An analysis of scholarly citations in Wikipedia

Supervisor
Prof. Alberto Montresor
Co-Supervisor
Cristian Consonni

Graduand
Alessio Bogon

University of Trento
Department of Information Engineering and Computer Science

Master Degree in Computer Science

March 24, 2016

Overview

Wikipedia

- Free-access, free-content Internet encyclopedia
- One of the most popular web sites
- Started in 2001
- Studied by many
  - Quality of citations [4]
  - Illness prediction [2]
  - Stock market moves strategy [3]

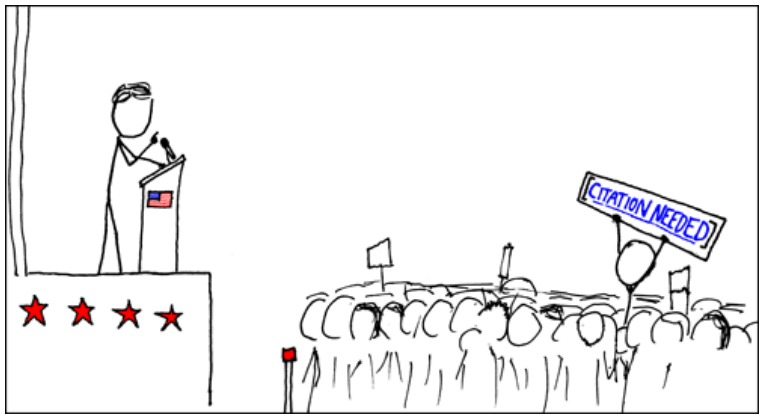


Figure: Wikipedian protester (`xkcd.com/285`)

Purpose of this work

- Analyze the quality of papers in Wikipedia, in term of:
  - Freshness of citations
  - Popularity of cited papers
  - Rank of cited journals
  - Lifetime of citations

Available datasets

- Microsoft Academic Graph: a dataset containing data about papers, authors, references, journals, conferences, etc.
- Wikipedia dumps: text of all page revisions since the beginning
- Wikimedia hourly page view statistics

Microsoft Academic Graph

- Dataset powering the Microsoft Academic Search engine
- Size: 96 GB
- Contains over 120M papers (1800 – 2016)
- Information about papers, authors, references, journals, conferences, keywords, etc.
- Papers identified by a Digital Object Identifier (DOI)
- Problems
  - Only *computer science* conferences
  - Some of papers' publication dates are incomplete
  - Not all the papers have a DOI (32% of them)

The missing pieces — Contributions

- History of papers appearing in Wikipedia (where and when)
- Page views dataset for large-scale article analysis

Extracting citations from Wikipedia

Problems

- Citations can be structured: *wikimarkup* templates
  - Many different variants
  - Anybody can use custom macros
  - Different templates for each language
- or unstructured: plain text
  - Recognize substrings that appear to be citations
- Entity disambiguation
- Dataset size: 13,3 TB as of September 1st, 2015

Solution

- Focus on publication identifiers (*DOI*, *PMID*, *arXiv*, *ISBN*)
- The **wikidump** framework

Wikidump

- Facility framework to extract **features** from Wikipedia XML dumps
  - Publication identifiers
  - Wikilinks
  - Page statistics
- Based on libraries by Aaron Halfaker
- Low memory consumption
- Highly parallelizable
- Written in Python

- Processed 445M page revisions (13,3 TB) in 21 hours

Type	Count
ISBN	1 153 330
DOI	651 199
PMID	372 939
arXiv	18 832

Table: Number of identifiers extracted

Wikimedia page views

Contains information about page visualization for all the Wikimedia projects, for all the languages.

Problems

- Dataset size: 23 TB (4,7 TB for the 2014)
- Aggregated and ordered by hour (8670 files per year)
- They need to be cleaned
- They need to be **reordered**
- Unfeasible on a single machine

Solution

- Clean and sort the dataset exploiting the UniTN Cisca Cluster
- 2014 dataset already published [1]

The Spark job

- Apache Spark: framework for large-scale data processing
- UniTN Cisca Cluster:
  - 125 workstations
  - 500 CPU cores in total
  - Available only at night and in the weekend
- Steps:
  - Normalize the content
  - Repartition the keyspace
  - Sort each partition locally

- Took one night to analyze and recreate the 2014 dataset
- Took many nights to get it to work

The Spark job — Workflow

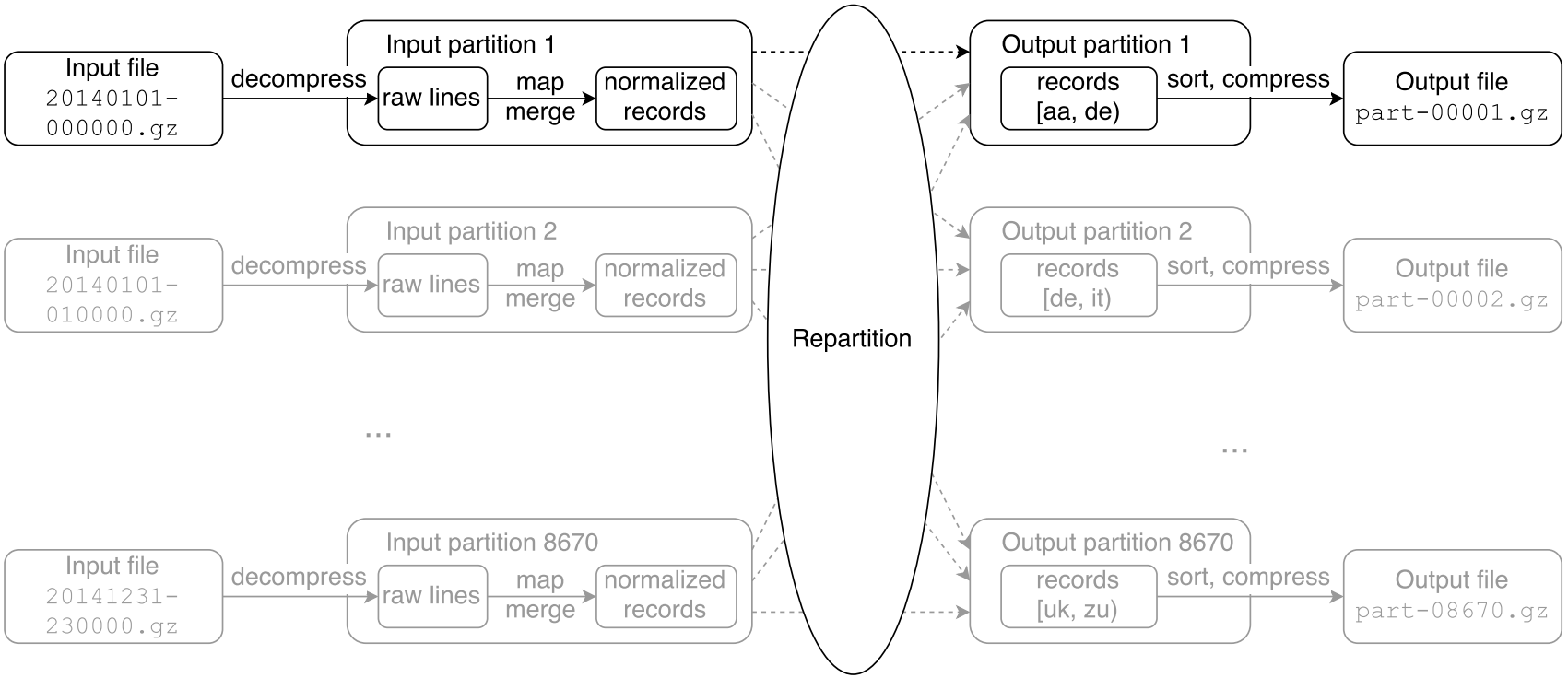
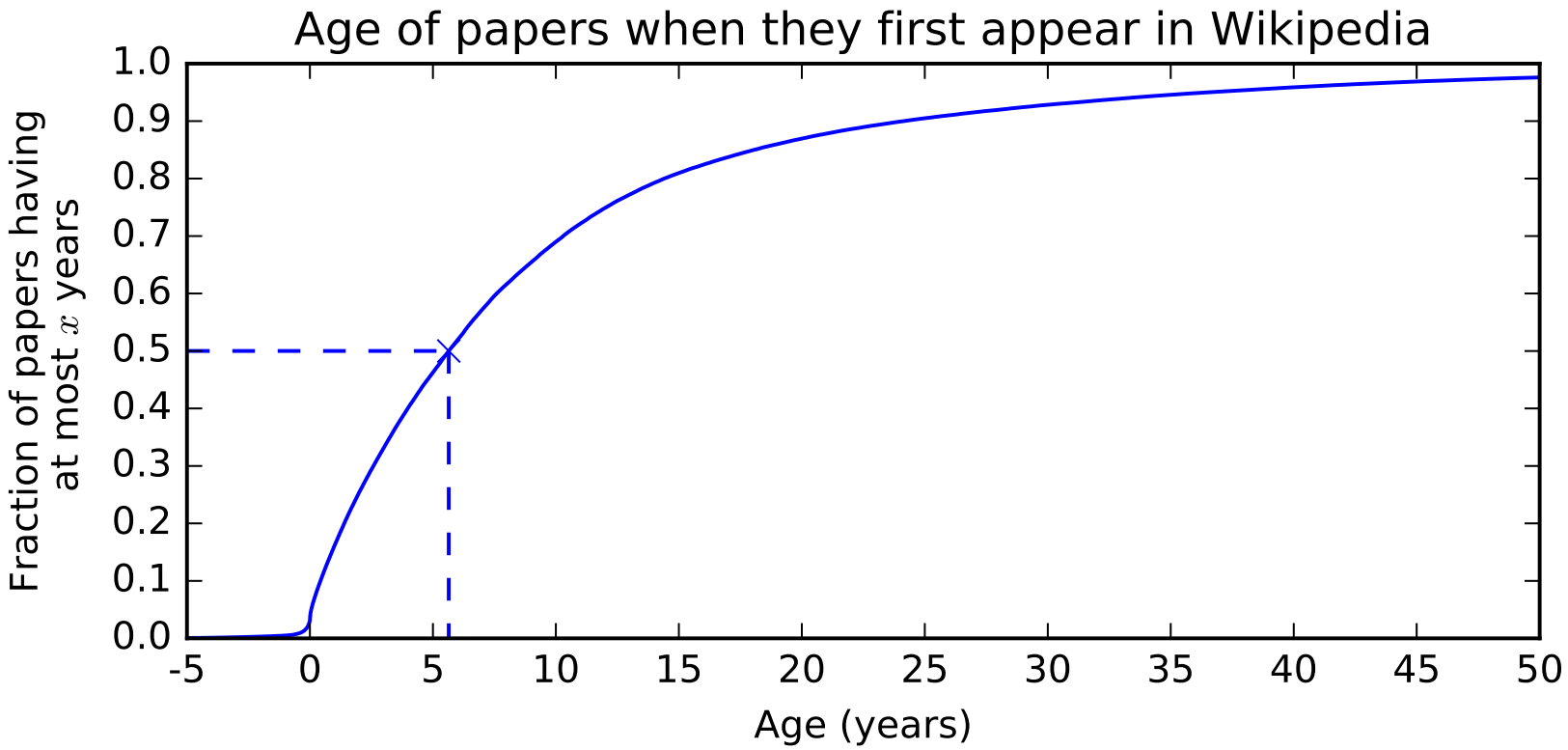


Figure: Workflow showing the processing of Wikimedia page views

Age of papers when inserted

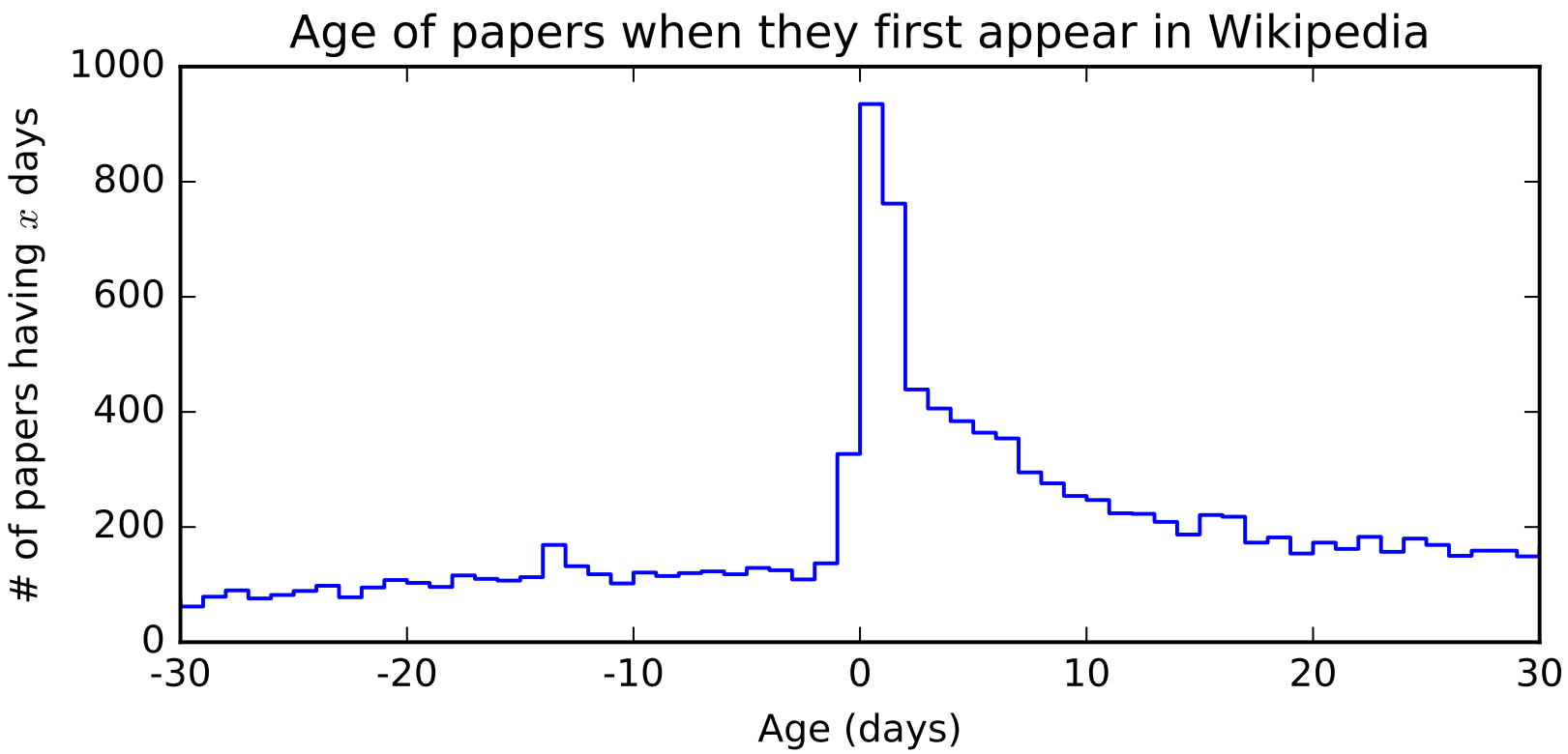
- How old is a paper when it is inserted in Wikipedia for the first time?
- Exploit the DOI mapping with the Microsoft dataset
- Interesting behavior of papers having few days

Age of papers when inserted



- Half of the papers cited are less than 5.5 years old

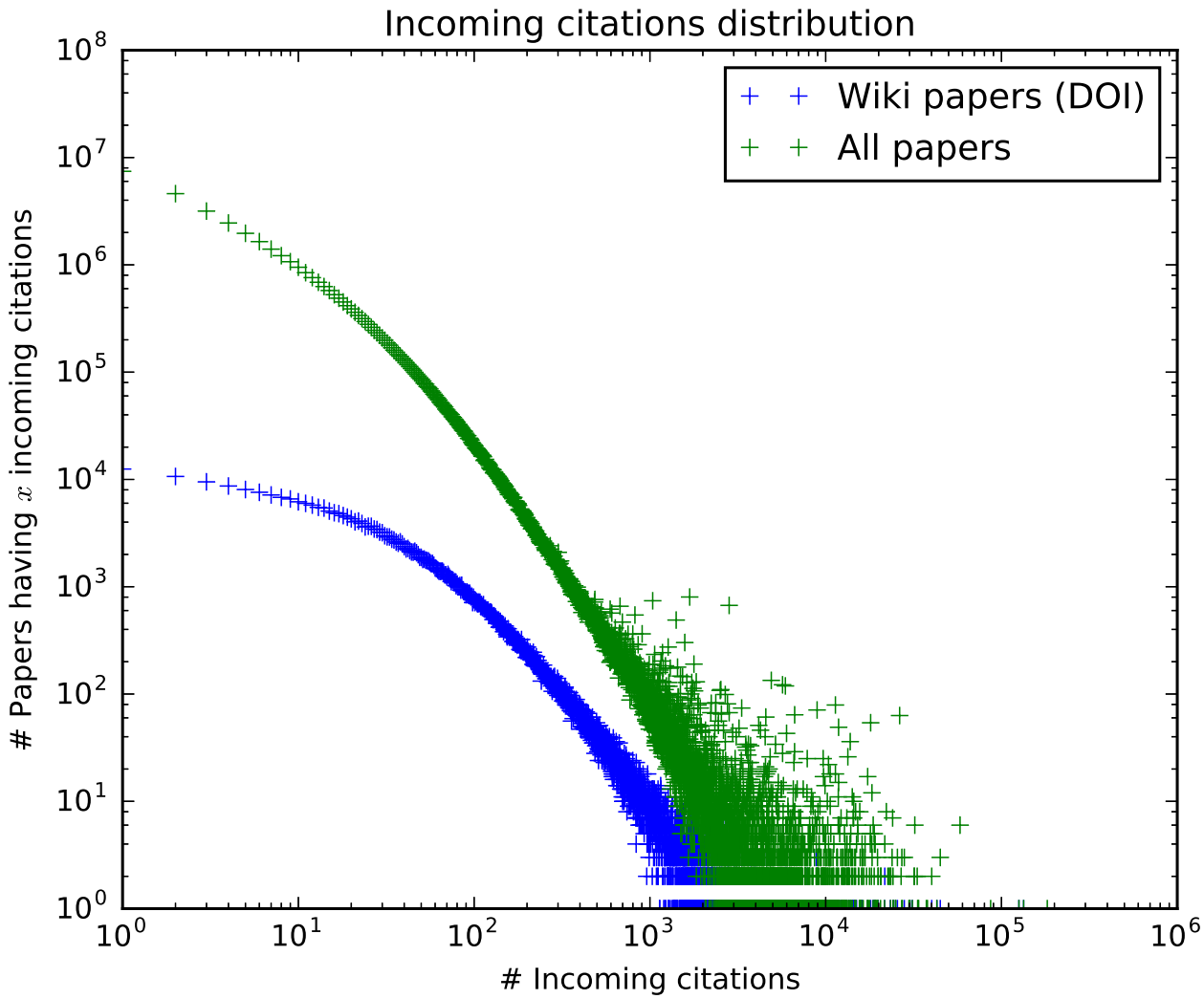
Age of papers when inserted — Detail



- 1.3% of papers are inserted within 7 days after the publication

Incoming citations distribution

- Arguably follows a power law [5]: $N(x) \sim x^{-\alpha}$
- How well papers in Wikipedia perform?



Incoming citations distribution

- Papers in Wikipedia behave like *Genome Research* and *PNAS*
- They outclass *Nature* and *Science*
- 74% of papers in Wikipedia have more than 10 incoming citations

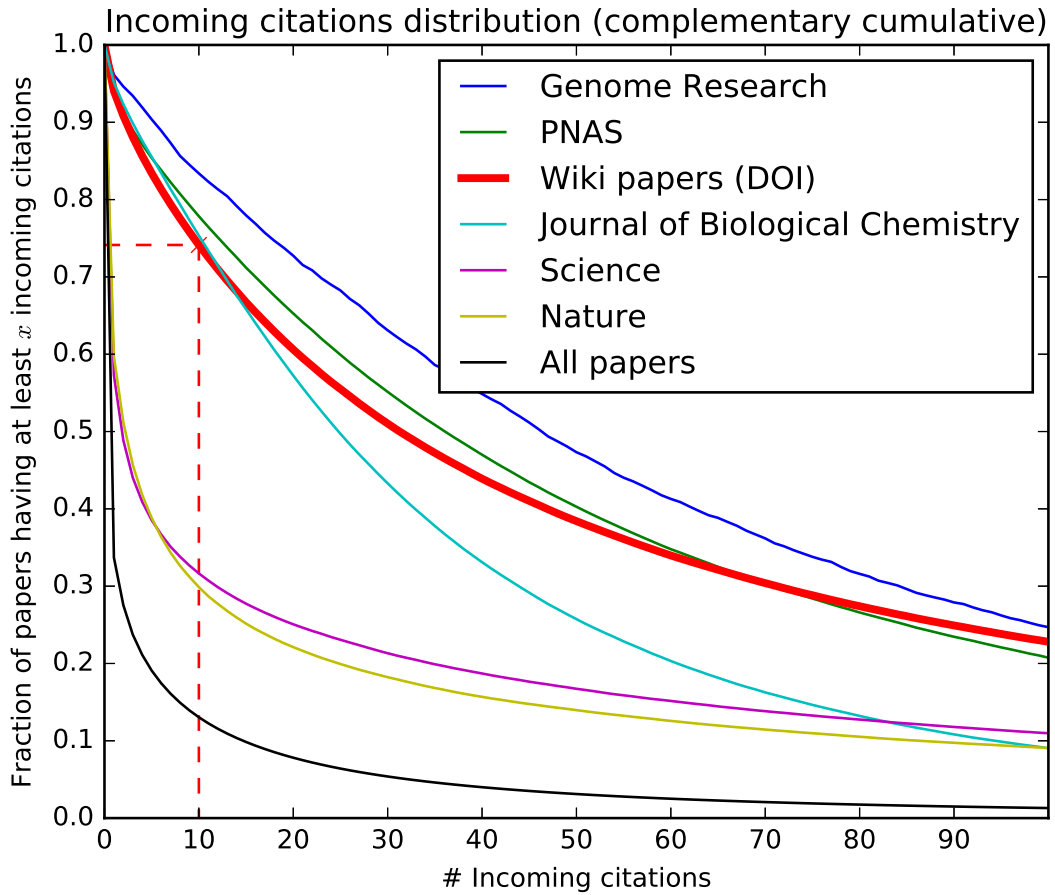


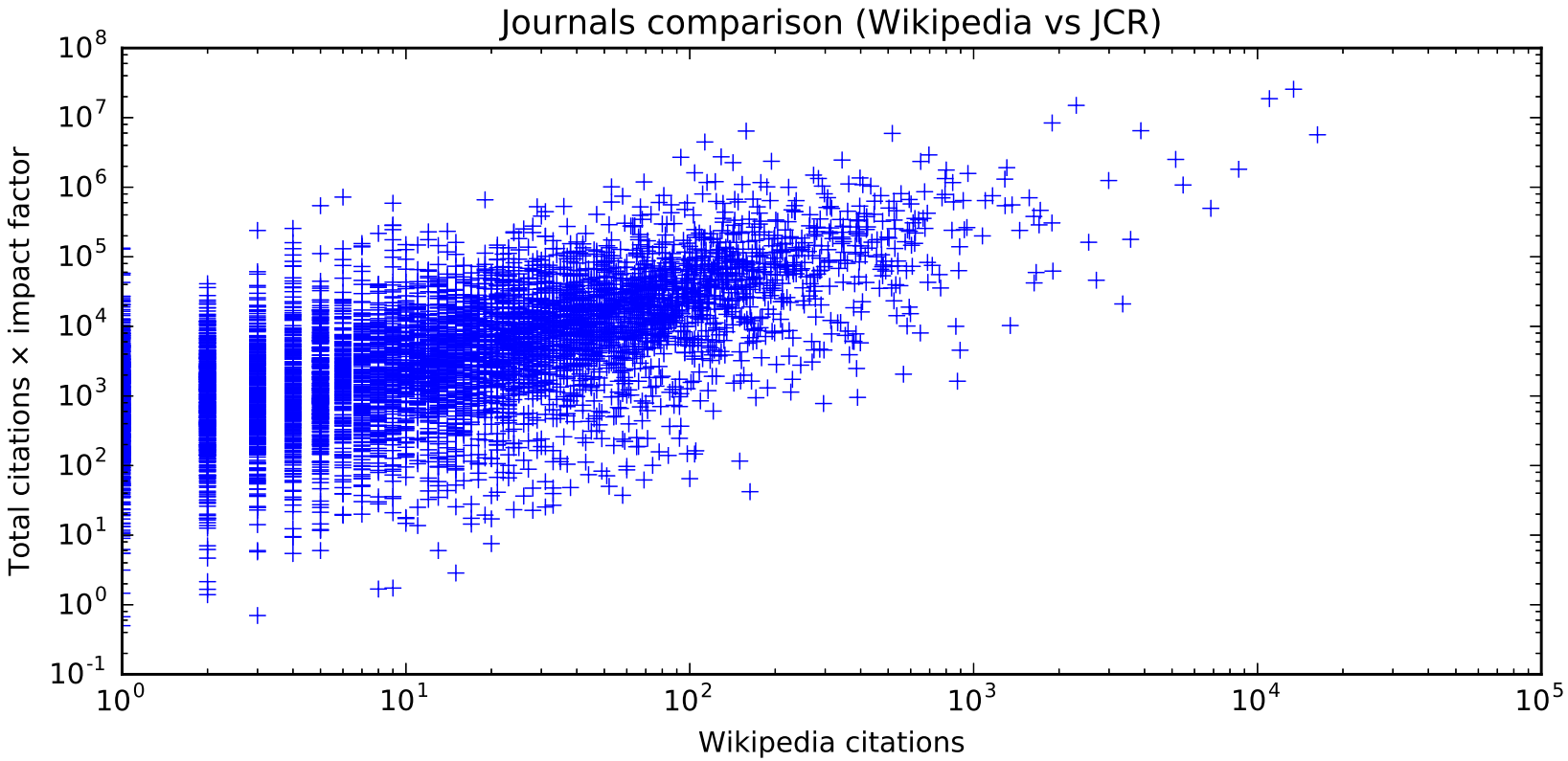
Figure: Papers cited in Wikipedia vs papers in top journals

Journals rank

- First proposed by Nielsen in 2007 [4]
- Most cited journals in Wikipedia are also the most important ones

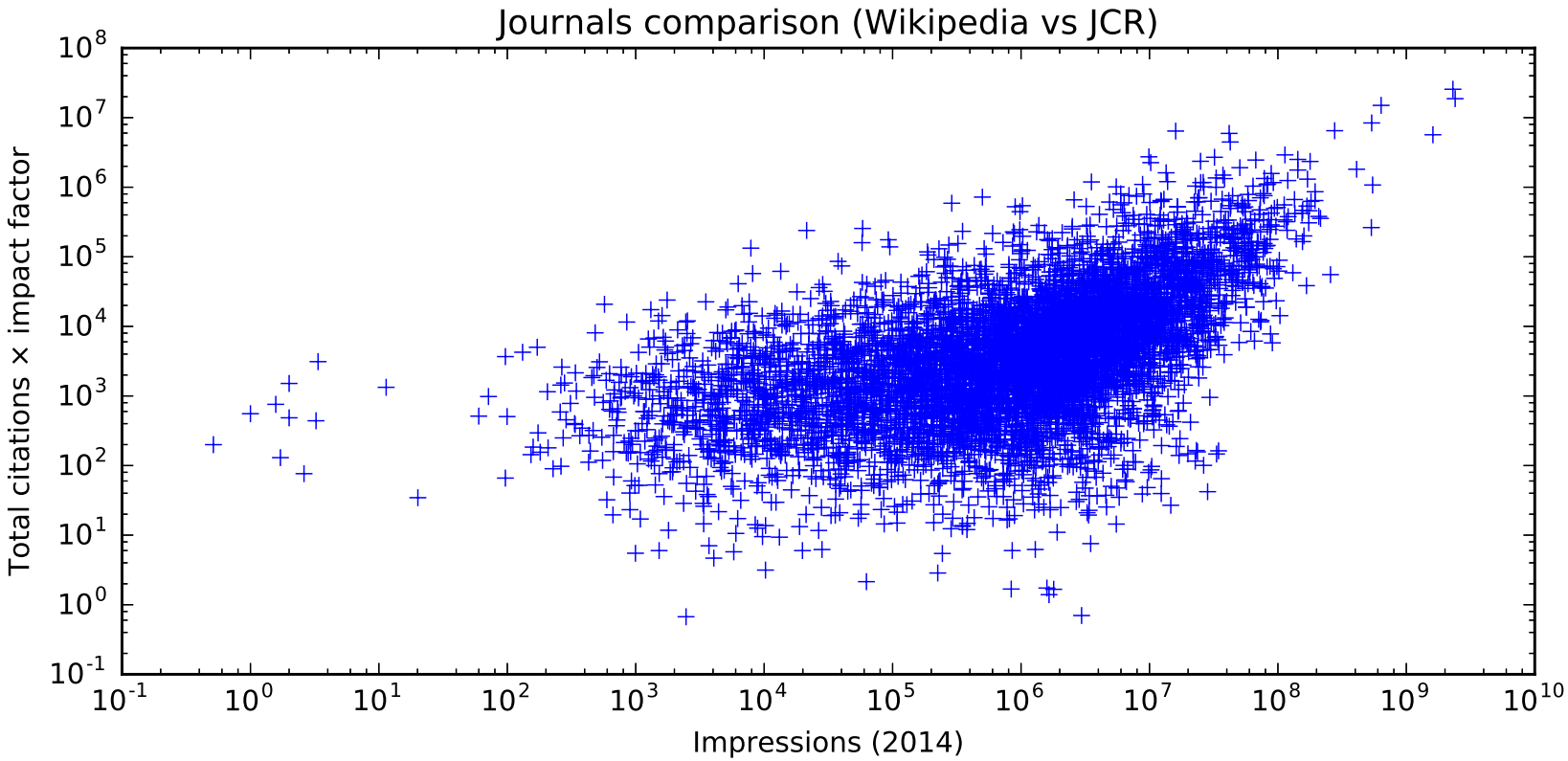
- Journals rank by impact factor (from JCR) versus:
  - citations in Wikipedia
  - visualizations in Wikipedia (in 2014)
- Measured in term of Kendall rank correlation coefficient ($-1 \leq \tau \leq 1$)

Journals impact factor vs Wikipedia citations



- Kendall rank correlation coefficient: 0.464

Journals impact factor vs Wikipedia impressions

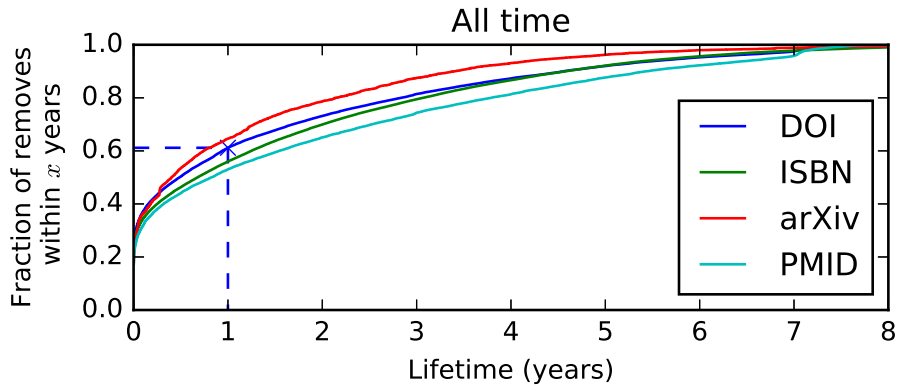


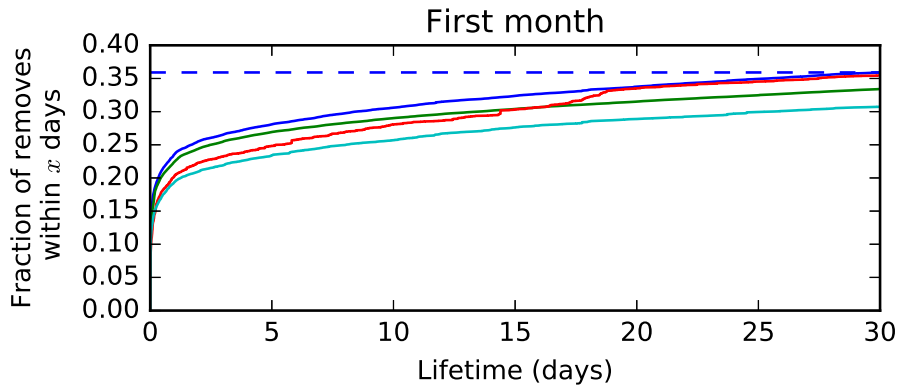
- Kendall rank correlation coefficient: 0.401

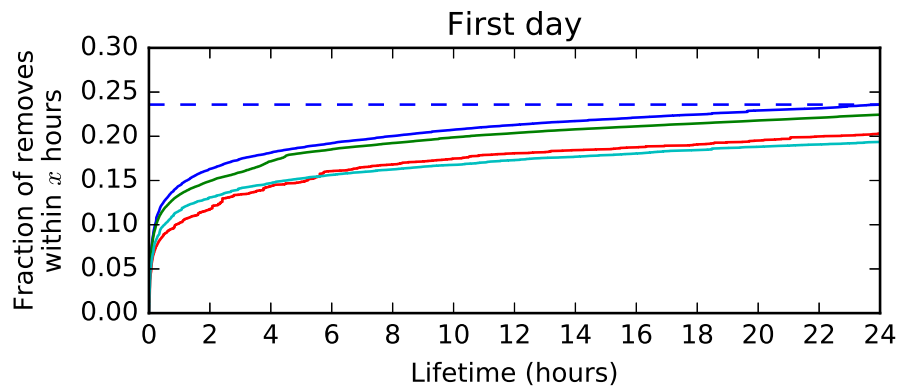
Lifetime of irrelevant citations

- How long does it take for a Wikipedia contributor to discover and remove an "irrelevant" paper from an article?
- An paper is "irrelevant" for an article if it appeared on that page and was then removed.

- Fraction of papers removed in time
- 61% of irrelevant DOI removed within one year
- 36% within one month
- 24% within one day







Further work

- Exploit the page views dataset and the framework
- Open questions:
  - If a paper appears in Wikipedia, will it become more popular in the scientific community?
  - In another way, do researcher use Wikipedia as their primary data source?
  - Predict whether a publication is going to stay on a page

References I


Alessio Bogon, Cristian Consonni, and Alberto Montresor.
Wikipedia pagecounts sorted by page (Year 2014).
February 2016.


David J. McIver and John S. Brownstein.
Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time.
PLoS Computational Biology, 10(4), 2014.


Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis.
Quantifying Wikipedia Usage Patterns Before Stock Market Moves.
Scientific Reports, 3:1–5, 2013.

References II

Finn Arup Nielsen.
Scientific citations in Wikipedia.
First Monday, 12(8):1–5, August 2007.

Sidney Redner.
How popular is your paper? An empirical study of the citation distribution.
The European Physical Journal B - Condensed Matter and Complex Systems, 4(2):131–134, 1998.