Can We Predict Scientific Impact with Social Media?

A Comparison with Traditional Metrics of Scientific Impact

Denis Helic

Knowledge Technologies Institute, Graz University of Technology

March 27, 2014

Question 1

(How) Will Social Media change scientific processes and/or influence scientific impact?

It is already happening...

- Qualitatively, we observe the following change
- Growing numbers of scholars discuss and share the research literature on Twitter, Facebook, etc.
- They organize articles in social reference managers like Mendeley
- Review it in blogs, on reddit, etc.
- The daily research work is moving online and is being put into the spotlight

Spotlight

- Traditionally, the spotlight was always almost exclusively on citations
- It is easy to quantify the scientific impact from citations, citations networks, etc.
- The citation count and derivatives such as h-index, PageRank, etc.
- Often criticized because it can not measure the invisible
- Discussion with colleagues, hallway talk, conference talks, and similar

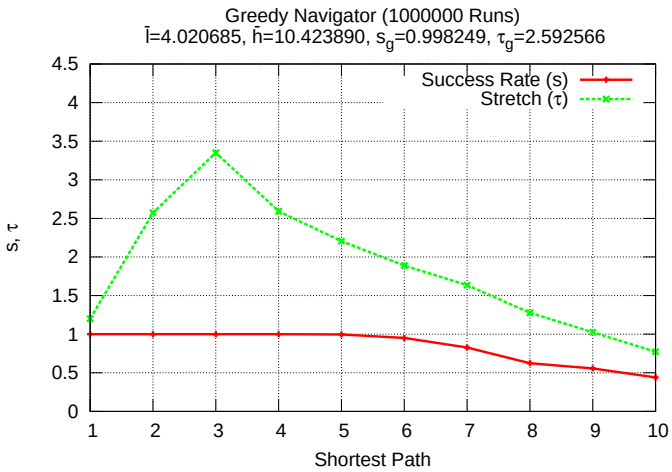
Question 2

Can we quantify the influence of Social Media on scientific processes?

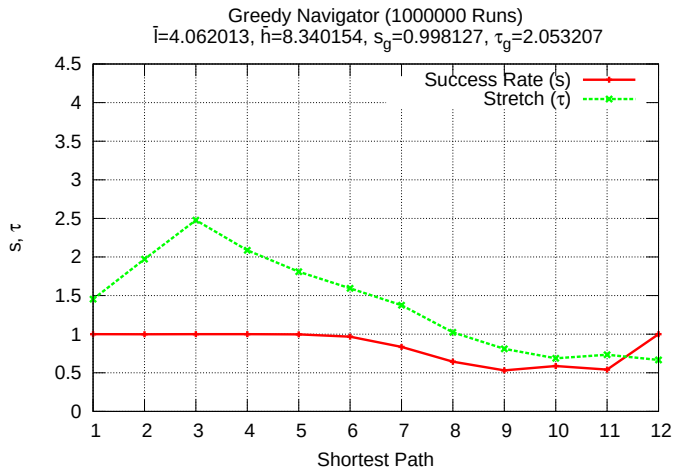
Example 1: Information Retrieval

- Can Social Media improve information retrieval?
- Allow scientists to access relevant articles more efficiently
- Traditionally, digital libraries will have subject catalogs, faceted navigation, or keyword search
- In a study with Mendeley tagging system we analyzed (hierarchical) navigational structures extracted from author keywords and readership tags

Example 1: Information Retrieval



(a) OK, $m = 20$



(b) OT, $m = 20$

Figure: Although the success rates remain excellent over all datasets, stretch increases slightly in keyword datasets. This results in path lengths that are on average longer by 1 or 2 in keyword networks.

Example 1: Information Retrieval

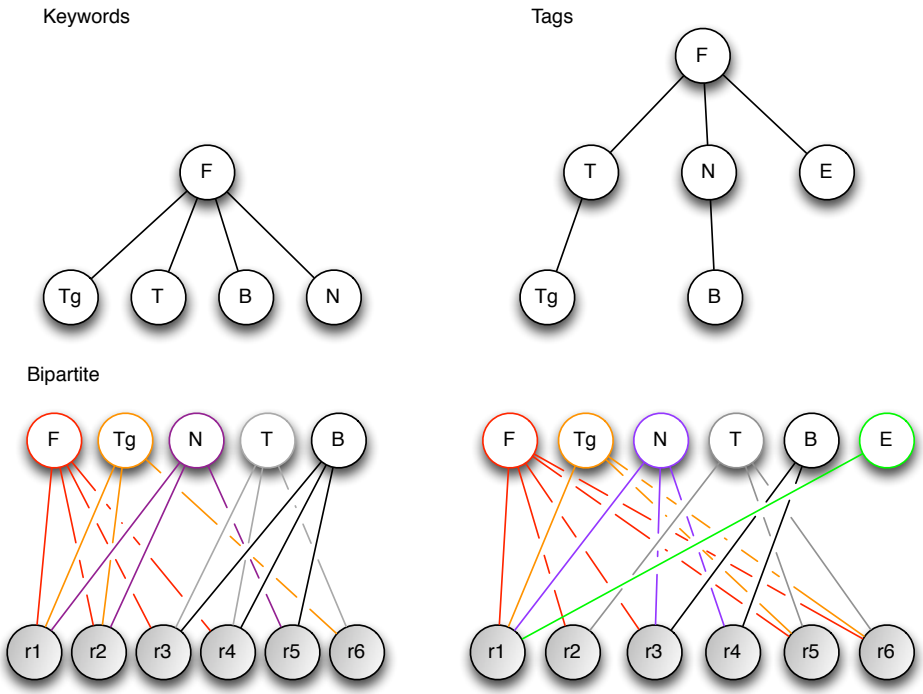


Figure: Keywords (left) and tags (right) with metadata "`folksonomy`" (**F**), "`tagging`" (**Tg**), "`tags`" (**T**), "`navigation`" (**N**), "`browsing`" (**B**), and "`entropy`" (**E**). Tag hierarchies are richer in structure than keyword hierarchies. Structurally richer hierarchies are more stable and robust to the negative effects of the user interface constraints.

Example 1: Information Retrieval

- Folksonomies and keyword hierarchies exhibit comparable quantitative properties
- We find interesting qualitative differences with regard to navigation
- Folksonomies create more efficient navigational structures
- They enable users to find target resources with fewer hops
- Reason: greater overlap between tags provides better options for users to switch between different parts of the network

Example 2: Citation Latency

- How early availability for accessing an article influences the citation latency?
- Citation latency: the time that it takes from the moment an article is accepted for publication until it is cited in other (published) articles
- Depending on the community, the process, the accessibility of the journal this may range anywhere from 3 months to 1-2 years
- Is the latency reduced by e.g. pre-print platforms
- `http://arxiv.org/` at Cornell

Paper

Tim Brody, Stevan Harnad, and Leslie Carr. 2006. Earlier Web usage statistics as predictors of later citation impact: Research Articles.

Example 2: Citation Latency

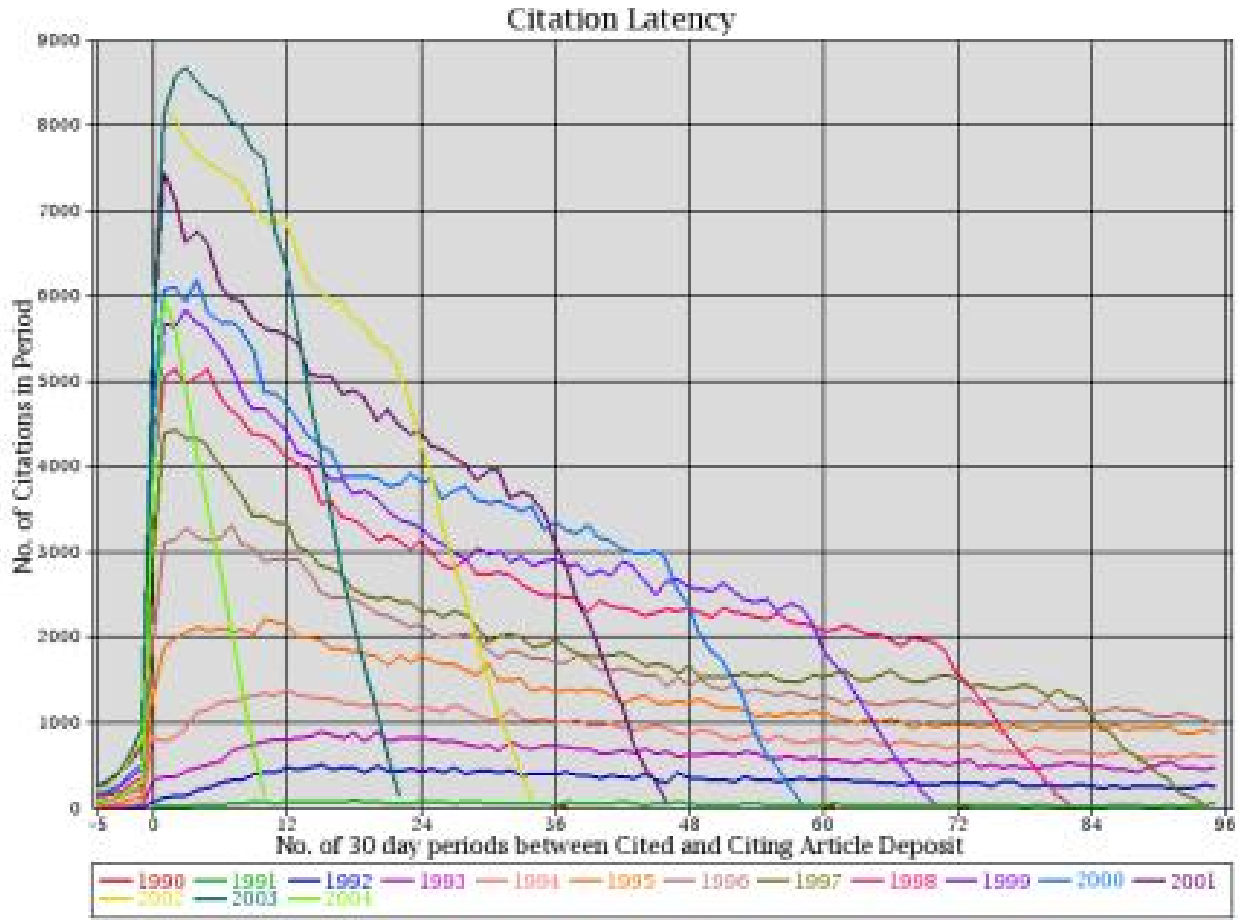


Figure: Changing distribution of latencies, e.g. for older articles the latency was approx. 12 months or more. Recently, latency decreased to seemingly nothing.

Example 2: Citation Latency

- The latency between an article being uploaded and later cited has reduced
- From a peak at 12 months to no or small delay at all to the peak rate of citations
- This can be biased because of the possibility to revise the paper
- However, it indicates that the authors are increasingly citing very recent work that has yet to be published
- Even new questions for the peer-review process?

Example 3: Download vs. citation vs. readership

- How downloads or an article compare to the number of citations that article obtains
- How readership data compares to the number of citations
- Readership data is e.g. a number of mentions in Mendeley user libraries
- How downloads and readership compare
- A study with Mendeley and Know-Center

Paper

Schlögl et al., Download vs. citation vs. readership data: the case of an information systems journal

Example 3: Download vs. citation vs. readership

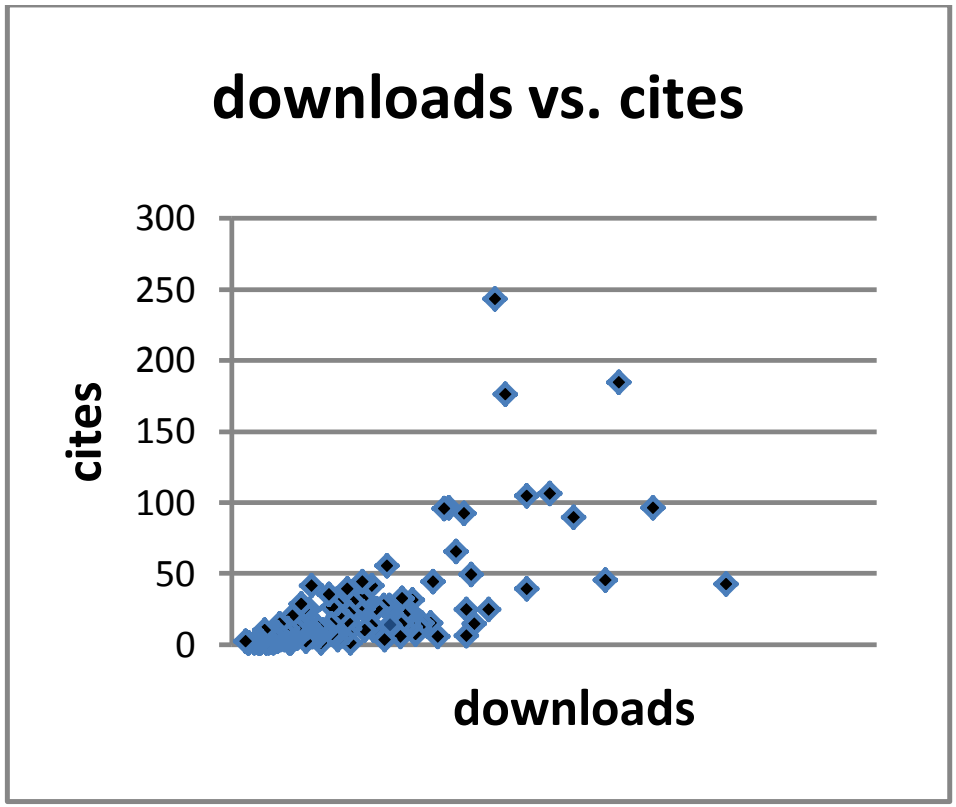


Figure: Spearman correlation r=0.77

Example 3: Download vs. citation vs. readership

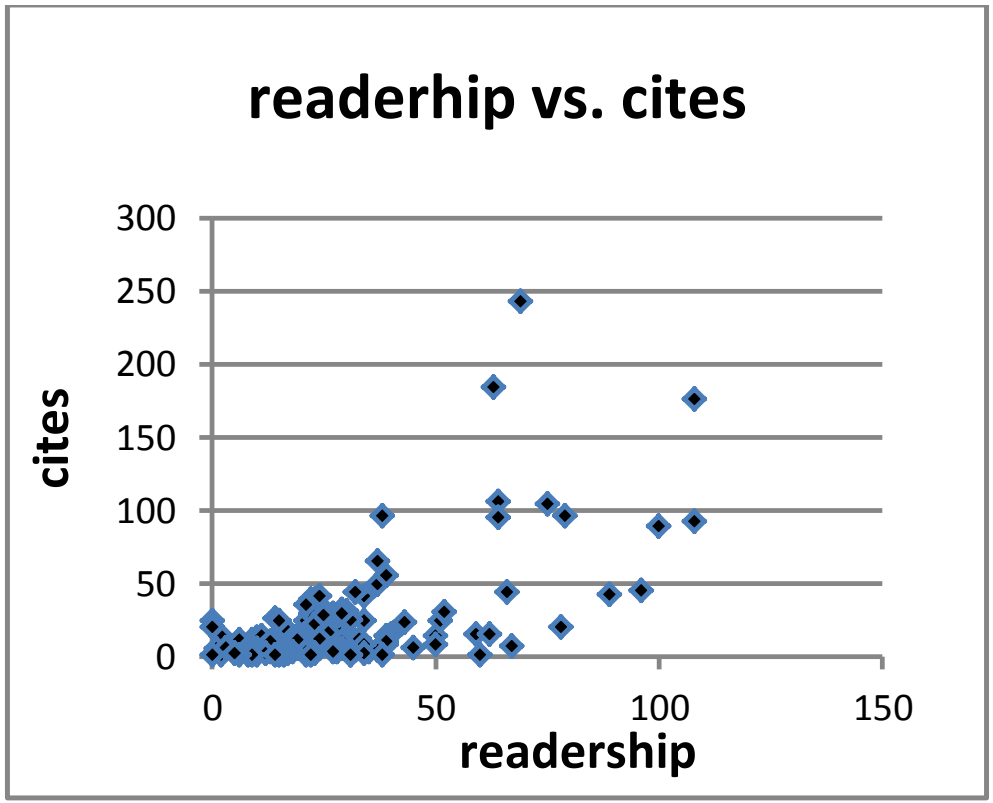


Figure: Spearman correlation r=0.51

Example 3: Download vs. citation vs. readership

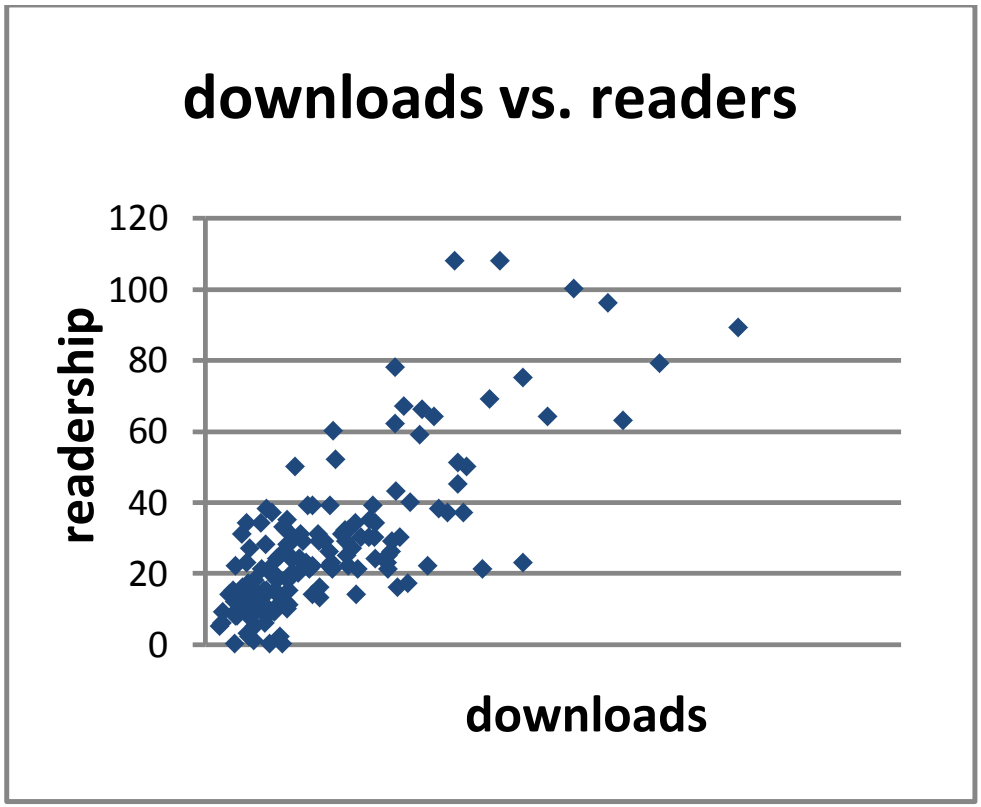


Figure: Spearman correlation r=0.73

Example 3: Download vs. citation vs. readership

- The results are in line with several other similar studies
- Correlations do however change depending on the source of citations
- Also depending on the journal or conference, scientific field, etc.
- Strongly time dependent
- Somewhat smaller correlation between readership and citations
- Mendeley a new young system?
- Mendeley user population?

Question 3

How should we quantify the influence of Social Media on scientific processes?

Methodology

- In all examples we measured a different thing and applied a different methodology
- Example 1: algorithmic approach to information retrieval
- Example 2: distribution of citation latency
- Example 3: non-parametric statistics with rank correlations
- Should we also apply other methods?

Time dependence

- Traditional as well as new metrics are strongly time dependent
- E.g. citation delay, time of the peak, etc.
- Downloads are strongly time dependent as a different function of time
- Social Media is even more sensitive to time and shorter time spans

Example 4: Response dynamics

- Now, we can include Social Media in the loop
- Ask questions such as what is the download latency for pre-prints
- How does Twitter influence the download latency?
- How does Twitter influence the citation count?
- A study with `http://arxiv.org/`

Paper

Shuai et al., How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations

Example 4: Response dynamics

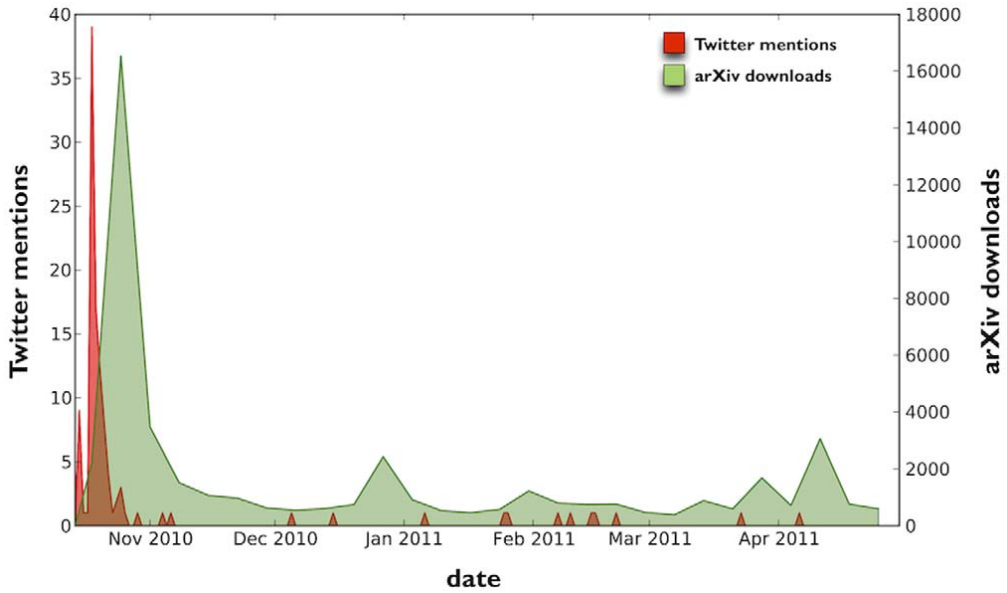


Figure: Twitter mentions spike shortly after submission and wane quickly, whereas downloads peak shortly afterwards but continue to exhibit significant activity many weeks later.

Example 4: Response dynamics

- Thus, we need an even more sophisticated methodology than simple correlation measurements
- Counting twitter mentions, downloads, and citations at different times can lead to varying correlations
- Time series analysis
- Multivariate regression methods, etc.
- Methodologically, a very interesting field!

Example 4: Response dynamics

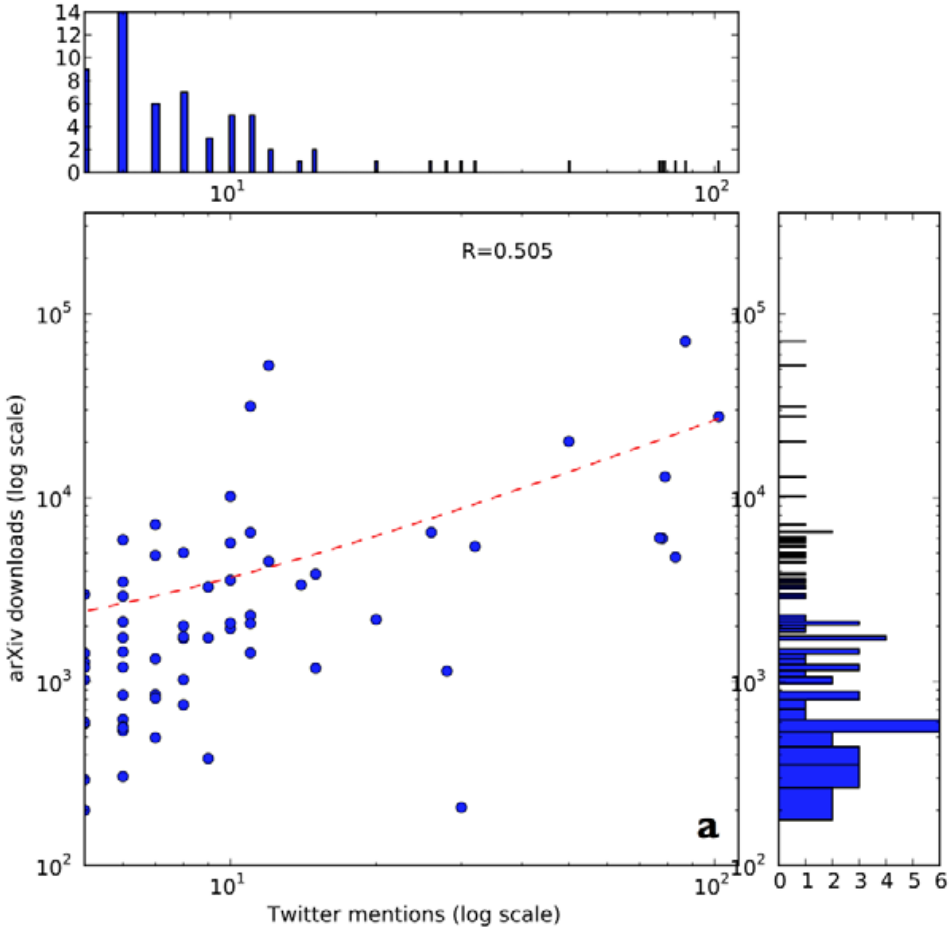


Figure: Pearson correlation R for 70 most mentioned articles

Example 4: Response dynamics

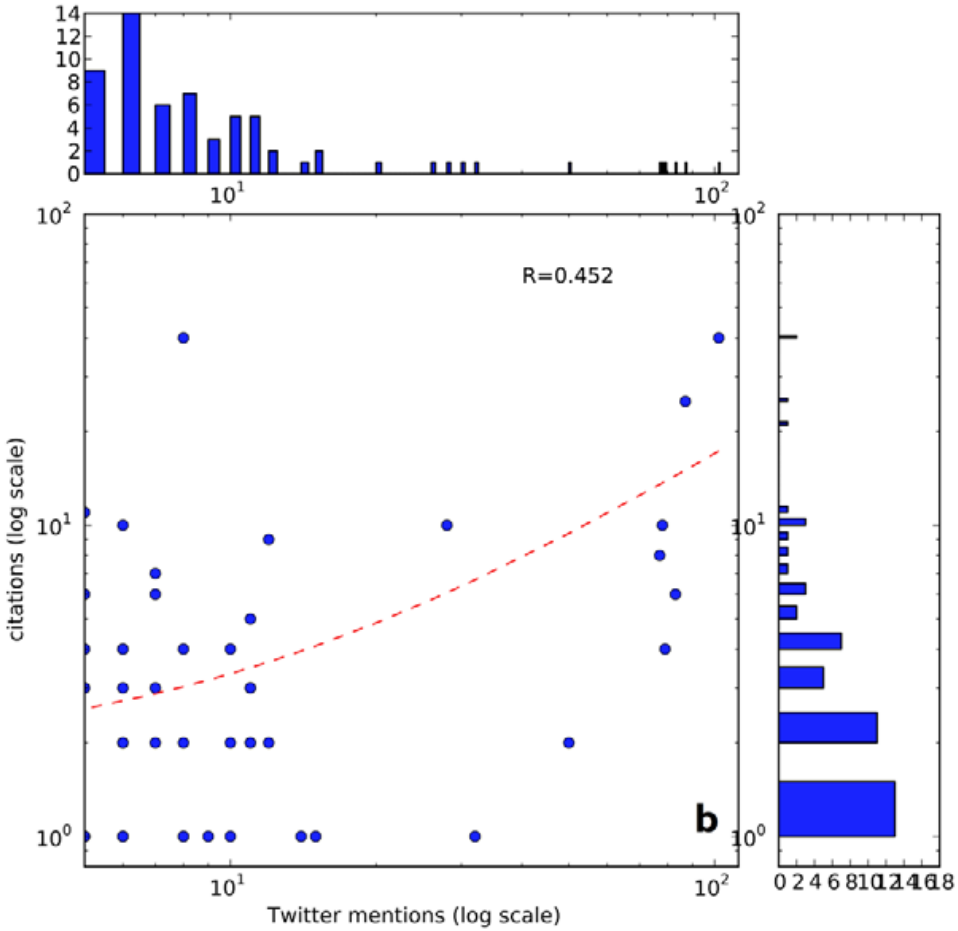


Figure: Pearson correlation R for 70 most mentioned articles

Example 4: Response dynamics

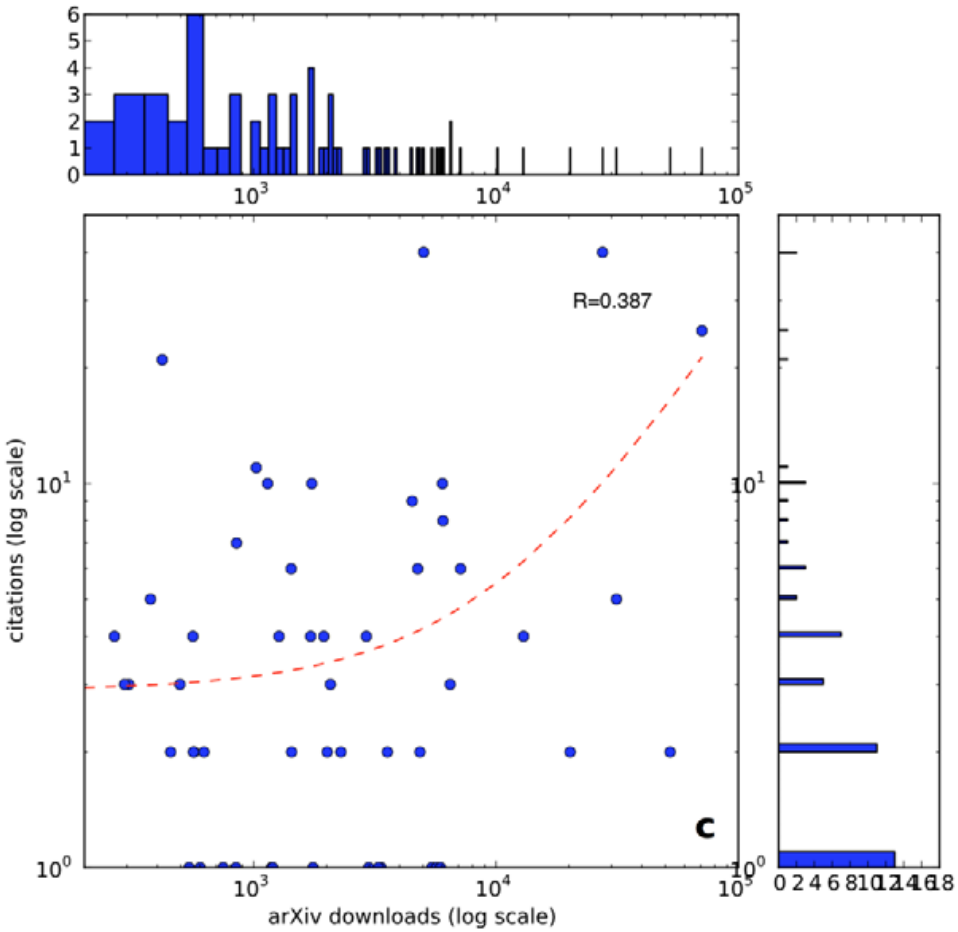


Figure: Pearson correlation R for 70 most mentioned articles

Example 4: Response dynamics

- The results are highly suggestive of a strong tie between social media interest, article downloads and even early citations
- There are two different temporal patterns of activity
- The volume of twitter mentions is statistically correlated with that of both downloads and early citations
- Two possible explanations: through exposition to twitter download and citation behavior is affected
- Second: intrinsic quality

Interpretation

- Causality vs. correlation
- A correlation between x and y may occur because:
  1. x influences y
  2. y influences x
  3. the influence is in both direction
  4. a third variable influences both x and y
- We need more analysis and interpretation

Interpretation and interdisciplinary approach

- Computer scientists are good at calculating things
- But we lack knowledge in user behavior
- We lack knowledge in community practices
- Only interdisciplinary teams can interpret the results in a satisfactory manner

Question 4

Can we move beyond quantification to modeling, predicting and understanding?

Hypotheses

- After observing, measuring and quantifying
- Can we formulate hypotheses which can be tested?
- Can such hypotheses explain the phenomena that we observe?
- Can we use these models to predict new phenomena, e.g. the scientific impact of an article?

Example 5: Long-term predictability

- Is there a long-term predictability in citation patterns?
- Are there universal laws governing citation process across the fields, authors, and journals
- What are the parameters of such a universal model?
- How does the model capture phenomena such as Social Media?

Paper

Wang et al., Quantifying Long-Term Scientific Impact

Example 5: Long-term predictability

- Extremely difficult because of impact heterogeneity
- E.g. power laws in the impact, citation, download, or twitter mentions distributions

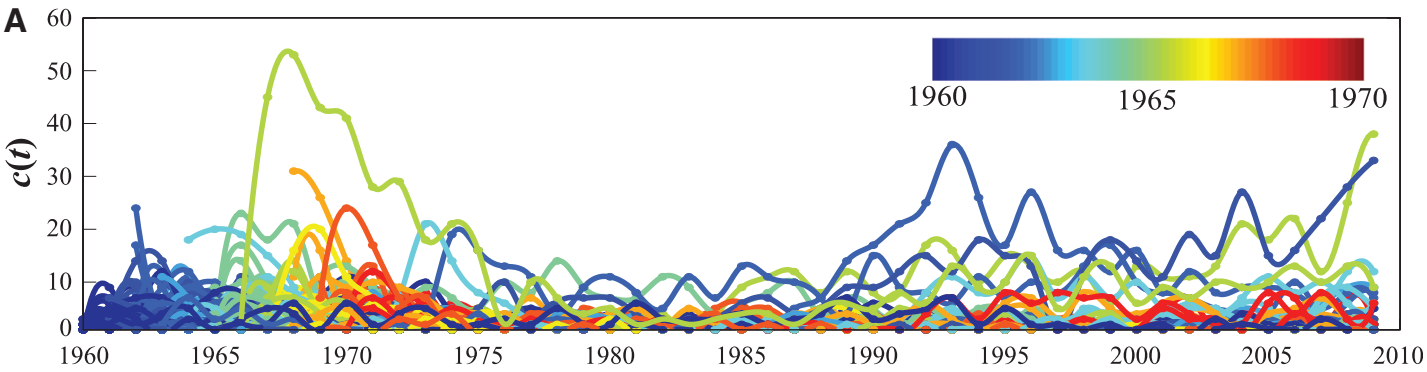


Figure: Yearly citations of randomly selected articles from the Physical Review

Example 5: Long-term predictability

- Three basic mechanisms that drive the citation history of individual papers:
  1. Preferential attachment (rich-get-richer phenomenon)
  2. New ideas are integrated in subsequent work: immediacy governs the time to citation peak and longevity captures the decay rate
  3. Fitness captures the intrinsic quality of a paper
- Novelty and fitness depend on the community response to the work, i.e. they capture also any influence coming from e.g. Social Media

Example 5: Long-term predictability

- Analytic solution of the model shows that the shape of citation distribution depends on immediacy, longevity and fitness
- But the long-term asymptotic behavior depends only on fitness
- After long time citation distributions of all papers with the same fitness converge regardless of their immediacy and longevity
- In other words, only the intrinsic paper quality as perceived in a particular community matters in the long run
- Empirical analysis also confirms these theoretical results

Example 5: Long-term predictability

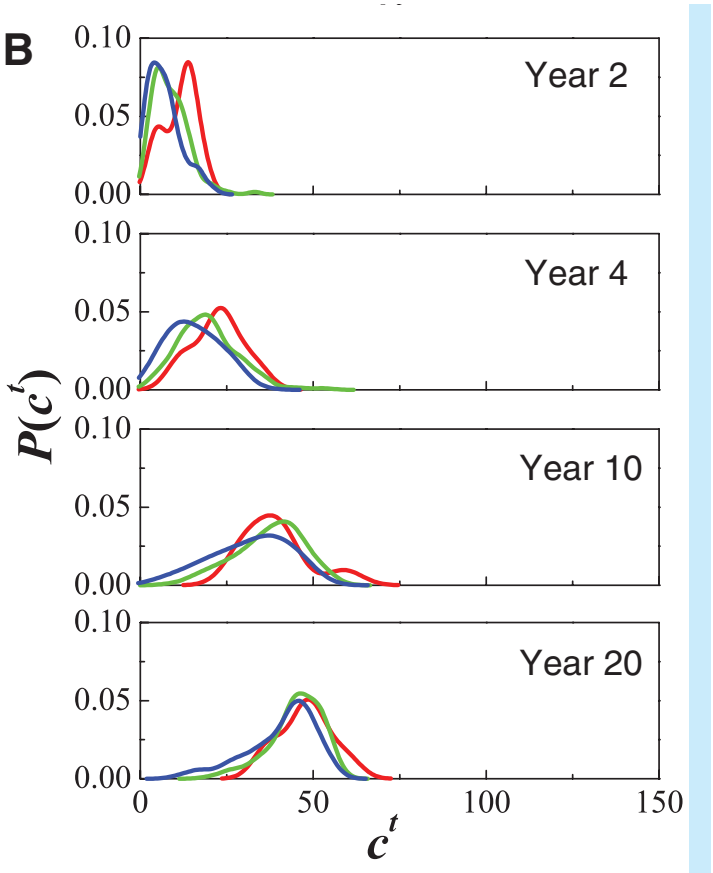


Figure: Convergence of citation distributions

Summary

- Social Media influences the scientific process
- We can quantify that impact in various ways
- We need sound methodologies for the quantification
- We need interdisciplinary research for interpretation of the results
- Recent results indicate that the intrinsic quality of a paper is the only indicator of its long-term impact
- Short term impact can be influenced by Social Media

Thank You!

Questions?