How biased is our Validation (Data) for AS Relationships?

Lars Prehn and Anja Feldmann
Max Planck Institute for Informatics
Saarland, Germany
<lprehn,anja>@mpi-inf.mpg.de

ABSTRACT

The business relationships between Autonomous Systems (ASes) can provide fundamental insights into the Internet's routing ecosystem. Throughout the last two decades, many works focused on how to improve the inference of those relationships. Yet, it has proven difficult to assemble extensive ground-truth data sets for validation. Therefore, more recent works rely entirely on relationships extracted from BGP communities to serve as "best-effort" ground-truth. In this paper, we highlight the shortcomings of this trend. We show that the best-effort validation data does not cover relationships between ASes within the Latin American (LACNIC) service region even though ~14% of all inferred relationships are from that region. We further show that the overall precision of 96-98 % for peering relationships achieved by three of the most prominent algorithms can drop by 14-25 % when considering only peering relationships between Tier-1 and other transit providers. Finally, we discuss potential ways to overcome the presented challenges in the future.

CCS CONCEPTS

• **Networks** → **Topology analysis and generation**; **Public Internet**.

ACM Reference Format:
Lars Prehn and Anja Feldmann. 2021. How biased is our Validation (Data) for AS Relationships?. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487552.3487825

1 INTRODUCTION

The Internet consists of many autonomous systems (ASes) that exchange reachability information (also known as routes). Which routes are made available to a neighbor often depends on business relationships. While actual business relationships are rather complex [25, 26], we often categorize them into three different types: (i) provider-to-customer (P2C), (ii) settlement-free peering partners (P2P), and (iii) relationships between ASes that belong to the same organization called sibling-to-sibling (S2S).

Many researchers rely on accurate relationship information for (i) simulations of routing incidents [48, 49, 59], (ii) IP-to-AS mapping [32, 46], or (iii) network (resource) management [40, 62]. Yet, there is no organisation or entity that can provide authoritative knowledge for those relationships. Over the last two decades, this lead to a large corpus of research focusing on inferring relationships from, e.g., routing information [22, 24, 26, 27, 36, 38, 43].

Yet, there are two major problems that those inferences suffer from: (i) limited visibility into the Internet's AS interconnection graph and (ii) lack of ground-truth validation data. The **visibility problem** is a well-known challenge in Internet topology research [3, 16, 29, 52]. While various partial solutions have been proposed (e.g., using data plane information [7, 18, 21], routing policy databases [9], or BGP community encodings at IXP route servers [28]), it is still a challenge to generate a comprehensive AS-level typology that also captures, e.g., private network interconnections [64].

The **lack of ground-truth validation data** has been pointed out as a challenge many times (e.g., [24, 43, 60]), yet recently proposed and evaluated algorithms (see, [36, 38]) rely entirely on "best-effort" validation data compiled from BGP communities—a technique initially introduced and used (among others) by Luckie et al. [43].

To better understand the implications of this trend, this paper focuses on the basic question: *How good is our "best-effort" validation (data)?* In particular, our work makes the following contributions towards answering this question:

- **Bias Analysis.** We analyze to which degree the geographical and topological biases within the sets of inferred and validated relationships match (§5). We uncover significant mismatches: While the "best-effort" validation data covers 31 % of all links between ASes in the ARIN region, it only covers less than 1 % of links in the LACNIC region. Yet, both regions contain roughly 15% of the inferred relationships.
- **Implication analysis.** We analyze how such bias mismatches may affect classification correctness for three (ASRank [43], ProbLink [36], and TopoScope [38]) classification algorithms[1] and uncover substantial drops in precision for certain groups of peering links (§6). In particular, we observe that the near-perfect precision of 96-98 % for the entire validation data set drops by 14-25 % (depending on the algorithm) for peering relationships between Tier-1 and transit providers.
- **Future outlook**: We discuss, in-depth, different approaches for compiling less biased and more complete validation data sets (§7) and highlight (i) the need for active discourse with operators and (ii) how the routing ecosystem's continuous change can be exploited to over-sample validation data.

To allow for the reproduction of our results and to facilitate the analysis of future validation efforts, we make our research code publicly available via:

https://gitlab.mpi-klsb.mpg.de/lprehn/imc2021_breval

[1] While we would have also analyzed UNARI [22], the authors do not provide publicly available artifacts.

IMC '21, November 2–4, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-9129-0/21/11.
https://doi.org/10.1145/3487552.3487825

2 WHY SHOULD WE CARE ABOUT BIAS?

Biases commonly arise in all forms of classifications—whether one looks at face detection [11], patient treatment [54], or criminal behavior [53]. While those disciplines may have stronger social impacts, the correctness of business relationships may have far-reaching and unintended consequences when studying the Internet's routing ecosystem. For instance, Müller et al. [50] recently proposed an algorithm that relies on the inferred relationships between Internet Exchange Point (IXP) members to identify spoofed packets (i.e., packets with a forged source address). The misclassification of a P2C as a P2P relationship could potentially result in many packets being falsely flagged as spoofed. If an IXP would publicly disclose, e.g., the number of spoofed packets per member, the reputation of certain members could sustain damage.

Yet, how did bias affect this example? IXPs are often built with the intention to keep local traffic local [3], i.e., they connect ASes within the same **geographical region**.[2] As most geographical regions have their own operator meetings, conferences, and communities—e.g., RIPE [57], NANOG [51], or APRICOT [5]—that release different recommendations on how to operate certain types of networks, the best practices for routing can differ among regions (and IXPs). For instance, Marcos et al. [45] recently reported that the usage patterns for AS path-prepending (a commonly used traffic engineering technique) vary strongly by region and over time. Similarly, **topological biases** can arise from how ASes of different sizes or locations within the Internet's hierarchy select their peering policies [42].

In summary, features such as the geographical or topological positioning of a network can greatly influence the routing decisions taken by its operators. This may become important when relationships are explicitly or implicitly[3] used in narrow contexts, e.g., only between members of an IXP. In such a case, the correctness estimates that were obtained from a potentially larger base of relationships may provide a false sense of safety which may result in economical consequences (as in the example above).

3 BACKGROUND

In this section, we first give a brief introduction to selected[4] relationship inference algorithms, then provide details on previously used techniques for obtaining validation data, and finally summarize the already-known sources of bias in validation data.

3.1 Classification Algorithms

Lixin Gao was the first to describe the Internet as a strict hierarchy in which customers receive transit from the providers "above" them and redistribute routes according to economically incentives [24]. Based on this hierarchy, she described the notion of a "valley-free" path—a path that travels strictly upwards, then to at most one AS of the same height, and then strictly downhill. Using this property, her proposed algorithm tries to maximize the number of valley-free paths.

Rather than maximizing the number of valley-free paths, more recent algorithms often first determine the clique of provider-free ASes at the "top" of the hierarchy and then iteratively infer relationships. In 2013, Luckie et al. [43] proposed ASRank—one of the most-used classifiers till today. ASRank utilizes AS-triplets, a new metric called "transit-degree", and an extensive list of heuristics to classify relationships. Giotsas et al. later modified the ASRank algorithm to adapt it to the IPv6 routing ecosystem [27].

In 2014, Giotsas et al. used routing information, IP paths, and geolocation data to infer two more complex types of AS relationships: partial-transit and hybrid relationships [26]. If a provider exports routes towards its customers and peers but not towards its own providers, then the provider and customer have a partial-transit relationship. Further, two ASes have a hybrid relationship if their observed relationships differ throughout various Points of Presence (PoPs).

In 2019, Jin et al. proposed ProbLink—a meta-classifier that builds upon an initial classification (e.g., from ASRank) [36]. The algorithm assigns a probability to each link to be of a certain type based on, e.g., the relationships of other nearby links, refines the selected relationship based on the highest probability, and iterates those two steps until convergence. UNARI [22] takes the idea of probability one step further and produces a measure of certainty for each link type as its outcome. TopoScope [38]—as the newest classification algorithm—applies machine learning techniques on a large set of link features to perform its classification. Notably, this algorithm also predicts additional AS links that, despite note being visible, might exist.

3.2 Validation Data

Compiling a set of ground-truth labels is crucial to properly evaluate any classification algorithm. Yet, this step has proven to be rather difficult for AS relationships. Before Luckie et al [43], only the works by Gao [24] and Dimitropoulos et al. [20] presented validation data from a Tier-1 and via operator surveys, respectively.

In 2013, Luckie et al. compiled their validation data from (i) directly reported relationships (e.g., by operators through a web interface), (ii) relationships extracted from routing policies encoded in WHOIS databases (more specifically, inside their `autnum` records) via the Routing Policy Specification Language (RPSL), and (iii) relationships extracted from BGP Community encodings within the Internet Routing Registry (IRR) databases or public documentation (e.g., ISPs that host such encoding on their website).

While relying on multiple databases allows for frequent recomputation of validation data, the sources (ii) and (iii) suffer from a set of well-known challenges. Most records within the WHOIS databases are added and maintained voluntarily, hence, some records get stale (i.e., become inconsistent with publicly visible routing information) over time [16].

While the same may be true for the publicly documented BGP community encodings, those, in addition, suffer from ambiguity problems. Simply put, BGP communities are just colon-separated value pairs[5] [14] that can be tagged onto routes. Which information is encoded into/decoded from a specific BGP community depends on the AS that sets/reads it. Ambiguity is introduced when a single BGP community represents different meanings to (potentially overlapping) sets of ASes, e.g., while the BGP community `3356:666`

[2] usually only a small fraction of ASes connect remotely [13].

[3] e.g., while using bdrmapit—a tool to map IPs to routers and ASes that relies on relationship inferences—on paths obtained from a limited number of vantage points

[4] based on significance to our work and recency.

[5] or triplets, see large BGP communities [31].

could be recognized as an attempt to blackhole a route [39], AS 3356 (Level3/CenturyLink/Lumen) uses it to tag peering routes [56].

Despite those challenges, the data compiled by Luckie et al. presents the first extensive source of validation information. Recent classification efforts rely solely on re-computations of their third data source—relationships from BGP communities [22, 36, 38].

3.3 Existing Insights into Validation Bias

Hard-to-Infer Links. Jin et al. [36] reported on sets of links for which it is challenging to infer them correctly. They describe those "hard" links as links with at least one of the following characteristics: (i) node-degree < 100, (ii) observed by $50 - 100$ vantage points, (iii) neither incident to a vantage point nor a clique AS, (iv) stub links for which there is no triplet containing two consecutive clique ASes, and (v) links for which a simple top-down classification results in a conflict. They further showed that even sophisticated algorithms like ASRank wrongly infer many of the relationships for hard links and that the validation data set is skewed towards links for which it is easy to infer them correctly.

Clique & Vantage Point Links. Luckie et al. [43] show that for their 2014 validation data set links incident to a clique AS are over-represented while links between stubs and non-clique ASes are under-represented. They also note that this disparity is mostly due to the significant biased introduced by the community-based data set—the validation data that has been used for the more recent validations. Similarly, they report that the community-based data set over-represents links incident to a vantage point over those only remotely visible.

Complex Relationships. As discussed in §3.1, AS relationships can differ based on the PoP the link is observed at. Giotsas et al. [26] reported that their improved algorithm exposed around 1k relationships as hybrid and around 3k relationships as partial-transit. As the inference of such relationships can be ambiguous, they should be handled separately during the validation process.

4 OBTAINING & CLEANING DATA

In this section, we first describe how we obtain validation and inference data (§4.1). Afterward, we take a closer look at the validation labels and identify entries that either need to be removed or handled carefully (§4.2).

4.1 Obtaining Validation Data & Inferences

Validation Data. While ASRank's validation data from April 2013 is publicly available at [12], ProbLink and TopoScope do not contain validation data in their public repositories [35, 37]. Upon request, we received the same validation data for both tools—12 snapshots unequally spread between January 2014 and April 2018. Each snapshot was generated using the community-based relationship extraction method described by Luckie et al. [43] for their ASRank validation.

Inference Data. The monthly generated inference snapshots that are publicly available for ASRank, ProbLink, and TopoScope only overlap throughout 2019. As this period is not covered by any of our validation snapshots, we requested (and promptly received) an inference snapshot for April 2018 generated by ProbLink. To produce comparable results for all three algorithms, we continue using the inference and validation snapshots for April 2018 throughout the remainder of the paper (unless explicitly specified otherwise). Notably, we use the term "inferred links" to refer to all AS links visible in the ASRank data set for April 2018.

4.2 Label Quality & Treatment

Spurious Labels. When taking a first look at the validation data, we notice 15 AS relationships formed with AS 23456. This AS is also known as "AS_TRANS" and is exclusively used to represent 32-bit ASNs for devices that only support 16-bit ASNs; therefore, AS_TRANS does not represent an actual network and hence can not have any business relationships. We further find 112 relationships involving reserved (e.g., for documentation or internal use, see [34]) ASes that should neither be publicly routed nor be used to validate business relationships.

Ambiguous Label Treatment. As briefly discussed in section 3, two ASes can have different relationships based on the PoPs they interconnect at [26]. In April 2018, the received validation data contains multiple labels for 246 relationships involving 233 different ASes. Arguably, those entries should be ignored for validation unless the classification algorithm explicitly infers or handles them; otherwise, it is ambiguous whether a simple relationship prediction is correct. Interestingly, we find that those validation entries are handled very differently in practice. If we treat an entry with multiple labels as P2P if it *starts* with P2P and otherwise as P2C, the number of P2P and P2C links in the validation data for 2017 and 2018 matches *exactly* those reported in the Toposcope paper [38]. We observe a similar match for the numbers reported for 2017 in the work by Jin et al. [36] if we treat an entry with multiple labels *always* as P2C.

Sibling Labels. Sibling (S2S) relationships represent links between two ASes that belong to the same organization and, hence, can use their resources interchangeably. When applying CAIDA's AS-to-Organisation data set [33], we find that 210 relationships in our validation data set and 2800 of the inferred relationships are actually sibling relationships and should be ignored during the validation process (unless specifically handled by the classification algorithm).

5 IS OUR VALIDATION DATA BIASED?

Regional Imbalance. As briefly discussed in section 2, how an AS routes traffic may depend on its geographic region. To analyze regional bias, we first map each ASN to a geographic service region using IANA's list of initial ASN assignments [34] and then refine the mapping based on the daily delegation files published by the Regional Internet Registries (RIRs) [2, 4, 6, 41, 58]. We abbreviate AFRINIC, APNIC, ARIN, LACNIC, and RIPE NCC as AF, AP, AR, L, and R, respectively. While IANA's list bootstraps the mapping for all ASes, the RIR delegation files correct the mapping for resources transferred between different regions after IANA's initial assignments [55]. Notably, no mapping from ASes to geographical regions is perfect; even with large amounts of active scanning, we would neither be able to reliably measure all IPs (and respectively infrastructure) that belong to an AS [8] nor would we be able to perfectly geolocate them [15]. Yet, we argue that our mapping—which relies on an AS' organizational service region rather than its

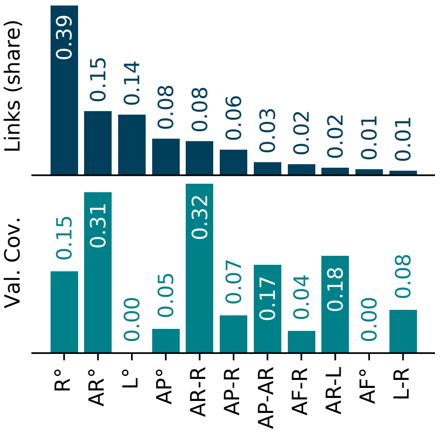

Figure 1: Regional imbalance: Fraction of links (top) and validation coverage (bottom) per geographical group with AF, AP, AR, L, and R denoting AFRINIC, APNIC, ARIN, LACNIC, and RIPE, respectively.

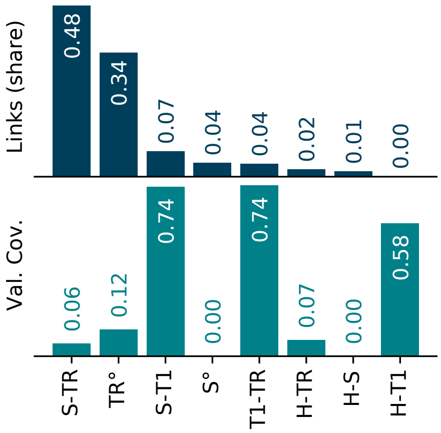

Figure 2: Topological imbalance: Fraction of links (top) and validation coverage (bottom) per topological group with H, S, T1, and TR denoting Hypergiants, Stub ASes, Tier-1 providers, and Transit providers, respectively.

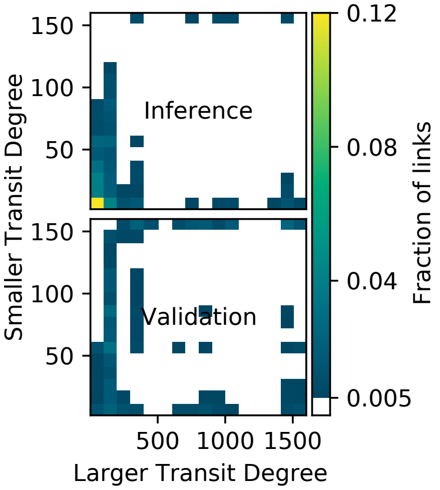

Figure 3: Transit degree imbalance for transit links: consistently colored heatmaps for inferred (top) and validatable (bottom) links, binned by the transit degree of their incident ASes.

infrastructure footprint—is still representative enough to provide hints on regional biases, if they really exist.

Using this mapping, we separate AS links into different link classes: If one of the involved ASes is reserved, we discard the link. If both ASes belong to the same region, we mark the link class as *<region>*° (e.g., *AF*° for links between two ASes in AFRINIC). If the ASes belong to different regions, we mark the link class as *<region*$_1$*>-<region*$_2$*>* where *<region*$_1$*>* is always the lexicographically smaller region, i.e., we treat AS links as undirected links.

Figure 1 shows the distribution of inferred relationships onto link classes as fractions (at the top) as well as the validation coverage (at the bottom), i.e., the fraction of links in a class for which we have validation labels. We observe that most (~79 %) of the relationships that we infer are between ASes of the same region. Yet, we observe drastic differences for the validation coverage among region-internal relationships: Even though we infer roughly the same number of *AR*° and *L*° relationships, we validate more than ~31 % of *AR*° links but less than 1 % of *L*° links.

Topological Imbalance. Next, we focus on whether the positioning of an AS in the Internet’s hierarchical structure yields a mismatch in bias. First, we classify each AS into either "Stub" or "Transit" based on whether the AS has at least one other AS in its customer cone (see CAIDA’s customer cone data set—available at [12]). Afterwards, we refine this basic mapping using two additional data sources: We re-classify ASes as (i) "Tier-1" providers based on a list from Wikipedia [63][6] and (ii) "Hypergiants" (i.e., the largest content providers) based on the list generated by Böttger et al. [10].
Figure 2 shows the topological balance based on those classes in a similar style as Figure 1. We observe that we only have substantial validation data for classes that involve Tier-1 ASes. While this insight in itself is not very new (compare [43] and [36]), we find its impact to be more drastic than previously reported: For our two majority classes, S-TR and TR°, that, in summary, contain 82 % of all inferred links, we can only validate 6 % and 12 % of relationships, respectively.

While most of the inferred links are in the S-TR class, this class is rather uninteresting as it largely consists of P2C relationships (67.8% according to validation data) for which all three classifiers are well-known to perform near-perfect. Thus, we drill deeper into our second largest class, links between Transit providers.

In particular, we want to understand whether the distribution of AS "size" matches between inferred and validated TR° links. Figure 3 shows a heatmap over all TR° links in the inferred data (top) and the validated data (bottom) where the x-axis shows the transit degree for the larger incident AS while the y-axis shows the transit degree for the smaller incident AS.[7] We observe that the vast majority of TR° links that we infer are between relatively small transit ASes (i.e., in the left-bottom corner). This mismatches with the more uniform distribution of our validation data. We further repeated this experiment with two alternative metrics: the provider-peer-observed customer cone—which relies on the correctness of the inferred business relationships and might hence be biased—and the node degree. The related figures (which can be found in Appendix B) suggest an even stronger mismatch.

6 IS OUR VALIDATION BIASED?

Now that we have a basic understanding of regional and topological bias mismatches in our validation data, we analyze how such mismatches translate to differences in classification correctness. For each of the tested classifiers, we calculate two confusion matrices

[6] which largely overlaps with the set of clique ASes inferred by ASRank.

[7] The row above 150 and the column to the right of 1500 catch all transit degree equal of larger than 150 and 1500, respectively. This prevents the few ASes with a substantially larger transit degree from distorting the plot.

Class	PPV_P	TPR_P	LC_P	PPV_C	TPR_C	LC_C	MMC
Total°	0.982	0.990	14216	0.996	0.992	30105	0.980
AP-AR	0.979	0.979	546	0.988	0.988	928	0.967
AP-R	0.985	0.987	892	0.968	0.965	338	0.952
AP°	0.992	0.992	502	0.994	0.994	648	0.986
AR-L	0.930	0.976	43	0.999	0.997	872	0.950
AR-R	0.956	0.978	1752	0.994	0.987	5707	0.957
AR°	0.926	0.954	617	0.998	0.996	12871	0.937
R°	0.990	0.996	9587	0.995	0.989	8318	0.985
S-T1	0.000	0.000	26	0.999	0.999	15533	-0.001
S-TR	0.994	0.988	2538	0.995	0.997	5334	0.987
T1-TR	0.839	0.955	641	0.996	0.985	7260	0.886
TR°	0.991	0.996	10219	0.980	0.952	1822	0.959

Table 1: Per group validation table for ASRank

Class	PPV_P	TPR_P	LC_P	PPV_C	TPR_C	LC_C	MMC
Total°	0.966	0.976	14216	0.988	0.983	30105	0.957
AP-AR	0.973	0.939	546	0.960	0.983	928	0.927
AP-R	0.973	0.995	892	0.986	0.927	338	0.940
AP°	0.976	0.989	502	0.991	0.981	648	0.969
AR-L	0.619	0.975	43	0.998	0.962	872	0.761
AR-R	0.953	0.951	1752	0.984	0.984	5707	0.936
AR°	0.951	0.859	617	0.993	0.998	12871	0.899
R°	0.971	0.988	9587	0.985	0.964	8318	0.954
S-T1	0.295	0.650	26	0.999	0.998	15533	0.437
S-TR	0.980	0.987	2538	0.994	0.991	5334	0.976
T1-TR	0.718	0.670	641	0.971	0.976	7260	0.667
TR°	0.982	0.996	10219	0.978	0.903	1822	0.930

Table 2: Per group validation table for ProbLink

Class	PPV_P	TPR_P	LC_P	PPV_C	TPR_C	LC_C	MMC
Total°	0.976	0.988	14216	0.995	0.989	30105	0.974
AP-AR	0.980	0.985	546	0.991	0.988	928	0.972
AP-R	0.983	0.994	892	0.985	0.959	338	0.961
AP°	0.986	0.992	502	0.994	0.989	648	0.980
AR-L	0.833	0.976	43	0.999	0.991	872	0.897
AR-R	0.947	0.975	1752	0.993	0.984	5707	0.950
AR°	0.930	0.943	617	0.997	0.997	12871	0.934
R°	0.984	0.993	9587	0.993	0.983	8318	0.976
S-T1	0.042	0.043	26	0.999	0.999	15533	0.041
S-TR	0.989	0.989	2538	0.995	0.995	5334	0.984
T1-TR	0.798	0.947	641	0.995	0.980	7260	0.858
TR°	0.989	0.996	10219	0.981	0.942	1822	0.954

Table 3: Per group validation table for Toposcope

(i.e., the number of True Positives, False Positives, True Negatives, and False Negatives) that result from treating either P2C links or P2P links as the "positive class."

Tables 1, 3, and 2 show the following classification correctness metrics for links of different classes[8]: (i) precision (PPV_X) and (ii) recall (TPR_x) when choosing P2P links ($X \to P$) or P2C ($X \to C$) links as positive class[9], the number of P2P ($X \to P$) and P2C ($X \to C$) links per class as LC_X, and Matthew's correlation coefficient (MCC) as symmetric evaluation metric[10].

Simply put, the MCC takes all values of the confusion matrix into account (i.e., it does not matter which class is treated as positive), is relatively robust against class imbalance (i.e., the fraction of validated P2P/P2C links in a class), and ranges between -1 and 1; values close to 1/-1 indicate positive/negative correlation between inference and validation while values close to 0 indicate correctness similar to an unbiased coin-toss [19].

Each table further colors differences between the classification correctness on the entire data set (Total°) as follows: If the per-class value is at least 1 % larger than the value for the entire data set, it is colored in green; if it is at least 1 %, 5 %, and 10 % lower, it is colored in yellow, orange, and red, respectively.

The tables first confirm common wisdom: All three algorithms perform near-perfect for P2C links. Yet, our evaluation further shows that all algorithms struggle with the same P2P link classes, namely AR-L, S-T1, and T1-TR. The low correctness for S-T1 links was already reported by [36], yet we disagree with their conclusion that "peering relationships between high-tier ASes and low-tier ASes are becoming more prevalent." We observe that most of those 26 links are formed with research ASes, anycast-based DNS providers, content delivery networks, and cloud providers, i.e., we observe that the problem lies in the broad aggregation of many diverse businesses models into a single "Stub" class, rather than a drastic change in policies. The overall correctness gap for P2P-based T1-TR relationships of up to 25 % shows that future classification efforts can still make substantial improvements for certain link classes. Yet, the increase of the correctness gap from ASRank to the two follow-up algorithms shows that following a strategy of simply improving the overall classification error can lead to substantial correctness degradation for classes that contain fewer links. Finally, the reduced correctness for AR-L relationships might hint towards unique routing policies in the LACNIC region that are not yet captured by algorithms that were constructed and validated almost exclusively on the policies present in the RIPE and ARIN regions.

6.1 Case Study: AS714 Cogent Communications

To better understand the low performance for the T1-TR class, we do a case study for AS714 (Cogent Communications). We chose AS714 as it is involved in around half (54 out of 111) of all the links that were wrongly inferred as P2P (i.e., those links that decreased PPV_P) by ASRank (which has the best precision and recall for this class). For the remainder of this section, we call those links "target links."

When analyzing the paths that include our 54 target links, we were unable to find any triplet "$C|AS714|X$" for which $AS714|X$ is a target link and C is another clique AS. This observation is critical as such triplets are necessary for ASRank to arrive at a P2C inference for $AS714|X$. While this provides us inside into what algorithmically

[8] we only show those classes that contained at least 500 relationships in summary

[9] As they only provide additional mixtures of precision and recall, we decided to not show (balanced) accuracy and f1-score.

[10] The Fowlkes–Mallows index—as the second prominent symmetric evaluation metric—showed slightly less numerical change, yet similar results.

caused the wrong inference, it does not explain why or how the routing phenomena that underpin those algorithms have changed.

To analyze target links beyond the public routing data, we focus on the 17 links that are also inferred to be P2P links in the most recent (Sept. 2021) snapshot. This allows us to directly trigger Cogent's looking glass to further investigate. We find that all the ASes involved in the 17 links consistently tag the routes they redistribute to *AS*714 with the BGP Community 174 : 991[11]. This community prevents Cogent from redistributing the received routes to other peers—including all of the other clique members.

We discussed the issue with few of the involved operators and also looked up the related RPSL routing policy objects via RADB. We found that there are two reasons why ASes tagged this community: Cogent only offers them partial transit (i.e., routes towards customers but not towards peers) and inaccurate validation data[12] (only 1 case).

7 DISCUSSION & OUTLOOK

Bias Mismatches. Throughout this paper, we demonstrated bias mismatches between inferred and validated relationships. While the features that we analyzed showed substantial mismatches, other features could introduce similar (or even greater) ones. Even though a more complex analysis of additional groups of "hard links" lies beyond the scope of this paper, we provide a list of twelve potential features for future analysis in the Appendix (§C).

Balance Through Sampling. While over-sampling of small classes or under-sampling of large classes are commonly used techniques to counteract biases, neither of them works (by default) well on AS relationships. Under-sampling prominent classes would result in a reduction of the already too small number of validated relationships. In contrast, simple over-sampling would bias the importance of specific error types (and often lead to over-fitting for ML-based classifiers). While there are more complex over-sampling methods (e.g., SMOTE [17], ADASYN [30], or MDO [1]) that synthetically (based on interpolation) produce new yet similar data points, theses techniques may introduce "incorrect" validation information when working with high dimensional data [23]. Yet, we might be able to leverage the heterogeneity and intrinsic, continuous change of the routing ecosystem to our advantage. If we understand for how long a certain set of relationships remains unchanged (e.g., via frequent exchange with network operators), we may be able to find a time frame after which the same AS can be re-sampled while still providing a unique-enough, new data point.

Future Validation Data. Most of our current validation data is passively obtained by scraping (poorly maintained) operator databases. We argue that compiling more extensive validation data requires active collaboration with network operators. In particular, we must clearly communicate incentives (e.g., services that they can benefit from) for why operators should accurately report (some of) their relationships through the channels they most commonly use (e.g., during operator meetings). A successful story using such a *do-ut-des* approach is the route collector project "Isolario." In only four years, the project acquired more peer ASes than RIPE RIS or Routeviews by partnering with HE.net. Whenever an AS connected to Isolario, HE.net would use the provided data to improve its statistics. The increase in reported size rendered the AS more attractive as a peering partner—a benefit that convinced many networks to continuously provide data.

Arguably, some operators may consider business relationships more sensitive than the routing information observed by a single (carefully selected) router. Yet, accurate information about a network's business relationships may be used to compile more valuable assets than simple statistics. One example would be router configurations generated by the Peerlock system. Peerlock utilizes relationship information to generate snippets of router configurations that prevent the redistribution of (accidental) route leaks [47].The mechanism's effectiveness may depend on the number of considered business relationships. Hence, operators might be willing to provide (and continuously update) their relationships in exchange for more secure and up-to-date Peerlock configurations. Similarly, relationship information may also be used to engineer recommendation systems for peering opportunities, i.e., rankings of beneficial IXPs (to peer at) and ASes (to peer with) for a given network.

Notably, the targeted interaction with operators could also counteract the current problem of missing validation data for an entire region that was reported in §5.

Future Research Efforts. Our analysis in §6 showed that (negligible) improvements in global classification correctness can severely impact the correctness for classes with potentially fewer links. In line with this finding, we argue that the current goal of negligibly improving the overall correctness actually hinders progress in this research space. Hence, we advocate that future efforts should be evaluated against more diverse goals. Further, given our findings from §4.2, we advocate for more careful and explicit handling of spurious labels, sibling relationships, and complex relationships during future validation efforts.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers and our shepherd, Cecilia Testart, for their valuable comments. We further thank Yuchen Jin and Zitong Jin for granting us access to the validation data for ProbLink and Toposcope.

REFERENCES

[1] Lida Abdi and Sattar Hashemi. 2015. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering* 28, 1 (2015), 238–251.

[2] AFRINIC. 2021. Extended Delegations, 20180405. https://ftp.apnic.net/stats/afrinic/2018/delegated-afrinic-extended-20180405. (2021). Last accessed: 24th April, 2021.

[3] Bernhard Ager, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. 2012. Anatomy of a large European IXP. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. 163–174.

[4] APNIC. 2021. Extended Delegations, 20180405. https://ftp.apnic.net/stats/apnic/2018/delegated-apnic-extended-20180405.gz. (2021). Last accessed: 24th April, 2021.

[5] APRICOT. 2021. Future APRICOTs. Available at https://www.apricot.net/. (2021). last-accessed: Tuesday, 20th April 2021.

[6] ARIN. 2021. Extended Delegations, 20180405. https://ftp.arin.net/pub/stats/arin/delegated-arin-extended-20180405. (2021). Last accessed: 24th April, 2021.

[7] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. Cloud provider connectivity in the flat internet. In *Proceedings of the ACM Internet Measurement Conference*. 230–246.

[11] Notably, this community is stripped before redistribution to customers; hence, it is rarely visible from the public routing infrastructure.

[12] i.e., contrary to the community-based validation data, the link is a P2P link rather than a P2C link.

[8] Shehar Bano, Philipp Richter, Mobin Javed, Srikanth Sundaresan, Zakir Durumeric, Steven J Murdoch, Richard Mortier, and Vern Paxson. 2018. Scanning the internet for liveness. *ACM SIGCOMM Computer Communication Review* 48, 2 (2018), 2–9.

[9] Giuseppe Di Battista, Tiziana Refice, and Massimo Rimondini. 2006. How to extract BGP peering information from the internet routing registry. In *Proceedings of the 2006 SIGCOMM workshop on Mining network data*. 317–322.

[10] Timm Böttger, Felix Cuadrado, and Steve Uhlig. 2018. Looking for hypergiants in peeringDB. *ACM SIGCOMM Computer Communication Review* 48, 3 (2018), 13–19.

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[12] CAIDA. 2018. The CAIDA AS Relationships Dataset, 1st April 2018. https://publicdata.caida.org/datasets/as-relationships/. (2018).

[13] Ignacio Castro, Juan Camilo Cardona, Sergey Gorinsky, and Pierre Francois. 2014. Remote peering: More peering without internet flattening. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. 185–198.

[14] R. Chandra, P. Traina, and T. Li. 1996. *BGP Communities Attribute*. RFC 1997. RFC Editor.

[15] Balakrishnan Chandrasekaran, Mingru Bai, Michael Schoenfield, Arthur Berger, Nicole Caruso, George Economou, Stephen Gilliss, Bruce Maggs, Kyle Moses, David Duff, et al. 2015. Alidade: Ip geolocation without active probing. *Department of Computer Science, Duke University, Tech. Rep. CS-TR-2015.001* (2015).

[16] Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J Shenker, and Walter Willinger. 2004. Towards capturing representative AS-level Internet topologies. *Computer Networks* 44, 6 (2004), 737–755.

[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[18] Kai Chen, David R Choffnes, Rahul Potharaju, Yan Chen, Fabian E Bustamante, Dan Pei, and Yao Zhao. 2009. Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*. 217–228.

[19] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining* 14, 1 (2021), 1–22.

[20] Xenofontas Dimitropoulos, Dmitri Krioukov, Marina Fomenkov, Bradley Huffaker, Young Hyun, KC Claffy, and George Riley. 2007. AS relationships: Inference and validation. *ACM SIGCOMM Computer Communication Review* 37, 1 (2007), 29–40.

[21] Adriano Faggiani, Enrico Gregori, Alessandro Improta, Luciano Lenzini, Valerio Luconi, and Luca Sani. 2014. A study on traceroute potentiality in revealing the internet as-level topology. In *2014 IFIP Networking Conference*. IEEE, 1–9.

[22] Guoyao Feng, Srinivasan Seshan, and Peter Steenkiste. 2019. UNARI: an uncertainty-aware approach to AS relationships inference. In *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies*. 272–284.

[23] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research* 61 (2018), 863–905.

[24] Lixin Gao. 2001. On inferring autonomous system relationships in the Internet. *IEEE/ACM Transactions on networking* 9, 6 (2001), 733–745.

[25] Phillipa Gill, Michael Schapira, and Sharon Goldberg. 2013. A survey of interdomain routing policies. *ACM SIGCOMM Computer Communication Review* 44, 1 (2013), 28–34.

[26] Vasileios Giotsas, Matthew Luckie, Bradley Huffaker, and KC Claffy. 2014. Inferring complex AS relationships. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 23–30.

[27] Vasileios Giotsas, Matthew Luckie, Bradley Huffaker, and Kc Claffy. 2015. IPv6 AS relationships, cliques, and congruence. In *International Conference on Passive and Active Network Measurement*. Springer, 111–122.

[28] Vasileios Giotsas, Shi Zhou, Matthew Luckie, and Kc Claffy. 2013. Inferring multilateral peering. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*. 247–258.

[29] Enrico Gregori, Alessandro Improta, Luciano Lenzini, Lorenzo Rossi, and Luca Sani. 2014. A novel methodology to address the internet as-level data incompleteness. *IEEE/ACM Transactions on Networking* 23, 4 (2014), 1314–1327.

[30] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328.

[31] J. Heitz, J. Snijders, K. Patel, I. Bagdonas, and N. Hilliard. 2017. *BGP Large Communities Attribute*. RFC 8092. RFC Editor.

[32] Bradley Huffaker, Amogh Dhamdhere, Marina Fomenkov, et al. 2010. Toward topology dualism: improving the accuracy of as annotations for routers. In *International Conference on Passive and Active Network Measurement*. Springer, 101–110.

[33] B. Huffaker, K. Keys, M. Fomenkov, and K. Claffy. [n. d.]. AS-to-Organization Dataset. http://www.caida.org/research/topology/. ([n. d.]). data range used: 20180401.

[34] IANA. 2021. Autonomous System (AS) Numbers. Available at https://www.iana.org/assignments/as-numbers/as-numbers.xhtml. (2021). last-accessed: Tuesday, 22nd April 2021.

[35] Yunchen Jin. 2019. ProbLink. GitHub repository: https://github.com/YuchenJin/ProbLink. (2019).

[36] Yuchen Jin, Colin Scott, Amogh Dhamdhere, Vasileios Giotsas, Arvind Krishnamurthy, and Scott Shenker. 2019. Stable and Practical {AS} Relationship Inference with ProbLink. In *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)*. 581–598.

[37] Ziting Jin. 2020. TopoScope. GitHub repository: https://github.com/Zitong-Jin/TopoScope. (2020).

[38] Zitong Jin, Xingang Shi, Yan Yang, Xia Yin, Zhiliang Wang, and Jianping Wu. 2020. TopoScope: Recover AS Relationships From Fragmentary Observations. In *Proceedings of the ACM Internet Measurement Conference*. 266–280.

[39] T. King, C. Dietzel, J. Snijders, G. Doering, and G. Hankins. 2016. *BLACKHOLE Community*. RFC 7999. RFC Editor.

[40] Thomas Krenc and Anja Feldmann. 2016. BGP prefix delegations: a deep dive. In *Proceedings of the 2016 Internet Measurement Conference*. 469–475.

[41] LACNIC. 2021. Extended Delegations, 20180405. https://ftp.lacnic.net/pub/stats/lacnic/delegated-lacnic-extended-20180405. (2021). Last accessed: 24th April, 2021.

[42] Aemen Lodhi, Natalie Larson, Amogh Dhamdhere, Constantine Dovrolis, and Kc Claffy. 2014. Using peeringDB to understand the peering ecosystem. *ACM SIGCOMM Computer Communication Review* 44, 2 (2014), 20–27.

[43] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and KC Claffy. 2013. AS relationships, customer cones, and validation. In *Proceedings of the 2013 conference on Internet measurement conference*. 243–256.

[44] MANRS. 2021. About MANRS. Available at https://www.manrs.org/about/. (2021). last-accessed: Tuesday, 20th April 2021.

[45] Pedro Marcos, Lars Prehn, Lucas Leal, Alberto Dainotti, Anja Feldmann, and Marinho Barcellos. 2020. AS-Path Prepending: there is no rose without a thorn. In *Proceedings of the ACM Internet Measurement Conference*. 506–520.

[46] Alexander Marder, Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, KC Claffy, and Jonathan M Smith. 2018. Pushing the boundaries with bdrmapit: Mapping router ownership at Internet scale. In *Proceedings of the Internet Measurement Conference 2018*. 56–69.

[47] Tyler McDaniel, Jared M Smith, and Max Schuchard. 2020. Flexsealing BGP against route leaks: peerlock active measurement and analysis. *arXiv preprint arXiv:2006.06576* (2020).

[48] Tyler McDaniel, Jared M Smith, and Max Schuchard. 2021. Peerlock: Flexsealing BGP. NDSS.

[49] Reynaldo Morillo, Justin Furuness, Amir Herzberg, Cameron Morris, Bing Wang, and James Breslin. 2021. ROV++: Improved deployable defense against BGP hijacking. NDSS.

[50] Lucas Müller, Matthew Luckie, Bradley Huffaker, Kc Claffy, and Marinho Barcellos. 2019. Challenges in inferring spoofed traffic at IXPs. In *Proceedings of the 15th International Conference on Emerging Networking Experiments and Technologies*. 96–109.

[51] NANOG. 2021. Future NANOG Meetings. Available at https://www.nanog.org/meetings/future/. (2021). last-accessed: Tuesday, 20th April 2021.

[52] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. 2009. The (in) completeness of the observed Internet AS-level structure. *IEEE/ACM Transactions on Networking* 18, 1 (2009), 109–122.

[53] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy* (first edition ed.). Crown, New York.

[54] Alice B Popejoy and Stephanie M Fullerton. 2016. Genomics is failing on diversity. *Nature News* 538, 7624 (2016), 161.

[55] Lars Prehn, Franziska Lichtblau, and Anja Feldmann. 2020. When wells run dry: the 2020 IPv4 address market. In *Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies*. 46–54.

[56] RADB. 2021. Query for AS3356. Available at https://www.radb.net/query?keywords=AS3356. (2021). last-accessed: Tuesday, 22nd April 2021, Internet Archive snapshot: https://web.archive.org/web/20210422074619/https://www.radb.net/query?keywords=AS3356.

[57] RIPE. 2021. Meetings and Events. Available at https://www.ripe.net/participate/meetings. (2021). last-accessed: Tuesday, 20th April 2021.

[58] RIPE NCC. 2021. Extended Delegations, 20180405. https://ftp.apnic.net/stats/ripe-ncc/2018/delegated\T1\textemdashripencc-extended-20180405.bz2. (2021). Last accessed: 24th April, 2021.

[59] Pavlos Sermpezis, Vasileios Kotronis, Petros Gigis, Xenofontas Dimitropoulos, Danilo Cicalese, Alistair King, and Alberto Dainotti. 2018. ARTEMIS: Neutralizing

BGP hijacking within a minute. *IEEE/ACM Transactions on Networking* 26, 6 (2018), 2471–2486.

[60] Lakshminarayanan Subramanian, Sharad Agarwal, Jennifer Rexford, and Randy H Katz. 2002. Characterizing the Internet hierarchy from multiple vantage points. In *Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, Vol. 2. IEEE, 618–627.

[61] Cecilia Testart, Philipp Richter, Alistair King, Alberto Dainotti, and David Clark. 2019. Profiling BGP serial hijackers: capturing persistent misbehavior in the global routing table. In *Proceedings of the Internet Measurement Conference*. 420–434.

[62] Martino Trevisan, Danilo Giordano, Idilio Drago, Maurizio Matteo Munafò, and Marco Mellia. 2020. Five years at the edge: Watching internet from the isp network. *IEEE/ACM Transactions on Networking* 28, 2 (2020), 561–574.

[63] Wikipedia. 2021. List of Tier 1 networks. (2021). Available at https://en.wikipedia.org/wiki/Tier_1_network Last accessed: March 30th, 2021.

[64] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. 2018. Leveraging interconnections for performance: the serving infrastructure of a large CDN. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 206–220.

A DOES PERFORMANCE CORRELATE WITH VALIDATION COVERAGE?

Some of the link classes for which the inference algorithms perform poorly have a higher validation coverage. In this section, we show that there is no correlation between these two metrics. We set up the following experiment: We uniformly sample a subset of relationships and track their evaluation performance using the metrics discussed in §6. We vary the subset size between 50 % and 99 % of the original set size by increments of 1 %. To get a more stochastically robust result, we repeat this process 100 times for each sample size. While we have done this analysis for all link classes mentioned in §5, we now discuss the results for the $T1 - TR$ class as it produced low-performance results while containing more than 600 peering relationships. Figure 4, 5, and 6 show the sample size on the x-axis against the precision (PPV_P), recall (TPR_P), and MCC on the y-axis. While the figures mark the individual performance measurements for each sampled set with a cross, they also show the interquartile-range (IQR) and median across all 100 sampled sets. Even though we observe that the variance increases with decreasing sample size, we neither observe an increasing nor a decreasing trend for the performance metrics. Notably, the other link classes (not shown) allow for similar conclusions.

B PLOTS FOR ALTERNATIVE METRICS

Figures 7, 8, and 9 show alternate variants of figure 3 for the customer cone size (CCS), CCS when ignoring links incident to route collector peers (i.e., vantage point ASes), and the node degree.

C POTENTIAL FEATURES

The following per-link metrics might help to identify additional groups of 'hard links':

(1) visibility over time
(2) number of prefixes redistributed via link
(3) number of addresses covered by those prefixes
(4) number of prefixes *originated* through the link
(5) number of addresses covered by those prefixes
(6) number of ASes that can observe (i.e., occur left from) the link
(7) number of ASes that might receive traffic via (i.e., occur right from) the link
(8) the relative difference in transit degree between the incident ASes
(9) the relative difference in PPDC size between the incident ASes
(10) and the number of common IXPs where both incident ASes are present
(11) and the number of common peering facilities where both incident ASes are present
(12) how the incident ASes behave, e.g., BGP serial hijackers [61] vs MANRS participants [44]

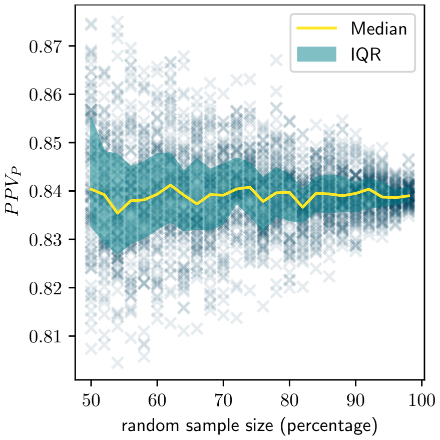

Figure 4: Correlation Analysis: Precision (P2P) for randomly drawn subsets of $T1 - TR$ links with increasing size.

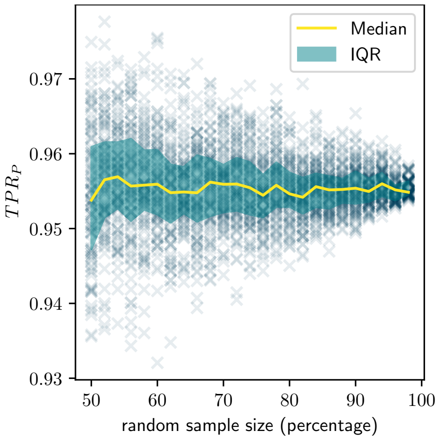

Figure 5: Correlation Analysis: Recall (P2P) for randomly drawn subsets of $T1 - TR$ links with increasing size.

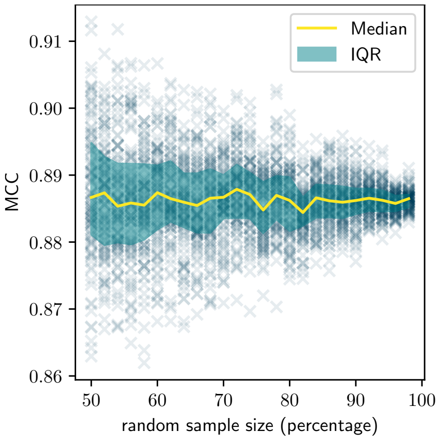

Figure 6: Correlation Analysis: MCC for randomly drawn subsets of $T1 - TR$ links with increasing size.

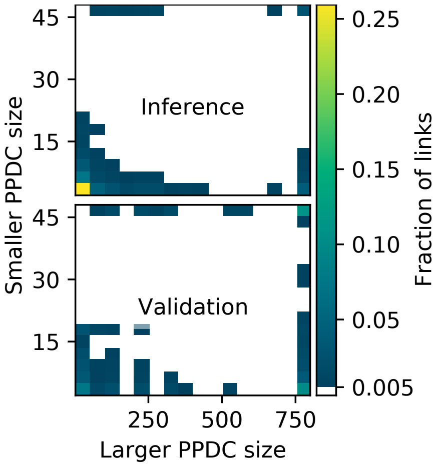

Figure 7: Customer Cone Imbalance for transit links

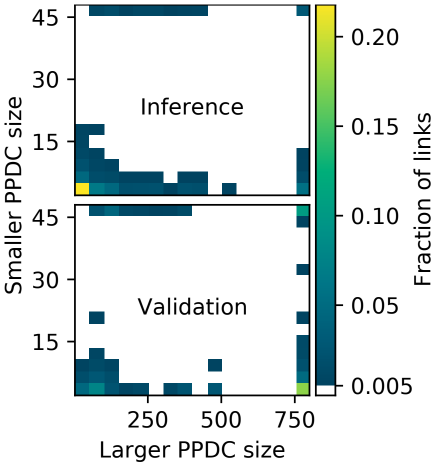

Figure 8: Customer Cone Imbalance for transit links (ignoring links with incident Route Collector Peers)

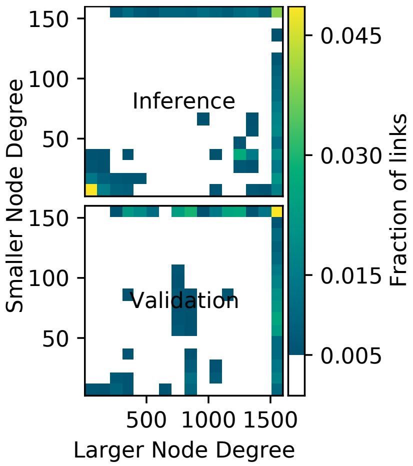

Figure 9: Node degree Imbalance for transit links