Measuring Web Centralization

DINRG Interim Meeting: Centralization on the Internet

June 03, 2021

Trinh Viet Doan (doan@in.tum.de)
Technical University of Munich

In the last decade, the Internet and the Web have been perceived to become increasingly centralized, as most of the traffic is delivered by so called "hypergiants". These hypergiants typically provide Content Delivery Infrastructures (CDIs) around the globe, covering cloud solutions as well as CDNs, among other distributed infrastructures such as public DNS services or content caches. Additionally, studies have shown that CDIs have moved very closely to the edge, resulting in a flattening of the Internet topology. Thus, such CDIs have become integral and potentially irreplaceable parts of the Internet ecosystem, being able to provide benefits in terms of availability, performance, and security to their end users. Due to these evolutions, the networking community has expressed increasing interest in (empirically) studying the motivation, extent, and effects of such centralization trends in the Internet ecosystem from technical, societal, and legal points of view. However, meaningful methods and metrics to understand Internet centralization (plus more empirical evidence) are lacking.

In light of this, we conduct a series of studies and use *CDI penetration* as one possible metric to investigate Web centralization in particular: Using a set of DNS measurements for `www.` subdomains of three Top-Level Domains (TLDs) (`.com`, `.net`, `.org`) and for Alexa Top 1M, we calculate the CDI penetration as the number of CDI-hosted websites relative to the total number of websites. To identify the number of websites hosted by CDIs, we use a heuristic based on CNAME records and IP-ASN mappings. For the dataset of the three TLDs, we find that CDI penetration has nearly doubled from around 8.2% (2015, 140M domains in total) to roughly 15% (2019, 165.5M domains) over the course of five years, which suggests a significant trend w.r.t. increasing centralization. Regarding the Alexa Top 1M measurement data, we observe that 23.4% of the pages are hosted on CDI over IPv4 as of 2019; in contrast, the CDI penetration over IPv6 is significantly higher with 81.9%. Similarly, when distinguishing by popularity based on the Alexa ranks, we notice that CDI penetration is higher for more popular websites. In particular, we see that for both datasets only a small set of CDIs/hypergiants (namely Cloudflare, Google, Amazon, Akamai, Fastly, and Microsoft) serves most CDI-hosted webpages. Overall, these observations indicate that these well-provisioned CDI host a large number of (popular and frequently visited) websites while also offering support for new protocols (and security extensions).

However, considering our findings, some questions are still left for discussion and future investigation: What other metrics can be used to determine and measure Internet and Web centralization? What features motivate the usage of centralized services? How dependent are users on such centralized services? What are tradeoffs between centralized services and decentralized alternatives?