.2.3 Inference from BGP Communities

As shown in figure 3.7, relationship Community values is the second most widely used Community type. Relationship Communities are used as a flexible way to implement routing policies such as the valley-free rule [92]. This is a strong incentive for AS operators to encode the correct relationship information in BGP Communities, otherwise their policies will be affected resulting in policy violations.

Relationship Communities are applied at the ingress points of an AS network and indicate the relationship type with the neighbour from which a path advertisement is received. To infer the AS relationships from the BGP Communities I parse the COMMUNITY and the ASPATH attributes of the collected BGP records to extract the Community values that encode relationship information and assign them to the corresponding AS links. For example, in the BGP record of figure 3.8, the COMMUNITY attribute contains two values that encode relationship type: 209:209 tags customers of AS209,

```
 Network          Next Hop      Metric LocPrf Weight Path
*i64.69.219.0/24  216.218.252.164  1     100     0  4323 19271
*i64.71.0.0/19    66.220.23.178    1     140     0  33544
*i74.117.167.0/24 213.248.67.105   48     70     0  1299 7018
```

Figure 3.9: Three entries from the routing table of Hurricane Electric (AS 6939), obtained by querying its Looking Glass server. In this case values of the LocPrf attribute indicate the relationship types with the first AS in the Path. For example the highest value 140 corresponds to customers, the value 100 to peers and the lowest value 70 is assigned to routes learned from providers.

and 4436:903 tags peers of AS4436. By mapping these communities to the AS path I infer that AS4436 – AS209 is a p2p link, while AS209 – AS19581 is a p2c link.

I am able to infer 41,542 different relationships from BGP Communities which include 58% transit links, 41.5% peering and 0.5% sibling links.

3.3 The Local Preference Attribute

Local Preference (or LocPref) is a local attribute and is not included in the BGP announcements received by RouteViews and RIPE monitors. LocPref values can be obtained by having a direct interface to a BGP router. Remote access to such interfaces is provided through public Looking Glasses that allow remote execution of non-privileged BGP commands. For example, the command `show ip bgp <prefix>` dumps the full router table which includes the LocPrf attribute. Figure 3.9 shows three entries from the routing table of AS6939 obtained by querying its public Looking Glass server.

I collected weekly table dumps from 28 public Route Servers (that belong to 26 large ISPs) in the same periods of time as above (August 2010 and February 2011 respectively). I accumulate 12,441 links which contain 5,839 ASes.

3.3.1 Analysing LocPref Attribute Values

In the simplest case, an AS uses only three LocPref attribute values; the largest value (most preferable) is for the c2p relationship, the smallest value (least preferable) is for the p2c relationship and the middle is for the peering relationship [195].

However I observe that most ASes use many LocPref values. An extreme example is illustrated in Figure 3.10. For example customers can use Communities values to request for up-scaling or down-scaling their LocPref value for traffic engineering pur-

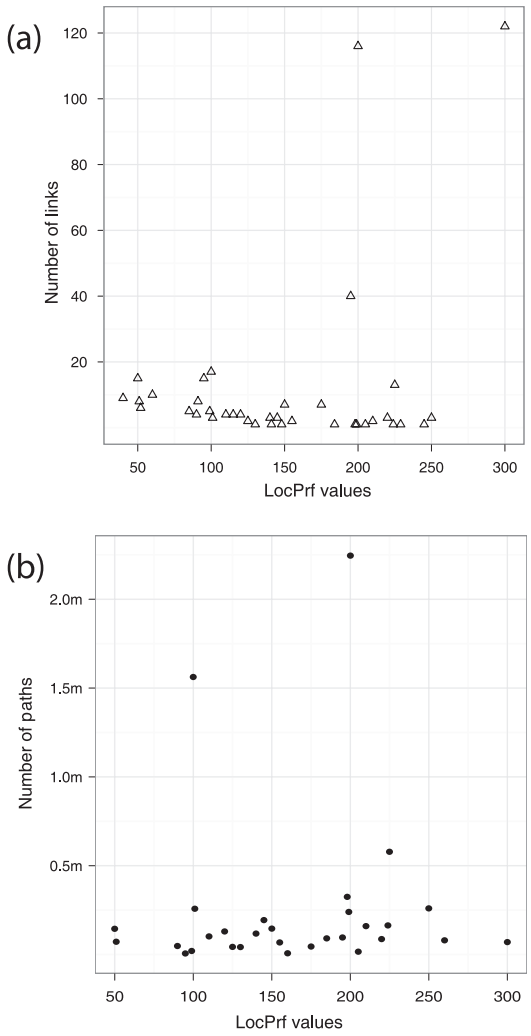


Figure 3.10: The appearance frequency of LocPref values of `AS 4436` in (a) AS links and (b) AS paths, respectively, in my BGP data.

poses. For each of such ASes, I try to identify the default LocPref values that are most frequently used:

1. For each LocPref value, I find out the number of links that the AS has assigned the value to. I also search for the number of AS paths in my BGP data that contain these links. I then calculate the distribution of links and paths, respectively, as a function of LocPref values (see Figure 3.10).

2. The LocPref values with the highest frequencies are chosen as the default values. I may choose more than three default values if their frequencies are significantly larger than the rest. This happens when two similar LocPref values are widely used for the same type of relationship with slightly different routing preference. In my work, I have chosen at most 5 default values.

3. I use the meaning of Communities attribute values obtained in the above to create a mapping between decide the relationship type of the default LocPref values. Usually the largest default value is for the c2p relationship and the smallest default value is for the p2c relationship.

In certain cases I can infer the meaning of more LocPref values based on the default values obtained from the above. For example if the majority of prefixes received from a peer AS are tagged with the default peer LocPref value and a few prefixes from the same AS are tagged with a slightly smaller LocPref value, and if this smaller value does not coincide the default transit value, I conclude that it is also a peer value with a reduced preference. I verify such conclusions against the Communities information and I discard any discrepancy. Note that the LocPref attribute values can only be used to infer transit and peering relationships.

3.4 Relation Inference based on BGP Attributes

A first application of the collected policy data from BGP attributes is to infer AS relationship directly from those attributes without involving further heuristics. I combined the inference results obtained from the Communities and LocPref attributes. As shown in Table 3.2, I am able to infer AS relationship for 43,155 links in total, which account for 39% of the links that are present in my BGP data. These links include all links among the Tier-1 ASes and most links between Tier-1 and Tier-2 ASes. A hybrid link is counted as both a transit link and a peering link. The partial-transit links and the backup links are included in the total number of transit links. When the relationship of a links is inferred from both BGP attributes, I only accept it if the two reach the same conclusion. It should be noted that the percentage of the inferred links decreases from 38.7% for August 2010 to 37.5% for February 2011. This happens because the topology growth between these two snapshots mainly happened with the addition of stub ASes at the edge of the AS topology while my Community data are mainly obtained by backbone ASes. This is manifested by the observation that in the BGP data for February 2011 there are 5,044 new ASes compared to August 2010, but only 5,208 new links.

I did not attempt to extract as many AS relationships as possible. Rather, my focus is to increase the certainty of the inferred AS relationships.

Table 3.2: AS Relationships Inferred From Routing Policies

	Aug 2010	**Feb 2011**
Number of paths	18,570,393	24,549,355
Number of AS links	111,511	116,719
Number of ASes	33,559	38,603
Number of inferred links	43,155	43,821
Number of ASes	16,877	16,918
Transit relationship	25,892	26,075
Peering relationship	17,996	18,603
Sibling relationship	176	177
Hybrid relationship	909	1,034
Indirect peering	708	811
Partial-transit relationship	1,526	1,828
Backup links	1,087	1,205
Inferred from Communities	36,340	38,130
Inferred from LocPref	12,441	12,602

- The Communities and LocPref attributes are configured by ASes themselves and are used by them in the BGP routing process. It is expected that ASes should use them to accurately reflect their business relationships.

- I collect the BGP data from the available sources that have been well studied and widely used. Some of the sources, e.g. the public Route Servers, are playing a crucial role in facilitating the Internet BGP routing.

- I cross-examine results obtained from different attributes or data sources. I discard any inconsistency or ambiguity from my results. This sometimes involves large amount of manual checks.

- I try to use as few heuristics as possible. When I have to use a heuristic, for example, to identify the default LocPref values, I make sure that the heuristic complies with engineering practice and supported by previous studies, and I impose safety checks.

The routing policy information enables me to reveal the following four special types of AS relationships that would not be discovered by existing inference heuristics.

3.4.1 The Hybrid Relationship

The traditional model of AS relationships assumes that two ASes have the same type of relationship for all the underlying physical connections. Hence, it is a 1-to-1 model that assigns one relationship type per AS pair. In reality AS interconnections can be more complex, resulting in a cases where two ASes agree different relationship types for different connections.

A hybrid relationship arises when two ASes agree to have both a peering relationship and a transit relationship. We identify two categories of hybrid links.

- IP version-dependent. Routing policies and paths for IPv4 traffic can differ significantly from those of IPv6. ASes often negotiate separate relationships for prefixes of different IP versions. Therefore two ASes may have a hybrid relationship if they are connected on both IPv4 and IPv6 planes.

- Location-dependent. The location of the Points-of-Presence (PoP) can affect AS relationships. Two ASes can have a hybrid relationship when they collocate at more than one private Network Access Points (NAP) or Internet eXchange Points (IXP). Figure 3.11 shows an example of a location-depended hybrid relationship.

- Some hybrid links are dependent on both IP versions and PoP locations. For instance, two ASes may have an IPv6 transit relationship at a private NAP and an IPv4 peering relationship at an IXP.

A hybrid relationship is identified when a same AS link is tagged with different sets of Communities values in *different* BGP Update messages. For example, consider the AS link AS3549 – AS3292. We observe that in a record from a RIPE monitor this link is tagged with the Communities 3549:2771 (route received from peer) and 3549:31208 (route received in Denmark), meaning that it is a peering relationship at a connection point in Denmark. Whereas in another record from a RouteViews monitor the same link is tagged with the Communities 3549:4354 (route received from customer) and 3549:30840 (route received in the USA), meaning that it is a transit relationship at a connection point in the USA.

It should be noted that if a link is tagged with different sets of Communities values in the *same* BGP Update message, we can not conclude it is a hybrid link. This

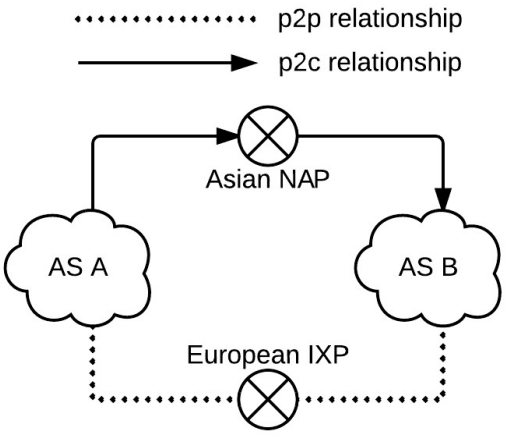


Figure 3.11: Example of hybrid relationships between AS A and AS B. The relationship for the AS link through the IXP is p2p, while the relationship for the link through the private NAP is p2c.

mainly happens when an AS specifies dual meanings to Communities values. For example, AS1273 uses the values 1273:1*** to tag customers (where 1*** means all numbers starting with 1) and it uses the values 1273:3*** to tag both providers and route prepending. When we observe a link tagged with both 1273:1*** and 1273:3*** in the same BGP record, we can only identify that it is not a hybrid link but a prepended p2c link after we learn (from the IRR and NOC data) that these Communities values are only settable by customers. Setting dual meanings for a Communities value is not a good practice but we observe thousands of such cases in my BGP data. When this happens, we only infer an AS relationship if sufficient extra information is available from other data. Otherwise we do not include it in my result.

As shown in Table 3.2, we discovered that 909 AS links have the transit/peering hybrid relationships in the August 2010 data. Although a small number, hybrid links are often between well-connected ASes. We observe that as high as 13% of AS links that carry both IPv4 and IPv6 traffic are hybrid links and more than 10% of all AS routing paths in the BGP data contain at least one hybrid link.

3.4.2 The Indirect Peering Relationship

The indirect peering relationship consists of two peering links, which together function as one 'virtual' peering link. It typically occurs when two ASes are peering with the same route server at an IXP such that they gain access to each other's network as if they have a peering link (without actually having a physical connection). We can

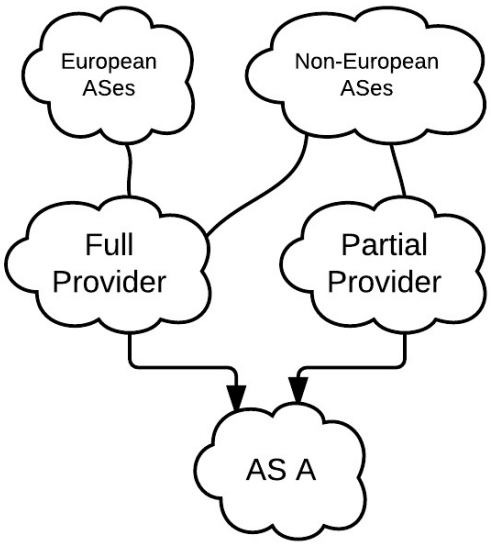


Figure 3.12: Example of the partial-transit relationship. AS A has a partial-transit relationship with the partial provider, which only transit its traffic to non-Europeean ASes.

detect this indirect peering relationship by the fact that both of the ASes tag the route server as a peering IXP.

Using the Communities data collected from BGP update messages, we discover that of the peering links, there are 708 peering links that can form 354 pair of indirect peering relationship. Each of the peering link can appear alone in their own routing paths. When two adjacent peering links form an indirect peering relationship, they do not violate the valley-free principle. From the prospect of Internet routing, these two peering links can be replaced by one peering link.

3.4.3 The Partial-Transit Relationship

A customer AS can multihome to more than one providers. The partial-transit relationship is a special case of the transit relationship where providers of a multihomed customer agree to offer transit within a limited geographical scope. A multihomed customer may use Communities values to instruct a national provider to serve traffic destined in the same country and an international provider to serve international traffic (Figure 3.12).

For example we observe AS3300 (as a provider) provides the customer-settable Communities value 3300:2100 which prevents a customer's route to be announced in Europe. A partial-transit link is only visible and used locally. Occasionally it can be fully activated (by the customer) if a provider of the customer fails (by setting relevant

Communities values.

3.4.4 The Backup Links

Backup links are usually invisible and they do not carry any traffic. When there is a disruption in the network, they are activated and become visible globally. But they will disappear once the network recovers. The backup links are not a new type of AS relationship. Rather they are *transit* links that have the backup function. Backup links are relevant to the Internet routing robustness and reliability. Backup links can be set in the following two ways. In my inference I identify both types of backup links.

When an AS has more than one available routes to a destination, it can set a default route and make other routes artificially less favourable. This is usually achieved by the traffic engineering technique of path prepending. AS path prepending is applied on at least 12% of the routes in the global routing table [128, 208], while it is possible that more routes with path prepending exist but are hidden because prepended paths become less preferable. The same technique can be used to create the backup links. The advantage is that such backup links can be automatically and instantly activated when the default route is disrupted. We identify a prepended backup link if the followings are satisfied: (a) it is a transit link; (b) the customer prepends the `AS_PATH` attribute such that the link is in an artificially longer path; and (c) we only observe the link for a short lifespan, e.g. less than 5 consecutive days in my monthly data.

Another technique to achieve backup links is the use of the Communities values of `NO-EXPORT` and `NO-ADVERTISE` that instruct a provider not to advertise the customer routes to anyone.

3.4.5 Analysis of Existing Algorithms

The following three existing algorithms are widely cited in the Internet research community.

1. PTE algorithm by Xia and Gao [202] The authors published their source code [20] but did not release any dataset. Here we examine a PTE dataset provided by [58].

2. CAIDA algorithm by Dimitropoulos et al. [76]. We examine three latest monthly datasets published by the authors for Nov. 2009 to Jan. 2010.

3. UCLA algorithm by Oliveira et al. [166]. We examine 31 daily snapshots published by the authors for August 2010.

The PTE and UCLA algorithms were based on AS topology data extracted from BGP data. The CAIDA algorithm was based on AS topology extracted from the traceroute data. They relied on various heuristics to infer AS relationships

3.4.5.1 Comparison.

I compare the AS relationship data inferred by my Attribute-Based (BAB) algorithm against those of the above algorithms. For each algorithm, I find the common links present in both my BAB data and their datasets, and I then evaluate their agreement for each type of AS relationships. I also consider the number of AS cycles in each dataset. If p2c links are represented as directed links, we can define an AS cycle as a sequence of p2c links where the last link leads to the first AS of the directed path. AS cycles should not exist on the Internet at all [129]. It is an indication of incorrect inference if AS cycles are present in the inference result. AS cycles create strongly connected components in the directed AS topology graph. I record the number of strong components and the size of the largest component. My results are shown in Table 3.3.

We can see the existing algorithms are not able to infer any of the special relationship types. Comparing with BAB data, the transit relationships inferred by CAIDA and PTE are only 43% and 70% correct, respectively. UCLA inflates the number of peering relationships and it contains a large number of AS cycles.

I also examine UCLA data from May to December 2009 (260 daily snapshots). I find that 46% links had changed their relationship type at least once in that period and 16% had changed more than once. It is possible that some ASes occasionally renegotiate and change their relationships, but intuitively such change should not happen to so many AS links so often.

The ability and correctness of the existing algorithms are inherently impaired by the fact that they solely reply on the AS path data, which does not contain sufficient useful information on AS relationships. Therefore they have to develop increasingly sophisticated heuristics to infer AS relationships from AS topology data. Here I show that the two important assumptions underlying their heuristics are problematic.

Table 3.3: Three past inference algorithms and comparison with the BGP attribute-based (BAB) algorithm

Existing algorithms (EA)		PTE	CAIDA	UCLA
	Number of ASes	31,843	33,499	38,269
	Number of AS links	143,102	80,128	141,651
	Transit links	94,632	73,246	79,037
	Peering links	48,470	6,654	62,614
	Sibling links	-	228	-
	Connected Comp. (max.size)	3(75)	3(694)	21(411)
Common links with BAB		25,552	25,168	38,409
Transit links	BAB	13,459	17,937	21,816
	EA	16,179	22,221	20,741
	agreement	12,490	17,777	20,084
Peering links	BAB	11,574	6,553	15,775
	EA	9,373	2,926	17,668
	agreement	8,247	2,816	14,068

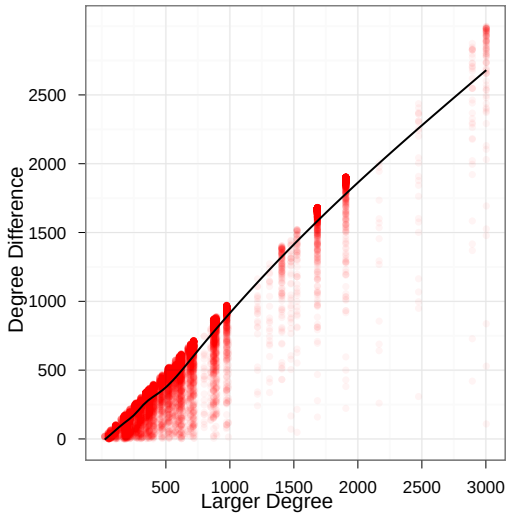


Figure 3.13: Distribution of the degree difference between ASes with the peering relationship.

3.4.5.2 The effect of degree difference on relationship misinferences.

Degree is defined as the number of links an AS has. The CAIDA and PTE algorithms assume that only ASes with "comparable" degrees can be possibly connected with a peering relationship; whereas ASes with 'very' different degrees tend to have a transit relationship.

Figure 3.13 shows the degrees of ASes with the peering relationship in my BAB data. It is obvious that peering ASes do not need to have similar degrees. In fact many

ASes with the peering relationship have large degree differences. This explains why CAIDA and PTE mistake many peering links as transit links. Also it indicates that the surprisingly large percentage (13%) of invisible customer links reported in [58] is probably an artifact of the inference methodology. These observations confirm the results of past studies that indicated limitations in traceroute datasets [117].

Figure 3.13 supports the intuition of the so-call 'Internet topology flattening' [99, 118], i.e. the connectivity of ASes are becoming independent from their relationships and hierarchy. Therefore, degree-based heuristics should be avoided as an AS relationship inference method.

3.5 Analysis of Valley-free violations

One of the most widely cited BGP properties is the valley-free rule [97], which is described in section 2.4. Fig.3.14) illustrates the patterns of valley-free and non valley-free paths.

To avoid loosing money, an AS usually avoids consuming network resources for transiting traffic between non-customer neighbours. The valley-free routing is a logical outcome of the economic model described by the AS relationships. It has been considered as a property of the entire Internet routing and it is only violated due to transient BGP configuration errors.

Detecting and analysing the valley-free violations is of particular importance for understudying the network economics and the inter-domain routing dynamics. The detection of non valley-free paths requires accurate AS relationship data which are usually not publicly disclosed.A fundamental assumption of the existing inference algorithms is the universality of valley-free paths. Most algorithms approach the relationship inference as an optimization problem where the goal is to infer AS relationships so that the number of valley-free paths is maximised [97, 41, 76], or try to extend some preknowledge about AS relationships assuming that all ASes along a path comply to the valley-free export policy [202, 197, 166]. As a result the available relationship datasets are biased towards valley-free paths and may introduce a significant number of false-negatives when used to detect valley-free violations.

I systematically mine the BGP Community attribute to construct a bias-free AS relationship dataset. I then use my AS relationship data to assess the valley-free path

violations. My assessment is conservative because I only identify the violations that can be confirmed by my AS relationship data, but there must be more violations that can not be identified yet. My results reveal a number of surprising findings.

- Valley paths are more than twice frequent than previously reported [173] .
- A large fraction of the valley-free violations (up to 50%) are not misconfigurations but intended policies from ASes that are either non-profit (e.g. research/educational networks) or follow a distinct economic model and establish relationships not described by the c2p/p2p/s2s model. Such valley paths persist for a long period of time, contrary to previous assumption that valley paths are resulted from transient configuration errors.
- IPv6 paths exhibit a disproportionately larger number of valley paths compared to IPv4 (5.4% in BGP tables and 22.3% in BGP Updates). The reasons for such large numbers of violations include the low traffic volumes, the configuration complexities and most of all the central role of non-profit/governmental organizations in todays IPv6 Internet.

Each of these finding is relevant to a better understanding of the BGP routing and should be appreciated in AS topology research and network operation.

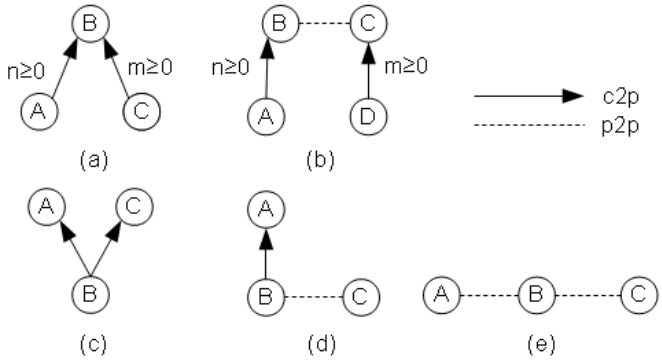


Figure 3.14: Patterns of valley-free paths (a, b) and examples of non valley-free paths (c, d, e). A valley-free path contains at most one p2p link, an 'uphill' leg of $n \geq 0$ c2p links and a 'downhill' leg of $m \geq 0$ p2c links. Any other patterns are non valley-free paths, or I called them valley paths.

3.5.1 Related Work

BGP routing and the AS topology have been widely studied research subjects. The valley-free rule has been coined by Gao in her seminal work on AS relationships [97].

A number of significant works followed up to propose relationship inference algorithms based on the valley-free rule [191, 41, 76, 202, 197, 166]. The most recent of these algorithms recognize that valley paths may exits as result of BGP configuration errors and propose filtering based on the persistency of the AS paths, although no standard methodology exists.

The first approach to study export policy violations has been provided by Mahajan et. al. [151] which was based on Gao's algorithm to infer the AS relationships. They presented the most common cases of valley-free violations due to transient misconfigurations. They also showed that their analysis contains only a limited number of false-positives due to relationship misinferences. However, their results may missed a large number of false-negatives in detecting the violations due to relationships misinferences. Gao's algorithm mistakes many p2p relationships as p2c, and as shown in section 3.5.3 a large number of valley paths would not be detected if p2p relationships were substituted by p2c. The last observation is confirmed by [84] which showed that the valley-free maximization problem can be solved using only p2c relationships.

Feamster et.al. [89] highlighted the need for a solution that will enable filtering of the erroneous valley announcements without revealing relationship information. Such a solution was proposed in [173] as a BGP extension, where the authors also provide a measurement of the valley paths based on Gao's algorithm. In [203] the authors observed that export misconfigurations may affect the levels of traffic received by an AS and therefore proposed a filtering mechanism based on the fluctuations of the traffic volumes.

Unlike the previous works, I do not approach the valley paths as a misconfiguration problem but I try to investigate whether alternative models of export configurations exist. My results complements some of the conclusions of [179] that there is no universality in Internet routing due to the complexity of the Internet ecosystem.

3.5.2 Methodology

Analysis of valley-free paths can be very sensitive to relationship misinferences depending on the centrality of the misinferred links. To ensure that my relationships data are as free from wrong relationships as possible, for the analysis of valley-free paths I only use relationships extracted from the BGP Communities attribute. As I

Table 3.4: Inference results based on BGP Community data

		IPv4	IPv6
BGP Tables	ASes	38,654	4,540
	AS links	112,501	18,210
	AS paths	6,170,769	355,366
	Identified valley paths	102,741	19,210
	Fraction of valley paths	1.7%	5.4%
BGP Updates	ASes	38,734	4,569
	AS links	122,828	19,217
	AS paths	16,623,077	902,618
	Identified valley paths	676,355	201,578
	Fraction of valley paths	4.0%	22.3%
AS relationships inferred from Community		38,433	9,620

mention in section subsec:discussion, Local Preference offers indirect relationships information while BGP Communities offer direct relationship information. Even though I apply a series of checks to sanitize the Local Preference values I want only direct relationship information for the assessment of the valley-free rule. For the BGP Communities we follow the inference methodology described in sections 3.4 and 3.2.3, for RouteViews and RIPE RIS data that span June 2011.

As shown in Table 3.4, to better understand the effects of the valley paths on BGP routing, I split the BGP data into two sets. The first contains AS paths and AS links extracted from BGP tables and the second contains the corresponding information extracted from BGP updates. BGP tables include only the best AS path towards a certain prefix, while the updates include additional AS paths that are not selected as active.

3.5.3 Results

By definition, single-hop paths can only be valley-free. I identify an AS path as a non valley-free path, or valley path, if the AS relationship of any two adjacent links in the path are known and they follow one of the following relationship patterns: (1) $p2c - c2p$ (2) $p2p - p2p$ (3) $p2c - p2p$ (4) $p2p - c2p$. Since my AS relationship dataset does not cover all the AS links, my assessment of the valley paths are conservative, i.e. the real number of valley paths must be even larger than what I have identified.

Table 3.4 shows the number of AS paths, the number of identified valley paths, and the ratio of valley paths to all paths. I observe a high number of valley-free violations present both in tables and updates datasets. The percentage of valley paths in IPv4 BGP Updates is more than twice of previously reported in [173] (4% compared to 1.4%), highlighting the effect of my bias-free relationship data in revealing more valley paths.

An interesting finding is the exceptionally large number of valley-free violations in IPv6 paths, especially in the BGP Updates. Currently, IPv6 is only partly deployed and the IPv6 traffic levels are less than 0.5% of all the Internet traffic [139]. The low popularity and traffic levels of IPv6 mean that IPv6 economics can be considerably different from IPv4, from which ASes generates most of their profit. Moreover, in order to encourage IPv6 connectivity a large number of large governmental and non-profit networks have been globally deployed. Such networks do not aim to maximize their profit from traffic exchange, and therefore they may choose BGP policies for improving IPv6 reachability instead of complying with the valley-free rule.

3.5.3.1 Persistency of Valley Paths

The valley paths have been considered as transient configuration errors that should last only for a limited time. To test this assumption I study the persistency (a) of all the AS path (Fig. 3.15), and (b) of the valley paths identified by my BGP Community data (Fig. 3.16).

The IPv4 valley paths follow roughly the same persistency pattern as all the AS paths. On the other hand, IPv6 valley paths exhibit a higher fraction of short-lived paths attributed to export misconfigurations. This difference between IPv4 and IPv6 in policy errors is expected due to the complexity and largely experimental nature of IPv6 networks. For the IPv6 BGP Table data, 30% of the valley paths persist for more than 24 days while 40% appear for less than four consecutive days. What is remarkable is that for the IPv4 BGP Tables data, more than half of the valley paths persist for the entire month while only 19% have a lifetime less than a week. This finding implies that a large fraction of valley announcements are the deliberate BGP policies followed by ASes whose primary concern is not commercial profit. I call the long-lasting valley paths the persistent valley paths.

To further investigate such paths, I infer the function of the relevant ASes by pars-

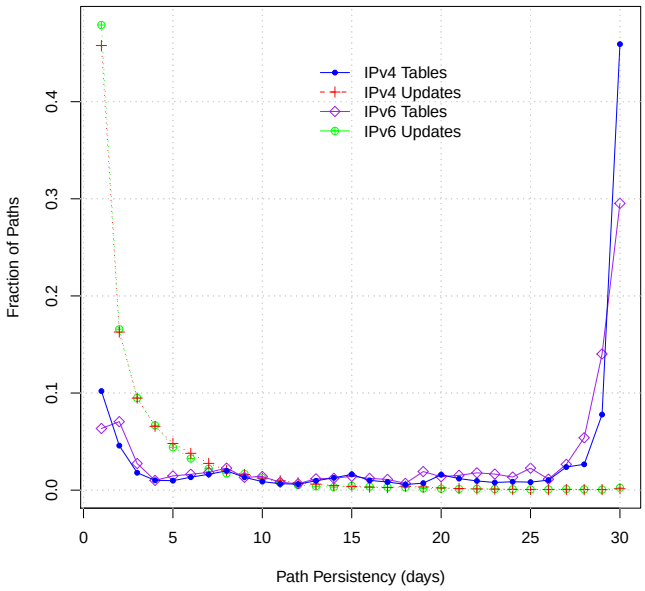


Figure 3.15: Distribution of the AS paths as a function of their persistency, where the persistency is measured as their largest number of consecutive days of appearance in my BGP data.

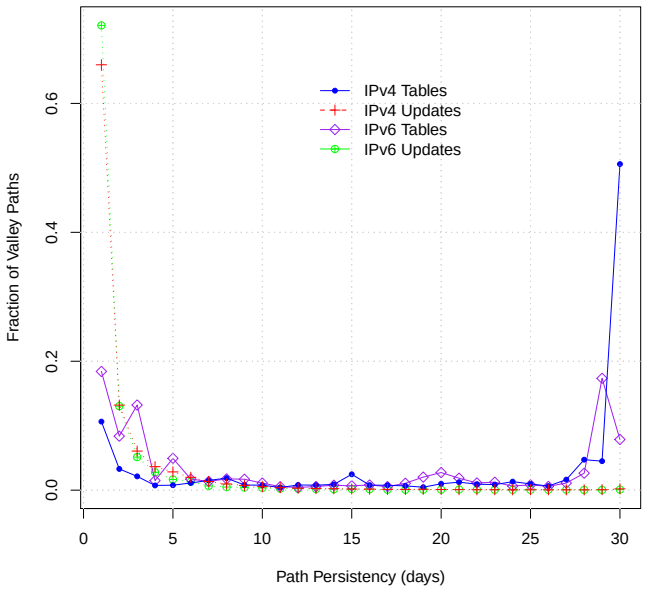


Figure 3.16: Distribution of the valley paths as a function of their persistency.

ing keywords in its WHOIS record [77]. There are four major types of functions: Internet provider, content provider, research/educational, and Internet Exchange Point (IXP). I find most persistent valley paths are due to the last two types of ASes. About

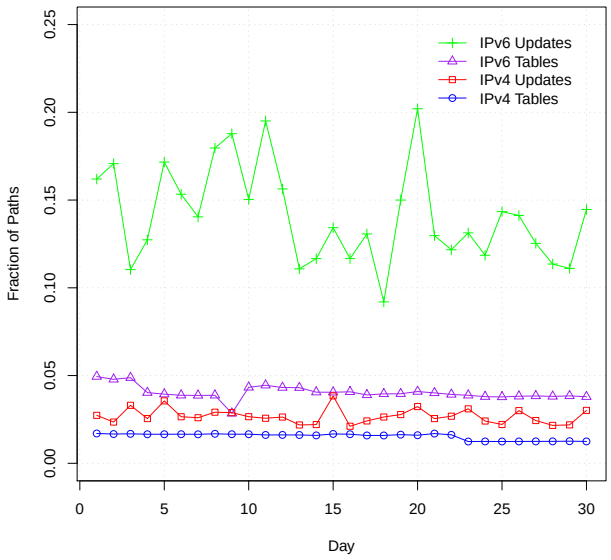


Figure 3.17: Fraction of valley paths per day. With sole exception the IPv6 Updates, the fraction of valley paths is stable throughout June 2011.

53% of the persistent valley routes are contributed by research/educational ASes, e.g. AS7575, AS11537 and AS7660, which are mostly non-profit with the purpose to unify university networks and facilitate their communication. Such ASes often establish a special type of AS relationship called *indirect peering*, where an AS functions as a mediate link between two other ASes who wish to peer but have common Point-of-Presence (PoP) only with the mediate AS but not between each other. Such relationships are often perceived as p2c by the previous relationship reference algorithms, which result in a large number of false-negatives when studying the valley paths. IXPs, e.g. AS4635 and AS6695, also contribute to a significant portion of the persistent valley paths (21%). Specifically, in many cases of public peering at IXP route servers, the AS number of an IXP is injected in the AS paths advertised to the IXP members. This is in fact another type of indirect peering since the IXP operates as a link between two ASes who are not directly linked but agree on a p2p relationship.

Valley paths also exhibit a relative stability over time, as shown in Fig. 3.17. Only the valley paths in IPv6 BGP Updates exhibit large fluctuations, indication of the high number of transient errors in IPv6 Updates. On the contrary, the fraction of valley paths in IPv4 and IPv6 Tables are remarkably stable, which highlights the baseline of consistent valley announcements.

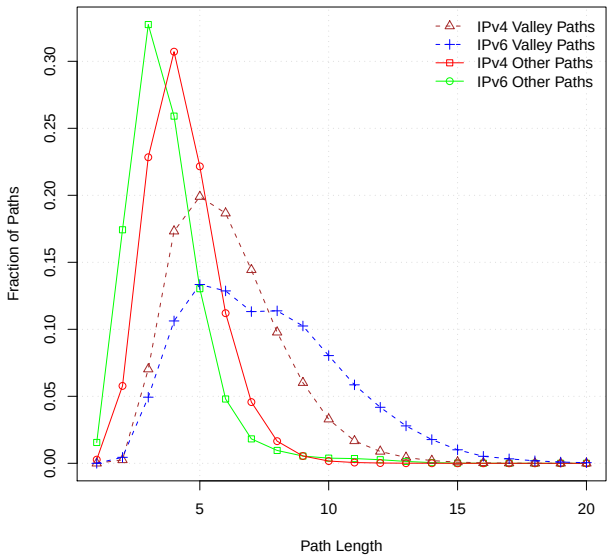


Figure 3.18: Distribution of paths as a function of path length (in hops). I compare the valley paths against all other paths (not identified as valley paths) for IPv4 and IPv6 respectively.

3.5.3.2 Length of Valley Paths

Figure 3.18 shows the distribution of (non-prepended) path length for the identified valley paths and all the other paths (not identified as valley paths). Valley paths are generally longer than valley-free paths. Paths with a length of over 10 hops are exclusively valley paths. The main reason is that the valley-free rule is essentially an export filter that blocks the advertisement of AS paths if they are received from specific types of neighbours. For instance, a customer will not forward a path from a peer to another peer if it obeys the valley-free rule. When the valley-free rule is absent, further forwarding of such routes is possible which results in paths with more hops.

The longer length of the valley paths provides an explanation of why BGP updates contain more valley paths than BGP tables . The path length is a BGP tie-breaking metric for route selection between routes that are equally preferable. For example when an AS receives two routes from the same category of neighbours with the same Local Preference value, it is more likely to choose a shorter valley-free path (which then appears in BGP table as the preferred route) than a longer valley path (in BGP update as an alternative route). Also, some of the valley paths in the BGP Updates are too short-lived to be captured by the BGP Table snapshots.

3.6 The IPv6 AS Relationships

The existing ToR algorithms analyze the IPv4 and IPv6 AS links using exactly the same principles. However, the AS links carrying IPv6 traffic may follow unconventional BGP routing policies, including relaxed peering requirements and even free IPv6 transit. These distinct policies may result in AS links with different relationship type between the IPv4 and the IPv6 Internet. Such relationships are called *hybrid* IPv4/IPv6 relationships and cannot be captured by the existing ToR algorithms. Hence, measurement artifacts are unavoidable under the current ToR inference approaches.

To rigorously analyse the IPv6 AS relationships and detect the hybrid relationships I rely on the BGP Communities relationship information, and the Local Preference (LocPrf) attribute. I utilize the metric of "customer tree" [76] to assess the impact of hybrid links on the IPv6 routing structure.

In August 2010, there are 346,649 IPv6 AS paths and 10,535 IPv6 AS links. 7,618 IPv6 AS links are also visible in the IPv4 topology. From the Community and the Local Preference attributes I am able to extract the actual AS relationship for 72% (7,651) of the all IPv6 links and for 81% (6,160) of the IPv4/IPv6 links. A number of interesting results can be observed from those data.

Firstly, 779 (or 13%) of the IPv4/IPv6 links have hybrid AS relationships. 67% of such hybrid links have a peering relationship for IPv4 and a transit relationship for IPv6; the rest are p2p for IPv6 and p2c for IPv4, except a single case where the two ASes have a p2c for IPv4 and a c2p for IPv6.

Secondly, the hybrid links usually happen among tier-1 or tier-2 ASes with large numbers of connections. As a result the hybrid links have a high visibility in IPv6 AS paths. More than 28% of the IPv6 paths contain at least one IPv4/ IPv6 link with hybrid AS relationships.

Thirdly, 13% of the IPv6 paths do not follow the valley-free rule (*valley* paths). The large number of IPv6 valley paths is a major reason underlying the inference errors of the existing ToR algorithms. My analysis indicates that 16% of the valley paths are due to the relaxation of the valley-free rule in order to expand the reachability of IPv6 prefixes. The IPv6 topology is partitioned in terms of valley-free routing [1]

[1] An example is the peering dispute between two transit-free ASes in the IPv6 plane, AS6939 and AS174, as described in [143]

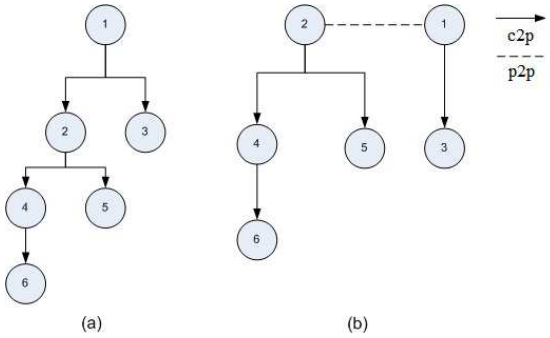


Figure 3.19: The change in the customer tree when the link 1–2 is (a) p2c or (b) p2p. In (a) AS1 can reach all the nodes through p2c links, while in (b) it can reach only AS3 through a p2c link.

and the relaxation of the valley-free rule is necessary in some cases to maintain IPv6 reachability.

The substantial existence of hybrid IPv4/IPv6 links suggests that the IPv4 and IPv6 Internet topologies should be studied separately. This is consistent with a recent study on the evolution of IPv4 and IPv6 AS topologies [207].

The ToR-annotated AS topology is very sensitive to misinference of AS relationship. An expressive metric to assess the impact of misinferred relationships is the "customer tree" of an AS (root), which contains all the ASes that the root can reach through p2c links. Figure 3.19 shows an example where the change of relationship between two ASes results in two different topologies. I find that when I replace the IPv6 AS relationships that were misinferred in [166] with the correct relationships inferred from the BGP Communities, the average length and the longest length (diameter) of the shortest valley-free AS paths of the union of the IPv6 customer trees are reduced from 3.8 to 2.23, and from 11 to 7 hops, respectively (Figure 3.20).

In summary, my results reveal substantial differences between the IPv4 and IPv6 relationships, including a significant number of IPv6 paths that are not valley-free, sometimes in exchange for better reachability. The IPv6 topology should be studied separately using new models that capture its distinct characteristics.

3.7 Summary

This section presented a new measurement framework to systematically extract BGP policy data from two attributes, the Communities which is used to tag BGP routes with additional meta-data, and the LocPref which is used to rank the AS paths. The

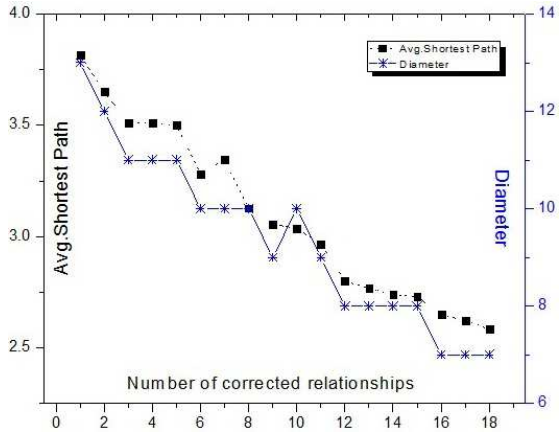


Figure 3.20: The change of the average shortest path and the diameter of the IPv6 AS customer trees as I gradually correct the misinferred relationship of the 20 hybrid AS relationships with the highest visibility in the IPv6 AS paths.

collected data provide a more informative data source that can be used to study the AS relationships and the derived policies without relying on heuristics. This methodology allows the inference of about 40% of the AS relationships that cover the majority of the AS links among the top-tier ASes.

The data extracted from the BGP attributes leads to a new AS relationships categorization and the extension of the valley-free rule, both of which are key features of inter-domain routing models.

- The existing model of AS relationships assign one relationship type to each AS link from three abstract relationship categories, customer-to-provider (c2p), peer-to-peer (p2p), and sibling-to-sibling (s2s). Each of these relationship types expresses a set of economic and routing rules and determine how paths are advertised and selected. However, these relationship abstractions cannot faithfully describe all the actual routing policies used and can introduce significant artifacts. Accordingly, an extended relationship categorisation has been proposed that can provide an accurate representation of complex routing policies. Additionally, I provided evidence that relationships to AS links do not always follow a one-to-one mapping, but often more than one relationship types can be agreed by two ASes for different geographical locations or for different types of traffic.

- The valley-free rule defines path patterns that allow ASes to minimize their routing costs through selective announcement of BGP routes. The valley-free rule has been widely perceived as a universal property of the Internet BGP routing

that is only violated due to transient configuration errors. Consequently, it has been used as the ground for studying the AS topology and inter-domain routing. Even though it is prevalent policy for advertising reachability information, I find that the valley-free pattern is intentionally violated by ASes that follow distinct economic models. Valley-free violations happen at least two times more often than previously estimated. Especially for the IPv6 topology, non valley-free paths appear in 20% of the paths in BGP Updates and more than 5% paths in BGP tables.

Chapter 4

Inference of Conventional AS Relationships

The Internet consists of thousands of independent interconnected organizations, each driven by their own business model and needs. The interplay of these needs influences, and sometimes determines, topology and traffic patterns, i.e., connectivity between networked organizations and routing across the resulting mesh. Understanding the underlying business relationships between networked organizations provides the strongest foundation for understanding many other aspects of Internet structure, dynamics, and evolution. This chapter presents a novel algorithm for the inference of conventional AS relationships that is based on the insights gained by the preliminary inference. The algorithm has been validated against three datasets to 99% accuracy. This algorithm will provide the foundation for the inference of complex relationships that will be presented in chapter 5.

4.1 Introduction

Business relationships between ASes, which are typically congruent with their routing relationships, can be broadly classified into two types: customer-to-provider (c2p) and peer-to-peer (p2p). In a c2p relationship, the customer pays the provider for traffic sent between the two ASes. In return, the customer gains access to the ASes the provider can reach, including those which the provider reaches through its own providers. In a p2p relationship, the peering ASes gain access to each others' customers, typically without either AS paying the other. Peering ASes have a financial incentive to engage in a *settlement-free* peering relationship if they would otherwise

pay a provider to carry their traffic, and neither AS could convince the other to become a customer. Relationships are typically confidential so must be *inferred* from data that is available publicly. This paper presents a new approach to inferring relationships between ASes using publicly available BGP data.

Measurement and analysis of Internet AS topologies has been an active area of research for over a decade. While yielding insights into the structure and evolution of the topology, this line of research is constrained by systematic measurement and inference challenges [179]. First, the BGP-based collection infrastructure used to obtain AS-level topology data suffers from artifacts induced by misconfigurations, poisoned paths, and route leaks, all of which impede AS-relationship inference. The algorithm incorporates steps to remove such artifacts. Second, AS topologies constructed from BGP data miss many peering links [28]. This lack of visibility does not hinder the accuracy of inferences on the observed links. Third, most AS-relationship algorithms rely on "valley-free" AS paths, an embedded assumption about the rationality of routing decisions that is not always valid [179], and which the algorithm does not make. Fourth, import and export filters can be complicated; some operators caveat their c2p links as being region or prefix specific. However, they still describe themselves as customers even if they do not receive full transit. Therefore, a c2p inference is made when any transit is observed between two ASes. Relationship inferences can still be c2p or p2p with the caveat that the c2p relationship may be partial. By developing filtering techniques the effects of such *hybrid relationships* are mitigated when computing an AS's customer cone, described later in this section. Finally, a single organization may own and operate multiple ASes; the inference of *sibling relationships* is left as future work because it is difficult to distinguish them from route leaks.

The proposed algorithm does not seek to maximize the number of valley-free (hierarchical) paths, since at least 1% of paths are non-hierarchical (section 4.2.6). Instead, inferences are made using AS path triplets (adjacent pairs of links) which allow to ignore invalid segments of paths that reduce the accuracy of valley-free path maximization approaches.

The algorithm relies on two key assumptions: (1) an AS enters into a c2p relationship to become globally reachable, i.e. their routes are advertised to their provider's providers, and (2) there exists a *clique* of ASes at the top of the hierarchy that obtain

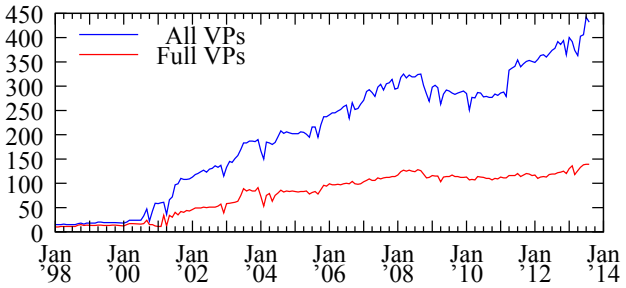


Figure 4.1: Number of ASes providing BGP data to Route Views and RIS over time. Currently, a third of all contributors provide a full view. The number of ASes providing a full view has not grown since 2008.

full connectivity through a full mesh of p2p relationships.

4.2 Data

This section presents the sources of data used: public BGP data, a list of AS allocations to RIRs and organizations, and multiple sources of validation data.

4.2.1 BGP Paths

BGP paths are derived from routing table snapshots collected by the Route Views (RV) project [21] and RIPE's Routing Information Service (RIS)[16]. Each BGP peer is a *vantage point* (VP) as it shows an AS-level view of the Internet from that peer's perspective. For each collector, one RIB file per day is downloaded between the 1st and 5th of every month since January 1998, and then the AS paths that announce reachability to IPv4 prefixes are extracted. Paths that contain AS-sets and compress path padding (i.e. convert an AS path from "A B B C" to "A B C") are discarded. All AS paths that are seen in any of the five snapshots are recorded, and their union is used to subsequently infer relationships. This means that all paths are used and not just "stable" paths because backup c2p links are more likely to be included if all AS paths are used, and temporary peering disputes may prevent a normally stable path from appearing stable in the five-day window of data.

Figure 4.1 shows the number of ASes peering with RV or RIS between 1998 and 2013, and the number providing full views (routes to at least 95% of ASes). RV is the only source that provides BGP data collected between 1998 and 2000, and while more than two thirds of its peers provided a full view then, it had at most 20 views during these three years. For the last decade, approximately a third of contributing

ASes provide a full view. Most (64%) contributing ASes provide routes to fewer than 2.5% of all ASes. The operators at these ASes likely configured the BGP session with the collector as p2p and therefore advertise only customer routes.

4.2.2 Allocated ASNs

To identify valid AS numbers assigned to organizations and RIRs, IANA's list of AS assignments [3] has been used. BGP paths that include unassigned ASes are filtered out, since these ASes should not be routed on the Internet.

4.2.3 Validation Data Directly Reported

CAIDA's website provides the ability to browse the relationship inferences and submit validation data. There are two separate datasets inferred from these corrections, one created in January 2010 and the other in January 2011 using an older relationship inference algorithm. This older algorithm did not produce a cycle of p2c links but assigned many more provider relationships than ASes actually had; 93% of the website feedback consists of p2p relationships, and 62% of that consists of correcting the inference of a c2p relationship to a p2p relationship. In total, feedback was received 129 c2p and 1,350 p2p relationships from 142 ASes through the website.

Follow up communication with the operators aimed to clarify unchanged inferences, as well as submissions that seemed erroneous compared to observations in public BGP paths. More than 50 network operators responded to the follow up emails. 18 relationships submitted through the website were later acknowledged by the submitting operator to have been inaccurately classified (9 by one operator) or to have changed subsequent to the submission. Additionally, based on email exchanges with operators, a file containing 974 relationships was assembled – 285 c2p and 689 p2p (contained within the "directly reported" circle of figure 4.2, described further in section 4.2.6).

4.2.4 Validation Data Derived from RPSL

Routing policies are stored by network operators in public databases using the Routing Policy Specification Language (RPSL) [29]. The largest source of routing policies is the RIPE WHOIS database, partly because many European IXPs require operators to register routing policies with RIPE NCC. The routing policy of an AS is stored as part of the *aut-num* record [29]. The aut-num record lists import and export

rules for each neighbour AS. An import rule specifies the route announcements that will be accepted from the neighbour, while an export rule specifies what routes will be advertised to the neighbour. The special rule ANY is used by an AS to import/export all routes from/to the neighbour, and is indicative of a customer/provider relationship. Using the RIPE WHOIS database from April 2012, a set of c2p relationships has been extracted using the following method: if X has a rule that imports ANY from Y, then a c2p relationship is inferred if a similar rule is observed in Y's aut-num that exports ANY to A. Limiting the collection process to records updated between April 2010 and April 2012 to ensure the "freshness" of data, RPSL records contributed 6,530 c2p links between ASes with records in RIPE NCC's database.

4.2.5 Validation Data Derived from Communities

AS relationships can be embedded in BGP community attributes included with each route announcement. Community attributes can be *tagged* to a route when it is received from a neighbour. Community attributes are optional transitive attributes; they can be carried through multiple ASes but could be removed from the route if that is the policy of an AS receiving the route [54]. The tagging AS can annotate the route with attributes of the neighbour and where the route was received, and the attributes can be used to control further announcements of the route. The commonly used convention is for the tagging AS to place its ASN, or that of its neighbour, in the first sixteen bits. The use of the remaining 16 bits is not standardized, and ASes are free to place whatever values they want in them. Many ASes publicly document the meaning of the values on network operations web sites and in IRR databases, making it possible to assemble a dictionary of community attributes and their policy meanings. I used a dictionary of 1286 community values from 224 different ASes assembled from [104] to construct a set of relationships from BGP data for April 2012; in total, there are 41,604 relationships in the collected dataset (16,248 p2p and 23,356 c2p).

4.2.6 Summary of validation data

Figure 4.2 uses a Venn diagram to show the size and overlap of the validation data sources. Overall, 2203 of 2225 relationships that overlap agree (99.0%), with multiple explanations for the discrepancies. For the directly reported source, some operators reported a few free transit relationships as peering relationships, i.e., they were

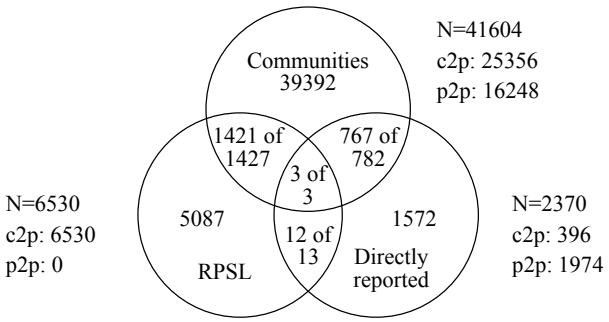
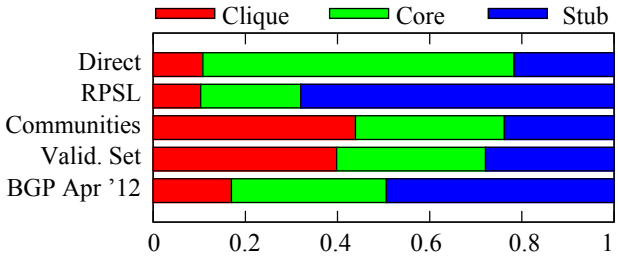


Figure 4.2: Summary of validation data sets collected and agreement among sets (first number inside intersections is number of overlapping relationships that agree). Overall, 2203 of 2225 relationships agree (99.0%) suggesting a limit on the accuracy of any source of validation data.

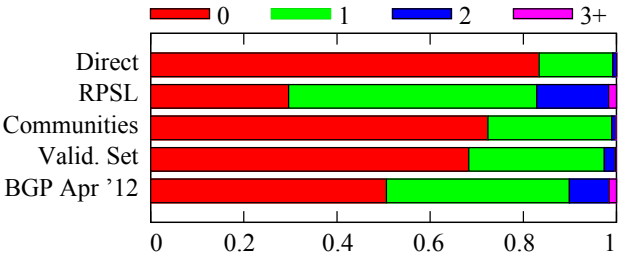
reported in the traditional economic sense rather than in a routing sense. For the RPSL source, some providers mistakenly imported all routes from their customers and some customers mistakenly exported all routes to their providers. For the BGP communities source, some customers tagged routes exported to a Tier-1 AS as a customer. While there are limits to the accuracy of all the sources of validation, the 99.0% overlap in agreement gives confidence in using them for validation.

A validation data set has been assembled by combining all four data sources in the following order: (1) directly reported using the website, (2) RPSL, (3) BGP communities, and (4) directly reported in an email exchange. Where a subsequent source classified a link differently, the classification is replaced; relationships acquired through email exchanges are trusted more than relationships submitted via the website. The validation data set consists of 48,276 relationships: 30,770 c2p and 17,506 p2p.

To estimate possible biases in the validation data set, their characteristics are compared with those of the April 2012 BGP dataset in terms of the link types and the minimum distance from a VP at which those links were observed. The closer the link is to a vantage point, the more likely it is to see paths that cross it. Links are classified as follows: *clique* (one endpoint is in the clique), *core* (both endpoints are not in the clique and are not stubs), and *stub* (one endpoint is a stub and the other endpoint is not in the clique). Figure 4.3(a) shows *clique* links are over-represented in the validation data set as compared to BGP data, while *stub* links are under-represented. This disparity is due to the validation data from BGP Communities, which mostly comes from large ASes. Figure 4.3(b) shows links directly connected to a VP (distance 0) are over-represented



(a) Link classification.



(b) Minimum link distance from VP.

Figure 4.3: Characteristics of validation data. Relative to BGP data, *clique* links and links directly connected to VPs are over-represented and *stub* links are under-represented.

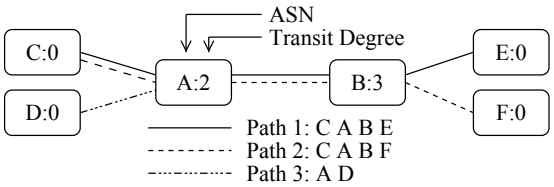


Figure 4.4: Computing the transit degree of ASes using paths. While the node degrees of ASes A and B are both three, A's transit degree is two because A it is not observed to announce D's prefixes to any neighbours. Nodes with a transit degree of zero (C, D, E, F) are stub ASes.

in the validation data relative to April 2012 BGP data, likely due to the Communities dataset, many of which involve ASes that provide VPs.

4.3 Inference Algorithm for Conventional AS Relationships

First I present the algorithm for inferring conventional relationships of type c2p and p2p. This algorithm will be extended in chapter 5 for the inference of complex relationships.

Two metrics of AS connectivity are used: the *node degree* is the number of neighbours an AS has; and the *transit degree* is the number of unique neighbours that appear on either side of an AS in adjacent links. Figure 4.4 illustrates the transit degree metric;

Algorithm 1 AS relationship inference algorithm.

Require: AS paths, Allocated ASNs, IXP ASes
1: Discard or sanitize paths with artifacts (§4.3.3)
2: Sort ASes in decreasing order of computed transit degree, then node degree (§4.3)
3: Infer clique at top of AS topology (§4.3.4)
4: Discard poisoned paths (§4.3.3)
5: Infer c2p rels. top-down using above ranking (§4.3.5)
6: Infer c2p rels. from VPs inferred not to be announcing provider routes (§4.3.5)
7: Infer c2p rels. for ASes where the provider has a smaller transit degree than the customer (§4.3.5)
8: Infer customers for ASes with no providers (§4.3.5)
9: Infer c2p rels. between stubs and clique ASes (§4.3.5)
10: Infer c2p rels. where adjacent links have no relationship inferred (§4.3.5)
11: Infer remaining links represent p2p rels. (§4.3.5)

ASes with a transit degree of zero are stub ASes. Transit degree is used to initially sort ASes into the order in which their relationships is inferred, breaking ties using node degree and then AS number. ASes inferred to be in the clique are always placed at the top of this rank order. Sorting by transit degree reduces ordering errors caused by stub networks with a large peering visibility, i.e., stubs that provide a VP or peer with many VPs.

4.3.1 Assumptions

The algorithm relies on three assumptions:

Clique: multiple large transit providers form a peering mesh so that customers (and indirect customers) of a transit provider can obtain global connectivity without multiple transit provider relationships.

A provider will announce customer routes to its providers. All ASes, except for those in the clique, require a transit provider in order to obtain global connectivity. It is assumed that when X becomes a customer of Y, that Y announces paths to X to its providers, or to its peers if Y is a clique AS. Exceptions to this rule include backup and region-specific transit relationships.

The AS topology can be represented in a directed acyclic graph. Gao *et al.* argue there should be no cycle of p2c links to enable routing convergence [98].

4.3.2 Overview

Algorithm 1 shows each high-level step in the conventional relationship inference

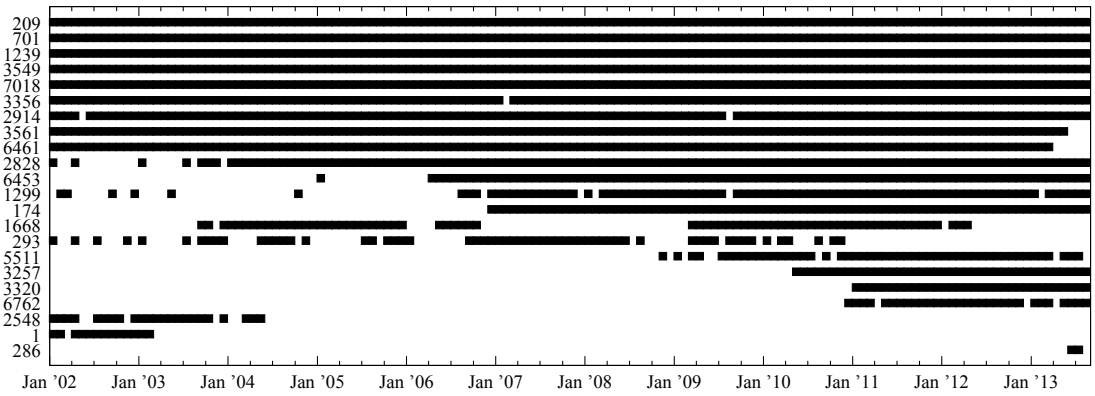


Figure 4.5: ASes inferred to be in the clique over time. The plot shows the 22 of the 26 ASes inferred to be in the clique at any time after January 2002. The clique's small size and consistent membership lend confidence in the inference methodology. As an artifact of the AS3356/AS3549 merger process in 2013, clique member AS6461 was inferred not to be in the clique.

algorithm. First, the input data are sanitised by removing paths with artifacts, i.e., loops, reserved ASes, and IXPs (step 1). The resulting AS paths are used to compute the node and transit degrees of each AS, and produce an initial rank order (step 2). Then, the algorithm infers the clique of ASes at the top of the hierarchy (step 3). After filtering out poisoned paths (step 4), a sequence of heuristics is used to identify c2p links (steps 5-10). All remaining links that are unclassified by this step are inferred as p2p. The algorithm contains many steps, a consequence of trying to capture the complexity of the real-world Internet and mitigate limitations of public BGP data [179]. The output from this process is a list of p2c and p2p relationships with no p2c cycles, by construction.

4.3.3 Filtering and Sanitizing AS Paths

The first step is to sanitize the BGP paths used as input to the algorithm, especially to mitigate the effects of BGP path poisoning, where an AS inserts other ASes into a path to prevent its selection. A poisoned path implies a link (and thus relationship) between two ASes, where in reality neither may exist. Poisoning is inferred in AS paths (1) with loops, or (2) where clique ASes are separated. Paths with AS loops are filtered out, i.e., where an ASN appears more than once and is separated by at least one other ASN. Such paths are an indication of poisoning, where an AS X prevents a path from being selected by a non-adjacent upstream AS Y by announcing the path "X Y X" to provider Z, so if the route is subsequently received by Y it will be discarded when the BGP process examines the path for loops [130]. For BGP paths recorded in April

2012, 0.11% match this rule. After the clique has been inferred in step 3, the next step is to discard AS paths where any two ASes in the clique are separated by an AS that is not in the clique. This condition indicates poisoning, since a clique AS is by definition a transit-free network. For BGP paths recorded in April 2012, 0.03% match this rule.

Additionally, paths containing unassigned ASes are also filtered out; BGP paths from April 2012 contain 238 unassigned ASes in 0.10% of unique paths. 222 of these ASes are reserved for private use and should not be observed in paths received by route collectors. In particular AS23456, reserved to enable BGP compatibility between ASes that can process 32-bit ASNs and those that cannot [194], is prevalent in earlier public BGP data and can hinder relationship inferences.

The algorithm also removes ASes used to operate IXP route servers because the relationships are between the participants at the exchange. To identify IXP route server ASes I utilized the Euro-IX IXP directory [6], to compile a list of 56 ASes known to operate route servers; these ASes are removed from paths so that the IX participants are adjacent in the BGP path.

Finally, all paths from 167.142.3.6 for May and June 2003, and from 198.32.132.97 between March and November 2012 have been discarded; the former reported paths with ASes removed from the middle of the path, and the latter reported paths inferred from traceroute [145].

4.3.4 Inferring Clique

This section describes how to infer the ASes present at the top of the hierarchy. Since Tier-1 status is a financial circumstance, reflecting lack of settlement payments, the focus is on identifying transit-free rather than Tier-1 ASes. First, the Bron/Kerbosch algorithm [46] is applied to find the maximal clique C_1 from the AS-links involving the largest ten ASes by transit degree.[1] Second, the algorithm tests every other AS in order by transit degree to complete the clique. AS Z is added to C_1 if it has links with every other AS in C_1 and it does not appear to receive transit from another member of C_1; i.e. no AS path should have three consecutive clique ASes. Because AS path poisoning may induce three consecutive clique ASes in a false BGP path "X Y Z", the algorithm

[1] Starting with ten ASes reveals most clique ASes and is small enough to prevent the incorrect inference of a clique below the top of the hierarchy. If there are multiple cliques, the clique with the largest transit degree sum is selected.

adds AS Z to C_1 provided there are no more than five ASes downstream from "X Y Z". A nominal value of five ASes will still capture clique ASes even in paths poisoned multiple times, yet is unlikely to wrongly place a large transit customer in the clique since a clique AS is likely to announce (and it is likely to observe [167]) more than five customers of large transit customers. If an AS would be admitted to C_1 except for a single missing link, that AS is added to C_2. Finally, because an AS might be in C_1 but not in the clique, the Bron/Kerbosch algorithm is re-used to find the largest clique (by transit degree sum) from the AS-links involving ASes in C_1 and C_2. The product of this step is a clique of transit-free ASes.

Figure 4.5 shows ASes inferred to be in the clique since January 2002. Nine ASes have been in the clique nearly every month, and ASes that are inferred to be in the clique are almost continuously present. The consistency of the inferred clique and feedback from operators increase the confidence in the proposed clique inference methodology. However, peering disputes and mergers of ASes can disrupt the inference of the clique. ASes may form alliances to prevent de-peering incidents from partitioning their customers from the Internet. If such a disconnection incident triggers activation of a backup transit relationship, a peer will disappear from the clique and instead be inferred as a customer of the allied peer. The process of merging ASes can can also result in peers being inferred as customers. For example, in 2013 Level3 (AS3356) gradually shut down BGP sessions established with Global Crossing (AS3549), shifting sessions to AS3356. In order to maintain global connectivity during this merger process, Level3 advertised customers connected to AS3549 to peers that were only connected to AS3356. As a result, ASes in the clique appeared to be customers of AS3356, when in reality they were peers. Specifically, in figure 4.5, AS6461 was not inferred to be a member of the clique because it had shifted all peering ports with Level3 to AS3356.

4.3.5 Inferring Providers, Customers, and Peers

The remainder of the algorithm infers p2c and p2p relationships for all links in the graph. Step 3 infers p2p relationships for the full mesh of links between clique ASes. The rest of this section uses figure 4.6 as reference.

AS path triplets: Inferences use only AS path triplets (adjacent pairs of links). Triplets provide the constraints necessary to infer c2p relationships while allowing to

Notation	Description
X < Y	X is a customer of Y
X − Y	X is a peer of Y
X ? Y	No inferred relationship

Table 4.1: Notation used to describe relationships.

ignore non-hierarchical segments of paths, and are more computationally efficient than paths. For example, in figure 4.6 path 1 is broken into two triplets: "1239 3356 9002" and "3356 9002 6846".

Notation: The notation in table 4.1 is used to describe relationships between ASes. A p2c relationship between X and Y is presented as "X > Y". The notation reflects providers as typically greater (in degree or size or tier of a traditional hierarchical path) than their customers. A triplet with no inferred relationships is presented as "X ? Y ? Z".

Sorting of ASes: ASes are sorted in the order according to which their c2p relationships are estimated. ASes in the clique are placed at the top, followed by all other ASes sorted by transit degree, then by node degree, and finally by AS number to break ties. Transit degree is used rather than node degree to avoid mistaking high-degree nodes (but not high-transit degree, e.g., content providers) for transit providers. Figure 4.6 shows the sorted order of the ASes in that graph.

Preventing cycles of p2c links: When a c2p relationship is inferred, the algorithm records the customer AS in the provider's customer cone, and adds the AS to the cones of all upstream providers. Any ASes in the customer's cone not in the cones of its upstream providers are also added to the cones of those providers. A c2p relationship will not be inferred if the provider is already in the AS's cone, to prevent a cycle of p2c links.

Step 5: Infer c2p relationships top-down using ranking from Step 2. This step infers 90% of all the c2p relationships, and is the simplest of all the steps. ASes are traversed top-down, skipping clique ASes since they have no provider relationships. When an AS Z is visited, the algorithm infers Y > Z if it observes "X − Y ? Z" or "X > Y ? Z". To have observed "X − Y", X and Y must be members of the clique (step 3). To have inferred "X > Y" by now, one must have visited Y in a previous iteration of this step. No cycle of c2p links can be formed because c2p relationships are assigned

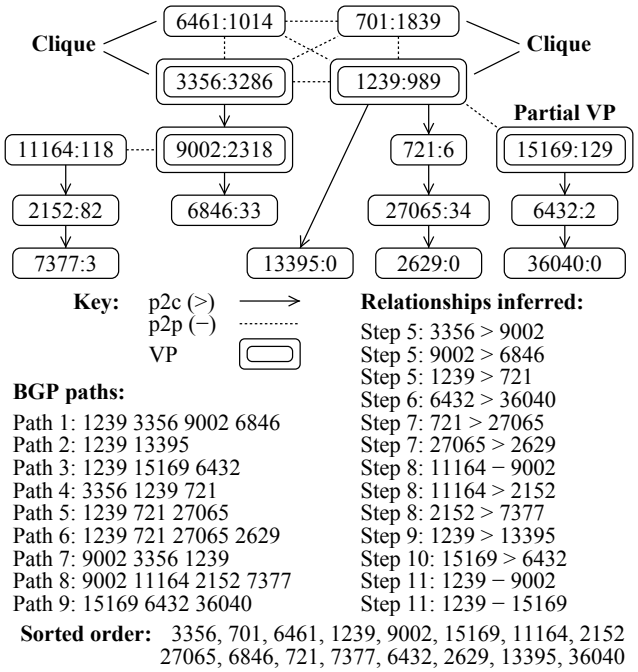


Figure 4.6: Inferring providers, customers, and peers. Each AS is labelled "AS Number:Transit Degree". VPs are in double squares, and (by definition) on the left side of all raw BGP paths. This set of paths to illustrates the inferences made at each step of the algorithm. Relationships listed use notation in Table 4.1.

along the *degree gradient*, i.e. no c2p relationship is inferred between two ASes where the provider has a smaller transit degree, a necessary condition to create a cycle. For example, in figure 4.6, the algorithm first considers (after the four clique ASes) c2p relationships for 9002 (3356), then 15169 (none), etc.

The order of the ASes in the triplet is important, at this step and for most of the remaining steps. To minimize false c2p inferences due to misconfigurations in one direction of a p2p relationship (an AS leaks provider or peer routes to peers), a c2p relationship is inferred when the provider or peer is closer than the customer to at least one VP in at least one triplet. This heuristic builds on the intuition that an AS enters a provider relationship to become globally reachable, i.e., at least one VP should observe the provider announcing the customer's routes. For example, when it is inferred that $3356 > 9002$ in figure 4.6, the algorithm uses the triplet "1239 3356 9002" from path 1 and not triplet "9002 3356 1239" from path 7 because 3356 appears before 9002 in path 1.

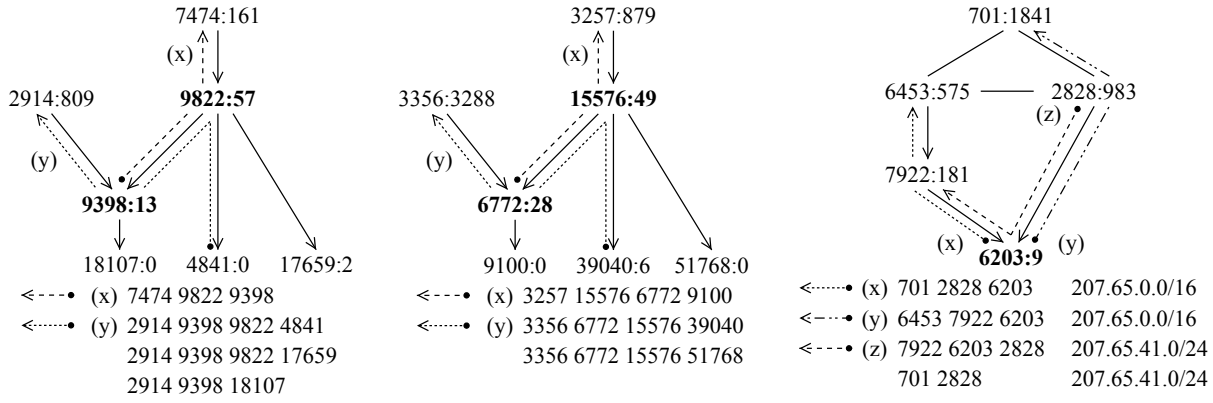
Step 6: Infer c2p relationships from VPs inferred to be announcing no provider routes. It is assumed that "partial VPs" providing routes to fewer than 2.5%

of all ASes either (1) export only customer routes, i.e., the VP has configured the session with the collector as p2p; or (2) have configured the session as p2c, but have a default route to their provider, and export customer and peer routes to the collector. Given a path "X ? Y ? Z" where X is a partial VP and Z is a stub, the link XY can either be p2c or p2p, requiring Y > Z. In figure 4.6, path 9 is used and the inference of 15169 as a partial VP to infer 6432 > 36040.

Step 7: Infer c2p relationships for ASes where the customer has a larger transit degree due to path poisoning. Given a triplet "W > X ? Y" where (1) Y has a larger transit degree than X, and (2) at least one path ended with "W X Y" (i.e. Y originates a prefix to X), then the algorithm assigns X > Y. It is assumed that c2p relationships against the degree gradient are rare, although they can arise from path poisoning. Condition (2) mitigates the risk of using poisoned paths because poisoned segments of a path do not announce address space. In figure 4.6 the algorithm infers 721 (X) > 27065 (Y) because path 5 shows 27065 announcing a prefix to 721. In the absence of path 5, it would be inferred that 2629 poisoned path 6 with 27065 and the algorithm would not assign a c2p relationship. When the relationship X > Y is assigned, Y > Z is also assigned where triplets "X > Y ? Z" are observed; in figure 4.6 the algorithm uses path 6 to infer 27065 > 2629.

Step 8: Infer customers for provider-less ASes. Provider-less ASes are visited top-down, skipping clique members because their customers were inferred in step 5. This step is necessary because steps 5 and 7 require an AS to have a provider in order to infer customers. Examples of provider-less ASes are some regional and research networks, e.g., TransitRail. For each provider-less AS X, the algorithm visits each of its neighbours W top-down. When a triplet "W X Y" is observed, the algorithm infers W − X because W was never observed announcing X to providers or peers in previous steps; therefore, X > Y. The condition in step 5 that the peer AS must be closest to the VP is removed, because provider-less ASes are mostly observed by downstream customers providing a public BGP view. In figure 4.6, 11164 (X) is a provider-less AS; path 8 is used to infer 9002 (W) − 11164 (X) and 11164 (X) > 2152 (Y). When the algorithm assigns X > Y, it also assigns Y > Z where triplets "X > Y ? Z" are observed; in figure 4.6 path 8 is used to infer 2152 (Y) > 7377 (Z).

Step 9: Infer that stub ASes are customers of clique ASes. If there is a link



(a) Siblings (9398 and 9822) (b) Mutual Transit (6772 and 15576) (c) Leaking or Poisoning (6203)

Figure 4.7: p2c-c2p valleys caused by unconventional routing policies between (a) siblings, (b) mutual transit, and (c) leaking/poisoning. Each AS is labelled with its transit degree, which influences the order of p2c inferences. An AS-to-organization map would resolve (a) but not (b) because the ASes in (b) are independent ASes. Leaking as in (c) results in paths with spurious p2c-c2p valleys when an AS leaks a route from a provider to another provider. Note that in (c) AS6203 could instead be poisoning to prevent AS2828 from selecting the more specific route (traffic engineering); (c) is an example of a prefix-list leak because AS2828 is the only origin AS observed for that prefix. All examples are present in April 2012 BGP data.

between a stub and a clique AS, it is classified as c2p. This step is necessary because step 5 requires a route between a stub and a clique AS to be observed by another clique AS before the stub is inferred to be a customer. Stub networks are extremely unlikely to meet the peering requirements of clique members, and are most likely customers. In figure 4.6, path 2 reveals a link between 1239 and 13395, but there is no triplet with that link in the set, perhaps because it is a backup transit relationship.

Step 10: Resolve triplets with adjacent unclassified links. The algorithm traverses again ASes top down to try to resolve one link as p2c from triplets with adjacent unclassified links. This step is necessary to avoid inferring adjacent p2p links in step 11, since adjacent p2p links imply anomalous behaviour, e.g., free transit or route leakage. The requirement in step 5 that the first half of the triplet must be resolved is loosened. When visiting Y, the algorithm searches for *unresolved triplets* of the form "X ? Y ? Z", and attempts to infer Y $>$ Z. For each unresolved triplet "X ? Y ? Z", the algorithm looks for another triplet "X ? Y $<$ P" for some other P. If one is found, the algorithm infers X $<$ Y (and Y will be inferred as a peer of Z in step 11). Otherwise it searches for a triplet "Q ? Y ? X", which implies Y $>$ X, and therefore both sides of the original unresolved triplet to X $<$ Y and Y $>$ Z would be resolved. Since there is

only confidence of resolving one side of the original triplet (embedding an assumption that most p2c links have already been resolved at earlier steps), no inferences are made in this case. Otherwise, the algorithm infers $Y > Z$ in this step, and $X - Y$ in step 11.

Step 11: Infer p2p links: A p2p relationships is assigned for all the links that have no inferred relationships.

4.3.6 Limitations of the Algorithm

Sibling Relationships and Mutual Transit: The above algorithm does not infer sibling relationships, where the same organization owns multiple ASes, which can therefore have unconventional export policies involving each sibling's peers, providers, and customers. Similarly, the algorithm does not infer mutual transit relationships, where two independent organizations provide transit for each other in a reciprocal arrangement. Both of these arrangements can lead to paths (and triplets) that violate the valley-free property, and in particular produce p2c-c2p valleys in paths. Gao's algorithm [97] inferred that two ASes involved in a non-hierarchical path segment were siblings, which maximizes the number of valley-free paths. Dimitropoulos *et al.* used WHOIS database dumps to infer siblings from ASes with similar organization names, because policy diversity among siblings makes it difficult to infer siblings from BGP data [76]. The present algorithm does not attempt to resolve these unconventional routing policies because it is difficult to accurately classify them; as a result, it produces p2c-c2p valleys in paths.

Figure 4.7 provides three examples of non-hierarchical path segments caused by siblings (figure 4.7(a)), mutual transit (figure 4.7(b)), and route leaks or path poisoning (figure 4.7(c)). In figure 4.7(a), ASes 9398 and 9822 are ASes owned by the same organization, Amcom Telecommunications, which implements complex export policies with these ASes. Specifically, customers of 9822 are exported to 9398's peers and providers, and routes originated by 9398 are exported to 9822's providers. These policies induce a p2c-c2p valley in a path, because 9398 is inferred as a customer of both 9822 (path x) and 2914, and it appears that 9398 announces customers of its inferred provider 9822 to its other inferred provider (path y). In figure 4.7(b), independently operated ASes 6772 and 15576 implement mutual transit, owing to complementary traffic profiles: AS6772 is an access provider with mostly inbound traffic, while AS15576 is a

content provider, with mostly outbound traffic. A WHOIS-derived database of sibling relationships does not help infer mutual transit arrangements. Finally, route leaks and path poisoning can also result in a p2c-c2p valley in a path. Figure 4.7(c) provides an example of a route leak from 6203, where it provides transit to a /24 prefix announced by one of its providers (AS2828) to another provider (AS7922). The prefix AS2828 announces is a more specific prefix of a /16 prefix that AS6203 announces, and has likely leaked because of a prefix list configured in AS6203. An alternative explanation is that AS6203 is poisoning that path so that AS2828 cannot select the more specific prefix. A route leak is the more plausible explanation because AS2828 is the only AS to originate the prefix. Without the use of prefixes, it is not possible to distinguish this valley from mutual transit or siblings. Detecting and validating route leaks and path poisoning remains an open problem.

To evaluate how the sibling links in were classified, I used sibling inferences derived from WHOIS database dumps [134]

These sibling inferences are not used in the inference algorithm because (1) there are no available supporting WHOIS databases going back to 1998, (2) in validation the sibling inferences contained a significant number of false-positives (where two ASes were falsely inferred to belong to an organization), and (3) they do not help distinguish mutual transit between independent ASes from other anomalies such as path poisoning and route leaks. In total, there were 4537 links observed between inferred siblings for April 2012; 4238 (93%) of these were inferred to be p2c. Because most of the siblings are inferred to have p2c relationships (i.e. a transit hierarchy) the sibling inferences were also used to examine if the ordering of ASes in paths supported the classification. Of the 312 organizations for which at least two siblings were observed in a path, 275 (88%) had a strict ordering; i.e. AS x was always observed in a path before sibling AS y. For example, there can be observed 21 Comcast sibling ASes in BGP paths with at least two siblings present; all of these ASes were connected beneath AS7922. It is possible that Comcast's siblings exported routes received from their peers (if any) to other siblings. However, no peering links can be seen from beneath AS7922 (perhaps due to limited visibility), and it would make more engineering sense to share peer routes among siblings by connecting the peer to AS7922.

Partial Transit, Hybrid Relationships, and Traffic Engineering: The algo-

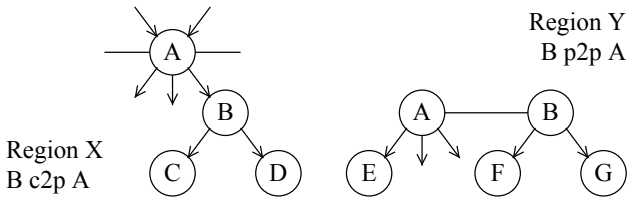


Figure 4.8: Two ASes may have hybrid relationships. In this example B is a customer of A in region X and B is a peer of A in region Y. However, the algorithm infers a single relationship between B and A: c2p. If A routes rationally, it will only advertise paths to F and G from B to customers.

rithm infers either a p2p or c2p relationship for all links. A partial transit relationship typically restricts the propagation (and visibility) of routes beyond a provider to the provider's peers and customers. Similarly, complex import and export policies can produce hybrid relationships [160]. Figure 4.8 depicts an AS pair with a c2p relationship in one region, and a p2p relationship elsewhere. Similarly, ASes A and B may enter into a p2p relationship, but B may not advertise all customers or prefixes to A, requiring A to reach those via a provider.

The presence of hybrid relationships can cause the algorithm to incorrectly infer a c2p link as a p2p link. In figure 4.8, if there cannot be observed providers of A announcing routes to E via A then the algorithm infers the c2p link between A and E as p2p. This inference occurs because step 10 collapses triplets with no relationships inferred for either link; i.e., it does not adjust the triplet "E ? A $>$ B". A separate step is necessary to distinguish partial transit relationships visible from a VP below the provider from peering relationships.

Paid Peering: An assumption behind c2p and p2p relationships is that the customer pays the provider and p2p relationships are settlement-free, as historically p2p relationships were viewed as mutually beneficial. Modern business relationships in the Internet are more complicated; ASes may enter into a paid-peering arrangement where an AS pays settlements for access to customer routes only. Multiple network operators confirmed several aspects of paid-peering: (1) approximately half of the ASes in the clique (figure 4.5) were paid-peers of at least one other AS in the clique as of June 2012; (2) paid-peering occurs between ASes at lower-levels of the AS topology; and (3) routes from paying and settlement-free peers have the same route preference. This last condition prevents the algorithm from distinguishing paid-peering from settlement-free peering using BGP data alone.

Step	Description	Validation (PPV)	Fraction
3	clique at top of AS topology	136 p2p @ 100%	153 (0.12%)
5	**c2p relationships top-down**	26664 c2p @ 99.8%	**71160 (56.4%)**
6	c2p relationships from VPs announcing no provider routes	116 c2p @ 99.1%	532 (0.42%)
7	c2p relationships to smaller degree providers	205 c2p @ 96.1%	2420 (1.92%)
8	relationships for ASes with no providers	120 c2p @ 93.3%	842 (0.67%)
		152 p2p @ 96.7%	333 (0.26%)
9	c2p relationships for stub-clique	422 c2p @ 95.0%	651 (0.52%)
10	collapse adjacent links with no relationships	524 c2p @ 94.7%	2474 (1.96%)
11	**p2p relationships for all other links**	15274 p2p @ 98.7%	**47517 (37.7%)**
		43613 @ 99.3%	126082 (100%)

Table 4.2: Validation of inferences (PPV) and number/fraction of inferences made at each step.

Algorithm	c2p			p2p		
	PPV (%)	TPR (%)	Errs (1/)	PPV (%)	TPR (%)	Errs (1/)
CAIDA	99.6	99.3	250	98.7	99.3	77
UCLA	99.0	94.7	100	91.7	98.8	12
Xia+Gao	91.3	98.6	11	96.6	81.1	29
Isolario	90.3	98.0	10	96.0	82.4	25
Gao	82.9	99.8	5.8	99.5	62.5	200

Table 4.3: The proposed AS relationship algorithm accurately classifies both c2p and p2p relationships, with high precision (PPV) and recall (TPR).

Backup Transit: A backup transit relationship occurs when a customer's export policies prevent their routes from being exported outside a provider's customer networks. The export policies used while the provider is in backup configuration are identical to peering; the difference between backup transit and paid peering is due to export filters instead of a contractual agreement. The present algorithm infers most backup transit relationships as peering.

4.3.7 Validation

This section presents an evaluation of the positive predictive value (PPV) and true positive rate (TPR, or *recall*) of the algorithm heuristics against the collected validation dataset (section 4.2.6). The AS relationships dataset consists of 126,082 links of which 43,613 (34.6%) were validated. Table 4.2 shows the PPV of inferences made at each step of the algorithm. Most relationship inferences are made in steps 5 (56.4%) and 11 (37.7%), and both of these stages have excellent PPV (99.8% and 98.7% respectively).

Table 4.3 compares the PPV of the inferences and those made by four other popular inference algorithms for April 2012 BGP paths. The algorithm correctly infers 99.6% of c2p relationships and 98.7% of p2p relationships. Unfortunately the authors

of SARK [191], CSP [160], and ND-ToR [197] did not respond to requests for the source code for their algorithm, or for a set of relationships inferred from April 2012 BGP paths. Gao's PPV for p2p relationships is the highest of the algorithms tested because it makes the fewest number of p2p inferences of all the algorithms, inferring many more c2p relationships than exist in the graph. The algorithm that performs closest is UCLA's; however, the improvements of the present algorithm result in six times fewer false peering inferences. I assembled additional historical validation datasets by extracting relationships from archives of the RIPE WHOIS database (RPSL, section 4.2.4) and public BGP repositories (BGP communities, section 4.2.5) at six month intervals between February 2006 and April 2012. The validation performance of Gao, UCLA, and CAIDA algorithms are quantitatively similar as shown in table 4.3.

I investigate the types of errors that these four algorithms produce, focusing on the two cases with significant occurrence: where the algorithm correctly infers c2p (p2p), but another algorithm mistakenly infers p2p (c2p). It should be noted that when the ground truth is p2p, Gao often infers the link as c2p, usually with the customer having a smaller degree than the provider. On the other hand, UCLA and Isolario produce errors where a p2p link is inferred to be c2p, often with the customer having a larger degree than the provider. The UCLA algorithm often infers c2p links to be p2p because it uses the visibility of a link from tier-1 VPs to draw inferences, and defaults to a p2p inference for links it cannot see (see section 3.5.1). Although a variant of this visibility heuristic used in the present algorithm, additional heuristics are also applied to accommodate for phenomena that inhibit visibility through tier-1 VPs, e.g., traffic engineering, selective announcements.

I compared the inferences with 82 partial transit relationships that were flagged by a community string. The algorithm correctly inferred 69 (84%) of them as p2c; 66 p2c inferences where made in step 10. In comparison, UCLA's dataset identified only 13 (16%) of the partial transit relationships as p2c. I also compared the inferences against a small set of 27 backup p2c relationships, of which only 2 were correctly identified as p2c. Validation data for partial and backup transit relationships is scarce because of their rarity and their limited visibility.

It is well-known that the public view misses a large number of peering links [28, 105]. While inferences can be made only for visible links, an important question is

whether the accuracy is affected by a lack of (or increasing) visibility. To understand this question, the following experiment was performed 10 times. Paths are selected from a random set of 25% of VPs, and successively more VPs are added to obtain topologies seen from 50%, 75% and all VPs. Then, the PPV of the inferences is calculated for each topology subset. The results show that the PPV of c2p inferences was consistently between 99.4% and 99.7% on all topology subsets. The PPV of p2p links varied between 94.6% and 97.7% with 25% of VPs, and 97.2% and 98.4% with 50% of VPs, indicating that the algorithm performs better when it has more data (VPs) available. Consequently, if the visibility of the AS topology increases in the future (e.g., due to new VPs at IXPs), the accuracy of the algorithm at inferring the newly visible links should not be affected.

4.4 Applications of AS Relationships

4.4.1 Assessing the market power of ASes

The customer cone is defined as the ASes that a given AS can reach using a customer (p2c) link, as well as customers of those customers (indirect customers) [76]. An AS is likely to select a path advertised by a customer (if available) over paths advertised by peers and providers because the AS is paid for forwarding the traffic. The most profitable traffic for an AS is traffic forwarded between customers, as the AS is paid by both.

The customer cone is a metric of influence, but not necessarily of market power. Market power requires the ability to restrict the mobility of customers; in general, an AS can enter into a provider relationship with whoever offers a suitable service. For large transit providers, particularly those in the clique where a full p2p mesh is required for global connectivity, the customer cone defines the set of ASes whose service might be disrupted if the AS were to have operational difficulty.

Due to ambiguities inherent in BGP data analysis, there are multiple methods to infer the customer cone of a given AS. This section will use the Provider/Peer Observed (PPO) customer cone which was found not to over-estimate the size of the customer cone and to be the least sensitive to complex agreements [146]. The PPO customer cone of an AS A is computed using routes observed from providers and peers of A, and not recursively as first defined in [76].

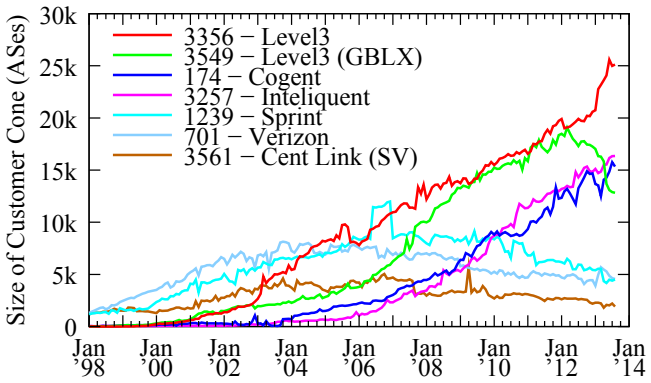


Figure 4.9: The size of the customer cones for the seven ASes that were among the three largest ASes between Jan. 1998 and Aug. 2013. The three largest ASes in Jan. 1998 (701, 1239, and 3561) are no longer in the top three.

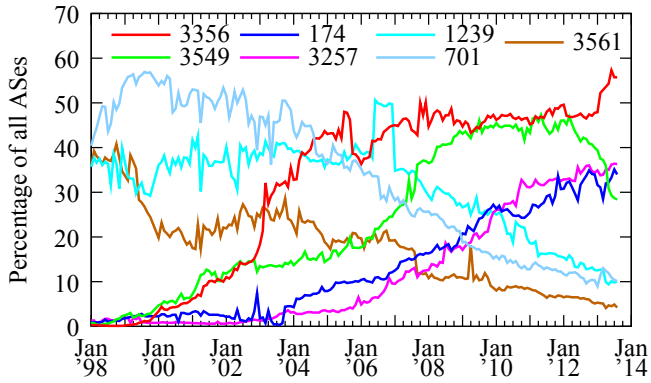


Figure 4.10: Relative size of provider/peer observed cone over time. 701 acquired part of 3561 in 1999 and moved customers across.

Figure 4.9 plots the seven ASes that ranked in the top three ASes by provider/peer observed customer cone size at any point from January 1998. Several interesting trends can be observed with just these seven ASes. First, the three ASes ranked in the top three for January 1998 (ASes 701, 1239, and 3561) are no longer in the top three. In absolute terms, the customer cone of 701 decreased in size between January 2002 and October 2012. The customer cone of 3356 reflects two other interesting events: (1) in early 2003, AS1 (Genuity/BBN) merged with 3356 to create the third largest network at the time, and (2) in late 2010, 3549 (the second largest AS by customer cone) was purchased by Level3 (the largest AS by customer cone). 3549's customer cone has since shrunk as new customers connect to 3356 and some of 3549's customers moved across.

Figure 4.10 plots the customer cone sizes for the same seven ASes, but as a fraction of the topology size: (1) ASes 701, 1239, and 3561 all had the same customer cone size

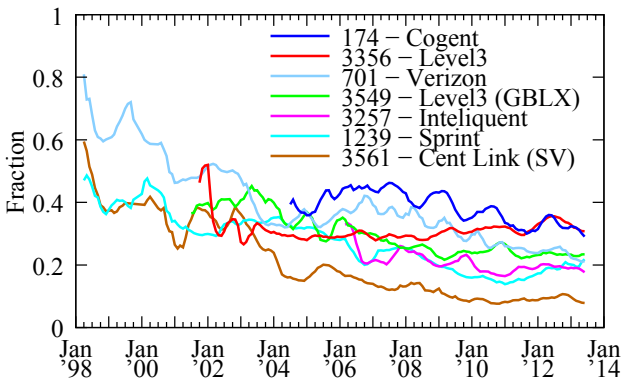


Figure 4.11: Fraction of ASes in X's customer cone that were reached via X from an AS in X's customer cone over time. Most ASes show a decline in the fraction of cone-internal paths.

in January 1998, (2) some customers of 3561 (MCI) shifted into 701 (Worldcom) due to the MCI-Worldcom merger in 1998, (3) 1239 held a third of the ASes in its customer cone for ten years until 2008, and (4) while 3356 had the largest customer cone in 2012, its relative cone size, i.e., as a fraction of the entire AS topology, was slightly smaller than AS701's was in January 2000. This last fact reflects massive growth in the Internet's AS topology since 2000, in addition to the consolidation undertaken by both ASes, yielding the largest customer cones of the two respective decades.

Since most companies providing Internet transit are by now also in other lines of business and do not report financial information specific to their transit business, it is not possible to correlate BGP consolidation with financial performance. But of the three ASes whose relative customer cone sizes have plummeted in the last decade (701, 1239, 3561), two of them (Verizon and Sprint) have moved into more profitable cellular service.

4.4.2 Topology Flattening

The introduction of CDNs and richer peering has resulted in a flattening of the Internet topology [99, 140] where ASes avoid sending traffic via transit providers. An intriguing question is how valid is the customer cone in a flattened Internet topology? How many paths still travel to the top of a given cone to reach destinations?

While public BGP data contains a small fraction of all peering links [28, 105] it is possible study shifts in routing behaviour from the paths of individual VPs because they reveal the peering links they use. For each VP that provides a full view to RV or RIS and is also in X's customer cone, the fraction of *cone-internal* paths, i.e., fraction

of paths from that VP that transit X (the cone's top provider) when reaching another AS in X's cone which is not also in the customer cone of the VP can provide a metric of measuring the flattening. Figure 4.11 shows the five-month moving average of this fraction for the seven ASes once they have at least 1000 ASes in their cone. Five of these networks show a steady decline in the fraction of cone-internal paths.

The AS relationship inferences also shed light on the continually increasing richness of peering in the Internet. As the number of full VPs has increased an order of magnitude since 2000 (from 12 to 125 in October 2012), the number of p2p links observable from these VPs jumped by two orders of magnitude (from about 1K to 52K), fraction of the entire graph from 10% (in 2000) to 38%. (even after the number of full VPs stabilized in 2008). This increase in peering (flattening) was not observed by individual VPs, most (75%) of which experienced almost no change in the *fraction* of links inferred as p2p. Instead, the increase in relative presence of p2p links in the graph is due to individual VPs seeing more unique p2p links.

4.5 Summary

This section presented, and validated to an unprecedented level, a new algorithm for inferring AS relationships using publicly available BGP data. The algorithm tolerates prevalent phenomena that previous algorithms did not handle. I validated 34.6% of the relationship inferences, finding the c2p and p2p inferences to be 99.6% and 98.7% accurate, respectively. Since even different sources of the validation data disagree by 1%, the algorithm reaches the limit of accuracy achievable with available data. The validation data set (excluding the feedback from operators) and the inferred relationships are publicly available at `http://www.caida.org/publications/papers/2013/asrank/`

Analysis of the Internet at the AS granularity is inherently challenged by measurement and inference in a dynamic complex network. A known concern is that public views of the AS topology capture only a fraction of the p2p ecosystem, since so few ASes share their full view of the Internet with BGP data repositories. Another challenge is the variety of complex peering relationships that exist. The next chapter explains how to infer such relationships by adapting this algorithm and by incorporating additional data from the control and data plane.

The accurate inferences shed new light on the flattening Internet topology, revealing a decline in the fraction of observed paths traversing top-level (clique) ASes from 2002 (over 80%) bottoming out in 2006 (just above 60%), followed by a slow rise back to 77% today, perhaps as these clique ASes adjust their peering strategies to try to recover some transit revenue.

Chapter 5

Inference of Complex Relationships

The traditional approach of modelling business relationships between ASes abstracts relationship types into three broad categories: transit, peering, and sibling. More complicated configurations exist, and understanding them would advance our knowledge of Internet economics as well as routing, but inferring them and validating the inferences with ground truth data is challenging. This chapter extends the conventional AS relationship inference algorithm to infer relationships that are complex in two dimensions: prefix granularity (relationships that differ across prefixes), and geography (relationships that differ across regions). Using this new algorithm, I find that 4.5% of the 90,272 provider-customer relationships I inferred for March 2014 were complex, including 1,071 hybrid AS relationships and 2,955 partial-transit relationships. I used two types of data to validate my inferences with 94.9% accuracy: feedback from operators and BGP communities. Using data from BGP and traceroute, I found that only 25% of the inferred hybrid relationships are established between different continents among large transit providers, while the majority of hybrid and partial transit relationships involve medium-sized European ISPs.

5.1 Introduction

The abstraction of AS relationships to three classes (and sometimes only the first two are used) provider-customer (p2c), peering (p2p) and sibling (s2), facilitates the development of inference heuristics at the cost of having a coarse-grained classification that may not accurately represent the actual relationship agreements.

This oversimplification ignores more complex relationships and may introduce artifacts into the study of inter-domain routing [179], such as spurious relationship

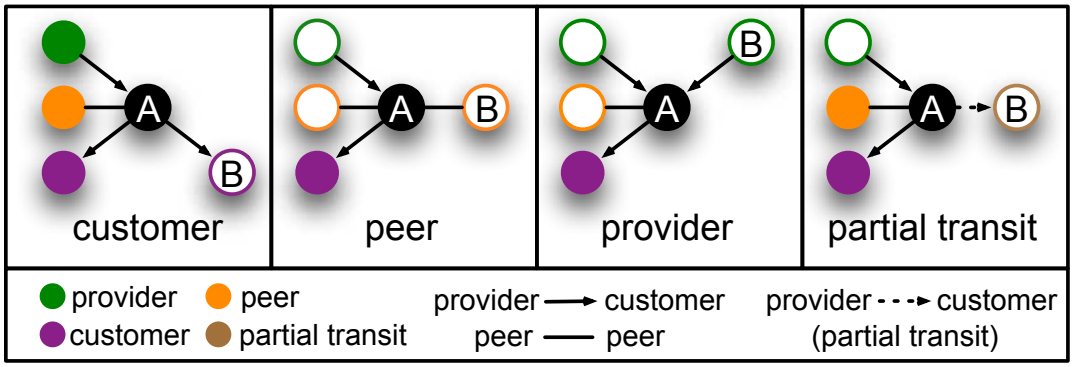


Figure 5.1: These graphs show how A's export policy to B changes based on A relationship to B, with the filled nodes exported to B.

cycles [78], artificial policy violations [173, 155], and generally inaccurate AS path inference [153, 70]. These problems have inspired the development of a relationship-agnostic AS topology model as an alternative [159, 160].

Specifically, AS relationships can vary in three dimensions: over time, over geographic region or by individual prefix. Inferring complex relationships requires inferring relationships at finer granularities, which imposes significant measurement and computational requirements, as well as the challenge of distinguishing the observable effects of complex relationships from the observable effects of traffic engineering policies [146].

Highlighting the challenge, a recent attempt to infer complex relationships by analysing policy anomalies at Points-of-Presence (PoPs) revealed only a single hybrid relationship [163]. As a result complex AS relationships remain obscure and the ability to infer them has been recently described as the *holy-grail* of relationship inference [163]. In a similar vein, Dimitropoulos et al. characterized the inference of complex relationships as a *formidable* task and hypothesized that that it requires direct access to the BGP configuration of border routers [76]

In this section I present a new algorithm to infer the two most common types of complex AS relationships: *hybrid peering* (or *dual peering/transit* [82]) relationships, which differ by interconnection location; and *partial transit* (or *regional* [83]) relationships, which restrict the scope of a provider-customer relationship to the provider's peers and customers (but not providers) [82, 83, 76, 88, 146].

These complex relationships can be defined as special cases of the traditional transit and peering types, which allows us to leverage the relationship inference algorithm

presented in section 4 instead of designing an entirely new algorithm. To achieve a more fine-grained relationship inference the algorithm combines passive BGP measurements, active measurements, and geolocation data. For March 2014 I inferred 1K hybrid relationships and 3K partial transit and regional transit relationships, used by not only tier-1 and tier-2 ASes, but also middle-sized multi-national European ISPs. The data also reveals how IXPs are driving the evolution of complex relationships as they alter the traditional symmetry of traffic flow between peering ASes. I use feedback from operators and BGP community information to validate to 94.9% accuracy the inferred complex relationships.

5.2 Background

As we have seen in the previous sections, to enable the development of inference heuristics an abstraction of AS relationships is required. Gao's seminal work [97] proposed a classification into three abstract relationship types: *provider-to-customer (p2c)* or *transit* relationship, *peer-to-peer (p2p)* relationship, and *sibling-to-sibling (s2s)* relationship.

Although the above classification captures the majority of AS interconnections, more complicated relationships can also exist. Norton lists two additional types of interconnections, *dual peering/transit* and *partial transit*, which happen predominantly in Europe [82, 83]. *Hybrid* relationships arise when two ASes agree to different relationship types at different inter-connection points. In the partial transit scenario (also defined as *regional*) a provider sells, for a discounted price, transit access to its customers and peers but not to its providers. Faratin et al. suggested that the AS relationships become increasingly complex because the growth of Content Distribution Networks (CDNs) and eyeball networks change the perceptions of symmetry in traffic delivery costs [88]. A survey by Dimitropoulos et al. [76], among AS administrators on the types of complex configurations, confirmed the existence of relationships that vary across different peering points and different prefixes.

Relationship abstractions coarser than the hybrid granularity may lead to artifacts in the study of inter-domain routing, such as spurious relationship cycles [78] or artificial policy violations [173, 155]. Mao et al. concluded that the existence of complex relationship types may be partly responsible for the inability to accurately infer AS-

level paths [153]. Similar conclusion has also been reached by the authors of [70]. Mühlbauer et al. proposed a relationship-agnostic AS topology model to capture the complex BGP policies that cannot be modelled based on the simple relationship abstraction [159, 160].

Unfortunately data on hybrid and partial transit relationships are very limited. Fine-grained relationship inference across two dimensions, prefixes and geography, imposes significant measurement and computational requirements that inhibited the development of appropriate algorithms [76]. On the other hand, inference at coarser granularity does not allow to distinguish complex relationships from traffic engineering policies that produce similar routing behavior [146].

Neudorfer et al. proposed a method to analyse the policy violations at different Points-of-Presence (PoPs) to enable the inference of complex relationship types [163]. However, they were only able to manually identify a single hybrid relationship which highlights the difficulties involved in the inference of complex relationships.

5.3 Data Sources

I combine three different data sources, to support per-prefix routing inferences at the Point of Presence (PoP) level of the AS connectivity[1]: BGP data, active measurement, and geolocation data

5.3.1 BGP measurements

I obtained per-prefix routing information from BGP table snapshots from RouteViews (RV) [21] and RIPE RIS [16], which collect BGP routing information by peering with globally distributed ASes (vantage points). Each BGP table snapshot provides an AS-level view of the Internet as seen by the corresponding vantage point (VP). For each RV and RIS collector I downloaded one RIB every five days for the first three months (January-March) of 2014. I extracted three attributes from each BGP entry, the AS_Path, the Network_Prefix and the Communities string. I discard AS paths that exhibit symptoms of misconfigurations or poisoning [146], e.g., loops, unassigned ASes, and non-adjacent Tier-1 ASes. I also discard prefixes longer than /24, which are commonly blackholed.

[1] A PoP is a physical location where ASes interconnect.

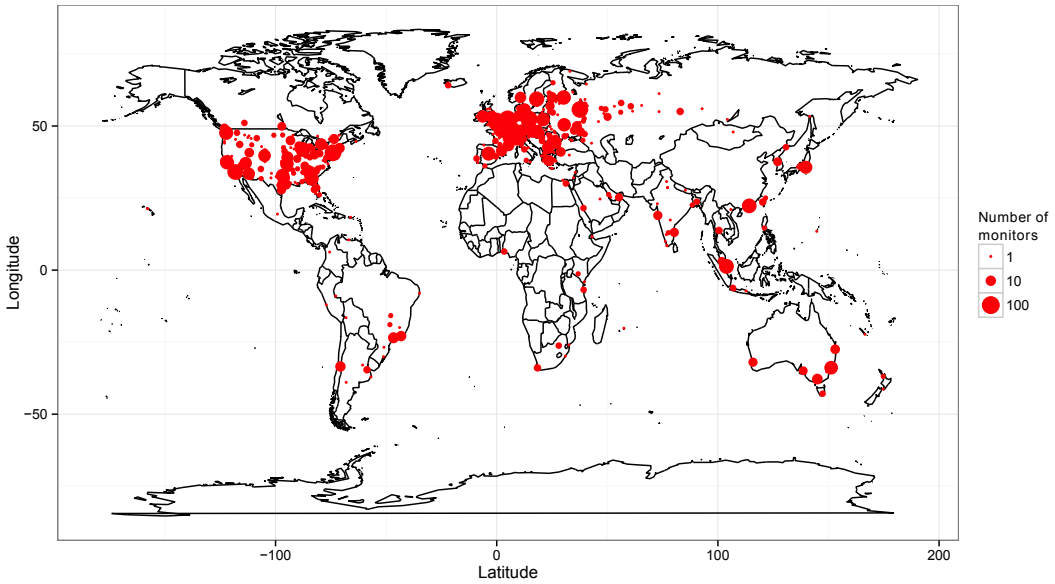


Figure 5.2: Geographical distribution of Ark nodes and traceroute servers used.

I use path and prefix attributes to identify candidate hybrid links that guide traceroute probing (Section 5.3.2). I use the Communities attribute to obtain additional information on collected paths. In total I used interpreted 6,001 Community values from 312 different ASes, that encode three types of information: geographic location of interconnection point (1,533 values from 117 ASes), route redistribution policies (2,966 values from 40 ASes), and type of relationship (1502 values from 281 ASes).

5.3.2 Active Measurements

To derive the geographical footprint of AS links, I aggregate interface-level paths derived from traceroute measurements. These paths are then processed to infer the PoP-level AS connectivity. I use two sources of active measurement data: CAIDA's IPv4 Routed /24 Topology Dataset March, 2014 [131] gathered by the Archipelago (Ark) [4] measurement infrastructure, and coordinated probing from thousands of public traceroute servers.

Archipelago supports continual Paris traceroute measurements from 94 monitors in 84 different ASes distributed in 39 different countries, probing to one random IP address in each /24 of the entire routed IPv4 space approximately every three days. To expand coverage beyond what these existing measurements capture, I develop an overlay interface to interact with 2,509 public traceroute hosts among 507 different ASes in 77 different countries. To avoid being blocked by traceroute servers for too frequent

path	**p**refix	**c**ommunity
4589 4436 20940 16625	23.211.232.0/23	4436:41718
4436 20940	72.246.99.0/24	4436:22220
4436 20940	72.246.99.0/24	4436:22122

Table 5.1: Paths tagged with geographic communities.

probing, I limit probing through this interface to one query every 10 seconds per host. Thus these measurements are carefully coordinated (using BGP data to inform target selection) to capture links and prefixes most likely to reveal complex relationships. Figure 5.2 shows the geographical distribution of traceroute hosts I use for my active measurements.

5.3.3 Geolocation data

To differentiate paths by geographic location, I use three sources of data, in order of preference: published BGP community information, strings founds in DNS hostnames, and a commercial database (NetAcuity).

Some ASes use geographic communities to indicate the ingress location of routes entering their network. Table 5.1 depicts two paths tagged with geographic communities. The first 16 bits of a 32-bit community value encode the AS that sets the community value; the second 16 bits encode a geographic location. For example, AS 4436 labels routes received in Los Angeles, Amsterdam, and London with community values 4436:41718, 4436:22220, and 4436:22122, respectively. This convention allows us to map community values to links in the AS paths and infer the entry point for the corresponding prefixes.

I use two data sources to map IP addresses to geographic locations: DRoP and Netacuity. DRoP is a constraint-based geolocation system developed by CAIDA [127] that depends on the IP address having a hostname containing a known geographically meaningful string such as an airport code (e.g., **l**ax in dc-lax-peer1–**l**ax-core1-ge.cenic.net) but then uses active measurements to validate whether the location inferred from the DNS string is consistent with RTT and TTL measurements taken from 66 monitors. If DRoP reveals no clear location, I use Netacuity [13], a commercial geolocation provider, to map the IP address to a geographic location. NetAcuity provides more coverage than the other methods, but it is optimized to accurately infer edge hosts (servers and end users) rather than router infrastructure IP addresses.

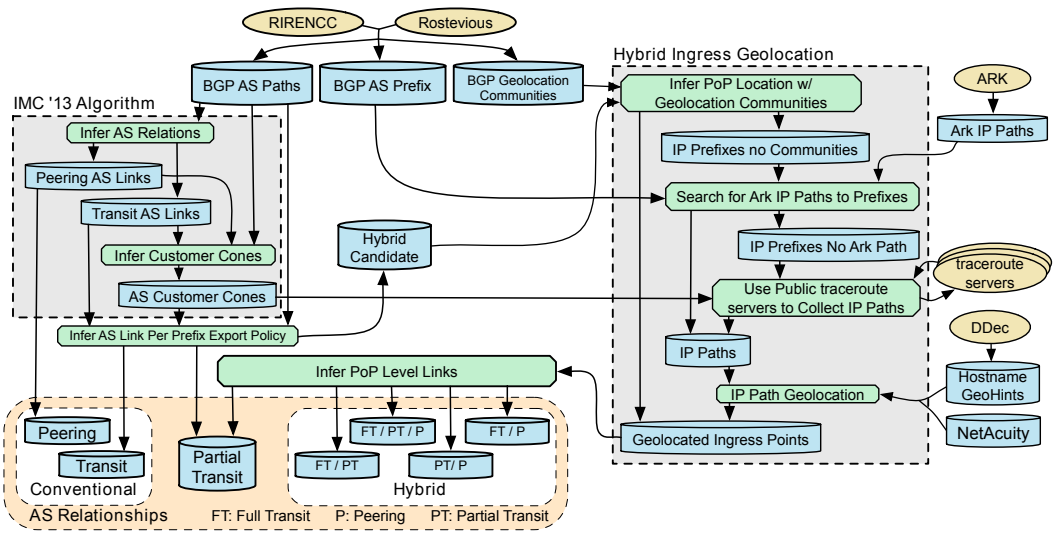


Figure 5.3: The process for inferring the hybrid and partial AS relationships.

5.4 Inference Methodology

I implement the method for inferring hybrid and partial transit relationships as an extension to existing relationship inference algorithm [146]. I first derive a base set of simple relationships between AS pairs that we will use in later steps. Figure 5.3 shows the entire inference process which this section explains in detail.

5.4.1 Classify AS link per Prefix Export Policies

In this section I will classify each p2c link as either: a conventional *provider-customer* , *partial transit*, or candidate *hybrid* relationship. First I infer the export policy for each p2c link by examining each prefix and the set of AS paths leading to them every five days throughout March 2014. I focus on p2c links because the baseline algorithm [146] always infers the transit part of a hybrid relationship.

For each prefix, I annotate each of its AS path links with its simple relationship inferred by the baseline algorithm. I discard AS paths that break the valley-free policy [146] and split the remaining paths into AS triplets. I iterate through the list triplets and tag the cp2 links according to the following criteria. If both links on the triplet are p2c, I tag the prefix as *full transit* for the link with its provider AS in the middle of the triplet regardless of its current tag, A→**B**→**C** or **A**←**B**←C. If the triplet contains a p2c link and a peer link, and the prefix has not yet been taged as *full transit* for the p2c link, A–**B**→**C** or **A**←**B**–C, I tag the prefix as *partial transit* for the p2c link. If a prefix has never been tagged for any p2c link in the triplet, I tag the prefix as *peering* for that p2c

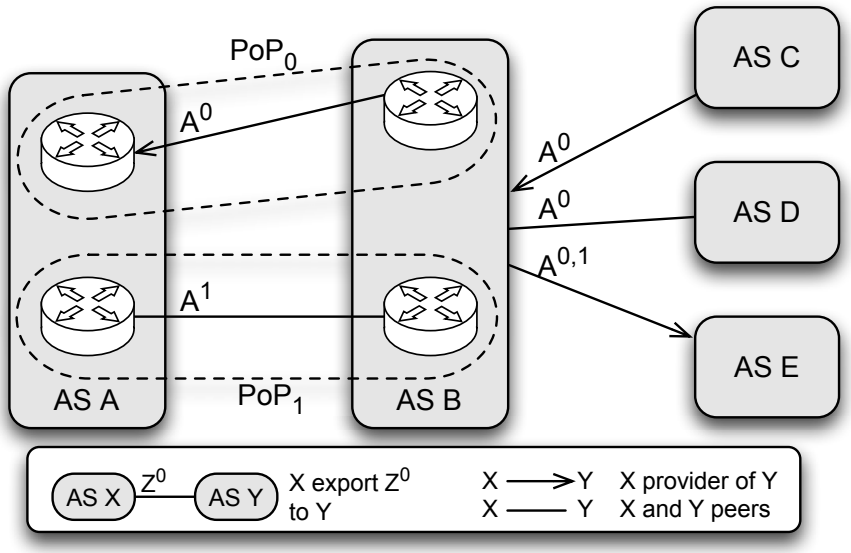


Figure 5.4: Hybrid relationship between ASes A and B. In PoP_0, A is a provider of B, while in PoP_1 they are peers. A's export policy depends on the relationship type at each PoP.

link.

If I found no $transit$ prefixes for a link, it was likely mis-inferred as p2c and is discarded. If all prefixes crossing a link are tagged as *full transit*, then the link is left as a conventional *provider-customer* relationship. If all the prefixes crossing a link are tagged as *partial transit*, the the link is classified as *partial transit*. If a link has a mixture of *full transit* and *partial transit* or has any *peering* prefixes, I tag it as a candidate *hybrid* and classify it in section 5.5.

Since this methodology requires the AS to have provider links, I am currently not able to infer complicated links among the ASes that are members of the Tier-1 clique. At the end of this process, I infer 2,955 partial transit and 6,682 candidate hybrid links.

5.5 Hybrid Relationships

Hybrid relationships differ by region. Figure 5.4 illustrates a hybrid relationship between ASes A and B. Assuming that B follows a valley-free export policy, it will export prefixes A_0 and A_1 differently at each location. B will export prefixes A_0 received in PoP_0 to its customers, providers and peers because it provides But B will export prefixes A_1 it receives at PoP_1 only toward A's customers. PoP-specific export policies will be observed only for B. If A also follows the valley-free policy it will export prefixes received by B only to its customers irrespective of the PoP. This example illustrates how I infer a hybrid relationship by associating different export patterns with

	community	DRoP	Netacuity
num. links	2,305 36.7%	2,931 46.7%	1,046 16.6%

Table 5.2: The number of links which were geolocated by each system. The systems are attempted in order: community, DRoP, and Netacuity.

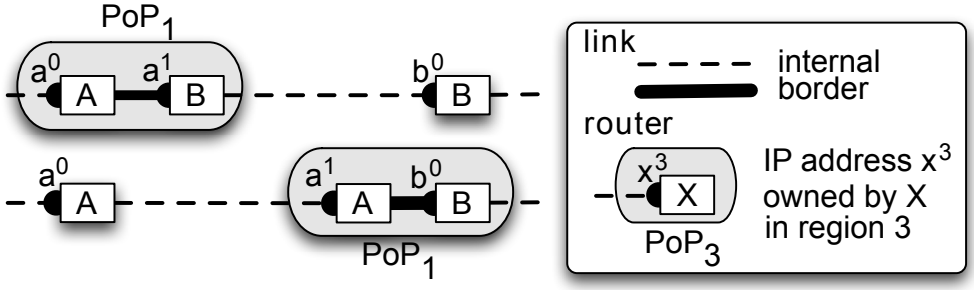


Figure 5.5: Regardless of which two routers form the interdomain link, interface a^1 is always part of the pair forming the interdomain link and so will be located in the same region.

different interconnections points (PoPs).

5.5.1 Geolocation of Ingress Points

Different export policies for different prefixes is an indication of hybrid relationships, but it may also reflect traffic engineering practices that restrict the scope of route advertisements. To distinguish between hybrid relationships and traffic engineering, my technique relies on the ability to differentiate paths based on geographic locations. To do so I attempt to geolocate to a city the ingress point of each prefix seen crossing a candidate *hybrid* link. Geographic locations can be expressed in different levels of granularity, from continent to street address. I choose city as the level of geolocation granularity because it offers the best trade-off between geolocation accuracy and level of detail. It is part of the future work to try to distinguish PoPs in the same city [68] to improve the geolocation accuracy. For each candidate *hybrid* link's prefixes I search for a BGP community that stores a geographic hint, which is the most trusted source of data and also minimizes the workload we impose on public traceroute servers. This first step resolves 36.7% of links.

For each remaining prefix I try to find a traceroute to the prefix across its candidate *hybrid* link in Ark traces. I map the IP path each Ark trace with a destination in the targeted prefix to an AS path, using the longest matching prefix in the original BGP table. If the resulting AS path contains the candidate *hybrid* link, I geolocate the IP address just before the AS changed, first by hostname string and then by Netacuity.

Although it is easy to identify the point where the AS changes, identifying the interdomain link is more difficult, since changing the position of the interconnect link does not affect the observed order of the IP addresses. Consider the observed IP path a^0 a^1 b^0 in figure 5.5: changing the interdomain link from a^0–a^1 to a^1–b^0 has no effect on the IP path. That is, although it is not clear whether the interdomain link was just before or after the AS changed in the path, it does not matter for geolocation. The IP address just before the change will be on one of the two routers connected to the interdomain link (figure 5.5) and so will be in the same location as the target link.

For the set of prefixes for which I find no usable Ark trace, I coordinate a distributed traceroute campaign among the public traceroute servers. I first determine which monitors *might* targeted prefixes through the candidate link, depending on their relationship to the provider of the candidate link. We divide the available vantage points into three sets.

The first set includes traceroute hosts under the customer cone of the provider of the candidate link, which may be able to reach any of the targeted prefixes through the candidate *hybrid* link. I name this group our *full-visibility* team of vantage points for the specific link.

The second group of vantage points includes only traceroute hosts that reside in ASes that peer with the provider of the candidate link, or in ASes under whose customer cone the provider belongs. I consider traceroutes from these hosts able to reach prefixes tagged as either *full transit* or *partial transit* through the candidate hybrid link, but not prefixes tagged as *peering*, since it would violate the valley-free rule. I group these hosts into the *limited-visibility* vantage points.

The third group of vantage points includes those monitors that, according to the valley-free rule, should not be able to reach any of the targeted prefixes; I consider these as the *zero-visibility* host group, and subsequently do not use them to probe the particular candidate link.

Considering the example in Figure 5.4, a vantage point inside AS E would belong to the full-visibility set. A vantage point inside AS C or AS D would belong to the limited-visibility set, while a vantage point in a provider of D would belong to the zero-visibility set. 93 of the treceroute monitors are in zero-visibility set for all of the candidate hybrid links; I do not use these monitors for my active measurements. Note

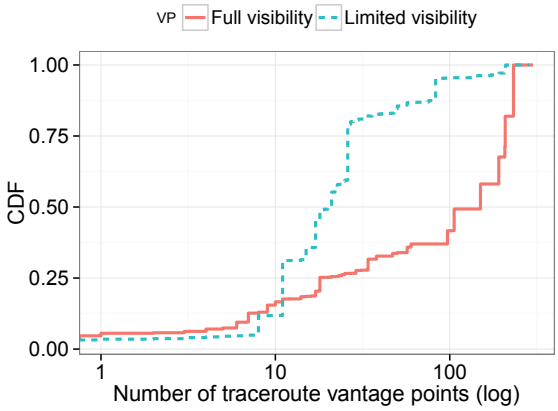


Figure 5.6: CDF of the number of traceroute vantage points for each candidate hybrid link per type of visibility. For 3% of the candidate hybrid links we have no full-visibility vantage point. For 85% of the candidate links we can distribute probing across more than 10 full- and limited- visibility vantage points.

that having a traceroute host in the full- or limited-visibility groups does not necessarily mean that paths originating from this host *will* cross the target prefixes, only that it *could*.

Completely measuring the PoP-level connectivity for a candidate hybrid link requires at least one vantage point of full visibility that reaches the target prefixes through the candidate link. Figure 5.6 shows that for 97% of the candidate hybrid links there is at least one traceroute monitor of full-visibility, and for 85% of the of the links there are more than 10 full-visibility monitors. For 462 of the 6,682 candidate hybrid links, we could not identify ingress points due to lack of both full-visibility vantage points and geographic communities.

After classifying vantage points I randomly select an IP address from each target prefix and distribute probing to it across vantage points. I sanitize the collected paths to remove incongruities. If a path contains loops, unresponsive or unresolved IP interfaces before or after the candidate link, I retain only the portion of the path that includes the candidate hybrid link. If an unresponsive interface lies between the ASes of the candidate hybrid link (e.g. "$B_1B_2 * A_1A_2$"), I discard the path. If an unresolved interface lies between the ASes of the candidate hybrid link (e.g. "$B_1B_2?A_1A_2$"), I use the PeeringDB's reverse DNS scan [14] to determine if it belongs to an IXP. If it does, I use the location of the IXP as the interconnection point, otherwise I discard the path.

After path sanitation, I have an IP address associated with the each AS link, figure 5.5. I first try to map this IP address to a hostname, so I can use the DNS string

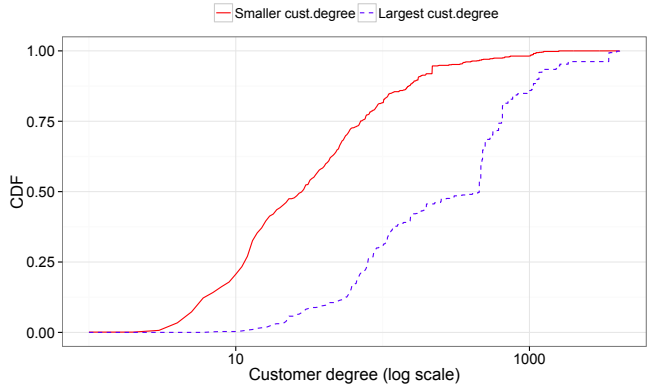


Figure 5.7: Number of customers for each AS involved in the inferred hybrid relationships.

rules to geolocate it; this step gave me locations for for an additional 46.7% of links. I resolved the remaining 16.6% of links by mapping their border IP address using the NetAcuity geolocation database.

Before doing the final AS link classification, I need to classify each of its PoPs, with the export policies observed across the link at that PoP. For each PoP, I examine the tagged prefixes that cross it using the same logic we used when we tagged its prefixes. If at least one *full transit* prefix crosses the PoP then it is added to the link's *full transit* PoP set. If no *full transit* prefix crosses the PoP, but at least one *partial transit* prefix crosses the PoP then we add that prefix to the link's *partial transit* set. If no *transit* type prefix crosses the PoP, but at least one *peering* prefixes does, I add that prefix to the link's *peering* set. If all the PoPs end up in the same type of set, I classify the link as that type. Note that it is impossible for a link to only have *peering* PoPs, since links with only *peering* prefixes are discarded. The remaining links we label as some type of hybrid, based on the combination of sets with PoPs (full/partial, full/peering, etc).

5.5.2 Results

Of the 90,272 provider-customer relationships inferred for March 2014, I infer 4,026 of these relationships as complex, including 1,071 hybrid AS relationships and 2,955 partial-transit relationships. Therefore, 4.5% of all the transit relationships inferred were are either hybrid or partial-transit. The inferred hybrid relationships include 969 full-transit/peering relationships, 72 partial-transit/peering relationships and 30 full-/partial-transit relationships. In four cases I encountered hybrid relationships with three relationship types across different PoPs (full-transit, partial-transit, and peer-

	Hybrid			Partial		
	TP	FP	FN	TP	FP	FN
Direct feedback	33	2	1	2	0	0
Communities	124	10	4	158	5	0
RPSL	45	-	-	38	-	-

Table 5.3: Validation results, showing true positives (TP), false positives (FP), and false negatives (FN). My algorithm has a 94.9% positive predictive value using direct and communities data.

ing).

The results indicate that hybrid relationships not only between large transit providers, but often between smaller ASes. Figure 5.7 plots the CDF of the number of customers (customer degree) for ASes involved in each hybrid link. In 25% of the hybrid peering relationships, the smaller AS of the relationship has fewer than 10 customers. The many hybrid links at the edge of the AS graph are mainly due to the prevalence of IXPs in Europe, which enables small ASes to achieve multinational presence within Europe and manage different relationship types for different countries. For 61.5% of the inferred hybrid links, the peering relationship crosses a European IXP. Moreover, a common practice among European IXPs is to mutually exchange routes, which further increases the density of international peering. I found that 12% of the intra-European hybrid relationships cross two interconnected IXPs.

The prevalence of IXP peerings within Europe also leads to the many observed partial-transit relationships. According to my findings, 88% of the inferred partial transit links are established between European ASes making partial-transit an interconnection practice which is almost exclusive in Europe. However, the number of customers does not necessarily indicate a network's size. For instance, I inferred 21 hybrid relationships that involve Akamai (AS 20940). Although Akamai has no customers it is one of the top 10 networks in terms of interdomain traffic volume [140]. According to self-reported network type in PeeringDB, 27% of the ASes involved in the inferred hybrid links consider themselves to be heavy-outbound content providers.

5.5.3 Validation

I use three sources of validation data: direct e-mail feedback, BGP communities that signal relationship type, and relationship types expressed in different Routing Policy Specification Language (RPSL) objects.

Direct feedback: I obtained feedback from seven operators (of twelve contacted) that previously contributed AS relationship corrections through CAIDA's web interface [146]. I sent each operator my inferences for their ASes and asked them to specify if they were correct, and asked them if they were involved in other hybrid or partial-transit relationships not included in my inferences.

BGP communities: I compiled a dictionary of 1,502 communities defined by 281 ASes, which I used to extract a set of 40,820 relationships for March 2014, as explained in [146]. Relationship communities enable the implementation of sophisticated routing policies, so operators have a strong incentive to configure community values with correct relationship annotations [92]. I considered five types of relationship communities: customer, partial-customer, peer, partial-provider, and provider. I used the partial-customer and partial-provider communities to obtain validation data for partial transit relationships. The algorithm captures hybrid relationships when it observes that an AS tags different inbound prefixes from the same neighbour with different community values depending on the ingress PoP. To mitigate transient misconfigurations I required that the same communities to be observed in all BGP snapshots I collected.

RPSL: I used RPSL objects to evaluate only true positives, since RPSL objects are commonly expressed at a high level and in most cases do not encode complex relationships.

Table 5.3 summarizes my validation results. Overall, I was able to confirm 202 hybrid relationships (19.4% of the total hybrid inferences), and 198 partial-transit relationships (6.7% of the total partial-transit inferences). For the validation datasets that allowed us to test both true- and false-positives the algorithm had the correct inference for 157/169 (92.8%) of the inferred hybrid relationships, and for 160/165 (97%) of inferred the partial transit relationships. The algorithm failed to infer five hybrid relationships present in my validation dataset. For four, both the transit and the peering PoP were located in the same city, and my city-level geolocation granularity was too coarse to identify the different PoPs.

5.5.4 Discussion

Despite the large number of complex relationships inferred, and positive validation, there are several limitations. I have not studied the time dimension of relation-

ships, which would require periodic execution of the inference algorithm. I also do not attempt to infer paid peering relationships because their routing behaviour is similar to unpaid peering [146]. The only difference of paid peering is that one of the peers pays to gain access to the other peer's customers and own networks, while in the conventional peering there is no cost involved for any of the two peers. However, in terms of export and import policies paid peering appears to follow identical rules with conventional peering. Therefore, paid peering follows a financial settlement similar to that of a transit relationship (one AS pays another AS to gain access to parts of its routing table), but the routing behaviour of a conventional peering relationship.

Generally, unconventional relationships can have arbitrary complexity that may not be expressible in terms of relationship types, which is why focused on two most common types of complex relationships according to operator feedback and the existing literature.

A second limitation relates to the well-known AS topology incompleteness problem [179]. The algorithm can only infer hybrid links for which the peering component is observable. To discover a hybrid transit/peering relationship the peering part should be visible to the available vantages points, Finally, the algorithm depends on the accuracy of external data such as the inference of conventional AS relationships and geolocation information. Errors in these data sources can corrupt my inference of complex relationships. Fortunately, CAIDA's AS relationship inference algorithm in [146] has been extensively validated and was found to exceed 99% accuracy.

Via feedback with AS operators I found that three of the inferred hybrid links were the result of misconfigurations, where the intended relationship was provider-customer. One operator followed up with a request for a semi-live system to alert operators of hybrid peering links, since they may be unintentional. Such misconfigurations are difficult to be detected because an accidental peering link does not cause violations of the valley-free rule and does not lead to sudden peaks of traffic volumes, as it happens when a peering or customer link is leaked as transit. The reported misconfigurations happened because an operator forgot to shut down a peering session, and because some IXP route server filters were not set properly to block the customer from receiving peering routes.

5.6 Summary

This chapter presented a new algorithm to infer the two most common types of complex AS relationships: hybrid and partial transit. I combined passive and active measurements along with geolocation data to achieve per-prefix relationship inference at the granularity of PoP-level connectivity. I inferred 1,071 hybrid and 2,955 partial-transit relationships for March 2014. I validated our inferences against a dataset of 417 AS relationships obtained through direct feedback from AS operators, and BGP communities, and found our results to be 94.7% accurate. I believe this is the first partly validated attempt to infer complex (fine-grained) AS relationships. My results reveal that complex relationships are more prevalent in the periphery of the AS topology than previously thought, while 61% of hybrid peering and 88% of partial transit relationships are between European ASes that leverage the Europe's extensive IXP ecosystem.

Chapter 6

Inference of Multilateral Peering

This section describes, implements, and validates a new method to reveal currently invisible AS peerings using publicly available BGP data sources. My approach infers peerings established over IXP-provided route servers. Most ASes at an IXP use a route server to implement multilateral peering (MLP) and maximize their peering. MLP is the prevalent peering paradigm in terms of number of p2p links [200, 57], therefore unearthing those links is a critical step towards more complete AS topologies. The default behaviour of route servers is to advertise all the routes they learn to all connected peers, though many IXPs allow their members to control how their prefixes are advertised by using a set of special-purpose BGP community values. I implement an algorithm to mine these community values, extract the route server participants, and infer their export policies. By combining BGP data from multiple sources I infer more than 206K p2p links from 13 IXPs for May 2013, 88% of which are not visible in publicly available BGP AS paths.

I validated 26K of these links using 70 publicly available looking glasses, finding at least 98.4% exist. In particular, these new methods find most of the peering at DE-CIX: 54K links. This IXP had more than 50K links in 2011 [28]. The peering not found by the algorithm are established bilaterally across the IXP peering fabric; this is a small fraction of the peering at the IXPs under study because my result for DE-CIX are comparable to that in [28], and the ASes that engage in bilateral peering are more likely to be selective or restrictive in their peering decisions, i.e. peer with a small fraction of members at an IXP.

I also analyse revealing topological characteristics of the discovered links. Notably, 25K (12.4%) of the links are between two stub ASes, making them impossible

to observe via BGP unless a vantage point is present in one. Similarly, 114K in my set (55.6%) involve at least one stub, reinforcing Ager's reports of dense peering connectivity at the edge [28]. I show that while some ASes publicly advertise a restrictive or selective peering policy, they engage in dense multilateral peering in selected geographical regions.

6.1 Introduction

Many findings in the AS topology research have been characterized as controversial due to the widely documented incompleteness of the existing topology datasets [59, 141, 65, 178, 58, 166]. The most widely-used data sources for compiling the AS graph use BGP and/or traceroute data. The incompleteness problem derives from the inability of these data sources to observe most p2p links due to policy restrictions imposed by operators that limit propagation of p2p links. The proliferation of Internet eXchange Points (IXPs) to support cost-effective dense peering amplifies the gap between what exists and what can be observed. Researchers have proposed a number of methods to find missing p2p links, such as more effective placement of vantage points [209, 186, 109], aggressive deployment of traceroute monitors at edges of the network [58, 184, 180], or combining different data sources including Internet Routing Registries (IRRs) and looking glass servers [55, 206, 124, 36]. In 2012, Ager *et al.* used sampled traffic (sFlow) data from a large European IXP to discover more peering links at this single IXP than were previously believed to exist in the whole Internet [28].

BGP data is the most widely used source of AS topology data. Route collectors operated by Route Views [21] and RIPE RIS [16] passively collect BGP messages and provide public archives of routing tables and update messages. Routing tables and update messages include an AS Path attribute which identifies the sequence of ASes visited before the route was received, also known as the *control path*. The AS path is the primary source of AS links and is generally considered a reliable source.

However, misconfiguration, route hijacks, and path poisoning may induce false links [166, 179]. Other sources of BGP data include Looking Glass (LG) servers that allow the remote execution of non-privileged BGP commands through a web interface or remote login. In the general case LG servers do not allow full BGP table dumps, and typically they are used in one-off queries and not periodic data collection due to

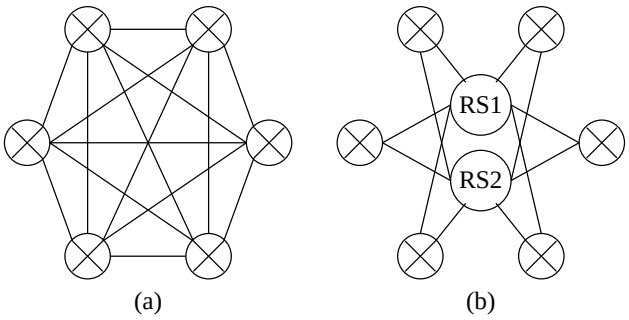


Figure 6.1: Bilateral (a) vs Multilateral (b) peering for a full-mesh connection between six ASes. Bilateral peering requires $n \cdot (n-1)/2$ BGP sessions, while multilateral peering requires only $c \cdot n$ sessions with c route servers.

the high cost of performing prefix-specific queries.

A second popular source of topology information is IP-level paths collected through globally distributed traceroute monitors. AS links can be inferred by mapping the collected IP addresses to ASNs. However, such mapping is a heuristic and can produce considerable artifacts arising from IP addresses returned by routers in paths that map to a third-party ASes [207].

A third source is the Internet Routing Registry (IRR), a publicly accessible database where AS administrators voluntarily and manually register adjacency and policy information. IRR data are frequently inaccurate, incomplete or intentionally false, although certain databases - notably RIPE - are more reliable. It has been shown that with the proper filtering techniques the IRR can provide a useful source of topology data [187, 42]. Other topology data sources exist, e.g. syslogs, but they are usually proprietary and not available to the research community.

6.2 Multilateral Peering

An increasing number of IXPs offer two interconnection paradigms, bilateral and multilateral peering, both of which are illustrated in figure 6.1. Bilateral agreements require establishing a new BGP session for every peering, which scales poorly at IXPs that mostly have participants with open peering policies who wish to maximize their peering. For IXPs with more than 50K reported peerings [28], managing a separate BGP session for each peer would involve considerable overhead. MLP offers a scalable way to support such dense peering; participants connect with one or more route servers which reflect routes learned from one participant to other participants. Further, some

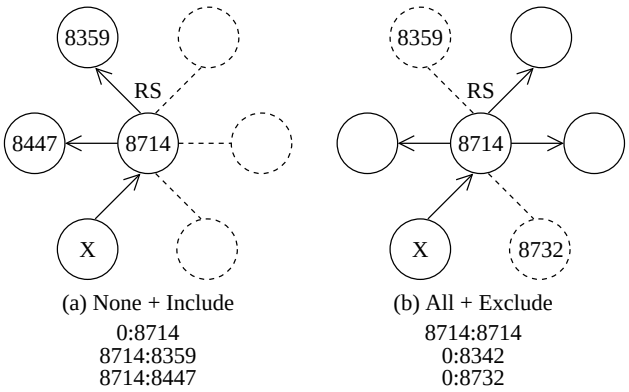


Figure 6.2: Controlling route advertisements in a route server using BGP communities. In (a), X advertises a route to two selected peers, while in (b) X advertises to all peers except two.

ASes will not enter into bilateral peering unless traffic requirements are met, but will advertise routes to a route server so that smaller ASes can reach them directly. The most notable example is Google, whose peering policy requires at least 100Mbps peak traffic to establish bilateral peering; they invite networks with less than 100Mbps traffic to peer with them via route servers [8]. Although connection to route servers is optional, a large percentage of IXPs' members opt in. For example, about 77% of the members of the two largest IXPs (DE-CIX and AMS-IX) are connected to the route servers. For the rest of this chapter I will use the term *RS members* to refer to route server members.

By default, routes sent to a route server are advertised to all RS members. However, members can control which networks receive their routes through the route server. Filtering mechanisms are essential for IXP participants because even ASes with very open routing policy may not wish to peer with everybody at a given IXP. There are several techniques to implement policy filters, but the most popular practice is through the use of BGP communities, an optional 32-bit BGP attribute used to encode additional information on a BGP route [125]. The values of BGP communities are not standardized, but IXPs clearly document the usage of their community values in IRR records or support pages. There are four common community types among all the IXPs I studied that define the following actions:

- ALL: routes are announce to all RS members. This is the default behavior.

- EXCLUDE: block an announcement toward a specific member. This action can be used in combination with the ALL community to exclude specific RS mem-

	DE-CIX	MSK-IX	ECIX
RS-ASN	6695	8631	9033
ALL	6695:6695	8631:8631	9033:9033
EXCLUDE	0:peer-asn	0:peer-asn	64960:peer-asn
NONE	0:6695	0:8631	65000:0
INCLUDE	6695:peer-asn	8631:peer-asn	65000:peer-asn

Table 6.1: Examples of patterns of community values for controlling announcements by a route server. Typically, members use ALL+EXCLUDE or NONE+INCLUDE to control announcements.

bers from receiving a route.

- NONE: Block an announcement toward all RS members. When a community type signals this action, no member receives the route unless they are listed with an INCLUDE community.

- INCLUDE: Allow an announcement toward a specific member. This action can be used in combination with the NONE community to allow only specific RS members to receive a route.

Table 6.1 shows some examples of community values for different IXPs. The `peer-asn` corresponds to the ASN of the RS member that will be included or excluded from receiving an advertisement. In this chapter I label route server community values *RS communities*. Because the `peer-asn` part of a community value is 16 bits wide, it is not possible to directly encode 32-bit ASNs. Many IXP operators map the 32-bit ASNs of their members to 16-bit ASNs in the private ASN range to enable filtering of 32-bit ASNs.

Figure 6.2 illustrates how operators use RS communities to control the announcement of routes by a route server, with rs-asn 6695. Figure 6.2a shows an example of the NONE+INCLUDE scenario listed in table 6.1. A route tagged with communities `0:6695 6695:8359 6695:8447` is advertised by the route server to ASes 8359 and 8447 only. Figure 6.2b shows an example of the ALL+EXCLUDE scenario. A route tagged with communities `6695:6695 0:5410 0:8732` is advertised to all members except ASes 5410 and 8732. Therefore, two ASes can peer via a route server if two requirements are satisfied: *connectivity* and *reachability*. Connectivity is enabled by establishing a session with the route server. Reachability is enabled by configuring

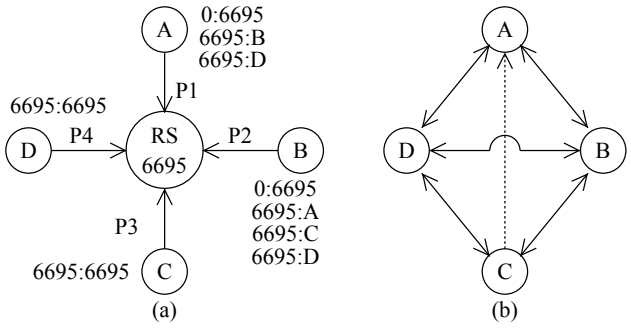


Figure 6.3: Inferring peering links over a route server using RS communities. The communities sent to the route server are shown in (a) while (b) shows the links that result. C's routes are received by A, but C blocks A from receiving its routes, so I do not infer a p2p link between A and C.

outbound filters using RS communities (when communities are used for advertisement control) and inbound AS-PATH filters.

6.3 Link Inference Algorithm

The discovery of IXP links is key to obtaining complete AS topologies. In this chapter, I explore how to discover IXP links on a regular basis, at low cost, with public or reproducible measurements. This section presents a framework for the inference of invisible IXP peering links through public BGP data. The key idea behind my methodology is that by obtaining both connectivity and reachability data via IXP infrastructure, it is possible to infer the peering links established with a route server without having to observe them in a BGP or traceroute path.

Connectivity data, namely which ASes are connected to a given route server, can be obtained from three sources of data: (i) looking glass (LG) servers that provide an interface to route servers, (ii) RPSL AS-SETs registered in IRRs by AS operators, and (iii) IXP websites that list connected networks. Information obtained from LGs is the most reliable as it explicitly reports the status of the route server routing table, although previous studies (and my own analysis) have found the other two sources to be accurate and current [124, 36]. Reachability data (RS community strings in route advertisements) can be extracted from public BGP data sources; these include passive BGP measurements (e.g. Route Views and RIS repositories) and active BGP queries of available IXP LG servers.

Before I describe the link inference methodology in detail I first illustrate an ex-

ample in figure 6.3 to explain the logic behind my algorithm. In figure 6.3 ASes A, B, C and D are connected to an IXP route server operated by ASN 6695. Each of these ASes advertises routes tagged with a set of RS community values. According to these RS communities, all ASes will receive each others' routes except C which is excluded by A. To infer peering links over the route server it is also required to know the import filters because an AS may filter some routes received. However, data on the import filters of all RS members are not available; I overcome this limitation by making the following *reciprocity* assumption: *If an RS member i does not exclude another RS member j from receiving its prefix advertisements, i will also not block the incoming advertisements from j.* Hence, a p2p link between two RS members is inferred if they have a reciprocal ALLOW export policy. In figure 6.3 (b) only A and C do not have a p2p link. In section 6.3.4 I validate the correctness of this reciprocity assumption against a 230 IRR-based import and export filters set by AMS-IX RS members.

6.3.1 Inference based on Active BGP Queries

Many IXPs provide public LG interfaces to their route servers which allow the use of non-privileged BGP commands to query the status of the route server routing table. The following steps describe the basic version of my algorithm for using the LG commands to infer the peering links over a route server:

Step 1: Obtain the ASNs and IXP IP addresses of the networks connected to the route server using the `show ip bgp` command. Let A_{RS} be the set of all connected networks on the route server.

Step 2: For each ASN $a \in A_{RS}$ collect the set of prefixes advertised to the route server using the `show ip bgp neighbor [address] routes` command. Let P_a be the set of advertised prefixes for an ASN $a \in A_{RS}$.

Step 3: For each ASN $a \in A_{RS}$ query the prefix information for a subset of its prefixes $P'_a \in P_a$ using the

`show ip bgp [prefix]` command. The prefix information will give us the set of RS community values $C_{a,p}$ applied by ASN a when it advertises a prefix $p \in P'_a$ to the route server.

Step 4: From steps 1-3, both the connectivity data (A_{RS}) and the reachability data ($C_{a,p}$) have been obtained, so I can infer the peering links established via the route

server. For each ASN $a \in A_{RS}$ I construct a set $N_a = \bigcap_p N_{a,p}$ with $N_{a,p} \subset A_{RS}$, which contains the route server participants toward which all of its routes are advertised. There are two cases depending on the type of RS communities used:

1. *ALL + EXCLUDE*: $N_{a,p} = A_{RS} - E_p$, where E_p is the list of the route server participants excluded by the RS communities, with $E_p \subset A_{RS}$.

2. *NONE + INCLUDE*: $N_{a,p} = I_p$, where I is the list of the route server participants included by the RS communities, with $I_p \subset A_{RS}$.

Step 5: For every pair of ASNs $(a, a') \in A_{RS}$ a p2p link between them is inferred if $a \in N_{a'}$ and $a' \in N_a$, i.e. only if both ASes a and a' allow each other to receive their routes.

If the IXP does not provide a LG, the RS communities can be obtained from third-party LGs of networks connected to the IXP, i.e., RS members. However, these third-party LGs cannot provide the full view of the RS communities for all the RS members, but only for those members that allow their routes to be advertised to the network that operates the LG.

6.3.2 Inference based on Passive BGP Data

In addition to active BGP querying via LG servers, my algorithm also works with passive BGP collections from Route Views and RIPE RIS archives. Using passive BGP collections offers three benefits. First, it allows us to extend my inference of p2p links, since not all IXPs provide LG interfaces to their route servers. Second, it can also reduce the number of LG queries necessary, which is explained in this section. Third, it allows us to infer historical peering trends, in conjunction with historical connectivity data through archived IRR records or website archiving services such as the Wayback Machine [19].

Although passive collections do not show the complete route server BGP table, it is still possible to obtain some RS community values. BGP communities are optional transitive attributes; they can be propagated by BGP speakers in the control plane. If at least one route server participant (or one of its customers) provides a vantage point to a Route Views or RIS route collector, it is possible to obtain the RS communities for a large number of IXP participants, depending on the number of RS peerings. For

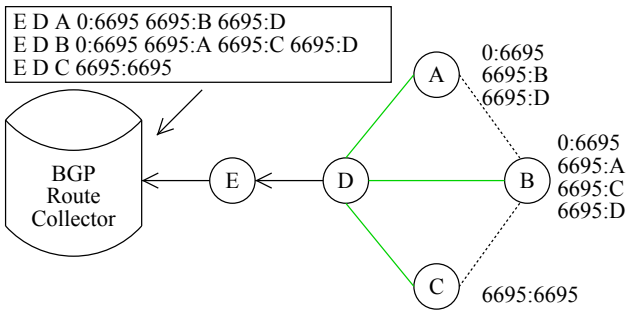


Figure 6.4: Inference of Route Server peering links through passive BGP measurements. Even though the AS paths with A–B and B–C links cannot be observed, their existence is inferred through the community values attached to routes received by E.

example, I was able to collect the RS communities for 101 of LINX's RS members from AS11666 which participates in LINX RS and contributes a BGP view to the Route Views EQIX collector. Figure 6.4 extends the example from figure 6.3. Suppose that AS E is a customer of AS D and contributes a BGP view to a collector. The IXP links that involve D (D–A, D–B, D–C) will be visible in AS paths archived by the route collector, although links (A–B, B–C) will not be seen in AS paths observed by E. However, these paths are also accompanied by the RS communities that B, C and D have applied if RS community values are propagated from the route server to AS D and then to E. Since D is the AS that provides a view of the IXP route server to the BGP collector, I will call it the *RS feeder*. Even if I have only one RS feeder from an IXP, if this feeder is densely connected I can obtain the RS communities for a large number of RS members.

One challenge with the inference of route server peerings using archived BGP data is to determine which AS applied the RS communities and at which IXP. The IXP can be determined based either on the upper or the lower 16 bits of the RS community values which typically encode the ASN of the route server. For example, it can be inferred that the RS community values in figure 6.4 were set at DE-CIX route servers because they contain DE-CIX's ASN (6695). However, sometimes there may be no RS community that encodes the ASN of the IXP. For example, consider the RS community values defined by MSK-IX in table 6.1. In the ALL+EXCLUDE scenario only the ALL community encodes 8631 which is the ASN of MSK-IX route servers. Since the ALL community is unnecessary because it is the default behaviour it may be omitted. Instead, the RS communities may only contain an array of EXCLUDE values of type

`0:peer-asn` which makes it difficult to determine the IXP route server as I described above. In such cases I infer the IXP by examining the excluded ASes; each of the excluded ASes may connect to route servers at different IXPs, but often the combination of ASes is only found at a single IXP.

After the RS communities have been extracted and the route server has been identified, I need to pin-point the AS that applied these communities, which I call the *RS setter*. I check every AS in the path against the list of the IXP's participants obtained either through RPSL objects or the IXP's website. I distinguish the following cases:

1. If the AS path contains less than two IXP participants I cannot pin-point the setter.

2. If the AS path contains two IXP participants, I identify as the RS setter the AS closest to the origin. For the first route in figure 6.4 I know that D and A are RS members; if AS E is not also a RS member then I infer A is the RS setter.

3. If the AS path contains more than two IXP participants, I need to determine which two have a p2p relationship (normally only one p2p relationship should be observed in an AS path, as explained in 2.4). For this purpose I use the AS relationships from [146] which have been shown to have over 99% accuracy for c2p relationships inferred. After I find the IXP participants with the p2p relationship, I identify the RS setter as the AS whose position in the path is closer to the IP prefix. For example, in figure 6.4, if all E, D and A are members of the route server I check the relationships of the links E – D, and D – A. Since E – D is of c2p type I infer A as the RS setter.

Having identified the RS setters S_{RS} and their community values $C_{s,p}$ I can then infer the route server peerings following steps 4 and 5 in section 6.3.1. When both active and passive measurement data are available for the same IXP route server I combine their data before executing steps 4 and 5.

6.3.3 Querying Cost

The cost of active measurements is expressed in terms of the number of queries that must be issued and processed [36]. Minimizing this querying cost facilitates more frequent collection. Keeping the cost low means that each measurement experiment

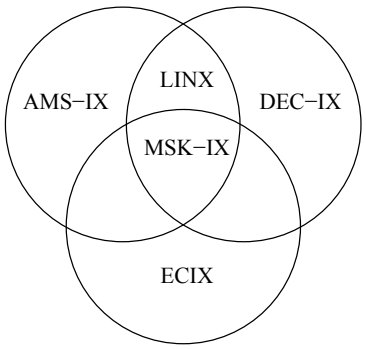


Figure 6.5: Cross examination of the excluded ASes can help us determine the IXP to which the exclusion RS communities pertain.

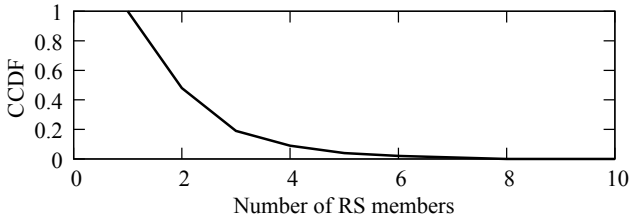


Figure 6.6: CCDF of the number of RS members advertising a given prefix to the DE-CIX route server. 48.4% of prefixes were announced by more than one member.

completes within a short time-frame that ensures the "freshness" of the obtained data. The cost c of my algorithm is given by the following equation:

$$c = 1 + |A_{RS}| + \bigcup_{a} P'_a \tag{6.1}$$

The total number of queries depends on the number of RS members ($|A_{RS}|$) and the number of prefix information commands issued for each prefix queried at step 3. During my five month analysis (January - May 2013) I found the community values applied by each RS member were remarkably consistent among their different prefix announcements toward a specific route server. In fact I found less than 0.5% of cases when a RS member had prefix announcements with different RS communities, and these differences were only found in less than 2% of their prefixes.[1] Therefore by randomly selecting 10% of the prefixes advertised by each RS member, with a maximum of 100 prefixes, I can obtain a consistent view of the RS communities.

I further reduce the querying cost by carefully selecting which IP prefixes to query. Figure 6.6 shows that 48.4% of prefixes received by the DE-CIX route server were

[1]Here I refer to prefix advertisements toward the same route server. When an AS is member of more than one IXP route servers its prefix announcements can differ significantly across IXPs, but my algorithm is applied on each route server separately.

advertised by at least two RS members. By strategically querying BGP information from an LG server for a prefix advertised by multiple RS members, I can obtain the RS communities attached by those members with a single prefix query. Consequently, I sort in decreasing order the prefixes in P_a according to the number of RS members m_p that advertise each prefix to the RS, and I start querying prefix information from the prefix with the highest m_p. In the case of the DE-CIX LG this optimization reduces the total number of queries to 8,400, which is the maximum cost I observed among all the route servers. Without these optimizations it would require 18x more queries. To even further reduce the number of LG queries, I exclude from the active queries the RS members for which the communities are collected through passive BGP measurements. Thus my optimized querying cost can be now expressed as:

$$c = 1 + |A_{RS} - A_{RS}^{passive}| + \bigcup_{a}(P'_a - P_a^{passive}) \tag{6.2}$$

where $A_{RS}^{passive}$ is the set of RS members whose RS communities can be obtained through passive measurements, and $P_a^{passive} = \emptyset$ for each $a \notin A_{RS}^{passive}$. Excluding those prefixes from the active queries reduces the total number of queries to 5,922. By conducting active measurement queries for different IXPs in parallel I can complete all measurements in less than 17 hours even with a rate limit of 1 query per 10 seconds.

6.3.4 Import and Export Filters

At the beginning of this section, I stated a *reciprocity assumption* according to which I infer the import filters: if a RS member is (not) blocked by the export filter it will also be (not) blocked by the import filter. In other words, if an AS is willing to send traffic to a RS member it will also be willing to accept traffic from it. To validate the correctness of this assumption I use import and export filter data from the members of the AMS-IX route servers. AMS-IX uses an IRR-based filtering mechanism from which the BGP configurations are automatically generated [2], so both import and export filters can be obtained for the AMS-IX members that utilize IRR filtering.

I extracted the filters of 230 AMS-IX RS members by parsing RIPE, ARIN and RADB databases from April 2013. For all of the AMS-IX members the import filters were at most as restrictive as the export filters, and often more permissive. None of

IXP	LG	ASes	RS	Pasv	Active	Links
AMS-IX	N	574	444	296	55	49249
DE-CIX	Y	483	369	113	256	54082
LINX	N	457	177*	137	39	14759
MSK-IX	Y	374	348	23	325	58501
PLIX	Y	222	211	37	174	21911
France-IX	Y	193	169	103	25	8117
LONAP	N	120	109	30	65	4458
ECIX	Y	102	83	33	50	2751
SPB-IX	Y	89	78	0	78	2828
DTEL-IX	Y	74	71	0	66	1725
TOP-IX	Y	71	52	19	33	1272
STHIX	N	69	42	4	23	340
BIX.BG	Y	53	52	0	52	950

Table 6.2: Results for the inference of MLP links per IXP. The *ASes* column shows the number of ASes at each IXP, and the *RS* column shows how many of these ASes are connected to the route server; LINX is marked with an asterisk because it does not provide a list of RS members either from its website or an AS-SET. I could only obtain partial data by searching the IRR records of LINX's members for AS8714, the ASN of LINX's route server. The *pasv* column shows the number of RS members whose community strings I obtain from passive BGP data, and the *active* column shows the number of RS members whose community strings I obtain via querying LGs, either directly from the IXP's LG (Y in the *LG* column) or from the LG of a member of that IXP. Finally, the *links* column shows the number of MLP links inferred for each IXP.

the IRR import filters blocked an AS that was not blocked by the corresponding export filter, confirming my assumption. However, about half of the import filters blocked fewer ASes than the export filters, meaning that many RS members are more open at receiving traffic, even from ASes which they do not wish to send traffic. Hence, my assumption is conservative, it does not introduce false-positives but it will miss asymmetric peering links where traffic flows in a single direction. Inference of asymmetric multilateral peering links is an open issue.

6.4 Results

I gathered the RS members and supported RS communities for 13 large European IXPs listed in table 6.2. I first collected RS communities through passive BGP data to minimize the querying load on LG servers. I accumulated daily BGP table dumps and update messages from Route Views and RIPE RIS repositories for 1-7 May 2013. I filtered out paths that contain (1) reserved, unassigned, and private ASNs (i.e. 23456 and

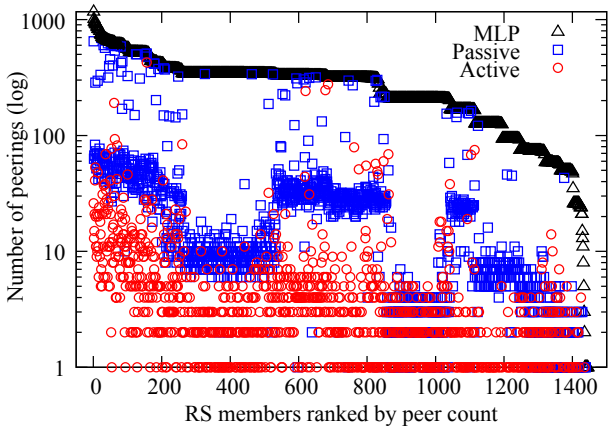


Figure 6.7: Comparison of the number of MLP links found through my algorithm against passive BGP data (Route Views, RIPE RIS, PCH) and active traceroute data (Ark, DIMES). The inferred MLP links have little overlap with links observed from current active and passive data sources.

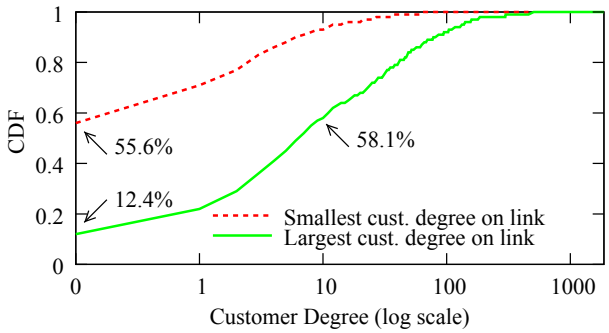


Figure 6.8: For each p2p link inferred from RS communities, the number of customers of the ASes involved. I plot separate lines representing smallest and largest degrees involved in each peering link. 55.6% of links involve one stub, 12.4% are between two stubs, and 58.1% involve ASes with fewer than 10 customers.

63488–131071) which should not have appeared in BGP advertisements and (2) path cycles that resulted from misconfiguration and poisoning. I also filtered out transient AS paths to avoid inferring short-lived links that may result from misconfigurations in setting community values. The next step is to query the available LGs for the RS communities of the remaining RS members. I wrote a script to automate this (HTTP) querying of LGs and parsing of responses. Nine of the IXPs provided an LG interface to their route servers[2]; for the remaining IXPs I use 11 LGs provided by their RS members. I can only obtain partial connectivity using only passive BGP data and third-party LGs to collect an IXP's RS communities; in the future I plan to integrate more data sources to expand our view of these IXPs. I filter out results with truncated output to

[2]France-IX's LG does not output RS communities, so I used Renater's LG which has a feed from France-IX RS.

avoid underestimating the number of the included or excluded RS communities. Truncated output occurs when certain LG interfaces do not print the entire response from the router because it exceeds a length threshold. From the 20 LGs I used in my measurements, only ECIX's LG has truncated output for 1 RS member, which I discarded. By combining connectivity information and the RS communities collected by passive and active measurements, I inferred 206,667 multilateral peering links between 1,363 different ASNs. The summation of links in table 6.2 is larger than 206,667 because 11,821 of the inferred links appear between ASes that co-locate in multiple IXPs. AMS-IX and DE-CIX have the largest link overlap with 7,502 common route server peerings, which is expected given that 123 of the ASes in my study are connected in both IXPs. However, I was only able to collect community data for a part of the AMS-IX and LINX members and therefore the actual link overlap may be larger.

Figure 6.7 compares the visibility of the MLP links in my dataset against topological data obtained from passive (BGP AS paths) and active (AS paths inferred from traceroute [4, 5]) measurements during the same period. I ordered the RS members by their inferred MLP links, so the line breaks correspond to clustering of IXPs; that is, the BGP-based visibility of each IXP is relatively consistent for all members of that IXP. In the Route Views and RIPE RIS data there were 58,952 visible peering links out of 153,837 total AS links. Only 24,511 (11.9%) of the p2p links are common between my dataset and the public BGP view. Hence, my measurements reveal 209% more peering links, and 18% more AS links than the public BGP view. The inferred links have very little overlap with the links visible to the existing publicly available traceroute topology data: I found only 3,927 links that also appeared in Ark-based and DIMES-based (traceroute-inferred) AS links for the same period [4, 5]. This minimal overlap can be explained by the fact that both Ark and DIMES do not infer links across IXP Route Servers, but report them as links between the RS members and the Route Servers. Therefore, it is likely that the common links between my dataset and the traceroute topologies are between ASes that peer both through Route Servers and bilateral connections.

To further explore the low visibility of these links in BGP and traceroute-derived AS paths, I examined the customer degrees of the ASes I infer established a p2p link via a route server. The customer degree expresses the number of customers to which

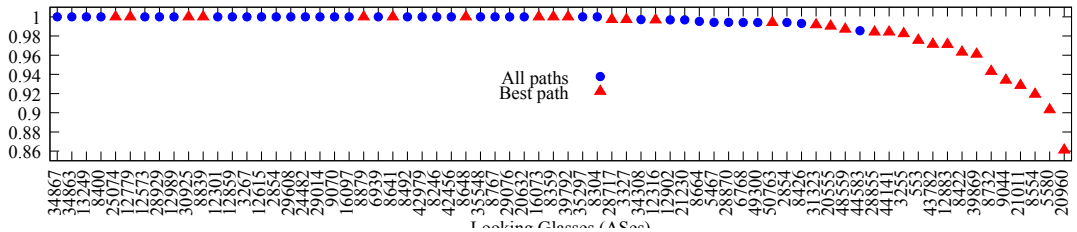


Figure 6.9: The fraction of successfully validated MLP links per AS, classified by the LG type used for validation. Circles correspond to ASes whose LG displays all paths, while triangles correspond to ASes whose LG displays only the best path. LGs that display only the best path may result in lower successful validation ratio when less preferred links are hidden.

an AS provides transit. The customer degree is not affected by the number of invisible links which are almost exclusively p2p links (see section 2.6.1). Figure 6.8 shows the degree distributions for the AS with the smallest degree on the link and for the AS with the largest degree. In contrast to what is observable in public BGP data, 12.4% of the p2p links in my MLP-inferred dataset are between two stub ASes at the edge of the network. Because these links occur at the edge they are visible in a path only from the ASes involved; unsurprisingly only 1.4% of these links are present in BGP AS paths at Route Views and RIPE RIS. Reinforcing the dense peering at the edge which is enabled by IXPs, 55.6% of all 206K links involve a stub, and 58.1% of the links involve ASes with at most 10 customers.

The small overlap between the MLP-revealed links and the publicly available topologies indicates that MLP links constitute the largest fraction of the invisible AS links in these IXPs.

6.4.1 Validation of Link Inference Algorithm

To validate the correctness of my link inference framework I test the agreement of the inferred links against connectivity information extracted from other public LG servers. By querying the PeeringDB database I collected the addresses of 70 LGs that are relevant to the inferred links, meaning the LG offers an interface to the collectors of an RS member or one of its customers. For every inferred link relevant to a particular LG, I try to confirm its existence by examining the AS paths returned from the command `show ip bgp [prefix]`. I use up to six different prefixes to ensure that path diversity due to traffic engineering policies will not cause the validation to miss existing links. I select the prefixes to be as geographically distant from each other as

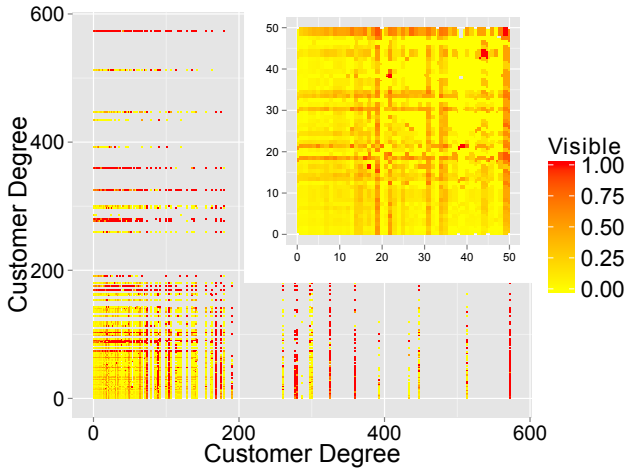


Figure 6.10: The fraction of the multilateral peering links that are visible in the topology extracted from Route Views and RIPE RIS. The inset graph zooms in the links between two ASes with customer degree less than 50. 98.6% of links between two stub ASes are invisible.

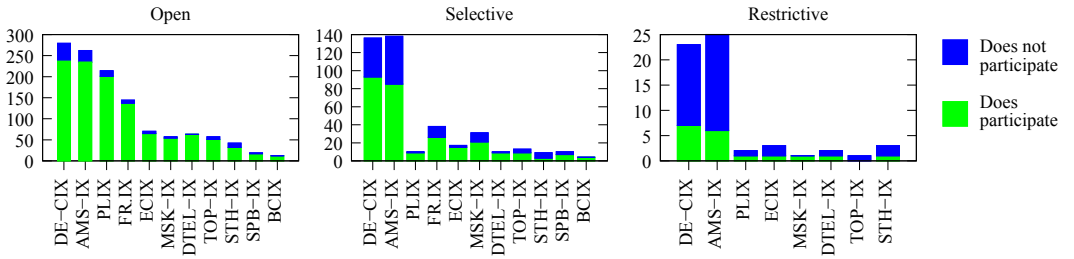


Figure 6.11: Participation in route servers compared to the self-reported peering policy. Most participants have an open peering policy and use the route servers. Route server participation is common among selective ASes, and rare with restrictive ASes.

possible, based on Maxmind's geolocation database [7].

I repeated the validation at two different time periods to ensure that the correctness of my algorithm is stable over time. In May 2013 I tested 18,100 links and I successfully confirmed the existence of 98% of those links. In October 2013 I tested 14,513 links and I was able to confirm 98.9% of them. In total I tested 26,392 different RS peerings, and I succeeded in confirming the existence of 98.4% of them. As shown in table 6.3 the validation results are consistently above 97% for all of the IXPs under study. In the intervening period between link inference and link validation some RS members were disconnected from the route servers or became idle. These cases were filtered out from the October 2013 results explaining the higher validation rate.

Observing a link in the BGP paths of an LG's output confirms the existence of this link, but the reverse does not necessarily hold (not observing a link does not necessarily

IXP	Links Validated		Links Confirmed	
DE-CIX	6250	11.6%	6152	98.4%
AMS-IX	6190	12.6%	6134	99.1%
MSK-IX	4171	7.1%	4122	98.8%
LINX	3597	24.4%	3492	97.1%
PLIX	2565	11.7%	2515	98.1%
France-IX	1162	14.3%	1126	96.9%
DTEL-IX	732	42.4%	726	99.1%
ECIX	553	20.1%	548	99.1%
LONAP	460	10.3%	455	98.9%
SPB-IX	425	15.0%	422	99.3%
TOP-IX	288	22.6%	288	100%
STHIX	77	22.6%	76	98.7%

Table 6.3: Validation of the inferred MLP links per IXP. I validate between 7.1% and 42.4% of the links inferred per IXP, and I confirm between 96.9% to 100% of the links tested.

mean that it does not exist). A path can be hidden from a LG if another path with higher local preference or lower hop count is available, and the LG displays only the active (best) path and not all the available paths. As a result the existence of links that are part of less preferred paths cannot be confirmed by querying LGs that only show active paths. Typically, paths learned from customers are assigned with higher local preference and may hide paths learned from peers. I also found that 14 ASes (out of the 70 used for validation) assigned a higher local preference value to bilateral peers than route server peers. Moreover, in 3 cases the ASN of the route server was not removed from the path making the path appear artificially longer. In all the October 2013 cases where a link failed validation a more preferred path existed. Figure 6.9 compares the validation results between the looking glass that display all the available paths against the looking glasses that display only the active path. I can see that LGs that only show the active paths restrict the validation effort.

6.4.2 Peering Policies of RS Members

As shown in table 6.2, on average 73% of an IXP's members chose to participate in MLP via an available route server. To explain this high participation rate, I collect self-reported peering policy information of IXP members from PeeringDB [14] or on the IXPs' websites where the information is available, e.g., PLIX. I was able to collect policy data for 904 out of the 1,667 IXP members, 72% of which reported an open

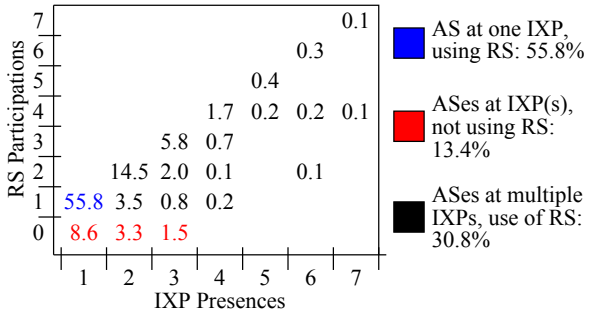


Figure 6.12: Number of IXPs an AS is connected to against route server participation at those IXPs. 55.8% of ASes in my set are at a single IXP and use its route server. 13.4% of ASes in my set do not use any route server at any of these IXPs (bottom row).

peering policy, 24% reported a selective peering policy, and 4% claimed a restrictive peering policy.

Figure 6.11 shows for each of the three self-reported peering policies of IXP members, the distribution of their participation in route servers at 11 IXPs (or only 8 IXPs that involved participants with restrictive peering policies). 92% of the ASes with open peering policies are connected to at least one route server, consistent with their self-reported interest in peering. Interestingly, 75% of the ASes with selective policy and 43% of the ASes with restrictive policy are also connected to at least one route server. These ASes decide to connect to a route server depending not only on their peering policy but also on their business strategy at a given IXP location and their desired relationships with that IXP's candidate peers.

Figure 6.12 compares the number of route servers where an AS is connected against the total number of the IXPs where the same AS is present. Most (55.8%) ASes in my set are at a single IXP and use its route server. 7.9% of the ASes with presence at multiple IXPs do not have consistent RS participation. For instance, AS9002, a large Ukrainian ISP with a selective peering policy, opts out of connecting to the route servers of Eastern-European IXPs (DTEL-IX and MSK-IX) where many of its customers co-locate, but appears to have an open peering policy in the route servers of Western-European IXPs (DE-CIX and AMS-IX). This suggests peering policies often can have local scope.

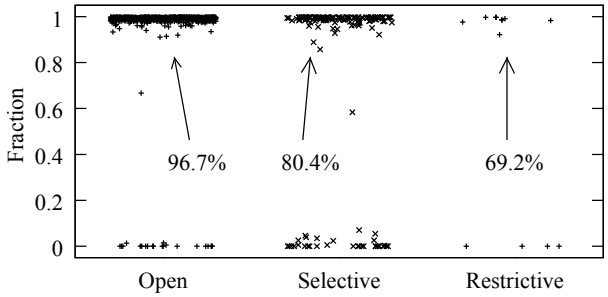


Figure 6.13: Fraction of RS members allowed to receive prefix announcements from an AS via the RS as a function of that AS's self-reported peering policy in PeeringDB. Nearly all ASes allow either (1) nearly all, or (2) only a small fraction of RS members to receive their routes; the use of RS communities does not scale well for implementing finer-grained filtering. Dots are scattered within each bin to allow visibility.

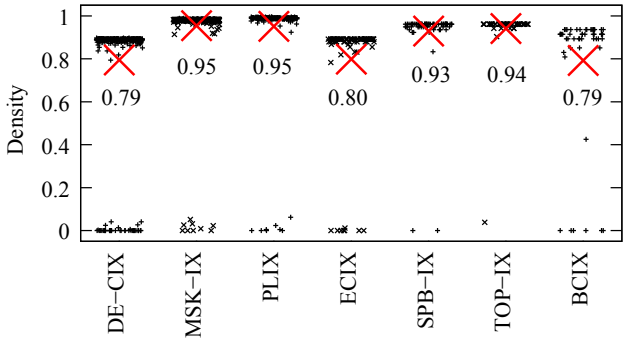


Figure 6.14: Density of peering links per RS member per IXP. For each member at each IXP, I plot the fraction of links the member established that it could have possibly formed using the IXP. The red crosses show the mean density of RS peering observed at each IXP. Dots are scattered within each IXP for visibility.

6.4.3 Route Filtering Patterns of RS Members

Export filters determine the set of RS members with which another RS member intends to peer. Figure 6.13 shows that there is a binary pattern for most ASes: either very few ASes receive routes, or the vast majority do. Specifically, almost all RS members block fewer than 10% or allow fewer than 10% of other RS participants from receiving their paths. This pattern is congruent with the nature of the most common RS community filters (ALL + EXCLUDE and NONE + INCLUDE); the use of these RS communities does not scale well for implementing finer-grained filtering over route servers, especially for IXPs with hundreds of members.

While the fraction of ASes that use an IXP route server is smaller for ASes that self-report their routing policies as selective and restrictive than for those self-reporting as open (figure 6.11), it is still the case that most ASes using route servers have an open

peering policy. Also, a network's observable MLP behaviour is not always consistent with its reported peering policy. Indeed, 69.2% of ASes self-reporting as restrictive and connecting to route servers are also open in establishing RS peerings, though perhaps in only certain regions. An example of dual peering policy is Blackberry which advertises open peering at Route Servers, but selective peering for bilateral arrangements. A more meaningful classification of RS participants' policies relies on observing their export filters, rather than the peering policy they self-report in PeeringDB. Studying individual IXPs reveals region-specific (differences in) peering policies. For example, AS12779, an Italian ISP, allowed only 3 RS members of TOP-IX (an IXP in Italy) to receive its prefixes, but blocked only 5 RS members in DEC-IX (an IXP in Germany).

6.4.4 Peering Density

Peering density can be expressed as the fraction of peering links that a RS member has established over the number of all possible peering links they could establish via the RS. Figure 6.14 plots this peering density metric for members of the route servers for which I have full connectivity data through the corresponding LG interfaces.[3] The density of RS peering links is between 80%-95%, depending on the number of RS members with open peering policies. Previous work has found the overall peering density of European IXPs to be around 70%, including multilateral and bilateral agreements [36, 28, 51], suggesting that the density of RS-based peering environments is higher than the density of bilateral peering environments.

6.4.5 Repellers

RS members that are blocked by the EXCLUDE communities can be described as *repellers* by the ASes that set those community values. Of the 1,363 RS members in all the IXPs, 570 are blocked by at least one AS. Figure 6.15 shows the number of times an AS is blocked compared to the geographic scope of its operations. Although it may seem surprising that global networks are the top repellers of routes by other ASes, these networks are connected to multiple IXPs and there is a significantly larger number of potential RS members they may repel. Of the 1,795 applications of EXCLUDE RS communities, only 12% are set by a provider to block a customer that co-locates in

[3] DTEL-IX is excluded because its LG server restricts queries for 5 RS members who do not wish to disclose their connectivity.

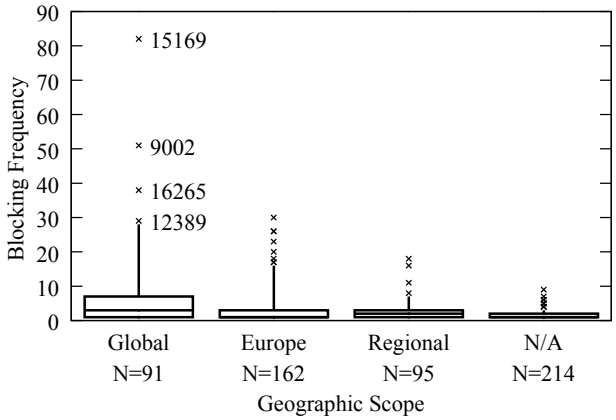


Figure 6.15: The distribution of blocking frequency using exclude community values, by geographic scope. ASes with an N/A geographic scope did not register their scope in PeeringDB. Google's AS (15169) is blocked 82 times by 75 different ASes that have another peering with Google that they prefer.

the same route server. To explore the rest of the exclusions, I calculate the customer cones for each RS member using the algorithm in [146]. The customer cone includes the set of ASes in the downstream path of a provider. I find that 77% of the EXCLUDE community values are used to block an AS that is part of the customer cone. Interestingly the most widely blocked network is AS15169 (Google), which is blocked 82 times by 75 different ASes. This filtering behavior is counter-intuitive for networks characterized by open peering, as AS15169 is an attractive peering partner due to the large volume of traffic that it carries. But in these cases, the AS that blocks AS15169 has a private peering with it in the same IXP or another PoP and prefers the use of the direct peering over the multilateral peering, which I confirmed through the available looking glasses and the IRR records of the RS members. The same behaviour is also observed for other large content providers like AS20940 (Akamai), which is blocked 14 times. Hence, the repulsiveness of an RS member is relative to the route server and not necessarily a global characteristic.

6.4.6 Hybrid Relationships

I observed that 1,230 of the RS links visible in passive BGP data are inferred as provider-customer by CAIDA's relationship inference algorithm [146]. I attempted to clarify whether these relationships are indeed hybrid p2p/p2c relationships, if they have been mistakenly inferred p2c, or if they are actual transit relationships over IXPs which are known to exist although IXP operators discourage such relationships. I collected

422 relationship-tagging and ingress-point tagging BGP community values, defined by 85 different ASes that are involved in 440 hybrid relationships. Combining the two community types, I can learn the relationship type at the different points-of-presence. I was able to verify 202 of these relationships as location-specific hybrid relationships. Moreover, I observed asymmetric routing for 16 of the links for which I obtained data from both ASes involved, namely the provider selected the path through the transit interconnection point while the customer preferred to route through the IXP. However, it is difficult to generalize these findings for all the links given the lack of additional data.

6.4.7 The Full Picture

My results reveal that the incompleteness problem is much larger than previously believed. Not only the publicly available data miss the majority of peering links but also past works on link discovery underestimated the peering density at IXPs (e.g. [36, 58]). For example [36] inferred a peering density of 26.8% for AMS-IX according to the published data [34], which is much lower than the observed peering densities both in this chapter and in [28]. To put into perspective the topology incompleteness problem and the contribution of my work, in this section I attempt to estimate the number of IXP peerings globally.

Past works have found that the peering density of Europeans IXPs ranges between 60% – 70% [28, 52]. These findings are supported by data obtained through public IXP peering matrices [11, 9, 17]. According to Cardona *et al.* [52] the peering density depends heavily on the pricing model. Flat-fee pricing encourages the establishment of more peerings and results in peering density close to 70%, while usage-based pricing leads to peering density close to 60%. The availability of route servers is a second important factor that leads to increased peering density by enabling multilateral peering [200, 57]. I utilise these findings to calculate the number of peering links within large European IXPs. For the 37 largest European IXPs with at least 50 members, I collected data from peering registries [6, 14] and individual IXP websites on the members, the pricing model, and the availability of route servers. For IXPs with a flat-fee pricing and available route servers I assume a peering density of 70%, for IXPs with usage-based pricing and available route servers I assume a peering density of 60%,

while for IXPs with no route servers I assume a peering density of 50%. Based on these assumptions I estimate the number of European IXP peerings to 558,291. In the case of the highest possible link overlap among those IXPs I estimate 399,732 unique AS peering links.

To expand my estimation I gathered data for all the IXPs globally with at least 50 members. In total I compiled data for 61 IXPs (37 in Europe, 14 in North America, 11 in Asia/Pacific, 1 in Latin America, and 1 in Africa), in which 8,577 different ASes are connected. I maintain the same assumptions on peering density except for the North American IXPs that have a primarily for-profit business model that results in lower peering density [36, 57]. Assuming a peering density of 40% for North American IXPs I estimate the global number of IXP peering links to 686,104, and the number of unique AS peering links to 510,870. For a more conservative estimation I assume that no IXP has peering density over 60%; in this case the global number of IXP peerings would be 596,011, while the total number of unique AS peerings would be 422,423.

6.4.8 Limitations of my methodology

Despite the substantial number of invisible links discovered, a large fraction of the AS topology is still missing. First, my methodology is limited to the discovery of multilateral peering agreements, and does not capture bilateral peering links. MLP is more prevalent in the European Internet, where I focused this work. IXPs in North America support mostly bilateral peerings, although there are notable exceptions such as Equinix, Any2 and Telx. Moreover, my algorithm requires that a route server utilize BGP communities to control path advertisements, and that these communities are not filtered out. Although this configuration is popular, alternative techniques exist. For example, the Vienna Internet Exchange (VIX) and the Hong Kong Internet Exchange (HKIX) provide a web portal to configure export filters, while the Netnod route servers strip out all community values before propagating paths to its RS members.

6.5 Summary

A variety of challenges in sustainable measurement instrumentation, the nature of the Internet's routing architecture, and the complexity of the ecosystem render topology incompleteness an inherent part of Internet science.

Using new techniques to mine IXP route server data with a mapping of BGP community values, I inferred 206K p2p links from 13 large European IXPs, four times more p2p links than are directly observable in public BGP data. My approach uses only existing BGP data sources, and requires only few active queries of LG servers, facilitating reproducibility of the results. I plan to apply the methods presented in Chapter 6 to additional IXPs and BGP data sources. I will also investigate how my algorithm can reduce the cost of traceroute measurements that target the discovery of IXP peering links, to enable sustainable operational measurement infrastructure using such methods.

Although I implemented and validated a new capability to analyse the dense establishment of peering at IXPs, I emphasize that a significant open question relates to how much traffic such IXP peering links carry, which would be an enlightening measure of their relative importance in the global topology.

Chapter 7

Conclusions

This thesis presented a set of novel measurement and inference algorithms to develop a new type of Internet inter-domain map that improves the state-of-the-art in inter-domain cartography. The produced map aims to accurately capture the diverse but complex AS connectivity annotated with the appropriate business relationships. The results contribute towards moving the field of inter-domain research beyond the coarse-grained representations of the AS topology as simple graphs, to a more realistic abstraction that is both mutli-graph and hyper-graph.

7.1 Summary of Research Work

When I begun my PhD research it had became evident that the existing approaches for the inference of business relationships were overly simplistic with questionable accuracy. Therefore they could not be utilised to tackle actual Internet engineering problems, such as the prediction of AS paths [153, 65, 179]. Despite this realisation, the research community lacked the data to develop more sophisticated inference techniques and instead it had been suggested that relationship-agnostic models should be preferred to mitigate the uncertainties of inference algorithms [159, 160]. However, agnostic models are inevitably more limited in terms of applications. At the same time, the different inference algorithms produced conflicting results despite claiming incompatible levels of accuracy [202, 76, 197].

I realised early in my Ph.D. that to achieve progress in relationship inference it was necessary to obtain ground-truth data on BGP policies in order to thoroughly understand the actual complexity of inter-domain routing and reduce the reliance on poorly validated heuristics. Although operators do not allow direct access to their router con-

figurations, BGP attributes that encode their policies can be found in publicly available BGP data. Benoit and Bonaventure [79] showed that BGP Communities indeed encode a rich set of policy data and urged their use in inter-domain research. However, interpreting the Communities to meaningful information requires an intimidating amount of manual effort to discover and parse relevant documentation sources [129]. As a result their use for data collection has been minimal in Internet research.

My first contribution (Chapter 3) was to develop a tool-chain for automatically collecting, interpreting and sanitising BGP Communities by processing IRR records and online policy documents found in PeeringDB. The result was to compile an extensive dictionary of Community values that provided an unprecedented amount of ground-truth data. This dataset was augmented with policy data encoded in Local Preference values. The ground-truth dataset provided a new tool to critically assess popular modelling heuristics, such as the valley-free "rule", the perceived symmetry in peering links, and Gao's abstraction of AS relationships.

My preliminary results attracted the attention of the Cooperative Association for Internet Data Analysis (CAIDA), who are the leading research groups of Internet measurements. CAIDA acknowledged the limitations of their past algorithm (section 3.4.5) and sought to develop a more accurate algorithm. Our collaboration offered me the chance to further develop my understanding on the problem of relationships inference, and provided me with access to valuable private datasets, including direct feedback from 142 AS operators. The outcome of our collaboration was the development of a new relationship inference algorithm, the complexity of which is indicative of our efforts to avoid generalisations and simplistic assumptions (Chapter 4). Even though the algorithm infers only conventional relationships, the inference process is composed by 11 different steps, each one designed to capture a different aspect of the observed complex behaviour of BGP policies. The algorithm's predictive policy value (PPV) has been validated to 99.6% for c2p links, and 98.7% for p2p links. The high level of accuracy allows the study of the transit IP market through the customer cones, and reveals the effects of topology flattening on the interconnection practices of the top providers.

A significant limitation of the above algorithm is the fact that it can simply annotate AS links with a single relationship type. Consequently, it can only infer the conventional relationship types of provider-to-customer (p2c) and peer-to-peer (p2p).

As shown in section 3.4 more complicated relationships exist, such as the hybrid and partial transit agreements. The geographical dimension of the complex relationships, and their restrictive export policies make it hard to distinguish them from traffic engineering applied on conventional relationship types. The PoP-level topology provides the appropriate resolution to model complex relationships [163], but there are great difficulties in obtaining the PoP-level connectivity of each AS link due to the limited coverage of the measurement infrastructures, and the inaccuracies in router geolocation.

To enable the inference of hybrid and partial transit relationships (Chapter 5), I combined an array of recent techniques in traceroute probing (Ark, distributed traceroute servers) and router geolocation (Communities, DDRoP and Netacuity). Combining different measurement techniques instead of treating them as disjointed allowed me to maximise their utility, since they offer different advantages in terms of accuracy and coverage. One of my priorities was to implement a measurement methodology that does not incur unrealistic computational and querying cost. By using hints from passive BGP data to orchestrate targeted traceroute probing I reduced the number of required traceroute measurements to fingerprint only the potential complex relationships. My results revealed that about 5% of the visible inferred p2c links are either hybrid or partial transit. A surprising finding to me was that complex relationships were not only observed among large providers who have more power to negotiate unconventional agreements, but 61.5% of the inferred hybrid relationships involved stub ASes. The large number of hybrid relationships at the periphery of the AS graph can be explained by the extensive IXP ecosystem in Europe, and the peering openness of many large ISPs that favour peering relationships even with their customers.

A second great challenge in inter-domain topology research has been the incompleteness of the visible topology [65, 28]. During my study of the routing and peering policies I realised that a particular IXP technology, the route servers, has been highly influential in the evolution of peering at the edge of the AS graph. Route servers operate as route reflectors among the members of an IXP, and allow them to switch their peering on and off through the use of simple redistribution policies. Hence, by observing these policies it is possible to infer the peering connectivity over route servers, given that we already know which ASes connect on the fabric of an IXP.

The above observation led me to develop a new topology inference algorithm that accurately discovers multilateral peering links at IXPs that support route server redistribution Communities. My results from 14 large European IXPs revealed massive peering meshes among the IXP participants and contributed 200% more p2p links in the visible AS topology. This approach succeeded in revealing such a large number of new links because it constitutes a completely new paradigm in topology discovery. Past works relied solely on the direct observation of AS links which requires the ever-increasing deployment of new vantage points [124, 36, 58], that leads to diminishing returns due to the nature of path propagation in inter-domain routing [149]. Furthermore, my algorithm relied solely on public data which have been shared with the research community, in contrast to recent works that were restricted by non-disclosure policies [28, 56].

7.2 Discussion

The practical use of internet cartography is part of an ongoing debate within the academic networking community regarding the applicability of topology research in real-world engineering problems. During my Ph.D. work I had the opportunity to establish direct communication with a large number of network operators from different geographic regions that manage diverse types of ASes. The interaction with the operators allowed me to understand how the industry utilises the topology and relationship datasets. I found that most network operators are highly interested in accurate inferences and are willing to initiate lengthy communication and spend considerable time to validate the data and suggest improvements. AS relationship data are actively used in policy making, network intelligence, and connectivity coordination. The existence of highly successful commercial services that rely on topology measurements to provide business intelligence such as Dyn Research (Renesys) [176] indicates that operators are not only willing to spend time but also money to obtain topological insights. However, business intelligence methods are private within each organisation and therefore they are hard to be appreciated by outsiders.

Another point of discussion is why over a decade of Internet topology research did not result in the deployment of a new routing protocol given the shortcomings of BGP. Nonetheless, this is an inherent problem with Internet research in general, due

to the cost involved in transitioning from an existing technology to a new one. The delays in the deployment of the IPv6 protocol are indicative of this problem: despite the depletion of IPv4 addresses has been predicted at least since 2000 [53], the deployment of IPv6 has been predominantly experimental until very recently [132, 66]. More importantly, the fact that past measurement and inference methodologies suffered from serious limitations led many researchers to distrust past findings which clearly hindered the adoption of proposed algorithms and protocols. I believe that extensive validation efforts like the ones presented in this thesis can significantly increase the confidence of Internet engineers in the outcomes of the Internet cartography research.

7.3 Future Directions

Although the field of inter-domain cartography has been active for over a decade, it has suffered from the very limited availability of topological and policy data that introduced bias and uncertainties in the results of many past studies. To tackle this fundamental challenge new research initiatives are funded to perform large-scale collection of network data, increasingly from the edge of the Internet topology [15, 93, 50, 192, 39, 180]. The new datasets create new opportunities for more accurate Internet mapping, but require careful handling to avoid the mistakes of the past.

Future research should strive to make measurement data publicly available, when not restricted by privacy and security concerns. Although my experience has been that data sharing incurs a significant maintenance overhead for the contributors of data, it can greatly benefit further research by enabling individual researchers with limited resources to access measurements from heterogeneous data sources. Furthermore, data sharing allows the re-examination and re-appraisal of past results which is a crucial aspect of robust scientific analysis [31]. A large part of the work carried out in this thesis was focused on re-examination of past assumptions and heuristics which proved to be valuable for developing new measurement techniques. Also, the topology datasets produced by my work have become publicly available and have been already downloaded by more than 100 times.

Additionally, since no single source of topological data can provide a complete topology, advancing our understanding of Internet topology will require extracting all

insights we can from all data sources available, including combining data sources where possible and appropriate. Today there exists many different measurement platforms, such as Ark [4], iPlane [148], Atlas [15] and Bismark [192], that collect the same type of data independently, often leading to overlapping and redundant measurements. This thesis and past works [124, 36] illustrated that synthesis of different routing data can provide new intuition that each dataset separately cannot provide. However, the many different formats and the lack of a centralised data source prevents researchers from fully utilising the available data and prevents the repeatability of experiments. Therefore, I believe that the orchestration of the different measurement platforms should be an important part of future works.

To raise the bar in Internet measurements it is necessary to develop sound validation strategies. Although the collection of validation data is difficult and time-consuming due to the confidential nature of AS connectivity that is often protected by non-disclosure agreements, thorough validation will assist in avoiding common mistakes that led the field of AS topology research in a precarious position [179, 198]. The work presented in this thesis involved a great effort to collect validation data from multiple sources, and to ensure the hygiene of the collected and produced datasets. The validation datasets have become publicly available to encourage a community effort to develop a repository of validation meta-data, as envisioned by Krishnamurthy et al. [138]

The future work should continue the research in extending the completeness of the AS map. Despite the progress achieved in this thesis, according to the estimations in section 6.4.7 about 50% of the AS topology remains invisible. Unearthing more links will also benefit the inference of hybrid relationships, since it would be possible to capture the peering part of potentially more hybrid relationships.

Further research should also be devoted in illuminating the role of IXPs in the evolution of AS connectivity practices. The IXP ecosystem becomes increasingly complicated by offering a range of different peering paradigms, such as bilateral, multilateral and remote peering. Despite recent works on IXPs [52, 28, 57, 56] the field is relatively undeveloped and more work is required to gain a better understanding of how IXPs influence the resiliency of the inter-domain connectivity, what are the regional differences in peering strategies and how the imbalance of the geographical distribution

of IXPs can lead to inefficiencies due to circuitous routes [115].

More work is also needed for the improvement of the AS relationships. In particular, future research should focus on developing relationship inference algorithms for the IPv6 topology. As explained in section 3.6, a large fraction of the AS links have different relationships between the IPv4 and IPv6 topologies. Despite the increasing congruity of the two topologies [74], the IPv6 market exhibits distinct economics - partly due to the significantly smaller traffic volumes. Therefore different heuristics should be developed to accurately capture the IPv6 relationships.

Finally, a possibly fertile area of future research would be to explore the predictive capabilities of customer-cone-related metrics on Internet evolution and dynamics, such as characteristics that correlate with an impending switch from a c2p to a p2p relationship. More metrics should be derived to correlate the influence of ASes in global routing with their relationship types. Study of traffic volumes can offer new insights in this area but unfortunately, there are very limited data publicly available. Therefore, future research efforts on AS ranking should focus on collecting traffic traces.

Appendix A

Author's Publications

Peer-reviewed publications

- GIOTSAS, V., AND ZHOU, S. Detecting and Assessing the Hybrid IPv4/IPv6 AS Relationships. ACM SIGCOMM Computer Communications Review, Volume 41, Number 4, August 2011. DOI=10.1145/2043164.2018501
- GIOTSAS, V., AND ZHOU, S. Valley-free violation in Internet Routing: Analysis based on BGP Community data. In 2012 IEEE International Conference on Communications (ICC 12), June 2012, Ottawa, Canada. DOI=10.1109/ICC.2012.6363987
- GIOTSAS, V., AND ZHOU, S. Improving the Discovery of IXP Peering Links through Passive BGP Measurements. In 16th IEEE INFOCOM Global Internet Symposium, April 2013, Turin, Italy. DOI=10.1109/INFCOM.2013.6567146
- GIOTSAS, V., ZHOU, S., LUCKIE, M., AND CLAFFY, K. Inferring Multilateral Peering. In 9th ACM SIGCOMM Conference on Emerging Networking Experiments and Technologies (CoNEXT 13), November 2013, Santa Barbara, California, US. DOI=10.1145/2535372.2535390
- LUCKIE, M., HUFFAKER, B., CLAFFY, K., DHAMDHERE, A., AND GIOTSAS, V. AS Relationships, Customer Cones, and Validation. In ACM SIGCOMM Internet Measurement Conference (IMC 13), October 2013, Barcelona, Spain. DOI=10.1145/2504730.2504735
- GIOTSAS, V., LUCKIE, M., HUFFAKER, B., CLAFFY, K. Inferring Complex Relationships. In 10th ACM SIGCOMM Internet Measurement Conference

(IMC 14), November 2014, Vancouver, Canada.

Invited Workshop Talks

- GIOTSAS, V. Exposing the Complexity of AS Relationships. CAIDA Workshop on BGP and Traceroute data. August 2011, La Jolla, California, US. DOI=10.1145/2317307.2317313

- GIOTSAS, V. Inferring AS relationships from BGP attributes. Tenth Mathematics of Networks (MoN10) meeting. September 2011, Loughborough University, UK.

- GIOTSAS, V. Fine Grained AS Relationship Inference. CAIDA mini-Workshop on BGP and Traceroute Measurements. May 2014, La Jolla, California, US.

Technical Reports

- GIOTSAS, V., AND ZHOU, S. Inferring AS relationships from BGP attributes. CoRR abs/1106.2417 (2011).

Bibliography

[1] About the Open Internet. European Commision. http://ec.europa.eu/digital-agenda/en/about-open-internet (Retrieved: 09/08/2014).

[2] AMS-IX Route Servers. https://www.ams-ix.net/technical/specifications-descriptions/ams-ix-route-servers#IRRdb.

[3] Autonomous System (AS) Numbers. http://www.iana.org/assignments/as-numbers/as-numbers.xml.

[4] CAIDA: Archipelago Measurement Infrastructure. http://data.caida.org/datasets/topology/ipv4.allpref24-aslinks/.

[5] DIMES AS Edges. http://www.netdimes.org/PublicData/csv/ASEdges4_2012.csv.gz.

[6] Euro-IX. https://www.euro-ix.net/.

[7] GeoLite IP Geolocaton Database. http://dev.maxmind.com/geoip/legacy/geolite.

[8] Google Peering Policy. https://peering.google.com/about/peering_policy.html.

[9] GR-IX Peering Matrix. http://www.gr-ix.gr/services/peering-matrix.shtml.

[10] iffinder alias resolution tool. http://www.caida.org/tools/measurement/iffinder/.

[11] INEX Peering Matrix. https://www.inex.ie/ixp/peering-matrix.

[12] Internet Routing Registries. http://www.irr.net/.

[13] Netacuity. `http://www.digitalelement.com/our_technology/our_technology.html`.

[14] PeeringDB. `http://www.peeringdb.com`.

[15] RIPE Atlas. https://atlas.ripe.net/.

[16] RIPE's Routing Information Service. `http://www.ripe.net/ris`.

[17] RONIX Connection Matrix. `http://www.ronix.ro/matrice.php`.

[18] Routing registry consistency check project. http://www.ripe.net/rrcc.

[19] The Wayback Machine. `http://archive.org/web/web.php`.

[20] Umass mnil software repository. http://rio.ecs.umass.edu/mnilpub/download.

[21] University of Oregon RouteViews Project. `http://www.routeviews.org/`.

[22] APNIC WHOIS database. http://wq.apnic.net/apnic-bin/whois.pl.

[23] RIPE WHOIS database. http://www.ripe.net/db/whois.html.

[24] The 32-bit AS number report. http://www.potaroo.net/tools/asn32/, May 2010.

[25] Internet topology collection. http://irl.cs.ucla.edu/topology/, May 2010.

[26] France Telecom Accused Of Holding YouTube Videos Hostage Unless It Gets More Money, January 2013. `https://www.techdirt.com/articles/20130102/02113921537/france-telecom-accused-holding-youtube-videos-hostage-unless-it-gets-more-money.shtml` (Retrieved: 09/08/2014).

[27] ACHLIOPTAS, D., CLAUSET, A., KEMPE, D., AND MOORE, C. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *J. ACM 56*, 4 (2009), 1–28.

[28] AGER, B., CHATZIS, N., FELDMANN, A., SARRAR, N., UHLIG, S., AND WILLINGER, W. Anatomy of a large European IXP. In *SIGCOMM 2012* (2012), pp. 163–174.

[29] ALAETTINOGLU, C., VILLAMIZAR, C., GERICH, E., KESSENS, D., MEYER, D., BATES, T., KARRENBERG, D., AND TERPSTRA, M. Routing policy specification language (RPSL), 1999.

[30] ALBERT, R., AND BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics 74*, 1 (2002), 47.

[31] ALLMAN, M. On changing the culture of empirical internet assessment. *ACM SIGCOMM Computer Communication Review 43*, 3 (2013), 78–83.

[32] ANDERSON, T., PETERSON, L., SHENKER, S., AND TURNER, J. Overcoming the internet impasse through virtualization. *Computer 38* (2005), 34–41.

[33] AUGUSTIN, B. IXP Mapping Project, September 2009. `http://www-rp.lip6.fr/~augustin/ixp/`.

[34] AUGUSTIN, B. Ixp mapping project, September 2009. `http://www-rp.lip6.fr/~augustin/ixp/#datasets` (Retrieved: 10/11/2014).

[35] AUGUSTIN, B., CUVELLIER, X., ORGOGOZO, B., VIGER, F., FRIEDMAN, T., LATAPY, M., MAGNIEN, C., AND TEIXEIRA, R. Avoiding traceroute anomalies with paris traceroute. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2006), ACM, pp. 153–158.

[36] AUGUSTIN, B., KRISHNAMURTHY, B., AND WILLINGER, W. IXPs: mapped? In *IMC '09* (2009), pp. 336–349.

[37] BAKER, F., AND SAVOLA, P. Ingress Filtering for Multihomed Networks. RFC 3704 (Best Current Practice), Mar. 2004.

[38] BARABÁSI, A.-L., AND ALBERT, R. Emergence of scaling in random networks. *science 286*, 5439 (1999), 509–512.

[39] BASSO, S., MEO, M., AND DE MARTIN, J. C. Strengthening measurements from the edges: application-level packet loss rate estimation. *ACM SIGCOMM Computer Communication Review 43*, 3 (2013), 45–51.

[40] BATTISTA, G. D., ERLEBACH, T., HALL, A., PATRIGNANI, M., PIZZONIA, M., AND SCHANK, T. Computing the Types of the Relationships Between Autonomous Systems. *IEEE/ACM Transactions on Networking* (2007).

[41] BATTISTA, G. D., PATRIGNANI, M., AND PIZZONIA, M. Computing the types of the relationships between autonomous systems. In *IEEE INFOCOM 2003 - IEEE International Conference on Computer Communications* (March 2003), pp. 156–165.

[42] BATTISTA, G. D., REFICE, T., AND RIMONDINI, M. How to extract BGP peering information from the Internet routing registry. In *MineNet '06* (New York, NY, USA, 2006), ACM, pp. 317–322.

[43] BIRD, S., KLEIN, E., AND LOPER, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, 2009.

[44] BLANCHET, M. Special-Use IPv6 Addresses. RFC 5156 (Informational), Apr. 2008. Obsoleted by RFC 6890.

[45] BRODKIN, J. Why YouTube buffers: The secret deals that makeand breakonline video. Ars Technica, 2013 July. `http://arstechnica.com/information-technology/2013/07/why-youtube-buffers-the-secret-deals-that-make-and-break-online-video/` (Retrieved: 09/08/2014).

[46] BRON, C., AND KERBOSCH, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *CACM* (1973).

[47] BU, T., AND TOWSLEY, D. On distinguishing between internet power law topology generators. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE* (2002), vol. 2, IEEE, pp. 638–647.

[48] BUSH, R., MAENNEL, O., ROUGHAN, M., AND UHLIG, S. Internet Optometry: Assessing the Broken Glasses in Internet Reachability. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference* (New York, NY, USA, 2009), IMC '09, ACM, pp. 242–253.

[49] BUTLER, K., FARLEY, T., MCDANIEL, P., AND REXFORD, J. A survey of BGP security issues and solutions. *Proceedings of the IEEE 98*, 1 (jan. 2010), 100 –122.

[50] CAIDA. Caida 2014-2017 program plan. `http://www.caida.org/home/about/progplan/progplan2014/` (Retrieved: 19/08/2014).

[51] CARDONA RESTREPO, J. C., AND STANOJEVIC, R. A history of an Internet exchange point. *SIGCOMM Comput. Commun. Rev. 42*, 2 (Mar. 2012), 58–64.

[52] CARDONA RESTREPO, J. C., AND STANOJEVIC, R. IXP traffic: a macroscopic view. In *LANC '12* (2012), ACM, pp. 1–8.

[53] CARPENTER, B. Internet Transparency. RFC 2775 (Informational), Feb. 2000.

[54] CHANDRA, R., TRAINA, P., AND LI, T. BGP Communities Attribute, 1996.

[55] CHANG, H., GOVINDAN, R., JAMIN, S., SHENKER, S. J., AND WILLINGER, W. Towards capturing representative AS-level internet topologies. *Comput. Netw. 44*, 6 (2004), 737–755.

[56] CHATZIS, N., SMARAGDAKIS, G., BÖTTGER, J., KRENC, T., AND FELDMANN, A. On the benefits of using a large ixp as an internet vantage point. In *Proceedings of the 2013 Conference on Internet Measurement Conference* (New York, NY, USA, 2013), IMC '13, ACM, pp. 333–346.

[57] CHATZIS, N., SMARAGDAKIS, G., FELDMANN, A., AND WILLINGER, W. There is more to IXPs than meets the eye. *ACM SIGCOMM CCR 43*, 5 (October 2013).

[58] CHEN, K., CHOFFNES, D. R., POTHARAJU, R., CHEN, Y., BUSTAMANTE, F. E., PEI, D., AND ZHAO, Y. Where the sidewalk ends: extending the Internet

AS graph using traceroutes from P2P users. In *CoNEXT '09* (2009), pp. 217–228.

[59] CHEN, Q., CHANG, H., GOVINDAN, R., AND JAMIN, S. The origin of power laws in Internet topologies revisited. In *INFOCOM 2002* (2002), vol. 2, pp. 608–617.

[60] CISCO SYSTEMS. Cisco Visual Networking Index: Forecast and Methodology, 2013 - 2018. Tech. rep., Cisco, June 2014. Available at `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf`.

[61] CLAFFY, K. Border gateway protocol (bgp) and traceroute data workshop report. *SIGCOMM Comput. Commun. Rev. 42*, 3 (June 2012), 28–31.

[62] CLAFFY, K., HYUN, Y., KEYS, K., FOMENKOV, M., AND KRIOUKOV, D. Internet mapping: From art to science. In *CATCH '09: Proceedings of the 2009 Cybersecurity Applications & Technology Conference for Homeland Security* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 205–211.

[63] CLAUSET, A., AND MOORE, C. Accuracy and scaling phenomena in internet mapping. *Phys. Rev. Lett. 94*, 1 (Jan 2005), 018701.

[64] COHEN, R., AND HAVLIN, S. Scale-free networks are ultrasmall. *Phys. Rev. Lett. 90*, 5 (Feb 2003), 058701.

[65] COHEN, R., AND RAZ, D. The Internet dark matter - on the missing links in the AS connectivity map. In *INFOCOM 2006* (April 2006), pp. 1–12.

[66] COLITTI, L., GUNDERSON, S. H., KLINE, E., AND REFICE, T. Evaluating IPv6 adoption in the Internet. In *PAM* (2010), pp. 141–150.

[67] COTTON, M., AND VEGODA, L. Special Use IPv4 Addresses. RFC 5735 (Best Current Practice), Jan. 2010. Obsoleted by RFC 6890, updated by RFC 6598.

[68] D. FELDMAN AND Y. SHAVITT AND N. ZILBERMAN. A structural approach for PoP geo-location. In *Computer Networks* (2012).

[69] DALL'ASTA, L., ALVAREZ-HAMELIN, I., BARRAT, A., VZQUEZ, A., AND VESPIGNANI, A. Exploring networks with traceroute-like probes: Theory and simulations. *Theoretical Computer Science 355*, 1 (2006), 6 – 24.

[70] DENG, W., MUHLBAUER, W., YANG, Y., ZHU, P., LU, X., AND PLATTNER, B. Shedding light on the use of as relationships for path inference. *Communications and Networks, Journal of 14*, 3 (2012), 336–345.

[71] DHAMDHERE, A. *Provider and Peer Selection in the Evolving Internet Ecosystem*. PhD thesis, Atlanta, GA, USA, 2009. AAI3364198.

[72] DHAMDHERE, A., AND DOVROLIS, C. The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh. In *ACM CoNEXT* (Dec 2010).

[73] DHAMDHERE, A., AND DOVROLIS, C. Twelve years in the evolution of the internet ecosystem. *IEEE/ACM Transactions on Networking 19*, 5 (October 2011).

[74] DHAMDHERE, A., LUCKIE, M., HUFFAKER, B., CLAFFY, K., ELMOKASHFI, A., AND ABEN, E. Measuring the deployment of ipv6: Topology, routing and performance. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference* (New York, NY, USA, 2012), IMC '12, ACM, pp. 537–550.

[75] D'IGNAZIO, A., AND GIOVANNETTI, E. Antitrust Analysis for the Internet Upstream Market: a Border Gateway Protocol Approach. *Journal Of Competition, Law and Economics* (2006).

[76] DIMITROPOULOS, X., KRIOUKOV, D., FOMENKOV, M., HUFFAKER, B., HYUN, Y., CLAFFY, K., AND RILEY, G. AS relationships: inference and validation. *SIGCOMM Comput. Commun. Rev. 37*, 1 (2007), 29–40.

[77] DIMITROPOULOS, X., KRIOUKOV, D., RILEY, G., AND CLAFFY, K. Classifying the types of autonomous systems in the internet. In *SIGCOMM '05* (New York, NY, USA, 2006), ACM.

[78] DIMITROPOULOS, X., SERRANO, M. A., AND KRIOUKOV, D. On cycles in as relationships. *SIGCOMM Comput. Commun. Rev. 38*, 3 (2008), 102–104.

[79] DONNET, B., AND BONAVENTURE, O. On BGP communities. *CCR 38*, 2 (2008), 55–59.

[80] DONNET, B., AND FRIEDMAN, T. Internet topology discovery: A survey. *Communications Surveys Tutorials, IEEE 9*, 4 (2007), 56 –69.

[81] DORIA, A., DAVIES, E., AND KASTENHOLZ, F. A Set of Possible Requirements for a Future Routing Architecture. RFC 5772 (Historic), Feb. 2010.

[82] DR.PEERING. Dual Transit/Peering. `http://drpeering.net/white-papers/Art-Of-Peering-The-Peering-Playbook.html#4`.

[83] DR.PEERING. Partial Transit (Regional). `http://drpeering.net/white-papers/Art-Of-Peering-The-Peering-Playbook.html#7`.

[84] ERLEBACH, T., HALL, A., AND SCHANK, T. Classifying customer-provider relationships in the internet. In *IASTED International Conference on Communications and Computer Networks* (Cambridge, Massachusetts, USA, November 2002), ACTA.

[85] FAGGIANI, A., GREGORI, E., LENZINI, L., LUCONI, V., AND VECCHIO, A. Network sensing through smartphone-based crowdsourcing. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems* (2013), ACM, p. 31.

[86] FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication* (New York, NY, USA, 1999), ACM, pp. 251–262.

[87] FANOU, R., AND FRANCOIS, P. Drawing the Map of West African Internet. Institute IMDEA Networks, February 2014. `http://www.networks.imdea.org/whats-new/news/2014/drawing-map-west-african-internet` (Retrieved: 10/08/2014).

[88] FARATIN, P., CLARK, D., BAUER, S., LEHR, W., GILMORE, P., AND BERGER, A. The Growing Complexity of Internet Interconnection. *Communications & Strategies*, 72 (2008).

[89] FEAMSTER, N., BALAKRISHNAN, H., AND REXFORD, J. Some foundational problems in interdomain routing. In *3rd ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets)* (San Diego, CA, November 2004).

[90] FERGUSON, P., AND SENIE, D. Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing. RFC 2827 (Best Current Practice), May 2000. Updated by RFC 3704.

[91] FLOYD, S., AND PAXSON, V. Difficulties in simulating the internet. *IEEE/ACM Trans. Netw. 9*, 4 (2001), 392–403.

[92] FOSTER, K. Application of BGP communities. *The Internet Protocol Journal 6*, 2 (2003).

[93] FOUNDATION, N. S. Nets: Large: Collaborative research:programmable inter-domain observation and control. Award Abstract 1413972. `http://www.nsf.gov/awardsearch/showAward?AWD_ID=1413972` (Retrieved: 20/08/2014).

[94] FRANCIS, W. N., AND KUCERA, H. Brown corpus manual. *Brown University Department of Linguistics* (1979).

[95] FULLER, V., AND LI, T. Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan. RFC 4632 (Best Current Practice), Aug. 2006.

[96] GALPERIN, H. Connectivity in Latin America and the Carribean: the Role of Internet Exchange Points. Tech. rep., The Internet Society, November 2013.

[97] GAO, L. On inferring autonomous system relationships in the internet. *IEEE/ACM Trans. Netw. 9*, 6 (2001), 733–745.

[98] GAO, L., AND REXFORD, J. Stable Internet routing without global coordination. *IEEE/ACM Transactions on Networking 9* (December 2001), 681–692.

[99] GILL, P., ARLITT, M., LI, Z., AND MAHANTI, A. The flattening internet topology: natural evolution, unsightly barnacles or contrived collapse? In *Proceedings of the 9th international conference on Passive and active network measurement* (Berlin, Heidelberg, 2008), PAM'08, Springer-Verlag, pp. 1–10.

[100] GILL, P., SCHAPIRA, M., AND GOLDBERG, S. Let the market drive deployment: a strategy for transitioning to BGP security. *SIGCOMM Comput. Commun. Rev. 41*, 4 (Aug. 2011), 14–25.

[101] GIOTSAS, V., LUCKIE, M., HUFFAKER, B., AND CLAFFY, K. Inferring Complex AS Relationships. In *Internet Measurement Conference (IMC)* (Nov 2014).

[102] GIOTSAS, V., AND ZHOU, S. Detecting and assessing the hybrid ipv4/ipv6 as relationships. *SIGCOMM Comput. Commun. Rev. 41*, 4 (Aug. 2011), 424–425.

[103] GIOTSAS, V., AND ZHOU, S. Inferring as relationships from bgp attributes. *CoRR abs/1106.2417* (2011).

[104] GIOTSAS, V., AND ZHOU, S. Valley-free violation in Internet routing: Analysis based on BGP Community data. In *IEEE ICC 2012* (June 2012), pp. 1193–1197.

[105] GIOTSAS, V., AND ZHOU, S. Improving the discovery of IXP peering links through passive BGP measurements. In *16th IEEE Global Internet Symposium* (Turin, Italy, Apr. 2013), pp. 3365–3370.

[106] GIOTSAS, V., ZHOU, S., LUCKIE, M., AND CLAFFY, K. Inferring multilateral peering. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies* (New York, NY, USA, 2013), CoNEXT '13, ACM, pp. 247–258.

[107] GOVINDAN, R., AND TANGMUNARUNKIT, H. Heuristics for internet map discovery. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE* (26-30 2000), vol. 3, pp. 1371 –1380 vol.3.

[108] GREGORI, E., IMPROTA, A., LENZINI, L., ROSSI, L., AND SANI, L. BGP and Inter-AS Economic Relationships. In *IFIP Networking* (2011).

[109] GREGORI, E., IMPROTA, A., LENZINI, L., ROSSI, L., AND SANI, L. On the incompleteness of the AS-level graph: a novel methodology for BGP route collector placement. In *IMC '12* (2012), pp. 253–264.

[110] GRIFFIN, T., AND HUSTON, G. BGP Wedgies. RFC 4264 (Informational), Nov. 2005.

[111] GRIFFIN, T. G., SHEPHERD, F. B., AND WILFONG, G. The stable paths problem and interdomain routing. *IEEE/ACM Trans. Netw. 10*, 2 (2002), 232–243.

[112] GUNES, M., AND SARAC, K. Importance of IP alias resolution in sampling internet topologies. In *IEEE Global Internet Symposium, 2007* (11-11 2007), pp. 19 –24.

[113] GUNES, M. H., AND SARA, K. Resolving anonymous routers in internet topology measurement studies. In *INFOCOM* (2008), IEEE, pp. 1076–1084.

[114] GUNES, M. H., AND SARAC, K. Resolving IP aliases in building traceroute-based internet maps. *IEEE/ACM Trans. Netw. 17*, 6 (2009), 1738–1751.

[115] GUPTA, A., CALDER, M., FEAMSTER, N., CHETTY, M., CALANDRO, E., AND KATZ-BASSETT, E. Peering at the Internets Frontier: A First Look at ISP Interconnectivity in Africa. In *Passive and Active Measurement* (2014), Springer, pp. 204–213.

[116] HADDADI, H., FAY, D., JAMAKOVIC, A., MAENNEL, O., MOORE, A. W., MORTIER, R., RIO, M., AND UHLIG, S. Beyond node degree: evaluating AS topology models. Tech. Rep. UCAM-CL-TR-725, University of Cambridge, Computer Laboratory, July 2008.

[117] HADDADI, H., FAY, D., UHLIG, S., MOORE, A., MORTIER, R., AND JAMAKOVIC, A. Mixing biases: Structural changes in the as topology evolution. In *Traffic Monitoring and Analysis*, F. Ricciato, M. Mellia, and E. Biersack, Eds., vol. 6003 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 32–45.

[118] HADDADI, H., FAY, D., UHLIG, S., MOORE, A., MORTIER, R., AND JAMAKOVIC, A. Mixing biases: Structural changes in the as topology evolution. In *Traffic Monitoring and Analysis*, F. Ricciato, M. Mellia, and E. Biersack, Eds., vol. 6003 of *Lecture Notes in Computer Science*. 2010, pp. 32–45.

[119] HADDADI, H., RIO, M., IANNACCONE, G., MOORE, A., AND MORTIER, R. Network topologies: inference, modeling, and generation. *Communications Surveys Tutorials, IEEE 10*, 2 (second 2008), 48 –69.

[120] HADDADI, H., UHLIG, S., MOORE, A., MORTIER, R., AND RIO, M. Modeling internet topology dynamics. *ACM SIGCOMM Computer Communication Review 38*, 2 (2008), 65–68.

[121] HANDLEY, M. Evolving the internet: Changing the engines in mid flight. Invited presentation at the ICSE 2004, Edingburgh, Scotland., May 2004. http://www.cs.ucl.ac.uk/staff/M.Handley/slides/icse.pdf.

[122] HANDLEY, M. Why the internet only just works. *BT Technology Journal 24*, 3 (July 2006), 119–129.

[123] HAWKINSON, J., AND BATES, T. Guidelines for creation, selection, and registration of an Autonomous System (AS). RFC 1930 (Best Current Practice), Mar. 1996. Updated by RFCs 6996, 7300.

[124] HE, Y., SIGANOS, G., FALOUTSOS, M., AND KRISHNAMURTHY, S. Lord of the links: a framework for discovering missing links in the Internet topology. *IEEE/ACM Transactions on Networking 17*, 2 (2009), 391–404.

[125] HILLIARD, N., E.JASINSKA, RASZUK, R., AND BAKKER, N. Internet Exchange Route Server Operations. `htp://tools.ietf.org/html/draft-ietf-grow-ix-bgp-route-server-operations-01`, Aug. 2013.

[126] HUFFAK, B., PLUMMER, D., MOORE, D., AND CLAFFY, K. Topology discovery by active probing. In *SAINT-W '02: Proceedings of the 2002 Symposium on Applications and the Internet (SAINT) Workshops* (Washington, DC, USA, 2002), IEEE Computer Society, p. 90.

[127] HUFFAKER, B., FOMENKOV, M., AND CLAFFY, K. DRoP:DNS-based Router Positioning. *CCR, to appear* (2014).

[128] HUI WANG, J., CHIU, D. M., LUI, J., AND CHANG, R. Inter-as inbound traffic engineering via aspp. *Network and Service Management, IEEE Transactions on 4*, 1 (June 2007), 62–70.

[129] HUMMEL, B., AND KOSUB, S. Acyclic type-of-relationship problems on the internet: an experimental analysis. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2007), IMC '07, ACM, pp. 221–226.

[130] HUSTON, G. Exploring Autonomous System Numbers. *The Internet Protocol Journal* (2006).

[131] HYUN, Y., HUFFAKER, B., ANDERSEN, D., LUCKIE, M., AND CLAFFY, K. C. The IPv4 Routed /24 Topology Dataset, 2014. `http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml`.

[132] KARPILOVSKY, E., GERBER, A., PEI, D., REXFORD, J., AND SHAIKH, A. Quantifying the extent of IPv6 deployment. In *PAM* (2009), pp. 13–22.

[133] KATZ-BASSETT, E., CHOFFNES, D. R., CUNHA, I., SCOTT, C., ANDERSON, T., AND KRISHNAMURTHY, A. Machiavellian routing: Improving internet availability with bgp poisoning. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (New York, NY, USA, 2011), HotNets-X, ACM, pp. 11:1–11:6.

[134] KEYS, K., AND HUFFAKER, B. Mapping autonomous systems to organizations: CAIDA's inference methodology.

[135] KHAN, A. Dataset: AS-level Topology Collection Through Looking Glass Servers, October 2013. `http://mmlab.snu.ac.kr/traces/lg/`.

[136] KHAN, A., KWON, T., KIM, H.-C., AND CHOI, Y. As-level topology collection through looking glass servers. In *IMC '13* (2013), pp. 235–242.

[137] KNIGHT, S., NGUYEN, H., FALKNER, N., BOWDEN, R., AND ROUGHAN, M. The Internet Topology Zoo. *Selected Areas in Communications, IEEE Journal on 29*, 9 (October 2011), 1765–1775.

[138] KRISHNAMURTHY, B., WILLINGER, W., GILL, P., AND ARLITT, M. A socratic method for validation of measurement-based networking research. *Computer Communications 34*, 1 (2011), 43–53.

[139] LABOVITZ, C. Six months, six providers and ipv6, April 2011.

[140] LABOVITZ, C., IEKEL-JOHNSON, S., MCPHERSON, D., OBERHEIDE, J., AND JAHANIAN, F. Internet inter-domain traffic. *SIGCOMM Comput. Commun. Rev. 41*, 4 (Aug. 2010).

[141] LAKHINA, A., BYERS, J., CROVELLA, M., AND XIE, P. Sampling biases in IP topology measurements. In *INFOCOM 2003* (2003), vol. 1, pp. 332–341.

[142] LEE, B. T. Comcasts deal with Netflix makes network neutrality obsolete. The Washington Post, February 2014. `http://www.washingtonpost.com/blogs/the-switch/wp/2014/02/23/comcasts-deal-with-netflix-makes-network-neutrality-obsolete/` (Retrieved: 09/08/2014).

[143] LIST, N. M. IPv6 internet broken, cogent/telia/hurricane not peering. `http://www.merit.edu/mail.archives/nanog/msg01006.html` (Retrieved: 20/08/2014).

[144] LOUGHEED, K., AND REKHTER, Y. Border Gateway Protocol (BGP). RFC 1105 (Experimental), June 1989. Obsoleted by RFC 1163.

[145] LUCKIE, M. Spurious routes in public bgp data. *SIGCOMM Comput. Commun. Rev. 44*, 3 (July 2014), 14–21.

[146] LUCKIE, M., HUFFAKER, B., CLAFFY, K., DHAMDHERE, A., AND GIOTSAS, V. AS Relationships, Customer Cones, and Validation. In *IMC '13* (2013), pp. 243–256.

[147] MA, R., CHIU, D. M., LUI, J.-S., MISRA, V., AND RUBENSTEIN, D. On cooperative settlement between content, transit, and eyeball internet service providers. *Networking, IEEE/ACM Transactions on 19*, 3 (June 2011), 802–815.

[148] MADHYASTHA, H. V., ISDAL, T., PIATEK, M., DIXON, C., ANDERSON, T., KRISHNAMURTHY, A., AND VENKATARAMANI, A. iplane: An information plane for distributed services. In *Proceedings of the 7th symposium on Operating systems design and implementation* (2006), USENIX Association, pp. 367–380.

[149] MAHADEVAN, P., KRIOUKOV, D., FOMENKOV, M., DIMITROPOULOS, X., CLAFFY, K. C., AND VAHDAT, A. The internet AS-level topology: three data sources and one definitive metric. *CCR 36*, 1 (2006), 17–26.

[150] MAHADEVAN, P., KRIOUKOV, D. V., FOMENKOV, M., HUFFAKER, B., DIMITROPOULOS, X. A., CLAFFY, K. C., AND VAHDAT, A. Lessons from three views of the internet topology. *CoRR abs/cs/0508033* (2005).

[151] MAHAJAN, R., WETHERALL, D., AND ANDERSON, T. Understanding BGP misconfiguration. In *SIGCOMM '02: Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications* (New York, NY, USA, 2002), ACM, pp. 3–16.

[152] MALECHA, G., AND SMITH, I. Maximum entropy part-of-speech tagging in nltk. *unpublished course-related report: http://www. people. fas. harvard. edu/gmalecha* (2010).

[153] MAO, Z. M., QIU, L., WANG, J., AND ZHANG, Y. On AS-level path inference. In *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems* (New York, NY, USA, 2005), SIGMETRICS '05, ACM, pp. 339–349.

[154] MARCHETTA, P., PERSICO, V., KATZ-BASSETT, E., AND PESCAPÉ, A. Dont trust traceroute (completely). In *ACM CoNEXT Student workshop* (2013).

[155] MAZLOUM, R., BUOB, M.-O., AUGE, J., BAYNAT, B., ROSSI, D., FRIEDMAN, T., ET AL. Violation of interdomain routing assumptions. *PAM'2014* (2014).

[156] MEDINA, A., LAKHINA, A., MATTA, I., AND BYERS, J. Brite: an approach to universal topology generation. In *9th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2001* (August 2001), pp. 346–353.

[157] MEYER, D. BGP Communities for Data Collection. RFC 4384 (Best Current Practice), Feb. 2006.

[158] MOTAMEDI, R., REJAIE, R., AND WILLINGER, W. A survey of techniques for internet topology discovery. *in submission to Communications Surveys and Tutorials* (2013). `http://mirage.cs.uoregon.edu/pub/Motamedi-topo-survey2013.pdf`.

[159] MÜHLBAUER, W., FELDMANN, A., MAENNEL, O., ROUGHAN, M., AND UHLIG, S. Building an AS-topology model that captures route diversity. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications* (New York, NY, USA, 2006), SIGCOMM '06, ACM, pp. 195–206.

[160] MÜHLBAUER, W., UHLIG, S., FU, B., MEULLE, M., AND MAENNEL, O. In Search for an Appropriate Granularity to Model Routing Policies. *SIGCOMM Comput. Commun. Rev. 37* (August 2007), 145–156.

[161] NAGESH, G. Comcasts deal with Netflix makes network neutrality obsolete. The Wall Street Journal, June 2014. `http://online.wsj.com/articles/fcc-probes-internet-traffic-slowdowns-1402679634` (Retrieved: 09/08/2014).

[162] NAGESH, G. Net-Neutrality Proposal Faces Public Backlash. The Wall Street Journal, July 2014. `http://online.wsj.com/articles/fccs-net-neutrality-proposal-faces-public-backlash-1405381100?mod=LS1` (Retrieved: 09/08/2014).

[163] NEUDORFER, L., SHAVITT, Y., AND ZILBERMAN, N. Improving as relationship inference using pops. In *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on* (2013), pp. 459–464.

[164] NEWMAN, M. E. The structure and function of complex networks. *SIAM review 45*, 2 (2003), 167–256.

[165] NORTON, WILLIAM. The 21st Century Internet Peering Ecosystem. DrPeering: The Internet Peering Playbook. `http://drpeering.net/core/ch10.2-The-21st-Century-Internet-Peering-Ecosystem.html` (Retrieved: 09/08/2014).

[166] OLIVEIRA, R., PEI, D., WILLINGER, W., ZHANG, B., AND ZHANG, L. The (in)completeness of the observed Internet AS-level structure. *IEEE/ACM Transactions on Networking 18* (February 2010), 109–122.

[167] OLIVEIRA, R. V., PEI, D., WILLINGER, W., ZHANG, B., AND ZHANG, L. In search of the elusive ground truth: the Internet's AS-level connectivity structure. In *SIGMETRICS 2008* (2008), pp. 217–228.

[168] PETERMANN, T., AND LOS RIOS, P. Exploration of scale-free networks: Do we measure the real exponents? *The European Physical Journal B - Condensed Matter 38* (March 2004), 201–204(4).

[169] PIRAVEENAN, M., PROKOPENKO, M., AND ZOMAYA, A. Y. Local assortativity and growth of internet. *The European Physical Journal B-Condensed Matter and Complex Systems 70*, 2 (2009), 275–285.

[170] POSTEL, J. User Datagram Protocol. RFC 768 (Standard), Aug. 1980.

[171] POSTEL, J. Internet Control Message Protocol. RFC 777, Apr. 1981. Obsoleted by RFC 792.

[172] QIU, J., AND GAO, L. AS Path Inference by Exploiting Known AS Paths. In *Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE* (27 2006-dec. 1 2006), pp. 1 –5.

[173] QIU, S., MCDANIEL, P., AND MONROSE, F. Toward valley-free inter-domain routing. In *IEEE ICC 2007* (June 2007), pp. 2009–2016.

[174] QUOITIN, B., AND UHLIG, S. Modeling the routing of an autonomous system with C-BGP. *Network, IEEE 19*, 6 (nov.-dec. 2005), 12 – 19.

[175] REKHTER, Y., LI, T., AND HARES, S. A border gateway protocol 4 (BGP-4). RFC 4271 (Draft Standard), jan 2006.

[176] (RENESYS), D. IP Transit Intelligence. http://dyn.com/ip-transit-intelligence/ (Retrieved: 24/10/2014).

[177] ROTHSCHILD, A. Peering disputes: Comcast, level 3 and you. INTERNAP, December 2010. http://www.internap.com/2010/12/02/peering-disputes-comcast-level-3-and-you/ (Retrieved: 09/08/2014).

[178] ROUGHAN, M., TUKE, S. J., AND MAENNEL, O. Bigfoot, sasquatch, the yeti and other missing links: what we don't know about the AS graph. In *IMC '08* (2008), pp. 325–330.

[179] ROUGHAN, M., WILLINGER, W., MAENNEL, O., PEROULI, D., AND BUSH, R. 10 lessons from 10 years of measuring and modeling the Internet's autonomous systems. *JSAC 29*, 9 (2011), 1810–1821.

[180] SÁNCHEZ, M. A., OTTO, J. S., BISCHOF, Z. S., CHOFFNES, D. R., BUSTAMANTE, F. E., KRISHNAMURTHY, B., AND WILLINGER, W. Dasu: Pushing experiments to the internet's edge. In *USENIX NSDI* (April 2013).

[181] SANDVINE. Global Internet Phenomena Report, 1H 2014. Tech. rep., 2014.

[182] SANGLI, S., TAPPAN, D., AND REKHTER, Y. BGP Extended Communities Attribute. RFC 4360 (Proposed Standard), Feb. 2006.

[183] SCHUMANN, R., AND KENDE, M. Lifting barriers to Internet development in Africa: suggestions for improving connectivity. Tech. Rep. Ref: 35729-502d1, The Internet Society, Analysis Mason, May 2013. http://www.internetsociety.org/sites/default/files/Barriers%20to%20Internet%20in%20Africa%20Internet%20Society_0.pdf (Retrieved: 10/08/2014).

[184] SHAVITT, Y., AND SHIR, E. DIMES: let the Internet measure itself. *CCR 35*, 5 (2005), 71–74.

[185] SHAVITT, Y., AND SHIR, E. DIMES: Let the internet measure itself. *CoRR abs/cs/0506099* (2005).

[186] SHAVITT, Y., AND WEINSBERG, U. Quantifying the importance of vantage points distribution in Internet topology measurements. In *INFOCOM 2009* (Apr. 2009), pp. 792–800.

[187] SIGANOS, G., AND FALOUTSOS, M. Analyzing BGP policies: methodology and tool. In *INFOCOM 2004* (2004), vol. 3, pp. 1640–1651.

[188] SIGANOS, G., FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. Power laws and the AS-level internet topology. *IEEE/ACM Trans. Netw. 11*, 4 (2003), 514–524.

[189] SPRING, N., MAHAJAN, R., WETHERALL, D., AND ANDERSON, T. Measuring ISP topologies with rocketfuel. *IEEE/ACM Trans. Netw. 12*, 1 (2004), 2–16.

[190] STROGATZ, S. H. Exploring complex networks. *Nature 410*, 6825 (2001), 268–276.

[191] SUBRAMANIAN, L., AGARWAL, S., REXFORD, J., AND KATZ, R. H. Characterizing the internet hierarchy from multiple vantage points. In *In Proc. IEEE INFOCOM* (Berkeley, CA, USA, 2001), University of California at Berkeley.

[192] SUNDARESAN, S., BURNETT, S., FEAMSTER, N., AND DE DONATO, W. Bismark: A testbed for deploying measurements and applications in broadband access networks. In *Proc. of. USENIX* (2014).

[193] VÁZQUEZ, A., PASTOR-SATORRAS, R., AND VESPIGNANI, A. Large-scale topological and dynamical properties of the internet. *Physical Review E 65*, 6 (2002), 066130.

[194] VOHRA, Q., AND CHEN, E. BGP Support for Four-octet AS Number Space, 2007.

[195] WANG, F., AND GAO, L. On inferring and characterizing internet routing policies. In *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2003), IMC '03, ACM, pp. 15–26.

[196] WANG, X. F., AND CHEN, G. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE 3*, 1 (2003), 6–20.

[197] WEINSBERG, U., SHAVITT, Y., AND SHIR, E. Near-deterministic inference of AS relationships. In *INFOCOM'09: Proceedings of the 28th IEEE international conference on Computer Communications Workshops* (Piscataway, NJ, USA, 2009), IEEE Press, pp. 377–378.

[198] WILLINGER, W., AND ROUGHAN, M. Internet topology research redux. *ACM SIGCOMM eBook: Recent Advances in Networking* (2013).

[199] WINICK, J., AND JAMIN, S. Inet-3.0: Internet topology generator. Tech. rep., Technical Report CSE-TR-456-02, University of Michigan, 2002.

[200] WOODCOCK, B., AND ADHIKARI, V. Survey of characteristics of Internet carrier interconnection agreements. Tech. rep., Packet Clearing House, May 2011.

[201] WYATT, E. F.C.C backs opening net-neutrality rules for debate. The New York Times, May 2014. `http://www.nytimes.com/2014/05/16/technology/fcc-road-map-to-net-neutrality.html?_r=0` (Retrieved: 09/08/2014).

[202] XIA, J., AND GAO, L. On the evaluation of as relationship inferences [internet reachability/traffic flow applications]. In *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE* (nov.-3 dec. 2004), vol. 3, pp. 1373 – 1377 Vol.3.

[203] XIONG, H., AND CHEN, M. Study on traffic characteristics of bgp misconfiguration. In *Communications, 2008. ICC '08. IEEE International Conference on* (may 2008), pp. 5751 –5755.

[204] YAO, B., VISWANATHAN, R., CHANG, F., AND WADDINGTON, D. G. Topology inference in the presence of anonymous routers. In *INFOCOM* (2003).

[205] YOUNG, D. Unbalanced Peering, and the Real Story Behind the Verizon/Cogent Dispute. Verizon Policy Blog, June 2014. `http://publicpolicy.verizon.com/blog/entry/unbalanced-peering-and-the-real-story-behind-the-verizon-cogent-dispute` (Retrieved: 09/08/2014).

[206] ZHANG, B., LIU, R., MASSEY, D., AND ZHANG, L. Collecting the Internet AS-level topology. *CCR 35*, 1 (Jan. 2005), 53–61.

[207] ZHANG, Y., OLIVEIRA, R., WANG, Y., SU, S., ZHANG, B., BI, J., ZHANG, H., AND ZHANG, L. A framework to quantify the pitfalls of using traceroute in AS-level topology measurement. *JSAC 29*, 9 (Oct. 2011), 1822–1836.

[208] ZHANG, Y., AND TATIPAMULA, M. Characterization and design of effective bgp as-path prepending. In *Network Protocols (ICNP), 2011 19th IEEE International Conference on* (Oct 2011), pp. 59–68.

[209] ZHANG, Y., ZHANG, Z., MAO, Z. M., HU, C., AND MAGGS, B. On the impact of route monitor selection. In *IMC '07* (2007), pp. 215–220.

[210] ZHOU, S., AND MONDRAGON, R. The rich-club phenomenon in the internet topology. *Communications Letters, IEEE 8*, 3 (March 2004), 180 – 182.

[211] ZHOU, S., AND MONDRAGÓN, R. J. Accurately modeling the internet topology. *Phys. Rev. E 70*, 6 (Dec 2004), 066108.