# Skylines: Demystifying Network Resource Islands with Virtual Landmarks


Marc Anthony Warrior
Northwestern University
warrior@u.northwestern.edu

Romain Fontugne
IIJ Innovation Institute
romain@iij.ad.jp

Randy Bush
randy@psg.com



## ABSTRACT

Existing mapping systems and measures used to describe Internet location fail to account for complexities introduced by DNS-based content allocation schemes. With DNS redirection, in particular, clients residing in the exact same network location may potentially be served by completely different content sources. Likewise, distant clients may be directed to the same servers. The existence of such scenarios challenges the validity of some common assumptions about network behavior: the relationship between a client's network location and its "view" of the Internet, independent of performance, is neither predictable nor consistent across clients.

In this paper, we introduce skylines, a distance measure that describes variation in Internet resource allocation between clients. TODO: finish list of contributions

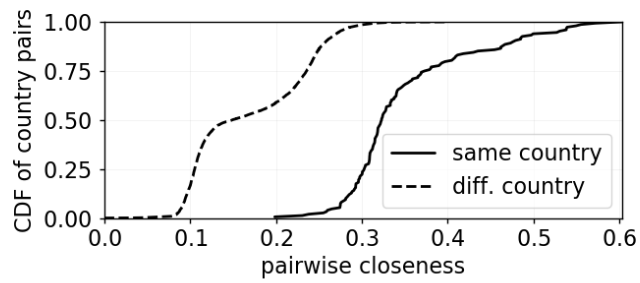




**Figure 1: Diagram illustration resource selection**

## 1 INTRODUCTION

"What's in a name?" — in the context of networking, quite a lot! IP addresses and prefixes indicate to routers where particular interfaces or subnets reside in the greater Internet [**?** ]. Domain names offer a simple, human-readable overlay for the complicated, often multiplexed addressing scheme underneath [**?** ]. Autonomous system (AS) numbers often help to distingish one network from another [**?** ]. Names and labels such as these, whatever form they take, allow us to organize immense network spaces with manageable and descriptive abstractions.

Unfortunately, a single name is often not enough. One machine can be described in terms of all of the examples listed above, and more. This is because it is important that names carry information relavant to the specific way in which they will be used. Routers are unable to direct traffic with domain names, just as humans cannot be expected to remember the plethora IP addresses they access every day. While it is intuitive that no labeling system is applicable across all possible dimensions, in practice, this fact is often taken for granted or neglected entirely.

In this project, I challenge the careless application of conventional client labeling schemes in Internet measurement experiments, particulary those subject to the effects of DNS redirection and CDN PoP (point of presence) catchment formation. I uncover and quantify the degree of misalignment between experimentally determined aggregate catchments and labels often assumed to indicate "similarity" between clients — country, AS number, announced BGP prefix, and /24 subnet — and use this to design the Skyline model, a grouping system which describes clients' relative distances from each other in terms of their *common network resource exposure* (CNRE). Describing clients in such terms highlights what should be a chief concern in large scale Internet measurement platforms and network optimization efforts: the sets of clients actually being directed the same resources.

In order to develop the Skyline model, I performed an exhaustive set of measurements to frame client experience on a per *site* basis. In my completed, preliminary work, I capture a snapshot of both DNS resolutions and latency measurements toward the 304 domains that appeared most frequently in the top 2441 most popular webpages. My measurements span over 10,000 unique clients spread across 185 countries and 3637 autonomous systems. I performed over 52 million pairwise comparisons with the results of these measurements to arrive at the foundation of what I have coined the "Skyline model".

With this project, I make the following contributions, including those I expect to stem from proposed work, which I have designated with ($p$)

- I perform a large exploration of client network performance on a per webpage level. My raw results are publically available on the RIPE Atlas platform.
- ($p$) I quantify the degree of misalignment between conventional grouping schemes and aggregate catchments.
- ($p$) I introduce the Skyline model, a client grouping scheme that reflects the extent of CNRE.

- (*p*) Using the Skyline model, I identify and analyze network resource islands — sets of clients with very high degrees CNRE.

## 2 PROBLEM SPACE AND RELATED WORK

This projects aims to gain an understanding of which clients are directed to the same set of resources across many distinct domains. Its most direct and immediate use case is influencing probe selection in large scale Internet measurements. For researchers, likely unaware of the relatively hidden allocation schemes of the wide array of CDN platforms and other large content distributors, it is difficult to determine, a priori, the degree of similarity between clients. Knowledge of whether there is a high probability that a pair of clients are being directed to altogether different resources may be significant to their experiment design. This approach to experiment design is in line with RIPE Atlas, one of the largest client based measurement platforms, which maintains an exhaustive set of tags on all of their clients in order to help researchers and network operators filter and refine the set selected for their experiment [**?** ]. Further, more abstract applications may include, but are not limited to, distributed denial of service mitigation [**?** ] and CDN node deployment [**? ?** ].

The most similar body of related work involves anycast CDN catchment analysis, which aims to investigate the set of clients routed towards particular CDN points of presence (PoPs) [**? ? ?** ]. My work differs significantly in scope: to my knowledge, I am the first investigate what I refer to as *aggregate catchments*, the joint behavior of many anycast CDN catchments and unicast CDN targets, spread across many content distribution platforms. Conversely, this related body work either focuses on individual platforms or specific services [**? ? ?** ].

Several authors have attempted to discover the topology of large CDN platforms through large scale measurement studies [**? ? ?** ]. While their findings are potentially of use in this project, their goals and contributions run parallel to what I aim to accomplish. They seek to identify the properties and locations of CDN resources; conversely, I seek to identify the target pools (sets of clients) of overlapping CDN resource catchments [**? ? ?** ]. Other work close to this space investigates the performance of a particular CDN deployment scheme [**?** ].

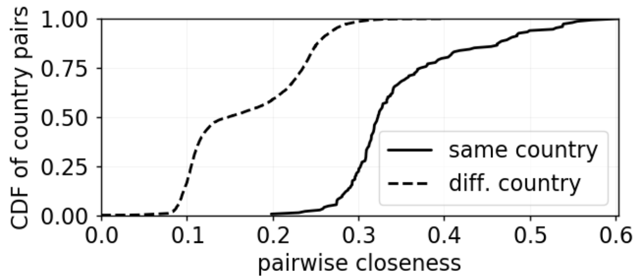



**Figure 2: diagram illustrating how domains were collected**

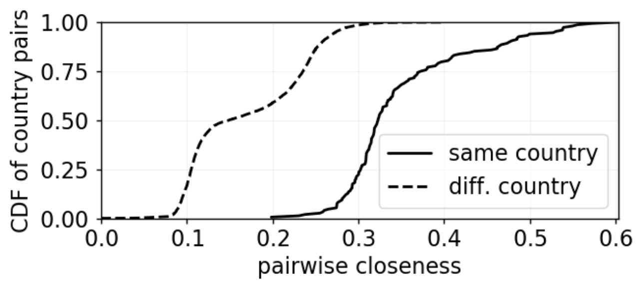



**Figure 3: # of sites touched vs # domains used (done)**

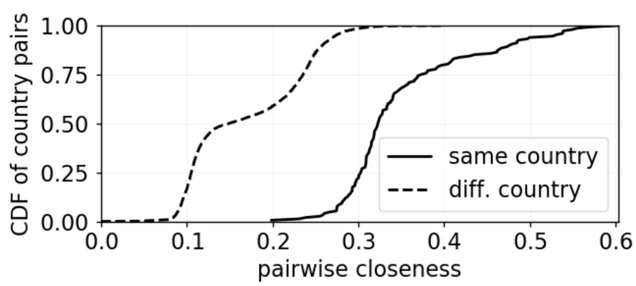



**Figure 4: % links covered per site vs # domains used (done)**

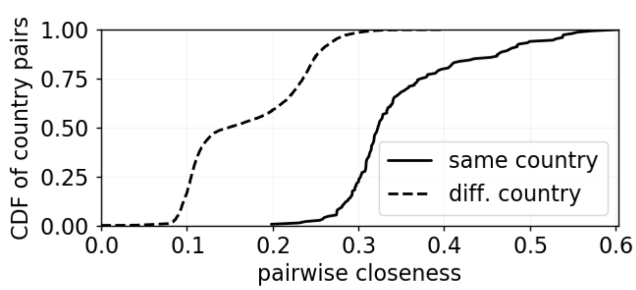



**Figure 5: diagram illustrating CRNE**

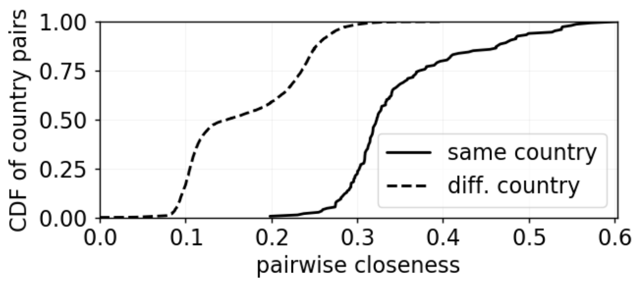



**Figure 6: "High" closeness (90th percentile?) vs # domains measured**

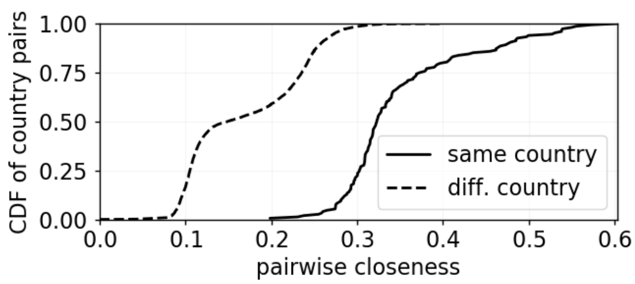



**Figure 7: Mean domain error vs # of distinct answers observed from domain (done)**

# 3 EXPERIMENT OVERVIEW

# 4 COMMON NETWORK RESOURCE EXPOSURE

# 5 FINDING CLUSTERS

# 6 CLUSTER ANALYSIS

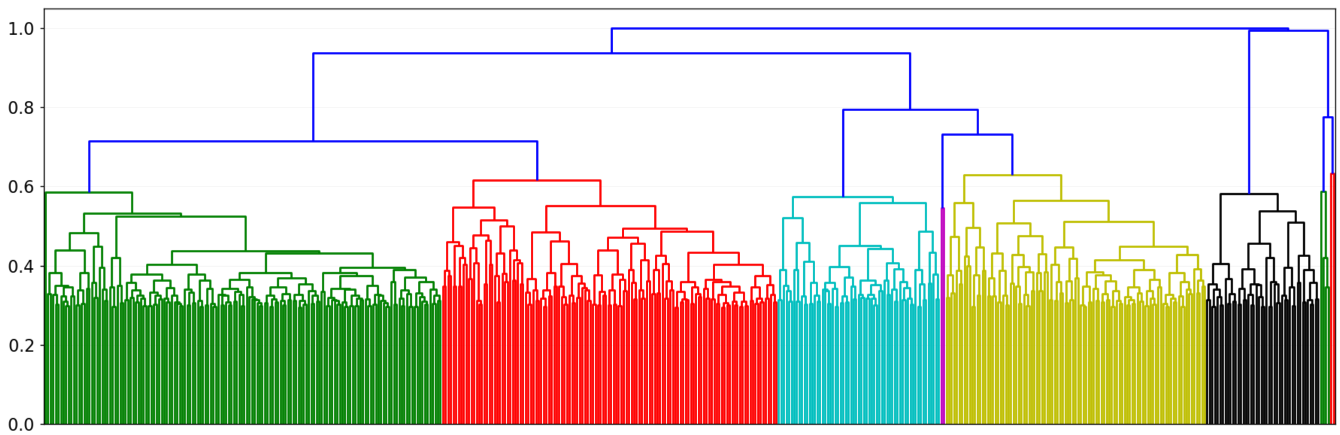
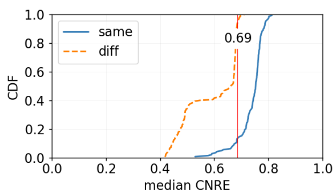
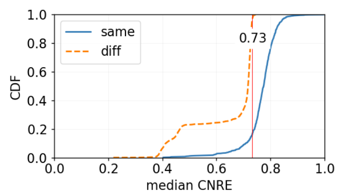


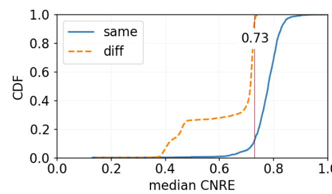

Figure 8: Dendrogram of CNRE distance across all client pairs



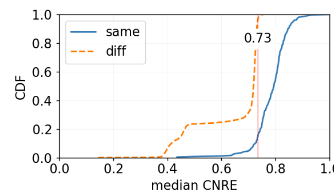

(a) country




(b) ASN




(c) BGP prefix




(d) IP /24 subnet

**Figure 9: CDFs of CNREs across client sets with matching (same) and non-matching (diff) labels. "Same" shows the CDF for the median CNRE distance across all client pairs matching a given label. "Diff" shows the CDF for the median CNRE distance from each label group toward all other labels.**

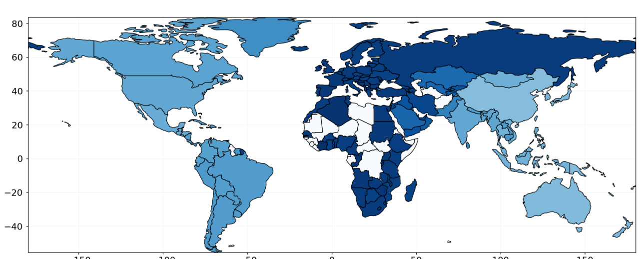
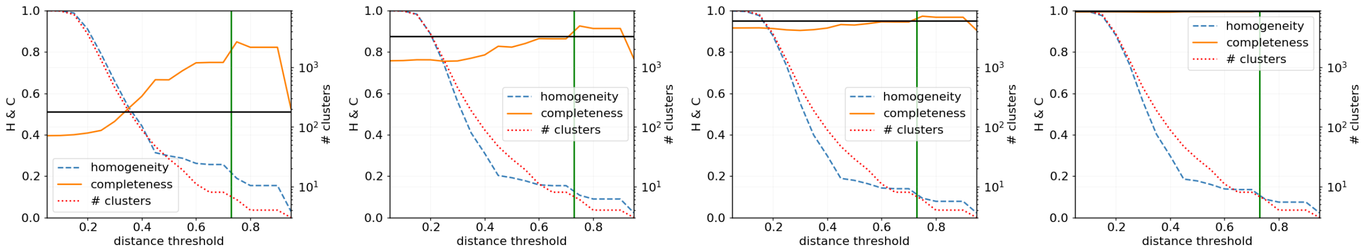


**Figure 10: Choropleth with each country shaded by its median CNRE distance from all other countries. TODO: add legend (I'm struggling to format it in geopandas...)**




**(a) country** **(b) ASN** **(c) BGP prefix** **(d) IP /24 subnet**

**Figure 11: Completeness and homogeneity (for each feature) vs clustering distance threshold. The vertical line marks 0.73, the CNRE distance at which clients with different labels become distinguishable.**

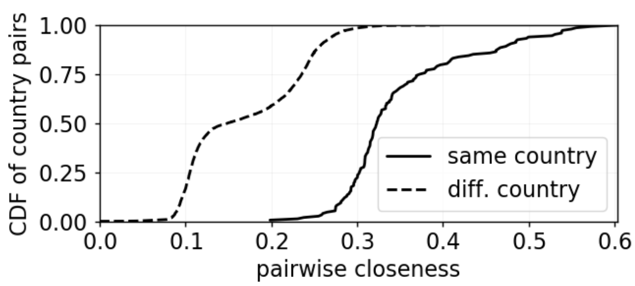



**Figure 12: CDF of # of outliers per cluster for geo distance, etc**

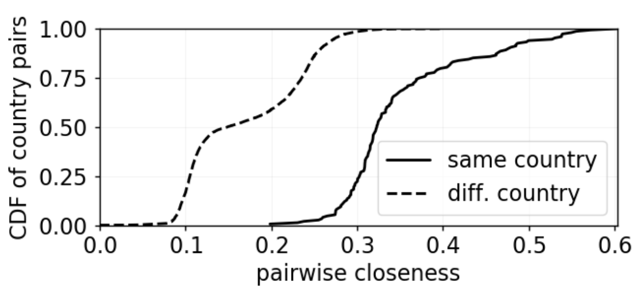



**Figure 13: Domain match alignment with clusters (what format?)**

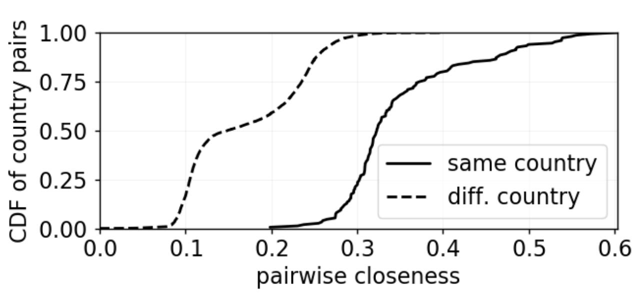



**Figure 14: Map of world with dot for each client and dot colors corresponding to each cluster**