Going down the rabbit hole: Understanding information seeking in Wikipedia

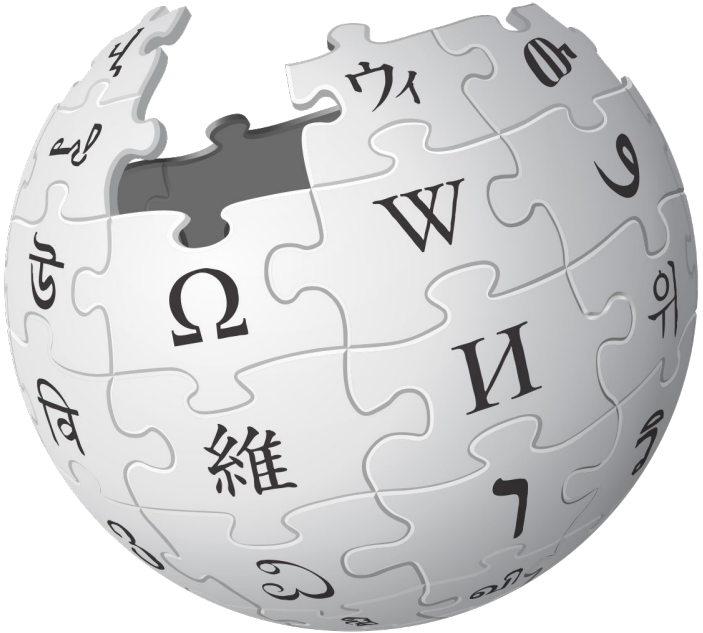
Martin Gerlach, Senior Research Scientist

WIKIMEDIA FOUNDATION

2023-06-07, CSS Seminar, Centre Marc Bloch

0.5M volunteer editors

60M articles



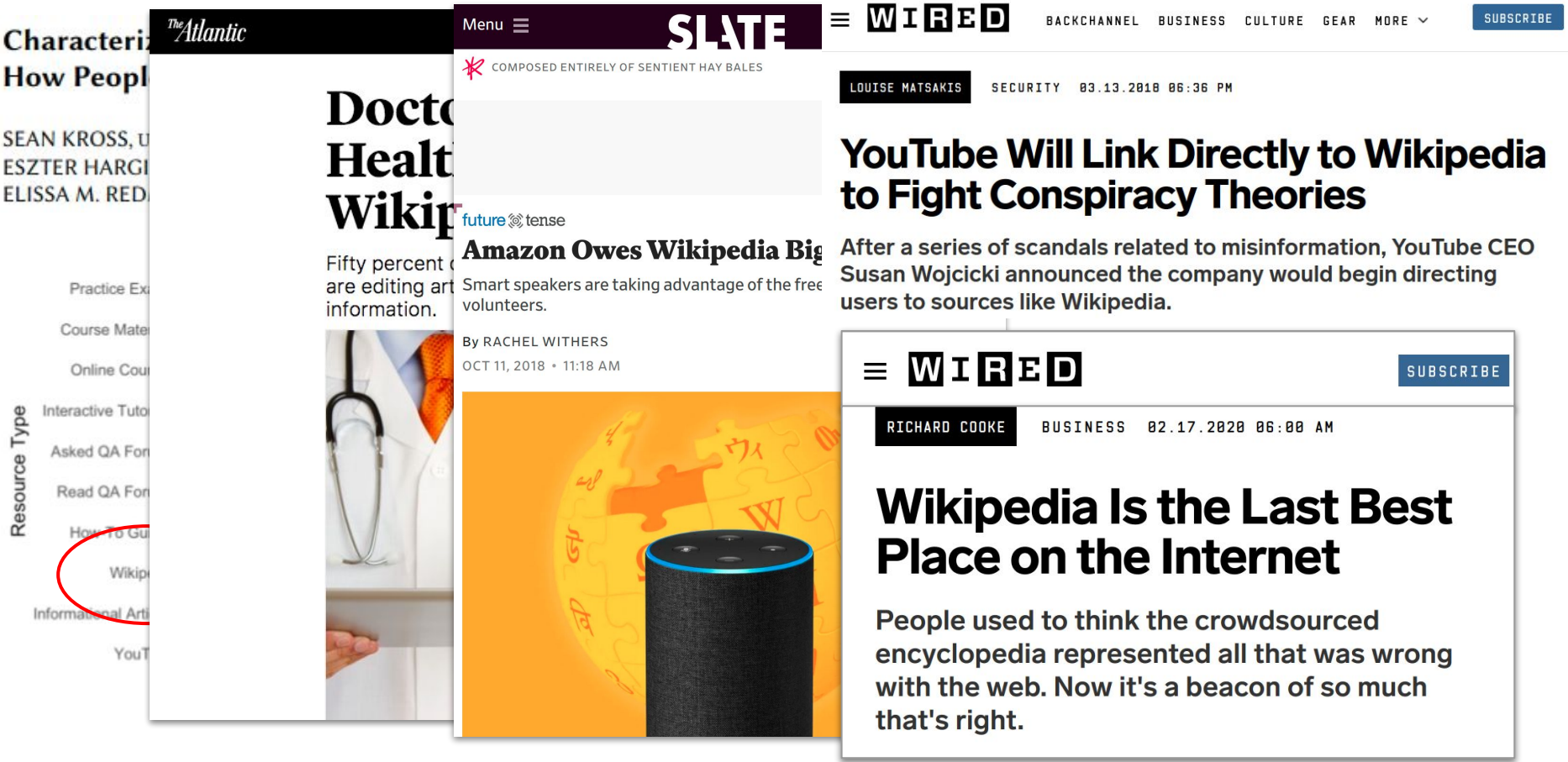
300+ languages

10M monthly edits

15B monthly pageviews

The largest encyclopedia

The importance of Wikipedia



Who operates Wikipedia?

Wikimedia Foundation

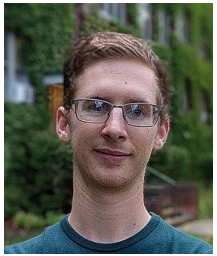
- It is a non-profit organization of ~700 staff
- It provides broad support to Wikimedia communities and projects: servers, data centers, legal and communications support, etc.
- It does not create or modify **content**.
- It does not define or enforce policies on the **projects**

Wikimedia Research Team

Leila Zia
Director, Head of Research

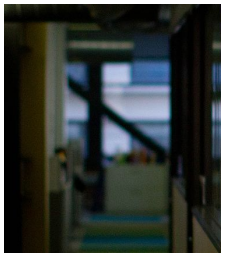
Pablo Aragón
Research Scientist

Martin Gerlach
Senior Research Scientist

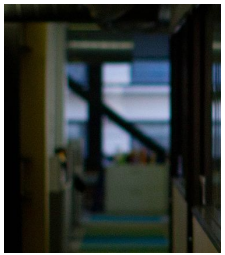


Isaac Johnson
Senior Research Scientist

Yu-Ming Liou
Lead Strategist



Caroline Myrick
Senior Analyst



Fabian Kaelin
Senior Research Engineer

Miriam Redi
Research Manager

Diego Sáez-Trumper
Senior Research Scientist

...and many formal collaborators:
https://w.wiki/_xgod

Research priorities

Addressing knowledge gaps

Improving knowledge integrity

Nurturing the research communities

White Papers: https://meta.wikimedia.org/wiki/Research:2030

Towards more Knowledge Equity

[from: Wikimedia 2030 Movement Strategy https://w.wiki/tg]

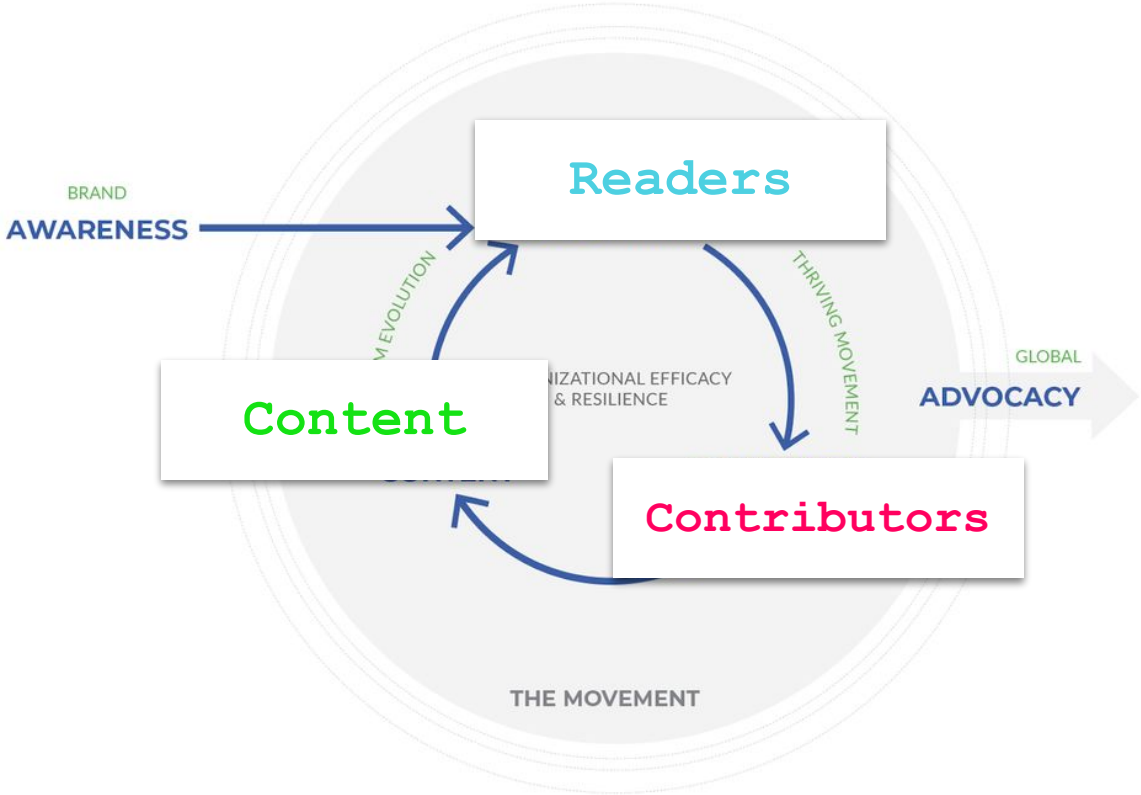
Knowledge equity: As a social movement, we will focus our efforts on the **knowledge and communities that have been left out by structures of power and privilege**. We will welcome people from every background to build strong and diverse communities. We will **break down the social, political, and technical barriers** preventing people from accessing and contributing to free knowledge.

Example Geography: English Wikipedia (950k articles)

Knowledge is socially constructed

Example: Wikipedia's gender gap

- **Content**: Less than 20% of biographies are about women (humaniki)
- **Contributors**: Less than 15% of editors identify as women (Community Insights Report 2021)
- **Readers**: Women comprise ~33% of regular* readers and account for ~28% of pageviews (Johnson et al. 2020); *using Wikipedia at least several times per week



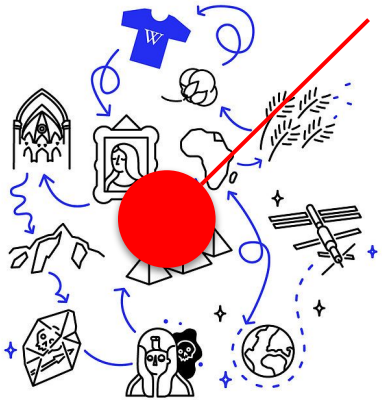
Redi, Miriam, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. "A taxonomy of knowledge gaps for wikimedia projects (second draft)." *arXiv preprint arXiv:2008.12314* (2020).

Research on readers

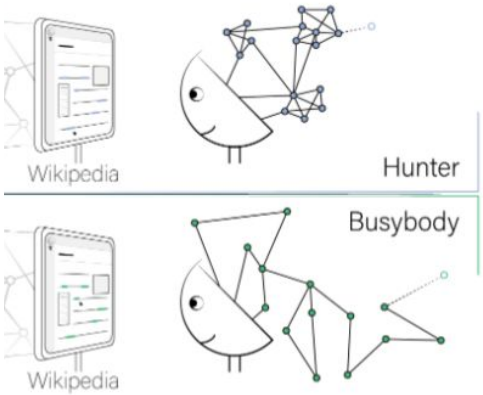
Readership Research

We are HERE

Jean-Honoré Fragonard, Public domain



Glenn Newcomer, CC BY-SA 4.0



1: Who and why?

Surveys on demographics and motivations of readers (2019-21)

2: Navigation

How are readers navigating content (articles, citations, images, etc)?
(2021-23)

3: Learning

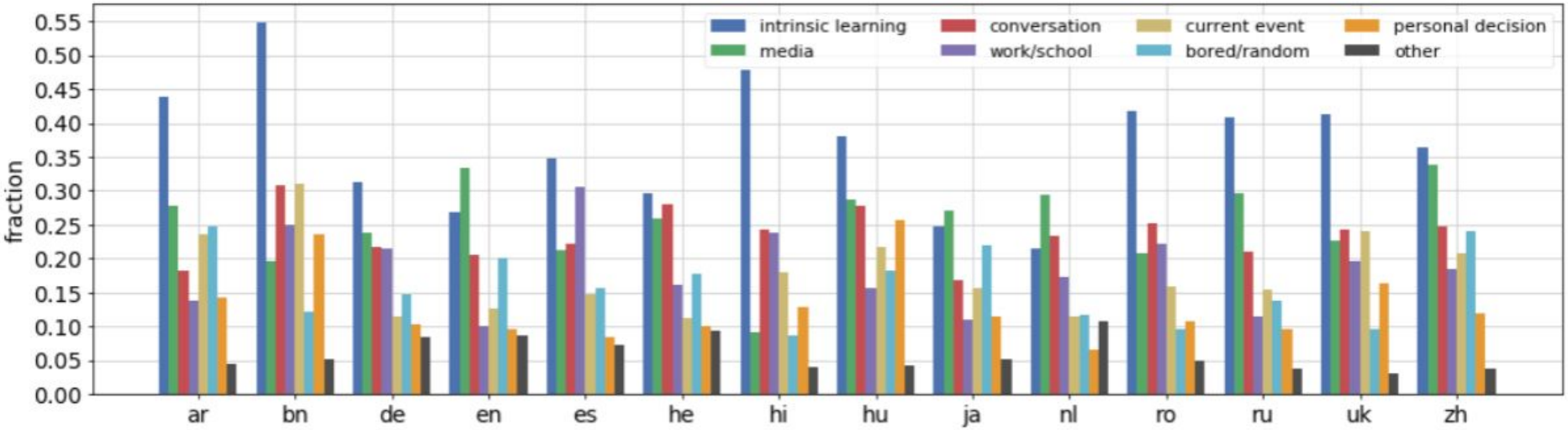
How do readers learn on Wikipedia? What makes readers curious/inquisitive?
(2023+)

Why readers visit Wikipedia? (Surveys)

Singer et al. **Why We Read Wikipedia**. *WWW'17*
Lemmerich et al. **Why the World Reads Wikipedia**. *WSDM'19*

Survey

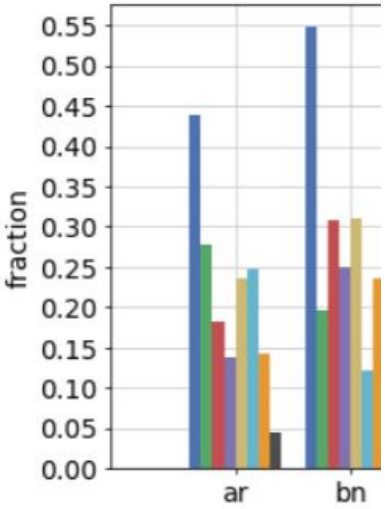
- Responses from 210K readers of 14 different languages
- Motivation: *I am reading this article because ...*

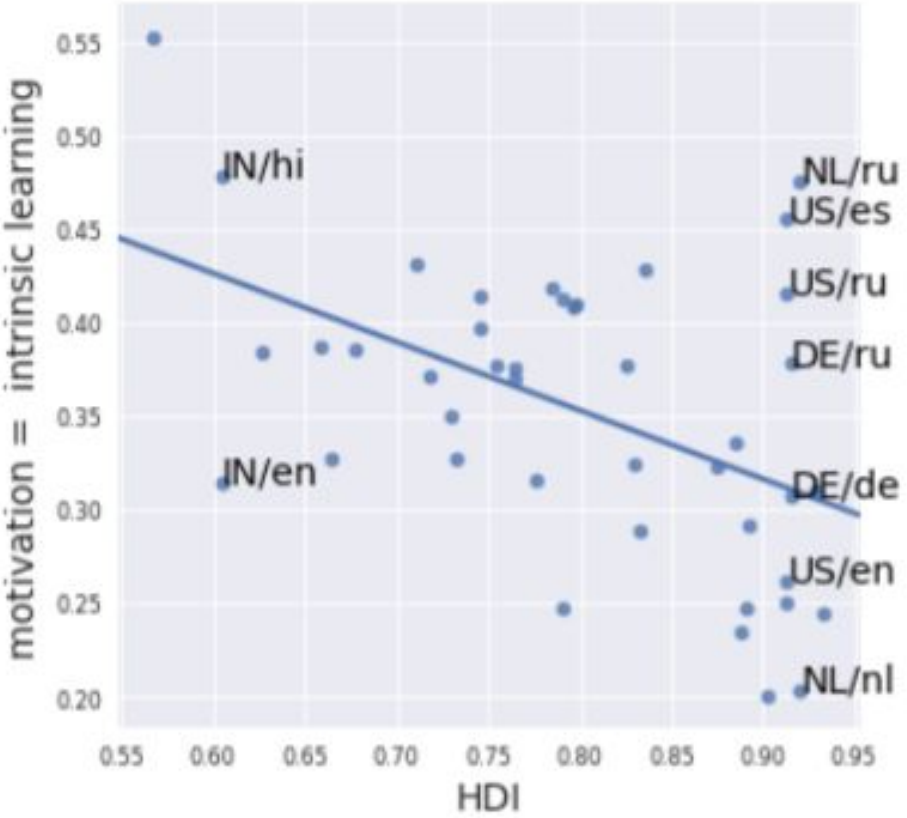


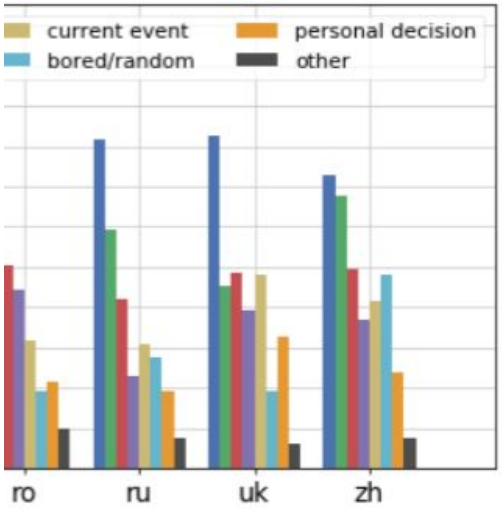
Survey

- Response
- Motivatic

ages







When readers visit Wikipedia?

Curious Rhythms: Temporal Regularities of Wikipedia Consumption *(arXiv:2305.09497)*

Log-based analysis

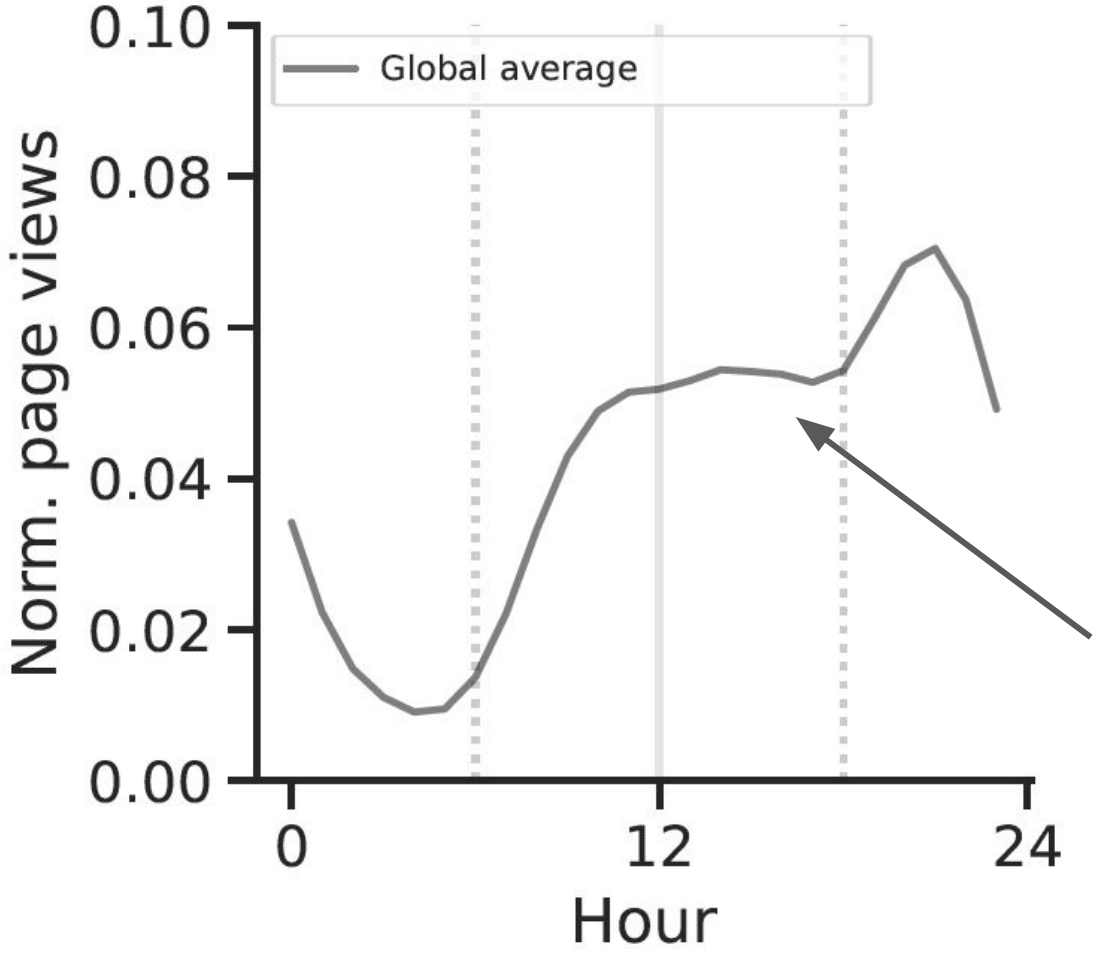
Data: one month of request to articles in English Wikipedia

-

Article	Country	Device	*Timestamp*	*TZ*	Local time
Bayesian inference	**USA**	**Desktop**	*2021-04-12 18:29:51*	*UTC-7*	**2021-04-12 11:29:51**
Avengers: Endgame	**Greece**	**Mobile**	*2021-04-12 18:30:26*	*UTC+3*	**2021-04-12 21:30:26**
Bayesian inference	**Mexico**	**Desktop**	*2021-04-12 18:30:51*	*UTC-6*	**2021-04-12 12:30:51**
Vikings	**USA**	**Desktop**	*2021-04-12 18:31:33*	*UTC-4*	**2021-04-12 14:31:33**
...	...	...	...	...	...

3.45B pageload events

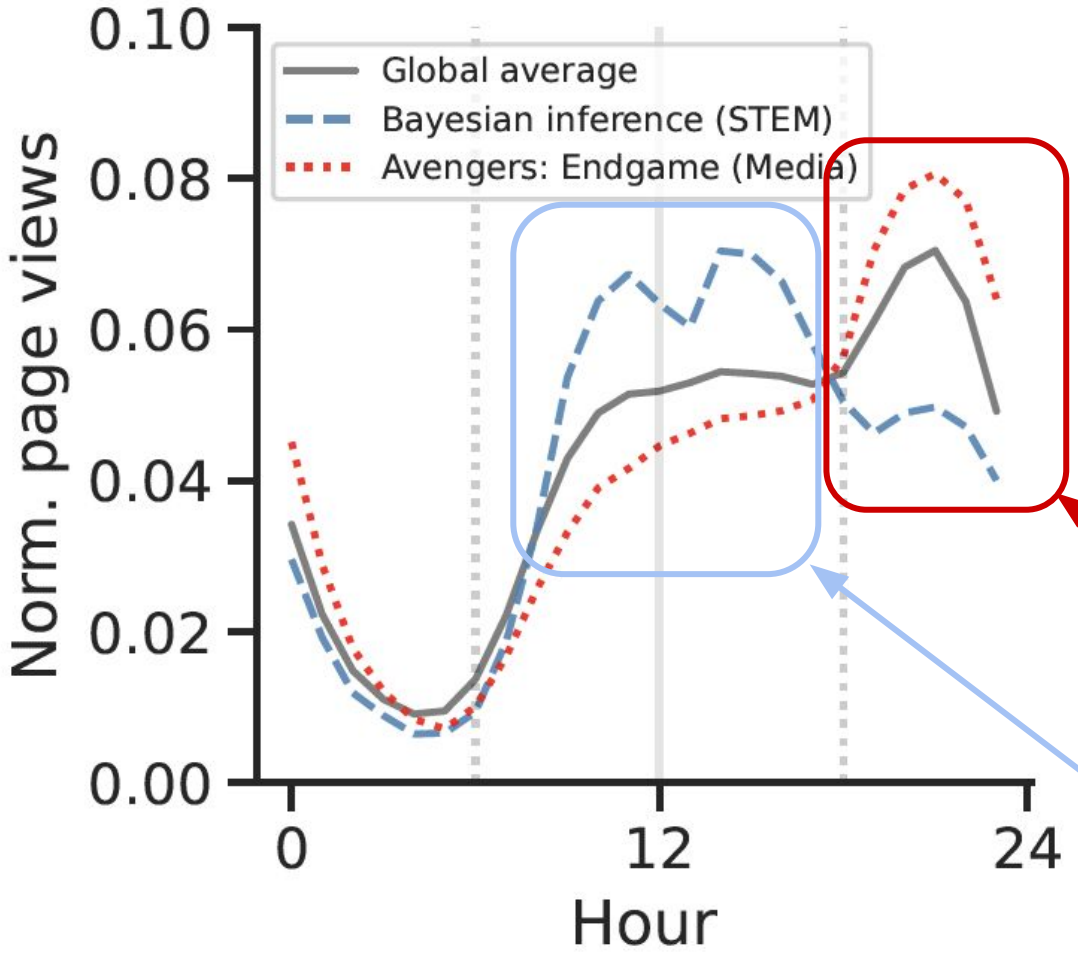
6.3M articles



Normalised daily pattern

$$\frac{\text{Pageloads in hour H}}{\text{Total pageloads}}$$

Circadian rhythm



Normalised daily pattern

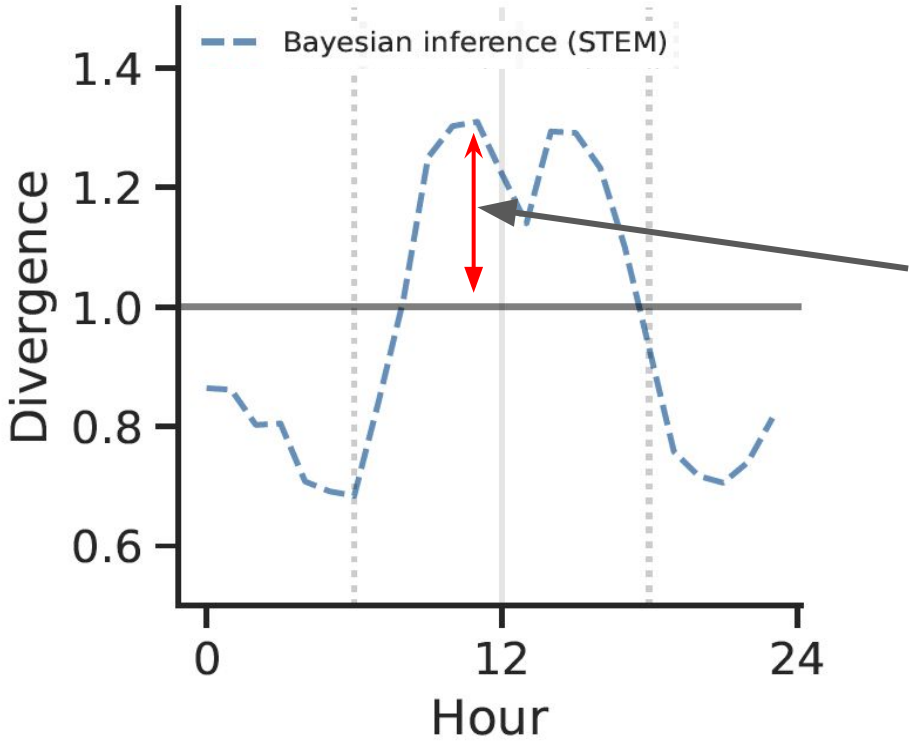
Pageloads in hour H for page P

Total pageloads for page P

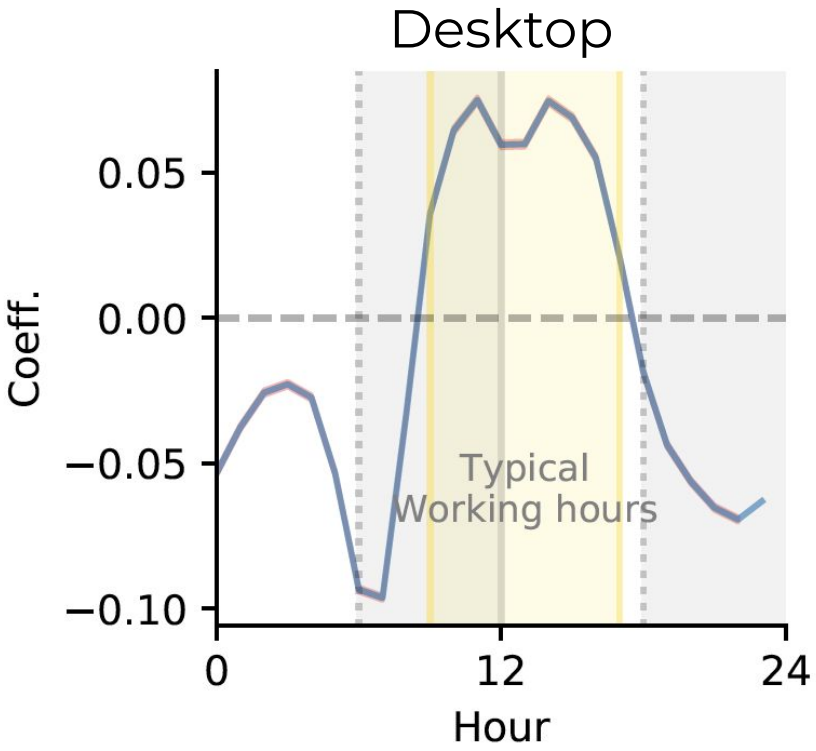
Media

STEM

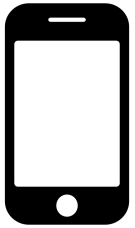
Linear regression



Predict the deviation from the pageload features

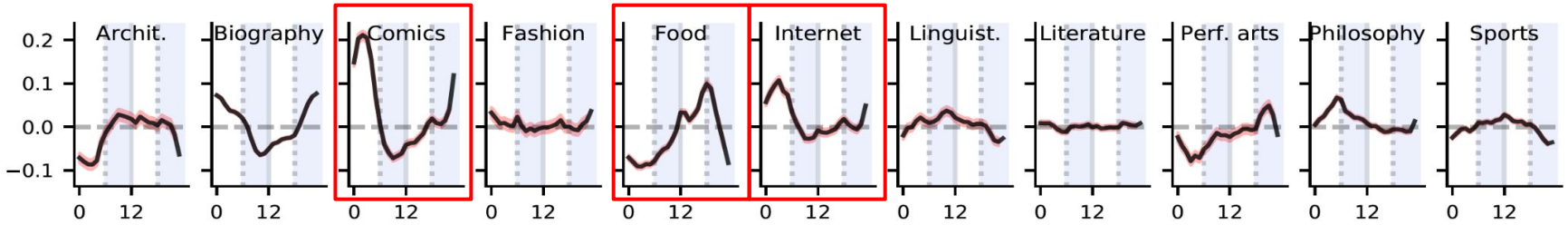


Mobile traffic higher during evening

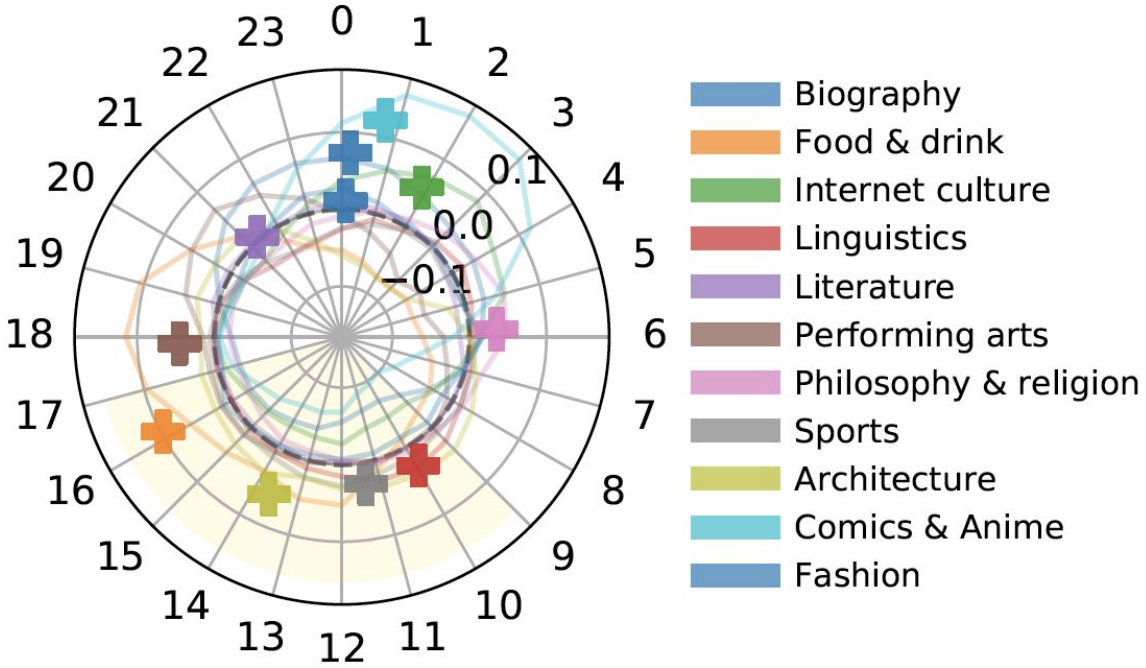


40% **60%**

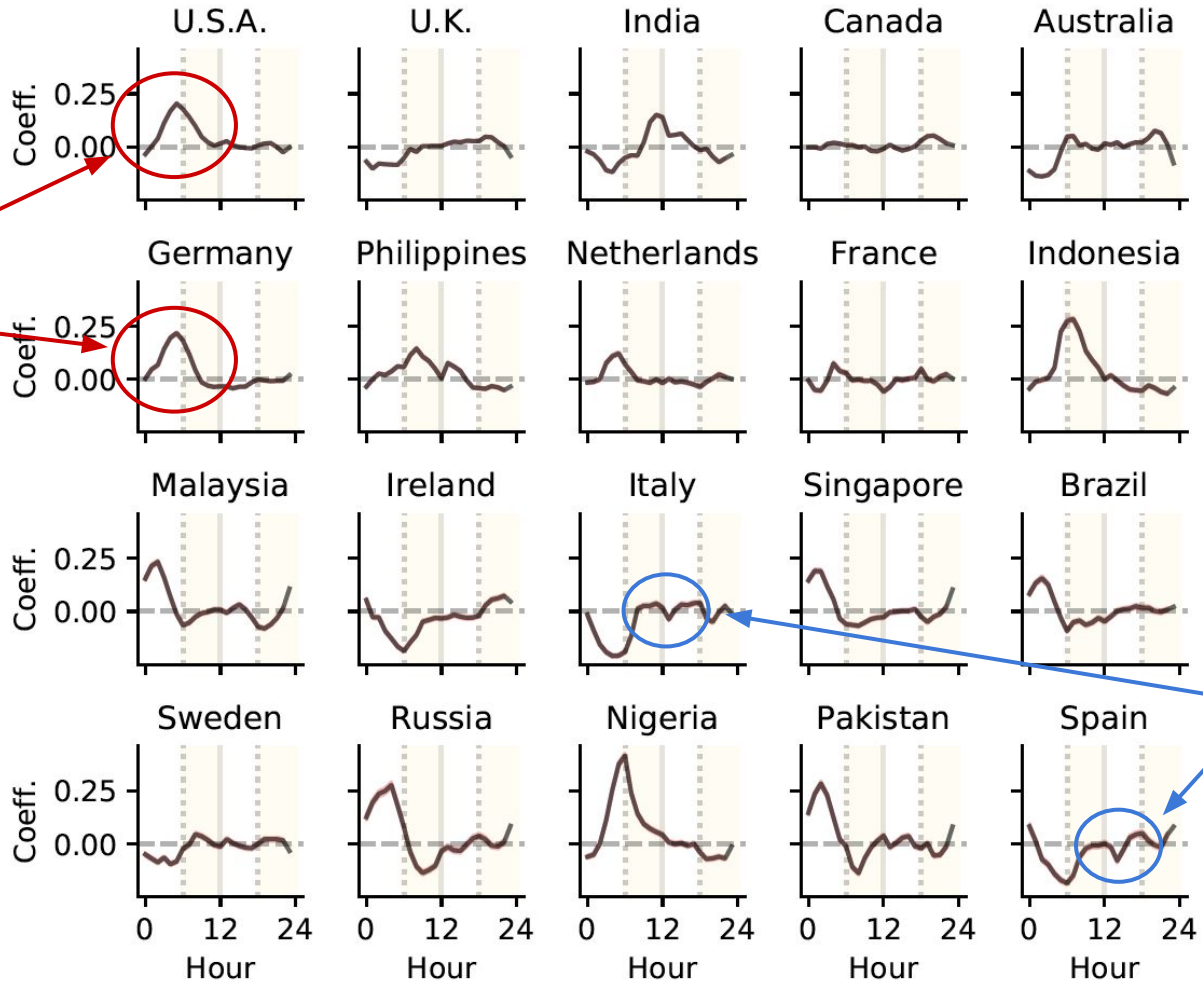
Device



Culture



Topics


U.S.A.
U.K.
India
Canada
Australia
Germany
Philippines
Netherlands
France
Indonesia
Malaysia
Ireland
Italy
Singapore
Brazil
Sweden
Russia
Nigeria
Pakistan
Spain
Coeff.
0.25
0.00
0
12
24
Hour
Similar consumption patterns
Similar consumption patterns
Country

How do readers navigate?

A Large-Scale Characterization of How Readers Browse Wikipedia *(ACM TWEB 2023)*
Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions *(WWW'22)*
Wikipedia Reader Navigation: When Synthetic Data Is Enough *(WSDM'22)*

Observing reader navigation from logs

Complexity: Reconstructing paths from individual pageloads

- Pseudo-user IDs from hash of IP+user_agent (no cookies)
- The referrer tells us where each request is from

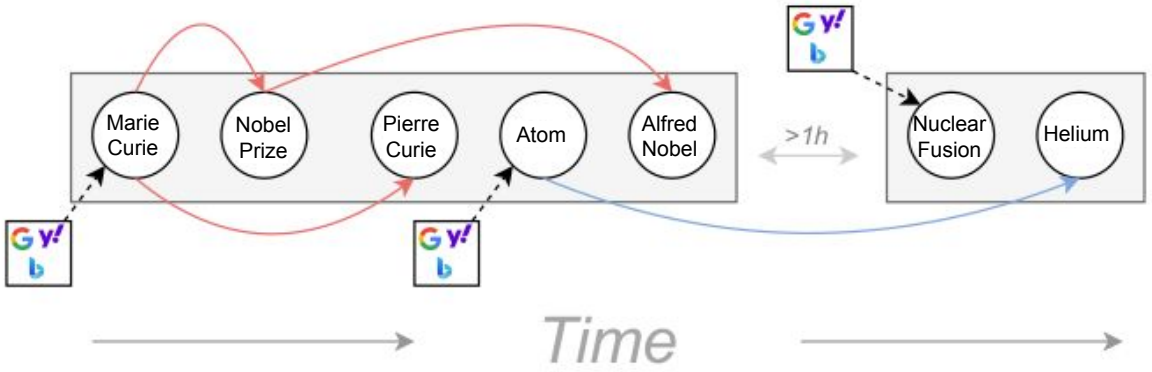
Time

User id	Timestamp	Article	Referrer
d6ni1i9fgl	2021-04-12 11:29:51	**A**	bing.com
d6ni1i9fgl	2021-04-12 11:31:26	**B**	WP: A
d6ni1i9fgl	2021-04-12 11:31:33	**C**	WP: A
d6ni1i9fgl	2021-04-12 11:36:16	**D**	WP: C
d6ni1i9fgl	2021-04-12 11:37:50	**E**	facebook.com
...	...	...	

Observing reader navigation from logs

Complexity: Reconstructing paths from individual pageloads

- 1.47B unique reading sessions

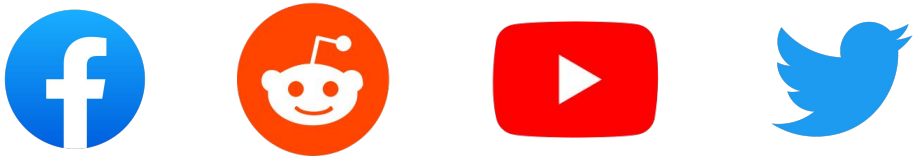


Reaching Wikipedia

Where reading sessions start?

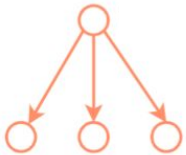
- 77% start from search engines

- 20% of external traffic unspecified/empty
- 1.5% from external websites



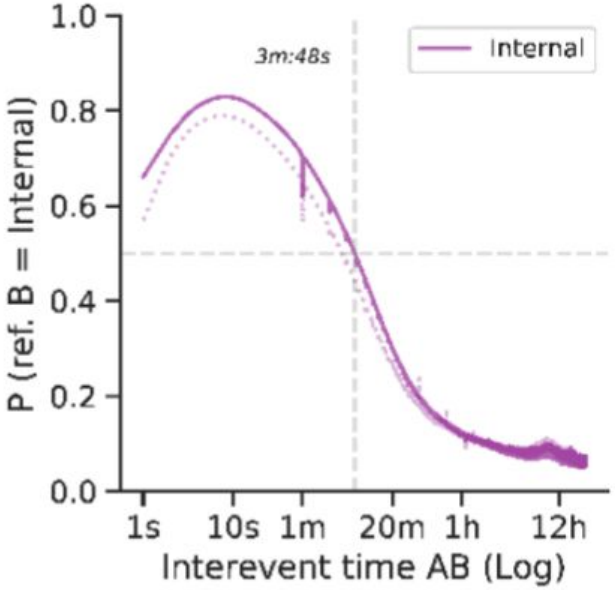
Structural properties

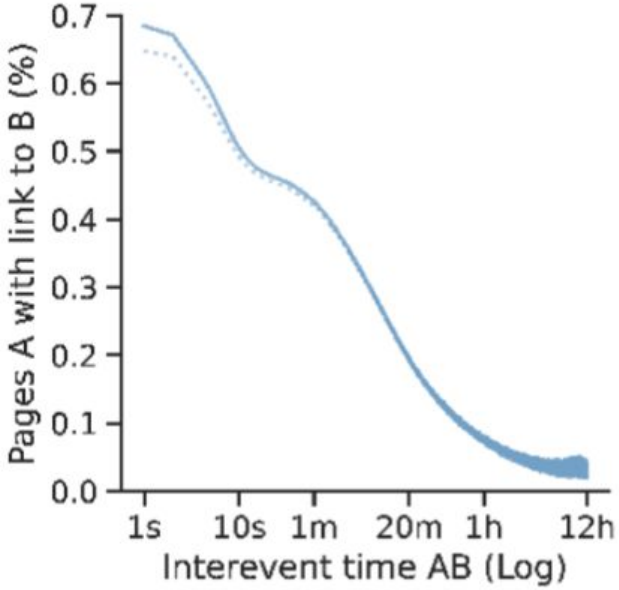
- Navigation is usually fast
  - median 74 s time between pageloads
- Navigation is short
  - ~73% with only 1 pageloads (90% have less than 4 pageviews)
- Depends on context and information need
  - Device: Longer sessions on desktop than on mobile (2.4 vs 1.99)
  - Topic:
    - Length: longer (entertainment) vs shorter (STEM)
    - Strategy: breadth (entertainment) vs depth (STEM)



Using external search for navigating Wikipedia

- 40% of pairs of consecutive pageloads: reader leave and re-enter via search engine
- in 30% of these cases internal link available





Targeted Navigation

Lab-based studies to understand human navigation

Wikispeedia

This game is easy and fun:

- You are given two Wikipedia articles* (or you choose two yourself).
- Starting from the first article, your goal is to reach the second one, exclusively by following links in the articles you encounter. (For the articles you are given this is always possible.)

Mission	*Clicks*	*Avg. hardness***	
Where Did Our Love Go >> Fine art	avg. 6, record 6	-	**Play!**
Windows XP >> Romania	avg. 5.8, record 3	3	**Play!**
Fertile Crescent >> Levee failures in Greater New Orleans, 2005	avg. 5, record 5	3.5	**Play!**
Corporation >> Mars Exploration Rover	avg. 5, record 5	2	**Play!**
Antananarivo >> Amsterdam	avg. 4.5, record 4	-	**Play!**

** On a scale from 1 (easy) to 5 (brutal); if you want to make this more reliable, just hit the *rate* button after each game.

Targeted navigation

Lab-based studies to understand human navigation

Wikispeedia

This game is easy and fun:

- You are given two Wikipedia articles* (or you choose two yourself).
- Starting from the first article, your goal is to reach the second one, exclusively by following links in the articles you encounter. (For the articles you are given this is always possible.)

Human wayfinding in information networks

Authors: Robert West, Jure Leskovec Authors Info & Claims

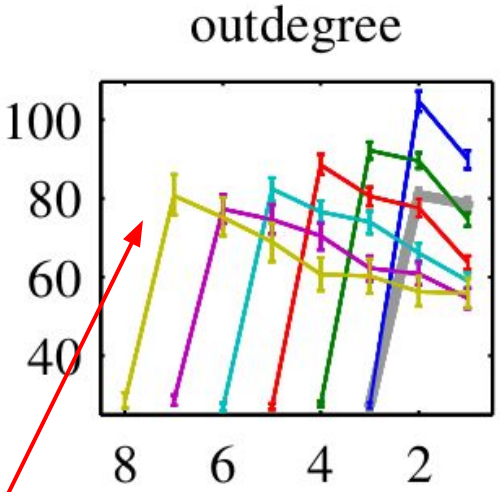
WWW '12: Proceedings of the 21st international conference on World Wide Web • April 2012 • Pages 619–628 • https://doi.org/10.1145/2187836.2187920

	*Avg. hardness***	
record 6	-	**Play!**
8, record 3	3	**Play!**
record 5	3.5	**Play!**
record 5	2	**Play!**
5, record 4	-	**Play!**

e *rate* button after each game.

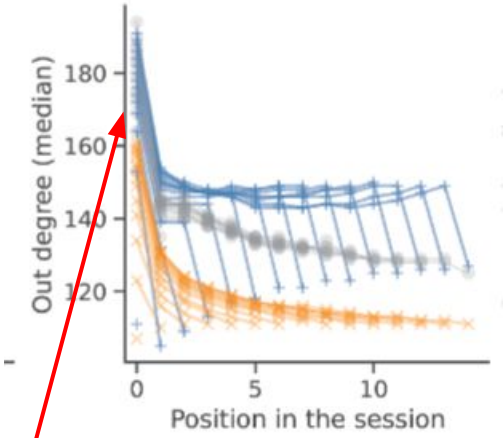
Targeted navigation

Targeted navigation



≠

Navigation “in the wild”

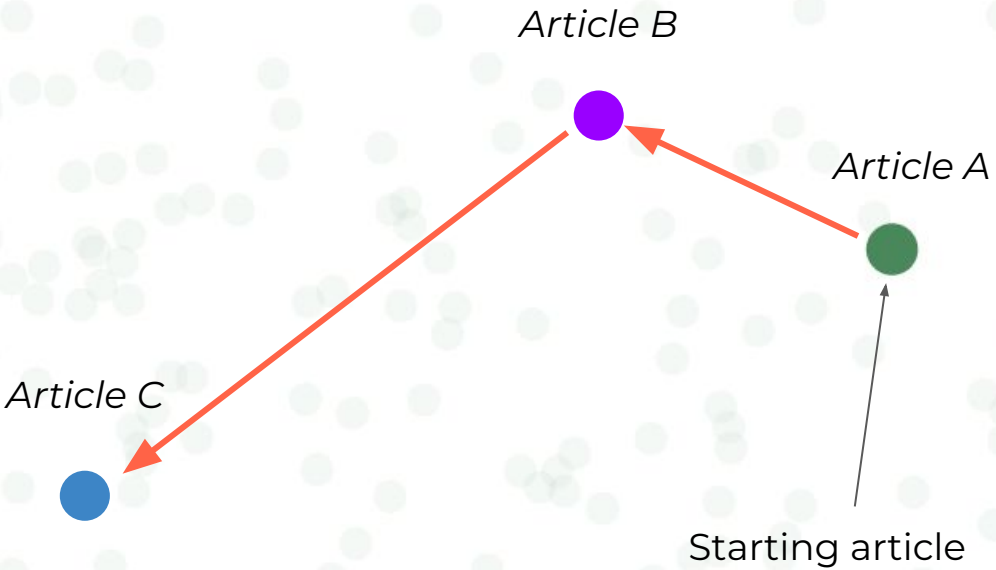


- Strategy: Use of hubs *after* first step

- Entry point with high out-degree; popularity of entry points
- Navigation after first step has roug out-degree

West & Leskovec. **Human wayfinding in information networks.** *WWW’12*

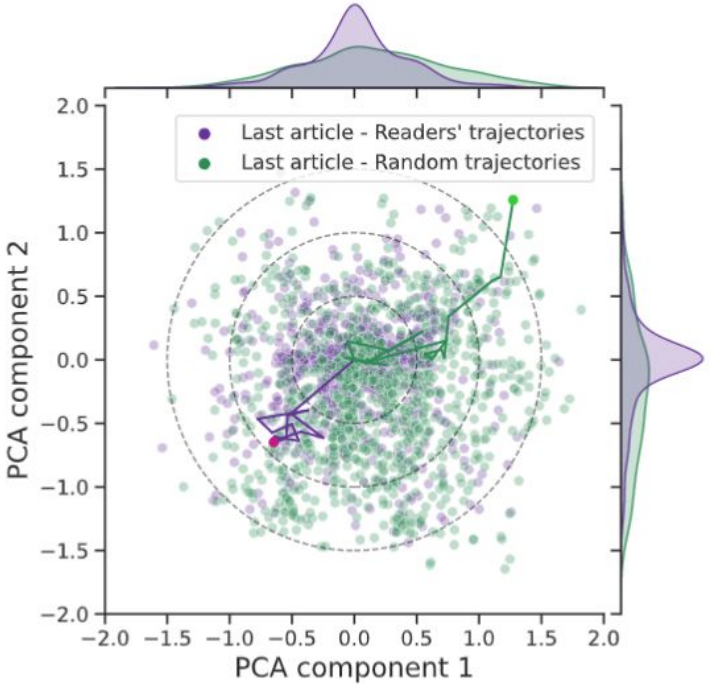
Diffusion in topic space

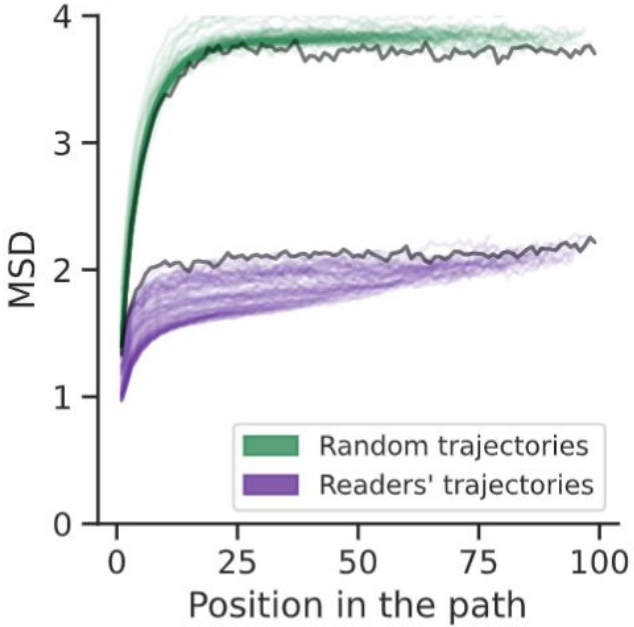


Embedding space (text)

Diffusion in topic space

Navigation “in the wild” is also different from random walks

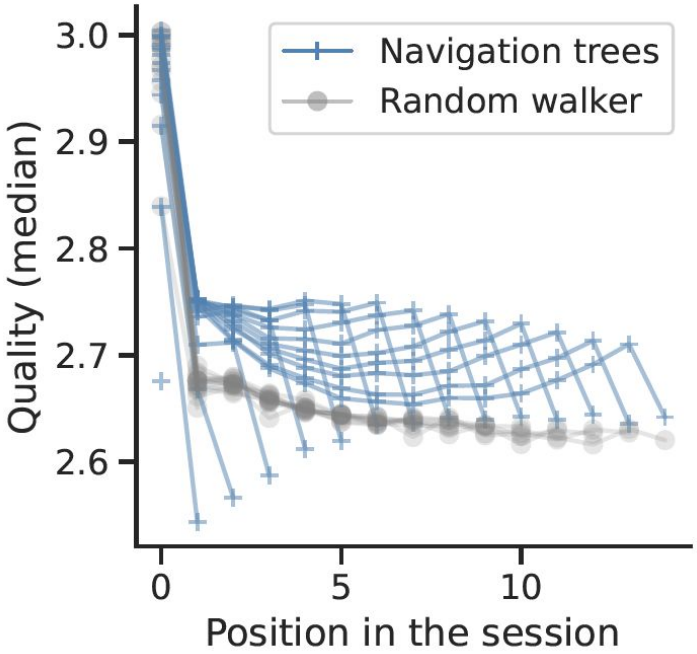




Encountering low-quality pages

Readers give up navigation when encountering low-quality pages

- On average, the last article of the session shows a drop in quality



How to make it easier to navigate?

Orphan articles: The dark matter of Wikipedia *(arXiv:2306.03940)*

Orphan articles

Def.: no incoming links

> This article **is an orphan, as no other articles link to it**. Please introduce links to this page from related articles; try the Find link tool for suggestions. *(March 2023)*

Expedition Medicine (sometimes known as **expeditionary medicine**) is the field of medicine focusing on providing embedded medical support to an expedition, usually in medically austere or isolated areas. Expedition medicine provides the physical and psychological wellbeing of expedition members before, during, and after an expedition. Expedition medicine may be practiced in support of commercial, non-governmental organizations, and government expeditions. Some medical governing bodies consider expedition medicine as a field within wilderness medicine, whilst others considered it be a separate discipline.[1][2]

Expedition medicine

Subdivisions	*Travel Medicine* *General environmental medicine* *Battlefield medicine*

History [edit]

This field of expedition medicine has ancient origins and has been practised almost since the advent of medicine and expeditions. Many ancient civilizations embedded medical staff with military units.[3]

During the Age of Discovery, expedition

Medical equipment used by Robert Falcon Scott on his 1910 Antarctic expedition

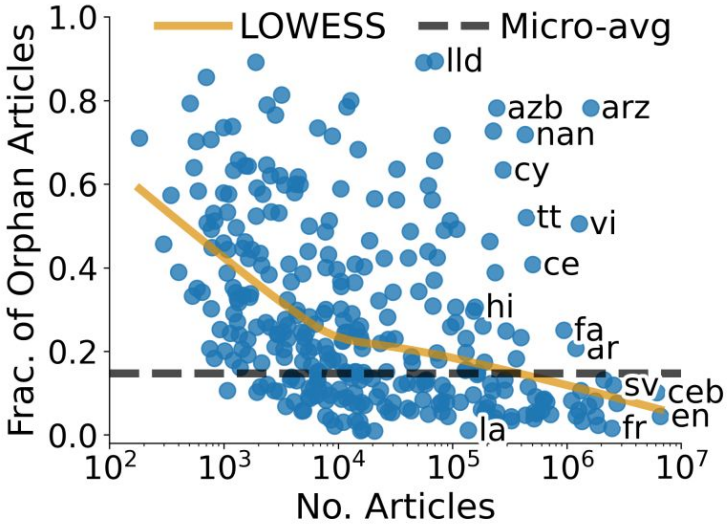
(In-) Visibility and knowledge gaps

- Links are crucial
  - "build the web" to enable readers to access relevant information on other Wikipedia pages easily. (**WP:BUILD**)
  - 38% of pageviews result from traffic via internal hyperlinks (Piccardi et al. 2023)
- Visibility as a structural bias
  - Biographies of women are less visible than biographies on men (Wagner et al. 2016)
    e.g. systematically lower scores for pagerank
- Communities are struggling to address this
  - campaigns are good at adding/improving the content about women
    however, they are less successful at addressing structural biases that limit their visibility (Langrock et al. 2022)

Orphan articles

Orphans are the dark matter of Wikipedia:

- Orphans are de facto invisible for readers navigating Wikipedia
- Orphans make up a large chunk of all content

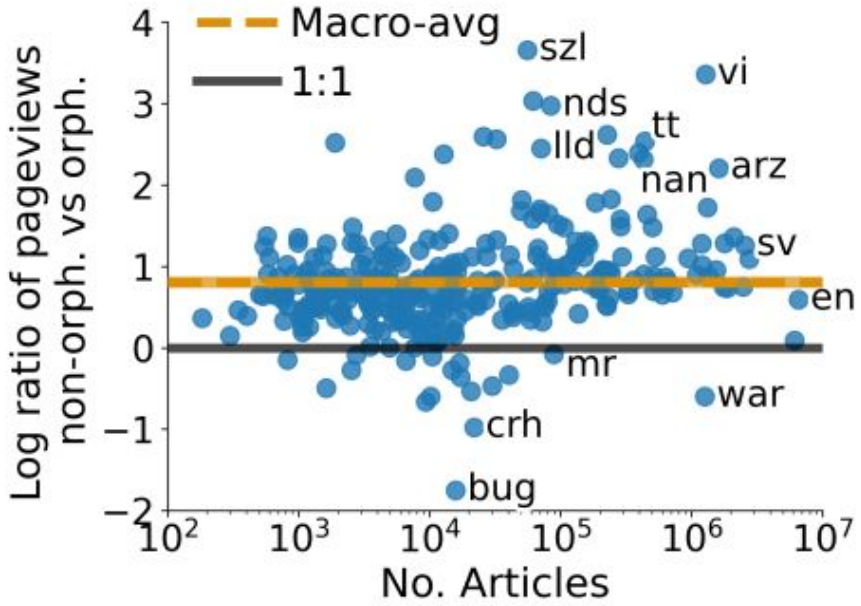


~15%: 8.9M / 60M articles

across 300+ language version

Orphans are less visible

Correlation: Orphans receive less pageviews than non-orphans



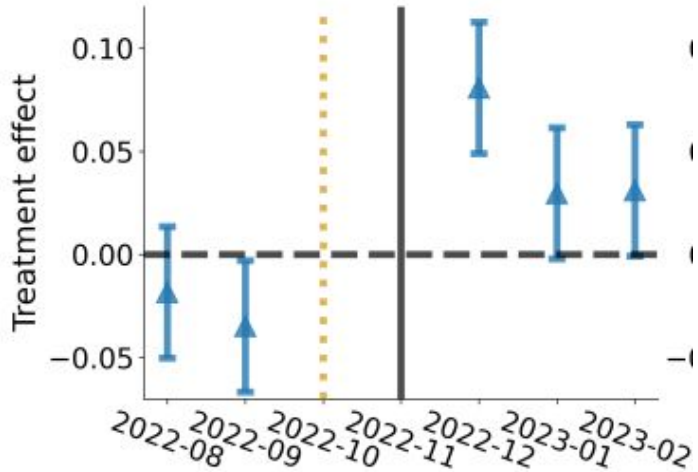
$$\frac{\text{median pageviews non-orphans}}{\text{median pageviews orphans}} > 2$$

Orphans are less visible

Establishing causality

- Treatment: Orphan article a in language w receives a new inlink
- Control: same orphan article a in language w!=w' remains orphan

Difference-in-differences



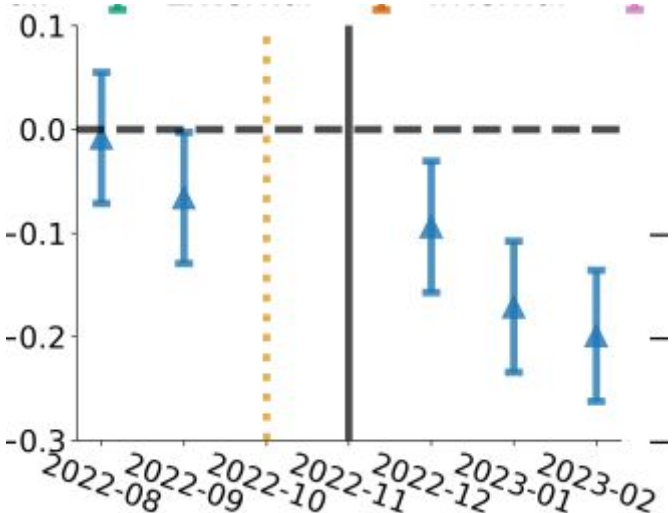
- 36K treatment-control pairs (192 languages)
- 6.5% increase overall (p<10^-10)
- Increase persists following months
- Driven by added internal links
-

Orphans are less visible

Establishing causality II - inverting the treatment

- Treatment: non-orphan article a in language w becomes an orphan
- Control: same non-orphan article a in language w!=w’ stays non-orphan

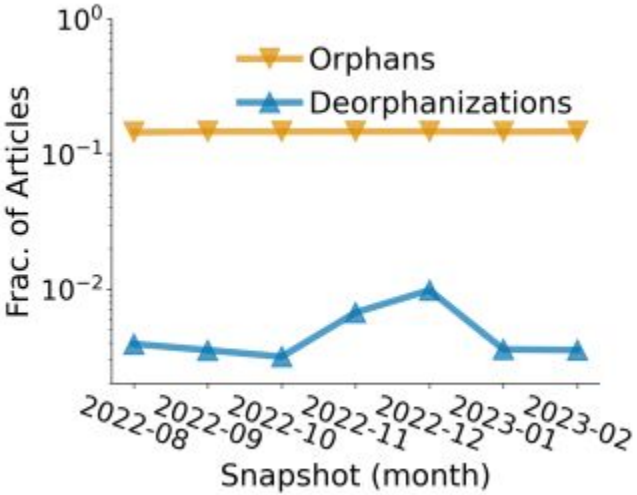
Difference-in-differences



- 12K treatment-control pairs (121 languages)
- 13% decrease overall (p<10^-10)
- Increase persists following months
- Driven by added internal links

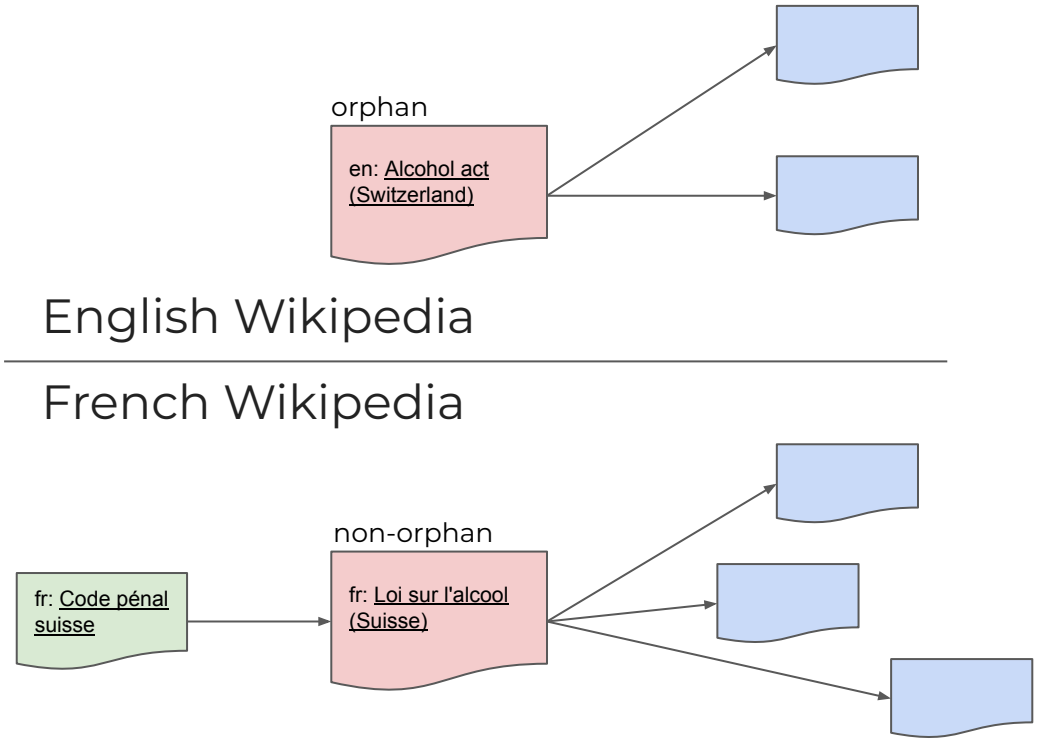
Challenges for editors

- Editors are struggling to add links to orphans
  - At the current rate, it would take editors >100 months to work through backlog of orphans



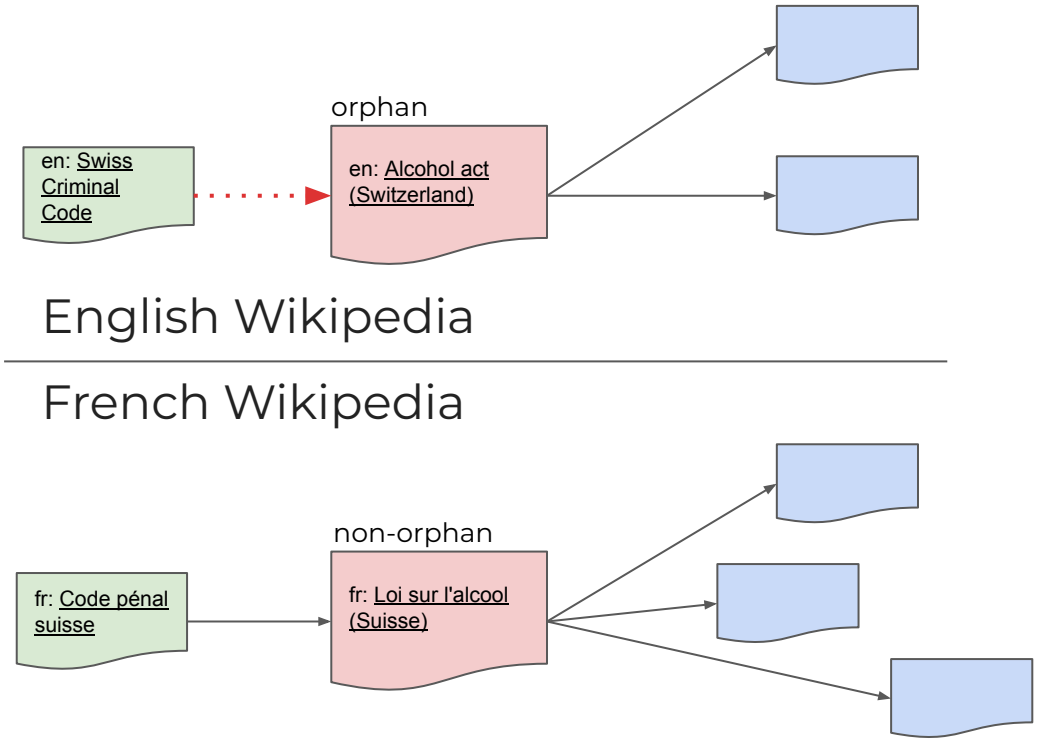
Opportunities: Link translation

- Developing automatic tools to support editors



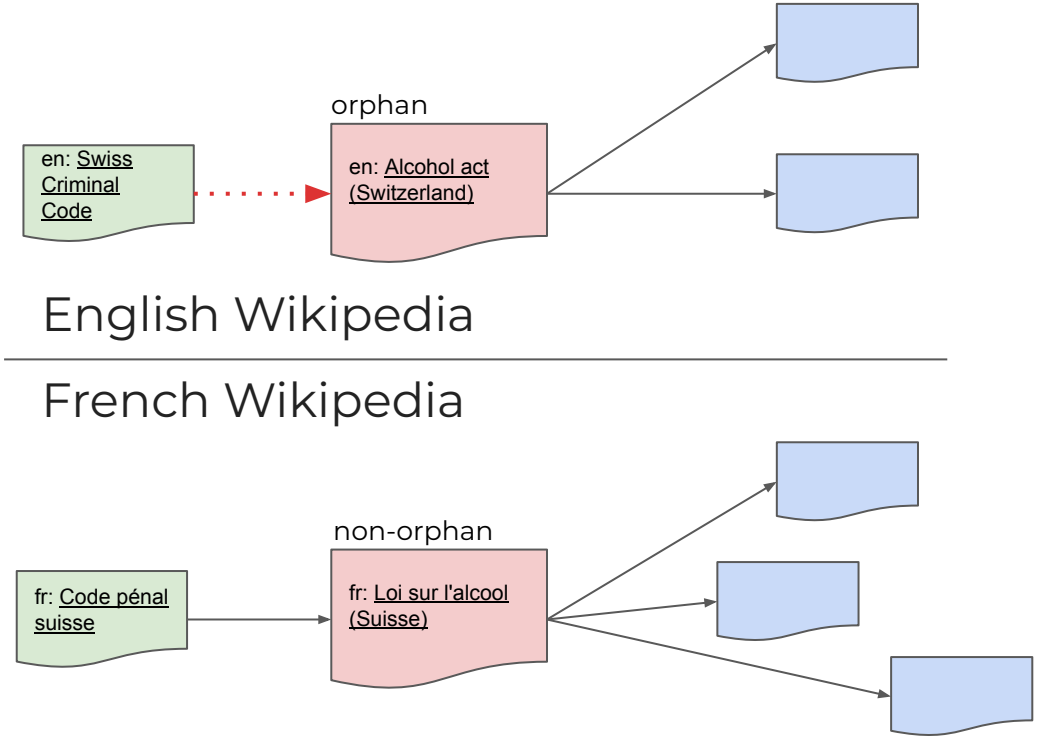
Opportunities: Link translation

- Developing automatic tools to support editors

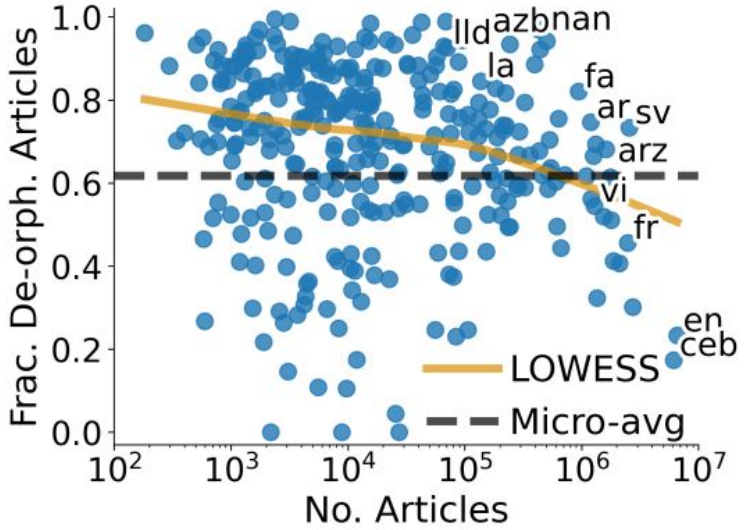


Opportunities: Link translation

- Developing automatic tools to support editors



Link Translation



Concluding remarks

- Readers are a crucial dimension in understanding knowledge gaps in Wikipedia
- Log-based analysis offers insights into information needs of readers
- Theoretical and practical implications from studying information seeking
  - Targeted navigation in lab-based settings
  - Interdependence with external search engines
- Improving navigation
  - Orphan articles as the dark matter of Wikipedia
  - Preferential attachment models for understanding network growth
- Maintenance vs Growth
  - Adding new content vs improving content (accessibility via links, quality, disinformation etc.)

Thank you!

Thanks to collaborators
- Akhil Arora (EPFL)
- Tiziano Piccardi (Stanford)
- Robert West (EPFL)

Reach out: mgerlach@wikimedia.org

Learn more: https://research.wikimedia.org/