IPv6 Data Collection and Anonymization Mask Length - Draft v1

(DRAFT) IPv6 Netflow Anonymization Policy v0.1 (DRAFT)

Author: ***Joe St. Sauver***

I. Introduction

As IPv6 becomes "more real," Internet2 should begin to collect and make available IPv6 netflow data for research and analysis much in the same way it currently makes anonymized IPv4 flow data available.

Flow data is useful for operational purposes, but also for performance studies and for security research, and if we don't have IPv6 netflow data available, it will be impossible to treat IPv6 traffic on par with IPv4.

At the same time, we have a legal and ethical obligation to respect the privacy of customer traffic. Thus, any netflow data collected from the Internet2 backbone must be anonymized before it is released for use by the community.

II. Current IPv4 Data Anonymization Policy

Currently, IPv4 data is sanitized by Internet2 by having the low order 11 bits of each IPv4 address zeroed before data is released for analysis, leaving 21 bits of each 32 bit IPv4 address intact.

For context, most sites have subnets somewhere in the /23-/25 range, which means that in general while it IS possible to use the masked IP addresses to tie a given netflow record to an institution, it is NOT possible to localize IPv4 data down to a unique subnet given data sanitized with an eleven bit mask.

III. Proposed IPv6 Data Anonymization Policy

In general (e.g., see https://www.arin.net/policy/nrpm.html at 6.4.3 and 6.5.4), the expectation is that:

-- LIRs will receive a /32 from ARIN (or RIPE, APNIC, etc.)
-- large sites will receive a /48 from of their LIR'S /32 while
  small sites needing only a few subnets over 5 years will get a /56
-- if one and only one subnet is neeed, you'll get a /64.

In keeping with the IPv4 policy, which allows data to be localized to a particular site, but not to a specific subnet, we'd need an IPv6 mask that will zero somewhere between:

-- the low order 65 bits (zero the low 64 to at least anonymize at the
  subnet level, then add one bit to insure that the subnet isn't unique)
-- and the low order 76 bits (assume we're going to anonymize an entire
  /56 assigned to a small site)

Within that range, I believe a 65 bit mask would potentially fail to adequately preserve customer privacy while a 76 bit mask would result in "over-anonymization," unnecessarily reducing the operational and research value of the resulting anonymized flow data.

Recommendation: Looking for a value between those two bounds, let's "split the difference" and zero the low order 69 bits, leaving the high 63 bits available for analysis.

IV. Discussion

Is 69 the right value?

By zero'ing the low order 69 bits, we insure that there will be at least 2^69 (or 590,295,810,358,705,651,712) potential IPs which might map to any IP address appearing in an anonymized flow record.

Similarly, there would be potentially be (69-64=5) 2^5=32 distinct possible subnets associated with any IP address appearing in any anonymized flow record.

That provides substantial protection against accidental invasion of any individual IPv6 user's privacy.

What if we view things from the other direction, from the perspective of a potential data analyst?

Zeroing the low order 69 bits means that for a given /56 allocation, there would be 63 bits for analysis, and thus a site that has a /56 will potentially "hash" or "bin" into (130-69-56=5) 2^5=32 unique chunks.

Similarly, if a site has a /48, that implies (130-69-48=13) 2^13=8,192 potential unique analysis "chunks" for a site of that size.

V. Conclusion

A 69 bit mask balances both the need for privacy, and the need to have IPv6 netflow data available at finer-than-site-by-site granularity for operational and research analysis purposes. Netflow data released for use by the community should first be anonymized by zeroing the low order 69 bits of all IPv6 addresses in that data.