BGP Communities: Analysis of Adoption and Dataset Creation

Andrei Pochmann Koenich[1], Arthur Vinícius Cunha Camargo[1], Leandro Marcio Bertholdo[1], Renan Paredes Barreto[2], Lisandro Zambenedetti Granville[1]

[1]Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil

[2]Universidade Federal do Rio Grande (FURG)
Rio Grande – RS – Brasil

{apkoenich, avccamargo, leandro.bertholdo, granville}@inf.ufrgs.br

renan.barreto@furg.br

Abstract. *The Border Gateway Protocol (BGP) communities enable autonomous systems (ASes) to perform traffic engineering, route selection, and routing coordination between their transit providers and peers. However, the heterogeneous and poorly standardized community use complicates routing automation efforts. In this paper, we present BGPScout, an application designed to automate the collection, interpretation, and consolidation of BGP community data from public sources. Using web scraping and natural language processing (NLP) techniques, BGPScout organizes information about traffic engineering communities, producing a consistent dataset that aims to support network automation and analysis activities.*

1. Introduction

The Border Gateway Protocol (BGP) is an essential routing protocol on the Internet because of its function of exchanging routing information between Autonomous Systems (ASes). BGP communities, in turn, represent an attribute with essential functionalities, as they allow the implementation of advanced routing policies, supporting applications such as traffic engineering [Donnet and Bonaventure 2008], outage detection [Giotsas et al. 2017], network geolocation [CAIDA 2021], and identification of networks under DDoS attacks [Giotsas et al. 2013, Streibelt et al. 2018].

Although well-defined standards for BGP communities exist, such as the well-known communities established in RFCs 1997 and 8642, their adoption remains largely non-standardized. Most ASes employ operator-defined communities [Brasil Peering Forum 2022], and this lack of standardization significantly hinders automation efforts through techniques involving software-defined networks (SDNs) [Bertholdo et al. 2020, Rizvi et al. 2022]—a challenge that largely affects content distribution networks (CDNs), which require dynamic and responsive traffic management.

In this context, this article presents BGPScout, a preliminary study aimed at analyzing the syntax and semantics of BGP communities using publicly available information. By combining web scraping techniques with large language models (LLMs), BGPScout extracts and organize relevant content from operator webpages, producing a dictionary of BGP communities for traffic engineering. This study represents an initial step toward the construction of a dataset that supports automated interpretation of BGP routing policies, contributing to future research on traffic engineering in the Internet.

2. Related Works

While the syntactic structure of BGP communities is standardized, their semantics are not. Each AS defines and applies community values according to its own operational goals, which makes it difficult to build automated systems that rely on them. This challenge has motivated research efforts aimed at inferring semantics from observable behavior. Existing work follows two main directions: inspecting operator-published documentation or analyzing public routing data collected by route collectors.

Early work on documentation-based extraction relies on lightweight web-scraping to obtain BGP community semantics from operator webpages [Razafindralambo 2018], trying to identify textual patterns associated with community definitions and to recover communities with missing or unknown semantic categories. Rather than using a machine-learning classifier, the method focuses on structuring the data.

Subsequent work expands this direction by introducing supervised learning to predict semantics [Werner 2020]. Using a manually created ground-truth dataset and a unified taxonomy of semantic categories, the authors train Multi-Layer Perceptron and Random Forest classifiers. By testing on BGP MRT data, the model recognized structural patterns embedded in community values and achieved up to 89.15% accuracy across the defined semantic subclasses.

A different line of research infers semantics directly from routing behavior by analyzing publicly available BGP updates and tables from route collectors [Silva Junior et al. 2024]. The authors classify communities into two broad groups: (i) *location communities*, which describe where a route was learned, and (ii) *action communities*, which encode traffic-engineering instructions. These methods, however, cannot recover the precise operator-defined meaning of each individual community.

More recently, Liu et al. [Liu et al. 2024] revisit documentation-driven extraction at Internet scale. Their system uses search engines to discover candidate webpages, applies machine-learning filters to remove irrelevant results, and employs regular expressions to extract structured information. By learning textual patterns through frequent itemset mining (Apriori algorithm), the system collects more than 513,000 URLs, of which 397 contain valid operator-defined BGP community semantics. Their work represents the current state of the art in semantic BGP community datasets.

Our approach takes a fundamentally different path by relying exclusively on self-reported semantics. All meanings are extracted directly from authoritative operator documentation, without applying any algorithmic inference or prediction. We intent to preserve any useful information provided by humans. Our goal is to build a dataset for further LLM processing, keeping track of all operator-provided information such as “Do not announce to Level3/CenturyLink (Brazil)” instead of coded informations such as“tag loc regular `^30[0-9]{3}$ 1-NA`” proposed by [Liu et al. 2024]. Table 1 provides a comparative overview that highlights these differences across existing approaches.

Table 1. Summary of BGP Community Semantics Extraction Approaches

Approach	Data Source	Core Technique	Semantic Type	Contribution
[Razafindralambo 2018]	Operator webpages	Scraping with structural pattern mining	Documented (self-reported)	Early structured extraction of BGP community semantics and recovery of missing categories
[Werner 2020]	Operator webpages	Supervised ML (RF + MLP)	Predicted taxonomy	Applies classification models to infer semantic categories from numeric patterns
[Silva 2024]	BGP updates + RIB dumps	Statistical inference of routing behavior	Inferred semantics	Distinguishes location vs. action communities through behavioral signal analysis
[Liu 2024]	Web sources + IRR data	Regex-based extraction + ML filtering	Documented (self-reported)	State-of-the-art large-scale documentation mining with Apriori-based pattern discovery
This Work	**Operator site (authoritative sources)**	**LLM parsing combined with adaptive scraping**	**Documented (self-reported)**	**Preserves original operator meaning; produces a high-quality dataset for downstream LLM processing**

3. Methodology

Our methodology takes into account the fact that traditional web scraping techniques lack the full capacity to handle the great heterogeneity of modern web pages. Therefore, LLMs represent a promising alternative, given their ability to adapt and understand data with little or no structure [Ahluwalia and Wani 2024]. This type of approach is capable of ensuring the processing of information based on meaning and context, surpassing methods that use regular expressions. However, LLMs also face problems like hallucinations and difficulties in handling specific data structures.

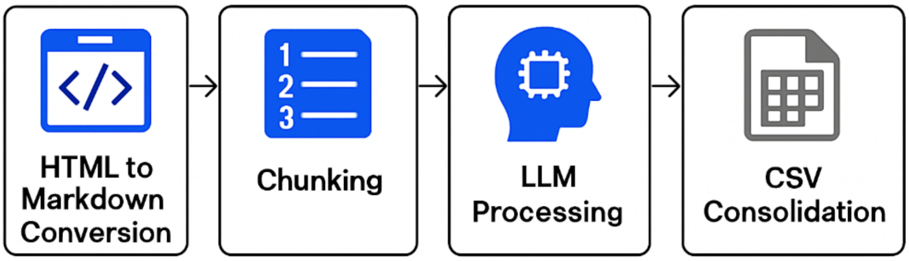

Figure 1. Methodology four-stage pipeline

In light of these factors, our proposed methodology operates through a four-stage pipeline, as shown in Fig. 1. In the first stage, the application uses Jina AI Reader [Wang et al. 2025] to convert the HTML content of a page into Markdown, preserving the page's semantic structure, removing noisy data, and emphasizing relevant information, resulting in LLM-friendly content [Mukherjee 2025]. Experimentally, the HTML to Markdown conversion provided by Jina AI proved to be significantly better in this context compared to other frameworks such as Docling [Auer et al. 2024], offering superior removal of unnecessary information and more consistent formatting.

In the second stage, the Markdown content is divided into chunks of approximately 15,000 characters, using page delimiters such as headings and section breaks as a reference for the division. This division limits incoherent content separations and allows for compliance with the processing limit imposed by the LLM used in the third stage. The reason for imposing a size limitation on the chunks is that chunk size is directly proportional to output size: the larger the chunk, the more output tokens the LLM model must generate and process. Therefore, it is important that the page includes a strategic division to ensure that the entire page can be processed and scraped successfully.

In the third stage, each chunk obtained from previous stage undergoes iterative processing using a LLM model. We employ prompt engineering, with detailed scraping instructions, the expected output format, and filter criteria, aiming to achieve exclusive traffic engineering community results. Furthermore, the prompt also determines the treatment that should be given to the parameterized communities.

Finally, the fourth stage consolidates all the information obtained into a single CSV file, which contains five fields: *AS number*, *AS name* (from the source page), *community value*, *community identifier/name*, and *functionality description*.

4. Partial Results

We developed BGPScout following the proposed methodology, and to evaluate its effectiveness, we applied this approach in several case studies, including a complex wiki page that compiles BGP communities used by Brazilian and international network operators [Brasil Peering Forum 2022]. This page serves as a representative example of the type of content targeted by BGPScout, as it provides a comprehensive overview of the local ecosystem, with heterogeneous formatting and parameterized community examples. We used this page as a source for qualitative analysis. For the experiments, we processed the resulting 9 chunks using the `Gemini-Flash-2.5` model, which was selected solely for evaluation purposes due to its availability to our research group and without affecting the generality or modularity of the pipeline design.

All tests were performed in a system equipped with an AMD Ryzen 7 5700U processor (8 cores, 16 threads) and 32GB of RAM. Across ten runs, the median execution time was 10 minutes and 34 seconds.

In our primary test case, the processing resulted in the extraction of 1,256 distinct BGP communities across 35 different ASes. Table 2 summarizes the total number of ASes and extracted communities by category.

Table 2. Distribution of communities by operator category

Category	ASes	Communities
National operators	15	571
International transit	10	579
Internet Exchanges	5	17
CDNs and global providers	3	52
Others	2	37
Total	**35**	**1,256**

Table 3. ASes and Community Coverage Across Datasets

Coverage	BGPScout	State-of-the-Art
Exclusive	20 ASes 781 communities	893 ASes 29,419 communities
Shared	15 ASes 475 communities	
Total	35 ASes 1,256 communities	908 ASes 29,894 communities

The results in Table 3 show that approximately *62% of the communities* and *57% of the ASes* extracted in our primary test case do not appear in the state-of-the-art semantic dataset [Liu et al. 2024]. Although the Liu et al. dataset is much larger in total size, it does not cover many operators and communities that BGPScout identifies from current documentation. This indicates that, despite its scale, the existing dataset still leaves significant room for further enrichment.

A closer inspection suggests that the missing coverage is related to regional operators or language constraints. The 20 ASes not included in the state-of-the-art semantic dataset are mostly associated with Latin American networks, whose documentation is often available only in Portuguese. These cases were likely not captured by Liu's methodology, which relies on English-language search terms and globally visible operators.

In addition, the state-of-the-art dataset [Liu et al. 2024] also omits/misses some community semantics, because it relies on outdated third-party documentation sources, including pages published as early as 2008. As a result, more recent semantics are not mapped. For example, BGPScout extracts communities in the 3549:700X range, which implement continent-based no-export decisions and are still active today, as confirmed through public looking glasses such as RIPE's. These communities do not appear in the Liu et al. dataset, illustrating how dependence on legacy sources leads to incomplete coverage and highlighting opportunities that our approach begins to address.

For the subset of communities shared by both datasets, the semantics reported by the two dictionaries are consistent. However, when we submit both dictionaries' semantic analysis by LLMs, the BGPScout dictionary proves to be significantly more LLM-friendly than the dictionary proposed by [Liu et al. 2024]. In Table 4, we illustrate BGPScout's explicit natural-language descriptions in contrast with the complex, nested JSON schema used in the state-of-the-art dataset. Using an LLM-as-a-Judge approach [Gu et al. 2024], where the models (`GPT-4o`, `Gemini 1.5 Flash`, `DeepSeek-V2`, and `Claude 3 Sonnet`) are tasked with interpreting the semantics of each dictionary, BGPScout's dictionary consistently demonstrated superior interpretability, indicating that its semantic structure is more LLM-friendly and that the specific function of each BGP community could be more easily understood by the LLMs. This result is further supported by the fully tabular organization of the BGPScout dictionary, which enables direct and intuitive lookup of each community, while the state-of-the-art dataset adopts a nested JSON structure that requires additional navigation.

Table 4. Comparison for AS2914 communities

Community	BGPScout	State of the Art
2914:411	Add 1 prepend for other NTT customers	prepend explicit 411 1
2914:429	Do not announce to any peer	sel_ann no-export explicit 429 all
2914:435	Change local-preference 50 (beyond country)	pref explicit 435 50
2914:4011	Add 1 prepend for all NTT customers in North America	sel_ann export regular `^401[1-3]$` customer 1-NA
2914:4229	Do not announce to any NTT peer in Europe	sel_ann no-export explicit 4229 peer,1-EU

5. Conclusion

This paper presented BGPScout, an application on LLMs, whose function is to build a dictionary of BGP communities using web scraping techniques. The proposed approach demonstrates the viability and effectiveness in extracting and consolidating information about BGP communities from multiple heterogeneous sources on the Internet.

Despite the promising initial results, BGPScout faces significant challenges. The extraction quality is highly dependent on the original structure of the source: well-organized pages with clear tables yield substantially better results than pages with inconsistent formatting or dispersed information. Another important challenge concerns the validation of the semantic correctness of the extracted communities. While the system ensures syntactic correctness, verifying that the functional description matches the community's actual behavior requires external validation, ideally through comparison with data observed in real routing tables.

The main contribution of this work is to demonstrate that LLMs, when correctly guided through meticulous prompt engineering, can automate tasks that traditionally require intensive human intervention. In the context of computer networks, this capability makes it possible to build and maintain knowledge bases essential for network automation, as well as the development of intelligent infrastructure management systems.

Promising directions include the processing of multiple sources with consolidation and conflict resolution, validation through comparison with data from RIPE RIS and RouteViews, incremental updates based on source change detection, and integration with systems such as BGPTunner [Bertholdo et al. 2020] and Anycast Agility [Rizvi et al. 2022], enabling automated traffic engineering policies. The continuation of these efforts aims to establish BGPScout as a reference tool for knowledge bases on BGP communities.

References

[Ahluwalia and Wani 2024] Ahluwalia, A. and Wani, S. (2024). Leveraging large language models for web scraping.

[Auer et al. 2024] Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Ramis, C. B., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., Lima, R. T. d., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., and Staar, P. W. J. (2024). Docling technical report. Technical report, AI4K Group, IBM Research.

[Bertholdo et al. 2020] Bertholdo, L. M., Ceron, J. M., Granville, L. Z., Moura, G. C., Hesselman, C., and Van Rijswijk-Deij, R. (2020). BGP anycast tuner: Intuitive route management for anycast services. In *16th International Conference on Network and Service Management (CNSM)*.

[Brasil Peering Forum 2022] Brasil Peering Forum (2022). Lista de communities BGP.

[CAIDA 2021] CAIDA (2021). BGP community dictionary dataset.

[Donnet and Bonaventure 2008] Donnet, B. and Bonaventure, O. (2008). On BGP communities. *SIGCOMM Computer Communication Review*, 38(2):55–59.

[Giotsas et al. 2017] Giotsas, V., Dietzel, C., Smaragdakis, G., Feldmann, A., Berger, A., and Aben, E. (2017). Detecting peering infrastructure outages in the wild. In *Proceedings of the ACM SIGCOMM Conference on Data Communication*, pages 446–459.

[Giotsas et al. 2013] Giotsas, V., Zhou, S., Luckie, M., and Claffy, K. (2013). Inferring multilateral peering. In *Proceedings of the 2013 ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, pages 247–258.

[Gu et al. 2024] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J. (2024). A survey on LLM-as-a-judge.

[Liu et al. 2024] Liu, Y., Wu, T., Wang, J. H., Wang, J., and Zhuang, S. (2024). Collecting self-reported semantics of BGP communities and investigating their consistency with real-world usage.

[Mukherjee 2025] Mukherjee, A. (2025). Boosting AI performance: The power of LLM-friendly content in markdown. Webex Developer Blog. Accessed on November 06, 2025.

[Razafindralambo 2018] Razafindralambo, N. (2018). Revisiting the BGP communities usage. Master's thesis, Faculté des Sciences appliquées. Master en sciences informatiques, à finalité spécialisée en "computer systems and networks".

[Rizvi et al. 2022] Rizvi, A. S., Bertholdo, L., Ceron, J., and Heidemann, J. (2022). Anycast agility: Network playbooks to fight ddos. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security)*, pages 4201–4218.

[Silva Junior et al. 2024] Silva Junior, B. A. d., Cunha, I. S., and Ferreira, R. A. (2024). *Automatic Inference of BGP Community Semantics*. PhD thesis, Federal University of Mato Grosso do Sul.

[Streibelt et al. 2018] Streibelt, F., Lichtblau, F., Beverly, R., Feldmann, A., Pelsser, C., Smaragdakis, G., and Bush, R. (2018). BGP communities: Even more worms in the routing can. In *Proceedings of the Internet Measurement Conference (IMC)*, pages 279–292.

[Wang et al. 2025] Wang, F., Li, Y., and Xiao, H. (2025). jina-reranker-v3: Last but not late interaction for listwise document reranking.

[Werner 2020] Werner, J. (2020). Chasing the unknown: A predictive model to demystify BGP community semantics. Master's thesis, Naval Postgraduate School.