Dissecting the Workload of a Major Adult Video Portal

Andreas Grammenos **Aravindh Raman** Timm Böttger

Zafar Gilani **Gareth Tyson**

Computer Lab / Alan Turing Institute

Introduction

- Streaming Content today is extremely popular
  - Popular services like Youtube, Netflix, and other thrive
    - Well studied over the years – predictable traffic patterns.
  - There is however, another side: Adult Video Streaming…
    - Lack of understanding of traffic type and patterns.

Methodology

- This paper: Present a large-scale analysis of access patterns for a major adult website
  - Focusing on understanding how individual viewer decisions (dubbed “journeys”) impact the workload observed.
- Gathered very granular data from a popular CDN – key points:
  - 1 hour of access logs for resources hosted served by the site.
  - 62K Users.
  - >20M access records.
  - >3TB of exchanged data.

Methodology (cont.)

- Web Scrape data
  - Retrieved metadata from source site based on CDN logs – i.e.:
    - Associated Categories
    - Associated hash-tags
    - View counters
    - Like/Dislike ratio
- In total we gathered metadata for near 5m videos covering over 91 % of the total requests observed in our CDN logs

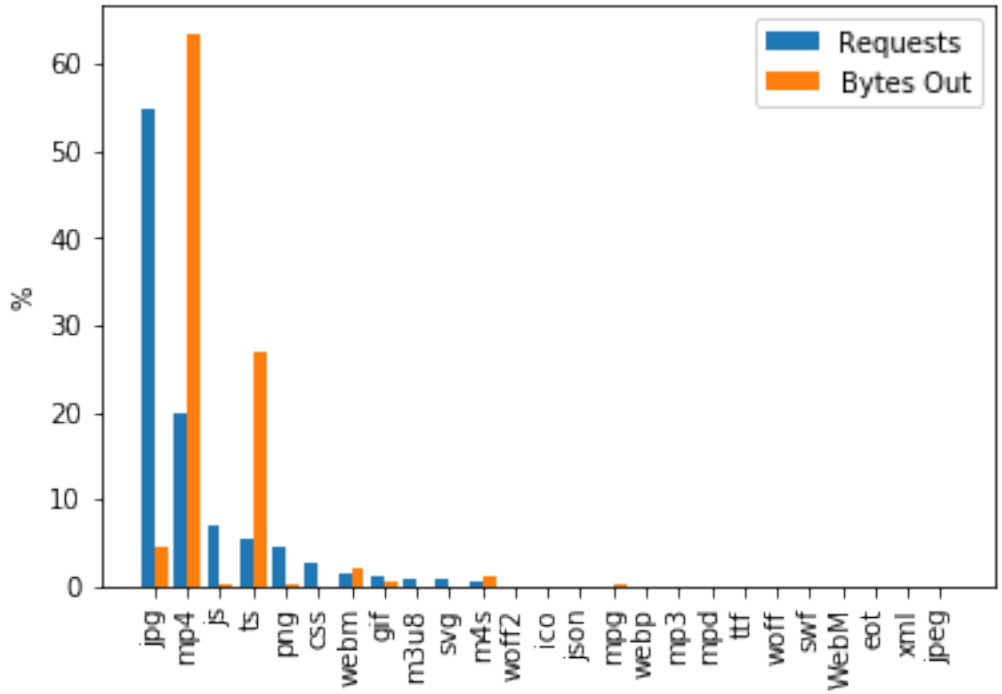
Characterization

- Initially, we perform a basic characterization of the following:
  - Corpus served
  - Overall site workloads (at the CDN level).

- We characterize the following:
  - Resource Type
  - Video Duration
  - View Counts
  - Category Affinity

Characterization: Resource Type

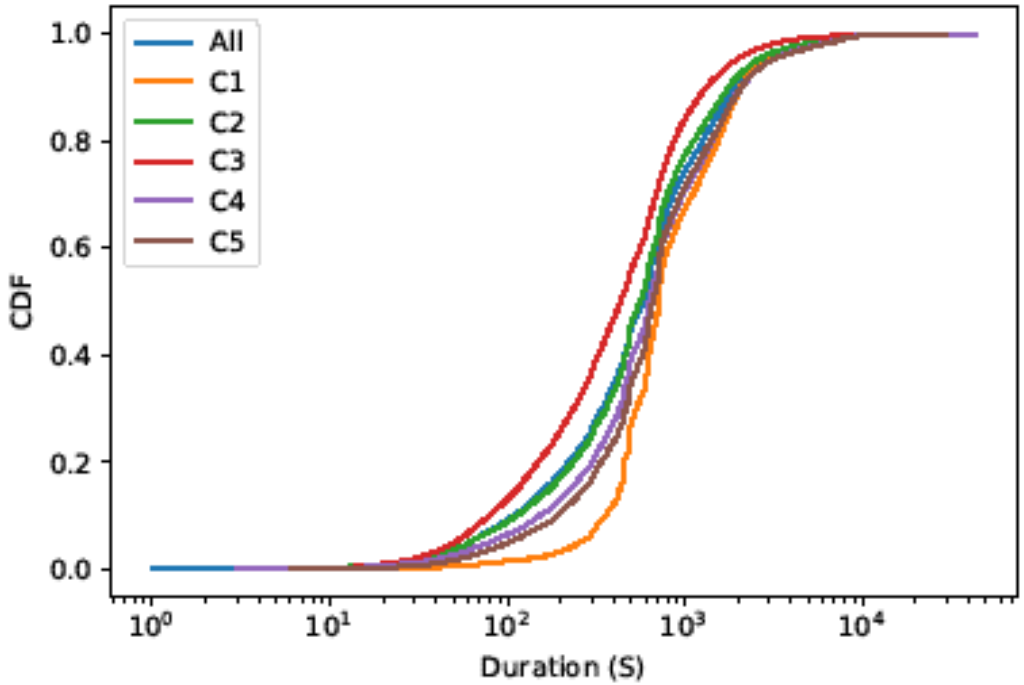
- Web sites consist of wide range of media – let's see in our case.



Fraction of requests to each resource type - shows distributions for both number of requests and number of bytes sent by the servers

Characterization: Video Duration

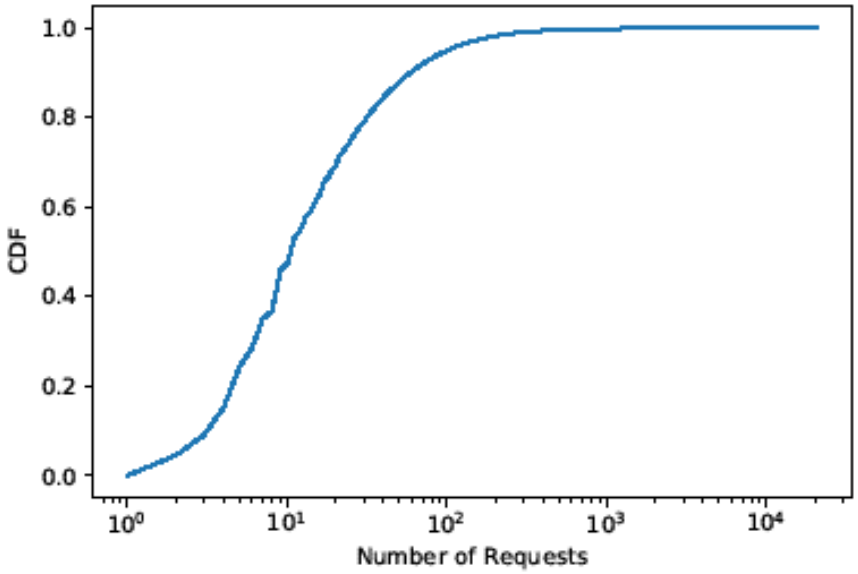
- Consequence: most accesses are driven by *non-video* content consumption



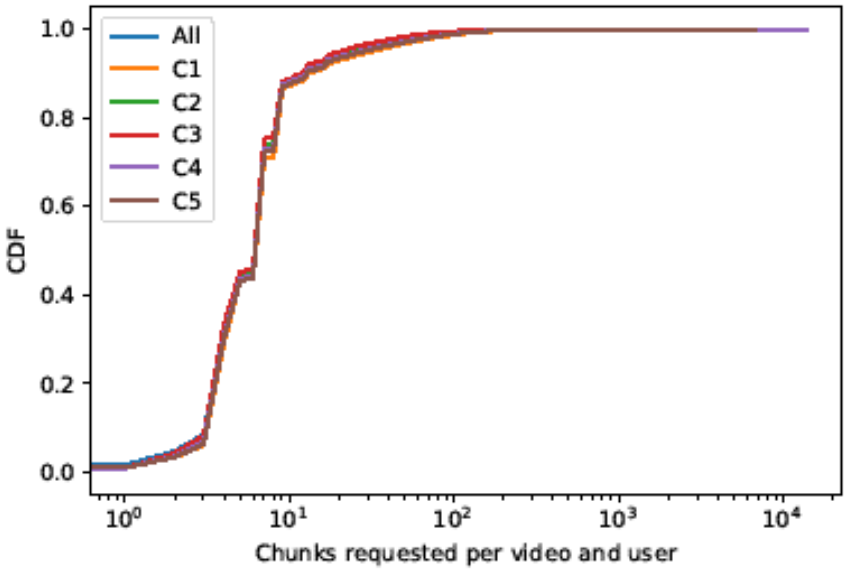
CDF of consumed video duration based on category using All and top-5 categories. Note "All" refers to all content within any category.

Characterization: View Counts

- Explore the popularity distribution of the resources, within our logs



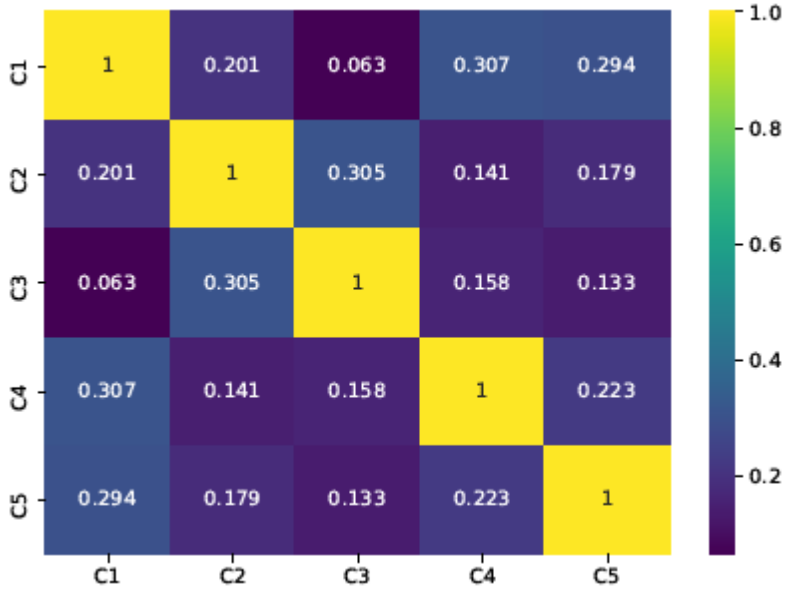
CDF Number of requests per object



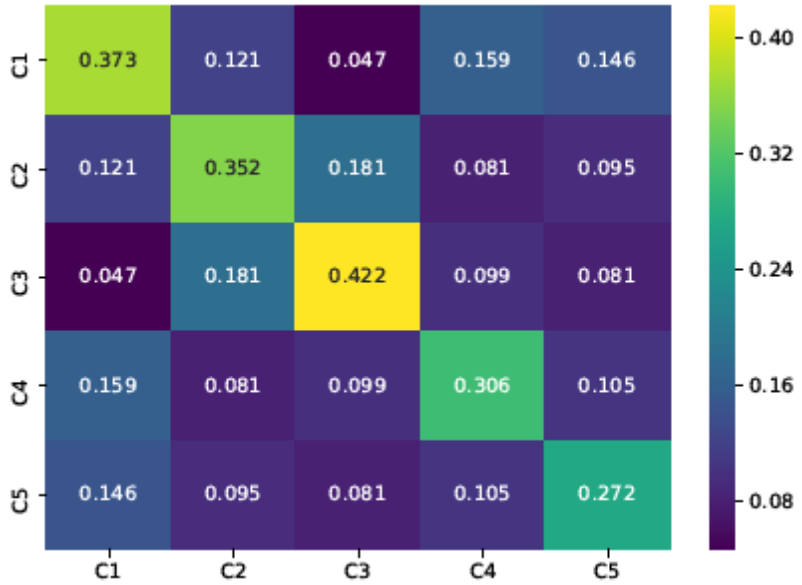
Distribution of video chunk per video request

Characterization: Category Affinity

- Subtle differences exist – exploring category co-location.



Heatmap showing the fraction of the pair-wise coexistence for the 5 most popular categories



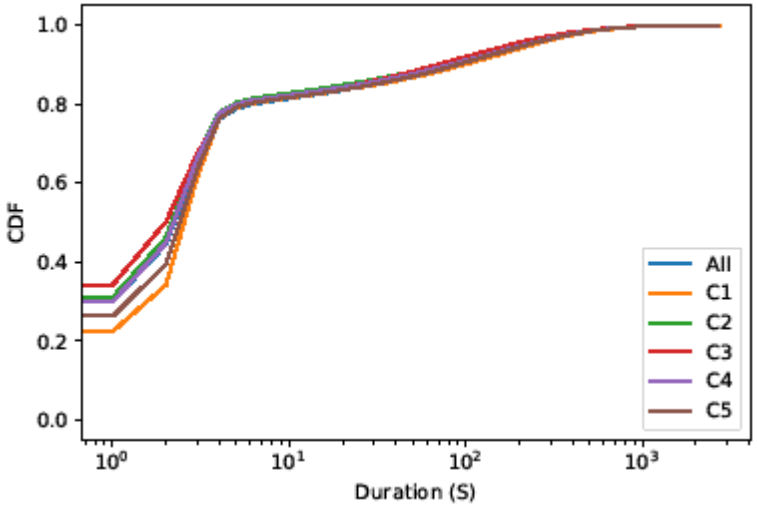
Heatmap normalised by the total number of videos(across all categories)

Characterization: Per-Session Journey

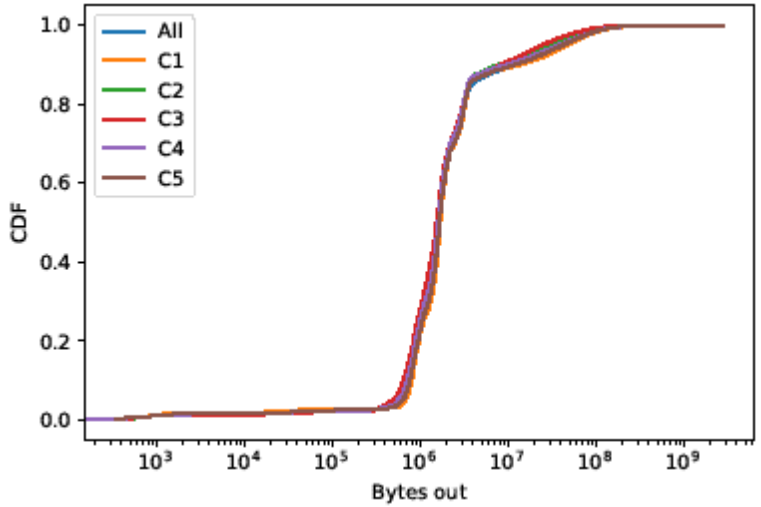
- Seen so far – workload dominated:
  - Image and video content
  - Patterns which suggest that users *rarely* consume *entire* videos
- Dive into individual sessions ("journeys") – we inspect:
  - Intra-Video Access Journeys
    - how users move between chunks within a single video?
  - Inter-Video Access Journeys
    - How individual sessions move between videos?

Intra-Video Journeys: Access Duration

- We explore the time each user dedicates to an individual video*.



CDF of the approximate consumption for each individual video across sessions for all and top-5 categories

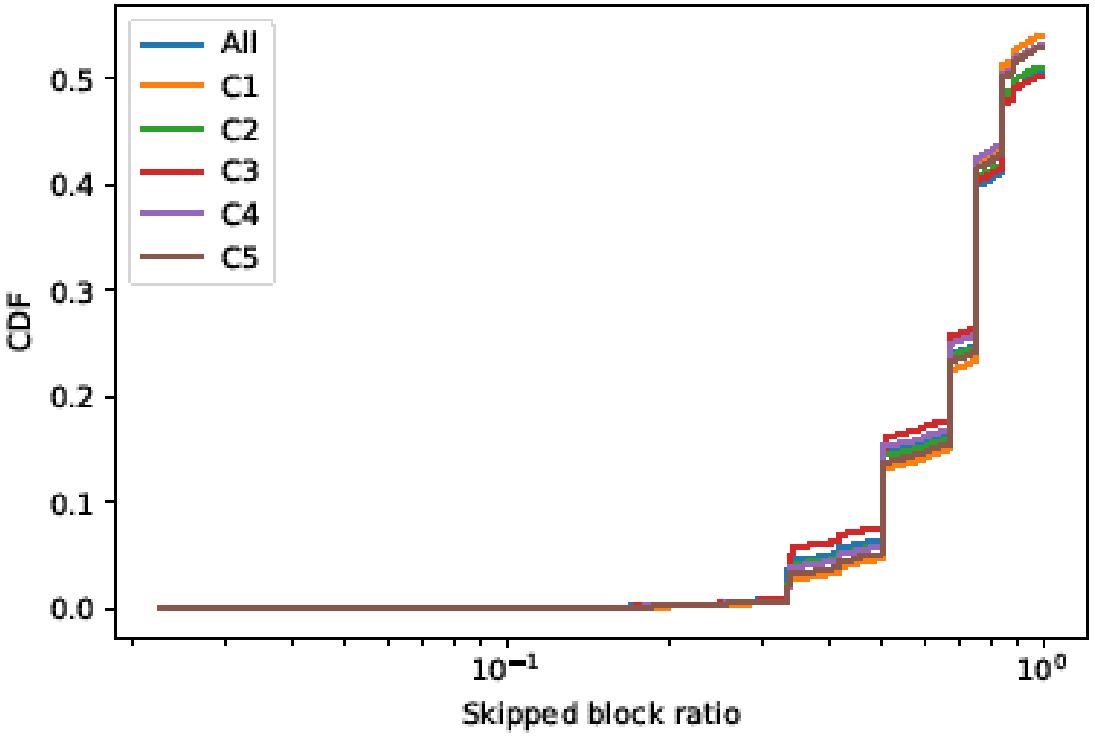


CDF of the bytes out per User/Video combination for all and top-5 categories

*Note that this is different to Figure in slide 7, which is based on the video duration, rather than the access duration.

Intra-Video Journeys: Cancellations and Skip Rates.

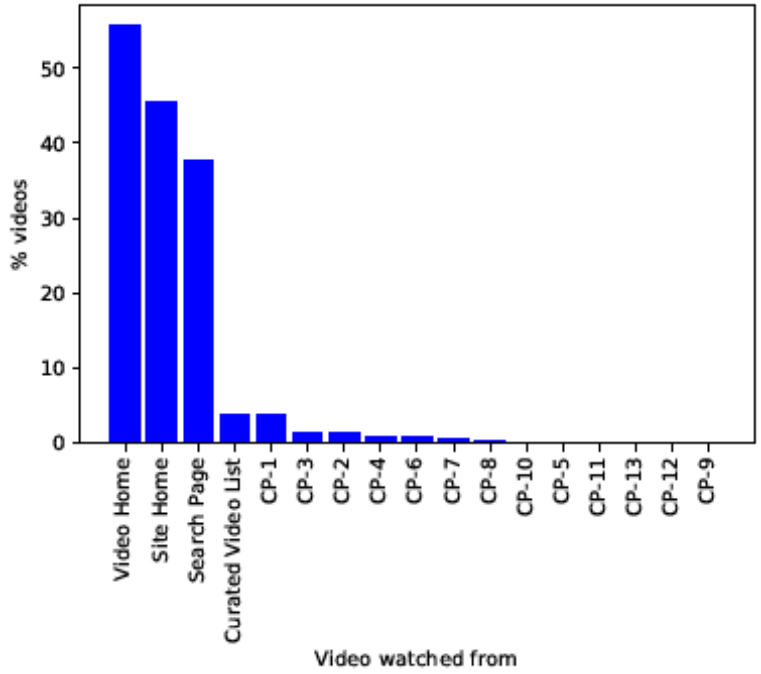
- Implications of incomplete video download: users *skip/cancel*!



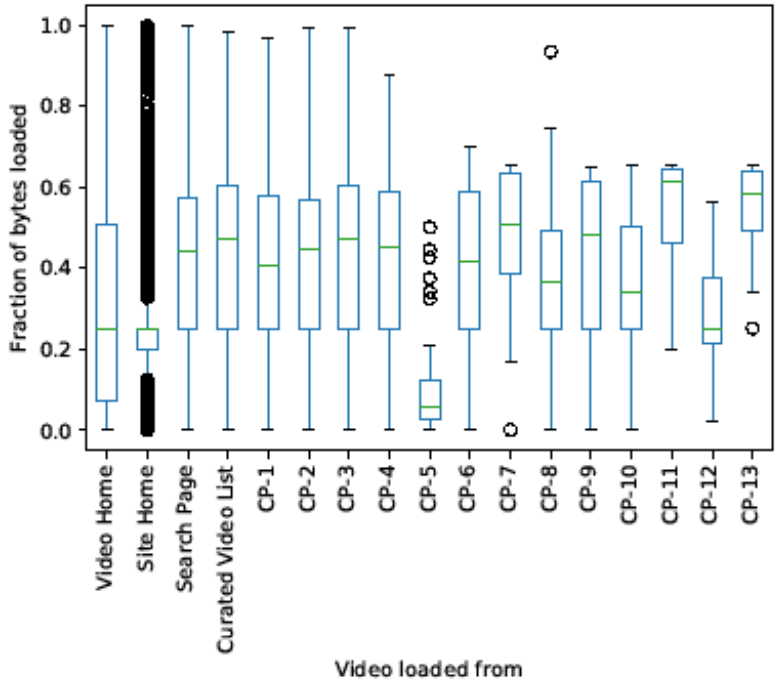
Skipped blocks for each category

Inter-Video Journeys: Video Load Points.

- Video Load Points – vantage points for website traffic origins?



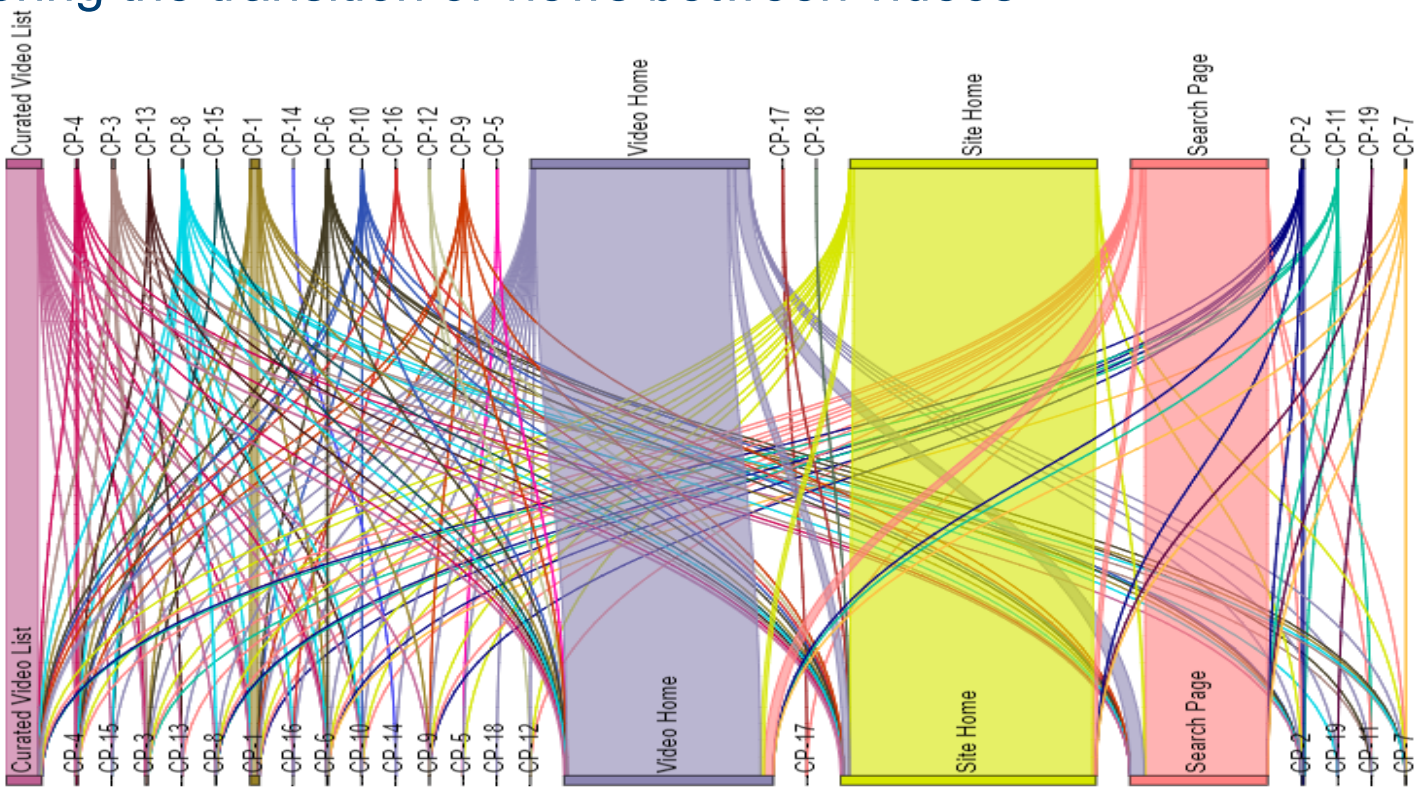
Where the videos are watched the most: >95% of videos are watched from either the main page of the video, homepage of the site, and



Where the videos are loaded the most: Y-axis gives the ratio of bytes out and total file size across users from various pages

Inter-Video Journeys: Inter Video Navigation

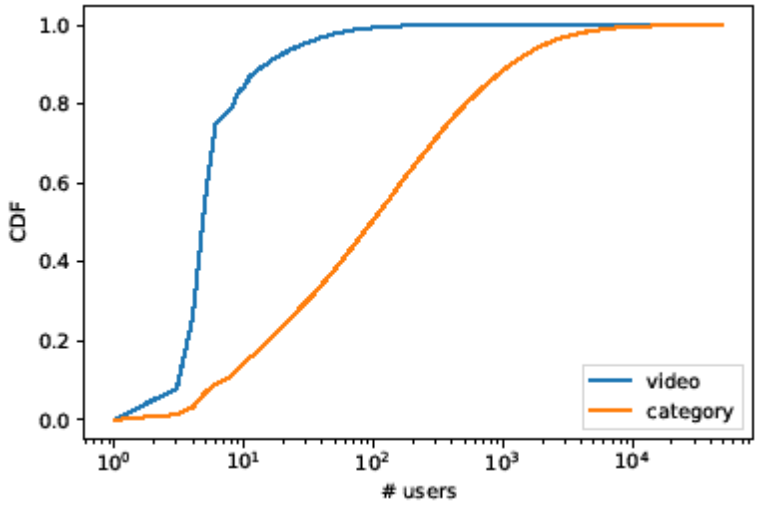
- Exploring the transition of views between videos



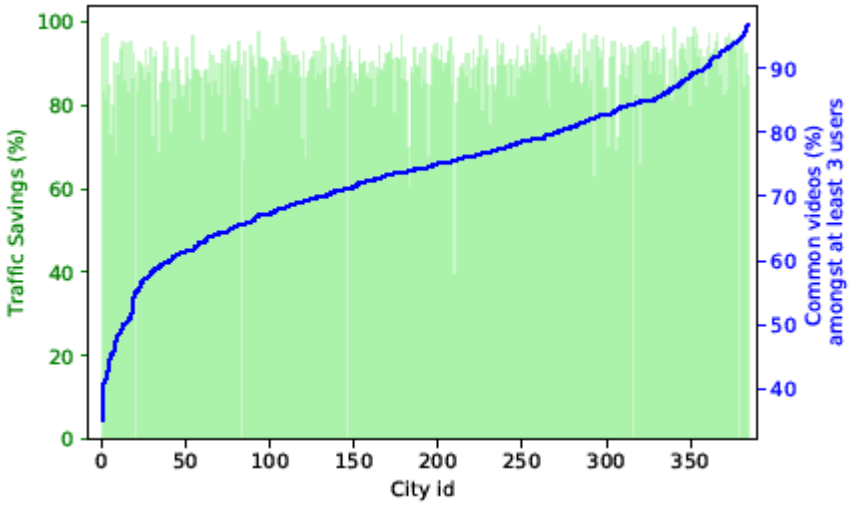
Sankey Diagram showing transitions between videos

Implications: Save BW

- Geo-Aware Caching: Can result in significant bandwidth savings!



CDF of number of users who have watched the same video in their city (blue) or a video from the same category in their city (orange)



Percentage of traffic saved at back-haul by implementing city-wide cache (Y-1) and the percentage of users who would have benefit by the scheme (Y-2).

Implications: Reduce Latency

- Predictive Loading
  - Predicting popular chunks
  - Pre-load thumbnail images as pre-loaded (cached) chunks
  - Explore behavioral trails to direct traffic to different caches – serving different interests.
  - Explore recommending videos on what resides in the cache.

Conclusions

- Concluding
  - Explored the characteristics of the traffic that a large adult-video portal has focusing on understanding in-session journeys.
  - Key take-aways:
    - Bulk of served objects are not videos!
    - Bulk of data is for video!
    - Small percentage of popular content.
  - This is just the start:
    - Possibility to explore/validate results against different portals
    - Develop optimized delivery systems and caching schemes

Questions?