How We Measure: ISP User Counts

In our physical world census information is used to inform the planning processes behind the provision of infrastructure, such as schools, hospitals, housing, and similar. It can be used to assess the impact of natural disasters, or to understand a society's needs in terms of food and energy security. Demographic data is also used to inform investment and business decisions.

You'd think that the Internet itself would be awash with similar information. After all, much of the Internet's economy is based on the aggregation of user profile data and then repackaging this data and selling it to advertisers in the form of ad placement capabilities. So, it's likely to be the case that similar census-related data is being continually gathered on the Internet. However, this data is a key commercial asset, owned by the corporate entities that gather the data. There is very little public data of a similar nature that relates the market positioning of Internet Service Providers (ISPs) in terms of the number of users of their services.

In our measurement work at APNIC Labs we were trying to relate our measurement data, based on a sampled subset of users, to the larger picture of user populations. If you had the information of the number of users of each Internet Service Provider (ISP), then if would be possible to derive data that can infer the level of adoption of a particular technology, such as IPv6, or DNS Security mechanisms.

This data would also be extremely useful in a number of areas. When a major ISP experiences a service failure what is the impact of service disruption where the ISP service has failed?

> There was an 8-hour service outage experienced by a major ISP in Australia, Optus, on the 8th November 2023. This provider is the second largest provider in the Australian ISP market, with an estimated 4 million users, so the outage was a major incident.
> https://en.wikipedia.org/wiki/2023_Optus_outage

The data would also be extremely useful in the area of public policy. How open is the market for the provision of Internet services within each country? How many users are served by each ISP? What's their respective market share?

Such information can also inform policy issues related to national security and resilience: How many local users are reliant on the services provided via a foreign platform?

Our response to this missing data set was to generate, on a daily ongoing basis, our estimate of the number of users per ISP for every ISP that we see on the Internet through the ad-based measurement platform. This report is published at the URL: https://stats.labs.apnic.net/aspop. As far as we are aware this is the only such public data set that encompasses the entirety of the public Internet.

Here I would like to explain how we calculate this data, and provide some responses to a recent presentation at the RIPE 89 meeting on this data set.

Data Generation

The process starts with the estimated current population in each country. The data we use is sourced from the United Nations Population Division. We use the mid-year population estimate from 2023 and apply the 2022-2023 growth rate to the period from mid 2023 to the present day to get an estimate of the current population of each country for this day.

The second data set we use is the proportion of the population of each country that are classed as Internet users. There are three possible sources for this data, the World Bank, the International Telecommunications Union (ITU) and the CIA World Factbook. We use the ITU data by preference, but the three data sets are well correlated in any case.

The combination of this data gives us an estimate of the current Internet user population per country. It should be noted that this is not the number of "subscriptions" to a service, as it attempts to include the number of users behind each subscription. It also is supposed to avoid "double counting", so where a user is part of a broadband service and also has a mobile service, then the user is still only counted once as an "Internet user".

The third component of the data is the ad presentation data of the APNIC measurement program. We use Google Ads to deliver some 25M individual ad impressions per day. We use the Maxmind geolocation database to map each user who received an ad impression to a country, and use a local default-free BGP routing table to also map each user to their "home" network. At this point we have now assembled a set of "home" networks (origin AS numbers) and the geo-located country for each presented ad.

Assumptions

Here we make two major assumptions. Both assumptions are somewhat questionable, but we've been forced to make them in the absence of generally available data.

The first assumption is that Google's ad placement algorithms apply to all users within a given country uniformly. In defining the ad campaigns, we attempt to make the placement definitions as generic as possible, so that within each country the ad placements are roughly equivalent to a random sampling drawn from all users in that country. The implication of this assumption is that if an ISP has twice the number of users than another ISP in the same country, then its users will receive twice the number of ad impressions. This could be stated as: The distribution of ad placement and the distribution of users across ISPs are assumed to correlate.

The second assumption is that each user uses a single ISP for Internet access. This is not necessarily the case. For example, a user may use a local mobile service provider for their mobile Internet access and Starlink for their broadband access. We also have a user in their workplace using their workplace's ISP and using a consumer ISP when they are at home. We are not able to account for such situations and in uniquely assigning each user to a single ISP in a country we tend to underestimate the user count for each ISP in consequence.

Due to the uncertainties the follow from these assumptions, the results we generate have an inevitable level of uncertainty. Some isolated comparisons of this data against other sources where we have access to ISP market share data in individual countries point to an overall level of uncertainty of around 20% or so in our estimates of users per ISP. Large consumer ISPs are still reported as having a large user population in the generated data, but the data for small networks is very uncertain.

The assumption of uniform distribution of ad placements across all ISPs within each country tends to fail where the number of placed ads in relation to the per-country user population is low. The best current example of this can be seen with the Russian Federation, where ad placement in this country has plummeted since February 2023 (a consequence of the hostilities between the Russian Federation and the Ukraine and associated western sanctions being placed on Russia).

The data for Norway highlights another assumption, namely that browsers do not use proxies. In the case of Opera this is not the case, and Opera performs many of the fetches from its own servers on behalf of Opera users. The result is that the system assumes that AS39832, the Opera AS, is the largest ISP in Norway, some four times the size of the next largest ISP, Telenor. (This Opera result is of course completely wrong, and I should remove Opera's AS from this data set!)

There is another assumption around the day of the week, and for holidays, where the analysis assumes that every day is much the same, whereas on business days the ad presentation into work-related ISPs is far higher than the presentation rate for the same ISPs in weekends and holidays.

As this is a measurement based on the placement of ads, the use of so-called "ad-blockers" can disrupt this measurement. Our assumption here is that like the ads themselves, the use of ad-blockers is also relatively uniformly distributed across all users in the country.

Conclusions

It's frustrating that this information is not generally collected in annual filings for national regulatory agencies and not collated internationally by the ITU-T, and this frustration has motivated us to use our measurement data to push out our estimates as a public data set. The conclusion from the recent RIPE presentation is that this method of estimation of the number of users for each ISP works well in countries with sufficient Google Ads presentations, a conclusion which correlates with our own experience in running this measurement for many years.

On the other hand, the generation of this data is based on a number of sweeping assumptions, which I've noted here, and numbers should be treated with some level of caution.

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.

Author

Geoff Huston AM, M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.

www.potaroo.net