YouTube Popularity Dynamics and Third-party Authentication

Niklas Carlsson
Linköping University, Sweden

Keynote at the 10th IEEE Workshop on Network Measurements (IEEE WNM @LCN), Nov. 2016
YouTube Popularity Dynamics, Edge Caching, Third-party Authentication, and Interactive Videos Streaming
“some topics I am very excited to talk about” ...
The work here was in collaboration ...

- **Including with students (alphabetic order):**
  - Youmna Borghol (NICTA, Australia)
  - Vengatanathan Krishnamoorthi (Linköping University, Sweden)
  - Siddharth Mitra (IIT Dehli, India)
  - Anna Vapen (Linköping University, Sweden)

- **... and non-student collaborators (alphabetic order):**
  - Martin Arlitt (HP Labs, USA, and University of Calgary, Canada)
  - György Dan (KTH, Sweden)
  - Derek Eager (University of Saskatchewan, Canada)
  - Anirban Mahanti (NICTA, Australia)
  - Nahid Shahmehri (Linköping University, Sweden)
Background: Research overview

Design, modeling, and performance evaluation of distributed systems and networks
Design, modeling, and performance evaluation of distributed systems and networks
Background: Research overview

Scalable content delivery

Design, modeling, and performance evaluation of distributed systems and networks
Background: Research overview

Scalable content delivery

Characterization, analytics, modeling

Efficiency and sustainability

Design, modeling, and performance evaluation of distributed systems and networks
Background: Research overview

Scalable content delivery

Characterization, analytics, modeling

Efficiency and sustainability

Network security

Design, modeling, and performance evaluation of distributed systems and networks
Background: Research overview

Scalable content delivery

Characterization, analytics, modeling

Efficiency and sustainability

Network security

Design, modeling, and performance evaluation of distributed systems and networks
In this talk ...
... model+understand popularity ...
... popularity dynamics and caching ...
... third-party authentication ...
Background: Research overview

Scalable content delivery

Characterization, analytics, modeling

Efficiency and sustainability

Network security

Design, modeling, and performance evaluation of distributed systems and networks
... innovative new streaming media ...
So let’s start …
Video streaming landscape
Video streaming landscape
Video streaming landscape

[Image of various video streaming platforms and devices]
Motivation

- Streaming services contribute to over 60% of the global Internet traffic currently.
- By 2020, this share is expected to be over 80%.
- Systems need to be well understood, scalable, and efficient to match growth projections.
The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity

*Proc. ACM SIGKDD 2012.*

Characterizing and Modeling Popularity of User-generated Videos

*Proc. IFIP PERFORMANCE 2011.*
Motivation

- Video dissemination (e.g., YouTube) can have widespread impacts on opinions, thoughts, and cultures
Motivation

• Not all videos will reach the same popularity and have the same impact
Motivation

• Not all videos will reach the same popularity and have the same impact
Motivation

• Not all videos will reach the same popularity and have the same impact
Aside ... Popularity distribution
Aside …
Popularity distribution
Aside ... Popularity distribution

IFIP Performance ‘11
Aside ... Popularity distribution

IFIP Performance ‘11, IPTPS ‘10
Aside ... Popularity distribution

IFIP Performance ‘11, IPTPS ‘10, PAM ’11
Aside ... Popularity distribution

Views (v)

Rank (r)

IFIP Performance ‘11, IPTPS ‘10, PAM ‘11, ACM TWEB
Zipf popularity...

... and long tails
Zipf popularity...
... and long tails

\[ v_r \propto r^{-\alpha} \]
Zipf popularity...
... and long tails

\[ v_r \propto r^{-\alpha} \]

\[ \log v_r = \log v_1 - \alpha \log r \]
Zipf popularity...  
... and long tails

\[ v_r \propto r^{-\alpha} \]

\[ \log v_r = \log v_1 - \alpha \log r \]
Zipf popularity...
... and long tails

\[ v_r \propto r^{-\alpha} \]

\[ \log v_r = \log v_1 - \alpha \log r \]

\[ y(x) = x_0 - \alpha x \]
Zipf popularity...
... and long tails

\[ v_r \propto r^{-\alpha} \]

\[ \log v_r = \log v_1 - \alpha \log r \]

\[ y(x) = x_0 - \alpha x \]
Zipf popularity...  
... and long tails

\[ v_r \propto r^{-\alpha} \]

\[ \log v_r = \log v_1 - \alpha \log r \]
Zipf popularity...

... and long tails

\[ \log v_r = \log v_1 - \alpha \log r \]
Zipf popularity...
... and long tails

E.g., ACM TWEB, PAM ‘11
IFIP Performance ‘11, IPTPS ‘10
Zipf popularity...
... and long tails

- Popularity distribution statistics and models
  - Across services (impact on system design)
  - Lifetime vs current
  - Over different time period (churn)
  - Different sampling methods

E.g., ACM TWEB, PAM ‘11,
IFIP Performance ‘11, IPTPS ‘10
Zipf popularity...  
... and long tails

- Popularity distribution statistics and models
  - Across services (impact on system design)
  - Lifetime vs current
  - Over different time period (churn)
  - Different sampling methods

E.g., ACM TWEB, PAM ’11, IFIP Performance ‘11, IPTPS ‘10
Rich-gets-richer ...

... and churn

E.g., IFIP Performance ‘11
Rich-gets-richer ... 

... and churn

- The more views a video has, the more views it is likely to get in the future

E.g., IFIP Performance ‘11
Rich-gets-richer ...
... and churn

- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary

E.g., IFIP Performance ‘11
Rich-gets-richer ... ... and churn

- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary

E.g., IFIP Performance ‘11
Rich-gets-richer ...

... and churn

- The more views a video has, the more views it is likely to get in the future
- The relative popularity of the individual videos are highly non-stationary
- Some long-term popularity

E.g., IFIP Performance ‘11
Motivation
Motivation

- Some popularity differences due to content differences
Motivation

- Some popularity differences due to content differences
Motivation

• Some popularity differences due to content differences
• But also because of other “content-agnostic” factors
  • The latter factors are of considerable interest but it has been difficult to accurately study them
Motivation

• Some popularity differences due to content differences
• But also because of other “content-agnostic” factors
  • The latter factors are of considerable interest but it has been difficult to accurately study them

*In general, existing works do not take content differences into account . . . (e.g., large number of rich-gets-richer studies)*
Methodology

• Develop and apply a methodology that is able to accurately assess, both qualitatively and quantitatively, the impacts of various content-agnostic factors on video popularity
Methodology

- Develop and apply a methodology that is able to accurately assess, both qualitatively and quantitatively, the impacts of various content-agnostic factors on video popularity.
Methodology

• Clones
  • Videos that have “identical” content (e.g., same audio and video track)
Methodology

• Clones
  • Videos that have “identical” content (e.g., same audio and video track)
Methodology

• Clones
  • Videos that have “identical” content (e.g., same audio and video track)
Methodology

• Clones
  • Videos that have “identical” content

• Clone set
  • Set of videos that have “identical” content
Methodology

• Clones
  • Videos that have “identical” content

• Clone set
  • Set of videos that have “identical” content
Methodology

• Clones
  • Videos that have “identical” content

• Clone set
  • Set of videos that have “identical” content
Methodology

- Clones
  - Videos that have “identical” content
- Clone set
  - Set of videos that have “identical” content
Methodology

- **Clones**
  - Videos that have “identical” content

- **Clone set**
  - Set of videos that have “identical” content

Clone sets allow us to control for content
Methodology

Clone sets allow us to control for content
Methodology

• Analyze how different factors impact the current popularity while accounting for differences in content
  • 1) Baseline: Aggregate video statistics (ignoring clone identity)
  • 2) Individual clone set statistics
  • 3) Content-based statistics
Methodology

Current popularity (e.g., views in week)

Some factor of interest
Methodology

- Focus on clone sets
Methodology: (1) Aggregate model

\[ Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{i,p} + \epsilon_i \]
Methodology: (2) Individual model

\[ Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{i,p} + \epsilon_i \]

(2) Individual model

Current popularity (e.g., views in week)

Some factor of interest
Methodology: (2) Individual model

\[ Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{i,p} + \epsilon_i \]
Methodology: (3) Content-based model

Current popularity (e.g., views in week)

(3) Content-based model

\[ Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{i,p} + \sum_{k=2}^{K} \gamma_k Z_{i,k} + \epsilon_i \]
Methodology: (3) Content-based model

\[ Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{i,p} + \sum_{k=2}^{K} \gamma_k Z_{i,k} + \epsilon_i \]

- **Content-agnostic factors**
- **Impact of content**
- **Scaled measured value**
- **Encoding:** 1 if clone \( k \); otherwise 0
- **Predicted value**
- **Error**
Data collection

- Identified large set of clone sets
  - 48 clone sets with 17 – 94 videos per clone set (median = 29.5)
  - 1,761 clones in total
- Collect statistics for these sets (API + HTML scraping)
  - Video statistics (2 snapshots ⇒ lifetime + weekly rate statistics)
  - Historical view count (100 snapshots since upload)
  - Influential events (and view counts associated with these)
Analysis approach

• Example question: Which content-agnostic factors most influence the current video popularity, as measured by the view count over a week?

• Use standard statistical tools
  • E.g., PCA; correlation and collinearity analysis; multi-linear regression with variable selection; hypothesis testing

• Linearity assumptions validated using range of tests and techniques
  • Some variables needed transformations
  • Others where very weak predictors on their own (but in some cases important when combined with others!!)
Which factors matter?

- Using multi-linear regression with variable reduction (e.g., best subset with Mallow’s Cp)
Which factors matter?

- *Using multi-linear regression with variable reduction (e.g., best subset with Mallow’s Cp)*
Rich-gets-richer

\[ \log P(v) \]

\[ \log v \]
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$

- Linear: $\alpha = 1$ (scale-free linear attachment)
- Sub-linear: $\alpha < 1$ (the rich may get richer, but at a slower rate)
- Super-linear: $\alpha > 1$ (the rich gets much richer)
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$

- **Linear:** $\alpha = 1$ (scale-free linear attachment)
- **Sub-linear:** $\alpha < 1$ (the rich may get richer, but at a slower rate)
- **Super-linear:** $\alpha > 1$ (the rich gets much richer)
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$

- **Linear**: $\alpha = 1$ (scale-free linear attachment)
- **Sub-linear**: $\alpha < 1$ (the rich may get richer, but at a slower rate)
- **Super-linear**: $\alpha > 1$ (the rich gets much richer)
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$

- **Linear:** $\alpha = 1$ (scale-free linear attachment)
- **Sub-linear:** $\alpha < 1$ (the rich may get richer, but at a slower rate)
- **Super-linear:** $\alpha > 1$ (the rich gets much richer)
The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$

- Linear: $\alpha = 1$ (scale-free linear attachment)
- Sub-linear: $\alpha < 1$ (the rich may get richer, but at a slower rate)
- Super-linear: $\alpha > 1$ (the rich gets much richer)
Rich-gets-richer

- The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$
  - Linear: $\alpha = 1$ (scale-free linear attachment)
  - Sub-linear: $\alpha < 1$ (the rich may get richer, but at a slower rate)
  - Super-linear: $\alpha > 1$ (the rich gets much richer)
Rich-gets-richer

<table>
<thead>
<tr>
<th>Slope estimate</th>
<th>Confidence intervals</th>
<th>Hypothesis testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\sigma$</td>
<td>90%</td>
</tr>
</tbody>
</table>

- **Individual**
- **Content-based**
- **Aggregate**

The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^{\alpha}$

- **Linear**: $\alpha = 1$ (scale-free linear attachment)
- **Sub-linear**: $\alpha < 1$ (the rich may get richer, but at a slower rate)
- **Super-linear**: $\alpha > 1$ (the rich gets much richer)
### Rich-gets-richer

<table>
<thead>
<tr>
<th></th>
<th>Slope estimate</th>
<th>Confidence intervals</th>
<th>Hypothesis testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\sigma$</td>
<td>90%</td>
</tr>
<tr>
<td>Individual</td>
<td>1.027</td>
<td>-0.091</td>
<td>0.988-1.065</td>
</tr>
<tr>
<td>Content-based</td>
<td>1.003</td>
<td>-0.014</td>
<td>0.98-1.027</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.932</td>
<td>-0.016</td>
<td>0.906-0.958</td>
</tr>
</tbody>
</table>

- The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$
  - **Linear**: $\alpha = 1$ (scale-free linear attachment)
  - **Sub-linear**: $\alpha < 1$ (the rich may get richer, but at a slower rate)
  - **Super-linear**: $\alpha > 1$ (the rich gets much richer)
**Rich-gets-richer**

<table>
<thead>
<tr>
<th></th>
<th>Slope estimate</th>
<th>Confidence intervals</th>
<th>Hypothesis testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\sigma$</td>
<td>90%</td>
</tr>
<tr>
<td>Individual</td>
<td>1.027</td>
<td>-0.091</td>
<td>0.988-1.065</td>
</tr>
<tr>
<td>Content-based</td>
<td>1.003</td>
<td>-0.014</td>
<td>0.98-1.027</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.932</td>
<td>-0.016</td>
<td>0.906-0.958</td>
</tr>
</tbody>
</table>

- The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$
  - Linear: $\alpha = 1$ (scale-free linear attachment)
  - Sub-linear: $\alpha < 1$ (the rich may get richer, but at a slower rate)
  - Super-linear: $\alpha > 1$ (the rich gets much richer)
- If accounting for content, close to linear preferential attachment
- If not accounting for content, sub-linear preferential attachment
Rich-gets-richer

<table>
<thead>
<tr>
<th></th>
<th>Slope estimate</th>
<th>Confidence intervals</th>
<th>Hypothesis testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha )</td>
<td>( \sigma )</td>
<td>90%</td>
</tr>
<tr>
<td>Individual</td>
<td>1.027</td>
<td>-0.091</td>
<td>0.988-1.065</td>
</tr>
<tr>
<td>Content-based</td>
<td>1.003</td>
<td>-0.014</td>
<td>0.98-1.027</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.932</td>
<td>-0.016</td>
<td>0.906-0.958</td>
</tr>
</tbody>
</table>

- The probability \( P(v_i) \) that a video \( v_i \) with \( v_i \) views will be selected for viewing follows a power law: \( P(v_i) \propto v^\alpha \)
  - Linear: \( \alpha = 1 \) (scale-free linear attachment)
  - Sub-linear: \( \alpha < 1 \) (the rich may get richer, but at a slower rate)
  - Super-linear: \( \alpha > 1 \) (the rich gets much richer)
- If accounting for content, close to linear preferential attachment
- If not accounting for content, sub-linear preferential attachment
### Rich-gets-richer

<table>
<thead>
<tr>
<th></th>
<th>Slope estimate</th>
<th>Confidence intervals</th>
<th>Hypothesis testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\sigma$</td>
<td>90%</td>
</tr>
<tr>
<td>Individual</td>
<td>1.027</td>
<td>-0.091</td>
<td>0.988-1.065</td>
</tr>
<tr>
<td>Content-based</td>
<td>1.003</td>
<td>-0.014</td>
<td>0.98-1.027</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.932</td>
<td>-0.016</td>
<td>0.906-0.958</td>
</tr>
</tbody>
</table>

- The probability $P(v_i)$ that a video $i$ with $v_i$ views will be selected for viewing follows a power law: $P(v_i) \propto v^\alpha$
  - Linear: $\alpha = 1$ (scale-free linear attachment)
  - Sub-linear: $\alpha < 1$ (the rich may get richer, but at a slower rate)
  - Super-linear: $\alpha > 1$ (the rich gets much richer)
- If accounting for content, close to linear preferential attachment
- If not accounting for content, sub-linear preferential attachment
Clone lessons ... (ACM SIGKDD 2012)

- Develop and apply a clone set methodology
  - Accurately assess (both qualitatively and quantitatively) the impacts of various content-agnostic factors on video popularity
- When controlling for video content, we observe a strong linear "rich-get-richer" behavior
  - Except for very young videos, the total number of previous views the most important factor; video age second most important
- Our findings also confirm that inaccurate conclusions can be reached when not controlling for video content
Ephemeral Content Popularity at the Edge and Implications for On-Demand Caching

*IEEE Transactions on Parallel and Distributed Systems (IEEE TPDS), 2016.*
Motivation and observations

- Ephemeral content popularity seen with many content delivery applications
  - At edge this results in many “one timers” (a.k.a. “one hit wonders”)
  - Makes indiscriminate on-demand caching highly inefficient, since many items added to the cache will not be requested again
Preliminary analysis

- YouTube request characteristics as observed at an edge network over a 20 month period
  - 2.3M videos and 5.5M views
Preliminary analysis

- YouTube request characteristics as observed at an edge network over a 20 month period
  - 2.3M videos and 5.5M views
Preliminary analysis

You Tube request characteristics as observed at an edge network over a 20 month period

- 2.3M videos and 5.5M views
- 71% of the requested videos are “one-timers”
Preliminary analysis

You Tube request characteristics as observed at an edge network over a 20 month period

- 2.3M videos and 5.5M views
- 71% of the requested videos are “one-timers”
Preliminary analysis

You Tube request characteristics as observed at an edge network over a 20 month period

- 2.3M videos and 5.5M views
- 71% of the requested videos are “one-timers”
- Demonstrate the need for selective caching policies
Preliminary analysis

- YouTube request characteristics as observed at an edge network over a 20 month period
  - 2.3M videos and 5.5M views
  - 71% of the requested videos are “one-timers”
  - Demonstrate the need for selective caching policies
  - Popularity follow power law (and Zipf)
Characterizing of “one timers”

- Using meta data about these videos, we take a closer look at one-timers and other videos receiving few views.
Characterizing of “one timers”

- Using meta data about these videos, we take a closer look at one-timers and other videos receiving few views
  - Fewer one-timers among movies, shows, and trailers
  - Strong (negative) correlation between global popularity and one-timers
Characterizing of “one timers”

- Using meta data about these videos, we take a closer look at one-timers and other videos receiving few views
  - Fewer one-timers among movies, shows, and trailers
  - Strong (negative) correlation between global popularity and one-timers
Cache modeling

Motivated by our power-law characterization and fittings, we use a Zipf model.
Cache modeling

\[ P(\text{cache miss}) = 1 - \frac{\sum_{i=k}^{\infty} i^{-\alpha} (i-k)}{\sum_{i=1}^{\infty} i^{-\alpha+1}}, \]

\[ P(\text{cache insertion}) = \frac{\sum_{i=k}^{\infty} i^{-\alpha}}{\sum_{i=1}^{\infty} i^{-\alpha+1}}. \]

- Motivated by our power-law characterization and fittings, we use a Zipf model
  - Cache on \( k^{\text{th}} \) request policy
Cache modeling

Motivated by our power-law characterization and fittings, we use a Zipf model

- Cache on k^{th} request policy
- Lower bound “oracle” policies
  - Exact knowledge (exact number of views)
Cache modeling

Motivated by our power-law characterization and fittings, we use a Zipf model

• Cache on k\textsuperscript{th} request policy
• Lower bound “oracle” policies
  • Exact knowledge (exact number of views)
  • Oracle with limited knowledge
    • Binary knowledge (above or below X views)
    • Knows total views, if more than X
    • Knows total views, if less than X
Cache modeling

\[ P(\text{cache miss}) = 1 - \frac{\sum_{i=k}^{\infty} i^{-\alpha}(i-1) - (k-1) \sum_{i=k}^{X-1} i^{-\alpha}}{\sum_{i=1}^{\infty} i^{-\alpha+1}}. \]

\[ P(\text{cache insertion}) = \frac{\sum_{i=k}^{\infty} i^{-\alpha}}{\sum_{i=1}^{\infty} i^{-\alpha+1}}. \]

- Motivated by our power-law characterization and fittings, we use a Zipf model
  - Cache on k\textsuperscript{th} request policy
  - Lower bound “oracle” policies
    - Exact knowledge (exact number of views)
  - Oracle with limited knowledge
    - Binary knowledge (above or below X views)
    - Knows total views, if more than X
    - Knows total views, if less than X
Cache modeling

\[
P(\text{cache insertion}) = \frac{\sum_{i=k+1}^{\infty} i^{-\alpha}}{\sum_{i=1}^{\infty} i^{-\alpha+1}},
\]

\[
P(\text{cache miss}) = 1 - \frac{\sum_{i=k}^{\infty} i^{-\alpha}(i-1)}{\sum_{i=1}^{\infty} i^{-\alpha+1}}, \quad \text{if } X \geq k
\]

\[
P(\text{cache miss}) = 1 - \frac{\sum_{i=k}^{\infty} i^{-\alpha}(i-k)}{\sum_{i=1}^{\infty} i^{-\alpha+1}}, \quad \text{otherwise.}
\]

- Motivated by our power-law characterization and fittings, we use a Zipf model
  - Cache on \(k^{th}\) request policy
  - Lower bound “oracle” policies
    - Exact knowledge (exact number of views)
    - Oracle with limited knowledge
      - Binary knowledge (above or below \(X\) views)
      - Knows total views, if more than \(X\)
      - Knows total views, if less than \(X\)
Evaluation

- Evaluation using both model and traces
  - Similar results
Evaluation

- Evaluation using both model and traces
  - Similar results
  - Limited knowledge
  - Noticeable gap if only knows total for videos with more than X
Evaluation

- Evaluation using both model and traces
  - Similar results
- Limited knowledge
  - Noticeable gap if only knows total for videos with more than X
  - Smaller gap if can predict one-timers (and ones with few views)
Evaluation

- Gap suggest room for improvement
Evaluation

- Gap suggests room for improvement
  - One-timer prediction may close the gap
Evaluation

- Gap suggests room for improvement
  - One-timer prediction may close the gap
- Also looked at SSD scenario
  - Read/write ratio vs cache miss rate
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)

*E.g., Inter-request time dependence ...*
Closing the gap

E.g., Inter-request time dependence …

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
  - Motivated simple baseline policies with imperfect knowledge
Closing the gap

Leverage biases in the probabilities that a request will be a one-timer

- Characterized the one-timers and their request patterns (see paper)
- Motivated simple baseline policies with imperfect knowledge
  - Inter-request Threshold Cache on kth Request

E.g., Inter-request time dependence …
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
  - Motivated simple baseline policies with imperfect knowledge
    - Inter-request Threshold Cache on kth Request
    - Age Threshold Cache on 1st Request

Fig. 15. The expected number of requests $E[x_i|t]$ for videos whose first inter-request time is at least $t/2$ and at most $2t$, plotted as a function of the logarithmic mid-point $t$. 

Number of requests to videos
with $|t-2|t|$ between first two requests

Average (over window)

Time $t$ between first two requests

0 1 sec 1 min 1 hour 1 day 1 mth 1 year

6 months 20 months
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
  - Motivated simple baseline policies with imperfect knowledge
    - Inter-request Threshold Cache on kth Request
    - Age Threshold Cache on 1st Request
  - Trace-driven analysis
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
  - Motivated simple baseline policies with imperfect knowledge
    - Inter-request Threshold Cache on kth Request
    - Age Threshold Cache on 1st Request
  - Trace-driven analysis
    - Some small improvements (but still a large gap …)
Closing the gap

- Leverage biases in the probabilities that a request will be a one-timer
  - Characterized the one-timers and their request patterns (see paper)
  - Motivated simple baseline policies with imperfect knowledge
    - Inter-request Threshold Cache on kth Request
    - Age Threshold Cache on 1st Request
  - Trace-driven analysis
  - Model to give delimiting insights for case when accurate prediction only possible for a subset of videos
Lessons for edge caching (TPDS paper)

- Collected and analyzed a longitudinal edge dataset
  - All YouTube video accesses over a 20-month period
  - Most videos receive few view (e.g., 71% one-timers)
  - Requests per video accurately modelled using power-law distribution

- Use novel workload model and trace-driven simulations to study the performance of alternative edge caching policies
  - Cache on kth request found able to greatly reduce the cache insertion rate, at the cost of relatively modest increases in cache miss rate

- Assess the potential room for improvements through use of content characteristics
  - Oracles suggest there is room for substantial improvements
  - However, would require the prediction of the number of future requests to the content items that are the least popular
  - This problem is both difficult and not well explored, as most research has focused on predicting the most popular contents ...
A Look at the Third-Party Identity Management Landscape

*IEEE Internet Computing, 2016.*

Information Sharing and User Privacy in the Third-party Identity Management Landscape

*Proc. IFIP SEC 2015*

Third-party Identity Management Usage on the Web, Proc

*Proc. PAM 2014*
Third-party Web Authentication

- Use an existing **IDP** (identity provider) account to access an **RP** (relying party)
- Fewer logins
  - Stronger authentication can be used
- Information sharing between websites
  - Privacy leaks!
Background

Third-party Authentication Scenario

Relying party (RP)

Redirect

Logged in

Identity provider (IDP)

Relationship between RP and IDP
Large-scale Crawling

- Popularity-based logarithmic sampling
  - 80,000 points uniformly on a logarithmic range
  - Pareto-like distribution
  - Capturing data from different popularity segments

3rd-party authentication
Large-scale Crawling

- Selenium-based crawling and relationship identification
- Able to process Web 2.0 sites with interactive elements
- Low number of false positives
- Validation with semi-manual classification and text-matching
IDPs vs Content Delivery Services

Content providers:
Import images, scripts etc. from other sites (third-party content providers)

IDPs are much more popular sites than content providers.

Relationships between RPs and IDPs from same region

Regional content service usage
Service-based Analysis

**Likely to be RPs**
- News, file sharing, info

**Likely to be IDPs**
- Social/portal

**Using IDPs from the social/portal category**
- File sharing, info

**Early adopters, using several IDPs**
- Video, tech

**Using IDPs from their own category**
- Commerce, tech

**Not RPs or IDPs**
- Ads, CDN

*3rd-party authentication*
Third-parties and Privacy Risks
App Rights and Information Flows

Privacy risks

App rights example

IDP

Actions:
- Read
- Write
- Update/remove

• Data being sent
• Risks related to
  - Data types
  - Combinations of types
Our Studies on Privacy Risks

- Categorization app-rights data
  - Manual study on the top 200 most popular websites
- Targeted login tests
- Longitudinal analysis of privacy risks
  - 200 websites over three years
Protocol Selection

- OpenID
  - Authentication protocol
  - Decreasing in popularity
- OAuth
  - RP may write/update info on IDP
  - Rich user data is shared
  - Increasingly popular
IDP Selection

- Top 200 April 2012: 69 RPs and 180 relationships
- Same sites, April 2015: +15 RPs and +33 relationships
- Many pairs and triples of popular IDPs
  - 75% of these RPs are selecting all their IDPs from the top 5 most popular IDPs

Privacy risks

Top IDPs:
- Facebook: 37%
- Twitter: 19%
- Sina: 12%
Risk Types

- Only a few relationships in the most privacy preserving category
- 2+ IDPs: More than half are using actions
  - Dangerous when having several IDPs
  - Potential multi-IDP leakage

News and file sharing RPs: most frequent users of actions
Multi-account Information Risks

- Cross account leakage
- Unwanted combinations of conflicting information
- RPs handle multi-IDP usage badly

Connecting several IDPs to an RP

Privacy risks
Structures in the RP-IDP Landscape

**High-degree IDP case**
- IDP having many RPs
- Top IDPs

**High-degree RP case**
- RP having many IDPs
- Specialized IDPs

**Hybrid case**
- Hybrids are both RP and IDP
**Privacy risks**

### RP-to-RP Leakage Example

- **Potential RP-to-RP leaks**
  - Data posted to IDP from RP1
  - Data read from IDP to RP2

<table>
<thead>
<tr>
<th></th>
<th>February 2014</th>
<th>April 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IDP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>645</td>
<td>473</td>
</tr>
<tr>
<td>Severe</td>
<td>150</td>
<td>66</td>
</tr>
<tr>
<td><strong>Facebook</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Severe</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td><strong>Google</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Severe</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Dataset with 44 RPs using Facebook, 14 using Twitter and 12 using Google
Contributions and Findings

- Large-scale RP-IDP study + methodology
  - Categorization of RP-IDP relationships
- Longitudinal changes in the RP-IDP landscape
  - Protocol analysis
  - Privacy risks and information sharing
- Simple web authentication often lack in user privacy
Quality-adaptive Prefetching for Interactive Branched Video using HTTP-based Adaptive Streaming

Empowering the Creative User: Personalized HTTP-based Adaptive Streaming of Multi-path Nonlinear Video
Proc. ACM FhMN@SIGCOMM 2013. (Also in ACM CCR). Best paper award

Bandwidth-aware Prefetching for Proactive Multi-video Preloading and Improved HAS Performance
Motivation

• Content personalization and personalized streaming
  • Regular web content is dynamic and personalized, while videos have remained largely unchanged
  • Viewer’s tastes vary significantly
  • Personalized streaming is relatively unexplored and several interesting questions remain open
We have all seen a movie that (in our taste) is...
We have all seen a movie that (in our taste) is...

too sad
We have all seen a movie that (in our taste) is...

**too sad**

**too violent**
We have all seen a movie that (in our taste) is...

too sad

too violent

too scary

...
We have all seen a movie that (in our taste) is...

too sad

too violent

too scary

... or where we may have wanted our favorite character to make a different choice...
We have all seen a movie that (in our taste) is...

too sad

too violent

too scary

... or where we may have wanted our favorite character to make a different choice...
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to selects between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to selects between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to selects between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive Branched Video

Allow user to select between multiple storylines or alternative endings

Clickable objects allow the user to interact with the player and influence the storyline
Interactive branched video

- Video personalization through user interaction
Interactive branched video

- Video personalization through user interaction
We have solved …

The problem of providing seamless playback in the presence of multiple branch options
We have solved …

The problem of providing seamless playback in the presence of multiple branch options
We have solved …

The problem of providing seamless playback in the presence of multiple branch options

- HTTP-based Adaptive Streaming
- Path and quality-aware prefetching
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD

- Multiple encodings of each fragment (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based adaptive streaming

![Diagram showing base video split into chunks over time]
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based **adaptive** streaming
  - Multiple encodings of each chunk (defined in manifest file)
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD

- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD
- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD

- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- **HTTP-based streaming**
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD

- **HTTP-based adaptive streaming**
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HTTP-based Adaptive Streaming (HAS)

- HTTP-based streaming
  - Video is split into chunks
  - Easy firewall traversal and caching
  - Easy support for interactive VoD

- HTTP-based adaptive streaming
  - Multiple encodings of each chunk (defined in manifest file)
  - Clients adapt quality encoding based on buffer/network conditions
HAS-based interactive branched video

- Branched video and branch points
  - The video can include branch points, with multiple branch choices
  - User selects which segment to play back next
• Branched video and branch points
  • The video can include branch points, with multiple branch choices
  • User selects which segment to play back next
• Our solution: Combine branched video and HAS
HAS-based interactive branched video

• Branched video and branch points
  • The video can include branch points, with multiple branch choices
  • User selects which segment to play back next
• Our solution: Combine branched video and HAS
• Goal: Seamless playback even if user decision at last possible moment
Problem description and constraints
Problem description and constraints

- Problem: Maximize quality, given playback deadlines and bandwidth conditions
Problem description and constraints

- Objective function:

\[
\text{maximize } \text{playback quality}
\]
Problem description and constraints

- Objective function:

\[
\text{maximize } \sum_{i=1}^{n_e} q_i l_i + \sum_{i=n_e+1}^{n_e + |\mathcal{E}^b|} \omega^b_e q_i l_i
\]

Current segment
Objective function:

\[
\text{maximize } \sum_{i=1}^{n_e} q_i l_i + \sum_{i=n_e+1}^{n_e+|\mathcal{E}^b|} w^b_e q_i l_i
\]
Problem Description and Constraints

- Download order: round robin (optimal)
Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (optimal)
• Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (optimal)
Problem Description and Constraints

- Download order: round robin (extra workahead)

Current segment first chunk next extra workahead
Problem Description and Constraints

- Download order: round robin (extra workahead)
Problem Description and Constraints

- Download order: round robin *(extra workahead)*

---

Current segment first chunk next extra workahead
Problem Description and Constraints

- Download order: round robin (extra workahead)

Current segment first chunk next extra workahead
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...

**Current segment first chunk next**
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...

Selected path
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...
Once branch point has been traversed, move on to next segment ...
• Once branch point has been traversed, move on to next segment ...
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment...
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...

```
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>8</th>
<th>9</th>
<th>22</th>
</tr>
</thead>
</table>
```

- current segment  
- first chunk next
Problem Description and Constraints

- Once branch point has been traversed, move on to next segment ...
Once branch point has been traversed, move on to next segment...
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls

Playback schedule

Download schedule
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - Current segment: e.g., 2 and 3
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - Current segment: e.g., 2 and 3

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if } 1 \leq i \leq n_e \]
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - Current segment: e.g., 2 and 3

\[ t^c_i \leq t^d_i = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if } 1 \leq i \leq n_e \]
Problem Description and Constraints

• Playback deadlines
  • for seamless playback without stalls
  • Current segment: e.g., 2 and 3

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if } 1 \leq i \leq n_e \]
Problem Description and Constraints

Playback deadlines

- for seamless playback without stalls
- Current segment: e.g., 2 and 3

\[ t_{i}^{c} \leq t_{i}^{d} = \tau + \sum_{j=1}^{i-1} l_{j}, \quad \text{if } 1 \leq i \leq n_{e} \]
Problem Description and Constraints

• Playback deadlines
  • for seamless playback without stalls
  • Current segment: e.g., 2 and 3

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if} \ 1 \leq i \leq n_e \]
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - Current segment: e.g., 2 and 3

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if } 1 \leq i \leq n_e \]
Problem Description and Constraints

Playback deadlines

• for seamless playback without stalls
• Current segment: e.g., 2 and 3

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{i-1} l_j, \quad \text{if } 1 \leq i \leq n_e \]

Startup delay

Playtime of earlier chunks
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - First chunks next segment: e.g., 4, 7, and 10
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - First chunks next segment: e.g., 4, 7, and 10

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{n_e} l_j, \quad \text{if } n_e < i \leq n_e + |\mathcal{E}^b| \]
Problem Description and Constraints

- Playback deadlines
  - for seamless playback without stalls
  - First chunks next segment: e.g., 4, 7, and 10

\[ t_i^c \leq t_i^d = \tau + \sum_{j=1}^{n_e} l_j, \quad \text{if } n_e < i \leq n_e + |\mathcal{E}^b| \]
Problem Description and Constraints

• Playback deadlines
  • for seamless playback without stalls
  • First chunks next segment: e.g., 4, 7, and 10

\[ t^c_i \leq t^d_i = \tau + \sum_{j=1}^{n_e} l_j \text{, if } n_e < i \leq n_e + |\mathcal{E}^b| \]

Time at which branch point is reached

Download completion times
Interactive Branched Video Contributions

- Designed and implemented branched video player that achieve seamless streaming without playback interruptions

- Designed optimized policies that maximize playback quality while ensuring sufficient workahead to avoid stalls

- Evaluation shows that solution effectively adapt quality levels and number of parallel connections so as to provide best possible video quality, given current conditions

---

- Extensions, generalizations, and variations include “multi-file prefetching for impatient users” [Proc. ACM Multimedia 2015]
Niklas Carlsson (niklas.carlsson@liu.se)
Research overview and pubs: www.ida.liu.se/~nikca/
Summary

Niklas Carlsson (niklas.carlsson@liu.se)
Research overview and pubs: www.ida.liu.se/~nikca/
Summary

Niklas Carlsson (niklas.carlsson@liu.se)
Research overview and pubs: www.ida.liu.se/~nikca/
Niklas Carlsson (niklas.carlsson@liu.se)
Research overview and pubs: www.ida.liu.se/~nikca/
Thanks for listening!

Niklas Carlsson (niklas.carlsson@liu.se)
Research overview and pubs: www.ida.liu.se/~nikca/