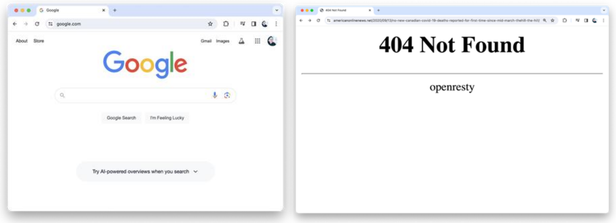
Some URLs Are Immortal, Most Are Ephemeral

Kritika Garg[1], Sawood Alam[2], Michele Weigle[1], Michael L. Nelson[1], Mark Graham[2], Dietrich Ayala[3]

[1]Old Dominion University, [2]Internet Archive, [3]Filecoin Foundation

MOTIVATION

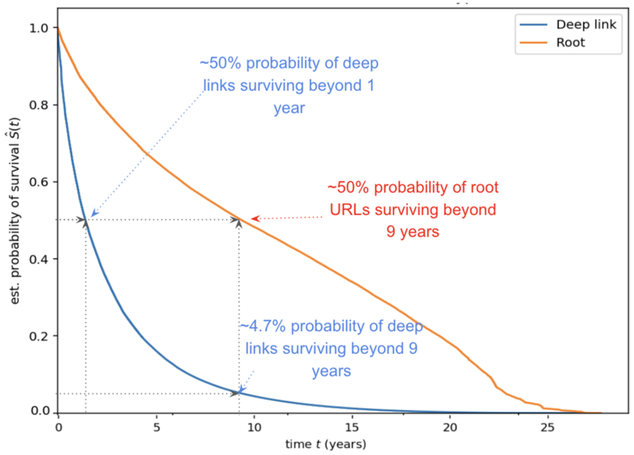
- Despite long-lasting domains like google.com, most web pages are ephemeral, as shown by frequent HTTP 404 errors.



- **"How long does a web page last?"** is often cited as "44 to 100 days", but the web has evolved since those numbers were first given in 1996.

SURVIVAL ANALYSIS RESULTS

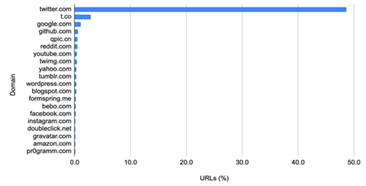
- We employed the Kaplan-Meier survival analysis to estimate the longevity of web pages.
  - The half-life of all URLs is 2 years
  - Root URLs had a longer half-life of 9 years, compared to one year for deep links.
  - URL half-life varied by decade: 1990s URLs had a 15-20 year half-life, early 2000s URLs lasted 6-7 years, and 2003-2021 URLs ranged from 6 months to 3 years.



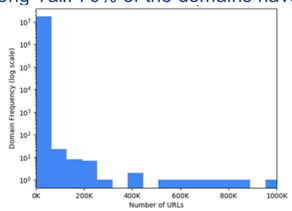
METHODOLOGY

- We comprehensively sampled URLs from the Internet Archive (IA) to curate a "sample of the web" consisting of TimeMaps for 27.3M URLs from 1996–2021, encompassing 7M unique hosts.
- Sampling the archived web can be challenging

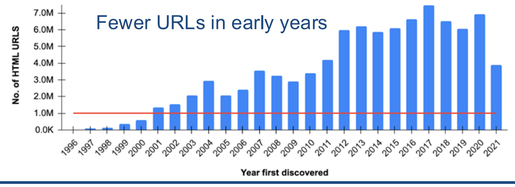
Popular domains are archived more



Long Tail: 70% of the domains have one

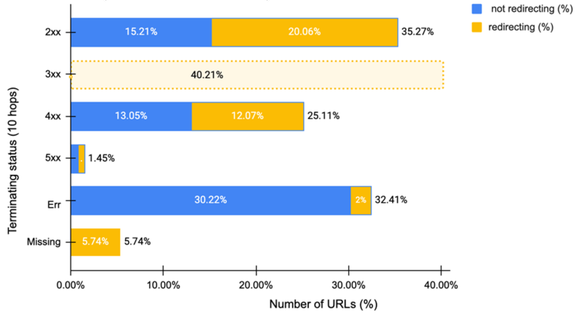


Fewer URLs in early years



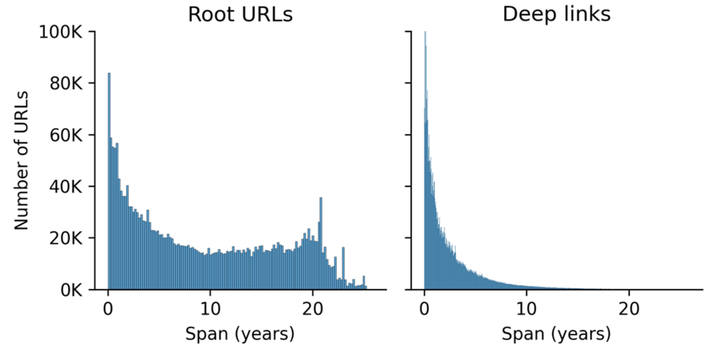
URL STATUS ANALYSIS

- In September 2023, we analyzed the URLs using IA's Heritrix crawler to assess their status as active or inactive, and also tracked redirecting URLs.
- 35% of URLs remain active, with nearly half of those first archived between 1996 and 2000 still accessible in 2023 (e.g., nasa.gov)
- 60% of URLs are are no longer accessible on the live web.
  - 27% of URLs had HTTP response errors
  - 9% of URLs (across 9.5% hosts) had HTTP connection errors
  - 23% of URLs (across 30% hosts) had DNS failures



WEB PAGE LIFESPAN RESULTS

- We analyzed the lifespan of 7.4M inactive URLs with multiple IA mementos, defining "birth" as the first successful archive date and "death" as the date after the last successful archive.
- We excluded the 30% of URLs that were never captured alive or had only 1 memento, suggesting an ephemeral existence.
- The average lifespan of a web page is 5.1 years, skewed by outliers. The median lifespan of a URL to be 2.3 years.
- The lifespan of root URLs follows a bimodal pattern:
  - 10% of root URLs die within a year
  - 20% thrive for over 20 years
- The lifespan of deep links exhibits an almost exponential decline.
  - 50% of deep links have a lifespan ≤ 6.6 months.
  - Only 4% thrive for over 20 years



	Count	Mean (years)	Median (years)	Max (years)
All URLs	7.5M	5.1	2.3	25.8
Root URLs	2.7M	9.9	8.8	25.8
Deep links	4.8M	2.5	1.3	25.1

REFERENCES

[1] https://github.com/oduwsdl/nypw/

ACKNOWLEDGEMENT

This work is supported by the Filecoin Foundation.

Contact Email: kgarg001@odu.edu