Why not to use popularity scores from platforms
The hidden biases of YouTube data

Merja Mahrt
Biases in platform data

Issues in CSS / big data literature

- Meaning of “likes,” “retweets,” etc. unclear (boyd & Crawford, 2012; Lomborg & Bechmann, 2014; Ruths & Pfeffer, 2014)
- “Hidden biases” due to social inequalities of usage (Crawford, 2013)
- Intransparent algorithms in platforms
  - E.g., Twitter’s “trending” algorithm (Gillespie, 2012)
Biases in platform data

Background: Research on digital fragmentation

- Fears about “echo chambers” (Sunstein, 2007) and “filter bubbles” (Pariser, 2011)
- Main factors
  - More content available online
  - More selective user behavior
  - Selection through platform design
- How to determine the reach of online content?
Digital fragmentation

A multi-method approach

- Platform data
  - Content structures
  - Popularity scores
- Usage data
- Survey
Popularity of YouTube content

Platform information: Popularity scores

- Content analysis
  - YouTube daily top 10
  - “Trends” in Germany
    - Trends Dashboard (now defunct)
  - March through August, 2014
- 1,164 unique videos
Top 10 YouTube videos

1,165 videos from German top 10, 03-08/2014

- Gaming
- Everyday life vlogging
- Music
- Humor
- Current events
- YouTube series, top 10 lists
- Sports
- Celebrities
- Commercials
- TV entertainment
- Other
Digital fragmentation

Popularity of YouTube content

- Survey
  - Online access panel (Respondi)
  - First week of September, 2014
  - 1,665 respondents, aged 18-69
  - Representative for German online users

- YouTube use
  - Reach of online video platforms
  - Recognition of popular YouTube videos
Popularity of YouTube content

Survey

- Video platforms highly popular
  - 50% of online users access video platforms at least once per week
  - [ARD/ZDF Online Study 2014: 45%; Koch & Liebholz, 2014]

- (Almost) daily users of video platforms
  - +24% aged 18-29
  - +11% male
  - +13% highschool graduates [Abitur]
Popularity of YouTube content

Survey

- Popularity of YouTube videos

- **Dagi Bee**
  - "Probleme jedes Mädchens #2"
  - 2%
  - 745,000 views

- **LeFloid**
  - "Über den Trend, sich selbst anzuzünden"
  - 5%
  - 1,100,000 views

- **Edeka**
  - "Supergeil"
  - 34%
  - 11,930,000 views
Popularity of YouTube content

YouTube usage behavior

- Clickstream data
  - Nielsen panel of 54,790 German users
  - YouTube usage in June, 2014
    - 8,147 users
    - 433,235 video views
    - 244,925 unique videos
      - 87% viewed by one user each
    - 109,093 channels
      - 80% viewed by one user each
Clickstream data

- Popularity of YouTube channels, June, 2014

- **Dagi Bee**
  - 1.1%
  - 68 videos watched by 90 users

- **LeFloid**
  - 2.6%
  - 90 videos watched by 211 users

- **Edeka**
  - 0.1%
  - 9 videos watched by 15 users
Popularity of YouTube content

Comparing platform and clickstream data

- 26 videos from top 10 with 1 million+ views (June, 2014)
Comparing platform and clickstream data

- 26 videos from top 10 with 1 million+ views (June, 2014)
Conclusion

Why not to use popularity scores

- Popularity scores and clickstream data capture different aspects of YouTube “use”—with different biases
- Over- and underrepresentation (age)
  - Validity of age in clickstream data doubtful
  - Small (sub)samples and idiosyncratic YouTube preferences
  - Limitations of assessing use on content level with survey
- (How) does YouTube favor advertisers and/or influencers?

- Which dataset captures reach of YouTube content best?
  - ...to ultimately assess digital fragmentation?
Thank you
Merja Mahrt
mahrt@hhu.de


