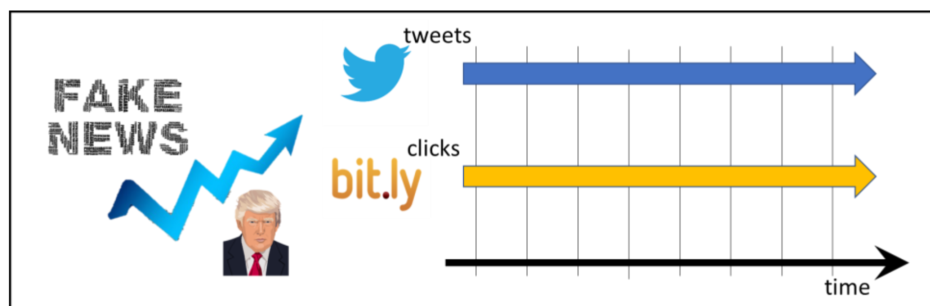
Do we Read what we Share? Analyzing the Click Dynamic of News Articles Shared on Twitter

Jesper Holmstrom
Daniel Jonsson
Filip Polbratt
Olav Nilsson
Linnea Lundstrom
Sebastian Ragnarsson
Anton Forsberg
Karl Andersson
Niklas Carlsson

LINKÖPING UNIVERSITY



Proc. IEEE/ACM ASONAM, Vancouver, Canada, Aug. 2019.

Motivation

- News and information spread over social media can have big impact on thoughts, beliefs, and opinions
  - Important to understand the sharing dynamics on these forums …

- Most studies trying to capture these dynamics rely only on Twitter's open APIs to measure how frequently articles are shared/retweeted
  - They do not capture how many users actually read the articles linked in these tweets ...

… here, we instead focus on the **clicks** leading to linked articles …

… and measure + analyze these **over time.**

Motivation

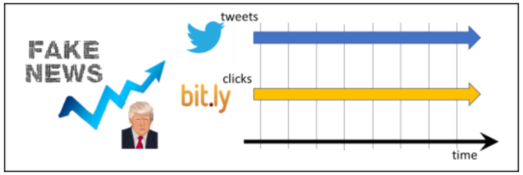
- News and information spread over social media can have big impact on thoughts, beliefs, and opinions
  - Important to understand the sharing dynamics on these forums …

- Most studies trying to capture these dynamics rely only on Twitter's open APIs to measure how frequently articles are shared/retweeted
  - They do not capture how many users actually read the articles linked in these tweets …

… here, we instead focus on the **clicks** leading to linked articles …

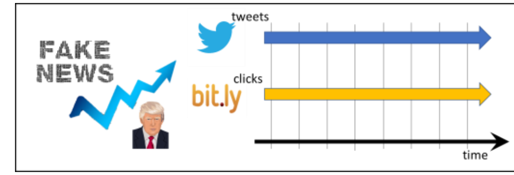
… and measure + analyze these **over time.**

Contributions at a glance

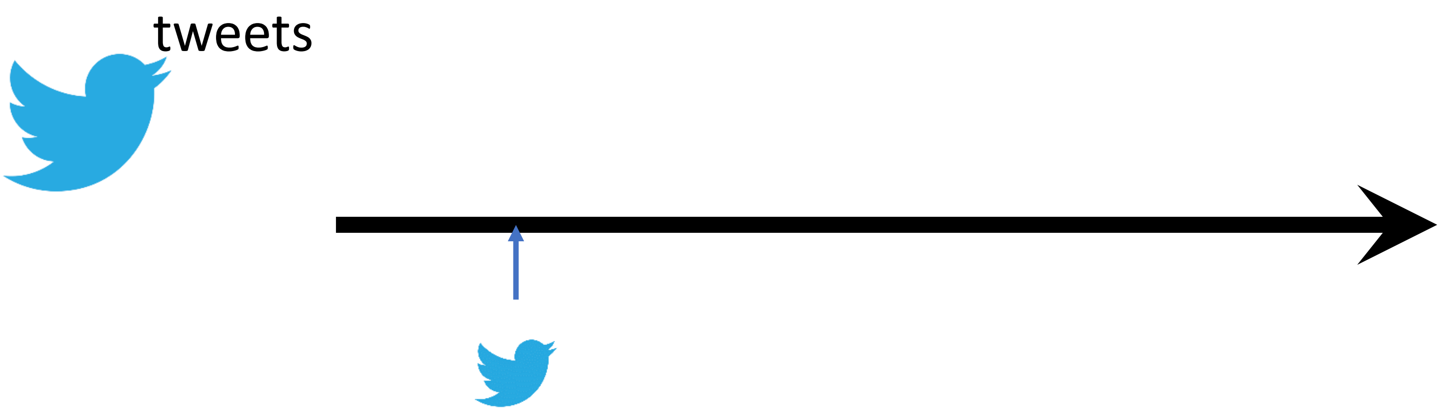


- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time

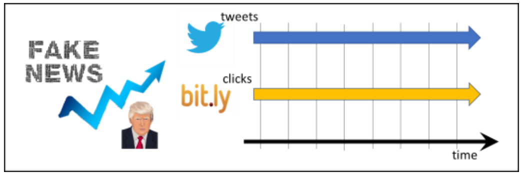
Contributions at a glance



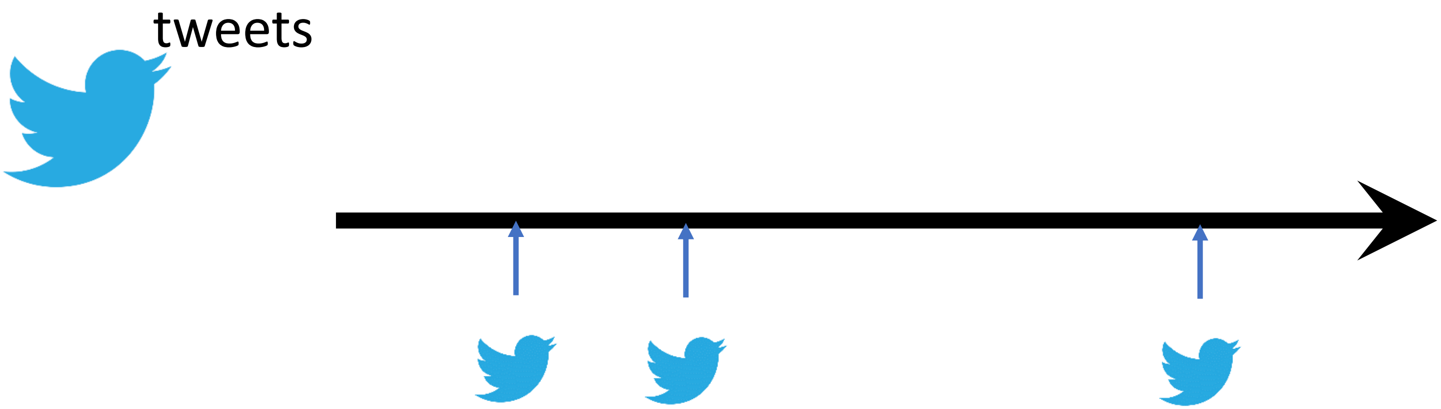
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



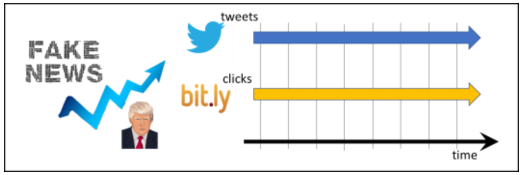
Contributions at a glance



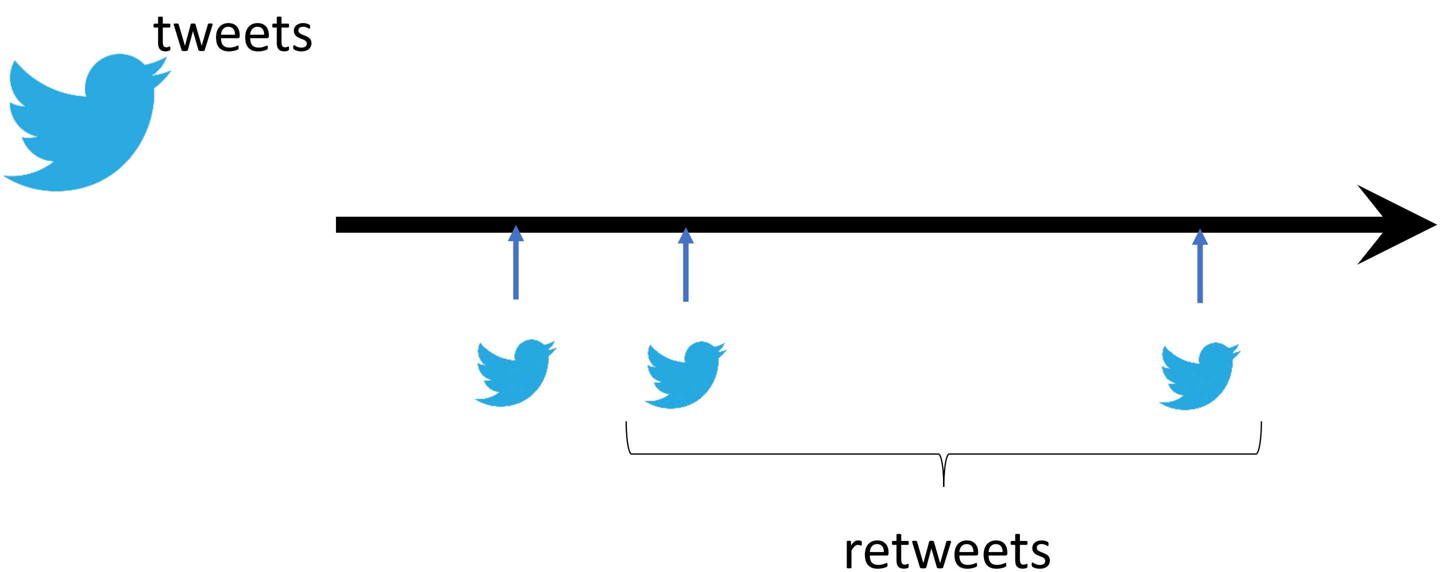
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



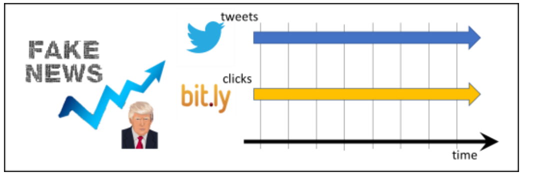
Contributions at a glance



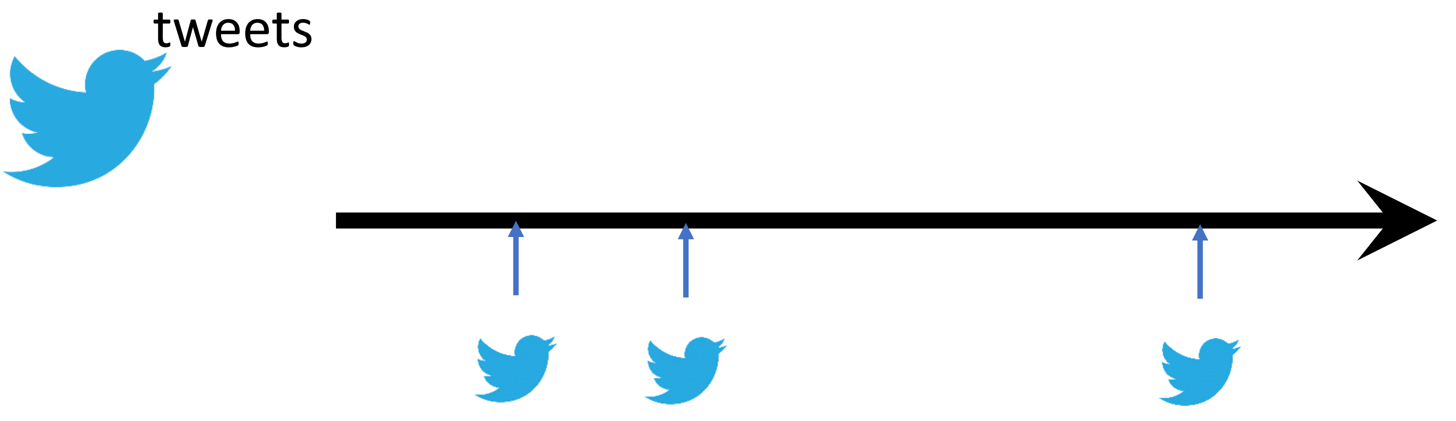
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



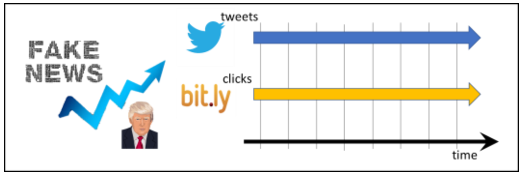
Contributions at a glance



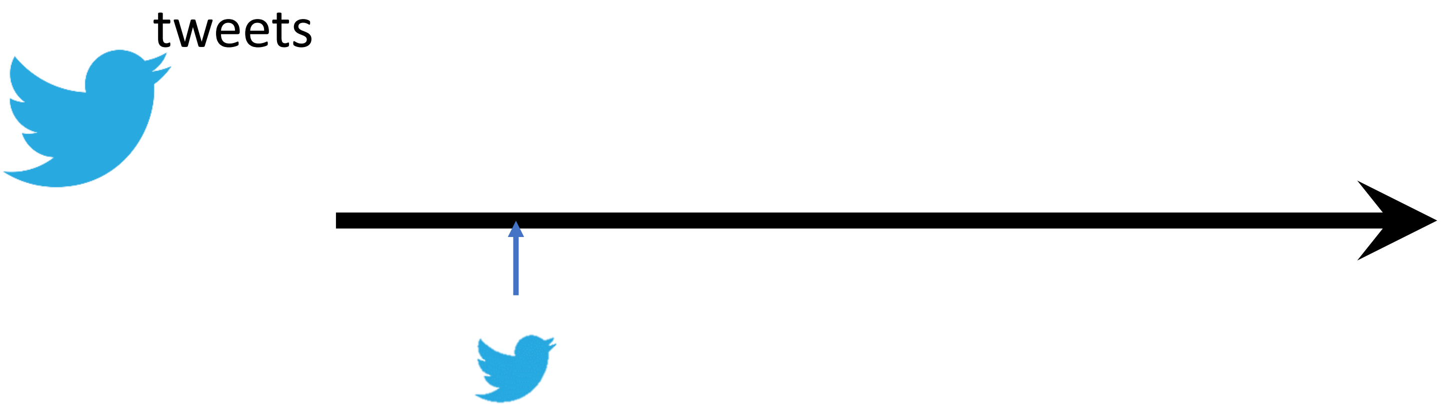
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



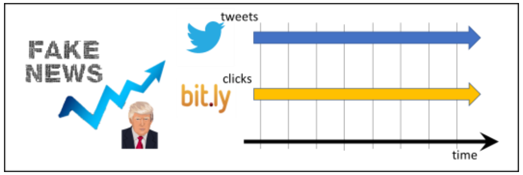
Contributions at a glance



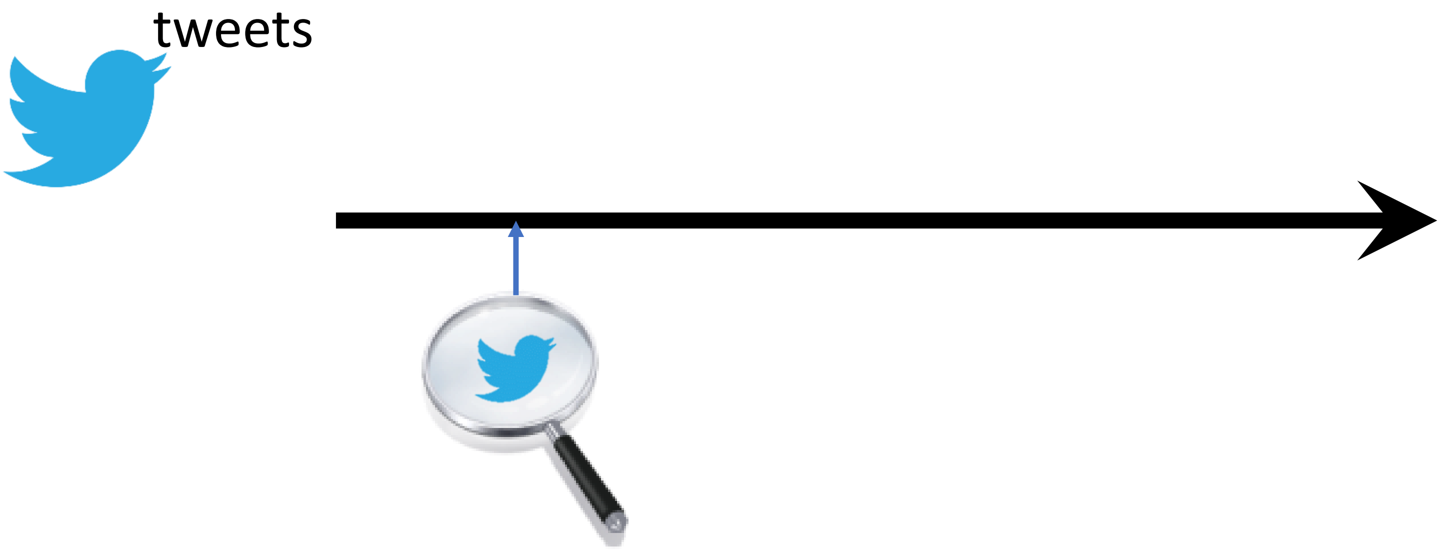
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



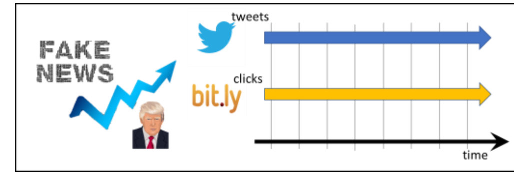
Contributions at a glance



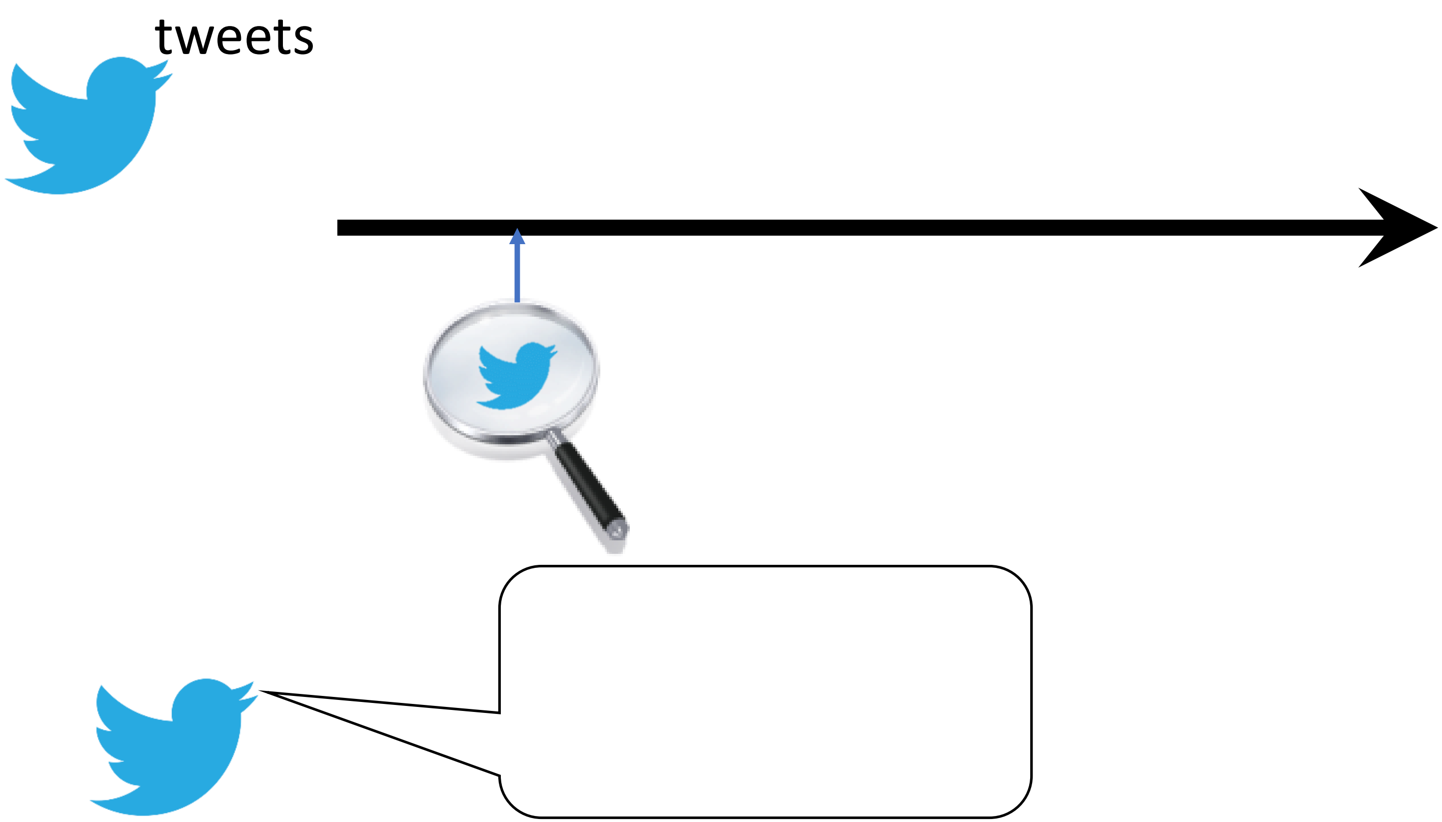
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



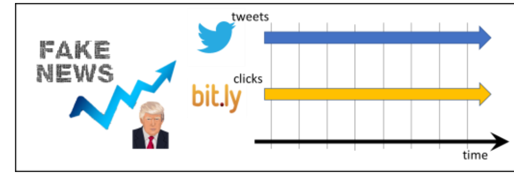
Contributions at a glance



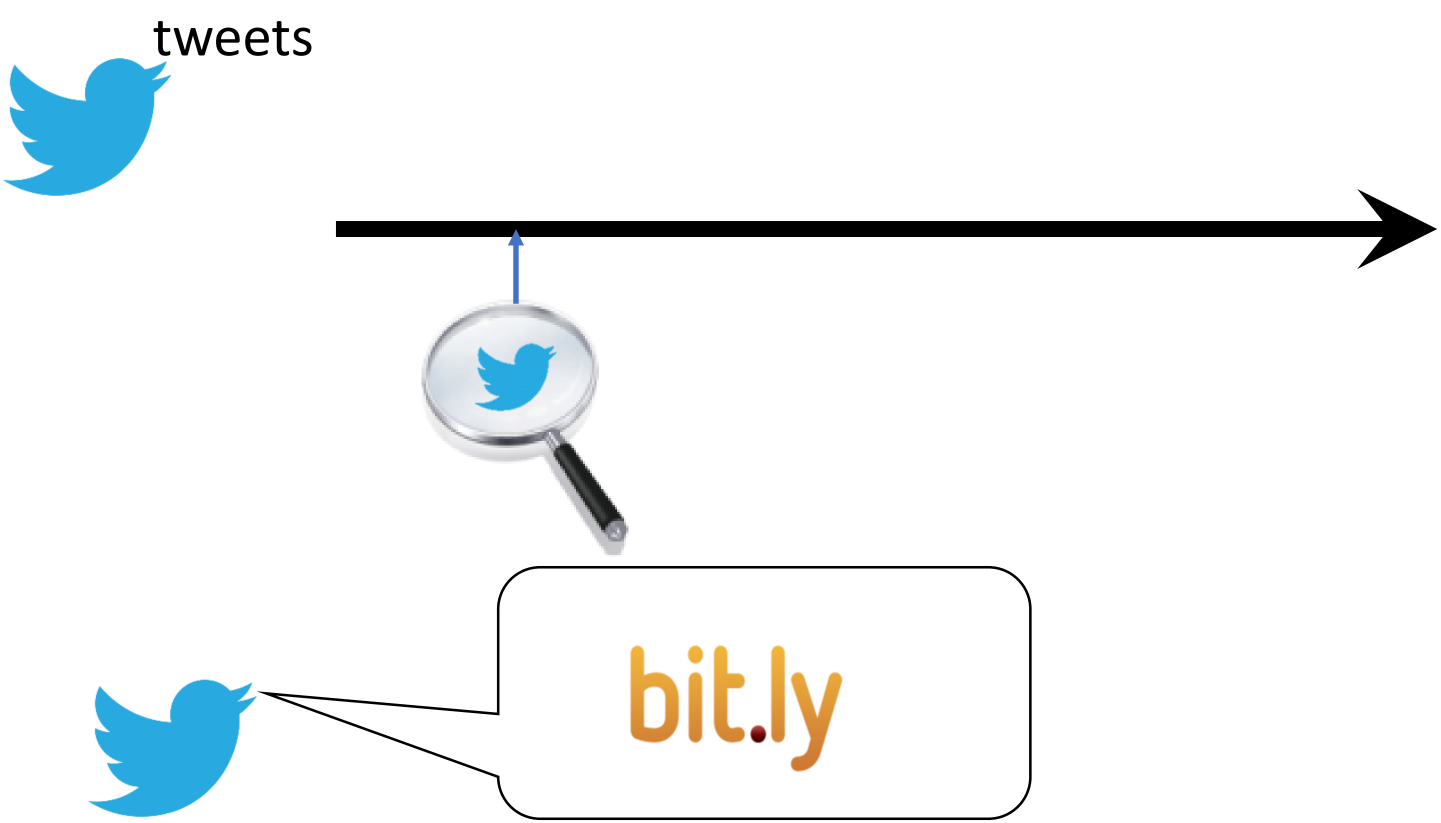
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



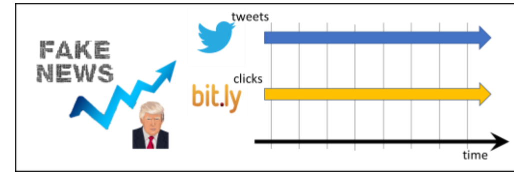
Contributions at a glance



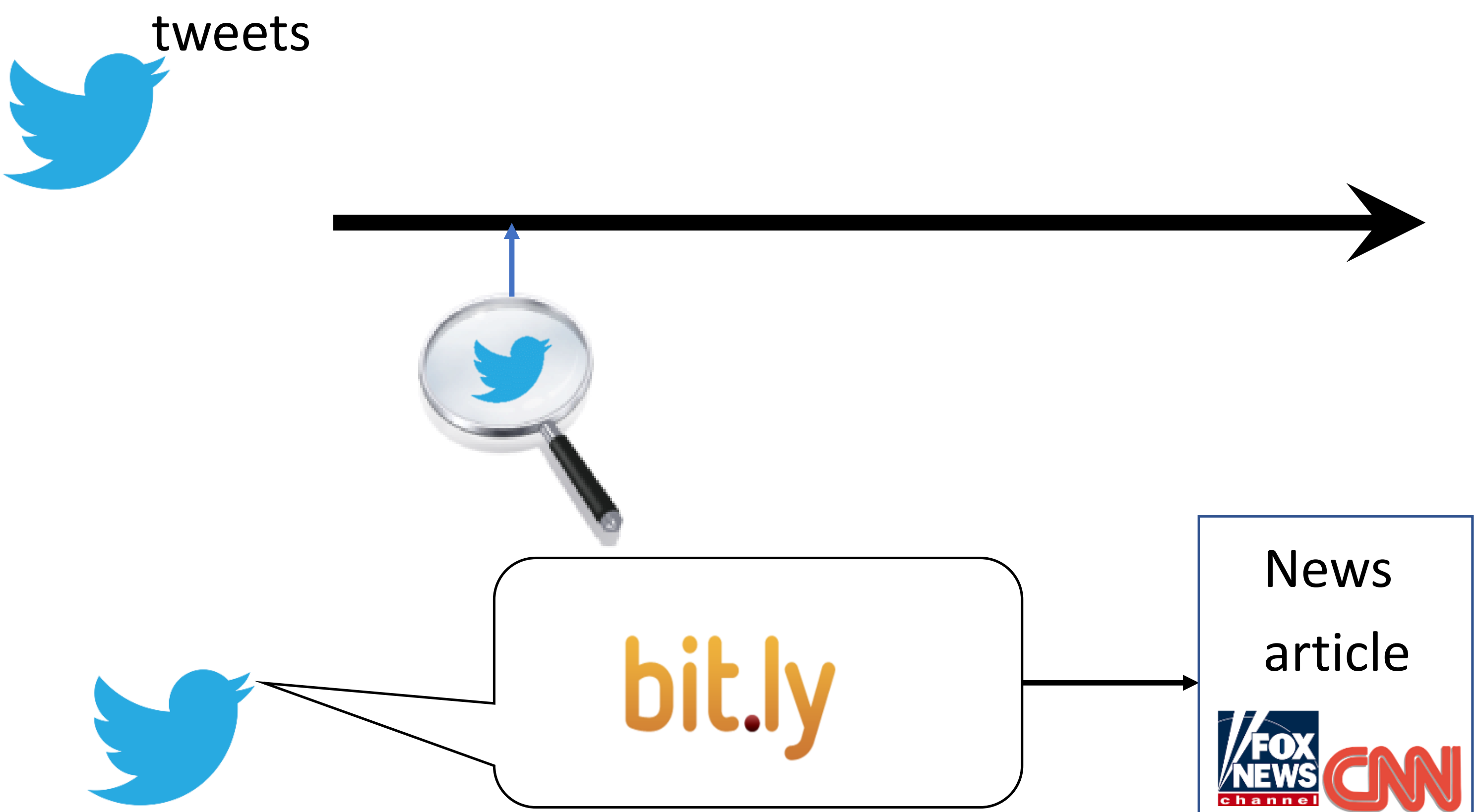
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



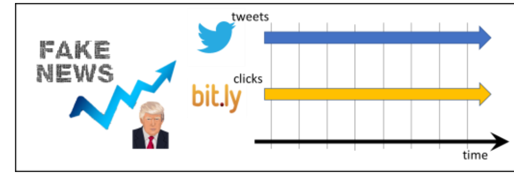
Contributions at a glance



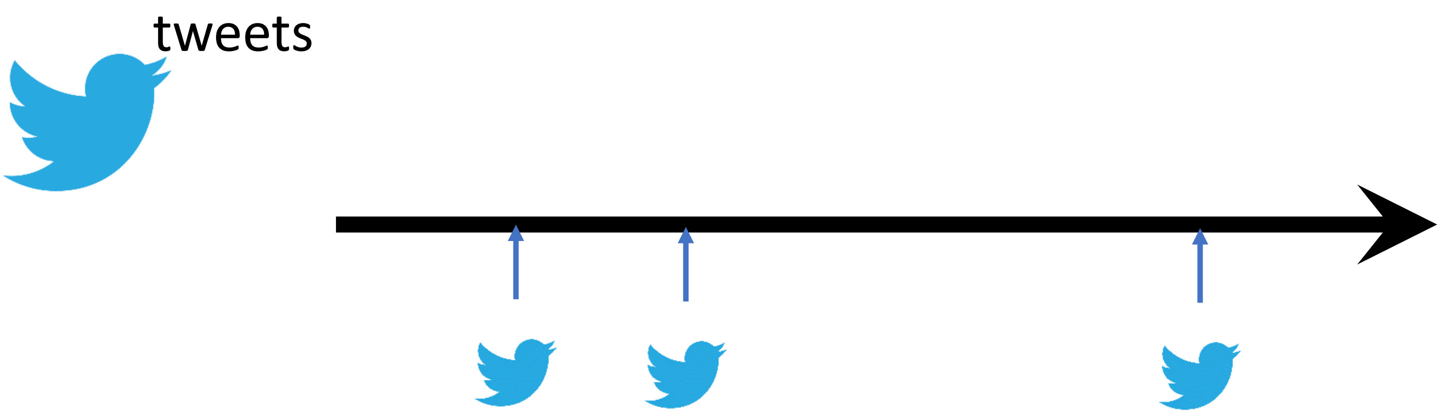
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



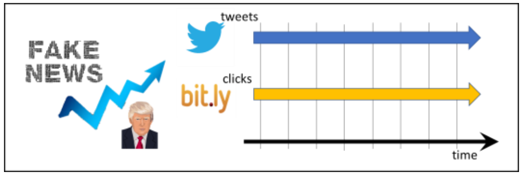
Contributions at a glance



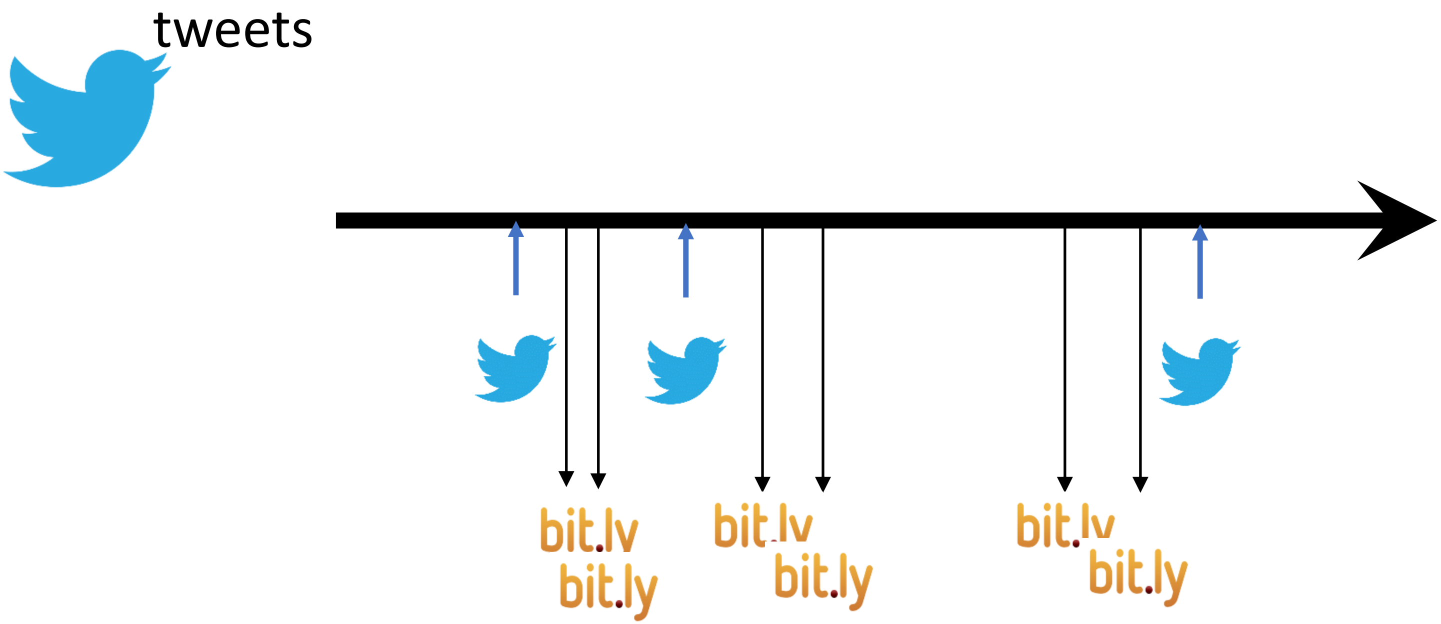
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



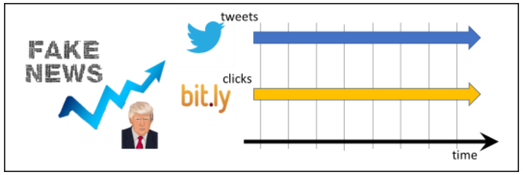
Contributions at a glance



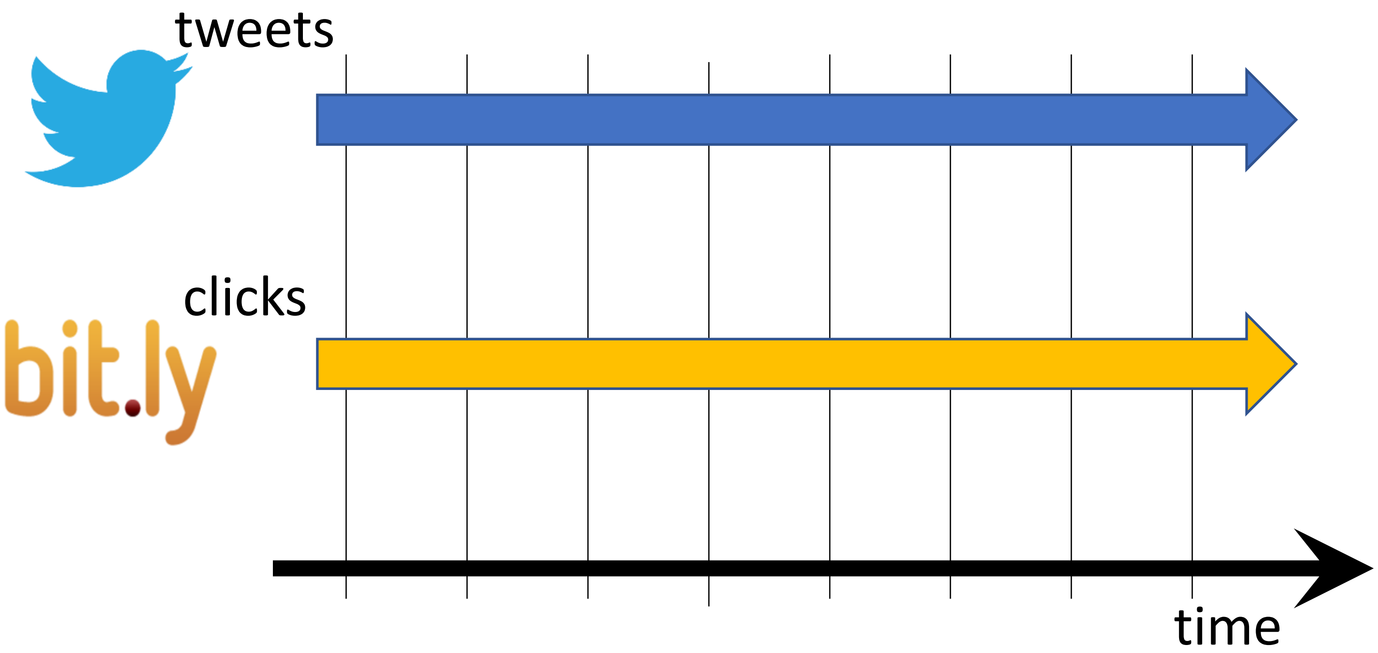
- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time



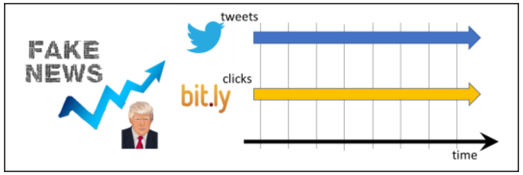
Contributions at a glance



- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time

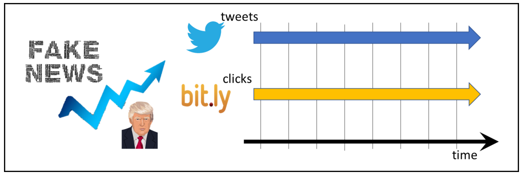


Contributions at a glance



- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time
- Example observations from temporal analysis
  - Noticeable differences in the relative number of clicks vs. retweets occurring at different parts of the news cycle
  - Retweet data often underestimates biases towards clicking popular links/articles
  - Significant differences in the clicks-per-tweets ratio, including (alarmingly) many links with more retweets than clicks
  - Significant age biases, including relatively high initial click rates for articles younger than a week and much more stable click rates for older and long-term popular articles
  - Insights into how age-dependent popularity skews and age-dependent churn impact the clicks observed by different classes of links

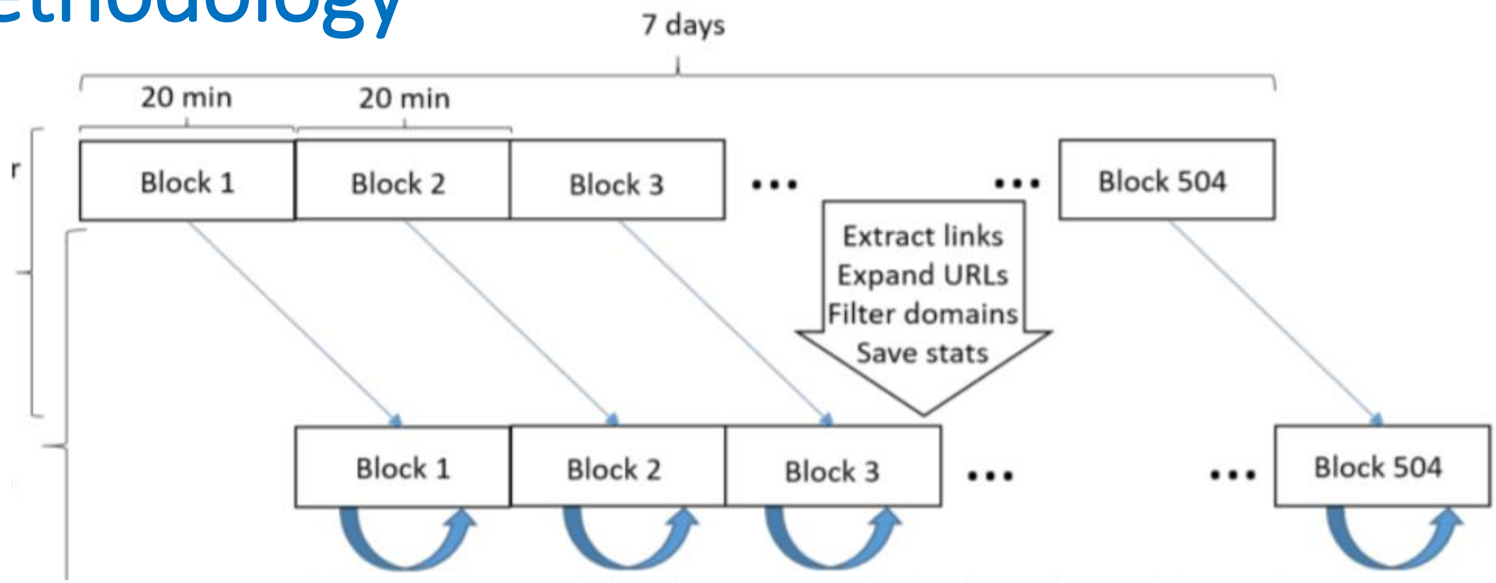
Contributions at a glance



- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time
- Example observations from temporal analysis
  - Noticeable differences in the relative number of clicks vs. retweets occurring at different parts of the news cycle
  - Retweet data often underestimates biases towards clicking popular links/articles
  - Significant differences in the clicks-per-tweets ratio, including (alarmingly) many links with more retweets than clicks
  - Significant age biases, including relatively high initial click rates for articles younger than a week and much more stable click rates for older and long-term popular articles
  - Insights into how age-dependent popularity skews and age-dependent churn impact the clicks observed by different classes of links
- We validate our findings (and identify invariants) using both data from May 2017 and a per-website-based analysis

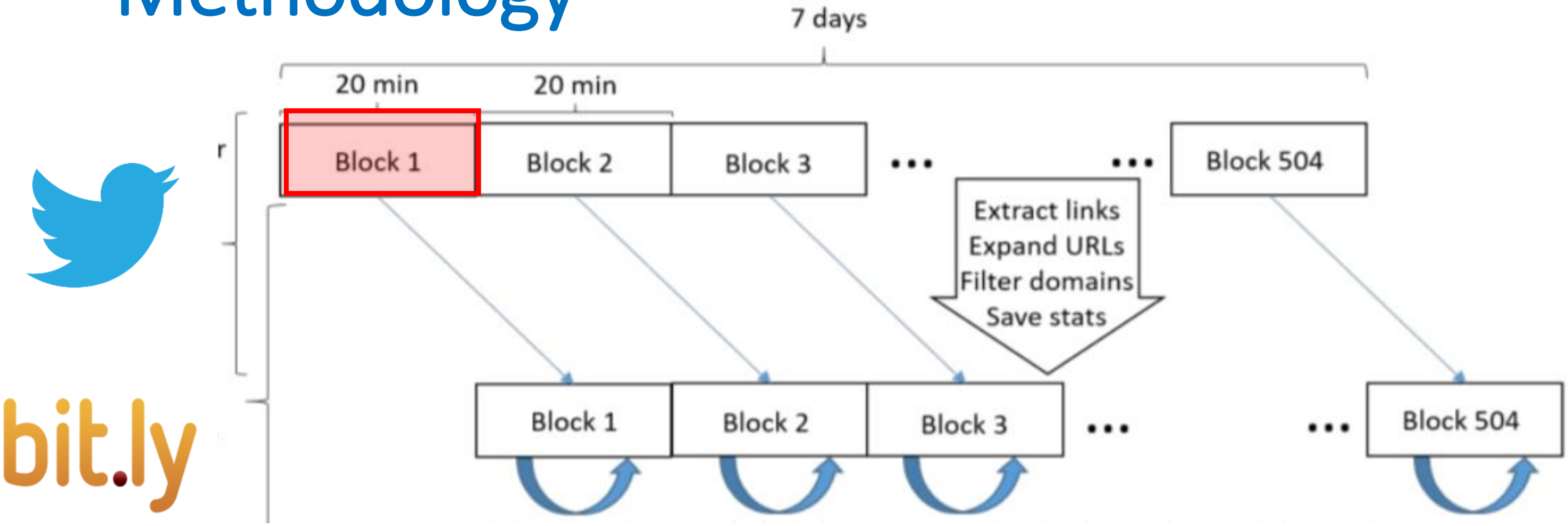
Methodology

Methodology



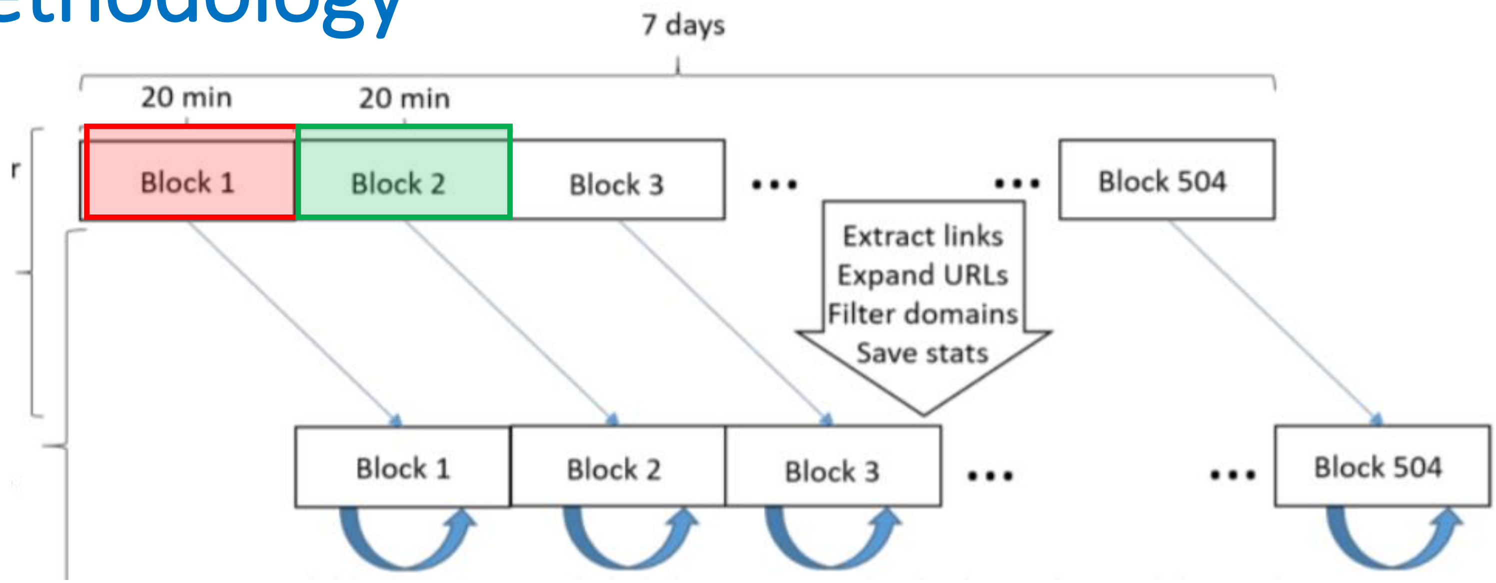
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days

Methodology



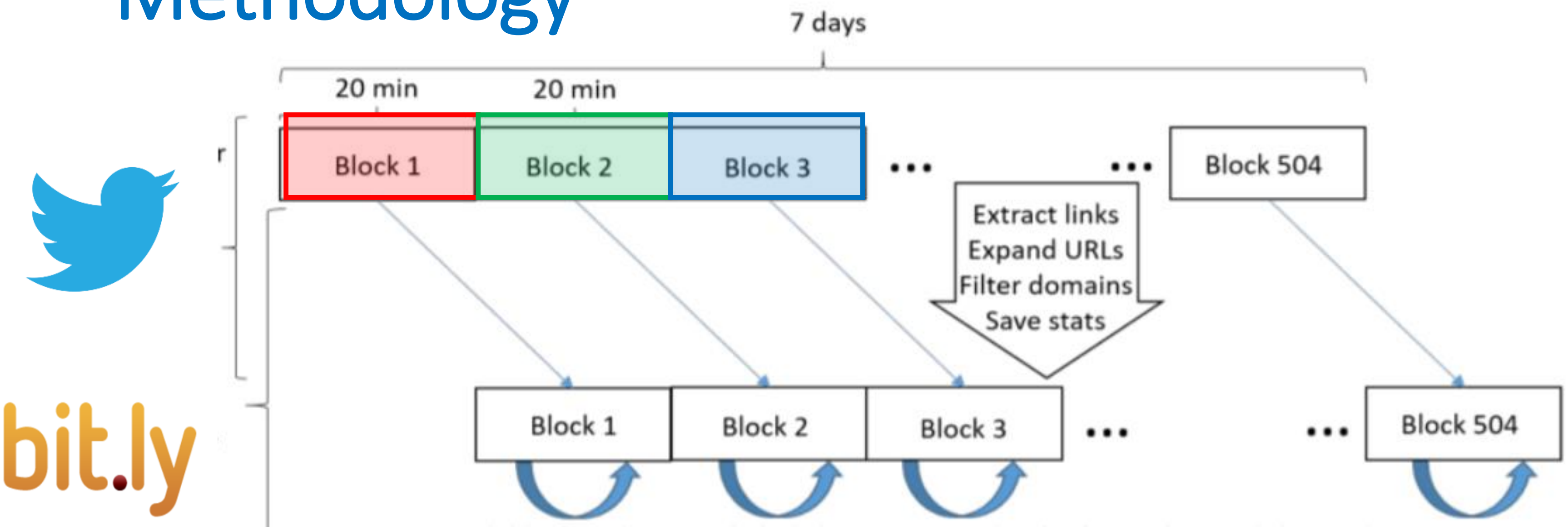
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days

Methodology



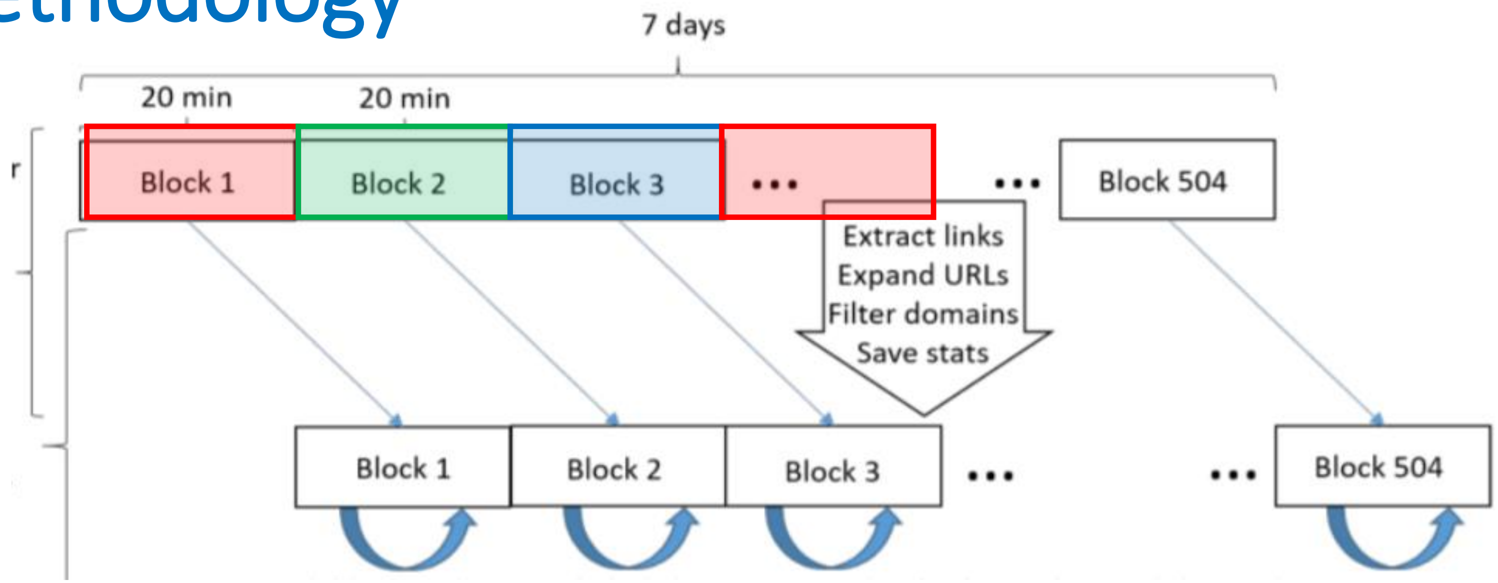
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days

Methodology



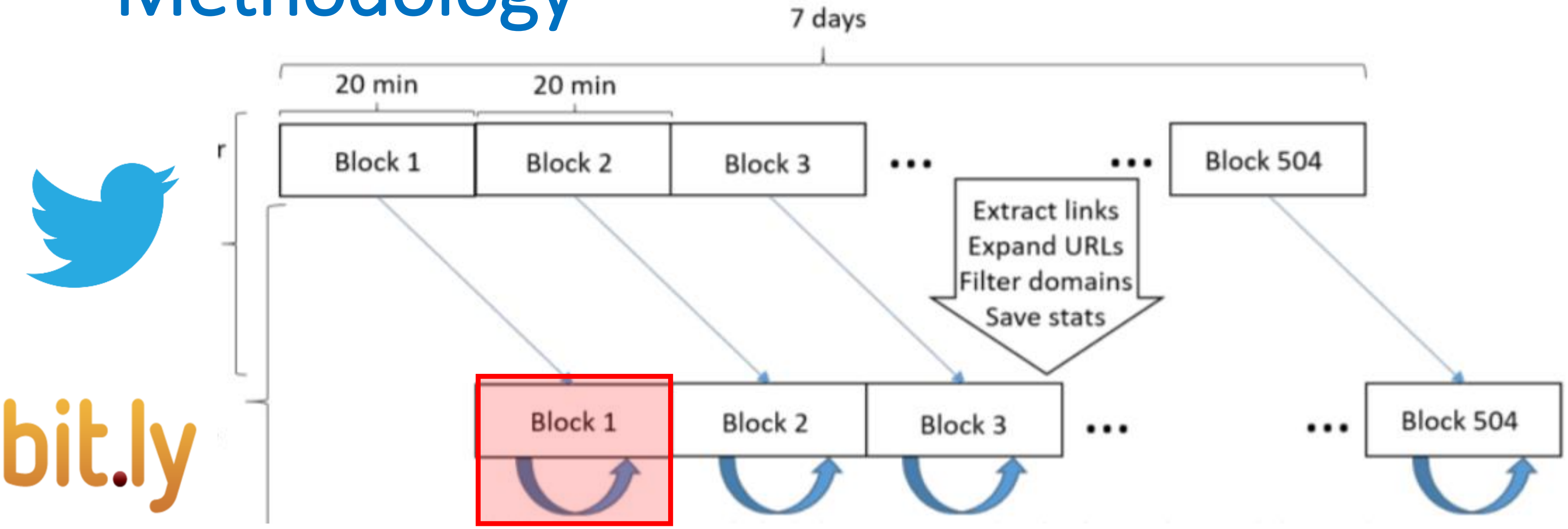
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days

Methodology



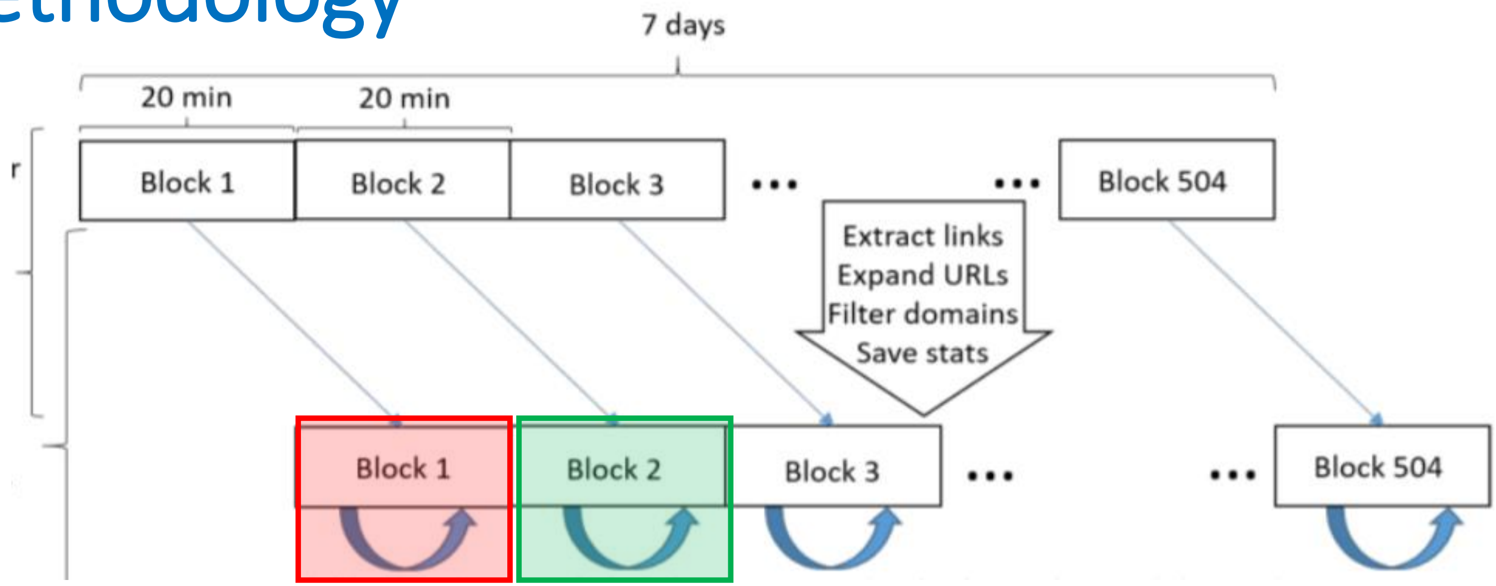
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days

Methodology



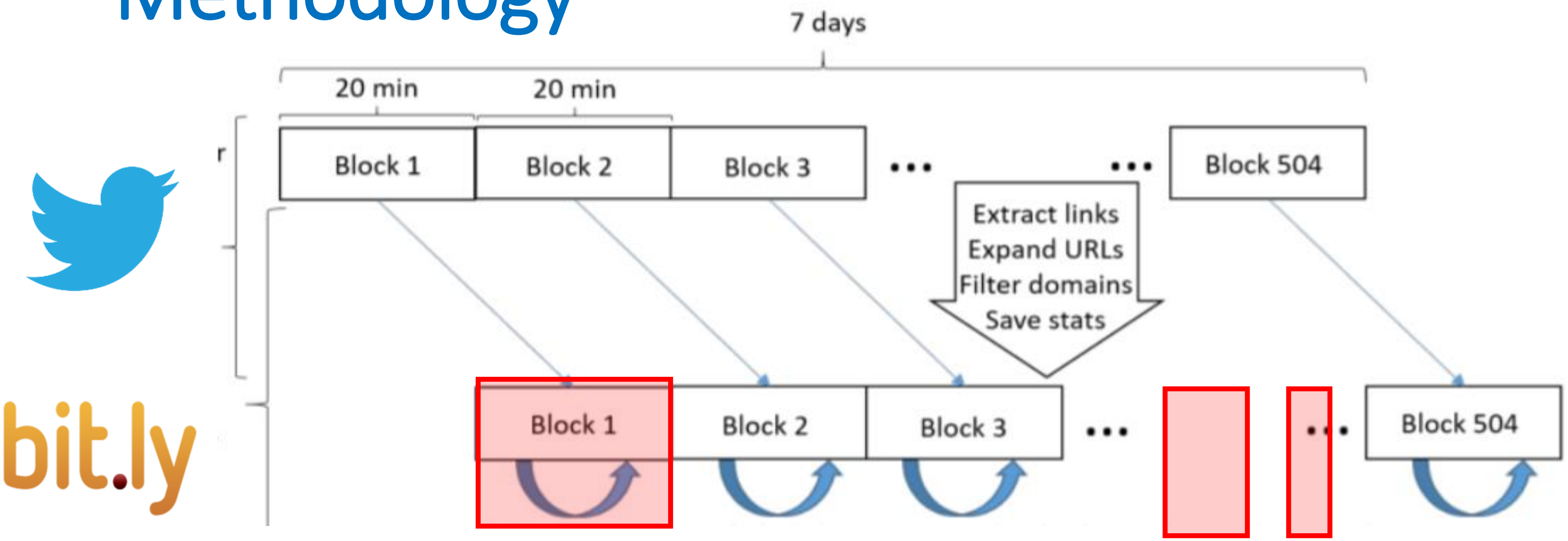
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days
- Longitudinal click statistics
  - Each block scheduled for collection every 2 hours for 5 days

Methodology



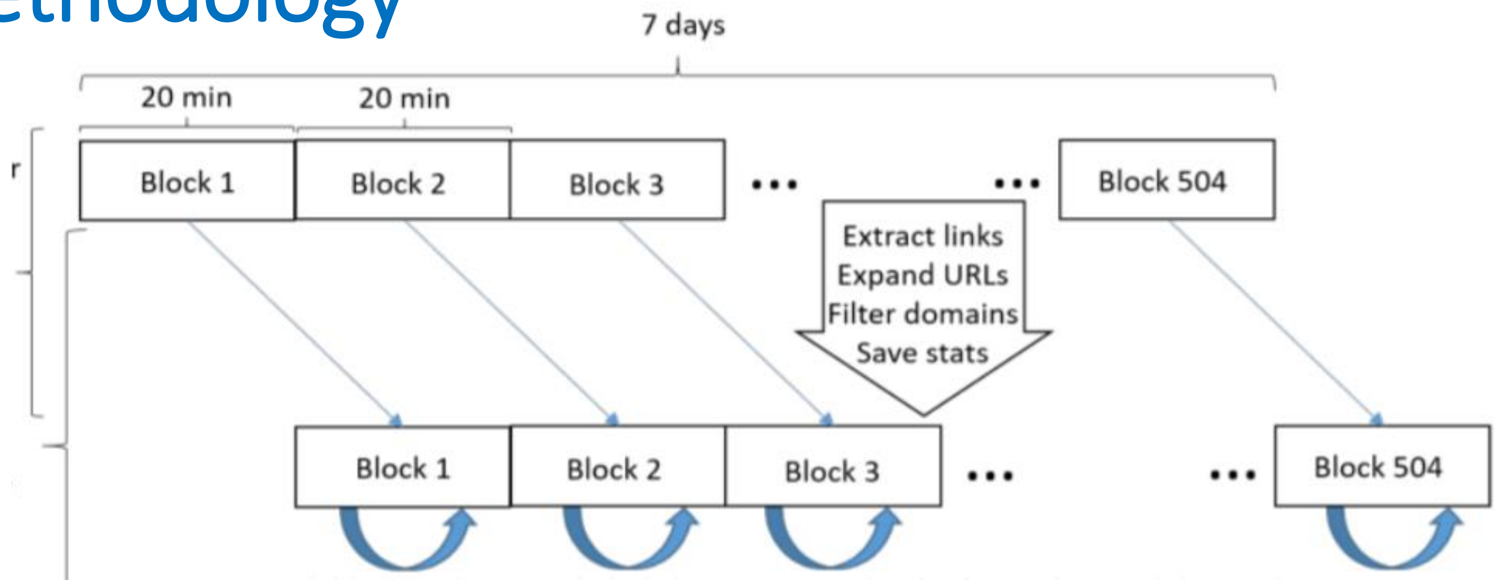
- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days
- Longitudinal click statistics
  - Each block scheduled for collection every 2 hours for 5 days

Methodology



- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days
- Longitudinal click statistics
  - Each block scheduled for collection every 2 hours for 5 days
- Careful sample frequency selection
  - To stay within rate limits, not all links sampled each time
  - Three sets per block: “top”, “random”, and “rest”
  - Sets sampled at different frequencies
  - Sets and their sizes designed to stay within rate limits (details/eqns. in paper)

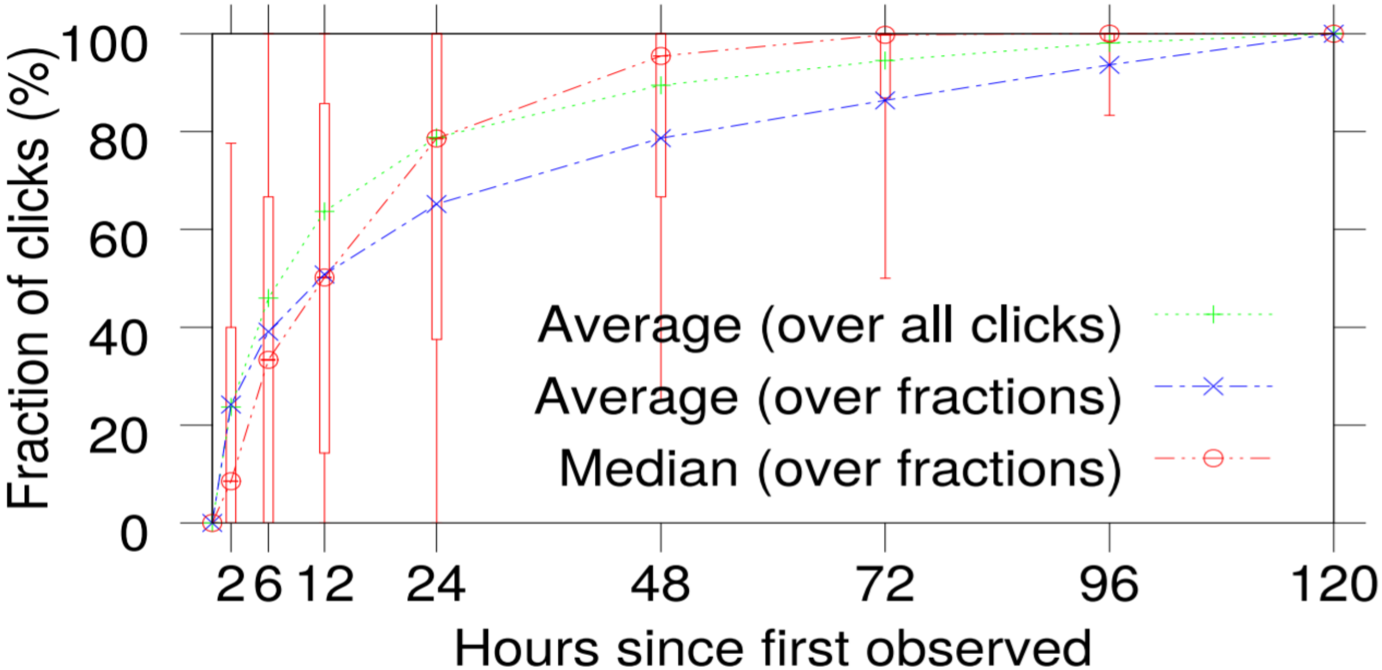
Methodology



- Collection of Bitly links to 7 pre-selected news website
  - 20-minute blocks (with latest tweets) collected over 7 days
- Longitudinal click statistics
  - Each block scheduled for collection every 2 hours for 5 days
- Careful sample frequency selection
  - To stay within rate limits, not all links sampled each time
  - Three sets per block: "top", "random", and "rest"
  - Sets sampled at different frequencies
  - Sets and their sizes designed to stay within rate limits (details/eqns. in paper)
- Complementing tweet statistics

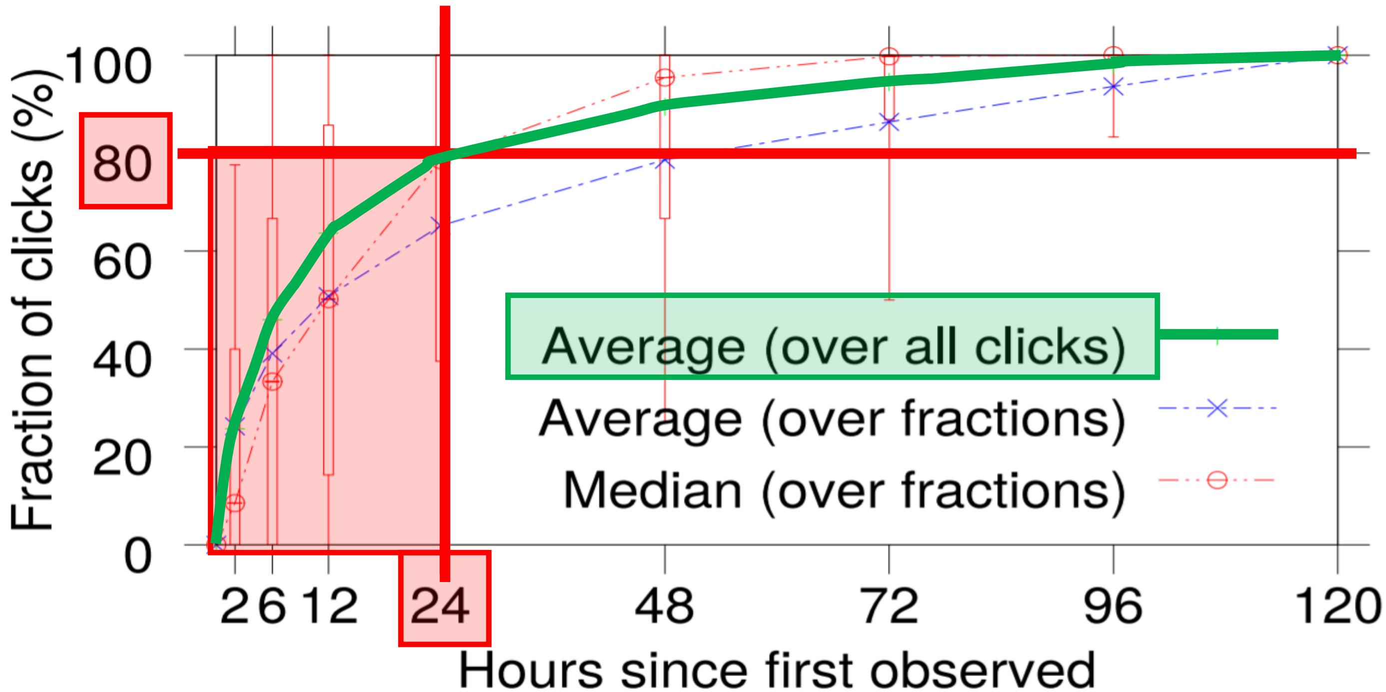
Clicks over time

Clicks over time



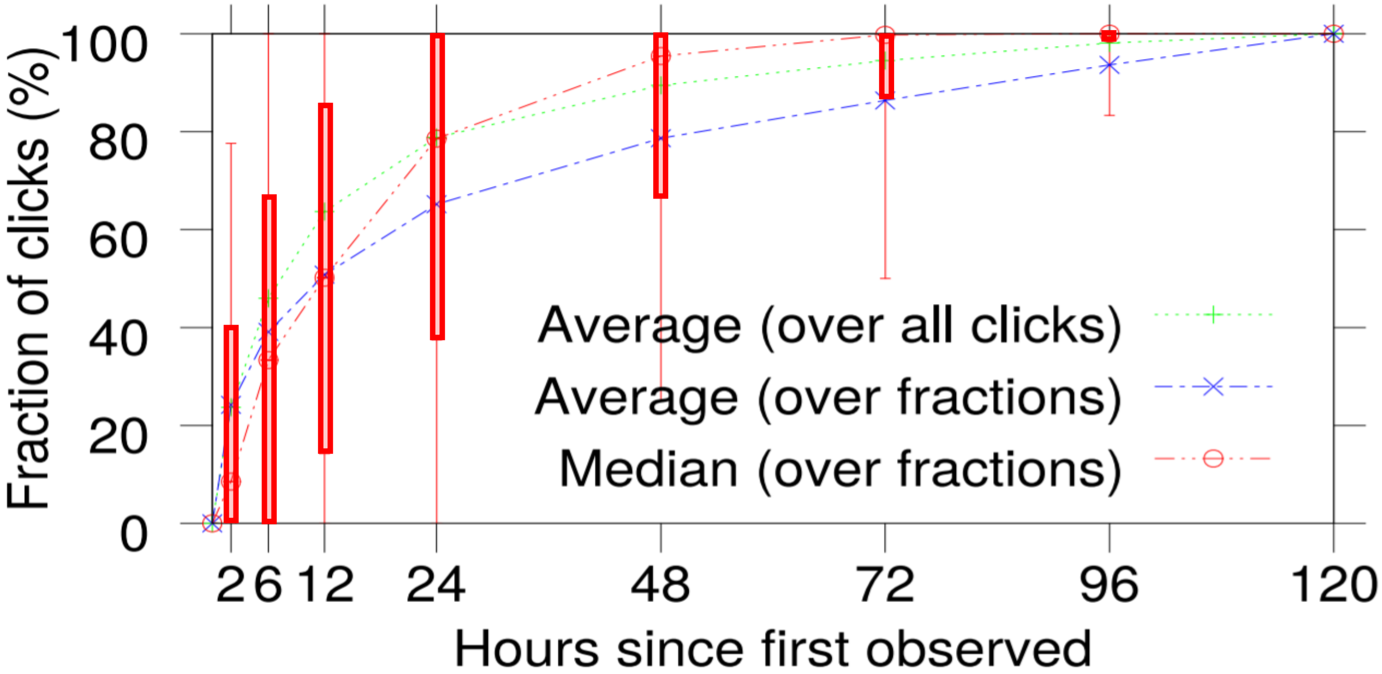
- 80% of all observed clicks occur within the first 24 hours

Clicks over time



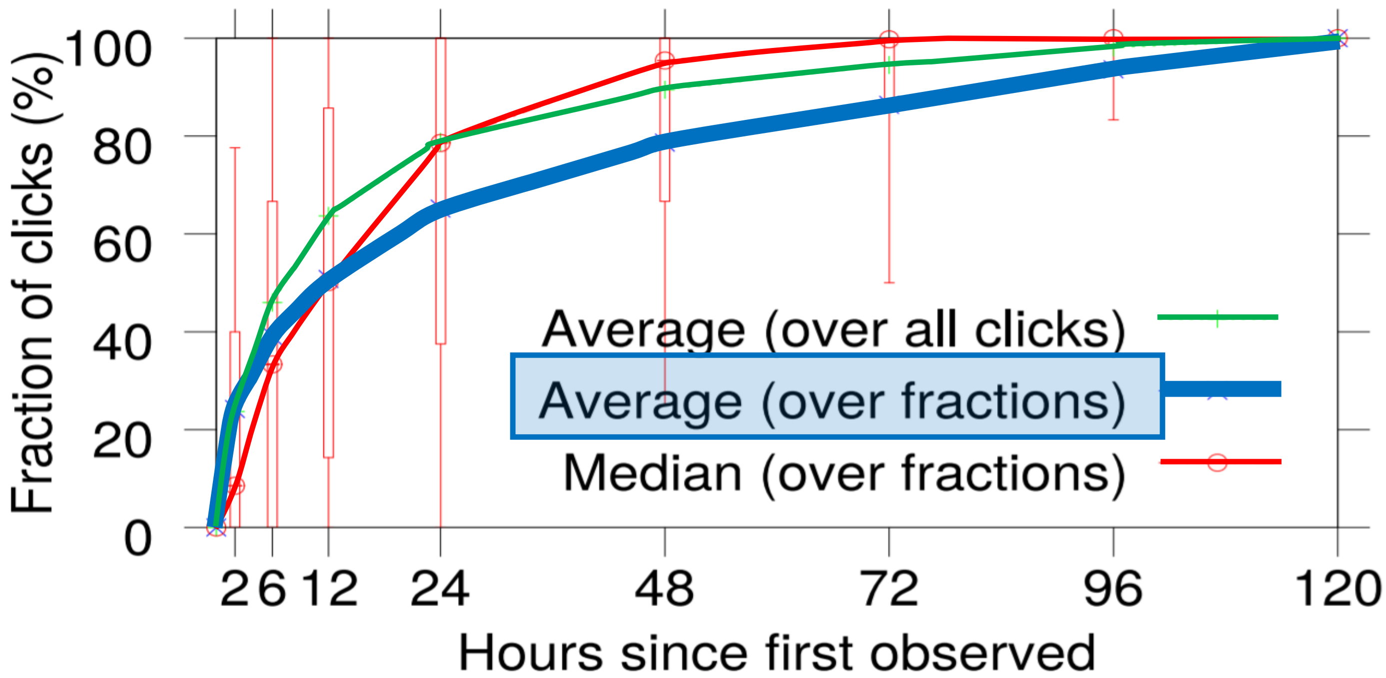
- 80% of all observed clicks occur within the first 24 hours

Clicks over time



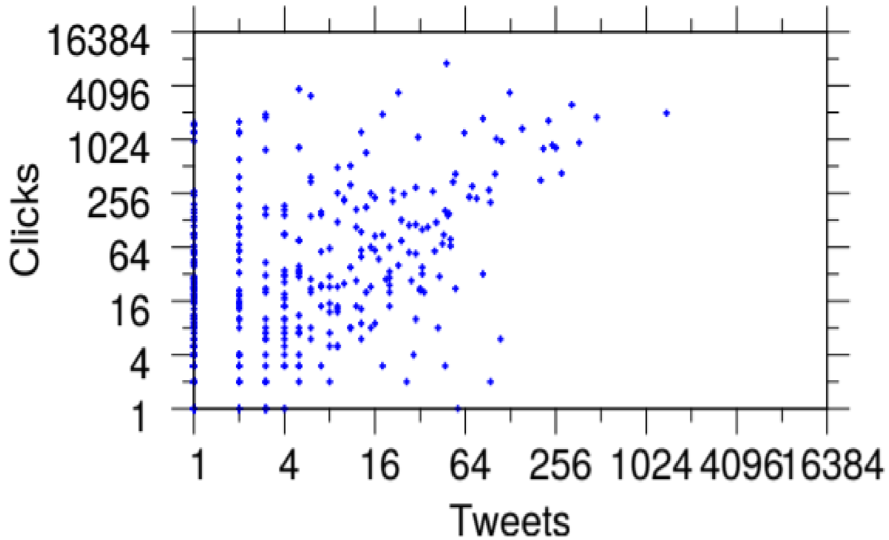
- 80% of all observed clicks occur within the first 24 hours
- Large variations across links

Clicks over time

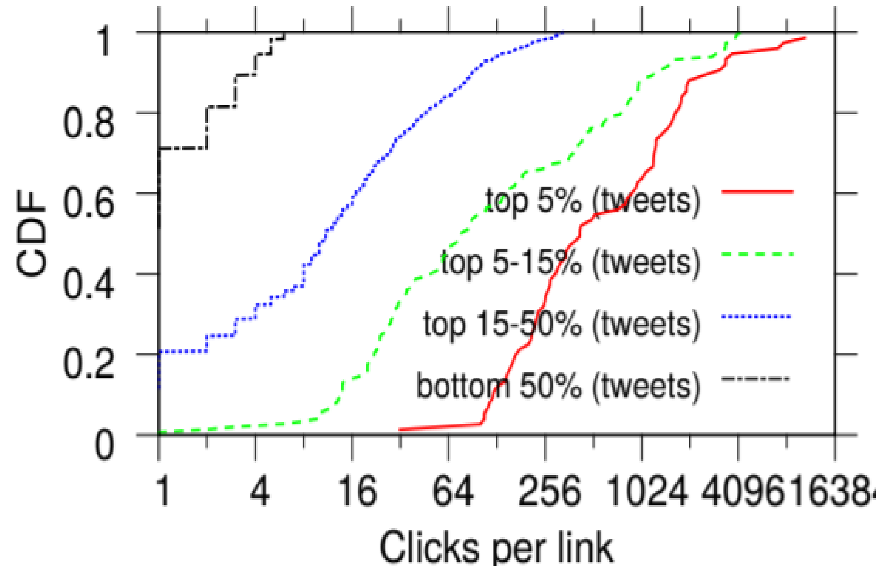


- 80% of all observed clicks occur within the first 24 hours
- Large variations across links
- Tail of less popular links sees a more even spread of clicks
  - Suggests that studies focusing only on popular articles (or tweets) may underestimate the duration of the news cycle and the time that many news articles are read after they are first published

Comparison with tweet data



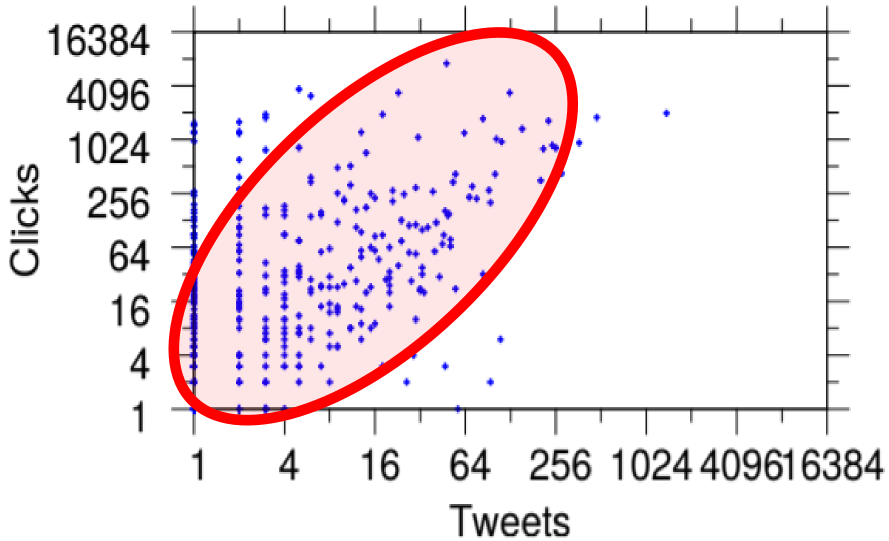
(a) Scatter



(b) CDFs

- Correlations between retweets and clicks, but significant differences

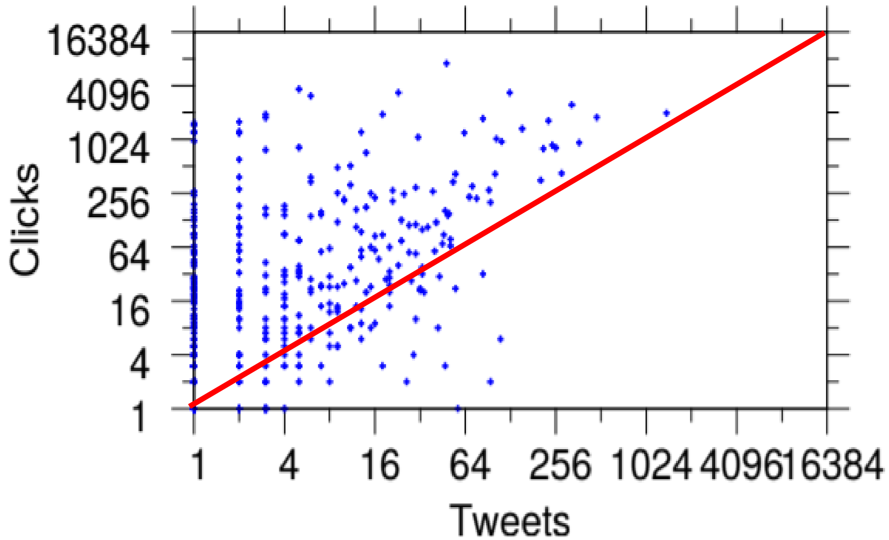
Comparison with tweet data



(a) Scatter

- Correlations between retweets and clicks, but significant differences

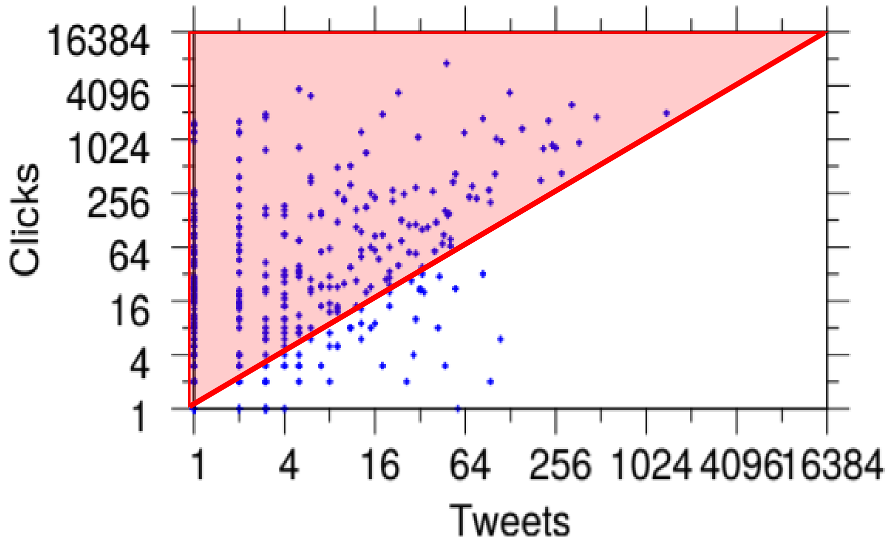
Comparison with tweet data



(a) Scatter

- Correlations between retweets and clicks, but significant differences
- Generally larger click volumes
  - Expected: Only subset of readers retweet what they read …
  - We may also miss earlier tweets (resulting in clicks)

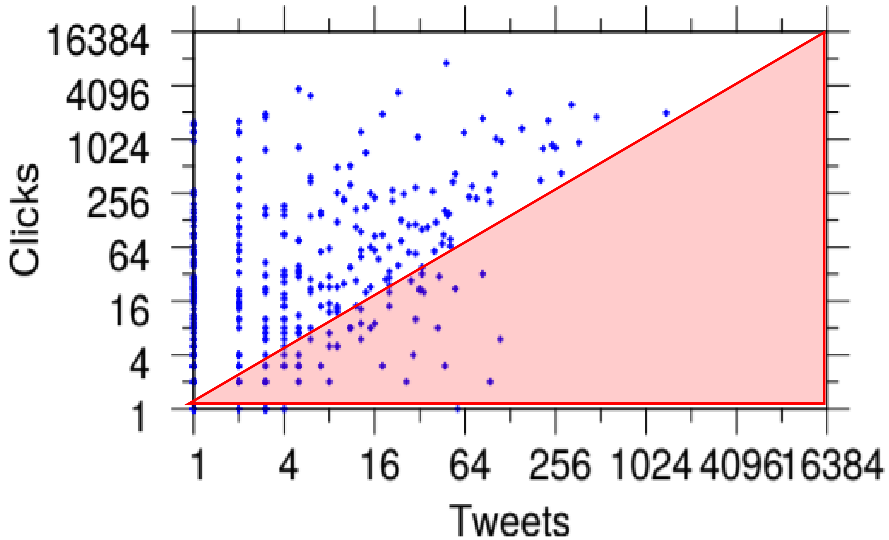
Comparison with tweet data



(a) Scatter

- Correlations between retweets and clicks, but significant differences
- Generally larger click volumes
  - Expected: Only subset of readers retweet what they read …
  - We may also miss earlier tweets (resulting in clicks)

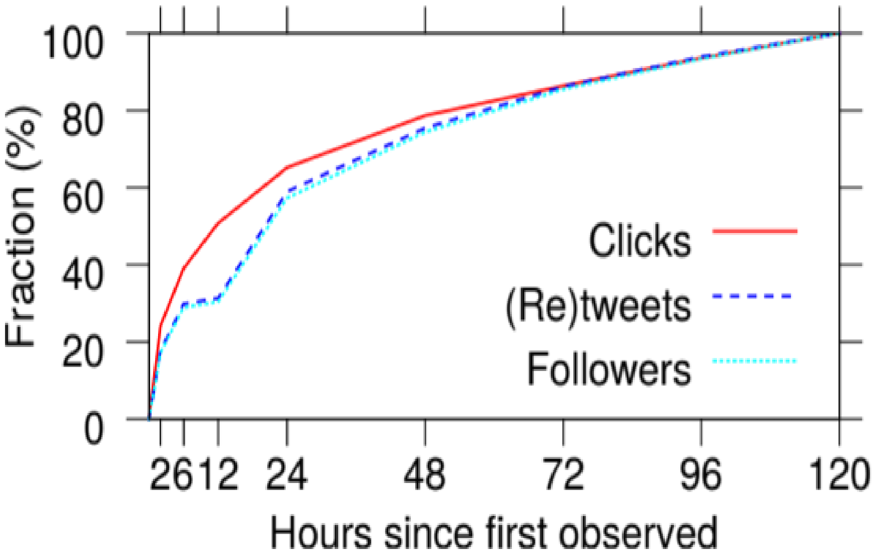
Comparison with tweet data



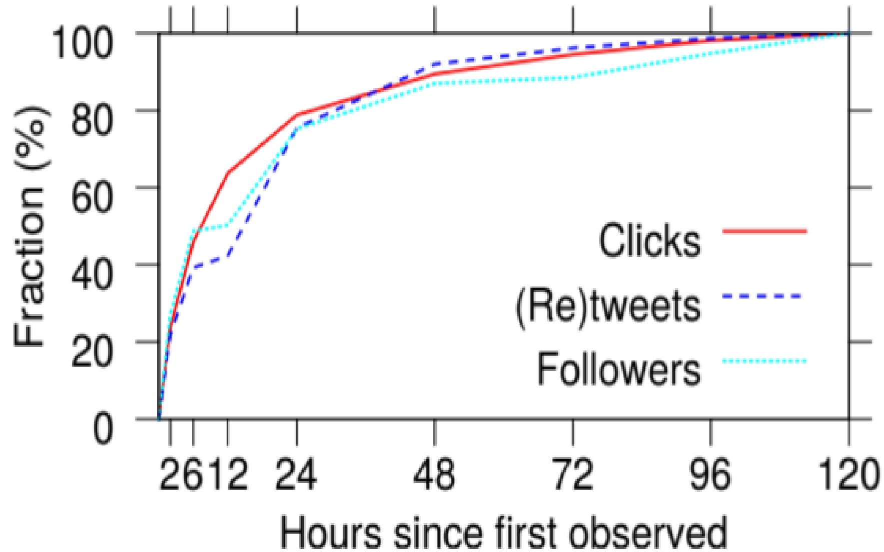
(a) Scatter

- Correlations between retweets and clicks, but significant differences
- Generally larger click volumes
  - Expected: Only subset of readers retweet what they read …
  - We may also miss earlier tweets (resulting in clicks)
- Significant set of tweets with more retweets than clicks
  - Some people (or bots) retweet the links without actually clicking the link.
  - This is clearly not good, as human sanity checking is an important tool to reduce the spreading of fake news

Comparison with tweet data



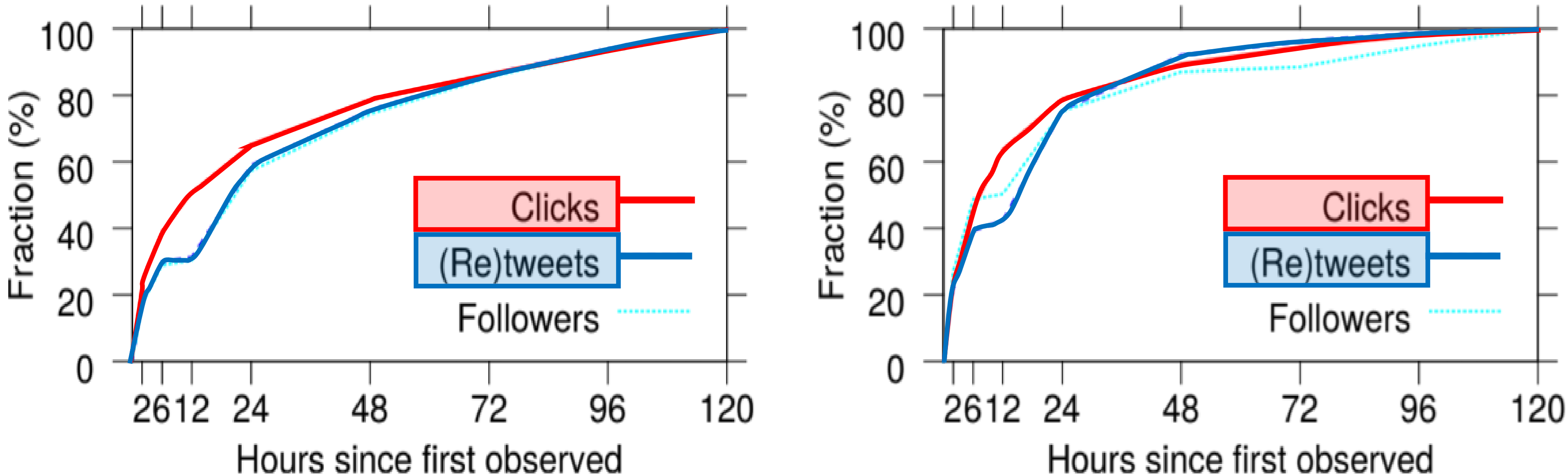
(a) Average (over fractions)



(b) Average (over all clicks)

- Clicks typically progress somewhat faster (at start) than retweets

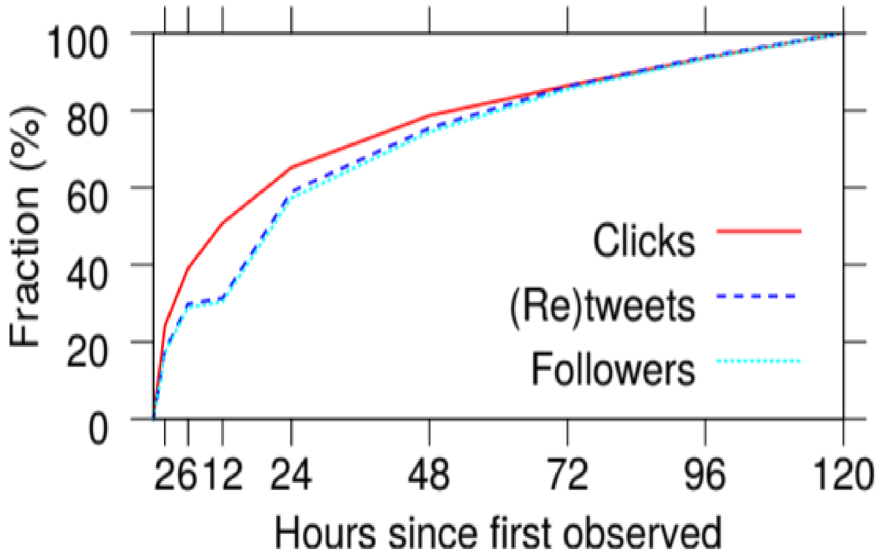
Comparison with tweet data



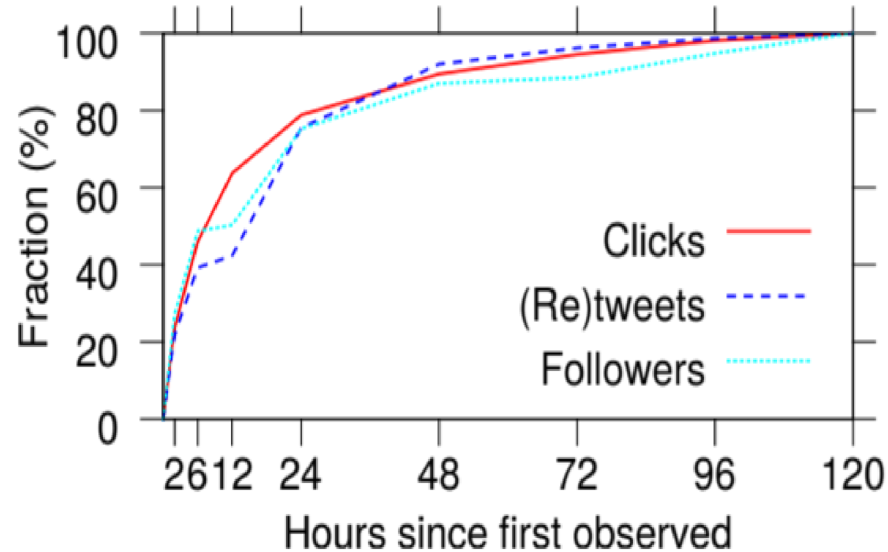
(a) Average (over fractions) (b) Average (over all clicks)

- Clicks typically progress somewhat faster (at start) than retweets

Comparison with tweet data



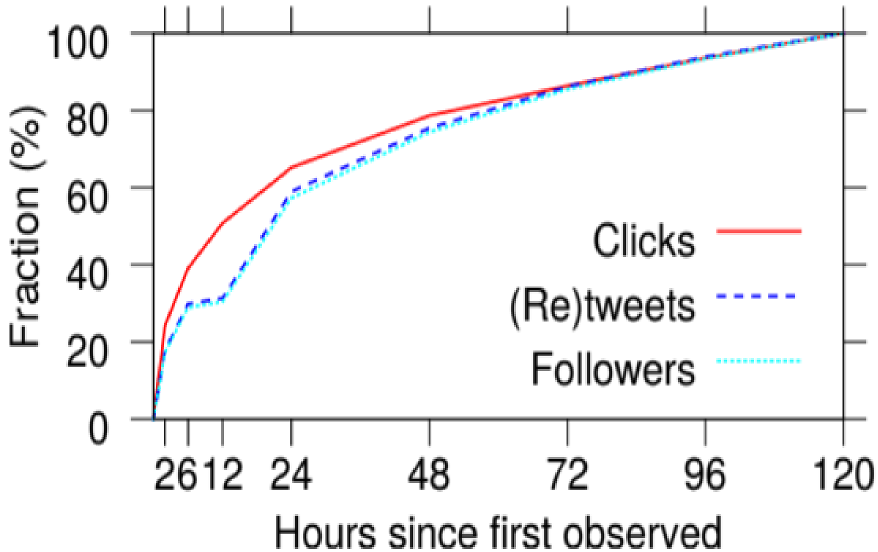
(a) Average (over fractions)



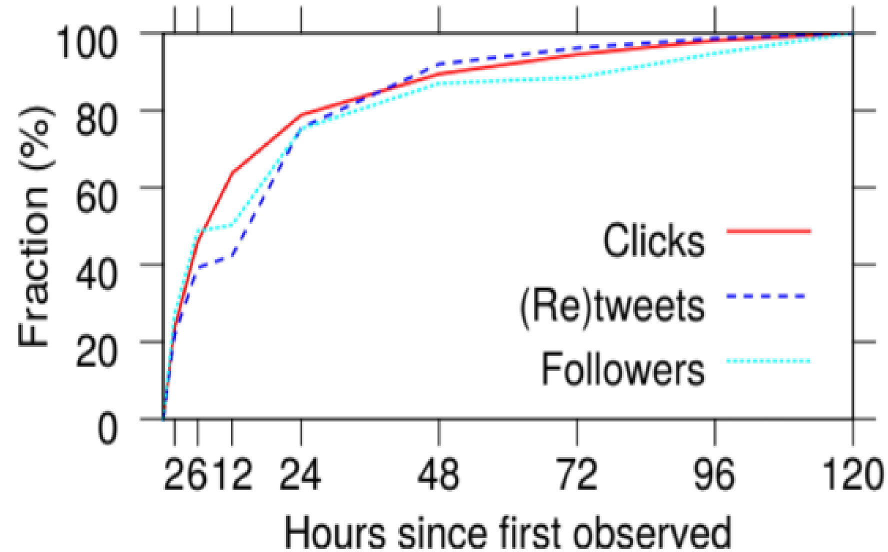
(b) Average (over all clicks)

- Clicks typically progress somewhat faster (at start) than retweets
- Also some subtle difference (see paper) that indicates that retweet data underestimate bias towards popular links/articles

Comparison with tweet data



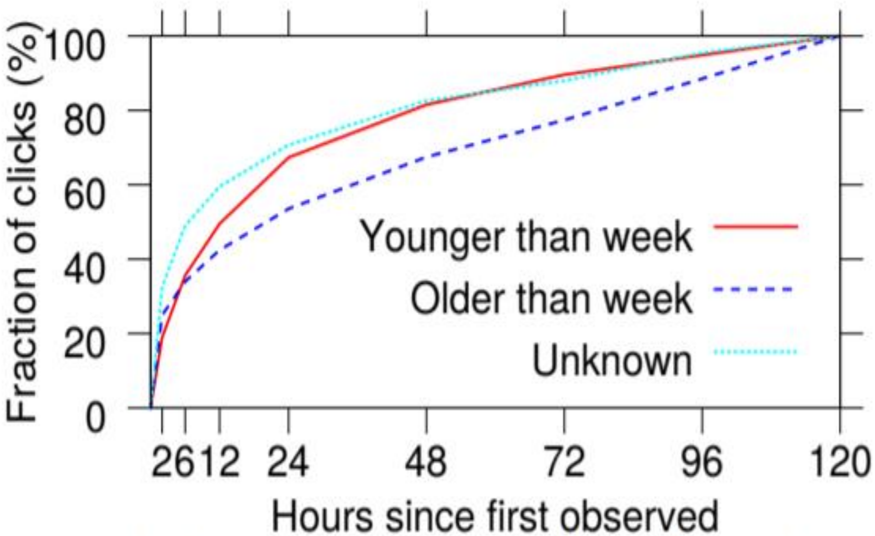
(a) Average (over fractions)



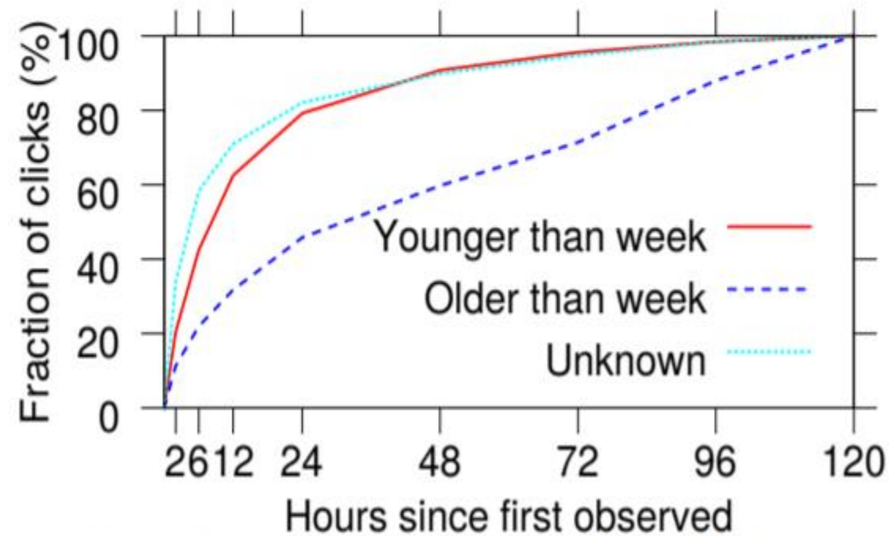
(b) Average (over all clicks)

- Clicks typically progress somewhat faster (at start) than retweets
- Also some subtle difference (see paper) that indicates that retweet data underestimate bias towards popular links/articles
  - These links are often highly shared in the beginning, but also accumulate reads/clicks (at a much slower rate) later …

Impact of age



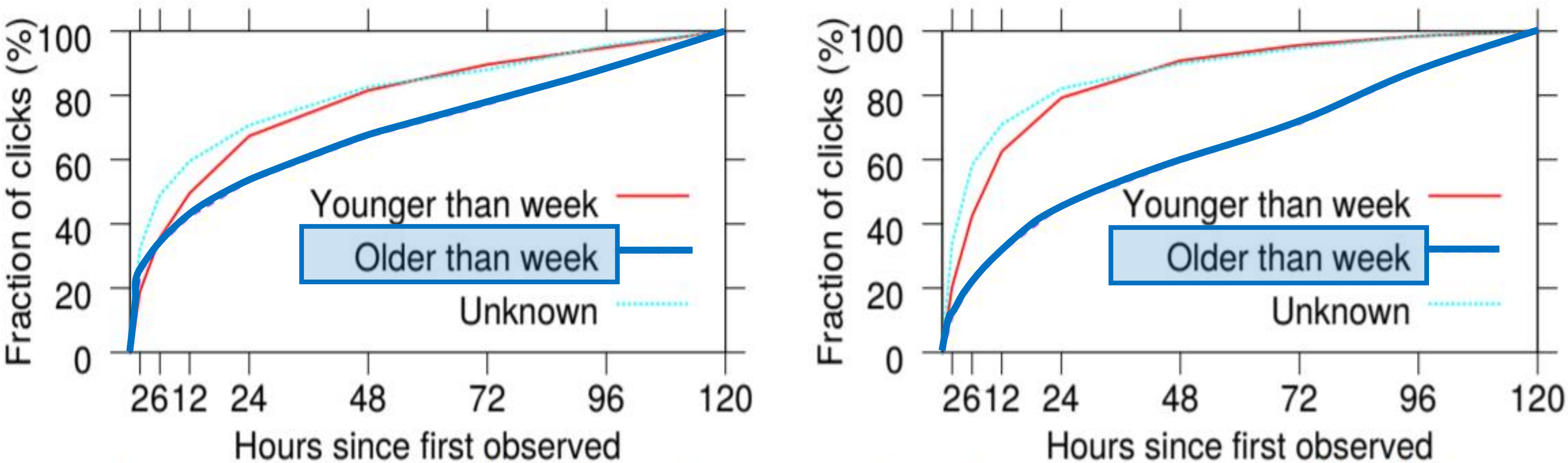
(a) Average (over fractions)



(b) Average (over all clicks)

- “Older” articles accumulating clicks at a much more uniform rate over the measurement duration

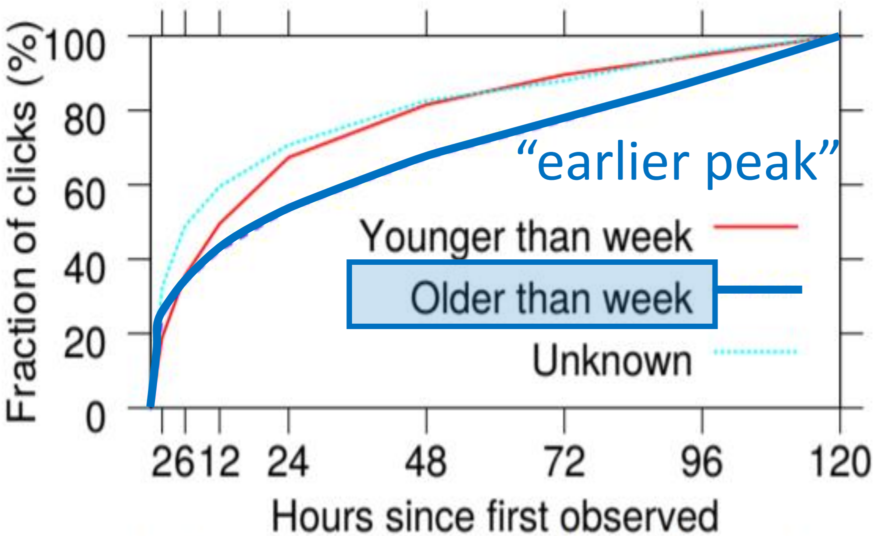
Impact of age



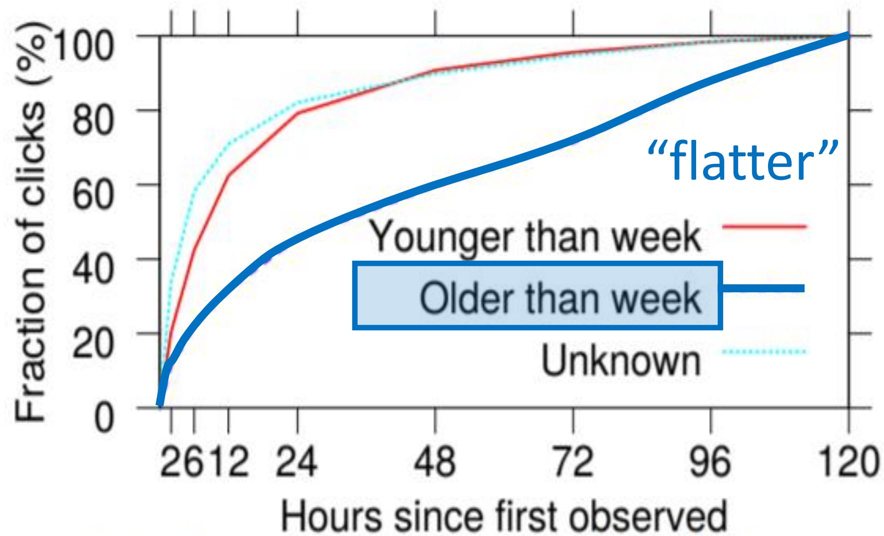
(a) Average (over fractions) (b) Average (over all clicks)

- “Older” articles accumulating clicks at a much more uniform rate over the measurement duration

Impact of age



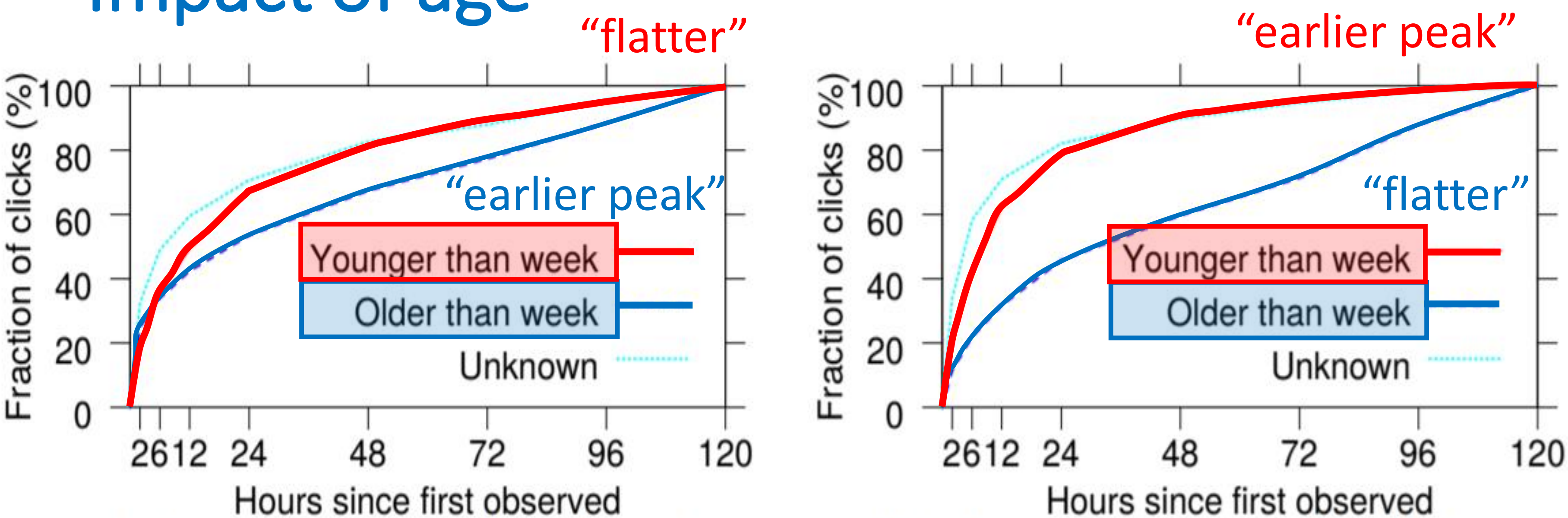
(a) Average (over fractions)



(b) Average (over all clicks)

- "Older" articles accumulating clicks at a much more uniform rate over the measurement duration
- For "older" articles, most of the clicks are associated with articles that do not appear to fade (as quickly) in popularity
  - The opposite is true for the "overall" set and for "younger" articles
  - Again, the "older" set contains more long-term popular articles

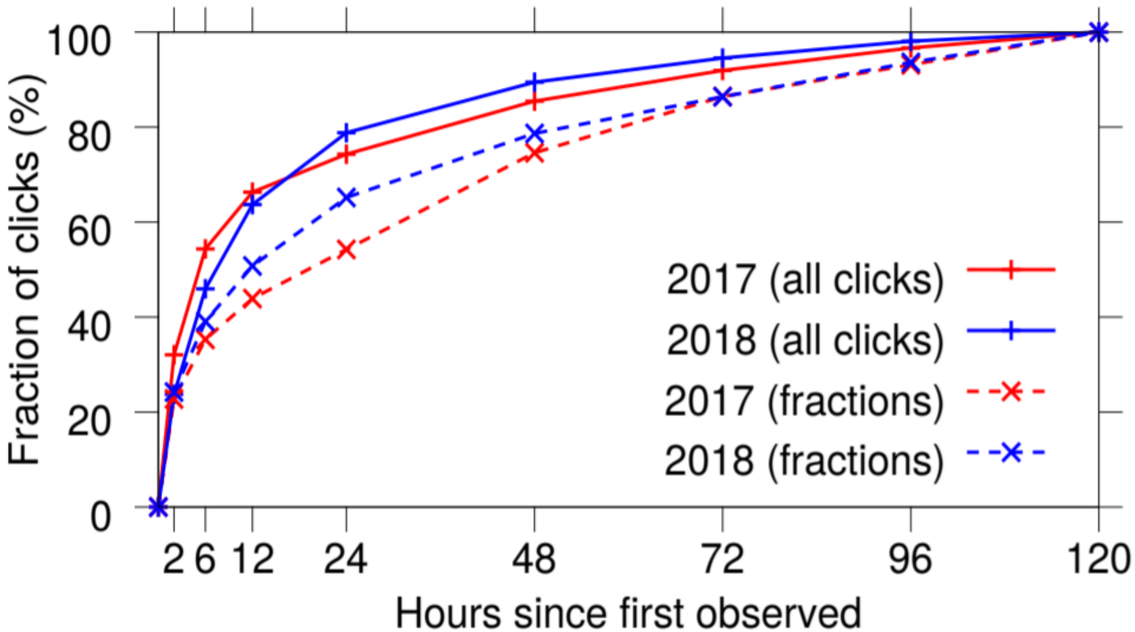
Impact of age

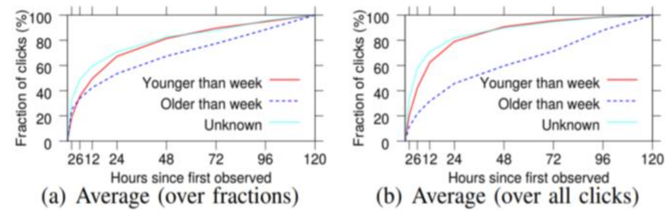


(a) Average (over fractions) (b) Average (over all clicks)

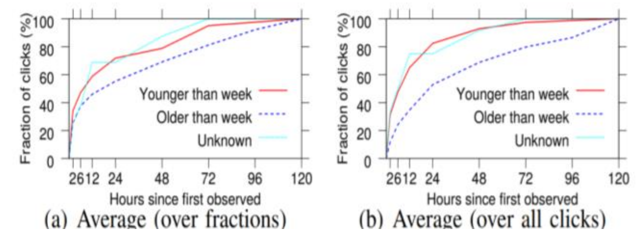
- “Older” articles accumulating clicks at a much more uniform rate over the measurement duration
- For “older” articles, most of the clicks are associated with articles that do not appear to fade (as quickly) in popularity
  - The opposite is true for the “overall” set and for “younger” articles
  - Again, the “older” set contains more long-term popular articles

Validation year-old data (from 2017)





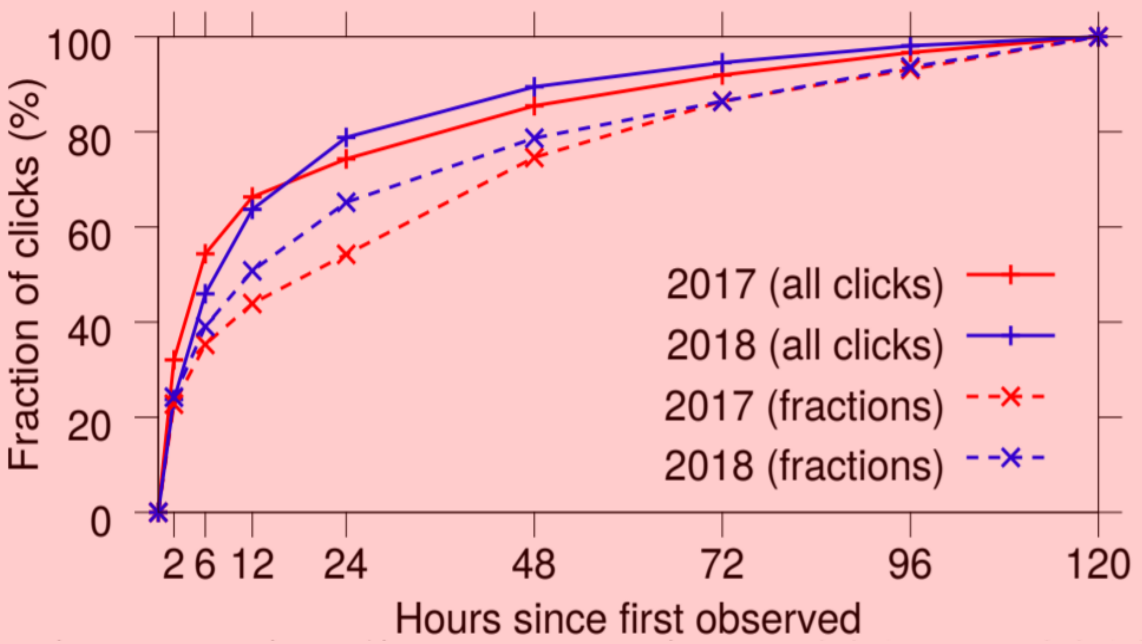
May 2018



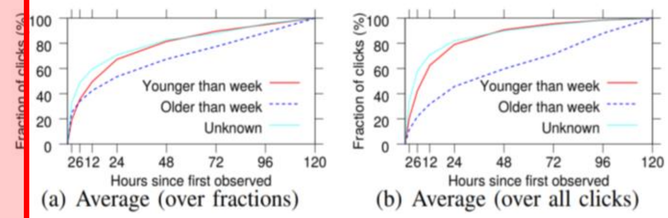
May 2017

- Identify invariants
  - The early peaks, the skew towards a subset of highly popular links, and the differences between links to articles of different age appear invariant

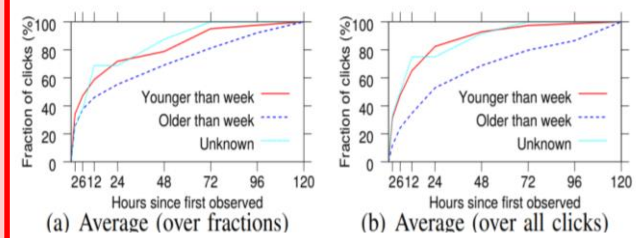
Validation year-old data (from 2017)



Overall



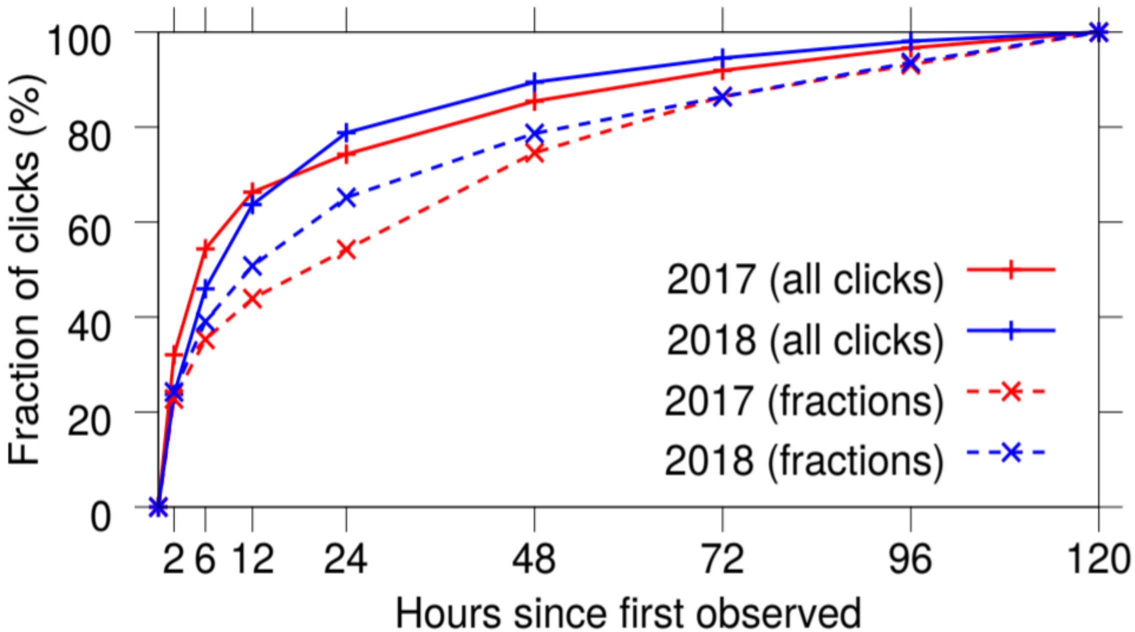
May 2018

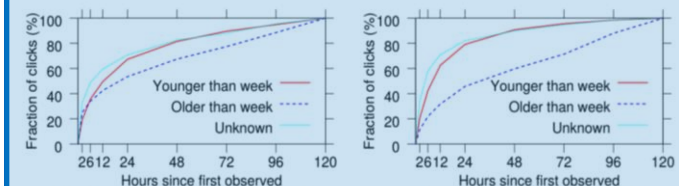


May 2017

- Identify invariants
  - The early peaks, the skew towards a subset of highly popular links, and the differences between links to articles of different age appear invariant

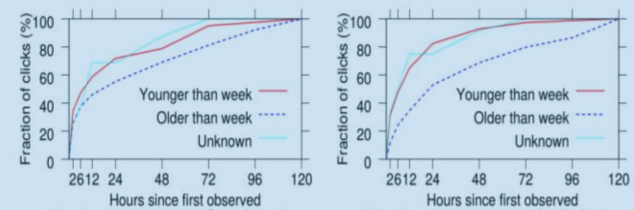
Validation year-old data (from 2017)





(a) Average (over fractions) (b) Average (over all clicks)

May 2018



(a) Average (over fractions) (b) Average (over all clicks)

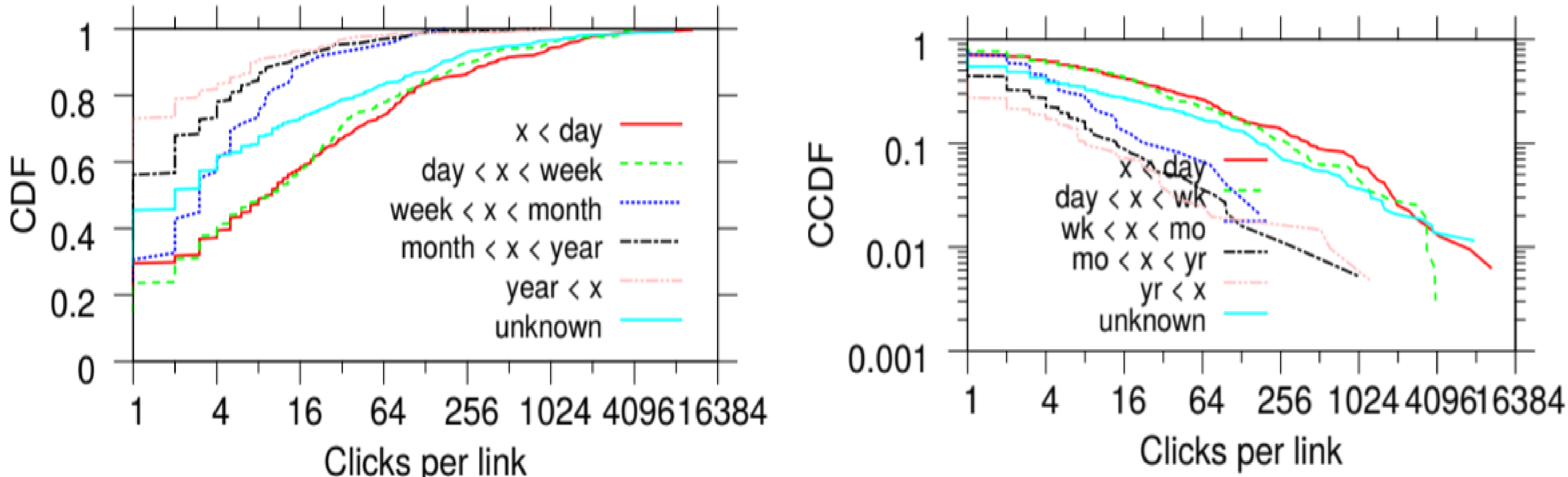
May 2017

Per-age group

- Identify invariants
  - The early peaks, the skew towards a subset of highly popular links, and the differences between links to articles of different age appear invariant

Age and long-term churn

Age-dependent popularity skew

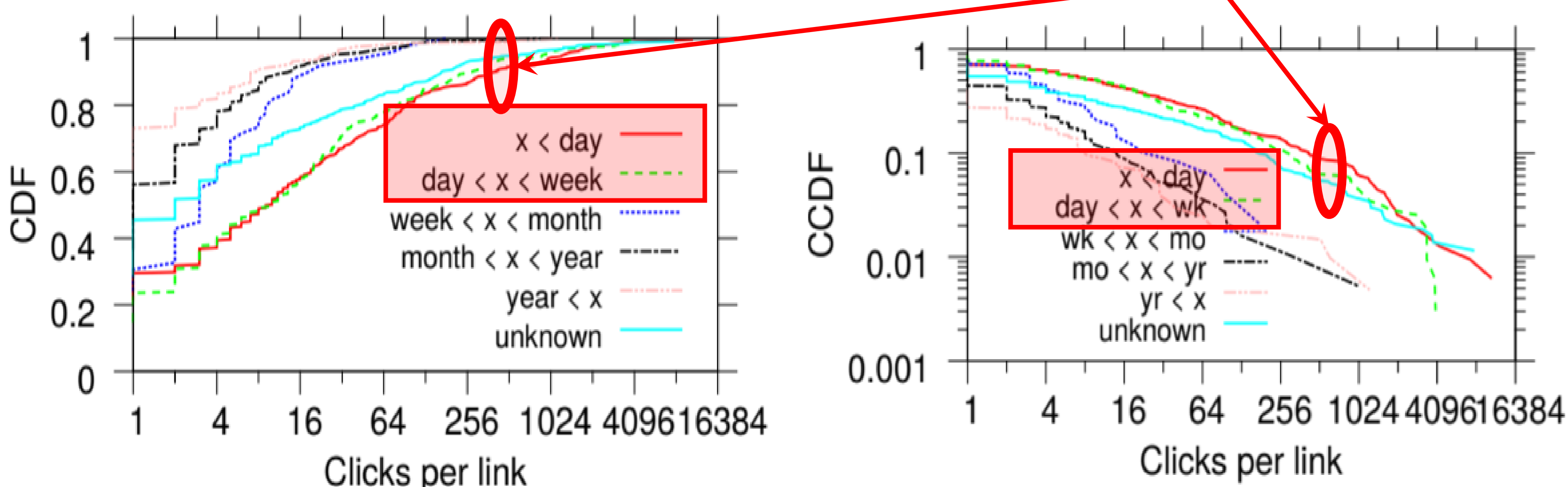


(a) CDF (b) CCDF

- While some differences within the two age-based sub-classes, the most substantial differences are between the categories themselves

Age-dependent popularity skew

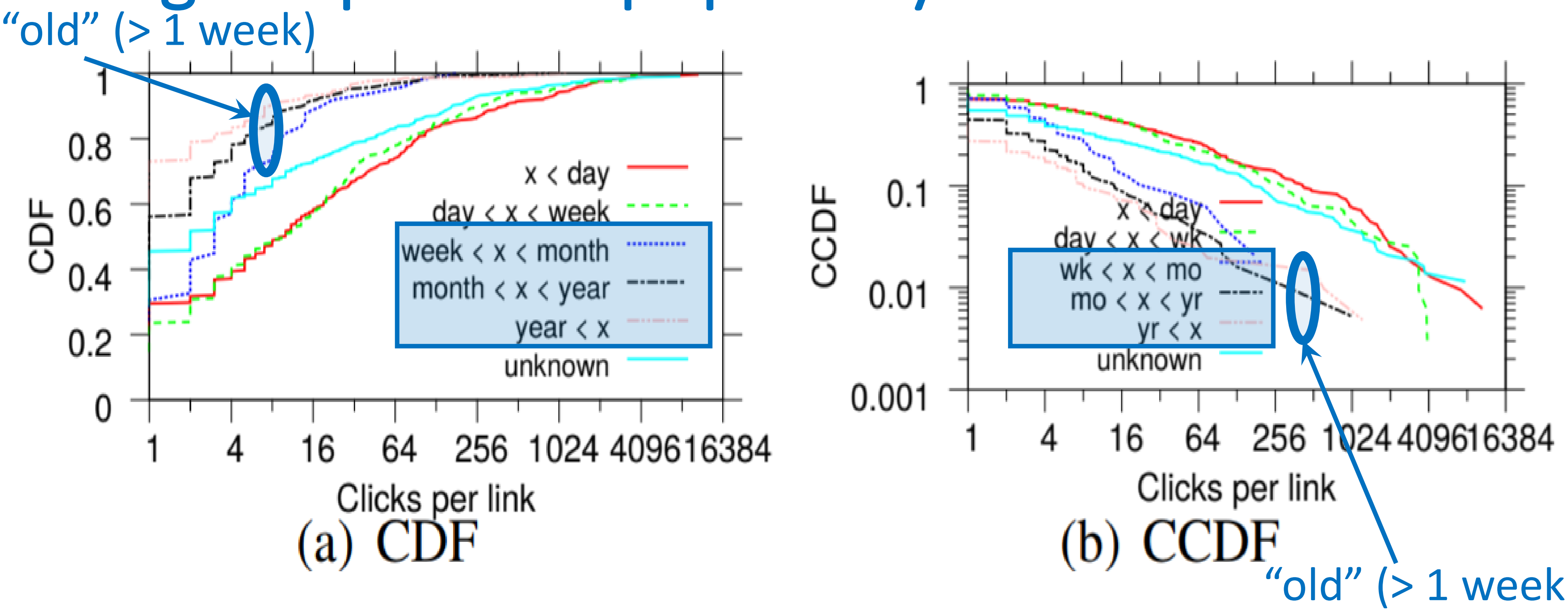
“young” (≤ 1 week)



(a) CDF (b) CCDF

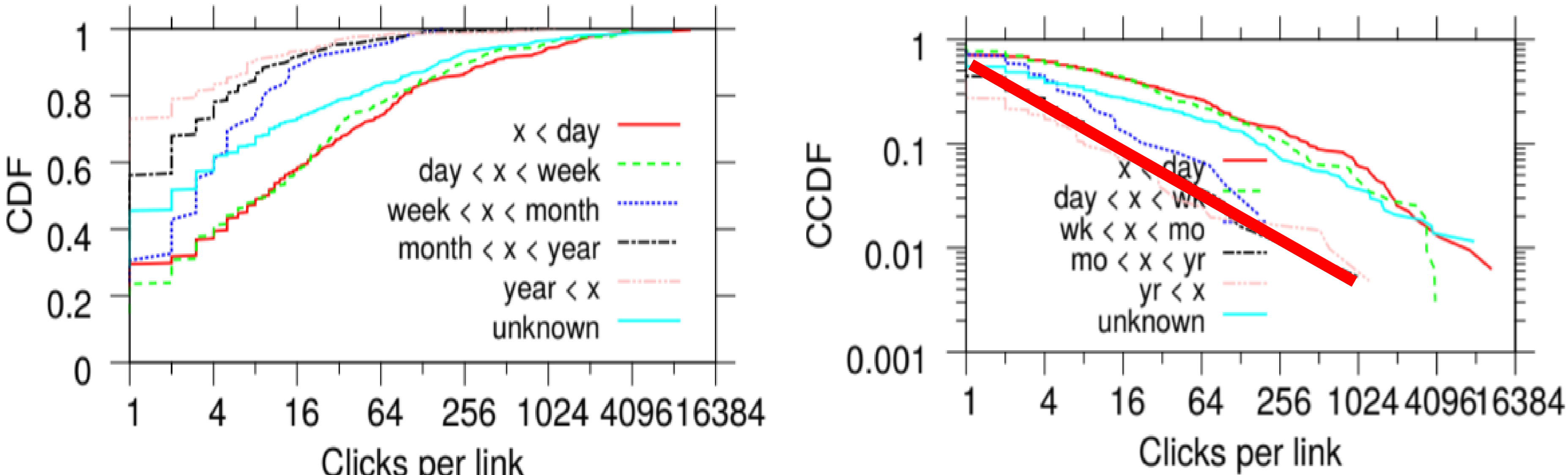
- While some differences within the two age-based sub-classes, the most substantial differences are between the categories themselves

Age-dependent popularity skew



- While some differences within the two age-based sub-classes, the most substantial differences are between the categories themselves

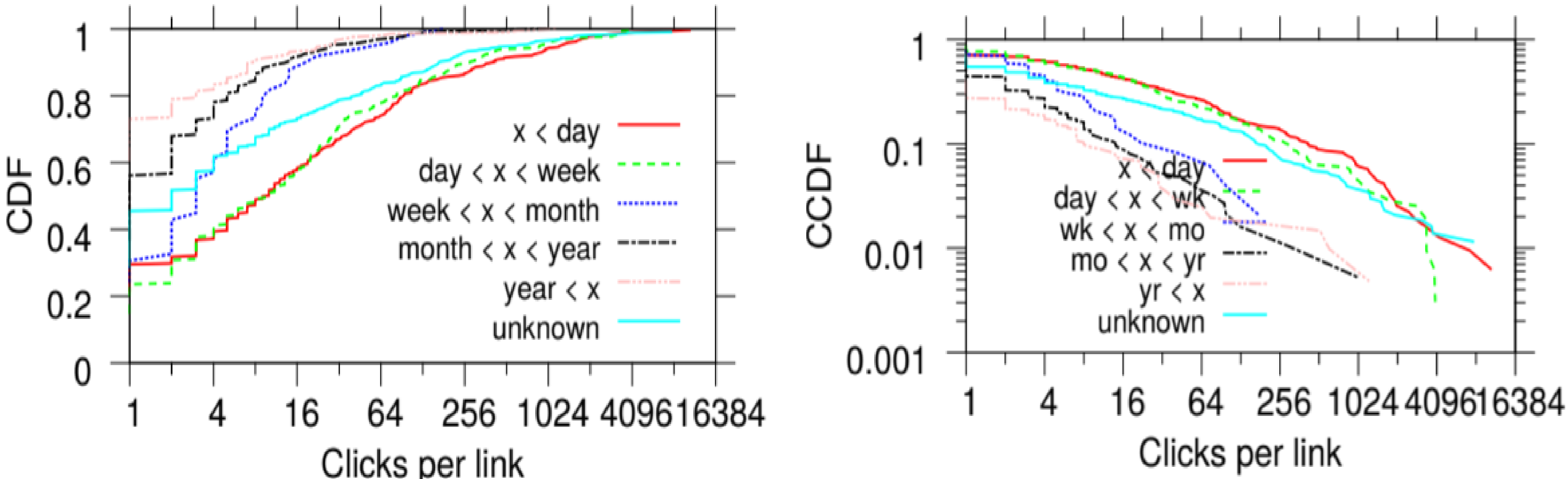
Age-dependent popularity skew



(a) CDF (b) CCDF

- While some differences within the two age-based sub-classes, the most substantial differences are between the categories themselves
- For the two “oldest” classes, the CCDFs shows relatively straight-line behavior, suggesting a power-law-like popularity skew

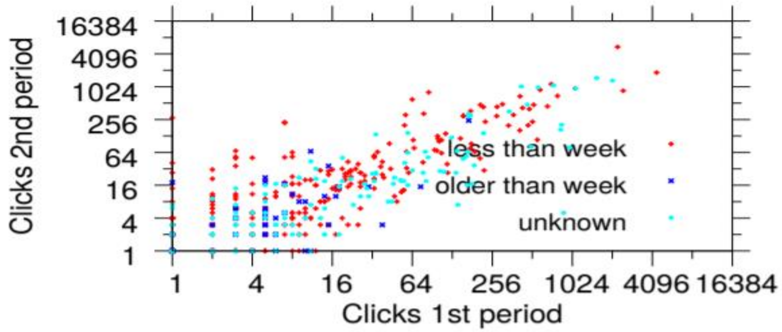
Age-dependent popularity skew



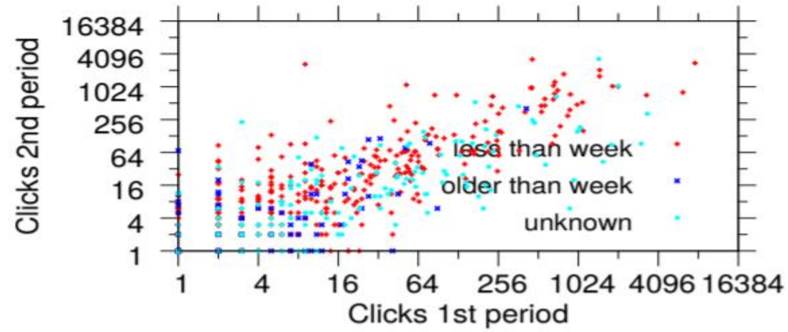
(a) CDF (b) CCDF

- While some differences within the two age-based sub-classes, the most substantial differences are between the categories themselves
- For the two "oldest" classes, the CCDFs shows relatively straight-line behavior, suggesting a power-law-like popularity skew
- For the "younger" articles, there is relatively higher popularity churn

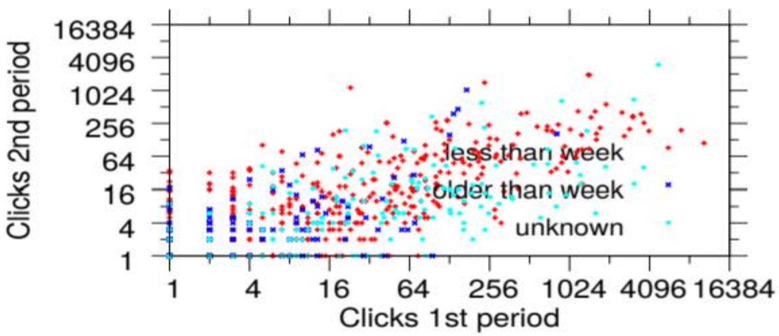
Age-dependent churn



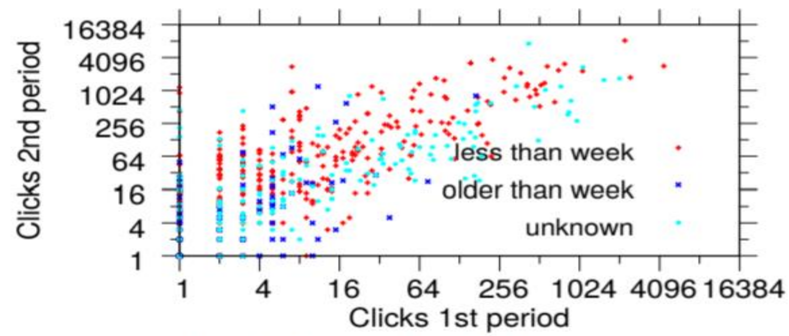
(a) 0-to-2 vs 2-to-6 hrs



(b) 0-6 vs 6-24 hrs



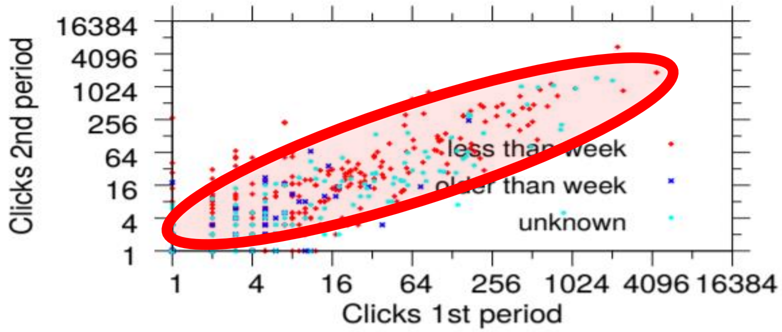
(c) 0-24 vs 24-120 hrs



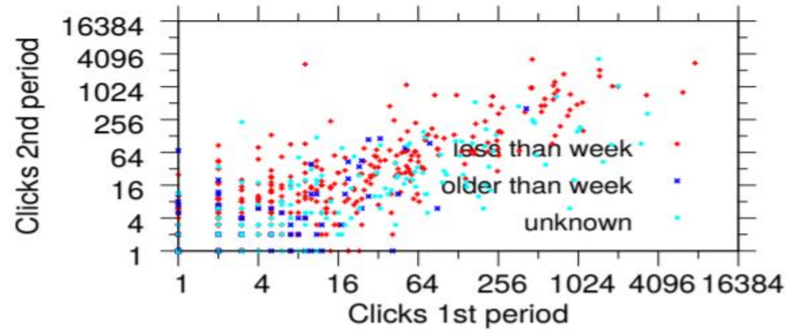
(d) 0-2 vs 2-120 hrs

- Increasing churn among both the “younger” and “older” articles
  - In contrast, for YouTube videos, long-term popularity has been found to reduce the churn over time [Borghol et al., 2011]

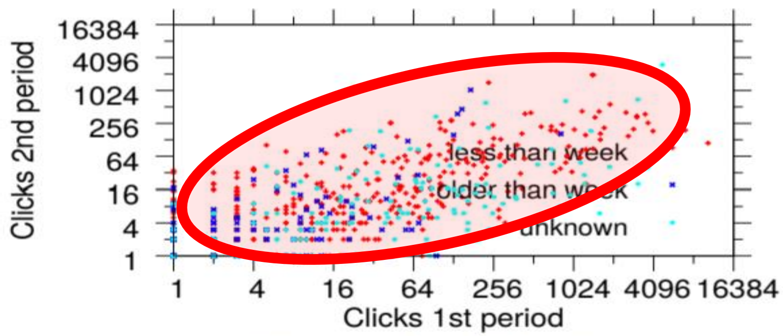
Age-dependent churn



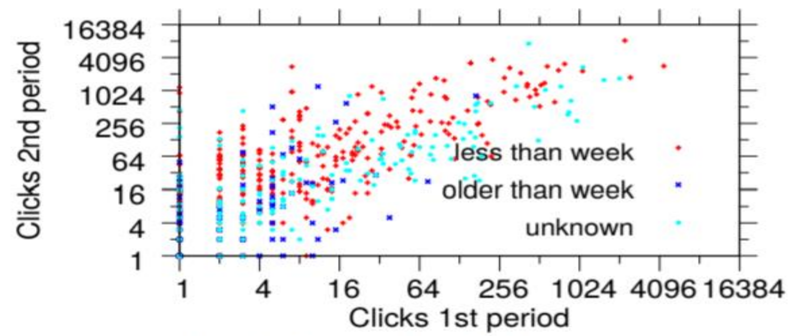
(a) 0-to-2 vs 2-to-6 hrs



(b) 0-6 vs 6-24 hrs



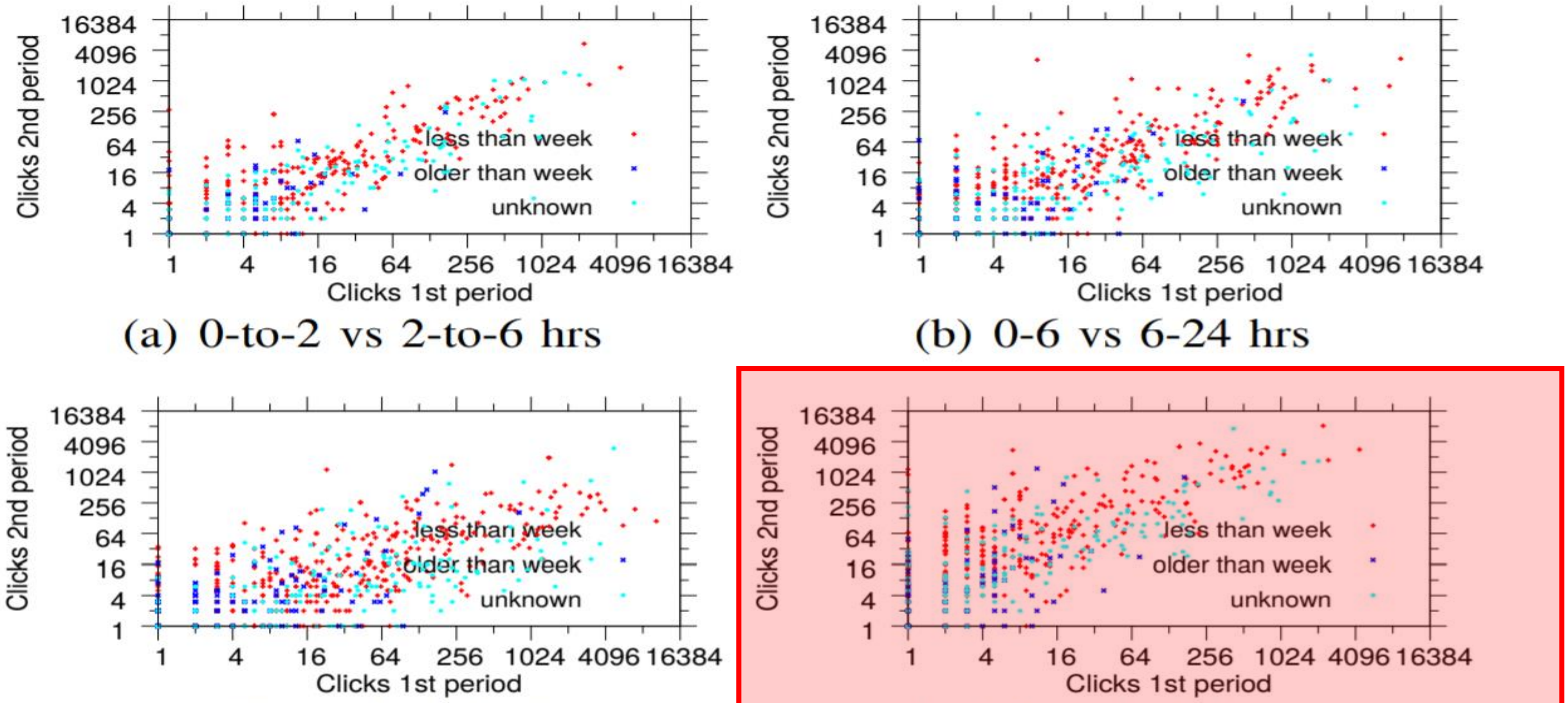
(c) 0-24 vs 24-120 hrs



(d) 0-2 vs 2-120 hrs

- Increasing churn among both the “younger” and “older” articles
  - In contrast, for YouTube videos, long-term popularity has been found to reduce the churn over time [Borghol et al., 2011]

Age-dependent churn



(a) 0-to-2 vs 2-to-6 hrs

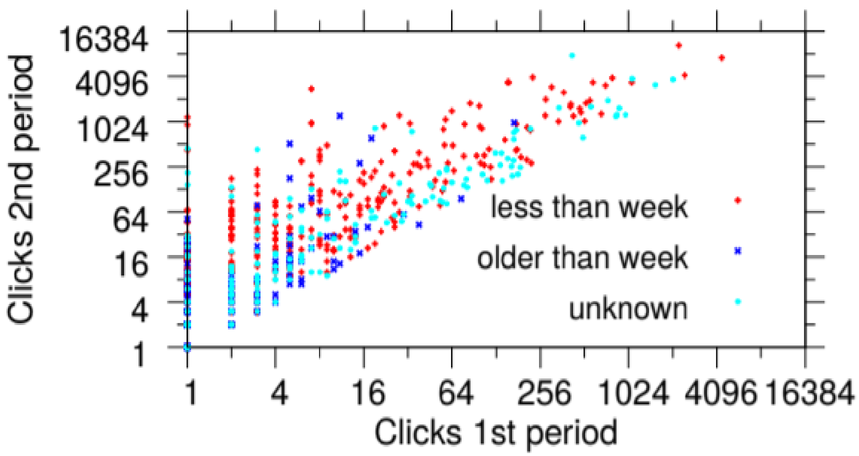
(b) 0-6 vs 6-24 hrs

(c) 0-24 vs 24-120 hrs

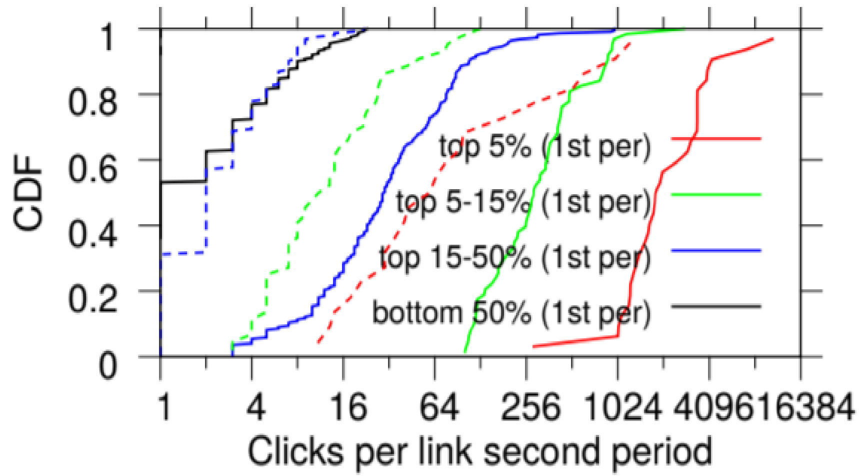
(d) 0-2 vs 2-120 hrs

- Increasing churn among both the “younger” and “older” articles
  - In contrast, for YouTube videos, long-term popularity has been found to reduce the churn over time [Borghol et al., 2011]
- A short initial interval have been found to be a good predictor of the clicks over the remainder of the time period

Age-dependent churn



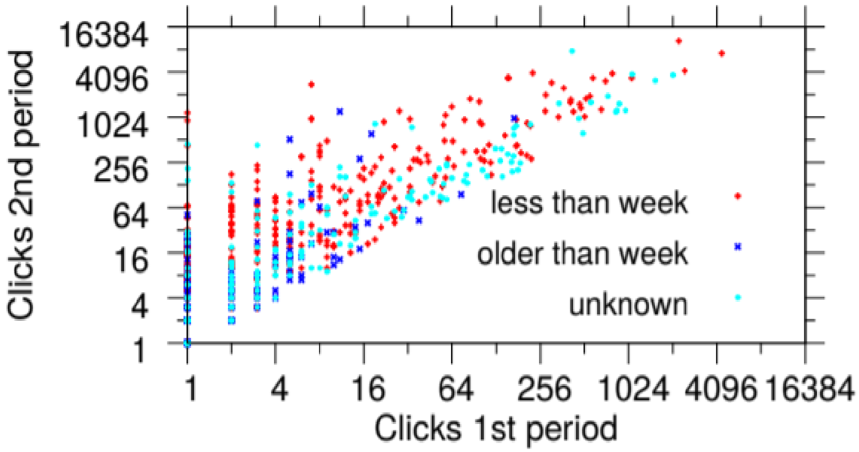
(a) Scatter



(b) Threshold CDFs

Even the clicks observed over a very short interval (e.g., 2 hrs) provides better insight into the actual information reach over a longer time period (e.g., 120 hrs) than the retweet do (even if using the same 120 hrs)

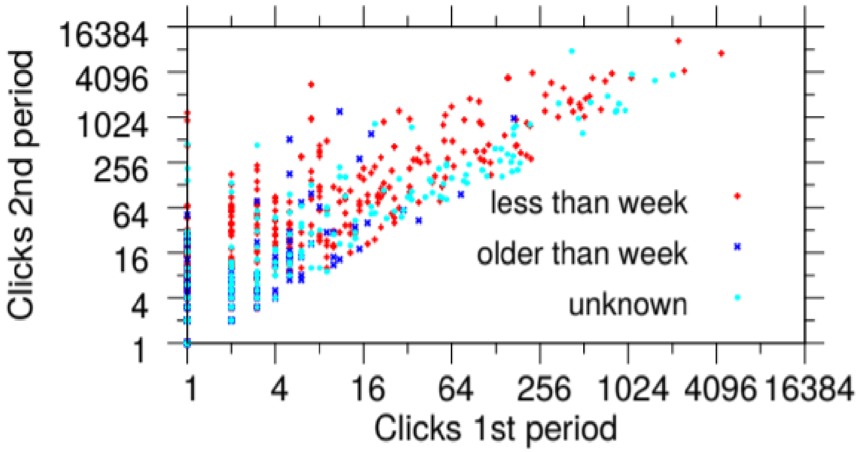
Age-dependent churn



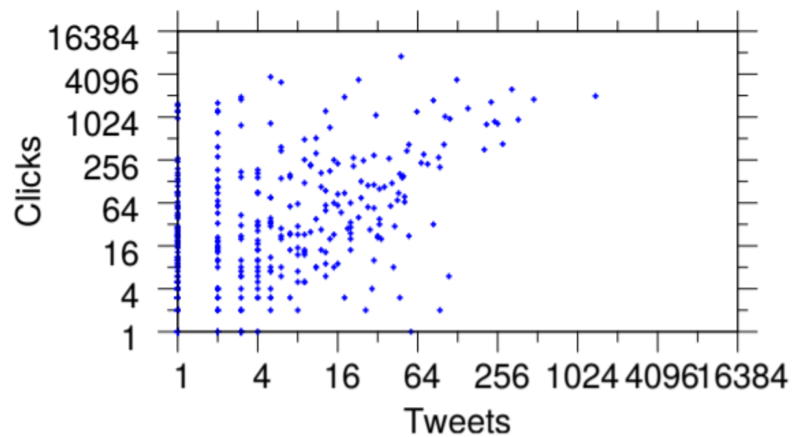
(a) Scatter

Even the clicks observed over a very short interval (e.g., 2 hrs) provides better insight into the actual information reach over a longer time period (e.g., 120 hrs) than the retweet do (even if using the same 120 hrs)

Age-dependent churn

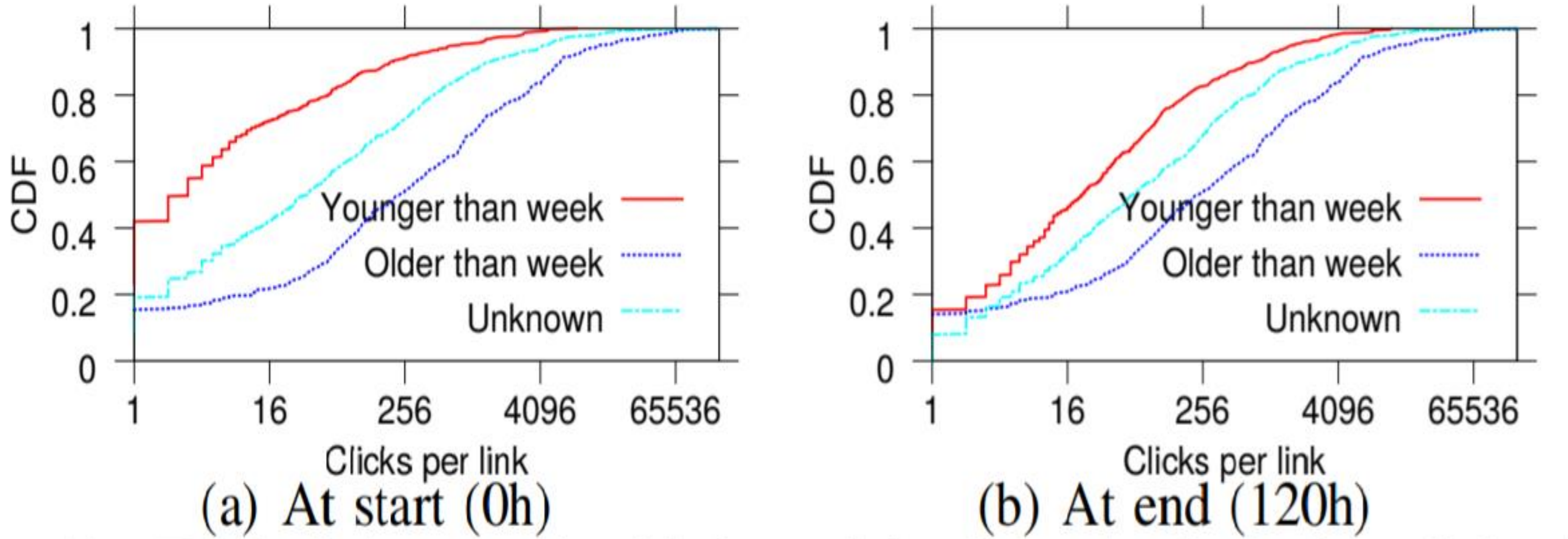


… compare with retweets vs clicks …



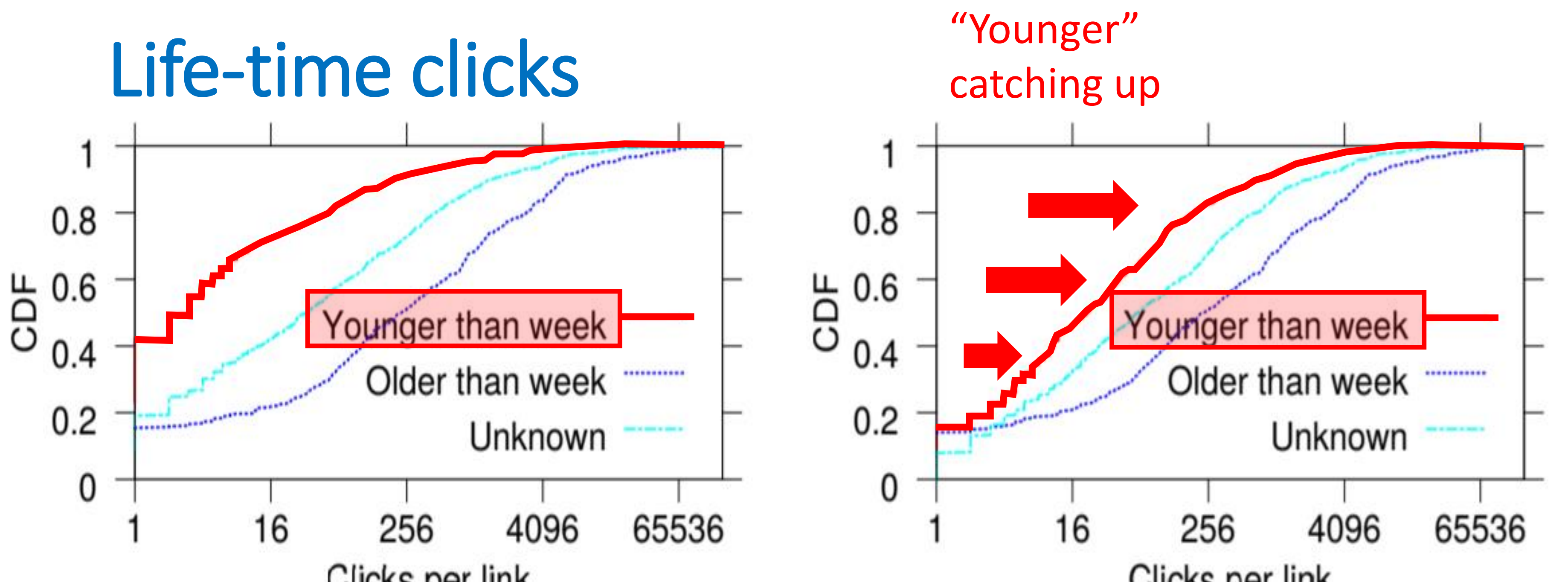
Even the clicks observed over a very short interval (e.g., 2 hrs) provides better insight into the actual information reach over a longer time period (e.g., 120 hrs) than the retweet do (even if using the same 120 hrs)

Life-time clicks



- “Younger” links gain on the lifetime clicks observed for the “older” links, closing the gap

Life-time clicks

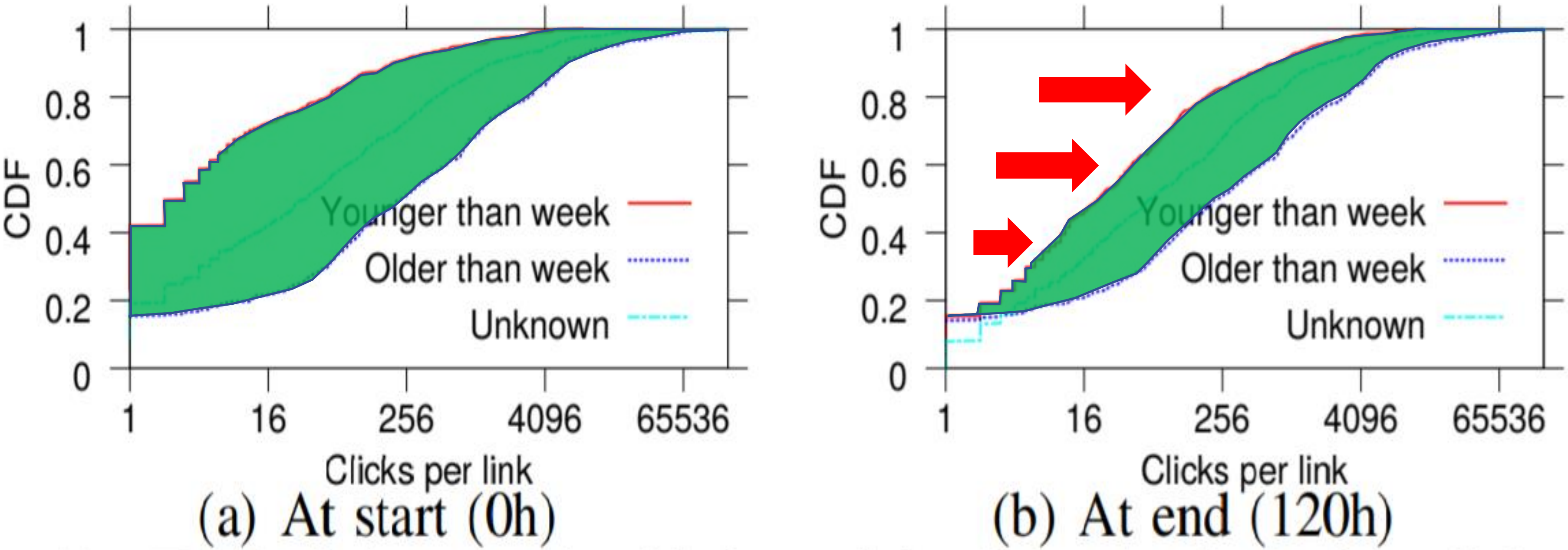


(a) At start (0h) (b) At end (120h)

- "Younger" links gain on the lifetime clicks observed for the "older" links, closing the gap

Life-time clicks

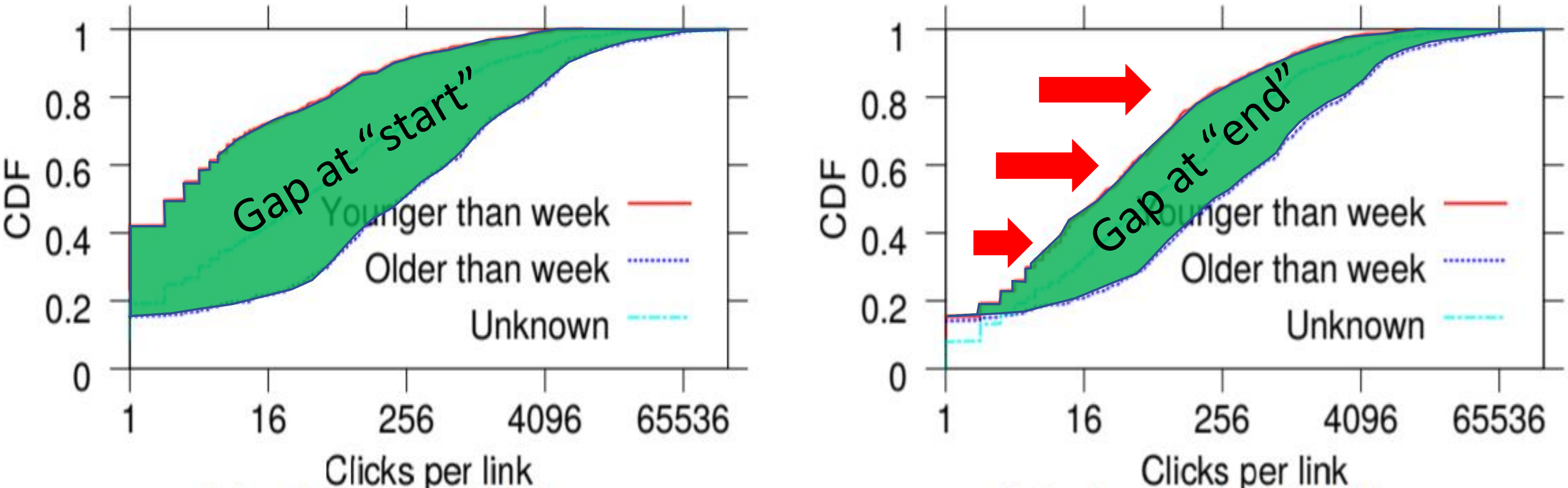
"Younger" catching up



- "Younger" links gain on the lifetime clicks observed for the "older" links, closing the gap

Life-time clicks

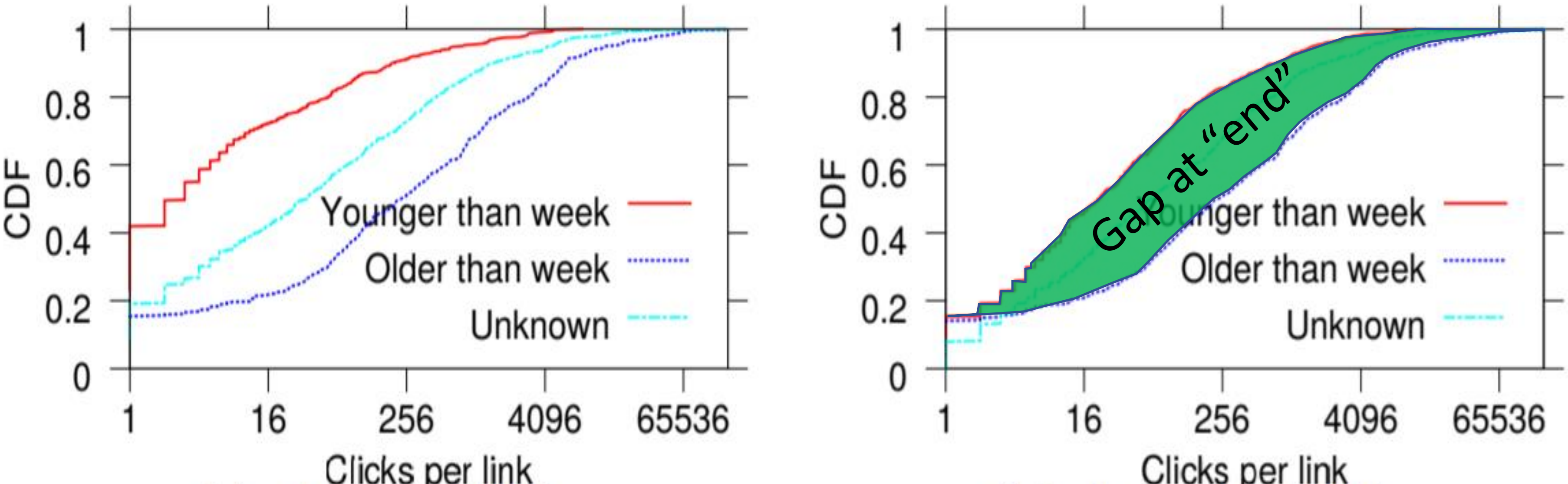
"Younger"
catching up



(a) At start (0h) (b) At end (120h)

- "Younger" links gain on the lifetime clicks observed for the "older" links, closing the gap

Life-time clicks

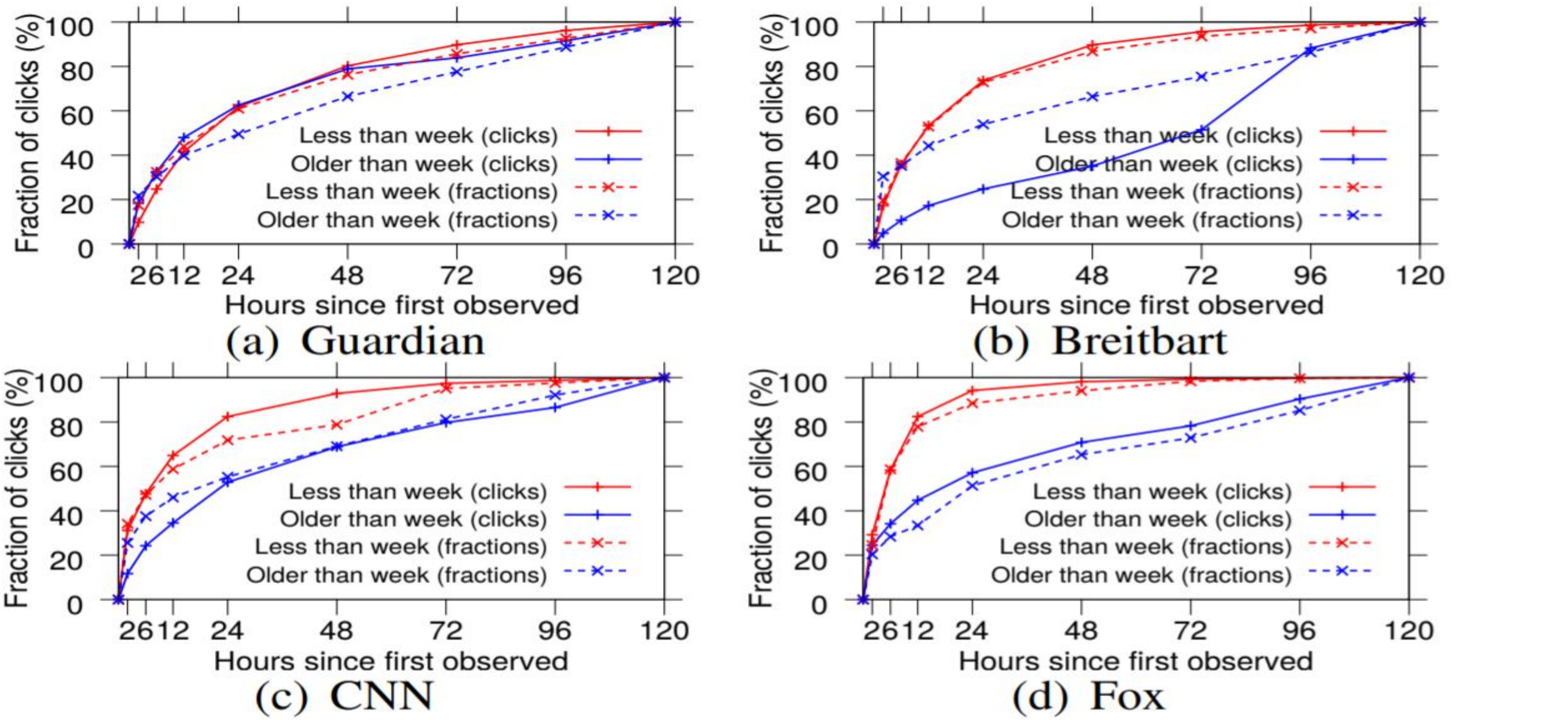


(a) At start (0h) (b) At end (120h)

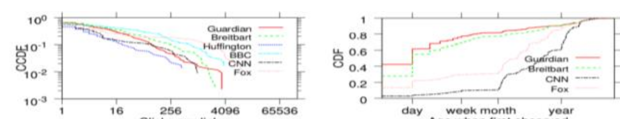
- "Younger" links gain on the lifetime clicks observed for the "older" links, closing the gap
- However, the gap is still substantial at the end of the 120 hour period
  - This highlighting that the "older" category includes many links to long-term popular articles

Per-site-based analysis

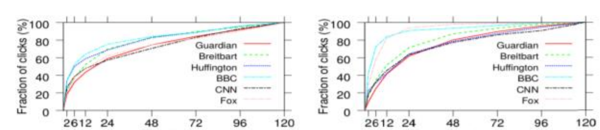
Invariants despite significant differences



- Age-based invariants (across websites) despite significant differences observed between the different websites (e.g., age, speed clicks are obtains, and click distribution)
- Our age-based conclusions are consistent for each of the news websites individually, further validating our previous claims

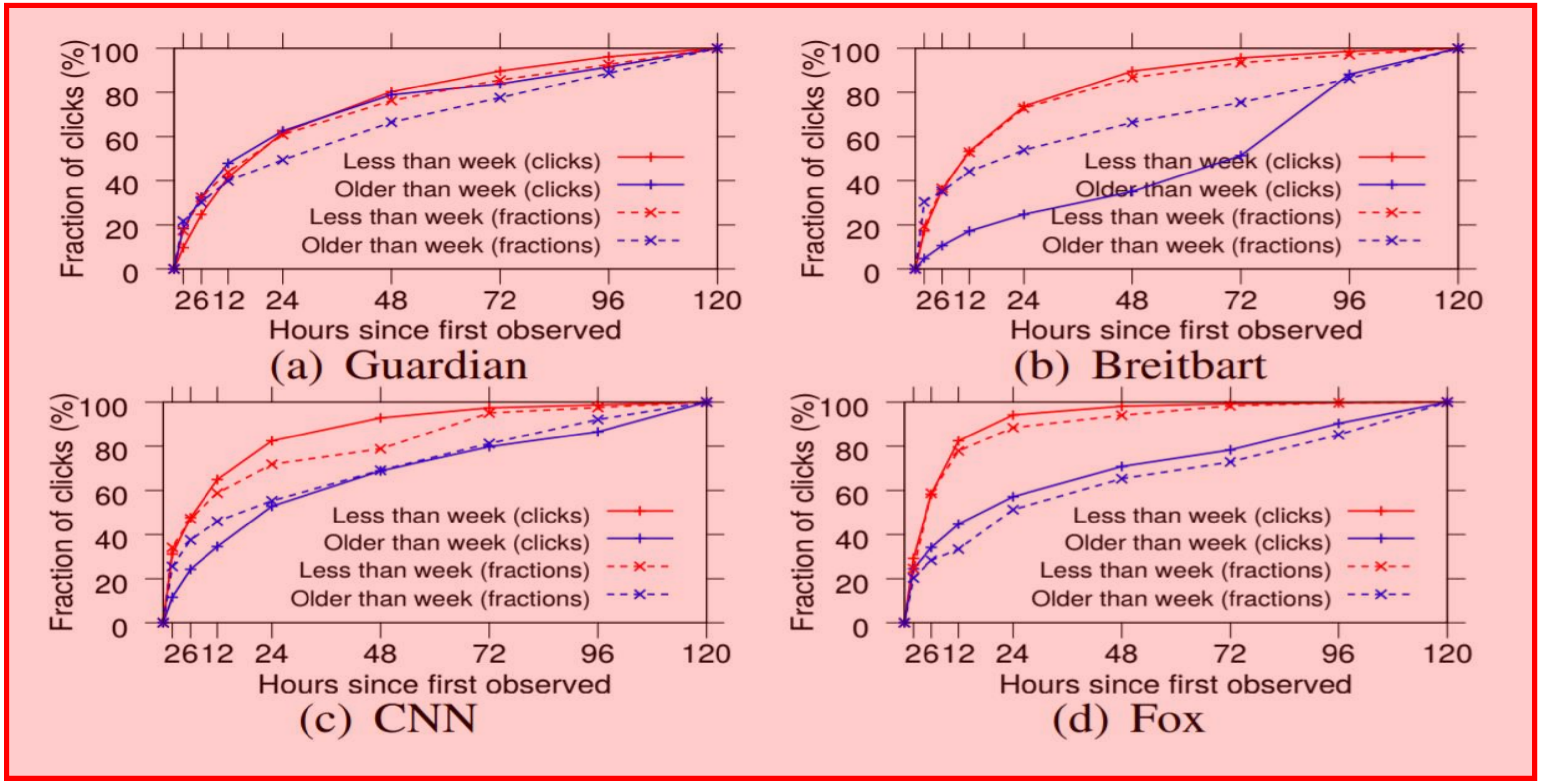


(a) CCDF clicks after discovery (b) CDF age at discovery
Fig. 13. Based on website (new only).

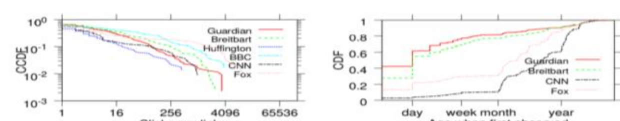


(a) Average (over fractions) (b) Average (over all clicks)
Fig. 14. Cumulative click over time for different news websites.

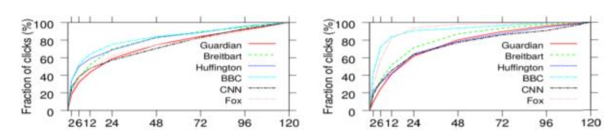
Invariants despite significant differences



- Age-based invariants (across websites) despite significant differences observed between the different websites (e.g., age, speed clicks are obtains, and click distribution)
- Our age-based conclusions are consistent for each of the news websites individually, further validating our previous claims

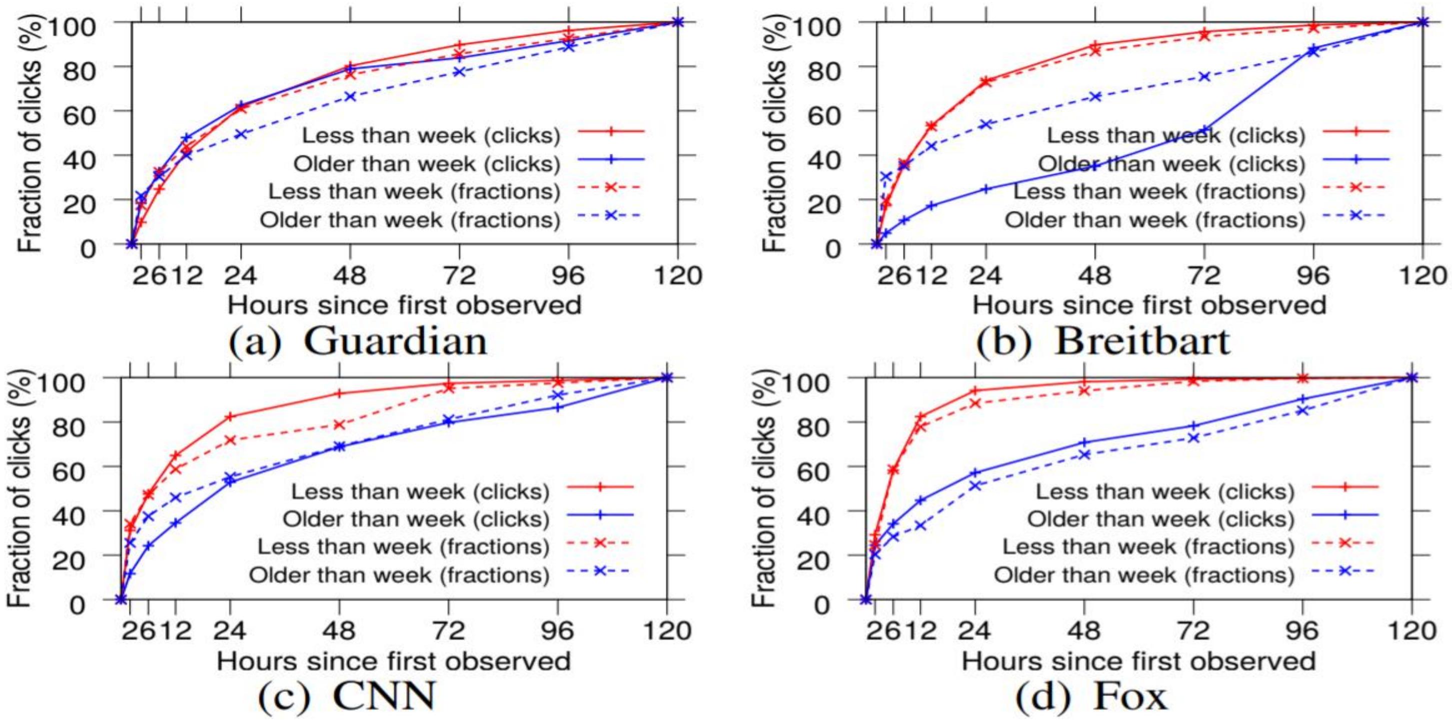


(a) CCDF clicks after discovery (b) CDF age at discovery
Fig. 13. Based on website (new only).



(a) Average (over fractions) (b) Average (over all clicks)
Fig. 14. Cumulative click over time for different news websites.

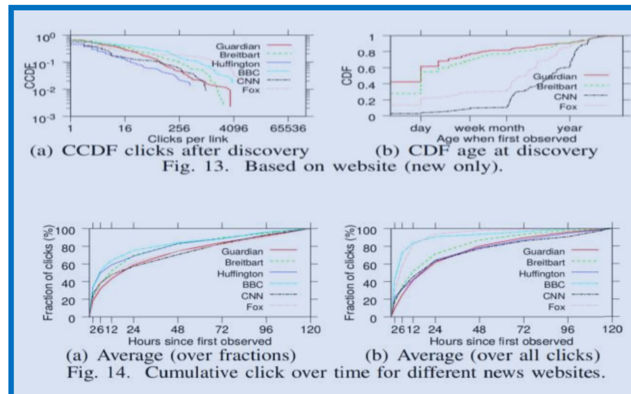
Invariants despite significant differences



(a) Guardian (b) Breitbart

(c) CNN (d) Fox

- Age-based invariants (across websites) despite significant differences observed between the different websites (e.g., age, speed clicks are obtains, and click distribution)
- Our age-based conclusions are consistent for each of the news websites individually, further validating our previous claims

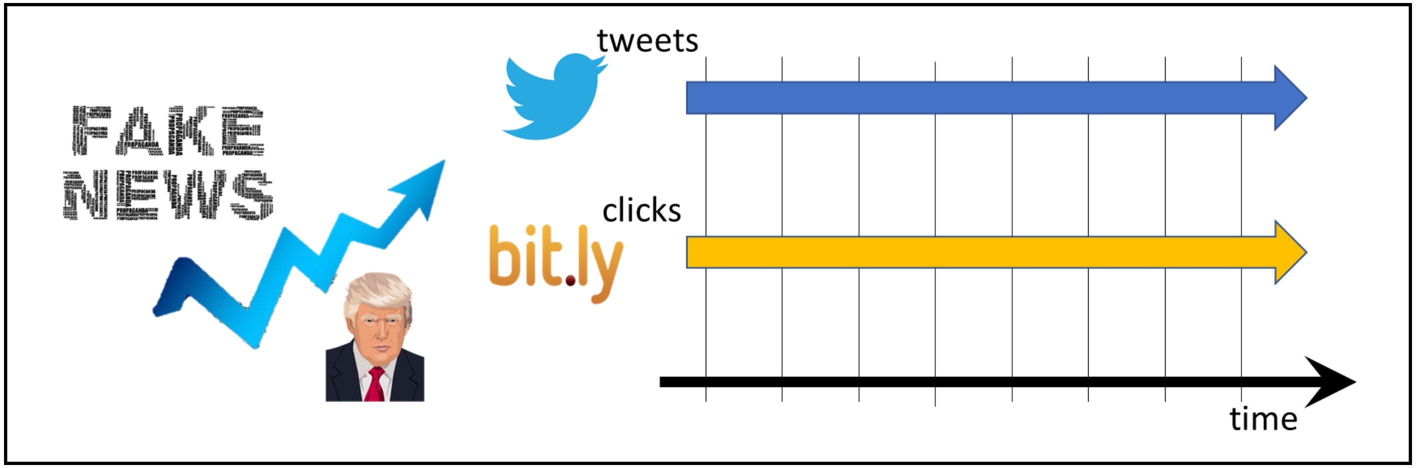


Conclusions

Conclusions and summary

- Two main contributions
  - A novel longitudinal measurement framework
  - The first analysis of how the number of clicks changes over time
- Example observations from temporal analysis
  - Noticeable differences in the relative number of clicks vs. retweets occurring at different parts of the news cycle
  - Retweet data often underestimates biases towards clicking popular links/articles
  - Significant differences in the clicks-per-tweets ratio, including (alarmingly) many links with more retweets than clicks
  - Significant age biases, including relatively high initial click rates for articles younger than a week and much more stable click rates for older and long-term popular articles
  - Insights into how age-dependent popularity skews and age-dependent churn impact the clicks observed by different classes of links
- We validate our findings (and identify invariants) using both data from May 2017 and a per-website-based analysis

Thanks for listening!



Do we Read what we Share? Analyzing the Click Dynamic of News Articles Shared on Twitter

Jesper Holmstrom, Daniel Jonsson, Filip Polbratt, Olav Nilsson, Linnea Lundstrom, Sebastian Ragnarsson, Anton Forsberg, Karl Andersson, and Niklas Carlsson

LINKÖPING UNIVERSITY

Niklas Carlsson (niklas.carlsson@liu.se)