Learning AS-to-Organization Mappings with Borges

Carlos Selmo
cselmo@itba.edu.ar
Instituto Tecnologico de Buenos Aires
Buenos Aires, Argentina

Esteban Carisimo
esteban.carisimo@northwestern.edu
Northwestern University
Evanston, IL, USA

Fabián E. Bustamante
fabianb@northwestern.edu
Northwestern University
Evanston, IL, USA

J. Ignacio Alvarez-Hamelin*
ihameli@fi.uba.ar
Universidad de Buenos Aires, Facultad de Ingeniería
Buenos Aires, Argentina

Abstract

We introduce Borges (Better ORGanizations Entities mappingS), a novel framework for improving AS-to-Organization mappings using Large Language Models (LLMs). Existing approaches, such as AS2Org and its extensions, rely on static WHOIS data and rule-based extraction from PeeringDB records, limiting their ability to capture complex, dynamic organizational structures. Borges overcomes these limitations by combining traditional sources with few-shot LLM prompting to extract sibling relationships from free-text fields in PeeringDB, and by introducing website-based inference using redirect chains, domain similarity, and favicon analysis. Our evaluation shows that Borges outperforms prior methods, achieving a 7% improvement in sibling ASN identification and an Organization Factor score of 0.3576. It also expands the recognized user base of large Internet conglomerates by 192 million users (≈ 5% of the global Internet population) and improves geographic footprint estimates across multiple regions.

CCS Concepts

• **Networks** → **Topology analysis and generation**.

Keywords

AS-to-Organization mappings

ACM Reference Format:
Carlos Selmo, Esteban Carisimo, Fabián E. Bustamante, and J. Ignacio Alvarez-Hamelin. 2025. Learning AS-to-Organization Mappings with Borges. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3730567.3732918

> In that Empire, the Cartographers Guilds made a Map of the Empire whose size was that of the Empire, and which coincided point for point with it
>
> *Jorge Luis Borges. On Rigor in Science. (1946)*

1 Introduction

The Internet is a vast network of interconnected Autonomous Systems (ASes) that various organizations around the world operate. Understanding the mapping between these organizations and their ASes is essential for network management, security, and policy enforcement.

Many studies have analyzed Internet topology through Autonomous Systems (ASes) and their relationships [20, 28, 34], using heuristics to infer these connections from public BGP data sources such as RouteViews [2] and RIPE RIS [1]. While AS-level research has advanced our understanding, this view is largely shaped by BGP data resolution. Shifting focus to the organizations managing these ASes can provide additional insights into peering disputes [19], legal actions [49], geopolitical influences, market concentrations [26, 29, 30, 32, 38], mergers, and address transfers.

Despite the value of an organization-level understanding of the network, a major challenge has been whether heuristics and data sources can fully capture the extent of organizations operating networks on the Internet.

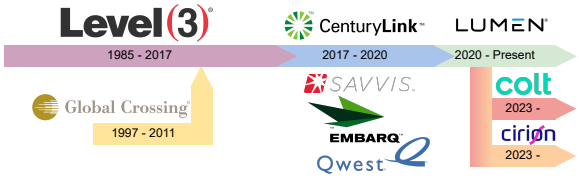


Figure 1: Timeline of Level3's mergers, demergers, acquisitions, and rebrandings.

The organizational structure of the Internet is constantly changing with frequent mergers and acquisitions – Arelion and Telia [3, 4], Sprint and T-Mobile USA [17], Vodafone and Cable & Wireless [48], to name a few. Among these, the history of Level3 Communications stands out as a compelling example of the speed and complexity of these shifts, as illustrated in Figure 1. In 2011, Level3 acquired Global Crossing, its closest competitor [43]. Just five years later, it was itself acquired by CenturyLink [14] – already in the midst of consolidating Qwest [12], Savvis [13], and Embarq [11] – before rebranding as Lumen [47] and subsequently spinning off regional assets to Cirion [35] and Colt [18]. These dynamic, multi-step changes highlight the need for a solution that is adaptive, context-aware, and capable of integrating multiple data sources to reflect real-world organizational relationships as they evolve.

*Also with CONICET — Universidad de Buenos Aires, INTECIN.

IMC '25, Madison, WI, USA

ACM ISBN 979-8-4007-1860-1/2025/10
https://doi.org/10.1145/3730567.3732918

The long-standing AS2Org method [8], introduced by Cai et al. [49] in 2010, relies on WHOIS records to create relationships between organizations and networks. More recent efforts [5, 15] improve on this by incorporating PeeringDB metadata and using heuristics to infer sibling ASes based on text fields like notes and "aka". However, both approaches depend on static identifiers and rule-based extraction methods that struggle with unstructured, inconsistent, or multilingual data—resulting in frequent false positives, missed relationships, and limited scalability.

We argue that these challenges—unstructured data, inconsistent formats, and lack of semantic context—are better addressed with Large Language Models (LLMs). Using few-shot prompting, LLMs can extract sibling AS relationships from messy, multilingual text without brittle rules or manual curation. Their flexibility makes them well-suited for mapping organizational structures in an Internet shaped by constant mergers, rebrandings, and regional variation [7, 16].

In this work, we introduce Borges (**B**etter **ORG**anizations **E**ntities mapping**S**) (§3) a new LLM-based approach to AS2Org.

Building on recent AS2Org systems that combine traditional WHOIS-based methods with PeeringDB data, Borges uses PeeringDB's Organizational ID (§4.1) and leverages LLMs to extract sibling relationships from embedded text fields via few-shot information extraction prompts [7] (§4.2). This eliminates the need for manual intervention required in prior work [5]. By relying on LLMs, Borges also extends AS2Org datasets to incorporate companies—websites as an additional signal for sibling inference: (1) identifying networks whose PeeringDB website fields resolve to the same final URL (§4.3.2), and (2) grouping networks whose websites share favicons and similar domain names (§4.3.3).

Our evaluation shows that Borges achieves high accuracy in extracting sibling information from the PeeringDB "notes" and "aka" fields (accuracy: 0.947), and in identifying when domain and favicon similarities indicate shared organizational control (accuracy: 0.986). By leveraging website data, Borges successfully maps Limelight Networks (LLNW-AS22822) and Edgecast (AS15133) under the same organization (both redirecting to www.edg.io); associates Sprint—after a series of redirects - with Cogent [24] (§4.3.2); and links Claro Chile and Claro Puerto Rico, which share a favicon but differ slightly in domain name (www.clarochile.cl and www.claropr.com).

To quantify how well an AS2Org approach captures the Internet's organizational structure, we introduce a new metric: the *Organization Factor* (§5.4). This metric ranges from 0—where each organization manages a single network—to 1, where all networks are grouped under a single organization. Using *Organization Factor*, we compare Borges, AS2Org, and *as2org+*, and find that Borges achieves a score of 0.3576, outperforming AS2Org and *as2org+* by 7% and 3.3%, respectively (§5.4). While no ground truth exists for organizational mappings, Borges's improvements in both the *Organization Factor* metric and accuracy make a strong case for the LLM-based approach.

We present Borges, a new system that integrates learning-based extraction and website inference to improve AS-to-Organization mappings. This paper makes the following contributions:

- **A learning-based extraction method** that uses few-shot prompting with Large Language Models to identify sibling ASNs from unstructured PeeringDB text fields (e.g., notes, aka), avoiding the limitations of regex-based approaches and manual validation.
- **A novel website-based inference module** that detects organizational relationships through redirect chains, domain similarity, and favicon analysis - enabling the discovery of sibling ASNs missed by existing methods.
- **A comprehensive evaluation** demonstrating that Borges outperforms AS2Org and AS2Org+, achieving a 7% improvement in sibling inference, expanding recognized organizational clusters by 192 million users (≈5% of the global Internet population), and increasing geographic coverage.
- **A new metric, the Organization Factor**, for quantifying how well an AS-to-Organization mapping reflects real-world organizational structure; Borges achieves the highest score to date (0.3576).
- **An open-source framework** by releasing the complete codebase[1] together with all prompts, enabling full reproducibility of our results and allowing the community to generate new mappings and improve the framework.

In the following sections, we detail the design of Borges, evaluate its effectiveness against existing methods, and discuss its broader implications for understanding the organizational structure of the Internet.

2 Background

In this section we present limitations of the traditional and the latest AS-to-Organization mapping techiques (§2.1) and discuss the opportunities of including both Large Language Models (LLMs) and networks' website information into the process (§2.2).

2.1 The State of the Art

Capturing the organization-level structure of the Internet has inspired several studies. Cai et al. [49] seminal work introduced the widely adopted AS2Org method. In recent efforts, some studies proposed to incorporate PeeringDB, given its value in related research problems [6, 22, 23, 33, 36, 50], as a valuable resource for obtaining more complete representations of the AS-to-Organization mappings. Next, we discuss the contributions of these studies and potential directions to continue developing more comprehensive methods.

AS2Org [49], long regarded as the standard for AS2Org mappings, uses organizational identifiers (Org IDs) from RIR allocation databases to group ASNs under the same entity. While these databases include other fields that could help infer sibling relationships, their limitations—such as outdated records and incomplete organization data—have discouraged broader use [49]. Despite AS2Org's broad coverage, it often fails to capture the full scope of an organization's network holdings.

Arturi et al. [5] proposed an enhanced AS-to-Organization mapping approach that incorporates additional information from PeeringDB. Their system, *as2org+*, extends AS2Org by extracting sibling relationships from unstructured text fields—such as `notes` and `aka`—using regular expressions. While this expands the available data, the reliance on simple regexes limits semantic understanding. As a result, *as2org+* frequently misclassifies numerical values (e.g., phone

[1] Our codebase is available at: https://github.com/NU-AquaLab/borges

numbers, years, addresses) as ASNs, leading to false positives. To mitigate this, *as2org+* combines filters with manual inspection, a process that is both labor-intensive and prone to inadvertently discarding correct inferences.

Chen et al. [15] followed a complementary path, identifying mismatches between CAIDA's AS2Org dataset and PeeringDB's records. Their method flags these discrepancies as candidates for reclassification and uses keyword matching along with semi-manual inspection to refine mappings.

Despite these advances, current techniques remain constrained by static heuristics and limited data sources. We identify two underexplored directions to improve AS-to-Organization mapping at scale: (1) Applying LLM-based, prompt-guided methods for more robust and semantically aware data extraction; and (2) Leveraging networks' websites as an additional source of evidence for inferring sibling relationships.

2.2 Completing AS2Org Mappings with Website Information

Using network websites as a signal for inferring shared organizational control remains largely unexplored in AS2Org mapping. Yet most networks—particularly those operated by commercial providers—maintain customer- or operator-facing websites. Large conglomerates often deploy standardized or "canned" web templates across subsidiaries to present a unified brand and reduce development costs. We posit that these visual and structural similarities can be leveraged to enrich AS2Org mappings.

A naive approach might rely solely on domain name similarity to identify sibling networks. While effective in some cases, this misses more complex organizational structures where branding and naming conventions vary across regions. For example, Telefonica operates under multiple names – Movistar, Telxius, O2—depending on the market. Orange, a French conglomerate, runs its transit division as Open Transit (AS5511). Even T-Mobile, despite consistent global branding, uses unrelated domains like http://www.telekom.sk for Slovak Telekom (AS6855) and http://www.t.ht.hr for Hrvatski Telekom (AS5391).

To address these complexities, we propose a more comprehensive sibling inference approach that goes beyond domain names. Specifically, we incorporate additional website attributes—such as shared logos, favicons, and banners—that are often replicated across networks under the same organizational umbrella.

3 Borges: System overview

In this section, we present Borges, a framework that combines AS2Org and PeeringDB data with three complementary techniques to infer sibling AS relationships. As shown in Figure 2, Borges is composed of three modules:

- **Organization Keys:** Leverages organizational identifiers from WHOIS and PeeringDB to cluster ASes under the same organization.
- **Named-Entity Recognition (NER)**: Uses LLM-based prompts to extract sibling relationships from unstructured text fields.
- **Web-based Inference:** Identifies sibling ASes through website similarity, including shared domains, redirects, and favicons.

Organization Keys: Borges uses organization identifiers from WHOIS and PeeringDB to group ASNs under the same entity. These datasets reflect different aspects of Internet operations—WHOIS captures allocation and delegation, while PeeringDB is operator-driven and often more current. By combining both, Borges provides a more complete view of organizational structure than either source alone.

Named-Entity Recognition: To extract sibling relationships from PeeringDB's unstructured fields (e.g., notes, aka), Borges applies few-shot prompting with Large Language Models (LLMs). This approach replaces the brittle, regex-based methods used in *as2org+* [5], which often required manual curation. In contrast, our LLM-based method offers greater flexibility and accuracy with minimal human intervention.

Web-based Inference: Borges introduces websites as a new signal for identifying sibling ASes. Using self-reported URLs in PeeringDB, we (1) identify networks pointing to the same final destination (directly or through redirects), and (2) infer sibling relationships from websites with similar characteristics—such as consistent branding in domain names or identical favicons. This technique captures relationships missed by registry data alone, particularly in cases where branding diverges across regions or subsidiaries.

4 Borges in detail

In this section, we provide a detailed explanation of the three main building blocks of Borges: (1) key-based clustering (§4.1), (2) the LLM-based Named-Entity Recognition module (§4.2), and (3) web-based sibling inference (§4.3).

4.1 Leveraging Entity-Relation Models

We leverage organizational identifiers (Org IDs) offered by both WHOIS and PeeringDB (PDB) data schemas. Both WHOIS and PDB describe allocations and operations with data objects for both ASes and Organizations, which are linked via a one-to-many relationship.

While the WHOIS Org ID (OID_W) may be imperfect to fully capture the AS-to-Organization mappings, this is still a valuable source of information. As discussed in §2, some organizations do not consolidate all their resources under a single entity, resulting in a partial representation of the overall organization. However, CAIDA's AS2Org, the most widely adopted source of AS-to-Organization mappings, still uses OID_W to generate most of the AS-to-Organization mappings. Despite its limitations, OID_W provides an AS-to-Organization mapping for all allocated networks, as each ASN must be assigned to an organization when allocated. Rather than using solely a single source, Borges incorporates OID_W as one of the sources to identify ASes under the same management.

In a similar way, PeeringDB replicates the WHOIS data structuring with networks and organizations' data objects linked by a relationship. We leverage this relationship, the PeeringDB Org ID (OID_P), to identify all ASes registered under the same organization. Unlike the WHOIS data structure, which is confined to legal and contractual boundaries (*e.g.,* treating subsidiaries as separate entities), PeeringDB provides an additional perspective from a network operations standpoint, allowing us to group resources under a common organization.

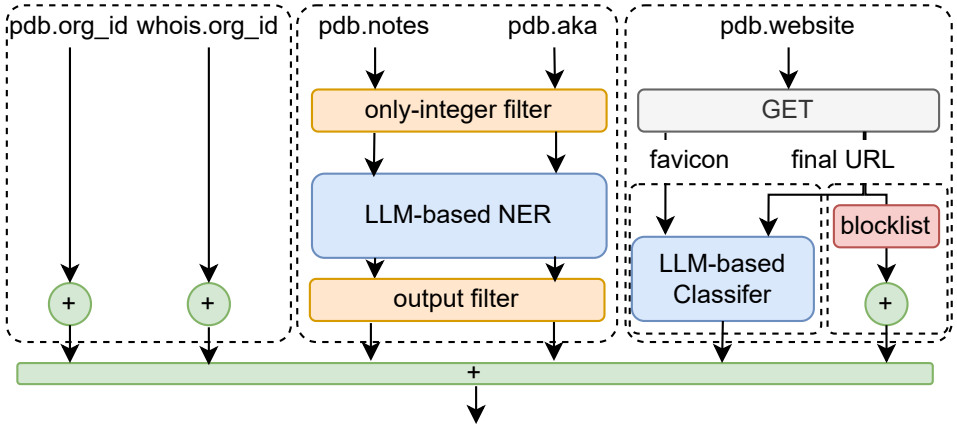


Figure 2: Diagram of the building block of Borges composed of three distinct modules for sibling inferences: (1) ***Key-based clustering*** **that utilizes organizational identifiers from AS2Org and PeeringDB datasets,** (2) ***Named-Entity Recognition*** **that uses LLM prompts extracts embedded information from text fields, and** (3) ***A web scraper + LLM-based Classifier*** **that generates sibling inferences based on the websites referenced in PeeringDB records.**

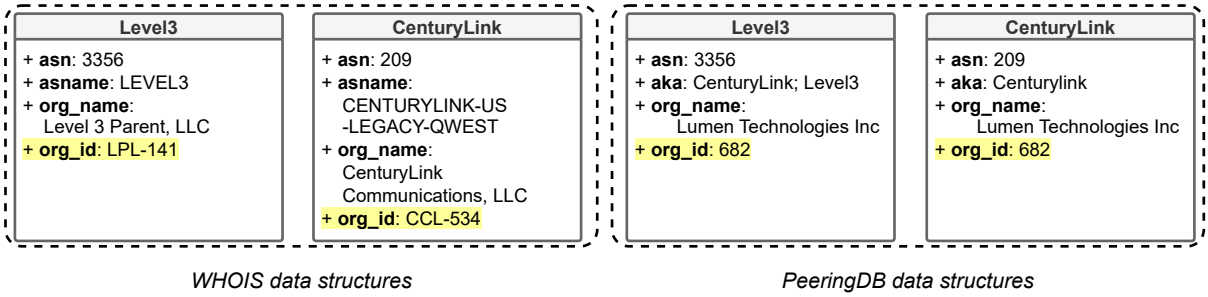


Figure 3: Example of a cluster obtained for Lumen (AS3356, formerly known as Level3) and CenturyLink (AS209), both under the same administration, when Organizational IDs from WHOIS records (left) and PeeringDB records (right) are applied. The example illustrates that WHOIS records fail to group both networks together, while PeeringDB data correctly indicates that both companies are under the same organizational umbrella.

Due to differing approaches by RIRs and PeeringDB in defining organizations, they often produce organizations with different compositions. To reconcile these discrepancies, we consolidate partially overlapping clusters into a single organization. For example, despite CenturyLink's acquisition of Level 3 nearly a decade ago, CenturyLink-AS209 and Level3-AS3356 are still assigned to separate clusters in the AS2Org datasets, as shown on the left in Figure 3. However, the right side of Figure 3 demonstrates that CenturyLink-AS209 and Level3-AS3356 are grouped under the same organization in PeeringDB, emphasizing the benefit of combining organizational IDs from both sources.

4.2 `notes` and `aka`: Leveraging Self-Reported Information

Borges builds upon the *as2org+* [5] framework by integrating learning-based methods to enhance and streamline inference processes. Operators often utilize text fields like `notes` and `aka` in the PDB data schema to share valuable information about *siblings* . As detailed in §2.1, *as2org+* employs a simple rule-based approach to extract and organize this sibling information. While straightforward, this method can lead to incomplete inferences and false positives, requiring expert human review to verify outputs and filter out inaccuracies. In contrast, Borges simplifies the extraction of *siblings* from text fields by applying Information Extraction (IE) techniques with modern Large Language Models, reducing false positives and eliminating the need for human intervention. In the next paragraphs, we detail all the steps involved in extracting sibling information from these unstructured text fields.

Input Filter: Borges begins the Information Extraction (IE) process by applying a dropout filter to enhance model accuracy by only considering text fields – either `notes` or `aka` entries – containing numbers on them. Although these fields are often used to share sibling information, this is not their most common use. Therefore, entries without numbers – and thus without potential Autonomous System Number (ASN) information – are removed from consideration.

Information extraction with Large-Large Models: Extracting Autonomous System Numbers (ASNs) from unstructured text fields like `notes` and `aka` falls within the scope of Named Entity Recognition (NER) in Natural Language Processing (NLP). NER techniques allow us to identify and extract ASNs, converting unstructured text into structured data through Information Extraction (IE). LLMs enhance this process by overcoming language barriers [25, 31], enabling us to capture ASNs independently of the language of the surrounding text.

Advancements in Few-Shot and Zero-Shot Learning with LLMs [7] highlight their potential for our problem domain. Task-agnostic pretraining in NLP reduces the need for task-specific fine-tuning, and zero-, one-, and few-shot settings can sometimes surpass state-of-the-art fine-tuned models [7]. This opens opportunities for enhancing IE, especially where manual annotation could be more practical due to extensive human effort and performance degradation with new annotation schemas. Zero-Shot IE systems leverage LLMs' inherent pre-trained knowledge for annotations [41], reducing the need for manual data labeling. We leverage these capabilities to extract ASNs embedded in natural text.

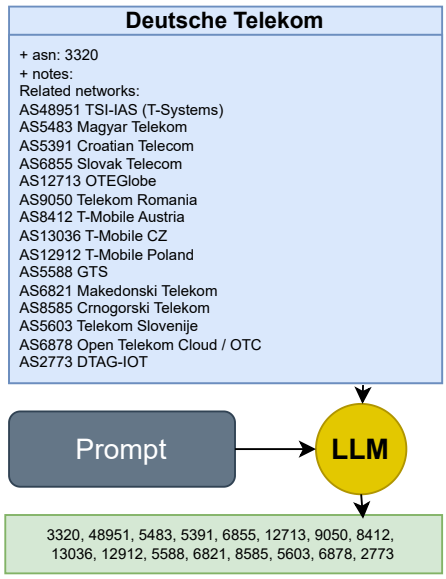


Figure 4: Example of the prompt-guided Information Extraction process from PeeringDB notes using LLMs. In this example, Deutsche Telekom (AS3320) reports its European subsidiaries in unstructured text, which can be successfully identified with Borges LLM-based approach.

We implemented this approach in Borges, utilizing OpenAI's GPT-4o-mini [40] with a `temperature` set to 0 and a `Top P` probability mass of 1. This setup ensures the model consistently produces the most probable next token, resulting in reproducible outputs unless the model weights are updated. Our prompt (fully detailed in Listing 2 in Appendix C) instructs the model to extract all sibling information embedded in the `notes` and `aka` fields, and to disregard all unrelated ASNs, such as those reported to be upstream providers, peers, or part of BGP communities. These restrictions on the ASNs to be extracted are particularly useful when dealing with networks such as Lattitude.sh-AS262287[2], which report their upstream connectivity. This represents a significant departure from *as2org+*, as its use of regular expressions required human inspection and a complex customer-to-provider relationship filter to exclude these cases. Figure 4 illustrates our implementation, highlighting Deutsche Telekom-AS3320's notes containing sibling information.

Output Filter To prevent hallucinations, we limit the output to only those number sequences that appear in the `notes` or `aka` fields.

4.3 The Web as a Source of Sibling Inferences

Borges advances the state of the art by incorporating networks' websites as a valuable resource for identifying networks operating under the same corporate structure. By leveraging the website field in the PDB data schema, we introduce a novel dimension for this identification. PeeringDB's authentication methods ensure that records are completed by actual network operators, enhancing the credibility and reliability of the provided websites despite occasional errors.

Our goal is to identify commonalities among websites associated with the same corporate entity, potentially indicating that the Autonomous Systems (ASes) linked to these websites share common administration. We assume that Internet providers under the same corporate umbrella tend to have similar websites.

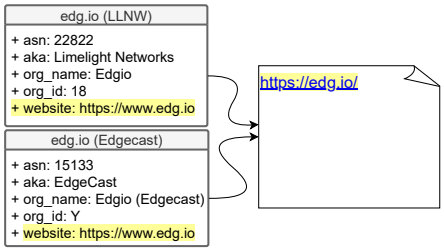
Given the complexity and richness of information available on network websites, our web module comprises three core components: (1) **Web Scraper**: Interacts with networks' reported websites to collect specific features (§4.3.1), (2) **Final URL Matching Module**: Identifies networks' websites that refer, directly or indirectly, to the same URL (§4.3.2), and (3) **LLM-Based Classification**: Detects networks reporting websites with similar domains and brand names, including those with identical favicons (§4.3.3).

4.3.1 Scraping the Web: Identifying Final URLs and Networks' Favicons. Borges utilizes `selenium` [44] to automate interactions with websites referenced in PDB records. By employing Selenium's headless browser functionality [37], we can fully render websites – including the execution of JavaScript – as if they were displayed in a regular browser. This live interaction is crucial for loading dynamic content and handling "refreshes and redirects" (R&R) encountered during scraping. As a result, Borges collects the final URLs of the websites referenced in PDB, enabling the detection of relationships across networks that are not visible in PDB records but become apparent after loading these pages.

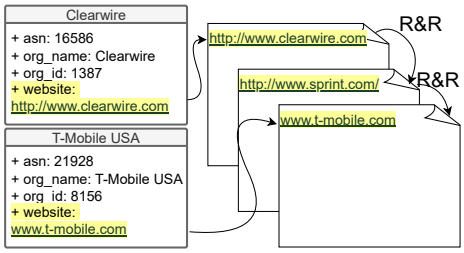
[2] An example of Lattitude.sh's entry is shown in Appendix B

Building on the list of final URLs, Borges collects favicons based on the hypothesis that networks under the same administration are likely to use identical or similar brand icons displayed as website favicons. To download and create a dataset of these icons, Borges utilizes Google's Favicon API[3].

4.3.2 Final URL Matching Module. Borges uses perfect URL matching from direct and indirect references as a first method for inferring sibling relationships based on website URLs. This approach leverages the information available in the field website of the PeeringDB data schema to identify networks registered under different organizations on PDB (i.e., networks with different OID_P) but being under the same company structure. In the following paragraphs, we explain both scenarios and detail how Borges effectively identifies these cases.



(a) Same website



(b) Refresh and Redirect (R&R)

Figure 5: Examples of PDB records representing networks under different organizations (each having different organization IDs), yet either directly (as depicted in Fig. 5a) or indirectly (following refreshes and redirects, as seen in Fig. 5b), lead to the same website.

The simplest scenario for inferring sibling relationships through the website field is when different organizations within the PDB register two networks that share the same website. An example is the recent merger of LimeLight Networks (LLNW, AS22822) and Edgcast (AS15133) into a single company named `edg.io` [21]. Despite the updates on PDB records of both networks, shown in Figure 5a, showing their new branding and website, they continue to fall under different PDB organizational umbrellas. By leveraging the information found in the website field, we can discover this sibling relationship.

Another scenario is when networks reference different websites within the PDB (and different organization too), but both ultimately lead to the same website through redirects. This situation becomes visible only when users interact with these websites, our in our dataset through scraping. Clear Wire (AS16586) is a prominent example of this as Sprint first acquired this company in 2012 [45], which T-Mobile subsequently acquired in 2020 [46]. As depicted in Figure 5b, in a 2021 PDB snapshot, the record for Clear Wire directed to `www.clearwire.com`, despite this website redirecting to Sprint (`www.sprint.com`), which in turn redirects to T-Mobile. More recently, in 2023, T-Mobile sold out the former Sprint fiber backbone to Cogent [24], again highlighting the dynamicity of mergers and acquisitions across Internet providers.

By utilizing the list of final URLs collected during the scraping process, Borges captures both organizations that report the same website and those that indirectly lead to the same final URL. As a final step, Borges applies a manually curated blocklist of domains that are typically included in PeeringDB records but are provided by unrelated companies. This situation commonly occurs with small companies that do not have their own websites and instead use mainstream communication channels (Facebook, LinkedIn, GitHub, Discord, etc.) to interact with their users and other operators. Given that there are only a handful of such examples, we manually curated this list, which is fully described in Appendix D.1.

4.3.3 LLM-Based Classification. Borges further extends its website-based inference by analyzing favicons and final URL similarity. We assume that companies under the same organization will use the same brand logos – displayed as favicons on their websites – and domain names that are variations of the parent company's name. Figure 6 provides an overview of the decision tree applied to determine that URLs associated to the same favicon actually belong to the same company.

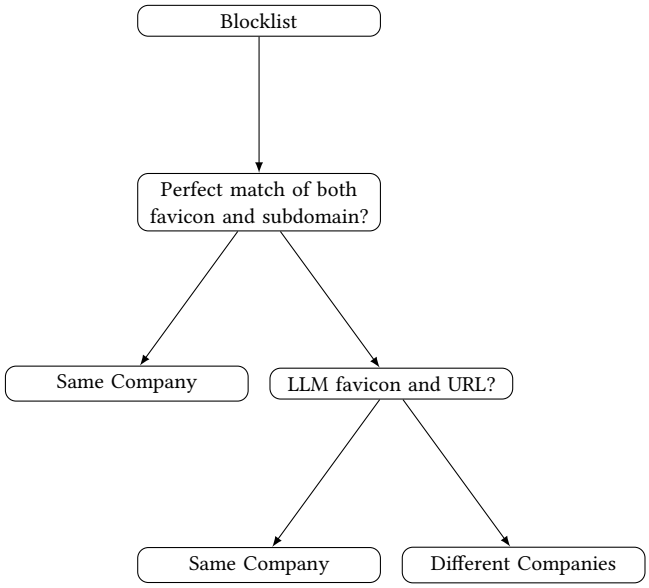


Figure 6: Decision Tree for Company Classification

As a first step, Borges, with the same criteria applied to create the blocklist from §4.3.2, applies a blocklist (see the full list in Appendix D.2) to exclude networks reporting mainstream communication

[3]Example of use of Google's Favicon API for Orange France: https://t3.gstatic.com/faviconV2?client=SOCIAL&type=FAVICON&fallback_opts=TYPE,SIZE,URL&url=https://www.orange.fr&size=16

Company	Logo	URLs
Digicel		https://www.**digicelgroup**.com/kn/en https://www.**digicelgroup**.com/cw/en https://www.**digicelgroup**.com/bb/en

Table 1: Digicel is a large operator in the Caribbean, whose subsidiaries' websites share both the favicon and subdomains.

Company	Logo	URLs
Claro		https://www.clarochile.cl/personas/ https://www.claro.com.do/personas/ https://www.claro.com.pe/personas/ https://www.claropr.com/personas/
Bootstrap		https://www.anosbd.com/ https://www.rptechzone.in/ https://bapenda.riau.go.id/ http://www.conexaointernet.com.br/ https://www.ramdiaonlinebd.com/

Table 2: Examples of domains sharing the same favicon. The first corresponds to Claro, a large operator with a presence in Latin America and the Caribbean, while the second is the default favicon of the Bootstrap web framework. Domain names help differentiate companies' brand logos from web frameworks.

platforms (e.g., Facebook, LinkedIn, GitHub, Discord) in the PDB website field, which creates links between unrelated companies.

After excluding these unrelated domains from consideration, Borges following decision rule involves grouping together all ASNs from entries that lead to the exact same favicon and share the same subdomain (e.g., www.**orange**.es and www.**orange**.pl). This rule assumes that all ASNs belong to the same company structure when they share identical subdomains and favicons, as illustrated in Table 1 for the case of Digicel, a conglomerate operating in the Caribbean.

To handle more complex cases, our next step reclassifies groups not previously considered under the same company by grouping all final URLs that display the same favicon. Table 2 provides two contrasting examples. The first is Claro, a cellular carrier operating in Latin America, whose local branches have slightly different final URLs and domain names but exhibit the same company logo in their favicons. However, as the second example illustrates, when websites use default favicons provided by web technologies like Bootstrap, WordPress, GoDaddy, or IXC Soft – a popular website developer for Brazilian networks – this can inadvertently group together unrelated final URLs, as they all share the same default favicon.

To address this ambiguity, Borges employs an LLM-based classifier. By inputting the favicon and the associated list of final URLs, Borges uses a prompt (described in Listing 3 in Appendix E) to query GPT-4o-mini to determine whether the information corresponds to (*a*) a specific company, (*b*) different companies.

5 Evaluating Borges

In this section, we evaluate the improvements introduced by Borges using PeeringDB and website data. We explain our experimental setup (§5.1), the results obtained from applying Borges (§5.2), validate the accuracy of both stages of the LLM (§5.3) and benchmark Borges against prior AS-to-Organization mapping techniques (§5.4).

5.1 Evaluation Setup

To evaluate Borges, we downloaded snapshots of PeeringDB and WHOIS information, as both serve as input sources. CAIDA provides an archive of both PeeringDB and AS2Org, albeit with different temporal granularities. Specifically, we used the PeeringDB snapshot from July 24, 2024, and CAIDA's AS2Org from July 1, 2024.

To complete the execution of Borges, we utilized the website information available in the PeeringDB snapshot and retrieved all favicons of these websites on July 30, 2024.

For our evaluation, we compared Borges with *as2org+* and AS2Org, whose mappings are also required. The mappings from AS2Org are publicly available and were downloaded from CAIDA. In contrast, *as2org+* does not offer any mappings and its methodology employs various regular expressions and requires substantial human intervention. For our evaluation, we built the mapping with a simple setup that uses only `pdb.org_id` (OID_P). Given that Borges works without human intervention, we also removed all manual steps from *as2org+* to compare both systems under the same fully automated conditions.

5.2 Borges's Features Contribution

Our first analysis focuses on understanding the individual contribution of each feature in identifying and retrieving Autonomous System Numbers (ASNs) and their role in forming network-based groupings. To quantify the impact of each feature, we analyze their performance in isolation before integrating them into a combined methodology. Due to the partially overlapping nature of the inferred organizations, combining features reveals their complementary effects, where each feature augments the others to create more comprehensive and larger organizational groupings within the network. Table 3 shows the individual contribution of each feature.

Source	Number of ASes	Number of Orgs
OID_P	30,955	27,712
OID_W	117,431	95,300
notes and aka	1,436	847
R&R	22,523	20,065
Favicons	1,297	319

Table 3: Summary of ASes and Organizations obtained from each of the features of Borges.

Organizational IDs. When examining both WHOIS and PeeringDB native fields in the data schema for linking networks to organizations, we find that our AS2Org snapshot contains 117,431 ASNs linked to 95,300 distinct organizations, while PeeringDB includes 30,955 ASNs mapped to 27,712 organizational IDs.

AS2Org provides an organization-level topology where organizations manage an average of 1.23 networks, with the largest organization being the US Department of Defense (DNIC-ARIN), which operates 973 networks. In contrast, organizations identified using PeeringDB manage an average of 1.12 networks, with the largest being ISC, operating 82 networks.

notes and aka. Our PeeringDB snapshot contains entries for 30,955 networks. Of these, 17,633 entries have non-empty fields, and only 2,916 entries contain numeric information in either the "aka" or "notes" fields – 1,038 in "aka" and 2,057 in "notes". From the "notes" and "aka" fields, Borges extracts 958 ASNs from 849 network entries, and when combined with the network entries' ASN, leads to a total of 1,436. Only considering these 1,436 from notes and aka, Borges obtains 847 different organizations.

Refresh and Redirects. From 30,955 network entries on PeeringDB, 26,225 contain website information, referencing 24,200 unique URLs. When accessing these websites, only 20,742 websites were available, and others involved a series of redirects, leading to a total of 20,094 unique final URLs. This web crawling enables an AS-to-organization mapping for 22,523 networks into 20,065 organizations.

Favicons. Our scraping process downloaded 14,516 unique favicons from 20,091 final URLs (3 final URLs did not lead to any favicon). From 14,516 unique favicons, 440 were shared for more than one final URL, with 1,260 unique URLs associated, where 281 of these shared favicons had the same subdomain (e.g., www.**orange**.es and www.**orange**.pl). This approach creates an AS-to-organization mapping for 1,297 networks into 319 organizations.

5.3 Validating LLM Stages

Our first step involves evaluating whether the LLM-based components of Borges –specifically, the Information Extraction (IE) and Classifier modules – are well-suited for the task of identifying sibling organizations. To assess the accuracy of our implementation, we applied human inspection to all entries containing numerical information and determined whether the outputs were correct. For the classifier, we manually inspected whether its decisions to label favicons and their associated domains as companies or web frameworks were accurate.

Information Extraction with LLMs Our evaluation of the accuracy of the Information Extraction Stage using LLM is a manual inspection of impressions across notes and aka fields of 320 entries, for which we manually extracted and structured the embedded information. In our evaluation, we define False Negatives (FN) as those ASNs included in either the notes or the aka but ignored by the LLM and not included as part of the inferred output. Our definition of False Positives (FP) considers two cases: (1) when the LLM misinterprets the presence of any other numerical expression (e.g., physical address, a phone number, or the maximum number of prefixes accepted) as an ASN, or (2) there is an actual ASN but the same organization does not manage it and instead is, for example, an upstream provider.

Metric	Value
True Positives (TP)	187
True Negatives (TN)	116
False Negatives (FN)	12
False Positives (FP)	5
Recall	0.94
Precision	0.974
Accuracy	0.947

Table 4: Accuracy, Precision and Recall of our LLM-based Information Extraction stage to recover sibling information embedded in notes and aka

Our evaluation of notes and aka's information extraction involved 320 PeeringDB records, as shown in Table 4. The results demonstrate that our LLM-based approach correctly extracted ASNs from 187 records and accurately disregarded numeric information not related to sibling ASNs in 116 records. In 12 records, our approach failed to recover ASNs embedded within these fields, and in 5 cases, it misinterpreted unrelated numbers as sibling ASNs. As a result, this stage achieves an accuracy of 0.947 with a precision and recall of 0.974 and 0.94, respectively.

We further explore some examples to familiarize ourselves with potential sources of these inaccuracies, For instance, AS7132 (AT&T) claims to belong to AS7018 (AT&T's largest network), yet the LLM does not return this relationship. Another case is AS10026 — formerly PACNET, now part of Telstra — which lists AS 2706 (HKBN) as part of its network. In this case, the LLM correctly extracts the reported data, but given that the data itself is wrong, the resulting inference is also incorrect.

Companies' Name Classification with LLMs Our evaluation of the accuracy of the Classification Stage using LLMs involves assessing whether the decision to label a tuple – composed of a favicon and a list of domains using this favicon – as either a company or a non-company entity (e.g., a framework like WordPress) is correct. In this context, False Positives (FPs) occur when a framework technology is incorrectly labeled as a company, while False Negatives (FNs) occur when a legitimate company is mistakenly labeled as a framework and thus ignored.

	Step 1	Step 2	All
True Positives (TP)	279	38	317
True Negatives (TN)	116	0	116
False Positives (FP)	1	0	1
False Negatives (FN)	43	5	5
Precision	0.996	1.0	0.997
Recall	0.8665	0.8837	0.984
Accuracy	0.9	0.8837	0.986

Table 5: Accuracy, Precision and Recall of our LLM-based Classifier in each of its two steps and as a whole.

We evaluated the LLM-based classifier by assessing the accuracy of the entire decision tree (see Fig. 6 in Sec. 4.3.3), as well as each individual step, with results presented in Table 5. Our evaluation involved manually reviewing 449 different favicons and their associated domain lists.

In the first step, the classifier considers PeeringDB entries of ASes that share the same favicon and subdomain in the final URL obtained after querying the reported website in PeeringDB. This step achieves an accuracy of 0.90, with a precision of 0.996 and a recall of 0.8665. However, due to its strict criteria, this stage produces 43 false negatives, which are intended to be reclassified in the next step.

The second decision rule relaxes the criteria by grouping together domains that share the same favicon, regardless of subdomain differences, and uses an LLM for this reclassification. This rule successfully reclassifies 38 of the 43 false negatives as true positives, reducing the number of false negatives to 5 and achieving a final accuracy of 0.8837 for this step.

Overall, the LLM-based classifier achieves an accuracy of 0.986, a recall of 0.984, and a precision of 0.997.

We compared our outputs with the annotated data and found that the method, while working as designed, misses many sibling networks. A remarkable example is DE-CIX and its subsidiaries, AQABA-IX in Jordan and Ruhr-CIX in the German region of Ruhr. Although they use the same favicon, their different domain names caused the LLM classifier to label them as unrelated companies.

5.4 Benchmarking Against AS2Org and *as2org+*

After evaluating the accuracy of our extraction methods, our next step is to compare the ability of Borges, *as2org+*, and the long-standing state-of-the-art approach AS2Org to group networks under the same organization.

We develop a metric called the Organization Factor (θ) to measure the capacity of each of these approaches to group networks under the same ownership as part of the same organization. The Organization Factor ranges between 0 and 1, corresponding to the hypothetical cases where all organizations manage a single network ($\theta = 0$) and where all networks are under the same organization ($\theta = 1$).

To compute this metric, we create a graph $G = (V, E)$ where the vertices V are all networks appearing in the WHOIS records – since this is a compulsory database for network delegations – and an edge connects two vertices v and u if both belong to the same organization. Under this definition, each organization forms a *clique* consisting of all networks inferred to be under the same organization, and there are no edges between networks of different inferred organizations.

Figure 7 illustrates how the Organization Factor (θ) is computed using two examples: (1) all organizations manage a single network, and (2) the organizations inferred by AS2Org. In this figure, the x-axis represents each organization inferred by each method, sorted in descending order based on the number of networks they contain, and the y-axis displays the cumulative sum of networks in each organization.

Since both methods include the same total number of networks (all delegated networks) but differ in the number of organizations

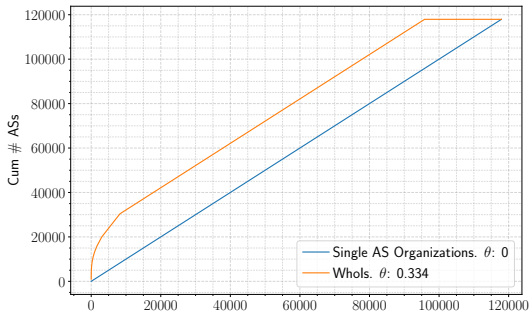


Figure 7: Illustrative example of the Organization Factor (θ) is computed using two examples: (1) all organizations manage a single network, and (2) the organizations inferred by AS2Org.

(as AS2Org infers fewer organizations because some manage multiple networks), we extend the shorter sequence with zeros to match the lengths. In this graphical representation, the Organization Factor (θ) is calculated as the normalized area under the cumulative distribution curve of network counts per organization.

Formally, let n be the total number of networks, and let k be the number of organizations inferred by a method. Let $s_1, s_2, \ldots, s_k$ be the sizes (number of networks) of these organizations, sorted in descending order. We define the cumulative sum $C_i = \sum_{j=1}^{i} s_j$ for $i = 1, \ldots, n$, where $s_j = 0$ for $j > k$ to fill the tail with zeros.

$$\theta = \frac{1}{n^2} \sum_{i=1}^{n} C_i - i = \frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} s_j \right) - i \tag{1}$$

Table 6 presents the Organization Factor (θ) for AS2Org, along with each of the four features of Borges – OID_P, notes and aka (N&A), Refresh and Redirect (R&R), and Favicons (F) – arranged across all possible combinations. Our results indicate that AS2Org achieves a θ score of 0.3343, which serves as the baseline, representing the state of the art for years. We also observe that *as2org+*, in the configuration used for our comparison (detailed in §5.1), scores 0.3467. Analyzing the contributions of Borges, we find that most features offer comparable improvements to *as2org+* relative to the baseline AS2Org. However, when all features are combined, Borges achieves a θ score of 0.3576, outperforming both AS2Org and *as2org+* by 7% and 3.3%, respectively.

To conclude this assessment, we emphasize that the Organization Factor (θ) cannot assess AS-to-Organization performance on its own; without conducting an accuracy check, as the Organization Factor (θ) does not distinguish between correct and incorrect mappings. We also recognize there is potential to apply θ beyond the AS-to-Organization context to other research on Internet structure.

6 Borges's Impact

In this section, we evaluate Borges's contribution to producing a more accurate representation of large network organizations, including access, transit, and content providers. We quantify the gains in organizational size across these categories (§6.1) and assess

AS2Org	OID_P	N&A	R&R	F	θ
✓					0.3343
✓	✓				0.3467
✓		✓			0.3386
✓			✓		0.3456
✓				✓	0.3384
✓	✓	✓			0.3503
✓	✓		✓		0.3520
✓	✓			✓	0.35
✓		✓	✓		0.3495
✓		✓		✓	0.3435
✓			✓	✓	0.349
✓	✓	✓	✓		0.3552
✓	✓	✓		✓	0.3533
✓	✓		✓	✓	0.3547
✓		✓	✓	✓	0.3527
✓	✓	✓	✓	✓	0.3576

Table 6: Organization Factor (θ) scores for individual and combined components of Borges: OID_P, Notes & AKA ()N& A), Refresh & Redirect (R & R), and Favicons (F). AS2Org serves as the baseline (θ = 0.3343), *as2org+* achieves 0.3467, and the full configuration of Borges reaches 0.3576.

how Borges improves visibility into the country-level footprint of international conglomerates operating on the Internet (§6.2).

6.1 Borges Contribution to Large Networks

We assess Borges's contribution to producing a more accurate representation of large networks on the Internet, focusing on three categories: (1) access networks, (2) transit providers, and (3) content platforms.

For access networks, we combine sibling inferences from Borges with population estimates from APNIC [27, 42], using data as of July 1, 2024. For transit providers, we evaluate the top 100, 1,000, and 10,000 networks based on CAIDA's AS-Rank, also as of July 1, 2024. For content platforms, we analyze the 16 largest hypergiants identified in prior work [6, 9, 10], including Akamai (AS20940), Amazon (AS16509), Apple (AS714), Facebook (AS32934), Google (AS15169), Netflix (AS2906), Yahoo! (AS10310), OVH (AS16276), Limelight (AS22822), Microsoft (AS8075), Twitter (AS13414), Twitch (AS46489), Cloudflare (AS13335), EdgeCast (AS15133), Booking.com (AS43996), and Spotify (AS8403).

Access Networks. Several international conglomerates operate prominent access networks, fixed-line, wireless, or both, across multiple countries. However, existing AS2Org approaches often fall short in grouping all of a conglomerate's networks under a single organizational entity. Our goal is to assess whether Borges improves this representation by consolidating these distributed access networks into unified organizational groupings.

To evaluate Borges's impact on grouping eyeball networks under common corporate ownership, we first compare the size of eyeball

	# Organizations	AS Population $\mathbb{E}$(AS2Org)	AS Population $\mathbb{E}$(Borges)
Changed	352	3,013,751	3,561,258
Unchanged	25105	117,805	117,805

Table 7: Comparison of the mean ($\mathbb{E}$) AS population between AS2Org and Borges in organizations with and without changes in the number of networks.

populations in Borges versus AS2Org. Table 7 summarizes the number of organizations whose user populations changed under Borges, along with the average number of users per organization.

Out of 25,457 total organizations, only 352 experienced changes in user population due to Borges's reconfiguration, while the remaining 25,105 remained unchanged. However, these changed organizations represent significantly larger user bases. The average number of users in modified organizations increased by approximately 500,000 from 3,013,751 in AS2Org to 3,561,258 in Borges', whereas unchanged organizations remained small, averaging just 117,805 users.

To further quantify the effect of Borges's reconfiguration, we compute the marginal growth in user population across modified organizations. For instance, if organization A in Borges merges B and C from AS2Org, with 300, 200, and 100 users respectively, the marginal growth is the increase over the largest prior group: 300 - 200 = 100 users.

Applying this metric across all reconfigured organizations, Borges achieves a total marginal growth of 193 million users, out of a global base of 4.21 billion, representing a 5% improvement in organizational coverage of the Internet's user population. In other words, Borges consolidates fragmented network entities in a way that more accurately reflects real-world corporate structure, especially among large access providers.

Our final analysis of access networks focuses on the top 20 organizations with the largest marginal user growth under Borges, as shown in Table 8. The results highlight Borges's ability to reconfigure and consolidate large international conglomerates operating across diverse regions, including Deutsche Telekom (T-Mobile), Telkom Indonesia, and Claro.

One of the most notable reconfigurations occurs for Deutsche Telekom, whose footprint expands by over 20 million users under Borges. Similarly, TIGO, a multinational provider active across Latin America, shows a marginal growth of 12 million users. These substantial increases underscore Borges's effectiveness in more accurately representing the scale and reach of global access providers.

Transit Netwokrs. Our next analysis focuses on Borges's impact on transit networks. Using CAIDA's AS-RANK of July 1, 2024, we examine the Borges's contribution to reshaping organizations with transit across the rank.

Our analysis of transit networks focuses on marginal growth in the number of ASNs managed by an organization, rather than user population. Because CAIDA's AS-Rank is computed at the ASN level, and recomputing it at the organizational level would require significant reprocessing, we assess marginal growth by measuring

Company	AS2Org	Borges	Difference
Deutsche Telekom	24,779,378	46,420,443	21,641,065
Telkom Indonesia	33,996,157	54,540,440	20,544,283
Charter	26,624,394	44,440,982	17,816,588
Virgin	11,539,556	25,973,469	14,433,913
TIGO	2,792,759	15,736,350	12,943,591
Claro	6,274,692	18,257,599	11,982,907
Orange	8,983,260	18,711,548	9,728,288
Cablevision Mexico	5,992,157	12,977,362	6,985,205
Free (Iliad)	7,085,849	13,183,971	6,098,122
Telefonica	11,147,816	17,239,924	6,092,108
LG Powercomm	6,689,237	12,683,677	5,994,440
Chunghwa Telecom	7,276,335	12,104,016	4,827,681
Telecom Hulum	12,875,363	17,124,563	4,249,200
Claro Brasil	16,912,676	20,917,350	4,004,674
ACT Fibernet	4,007,919	7,925,537	3,917,618
J:COM (Japan)	4,945,904	7,905,008	2,959,104
Telia	3,159,568	5,713,328	2,553,760
BRM (Brasil)	10,055,599	12,248,262	2,192,663
GigaMais Telecom	1,071,147	3,134,677	2,063,530
Telenor	2,415,632	4,415,607	1,999,975

Table 8: Top 20 marginal AS population growths.

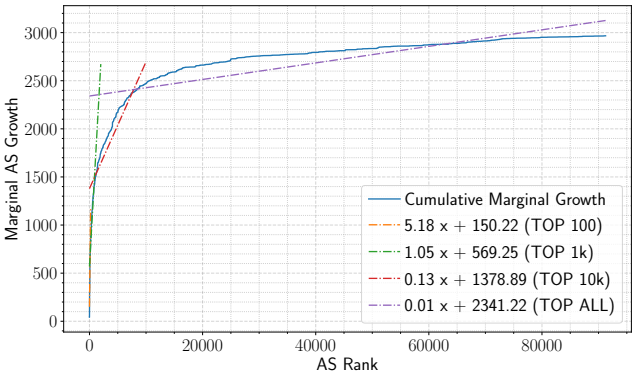


Figure 8: Marginal network growth of organizations sorted by their AS-RANK.

how many additional networks are associated with an organization, relative to its highest-ranked ASN.

Figure 8 presents this analysis, showing the cumulative marginal growth of networks per organization when comparing Borges with AS2Org. The figure also includes linear regression fits for the top 100, 1,000, and 10,000 ASNs in AS-Rank.

Our results reveal that the highest-ranked networks see the greatest consolidation: the top 100 networks gain, on average, 5 additional ASNs under Borges, indicating a substantial reconfiguration of large transit providers. This effect extends through the top 1,000 (slope ≈ 1), but tapers off in the long tail, where networks are more likely to be stand-alone and not part of larger organizational entities.

Hypergiants. Our final analysis examines hypergiant networks, which often span multiple business units (e.g., Google Cloud and YouTube) or reflect historical mergers and acquisitions (e.g., Akamai and Prolexic). We evaluate whether Borges more fully captures the organizational footprint of these companies by consolidating their disparate ASNs under a single entity.

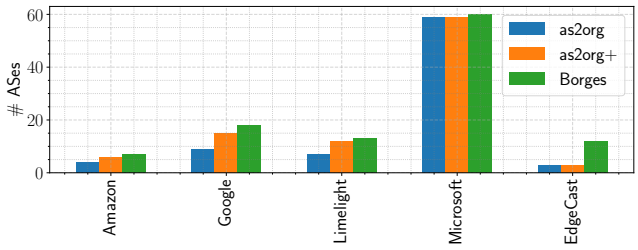


Figure 9: Comparison of the organization size of Hypergiants using AS2Org, *as2org+* and Borges.

Our evaluation shows that Borges improves the organizational representation of 5 hypergiant networks, as illustrated in Figure 9. The most significant change is for Edgecast, which gains 9 additional networks through its consolidation with Limelight. Other hypergiants also benefit: Microsoft, Google, and Amazon see increases of 1, 3, and 1 networks, respectively, compared to AS2Org.

6.2 Conglomerates' Footrprints

Our last analysis expands our analysis of the footprint of international access providers operating in various countries, but now focuses on Borges contribution to drawing a better country-level.

Our analysis shows that Borges expands the country-level footprint, defined as the number of countries where APNIC estimates identify users for 101 organizations. Among these, the average marginal increase is 2.37 countries. Table 9 lists the top 20 organizations with the largest expansions. Notable examples include Digicel, which grows from 4 to 25 countries; Deutsche Telekom (T-Mobile), from 3 to 14; and Claro, from 1 to 5. These results demonstrate Borges's effectiveness in capturing the global reach of multinational network operators.

7 Discussion

While our approach expands the scope and accuracy of AS-to-Organization mappings, several limitations remain. Borges's key contribution is leveraging websites as a new signal for sibling inference. However, there is no longitudinal archive of websites referenced in PeeringDB, which prevents us from analyzing how organizational structures evolve over time.

PeeringDB itself provides only partial coverage of the AS ecosystem, as registration is voluntary and many ASes remain undocumented. For example, prior work by Moura et al. [38] manually identified several Microsoft ASNs that are absent from PeeringDB.

Company	AS2Org	Borges	Difference
Digicel	4	25	21
Zscaler	16	28	12
Deutsche Telekom	3	14	11
NTT	2	11	9
PacketHub	61	70	9
Columbus Networks	5	13	8
TIGO	2	9	7
Cable & Wireless	7	14	7
MainOne	3	9	6
Cogent	18	24	6
Leaseweb	3	9	6
Claro	1	6	5
Latitude Sh	16	21	5
xTom GmbH	4	9	5
Contabo	15	20	5
SoftLayer	7	11	4
UNINETT	1	5	4
IBOSS	3	6	3
Orange	2	5	3
Misaka	2	5	3

Table 9: Top 20 organizations' country-level footprint growths.

Even among registered networks, entries can be incomplete or inconsistent'fields may omit sibling information, lack website URLs, or contain outdated or inaccurate data.

In addition, our method is not designed to capture complex, multi-layered ownership structures that span distinct regions and brands. For instance, América Móvil operates both Claro in Latin America and A1 in Europe [39], yet public records rarely make such relationships explicit. Capturing these deep, often intentionally opaque corporate ties remains an open challenge for future work.

8 Conclusions

The structure of Internet organizations continues to evolve through mergers, acquisitions, rebrandings, and regionally distinct branding strategies. Mapping these dynamic relationships to Autonomous Systems is essential for understanding Internet topology, assessing infrastructure resilience, and informing policy, yet existing AS-to-Organization methods struggle with outdated records, unstructured metadata, and organizational ambiguity.

We introduced Borges, a new framework for inferring sibling ASNs using large language models and website-based signals. Borges extracts relationships from noisy PeeringDB text fields using few-shot prompting and expands coverage through website redirection, domain clustering, and favicon analysis. Our results demonstrate that Borges improves both accuracy and coverage over existing systems, adding millions of users to organizational clusters and improving global visibility into network operators' real footprint.

By blending structured data with semantically aware extraction, Borges opens a path toward more complete, adaptive Internet measurement.

As LLMs continue evolving, Borges opens a path for exploration with future, more complex LLM models, and alternative models to the ones used in this work, such as Meta's Llama and DeepSeek's R1.

9 Acknowledgements

This work was supported by National Science Foundation grant CNS-2107392 and UBACyT 20020220100053BA. The opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of the funding sources. The authors would like to thank the reviewers for their valuable feedback and our shepherd, Oliver Gasser, for their comments on early versions of this manuscript.

References

[1] 2020. RIPE Routing Information Service (RIS). https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris.
[2] 2020. RouteViews. http://www.routeviews.org/routeviews/.
[3] Arelion (formerly Telia Carrier). 2021. Telia Carrier Divestment Completed. https://www.arelion.com/about-us/press-releases/telia-carrier-divestment-completed.
[4] Arelion (formerly Telia Carrier). 2022. Telia Carrier Rebrands as Arelion. https://www.arelion.com/about-us/press-releases/telia-carrier-rebrands-as-arelion.
[5] Augusto Arturi, Esteban Carisimo, and Fabián E. Bustamante. 2023. as2org+ : Enriching AS-to-Organization Mappings with PeeringDB. In *Proc. of PAM*.
[6] Timm Böttger, Felix Cuadrado, and Steve Uhlig. 2018. Looking for hypergiants in peeringDB. *ACM SIGCOMM Computer Communication Review* (2018).
[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* (2020).
[8] CAIDA. 2022. Mapping Autonomous Systems to Organizations: CAIDA's Inference Methodology. https://www.caida.org/archive/as2org/.
[9] Esteban Carisimo, Carlos Selmo, J. Ignacio Alvarez-Hamelin, and Amogh Dhamdhere. 2018. Studying the Evolution of Content Providers in the Internet Core. In *Proc. of TMA*.
[10] Esteban Carisimo, Carlos Selmo, J Ignacio Alvarez-Hamelin, and Amogh Dhamdhere. 2019. Studying the evolution of content providers in IPv4 and IPv6 internet cores. *Computer Communications* (2019).
[11] CenturyLink. 2009. CenturyTel and EMBARQ Complete Merger. https://ir.lumen.com/news/news-details/2009/CenturyTel-and-EMBARQ-Complete-Merger/default.aspx.
[12] CenturyLink. 2010. CenturyLink and Qwest Agree to Merge. https://news.lumen.com/2010-04-22-CenturyLink-and-Qwest-Agree-to-Merge.
[13] CenturyLink. 2011. CenturyLink and Savvis Complete Merger. https://news.lumen.com/2011-07-15-CenturyLink-and-Savvis-Complete-Merger.
[14] CenturyLink. 2016. CenturyLink to Acquire Level 3 Communications. https://news.lumen.com/CenturyLink-to-acquire-Level-3-Communications.
[15] Zhiyi Chen, Zachary S Bischof, Cecilia Testart, and Alberto Dainotti. 2023. Improving the Inference of Sibling Autonomous Systems. In *Proc. of PAM*.
[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* (2023).
[17] CNN. 2020. Sprint Brand to Be Phased Out as T-Mobile Takes Over. https://www.cnn.com/2020/08/03/tech/sprint-tmobile-brand/index.html.
[18] Colt. 2022. Colt and Lumen EMEA. https://www.colt.net/resources/colt-lumen-emea/.
[19] Amogh Dhamdhere, David D. Clark, Alexander Gamero-Garrido, Matthew Luckie, Ricky K. P. Mok, Gautam Akiwate, Kabir Gogia, Vaibhav Bajpai, Alex C. Snoeren, and Kc Claffy. 2018. Inferring Persistent Interdomain Congestion. In *Proc. of ACM SIGCOMM*.
[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] Edgio. 2022. Limelight Completes Acquisition of YahooâĂŹs Edgecast, Rebrands as Edgio. https://edg.io/resources/blog/limelight-completes-acquisition-

of-yahoos-edgecast/.
[22] Vasileios Giotsas, Christoph Dietzel, Georgios Smaragdakis, Anja Feldmann, Arthur Berger, and Emile Aben. 2017. Detecting peering infrastructure outages in the wild. In *Proc. of ACM SIGCOMM*.
[23] Vasileios Giotsas, Georgios Smaragdakis, Bradley Huffaker, Matthew Luckie, and KC Claffy. 2015. Mapping peering interconnections to a facility. In *Proc. of CoNEXT*.
[24] Diana Goovaerts. 2023. T-Mobile sells Sprint fiber assets to Cogent for 1B. https://www.fierce-network.com/telecom/t-mobile-strikes-1-deal-sell-sprint-fiber-assets-cogent.
[25] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv* (2023).
[26] Ralph Holz, Jens Hiller, Johanna Amann, Abbas Razaghpanah, Thomas Jost, Narseo Vallina-Rodriguez, and Oliver Hohlfeld. 2020. Tracking the Deployment of TLS 1.3 on the Web: A Story of Experimentation and Centralization. *ACM SIGCOMM Computer Communication Review* (2020).
[27] Geoff Huston. 2014. How Big is that Network? https://labs.apnic.net/?p=526.
[28] Yuchen Jin, Colin Scot, Amogh Dhamdhere, Vasileios Giotsas, Arvind Krishnamurthy, and Scott Shenker. 2019. Stable and Practical AS Relationship Inference with ProbLink. In *Proc. of USENIX NSDI*.
[29] Aqsa Kashaf, Vyas Sekar, and Yuvraj Agarwal. 2020. Analyzing Third Party Service Dependencies in Modern Web Services: Have We Learned from the Mirai-Dyn Incident?. In *Proc. of IMC*.
[30] Rashna Kumar, Sana Asif, Elise Lee, and FabiÃąn E. Bustamante. 2023. Each at its own pace: Third-party Dependency and Centralization Around the World. In *Proc. of ACM SIGMETRICS*.
[31] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv* (2023).
[32] Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. 2021. Who's Got Your Mail? Characterizing Mail Service Provider Usage. In *Proc. of IMC*.
[33] Aemen Lodhi, Natalie Larson, Amogh Dhamdhere, Constantine Dovrolis, and Kc Claffy. 2014. Using peeringDB to understand the peering ecosystem. *ACM SIGCOMM Computer Communication Review* (2014).
[34] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and KC Claffy. 2013. AS relationships, customer cones, and validation. In *Proc. of IMC*.
[35] Lumen. 2022. Lumen Closes Sale of its Latin American Business to Stonepeak. https://news.lumen.com/2022-08-01-Lumen-Closes-Sale-of-its-Latin-American-Business-to-Stonepeak.
[36] Fabricio Mazzola, Pedro Marcos, Ignacio Castro, Matthew Luckie, and Marinho Barcellos. 2022. On the latency impact of remote peering. In *Proc. of PAM*.
[37] Diego Molina. 2023. Headless is Going Away! https://www.selenium.dev/blog/2023/headless-is-going-away.
[38] Giovane CM Moura, Sebastian Castro, Wes Hardaker, Maarten Wullink, and Cristian Hesselman. 2020. Clouding up the internet: How centralized is dns traffic becoming?. In *Proc. of IMC*.
[39] America Movil. 2022. 2022 ANNUAL REPORT FORM 20-F. https://s22.q4cdn.com/604986553/files/doc_financials/2022/ar/20F-2022-FINAL.pdf.
[40] OpenAI and Josh Achiam et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[41] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv* (2023).
[42] Loqman Salamatian, Calvin Ardi, Vasileios Giotsas, Matt Calder, Ethan Katz-Bassett, and Todd Arnold. 2024. What's in the Dataset? Unboxing the APNIC per AS User Population Dataset. In *Proc. of IMC*.
[43] SEC. 2011. Level 3 to Acquire Global Crossing. https://www.sec.gov/Archives/edgar/data/794323/000119312511093638/dex991.htm.
[44] Selenium. 2023. Selenium. https://www.selenium.dev.
[45] Aaron Souppouris. 2012. Sprint acquires majority stake in Clearwire. https://www.theverge.com/2012/10/18/3520500/sprint-majority-stake-clearwire.
[46] T-Mobile. 2020. T-Mobile Completes Merger with Sprint to Create the New T-Mobile. https://www.t-mobile.com/news/un-carrier/t-mobile-sprint-one-company.
[47] Telecompetitor. 2020. CenturyLink Rebrands as Lumen, Sort of. https://www.telecompetitor.com/centurylink-rebrands-as-lumen-sort-of/.
[48] Vodafone. 2012. Vodafone to Acquire Cable & Wireless Worldwide. https://www.vodafone.com/news/corporate-and-financial/cww.
[49] Xue Xue, John Heidemann, Balachander Krishnamurthy, and Walter Willinger. 2010. Towards an AS-to-organization Map. In *Proc. of IMC*.
[50] Bahador Yeganeh, Ramakrishnan Durairajan, Reza Rejaie, and Walter Willinger. 2019. How cloud traffic goes hiding: A study of Amazon's peering fabric. In *Proc. of IMC*.

A Ethics

This work does not raise any ethical issues

B notes to report upstream rather than siblings

Listing 1 provides an example of a network that uses its notes field to report the ASNs of its upstream connectivity rather than indicating sibling relationships. Without contextual information, these ASNs could be misinterpreted as siblings.

Listing 1: Example of the `net` entry for Maxihost (AS262287) in the PDB snapshot of June 2021.

```
Through the Bare Metal Cloud proprietary platform,
    Maxihost deploys high-performance physical servers
    in multiple regions around the globe. Maxihost owns
    a Tier 3 compliant Datacenter in Sao Paulo, where
    its headquarter is located. See more at https://www.
    maxihost.com/

We connect directly with the following ISPs,
- Algar (AS16735)
- Sparkle (AS6762)
- Voxility (AS3223)
- GTT (AS3257)
- Cogent (AS174)
- FL-IX (Florida Internet Exchange)
- IX.br (Brazilian Internet Exchange)
- Equinix Exchange
- Any2 California (CoreSite Exchange)
- DE-CIX New York
- DE-CIX Dallas
- NSW-IX (Australia Internet Exchange)
```

C Infromation Extraction with LLMs

Listing 2 presents the prompt used to extract embedded sibling information from the notes and aka fields.

Listing 2: Prompt implemented to extract embedded sibling information in notes and aka fields.

```
prompt = """
You are a network topology expert who wants to find
    Autonomous Systems(ASs) that belongs to the same
    organization by reading the peeringdb information.

Please inform the ASs that are peering with the original
    AS.
Don't inform the AS that the original AS is connected to,
     inform the one that are peering as the same
    organization.
If some AS number is mentioned in the 'as-in' and 'as-out
    ' sections in the Notes field, it doesn't mean that
    they belong to the same organization.

The PeeringDB information for the ASN {asn} is:

Notes: {notes}

AKA: {aka}

{format_instructions}

Just inform an AS if it is number is explicitly written
    in the AKA or Notes fields provided.
Yo don't know the relation between a company name and its
     AS number.
```

```
Also explain why you choose the ASs informed.

"""
```

D Blocklists

Borges utilizes blocklists to filter out domains reported in the networks' website field that do not point to the company's official website but instead redirect to mainstream communication platforms.

D.1 Subdomain Blocklist

Table 10 shows the manually curated list of subdomains removed from consideration when inferring sibling across networks reporting the same subdomain.

No.	Blocked Domain
1	myspace
2	github
3	he
4	facebook
5	instagram
6	linkedin
7	bgp.tools
8	oracle
9	discord
10	peeringdb

Table 10: Blocklist of Website Domains

D.2 Final URL Blocklist

Table 11 presents the manually curated list of domains excluded from sibling inference when they are used along with favicons.

No.	Blocked Domain
1	example.com
2	github.com
3	linkedin.com
4	facebook.com
5	discord.com

Table 11: Blocklist of Website Domains

E LLM-based Classifier

Listing 3 displays the prompt to identify whether a favicon and a list of final URLs correspond to networks operating under the same corporate umbrella.

Listing 3: Prompt implemented to determine a favicon and a list of final URLs correspond to networks under the same corporate umbrella

```
message = HumanMessage(
    content=[
        {"type": "text", "text": f"Accessing these URLs {x['final_url']} returned the attached favicon. If it is a telecommunications company, what is the company's name? If it is a subsidiary, provide the parent company's name. If it is not a telecommunications company, is it a hosting technology? Reply only with the name of the company or technology. If it is none of the above, reply 'I don't know'."},
        {
            "type": "image_url",
            "image_url": {"url": f"{image_data}"},
        },
    ],
)
```