# A Large-Scale Characterization of How Readers Browse Wikipedia

Tiziano Piccardi, Martin Gerlach, Akhil Arora, Robert West

EPFL

WIKIMEDIA FOUNDATION

## Summary - TL;TR

1. The **majority** of the incoming **traffic** originates from **search engines**.
2. The **majority of navigation** on Wikipedia is **on the surface**, with short sessions.
3. Access from **desktop** and **mobile** devices have a **different temporal pattern**.
4. Different **topics** are associated with **different navigation patterns**.
5. Navigation is **interrupted** on **low-quality pages**.

Server logs from English Wikipedia

One month 6.52B load events

| User id | Timestamp | Device | Referrer | Article |
|---|---|---|---|---|
| d6nil9fgl | 2021-04-12 11:29:51 | Desktop | bing.com | A |
| d6nil9fgl | 2021-04-12 11:31:26 | Desktop | WP: A | B |
| d6nil9fgl | 2021-04-12 11:31:33 | Desktop | WP: A | C |
| d6nil9fgl | 2021-04-12 11:36:16 | Desktop | WP: C | D |
| d6nil9fgl | 2021-04-12 11:37:50 | Desktop | facebook.com | E |
| ... | ... | | ... | ... |

## Landing on Wikipedia

### *Incoming traffic origin* (1)

**77%** *Search engines*

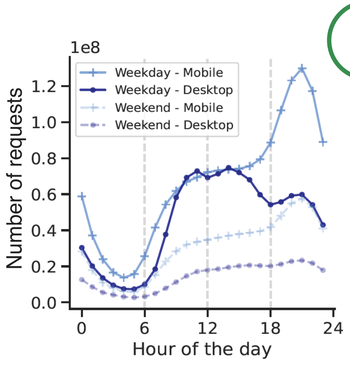
**20%** *Empty referrer*

*Revisit or bookmark*

**1.5%** *External websites*

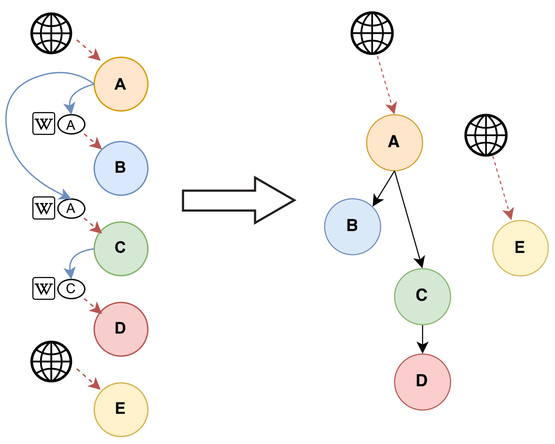
### *Daily pattern* (2)

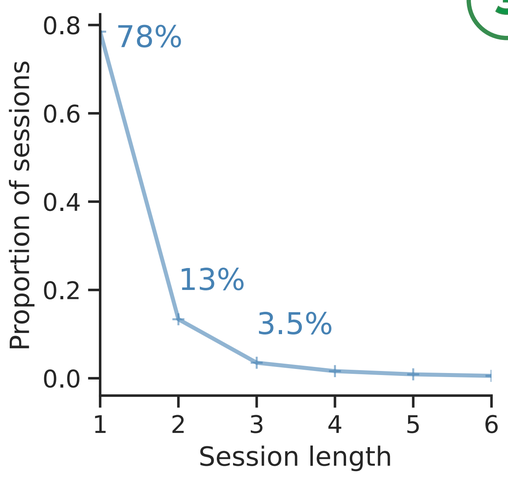
*Differences between Desktop and Mobile*



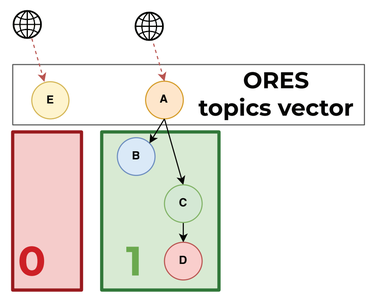

## Beyond the first article

### *Aggregate the sessions* (3)



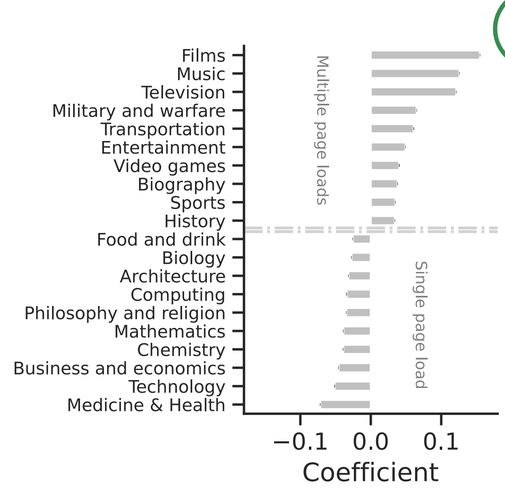




### *Topics* (4)




Logistic regression

*Predict deep navigation*




### *Evolution of the sessions* (5)