The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity

Youna Borghol
Sebastien Ardon
Niklas Carlsson
Derek Eager
Anirban Mahanti

UNSW & NICTA
NICTA
Linköping University
University of Saskatchewan
NICTA

August 15, 2012
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
Not all videos will reach the same popularity and have the same impact.

Some popularity differences due to content differences.
Popularity differences arise not only because of differences in video content, 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

- Popularity differences arise not only because of differences in video content, 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)
Motivation

- Popularity differences arise not only because of differences in video content, 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

For example, videos uploaded by users with large social networks may tend to be more popular because they tend to have more interesting content, not because social network size has a substantial direct impact on 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

- 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

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

Some factor of interest
Methodology

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

(1) Aggregate model

- Ignore clone “identity” (or content)
  - Can be used as a baseline ...
Methodology: (1) Aggregate model

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

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

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

(2) Individual model

- Current popularity (e.g., views in week)
- Some factor of interest
- Predicted value
- Error
Methodology: (3) Content-based model

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

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

- **Scaled measured value**
- **Encoding:** 1 if clone k; otherwise 0
- **Content-agnostic factors**
- **Impact of content**
- **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!!)
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information…
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information …
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information …
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information...
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information …
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information …
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information ...
Preliminary analysis

- A closer look at correlations between factors and identifying groups of variables that provide redundant information …
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)
### Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for
## Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for
Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for
Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

$\Delta = 0.114$

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for
### Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

$$\Delta = 0.063$$
### Impact of content identity

<table>
<thead>
<tr>
<th></th>
<th>View count (1 var.)</th>
<th>+ age (2 var.)</th>
<th>+ followers (3 var.)</th>
<th>All (15 var.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual (e.g., 41)</td>
<td>0.861</td>
<td>0.870</td>
<td>0.874</td>
<td>0.895</td>
</tr>
<tr>
<td>Content-based</td>
<td>0.792</td>
<td>0.850</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.707</td>
<td>0.808</td>
<td>0.808</td>
<td>0.821</td>
</tr>
</tbody>
</table>

- View count by itself explain a lot of the variation
- The relative importance of age, followers etc. over estimated if content is not accounted for

\[ \Delta = 0.114 \]
\[ \Delta = 0.063 \]
## 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>
<tr>
<td>$H_0$: $\alpha = 1$</td>
<td>$H_0$: $\alpha \geq 1$</td>
<td>$H_0$: $\alpha \leq 1$</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>α</td>
<td>σ</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 $i$ 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>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>
<tr>
<td>Individual</td>
<td>1.027</td>
<td>-0.091</td>
</tr>
<tr>
<td>Content-based</td>
<td>1.003</td>
<td>-0.014</td>
</tr>
<tr>
<td>Aggregate</td>
<td>0.932</td>
<td>-0.016</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
First-mover advantage

- **Significant first-mover advantage**
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

<table>
<thead>
<tr>
<th>Ratio of view counts (x)</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner uploaded</td>
<td>27.1</td>
<td>12.5</td>
<td>8.3</td>
<td>6.3</td>
<td>6.3</td>
<td>39.6</td>
</tr>
<tr>
<td>Winner searched</td>
<td>66.7</td>
<td>8.3</td>
<td>0.0</td>
<td>8.3</td>
<td>8.3</td>
<td>8.3</td>
</tr>
</tbody>
</table>
First-mover advantage

- Significant first-mover advantage
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

<table>
<thead>
<tr>
<th>Ratio of view counts (x)</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner uploaded</td>
<td>27.1</td>
<td>12.5</td>
<td>8.3</td>
<td>6.3</td>
<td>6.3</td>
<td>39.6</td>
</tr>
<tr>
<td>Winner searched</td>
<td>66.7</td>
<td>8.3</td>
<td>0.0</td>
<td>8.3</td>
<td>8.3</td>
<td>8.3</td>
</tr>
</tbody>
</table>
First-mover advantage

- Significant first-mover advantage
- First-mover often the “winner”; even when not the winner, it is not far behind (e.g., 50% of the first movers are within a factor 10 of the “winner”)
- The first video discovered through search have even better success rate

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner uploaded</td>
<td>27.1</td>
<td>12.5</td>
<td>8.3</td>
<td>6.3</td>
<td>6.3</td>
<td>39.6</td>
</tr>
<tr>
<td>Winner searched</td>
<td>66.7</td>
<td>8.3</td>
<td>0.0</td>
<td>8.3</td>
<td>8.3</td>
<td>8.3</td>
</tr>
</tbody>
</table>
Initial popularity

<table>
<thead>
<tr>
<th></th>
<th>Aggregate</th>
<th></th>
<th></th>
<th></th>
<th>Content-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1d</td>
<td>3d</td>
<td>7d</td>
<td>14d</td>
<td>1d</td>
</tr>
<tr>
<td>View Count</td>
<td>0.44</td>
<td>0.42</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>Keywords</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>Video quality</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td>0.35</td>
</tr>
<tr>
<td>Upl. View cnt.</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td>0.64</td>
</tr>
<tr>
<td>Upl. Followers</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td>0.58</td>
</tr>
<tr>
<td>Upl. Contacts</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td>0.42</td>
</tr>
<tr>
<td>Upl. Video cnt.</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td>0.38</td>
</tr>
</tbody>
</table>

Age-based analysis
- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content
### Initial popularity

#### Age-based analysis
- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content

<table>
<thead>
<tr>
<th></th>
<th>Aggregate</th>
<th>Content-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1d</td>
<td>3d</td>
</tr>
<tr>
<td>View Count</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>Video quality</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Upl. View cnt.</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Upl. Followers</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Upl. Contacts</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Upl. Video cnt.</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>
### Initial popularity

<table>
<thead>
<tr>
<th></th>
<th>Aggregate</th>
<th>Content-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1d</td>
<td>3d</td>
</tr>
<tr>
<td>View Count</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>Video quality</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>Upl. View cnt.</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Upl. Followers</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Upl. Contacts</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Upl. Video cnt.</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

**Age-based analysis**

- Uploader popularity a good initial predictor
- After about a week, the view count catches up
- Factors such as keywords relatively (much) more important when taking into account the content
Contributions

- 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

- Analyze a number of phenomena that may contribute to rich-get-richer, including the first-mover advantage, and search bias towards popular videos

- For young videos, factors other than the total number of previous views become relatively more important
  - E.g., uploader characteristics and number of keywords

- Our findings also confirm that inaccurate conclusions can be reached when not controlling for video content
Thank you!

- Youmna Borghol  UNSW & NICTA
- Sebastien Ardon  NICTA
- Niklas Carlsson  Linköping University
- Derek Eager  University of Saskatchewan
- Anirban Mahanti  NICTA