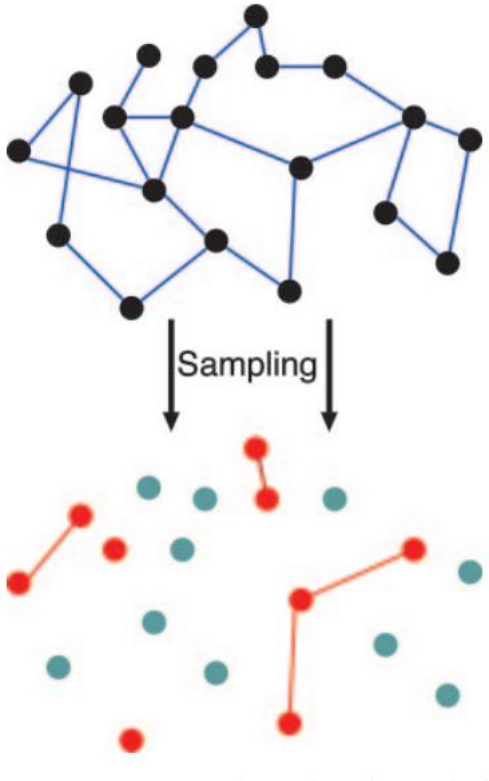
NETWORK SCIENCE OF ONLINE INTERACTIONS



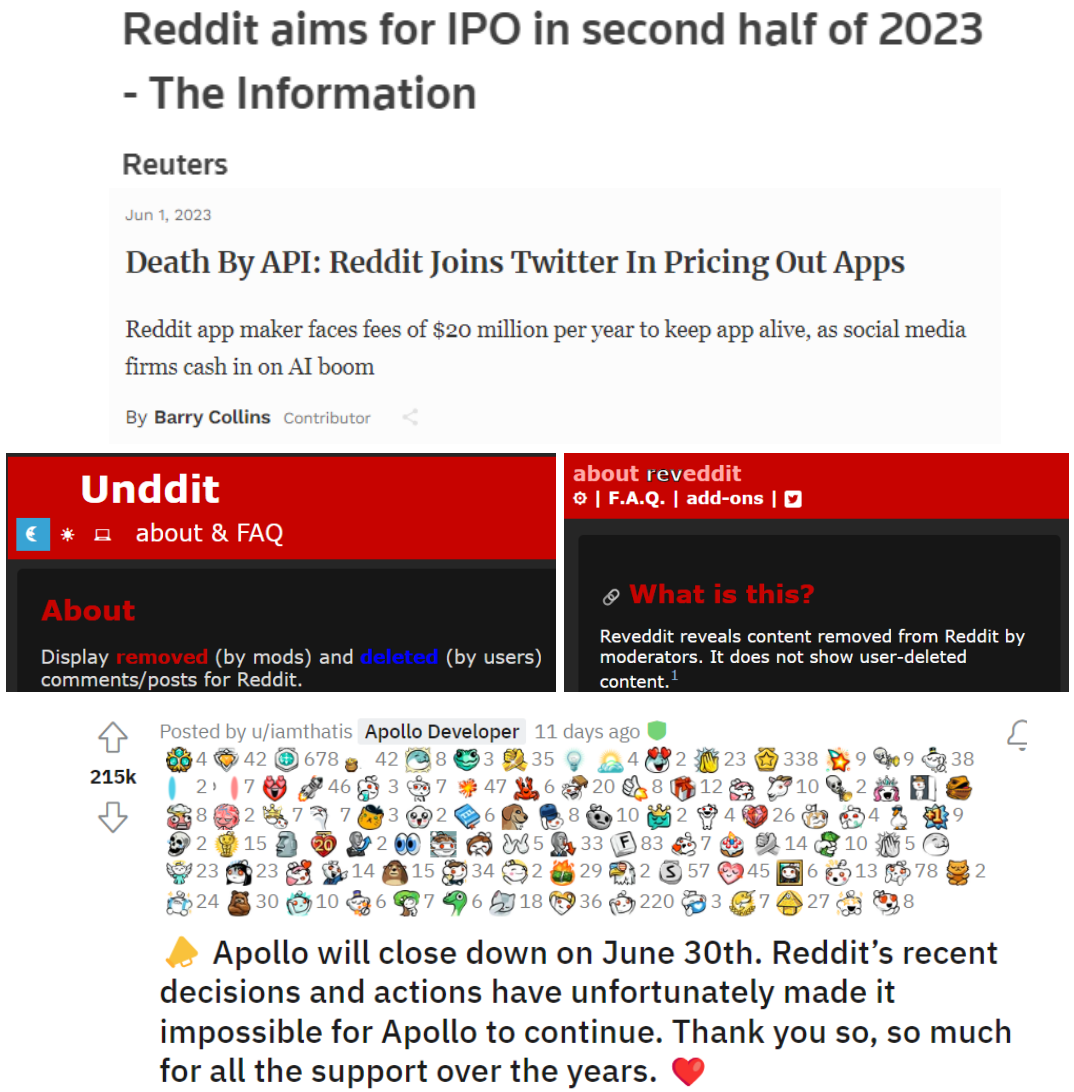
Reddit update + Sampling bias

Joao Neto

21/Jun/2023

WHAT IS GOING ON?

- Reddit will have an IPO later this year, needs to raise money
- Reddit's data is *very* valuable for LLMs
  - GPT-2 was trained on 2017 Reddit (WebText)
  - GPT-3 was 22% 2020 Reddit (WebText2)
- The problem: this data is freely available
  - Pushshift API and data dumps
  - Reddit API has generous limits
- The solutions
  - Kill Pushshift
  - Make large-scale API (very) expensive
- The unintended (intended?) consequences
  - Destroys the ecosystem of Pushshift-based modtools
  - Destroys **all** 3rd-party Reddit clients (RIF, Apollo, etc)
- The blowback: biggest social media protest ever (in content)



WHAT HAPPENED?

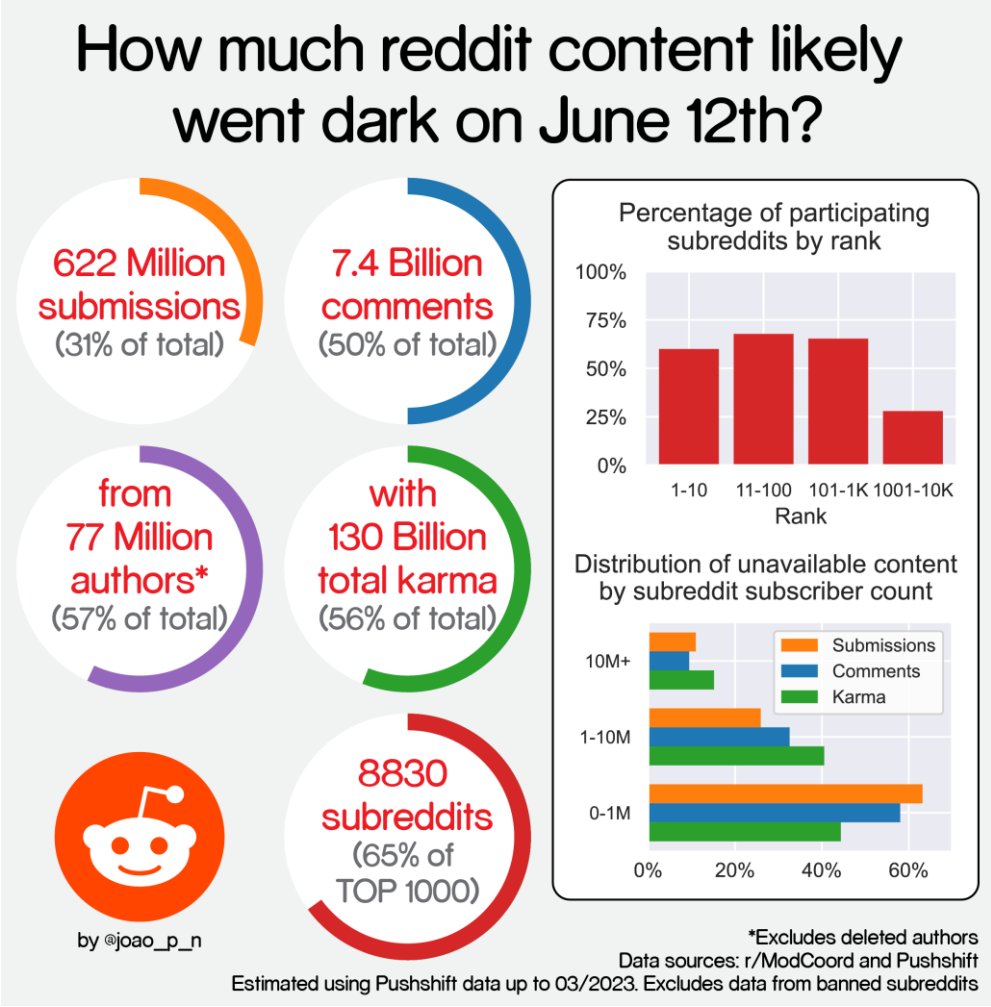
- 8830 subreddits went dark for 48h on June 12th
  - Some decided by the mods
  - Some consulted with the community (pool)
- As of now, 3119 are still down
- 600+ promised to stay down indefinitely
- 65% of TOP1000 subreddits down
- Broke "+reddit" Google search

BREAKING | Jun 16, 2023

Reddit's Blackout Has Broken A Popular Google Search Hack. Can Bard And ChatGPT Offer An Alternative?

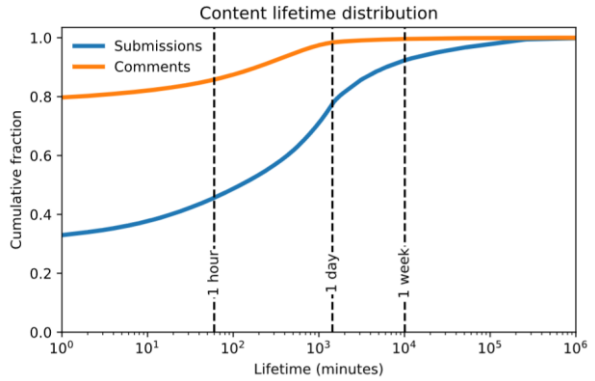
A popular Google hack that involves adding the term "reddit" or "+reddit" to a search query for better results isn't working due to the ongoing subreddit blackout.

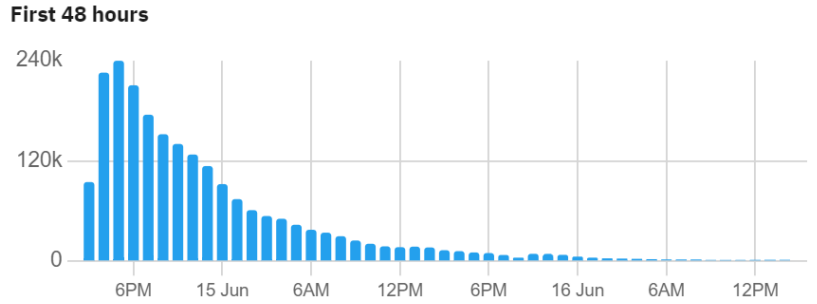
By **Siladitya Ray** Forbes Staff



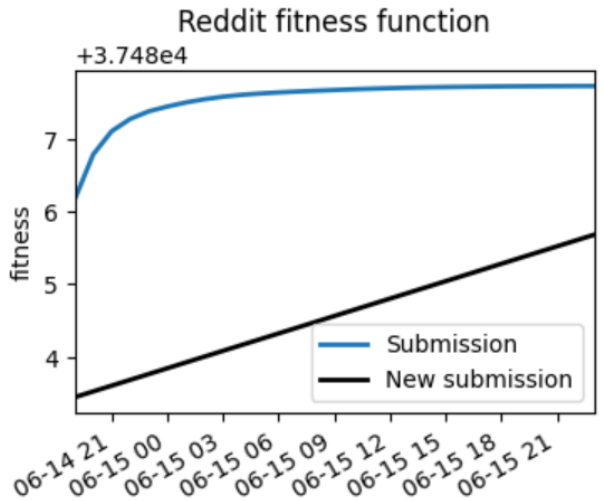
SPREADING EXPERIMENT

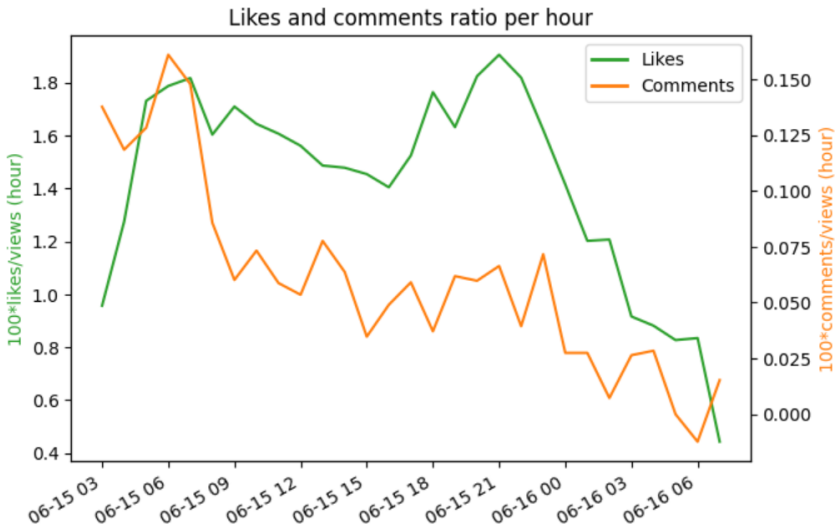
- Views and score over time
  - Non-gaussian, lasted about a day
  - High ratio of comments at peak views, half later
  - Consistent ratio of likes
  - Both crater after a day
- Effect of Hot sorting:
  - Fitness $f = \log_{10} score + t_{submission}/45000$
  - Soon, even a new submission will be ranked higher





2.2m
Total views





TIMELINE IN FORBES NEWS

Jun 1, 2023

Death By API: Reddit Joins Twitter In Pricing Out Apps

Reddit app maker faces fees of $20 million per year to keep app alive, as social media firms cash in on AI boom

By **Barry Collins** Contributor

BREAKING | Jun 13, 2023

Reddit Stands By Controversial API Cha… Continues

Changes to Reddit's API—which includes new fees for thir… met with pushback from thousands of subreddits, many o… will last until Wednesday.

By **Antonio Pequeño IV** Forbes Staff

BREAKING | Jun 15, 2023

Reddit Blackout Rolls On For More Than 5,000 Subreddits Past Planned End Date—Some Of Which Plan To Stay Dark Indefinitely

A major Reddit blackout in protest of API pricing changes was supposed to end Wednesday for most subreddits.

By **Antonio Pequeño IV** Forbes Staff

BREAKING | Jun 15, 2023

Reddit CEO Pushes Back Against Blackout—Will Consider Letting Users Vote Out Moderators

Reddit CEO Steve Huffman is looking to put an end to the site's blackout.

By **Antonio Pequeño IV** Forbes Staff

…Moderators If They Don't Open Their Pages,

…eager to end the protests and reopen the shuttered subreddits

…aff

Jun 18, 2023

If Your Reddit Feed Is Full Of John Oliver, Here's Why

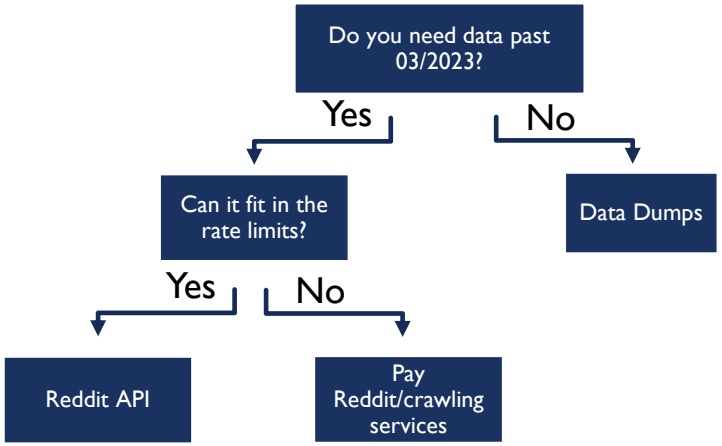
Two major Reddit communities call off the strike – but now only allow pictures of comedian John Oliver

By **Barry Collins** Contributor

WHAT ABOUT APIS?

- Pushshift is dead out of "approved mods"
- Reddit API has **higher** limits for authenticated users
  - 60 → 100 calls per minute
- **Much lower** for non-authenticated users
  - 30 → 10 calls per minute
- Anything beyond that is **very** expensive
  - $0.12 for 1000 calls
- Academic access on a "contact us" basis, no idea how it is going

- Structuring a project:

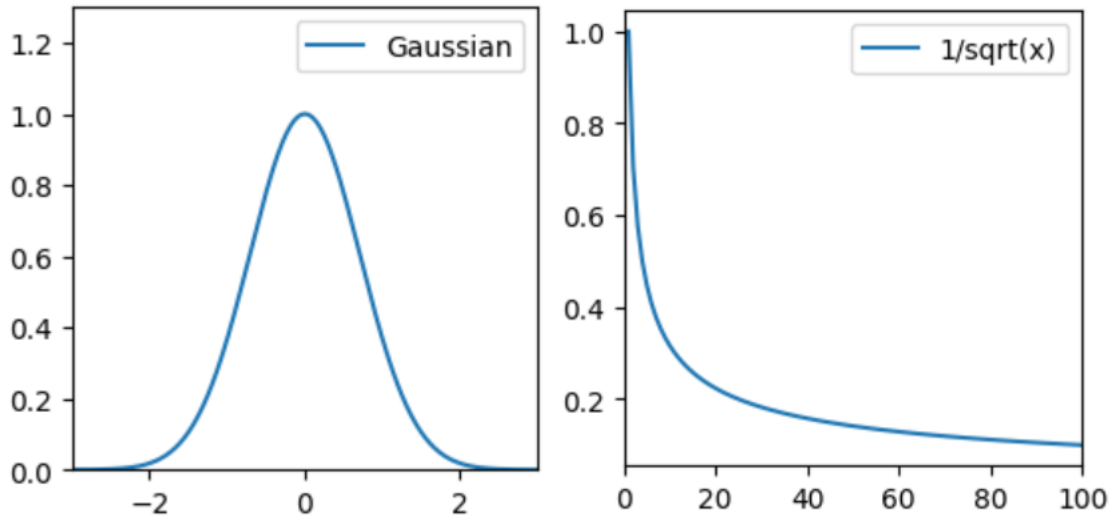


REDDIT SUMMARY

- Still ongoing
- Subreddits opening
- Moderators being removed
- All moderation goodwill burned
- Expect quality in drop over time (new Digg?)
  - Reddit may go full AI moderation in the future
- A bunch of federated alternatives popping up
  - Lemmy (most popular)
  - TrustCafe (by Wiki guy)
- **Lots** of interesting scientific questions
  - Mod coordination
  - User sentiment

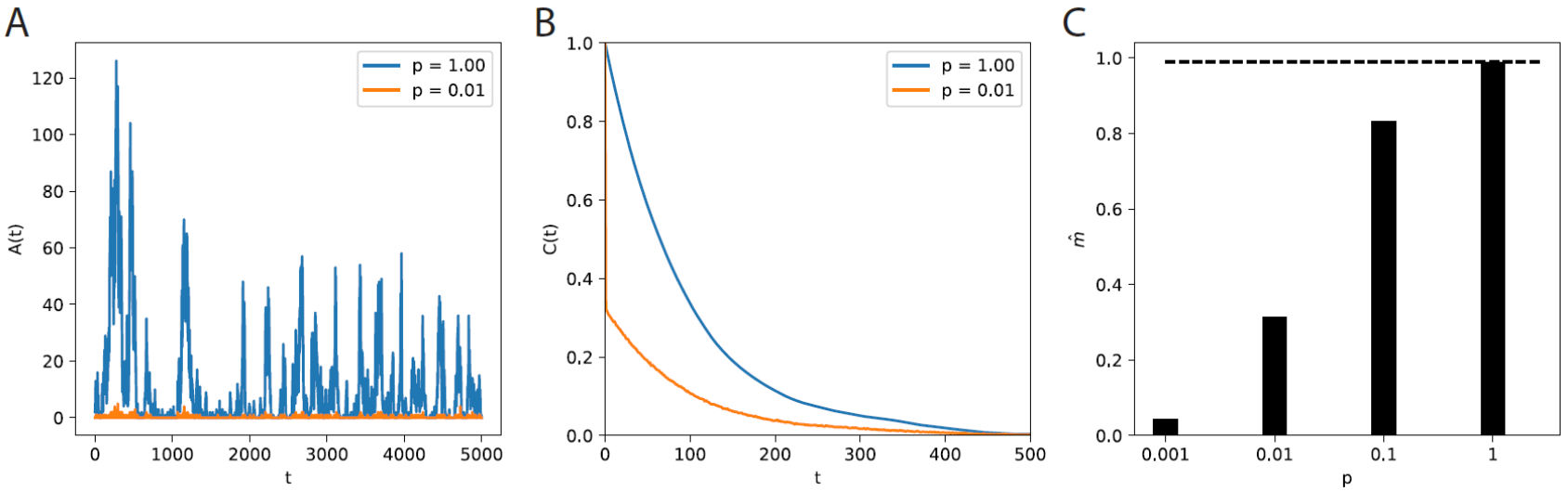
WHAT IS SAMPLING BIAS?

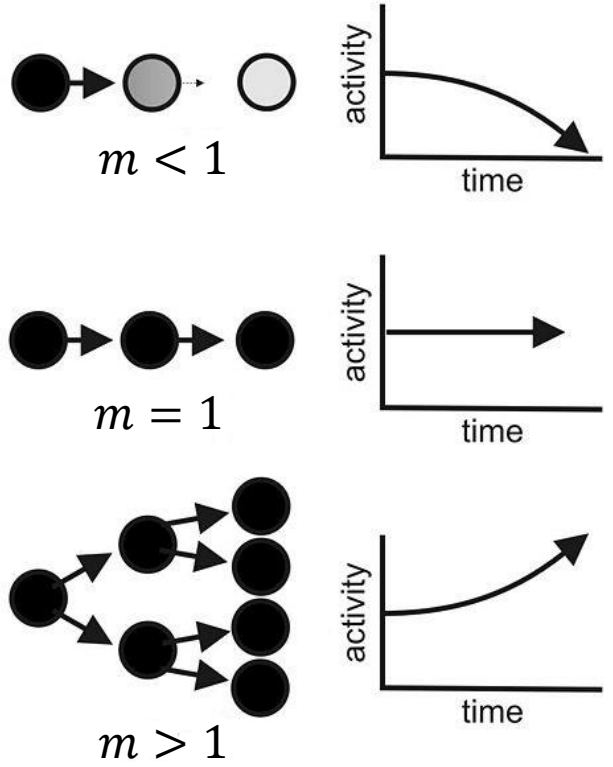
- Rarely you get the full data of something
- Usual assumption:
  - Samples are independent, identically-distributed random variables (i.i.d)
- Then: Central Limit Theorem
  - Value distribution is Gaussian
  - Sampled mean converges monotonically to true mean
  - Berry-Essen theorem: convergence rate is $\sim 1/\sqrt{n}$
  - 10-100 samples should be enough to estimate it
- What if samples are not i.i.d?
  - Dynamical system: samples are correlated in time
  - Distribution is not stationary (e.g. growing system)



EXAMPLE: BRANCHING PROCESS

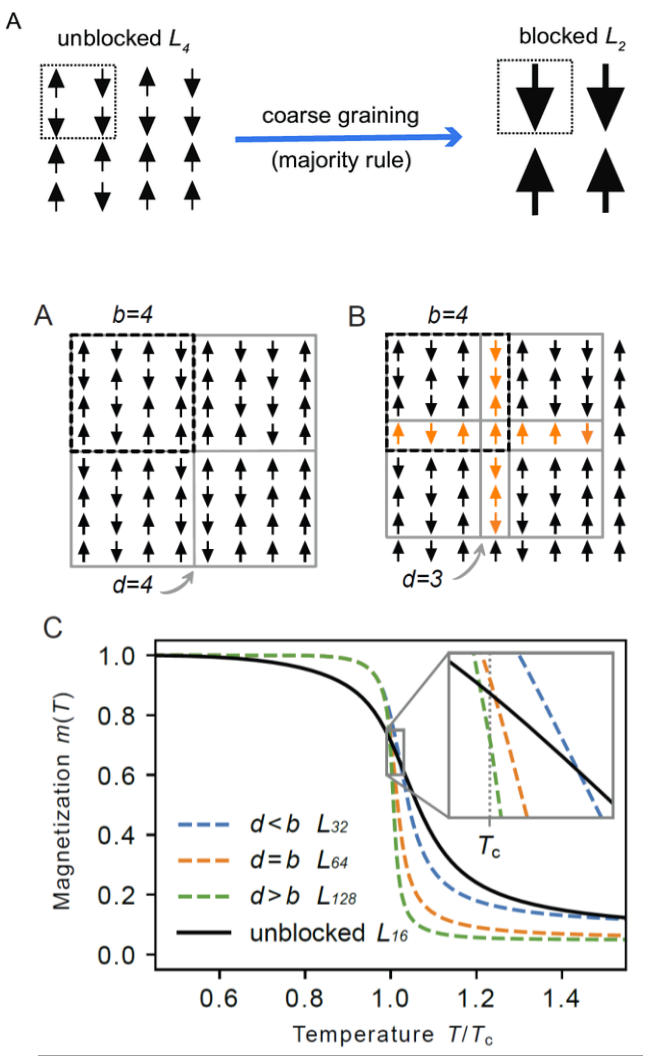
- Branching process with drive h
  - $A(t+1) = mA(t) + h$
  - $m = 0.99, h = 0.01$
  - Highly correlated dynamics with some spontaneous activity
  - Sample active nodes with probability p
  - Sampling heavily bias measures of activity, correlation and branching parameter





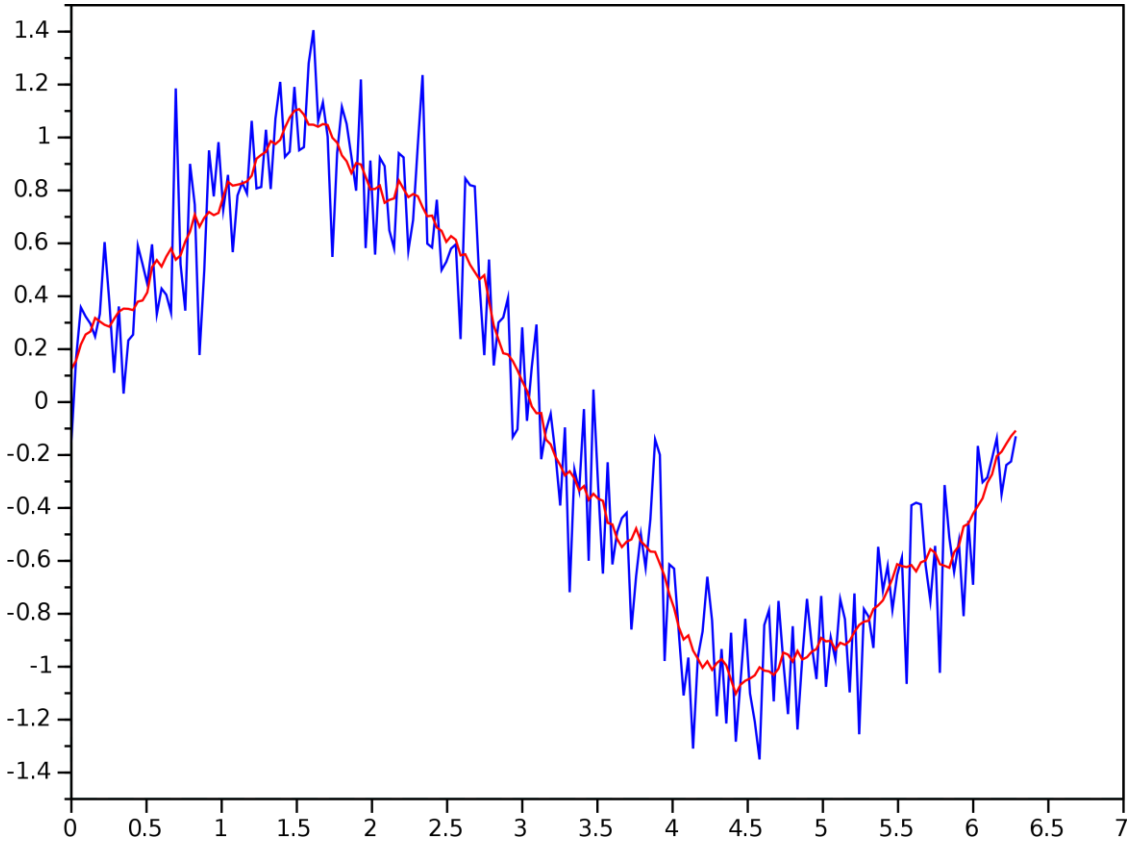
EXAMPLE: ISING MODEL

- Model of binary state with phase transition
- Renormalization: model retain properties on rescaling
  - Akin to cluster sampling in statistics
- What if you don't cluster properly?
  - Overlapping units → overestimate correlations
  - Missing units → underestimate correlations
  - Both bias measurements (e.g. critical point)



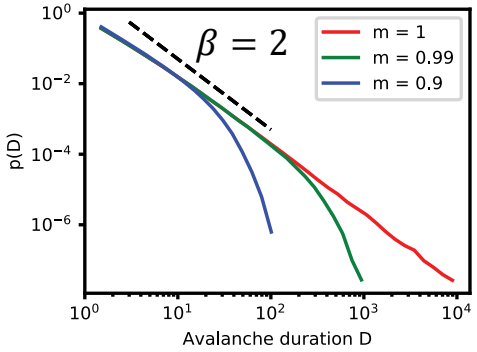
TEMPORAL BIAS

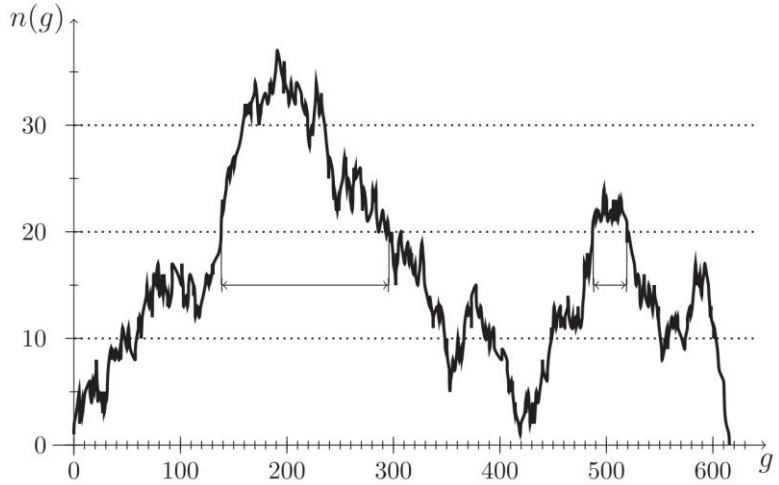
- Typical case: analysing event timeseries
  - Events are very noisy
  - Apply some smoothing/temporal averaging
- The issue:
  - Averaging introduces temporal bias
  - Hides extreme events
  - Overestimate temporal correlations
- Correlation metrics should not be naively calculated on averaged/binned data

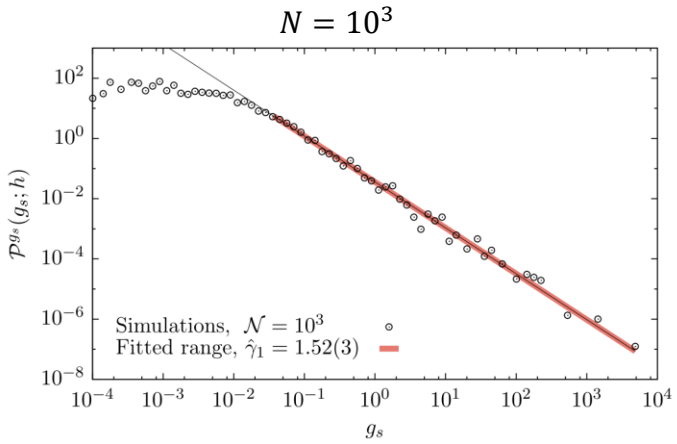


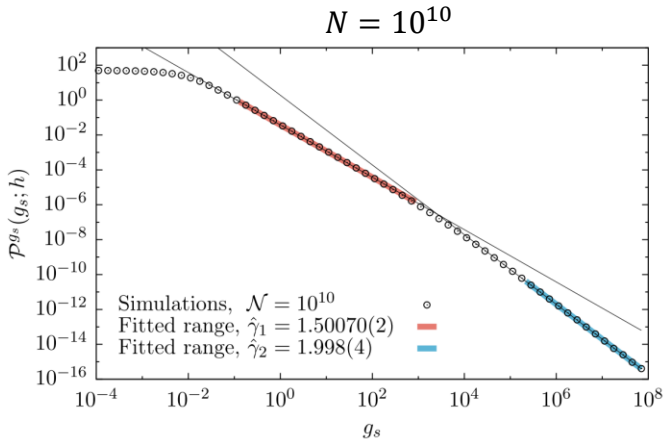
THRESHOLDING BIAS

- Signal thresholding: only dynamics above a threshold h matter
- Model: branching process
  - Duration distribution $p(D) \sim D^{-2}$
- Results
  - Thresholding changes the scaling exponent
  - $h = 100$: exponent $2 \rightarrow 1.5$
  - Would mischaracterize the dynamics
  - True exponent is recovered for much larger statistics





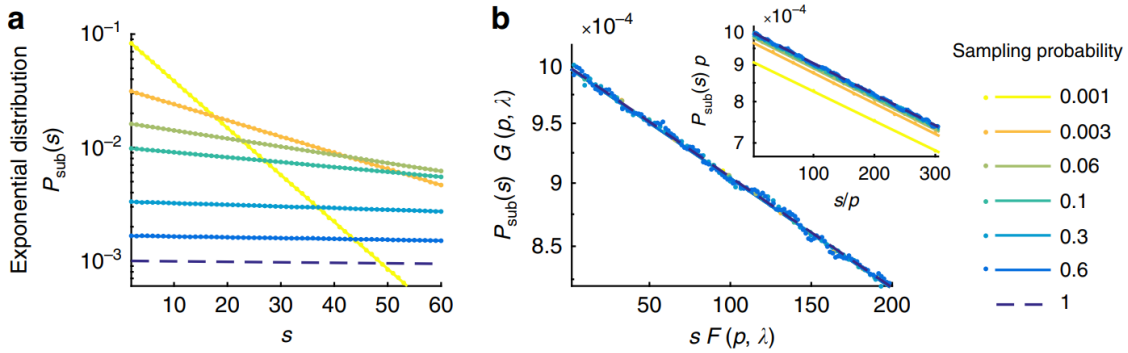




Font-Clos, et al. (2015). New Journal of Physics, 17(4), 043066.

SUBSAMPLING DISTRIBUTIONS

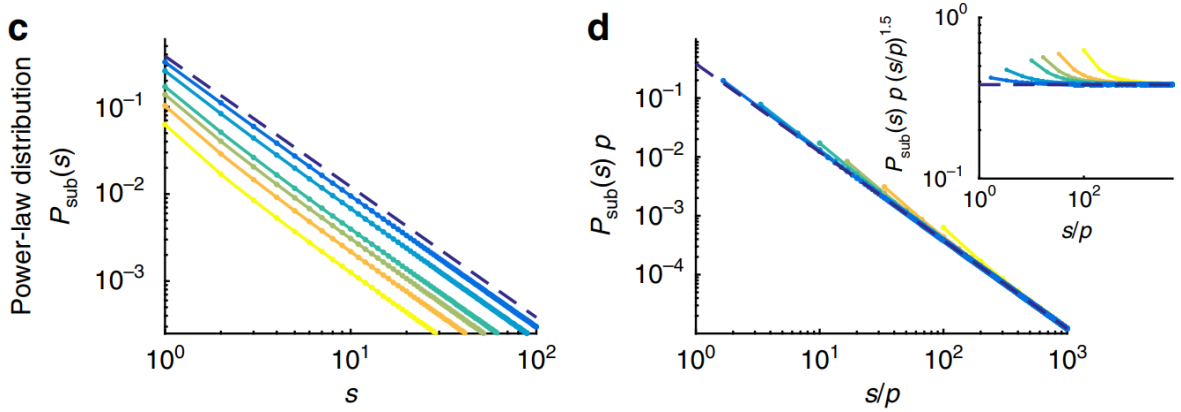
- You know the fraction p you are sampling
- Subsampling scaling
  - Obtaining $P(s)$ from $P_{sub}(s)$
- For exponential $P(s) = C_\lambda e^{-\lambda s}$
  - $\lambda = \ln\left(\left(e^{\lambda_{sub}} - 1\right)p + 1\right)$
  - $P(s) = \left(1 - e^{-\lambda} + pe^{-\lambda}\right)P_{sub}\left(\frac{\lambda}{\ln\left(\frac{e^{\lambda}+p-1}{p}\right)}s\right)$
- Subsampling of an exponential is an exponential

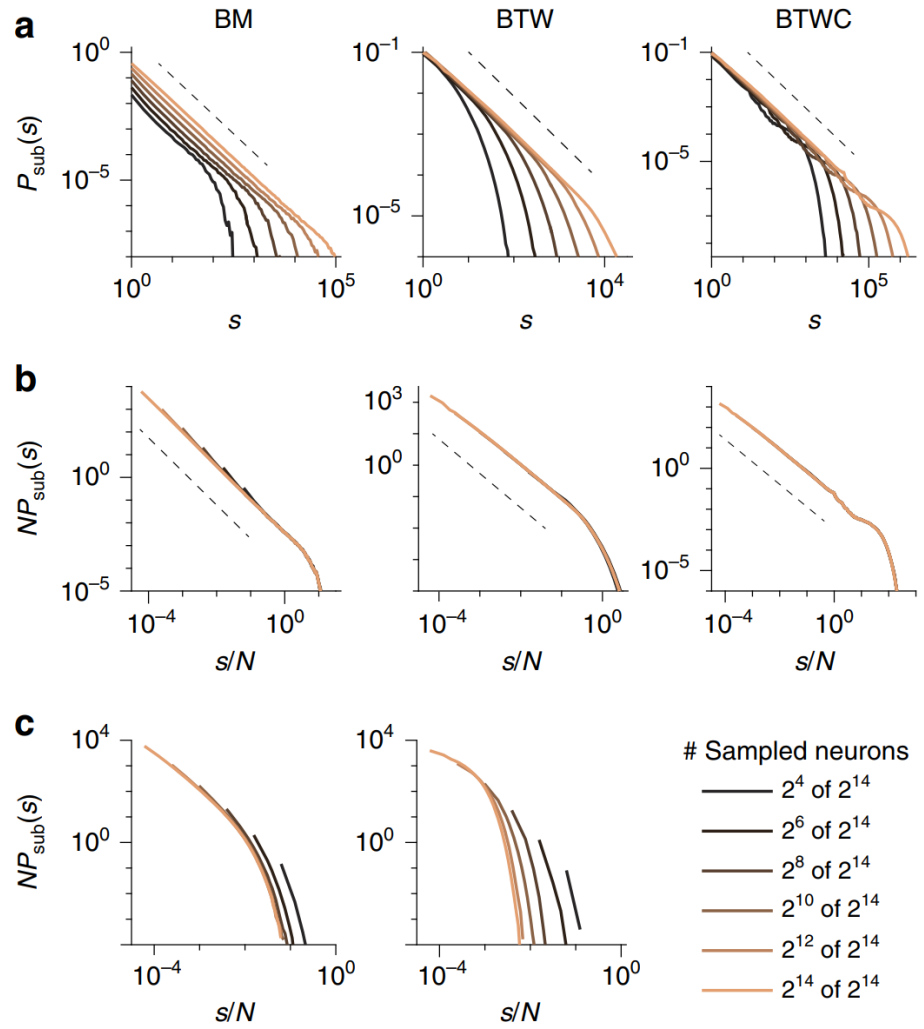


Levina, A., Priesemann, V. Subsampling scaling. Nat Commun 8, 15140 (2017).

SUBSAMPLING DISTRIBUTIONS

- Power-law distributions $P(s) = C_\gamma s^{-\gamma}$
  - Generally do not collapse
  - Sampling of a power-law is not exactly a power-law
  - Tail can be collapsed if distribution is long enough
  - $P(s) \approx pP_{sub}(ps)$
  - Fails if large exponential cut-off





SUMMARY

- Sampling bias can come in many forms
  - Sampling
  - Averaging
  - Thresholding
- Dealing with it is
  - Gaussian: usual
  - Exponential: possible
  - Power-law: tricky
    - But possible