Poster: Collection and Sharing of A Residential Dataset

Shuyue Yu
syyu@cs.columbia.edu
Columbia University
New York, United States

Thomas Koch
tak2154@columbia.edu
Columbia University
New York, United States

Ilgar Mammadov
im2703@columbia.edu
Columbia University
New York, United States

Hangpu Cao
hc3346@columbia.edu
Columbia University
New York, United States

Gil Zussman
gil@ee.columbia.edu
Columbia University
New York, United States

Ethan Katz-Bassett
ethan@ee.columbia.edu
Columbia University
New York, United States

Abstract

Given the increasing residential Internet use, a thorough understanding of what services are used and how they are delivered to residential networks is crucial. However, access to residential traces is limited due to their proprietary nature. Most prior work used campus datasets from academic buildings and undergraduate dorms, and the few studies with residential traces are often outdated or use data unavailable to other researchers. In our SIGMETRICS 2025 publication [15], we introduced a new residential dataset—we have been collecting traffic from ~1000 off-campus residences that house faculty, postdocs, graduate students, and their families. Although our residents are university affiliates, our dataset captures their activity at home, and we show that this dataset offers a distinct perspective from the campus and dorm traffic. We also investigate the serving infrastructures and services accessed by the residences. Extending this published work, since May 2025, we have improved our pipeline efficiency to enable continuous 24/7 data collection and scale up to ~1500 residences, providing a more complete view of the residential network. We also make the dataset available for research use upon request, with the goal of motivating and supporting future research.

ACM Reference Format:
Shuyue Yu, Thomas Koch, Ilgar Mammadov, Hangpu Cao, Gil Zussman, and Ethan Katz-Bassett. 2025. Poster: Collection and Sharing of A Residential Dataset. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025, Madison, WI, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3730567.3768615

1 Introduction

Residential Internet usage has rapidly increased and morphed, making it crucial to thoroughly understand which services users access and how they are delivered.

In general, the community lacks access to *open residential traces.* Prior work mapped serving infrastructures of large providers via active measurements [3, 4], but they lacked visibility into how traffic was delivered from the infrastructures. Other studies collected user traffic via crowdsourcing [5], but what could be measured was limited since they required explicit action from participants who deployed the kits. These datasets usually included few users (~100 households) and were hard to scale and maintain. Most datasets available to researchers were from campus networks and captured behavior in undergraduate dorms, classrooms, and offices [11, 13]. We show that even with dorms, on-campus traffic still differs from off-campus residential usage.

Studies using industry traces illustrated the serving infrastructures from their perspectives [9, 12], but they may not apply to services beyond those offered by those providers. Some reports showed global service usage patterns [8], but they did not offer peer-reviewed methodologies and usually lack details. A few academic papers used traces from residential networks [2, 7, 10], but they often did not provide a broad understanding of Internet use and are now outdated. Further, the datasets were not available to other researchers, limiting the set of questions answered.

In our recent paper published at SIGMETRICS 2025 [14], we introduced a dataset collected for four hours per day from 1000 off-campus residences that house *graduate students, faculty, and their families.* Even though our residents are university-affiliated, our dataset is residential, capturing Internet usage at homes. While our network differs from many regular residential networks operated by large commercial ISPs, our dataset remains a valuable new source. To our knowledge, this is the largest and most detailed residential dataset shared to date. We also present a *detailed, broad view of Internet service usage and delivery* for this network.

IMC '25, Madison, WI, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1860-1/2025/10
https://doi.org/10.1145/3730567.3768615

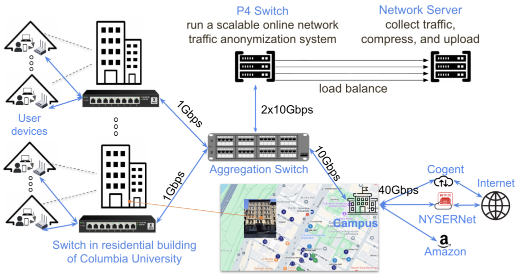
P4 Switch
run a scalable online network traffic anonymization system
Network Server
collect traffic, compress, and upload
load balance
2x10Gbps
User devices
1Gbps
1Gbps
Aggregation Switch
10Gbps
40Gbps
Cogent
Internet
NYSERNet
Campus
Amazon
Switch in residential building of Columbia University

Figure 1: Our network topology and data collection pipeline.

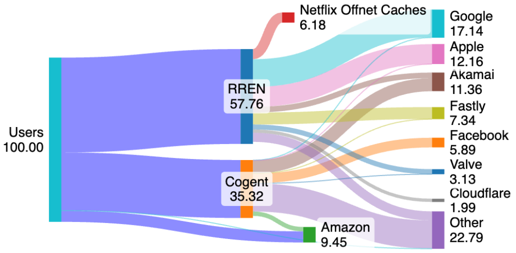

Figure 2: Key organizations on the routes between our users and remote hosts.

Following this publication, we have improved our pipeline to enable 24/7 collection, expanding coverage to ~1500 off-campus residences. We have also been sharing the dataset with other researchers for studies that have undergone IRB review and follow our Acceptable Use Policy. With our broad analysis and data sharing, we hope to enhance understanding of Internet service delivery and support future research.

2 An Open Residential Dataset

Columbia University owns many off-campus residential buildings, which house different resident types: (1) graduate students (mainly PhD students), (2) postdocs, and (3) faculty and their families. With data collected from our campus, we demonstrate that Columbia's off-campus residential traffic significantly differs from its on-campus traffic.

Our residential dataset consists of anonymized, unsampled packet traces from over 400 residential apartments collected four hours per day for more than three years, and from approximately 1000 residences since early 2024. Initially, our data collection pipeline ran on standard commodity hardware with limited processing and storage capacity, which restricted us to capturing only four hours of data per day until May 2025. Since then, we have deployed an upgraded pipeline capable of continuous, 24/7 residential data collection and scaling to approximately 1500 residences.

The upgraded pipeline is built around an Intel Tofino programmable switch, which runs a scalable online network traffic anonymization system [6]. This system efficiently anonymizes IP and MAC addresses and flexibly removes packet payload depending on protocol, while preserving key information such as DNS A records and TLS Server Name Indication (SNI). Figure 1 shows the details of our network topology and data collection pipeline, and Figure 2 visualizes the flow of traffic between our residential users and the organizations of the remote hosts with a Sankey diagram.

3 Data Sharing

The pipeline was declared exempt from Human Subjects Research by our Institutional Review Board (IRB) and approved by our university IT. We make the dataset available to researchers upon request for studies that have undergone IRB review and agreed with our Acceptable Use Policy. Both the existing dataset and future collections—which are not easily available within campus networks and require significant efforts to collect—will be accessible. Additional details, including metadata, access request procedures, and the Acceptable Use Policy, are provided on our project website [1].

Acknowledgements

This work is supported in part by NSF grants OAC-2029295 and CNS-2212479.

References

[1] 2025. Columbia University Residential Datasets. https://wimnet.github.io/CUResidential/.

[2] M. Allman. 2020. Putting DNS in Context. In *IMC 2020*.

[3] M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan. 2013. Mapping the Expansion of Google's Serving Infrastructure. In *IMC 2013*.

[4] P. Gigis, M. Calder, L. Manassakis, G. Nomikos, V. Kotronis, X. Dimitropoulos, E. Katz-Bassett, and G. Smaragdakis. 2021. Seven Years in the Life of Hypergiants' Off-Nets. In *SIGCOMM 2021*.

[5] S. Grover, M. Park, S. Sundaresan, S. Burnett, H. Kim, B. Ravi, and N. Feamster. 2013. Peeking Behind the NAT: An Empirical Study of Home Networks. In *IMC 2013*.

[6] H. Kim and A. Gupta. 2019. ONTAS: Flexible and Scalable Online Network Traffic Anonymization System. In *Proceedings of the Workshop on Network Meets AI and ML 2019*.

[7] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. 2010. Internet Inter-Domain Traffic. In *SIGCOMM 2010*.

[8] Sandvine. 2023. https://www.sandvine.com/global-internet-phenomena-report-2023

[9] B. Schlinker, H. Kim, T. Cui, E. Katz-Bassett, H. Madhyastha, I. Cunha, et al. 2017. Engineering Egress with Edge Fabric: Steering Oceans of Content to the World. In *SIGCOMM 2017*.

[10] M. Trevisan, D. Giordano, I. Drago, M. Mellia, and M. Munafo. 2018. Five Years at the Edge: Watching Internet from the ISP Network. In *CoNEXT 2018*.

[11] A. Ukani, A. Mirian, and A. Snoeren. 2021. Locked-in during Lockdown: Undergraduate Life on the Internet in a Pandemic. In *IMC 2021*.

[12] F. Wohlfart, N. Chatzis, C. Dabanoglu, G. Carle, and W. Willinger. 2018. Leveraging Interconnections for Performance: The Serving Infrastructure of a Large CDN. In *SIGCOMM 2018*.

[13] B. Yeganeh, R. Rejaie, and W. Willinger. 2017. A View from the Edge: A Stub-AS Perspective of Traffic Localization and Its Implications. In *TMA 2017*.

[14] S. Yu, T. Koch, I. Mammadov, H. Cao, G. Zussman, and E. Katz-Bassett. 2025. Internet Service Usage and Delivery As Seen From a Residential Network. *Proc. ACM Meas. Anal. Comput. Syst.* 9, 2, Article 41 (2025).