Mastodon Network Analysis

Ved Kothavade
University of Maryland, College Park
College Park, MD
vedk@terpmail.umd.edu

Phan Ngyuen
University of Maryland, College Park
College Park, MD
phan2003@terpmail.umd.edu

***Abstract*—The Mastodon network is a large area of research for sociologists and computer scientists alike, due to its distributed nature and unique audience. In this report, we crawl Mastodon network and measure statistics about the nature of the instances that make it up, and the connections between them. There exists research done on the content of the posts on Mastodon instances, but there is none to be found regarding the nature of the servers that run instances themselves, nor their connectivity with each other.**

TODOs

I. Introduction

We want to explore characteristics of the Mastodon network related to the geographical distribution of nodes, cloud providers nodes are hosted on, and more. This paper is largely inspired by "Design and evaluation of IPFS: a storage layer for the decentralized web", which demonstrates the kinds of measurments that could be done on a decentralized network [1].

II. Methodology

While writing a crawler is simple in principle, to mitigate time constraints we began with a seed list of alive nodes created by an external, open-source Fediverse crawler. **TODO: cite crawler** This crawler is updated every 6 hours. From the list created by this crawler, we filter down to Fediverse nodes hosting Mastodon or Mastodon-compatible software (as opposed to, for example, WordPress), and then to nodes which supported the `peers` API we leverage to determine node connectivity.

1: cite crawler

From this filtered list of nodes, we then collected data about each node leveraging the MaxMind GeoIP databases and by resolving the IPs for the domains, and looking up their ASN, AS organization names, country codes. **TODO: cite MaxMind** Furthermore, we use an `instance` API to collect user and post counts for all instances.

2: cite MaxMind

Upon collecting this data about nodes in a database, we use the neo4j graph database to insert all nodes, the properties of these nodes, and finally create `PEERS_WITH` relationships between peers as defined by the aforementioned `peers` API.

We believe that this is a representative dataset of the Mastodon network, and thus feel that the analyses we conduct from this graph dataset represent the true network well.

III. Analysis

A. Location of Instances

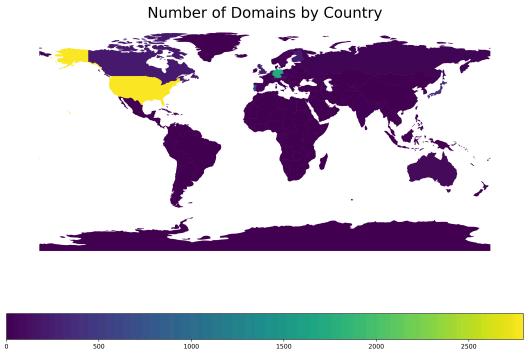

Fig. 1: Heatmap of Instance Locations

TABLE I: Instance Counts in Top 10 Countries

Rank	Country	Count	Share
1	US	2795	30.72%
2	FR	1735	19.07%
3	DE	1700	18.69%
4	PT	652	7.17%
5	JP	415	4.56%
6	GB	249	2.74%
7	FI	234	2.57%
8	CA	229	2.52%
9	NL	181	1.99%
10	SG	138	1.52%
Total		9097	

As Table I shows, the majority of instances are in the United States, followed by France, Germany, Portugal, and Japan. In fact, the top 5 countries alone account for 80.21% of all instances.

B. Cloud Providers

TABLE II: Instance Counts by Cloud Provider

Rank	Cloud Provider	Count	Share
1	None	6200	59.33%
2	OVH	2245	21.48%
3	Hetzner	1207	11.55%
4	DigitalOcean	493	4.72%
5	AWS	200	1.91%
6	GCP	68	0.65%
7	Azure	37	0.35%
Total		10450	

Most instances are not hosted on cloud providers, and among those that are, surprisingly the most common are none of the big 3, but instead OVH, Hetzner, and DigitalOcean. In total, we see that 40.67% of instances are hosted on cloud providers, indicating that the majority of instances are run on personal servers. From this, we can infer that many Mastodon administrators are

TODO: smth about relevance of non-cloud

3: smth about relevance of non-cloud

C. Autonomous Systems

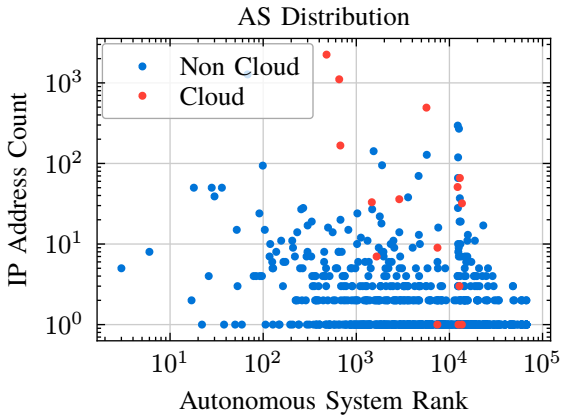

Fig. 2: Distribution of IPs across ASes according to their size (measured by their AS rank).

We mapped IP addresses to Autonomous Systems using GeoLite2, found the CAIDA AS Rank for each AS. According to CAIDA, the AS Rank is a measure of the influence of an AS to the global routing system, as calculated by their sizes, peering agreements, and more.

We then plotted the number of IP addresses per AS, colored by whether the AS is a cloud provider or not. We see that the cloud providers are disproportionately large, and that there are many more IP addresses in the cloud providers than in the non-cloud providers.

TABLE III: Autonomous Systems covering > 50% of all found IP Addresses

ASN	AS Rank	AS Name	Count	Share
16276	479	OVH	2242	21.45%
13335	68	CLOUDFLARENET	1273	12.18%
24940	655	HETZNER-AS	1108	10.6%
14061	5664	DIGITALOCEAN-ASN	493	4.72%
63949	12281	AKAMAI-LINODE-AP	295	2.82%
Total			10450	51.78%

D. Most Peered Instances

TABLE IV: Top 10 Most Peered Instances

Instance	Peers	Users	Posts
mastodon.social	16821	2734739	131726225
mastodon.online	16464	189377	10608587
mstdn.social	16439	261299	20088443
mas.to	16421	183971	10976052
fosstodon.org	16300	62442	4240866
chaos.social	16274	13005	8360180
hachyderm.io	16274	56449	4074546
infosec.exchange	16215	75993	4361671
mastodon.gamedev.place	16188	33514	1721158
social.tchncs.de	16178	22857	3573231
mastodon.world	16171	191803	7250207

From Table IV, we see that the most peered instance is `mastodon.social`, with 16464 peers. There appears to be a trend that the more peered an instance is, the more users and posts it has, but this is not strict.

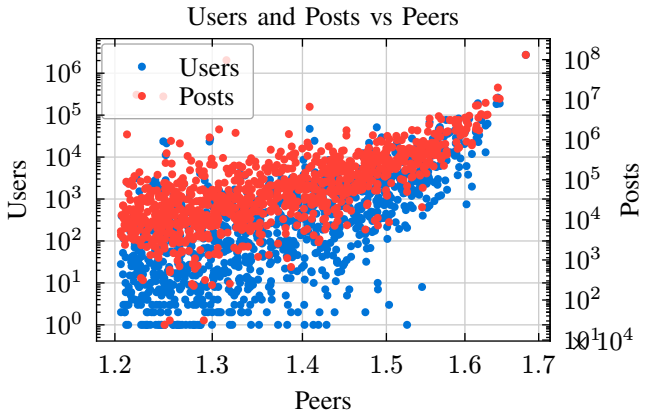

Fig. 3: Users and Posts vs Peers for 1000 Most Peered Instances

We see that there is a positive correlation between the number of peers and the number of users and posts in Fig. 3.

References

[1] D. Trautwein *et al.*, "Design and evaluation of IPFS: a storage layer for the decentralized web," in *Proceedings of the ACM SIGCOMM 2022 Conference*, in SIGCOMM '22. Amsterdam, Netherlands:

Association for Computing Machinery, 2022, pp. 739–752. doi: 10.1145/3544216.3544232.