Measuring the Internet's edge with Illuminati

Martin Casado and Michael J. Freedman
http://illuminati.coralcdn.org/

Internet systems increasingly rely on a client's network attributes to inform security- and performance-sensitive decisions. There are many examples: Some content distribution networks (CDNs) use the public IP address of clients' DNS resolvers to choose a server replica "close" to the client, while others rely on a client's public IP address. Financial services help minimize fraud by detecting anomalies between their client's expected and current locations. Many news and academic journal services perform admission control for institutional subscribers by explicitly whitelisting IP ranges.

Unfortunately, the Internet's edge is darkening. NATs, proxies, and other indirection points increasingly occlude a server's view of its clients. And while it is generally understood that these obstacles continue to chip away at edge transparency, there has been little work to quantify their effect on the accuracy of decisions based on public network parameters.

These questions are not merely academic. We have personal experience with clients using CoralCDN to circumvent admission control based on the public IP address's country. If clients of CDNs using DNS redirection use resolvers that are actually quite distant from them, these systems can make bad server selections, resulting in poor performance. If an online advertiser blacklists an IP address because of suspected click fraud, they can reject legitimate traffic from multiple users behind the same NAT or proxy. Similarly, if clients use web proxies, they may circumvent locality-based access restrictions that enforce regulatory compliance, *e.g.*, to blackout major-league baseball games within certain regions or to prevent access to Nazi memorabilia in France.

In this paper, we present the methodology and results of our efforts to bound the accuracy one can expect when inferring the location or uniqueness of clients given a set of HTTP requests. In particular, we ask two central questions, posed from the viewpoint of a server operator:

- *How well can a client's IP address and other public information serve as a unique notion of identity?*
- *How well can one predict a client's location, based on its IP address and public parameters?*

In addition, in the process of answering these questions, we discovered a technique for estimating the distribution of NAT'd network sizes. We also present a method for detecting unique clients given multiple requests from the same public IP and an estimation of whether they are behind a proxy or NAT.

Our project, hereafter called *illuminati* for its attempt to "illuminate the shadows of the Internet," opportunistically measures oblivious web clients, their proxies, and their DNS resolvers. We will present an analysis of data collected over a period of six months from, to date, more than 6.5 million unique hosts from 213 countries worldwide.

To address the question of locality, we determine the location of IP addresses using a commercial-grade geolocation database from Quova. We associate clients with their resolvers using modified DNS nameservers and web servers that synthesize random URLs per client request. We determine the IP addresses of clients behind web proxies (and thus their actual locations) by having clients run a specialized Java applet (in their unmodified web browsers).

To address the question of identity, we use a number techniques including local-vs-public IP address comparisons (via a Java applet), SYN fingerprinting, and HTTP header analysis. Our results indicate that while most hosts are behind NATs—73.5% of those we measured—the majority of NAT'd networks are comprised of only a few hosts. In fact, our measurements suggest that the sizes of NAT'd networks follow an exponential distribution.

Additionally, this applet enables us to differentiate between non-NAT'd hosts and those behind middleboxes. We use this information as "ground truth" to develop passive heuristics—based only on HTTP headers and SYN fingerprints—by which a server can detect unique clients with the same public IP address and whether they are behind a NAT or proxy. We find that, in general, proxies serve significantly more clients than standard NATs.

We will demonstrate our collection architecture operating in real-time, how we collect various information about our clients, and a map-based visualization of proxies and the location of clients they serve.