Science is Shaped by Wikipedia

Evidence From a Randomized Control Trial

DOUGLAS HANLEY (PITT)

NEIL THOMPSON (MIT/SLOAN)

EIEF 2018

Wikipedia

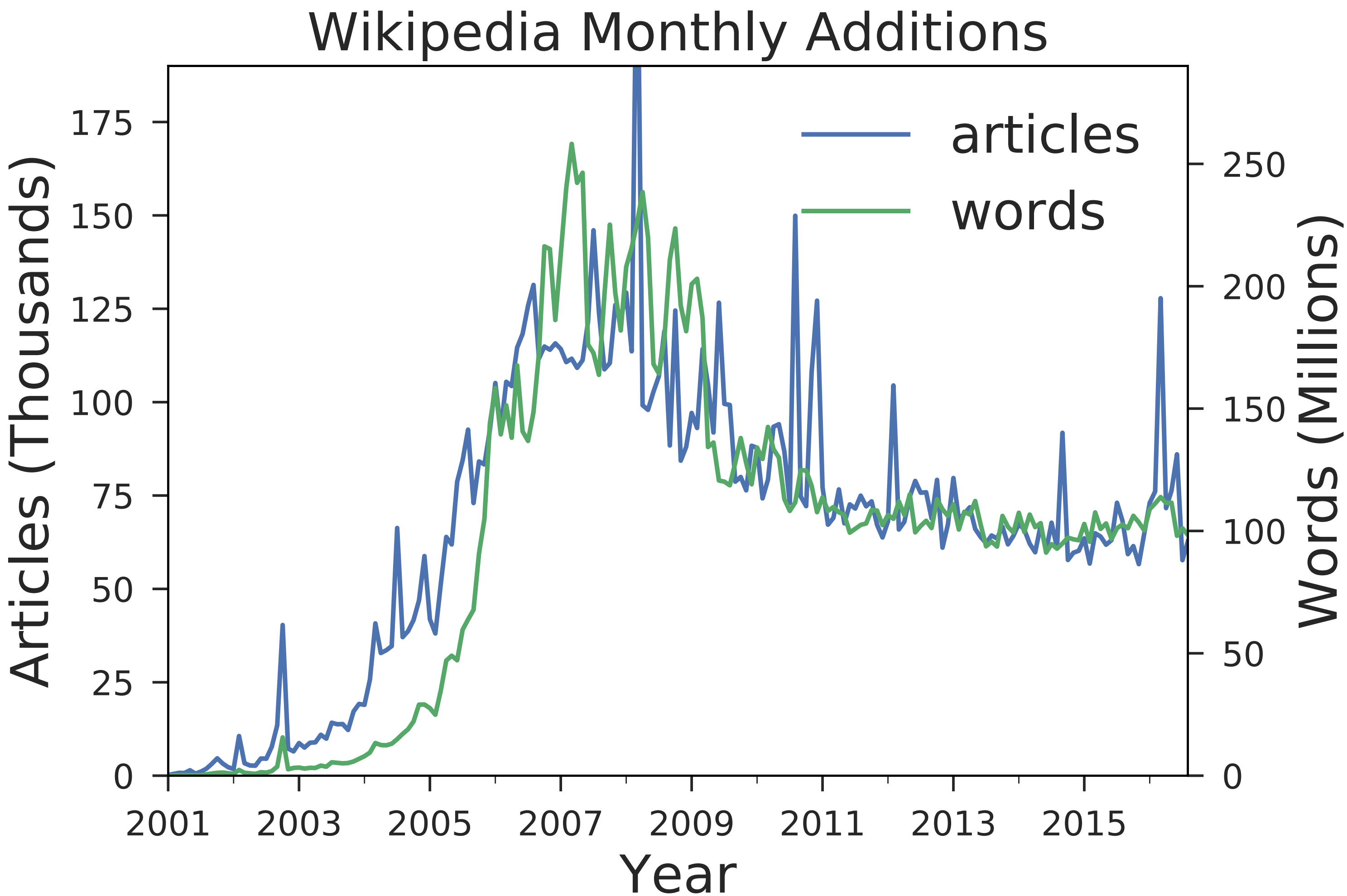
Since 2001, we've seen the creation of a free, searchable, crowd-sourced, online encyclopedia

Wikipedia is the 5th most-visited website on the Internet

Aggregate statistics

- 13 million articles
- 18 billion page views/month
- 500 million unique visitors/month

Rise of Wikipedia



Collaboration

Wikipedia was an early player in the open collaboration

- open source software (GNU/Linux, etc)
- GitHub
- StackExchange
- Quora
- Polymath*

Open both in the acceptance of contributions and in the dissemination of results

- stark difference from old coprorate lab model and academic lab model

Motivation

A large proliferation of *theories* of knowledge diffusion recently

- Lucas and Moll (2014)
- Perla and Tonetti (2014)
- Benhabib, Perla, and Tonetti (2014)
- Buera and Oberfield (2017)

Little direct evidence on diffusion and its mechanisms

- We want to undertand Wikipedia but also use it as a tool to understand diffusion

Agenda

Question: What is the effect of Wikipedia on scientific progress and on economic growth?

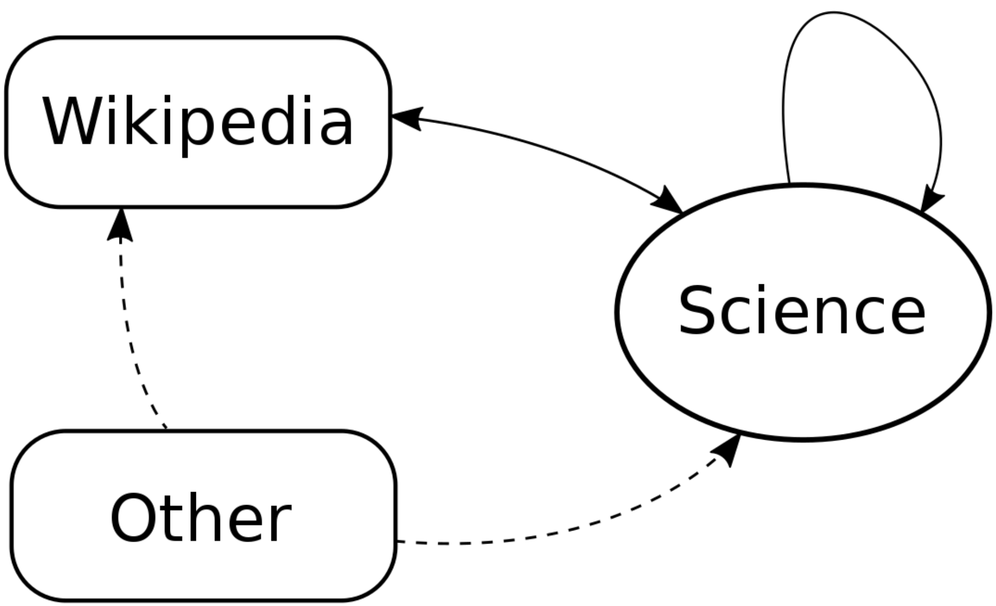
We take both an observational and experimental approach to answering this question.

Today's talk:

- Observational results from editing history
- Description of experimental approach
- Estimation of causal impact of Wikipedia
- Discussion of possible mechanisms

Causality

We are interested in the direct causal effect of Wikipedia, but with observational data we can frame the magnitude of the effect



Observation

The rise of Wikipedia could have various effects

- diffusion of frontier research or reviving old ideas
- diffusion between different fields or subfields
- diffusion between people or countries

Relationship with science is bidirectional

- science → Wikipedia: can be observed through citations
- Wikipedia → science: measure with document similarities

Datasets

Full editing history of all **Wikipedia** articles (353 million edits, 20 TB)

- username of editor and date/time of edit
- full article text after each edit
- user-generated category structure
- daily page views since 2007

Metadata and text of all **Elsevier** journal articles (2,061 journals)

- authors and publication date
- journal, volume, issue number
- full article text

Document Statistics

We look at all Wikipedia articles and scientific articles after 2000, with special focus on the field of chemistry.

	Wikipedia Total	Wikipedia Chemistry	Science Total	Science Chemistry
Journals	-	-	2K	50
Issues	-	-	~1M	19K
Articles	13M	150K	8.5M	290K
Words	18B	1B	~3B	636M

Average chemistry publication lag: **8 weeks**

Wikipedia editing

Initial "Stub"

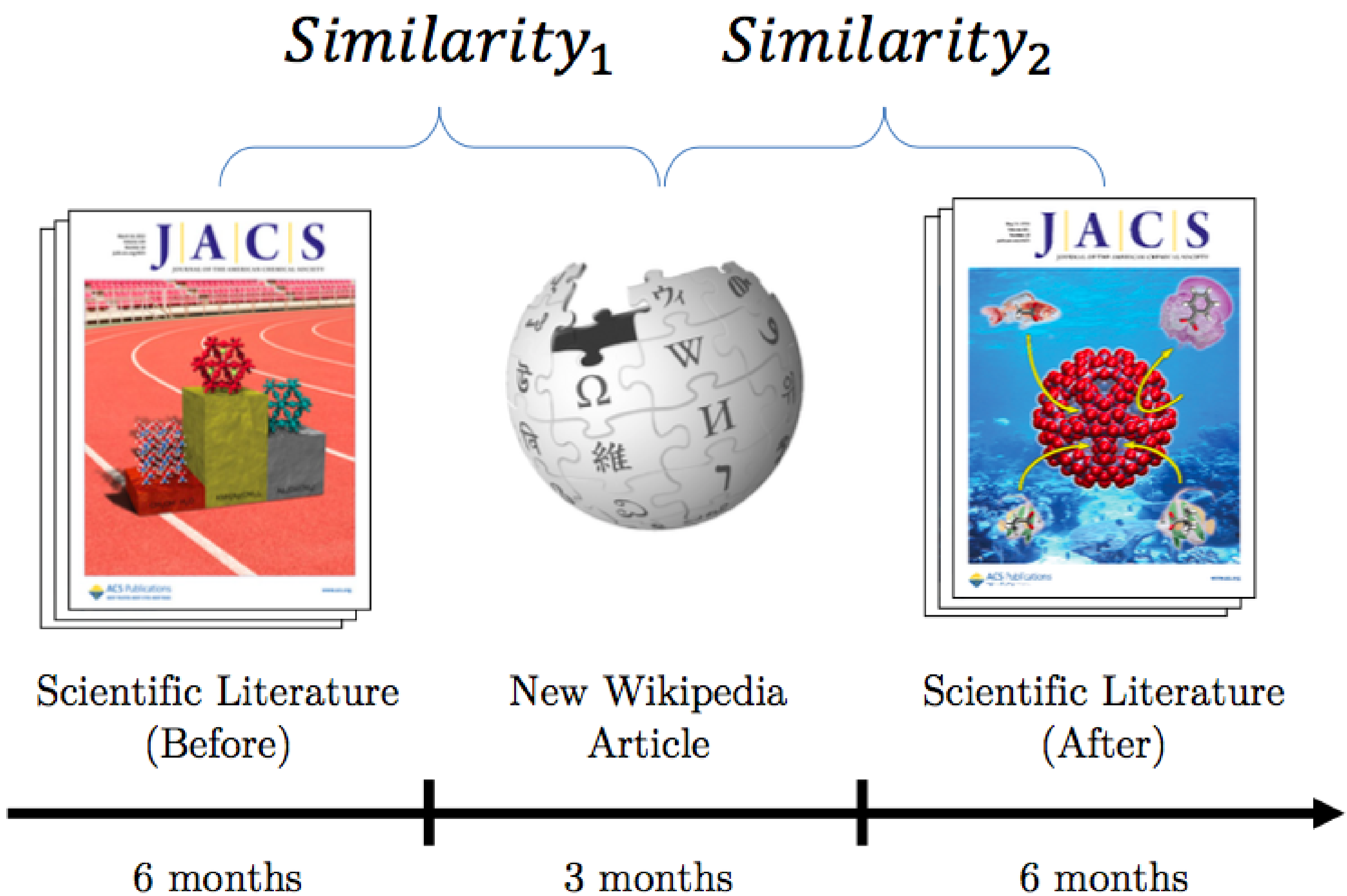
"Magnesium sulfate," $MgSO_4$, (commonly known as Epsom salts) is used as a therapeutic bath.

Edits

"Magnesium sulfate~~,~~" ~~$MgSO_4$,~~ (commonly ~~known as~~ called "Epsom ~~salts~~ salt" in hydrated form) is ~~used as a therapeutic bath~~ a chemical compound with formula $MgSO_4$.

Epsom salt was originally prepared by boiling down mineral waters at Epsom, England and afterwards prepared from sea water. In more recent times, these salts are obtained from certain minerals such as siliceous hydrate of magnesia.

Wikipedia Effect



Document Similarity

How can we quantify the similarity between two documents? We use the **cosine similarity metric**

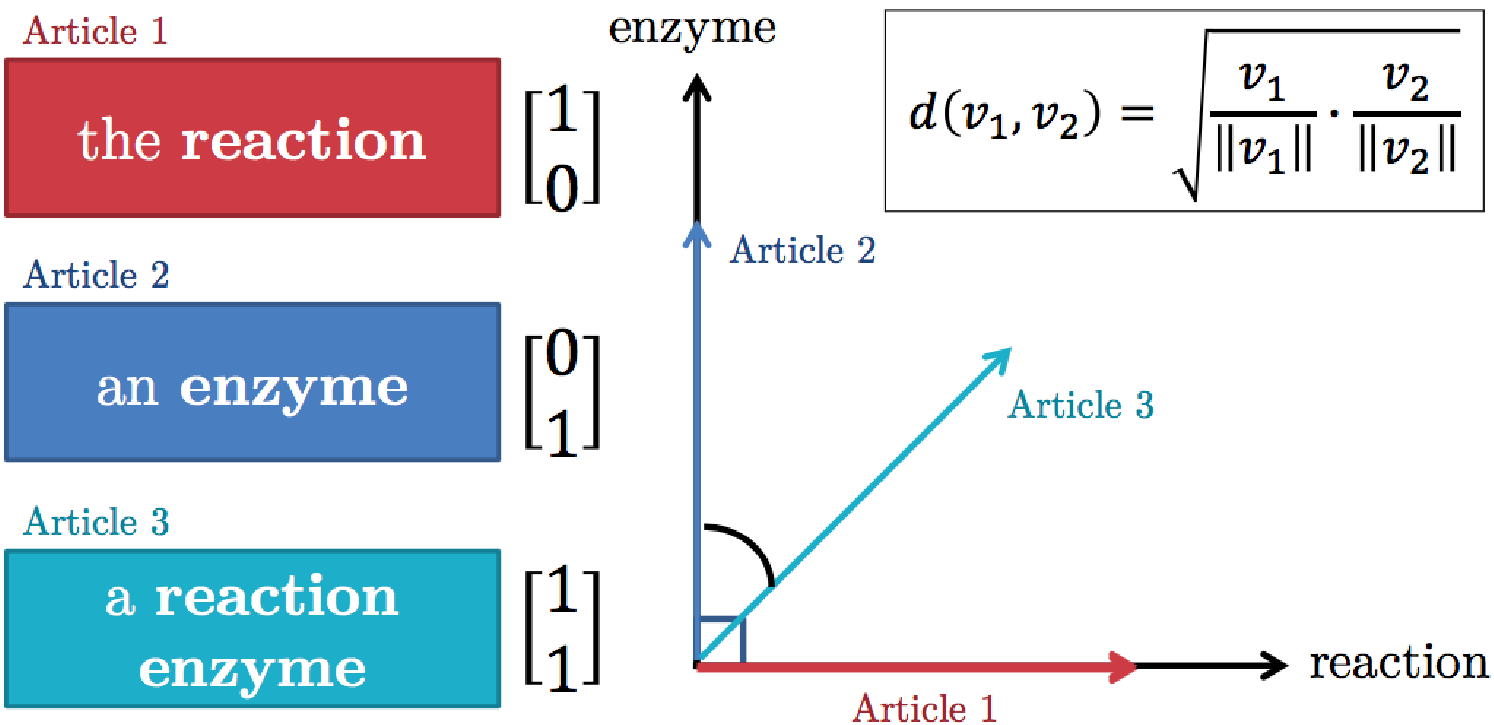
Convert each document into a word frequency vector. If there are N possible words, each document is an N dimensional vector

The similarity between two documents is simply their normalized vector product

$$d(v_1, v_2) = \sqrt{\frac{v_1 \cdot v_2}{||v_1||\ ||v_2||}} \in [0, 1]$$

where $||v|| = \sqrt{v \cdot v}$

Cosine Distance



$$\begin{bmatrix}1\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\end{bmatrix} \quad \begin{bmatrix}1\\1\end{bmatrix}$$

$$d(v_1, v_2) = \sqrt{\frac{v_1}{\|v_1\|} \cdot \frac{v_2}{\|v_2\|}}$$

Metric Refinements

What about very common words like "the" or "and"? We downweight words by the fraction of documents that they appear in (TF-IDF)

$$w_i = \log\left(\frac{D+1}{d_i+1}\right) + 1 \geq 1$$

Errors and misspellings are unavoidable for such a large dataset. We ignore words that appear in fewer than 5 documents in total

- Otherwise mispellings would be highly weighted

Two Approaches

Naive observational

- Look at organic creation of articles over entire history
- Will confound causal effect and selection on "hot" topics

Experimental

- Publish our own Wikipedia articles
- Yields causal effect of "marginal" Wikipedia articles

Observational

Identify all chemistry articles in Wikipedia and extract their text 3 months after they are "born"

- delay is due to fact that many articles start as tiny "stubs"

Focus on top 50 ranked chemistry journals in our Elsevier sample

What is the "effect" of the introdution of a Wikipedia article?

Fields and Categories

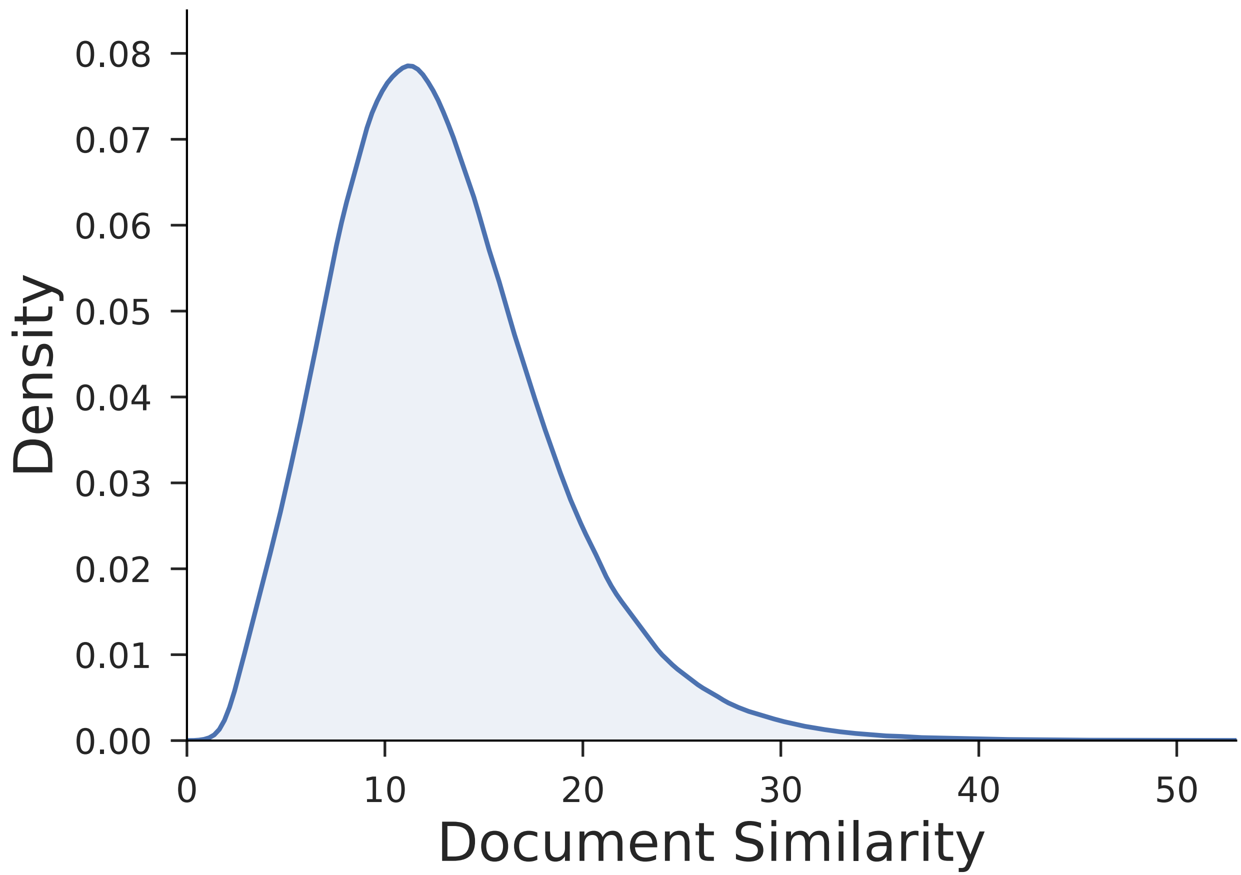
What constitutes a chemistry article? We use the built-in category hierarchy of Wikipedia.

Chemistry → Organic chemistry → Organic compounds → Aromatic compounds → Graphene

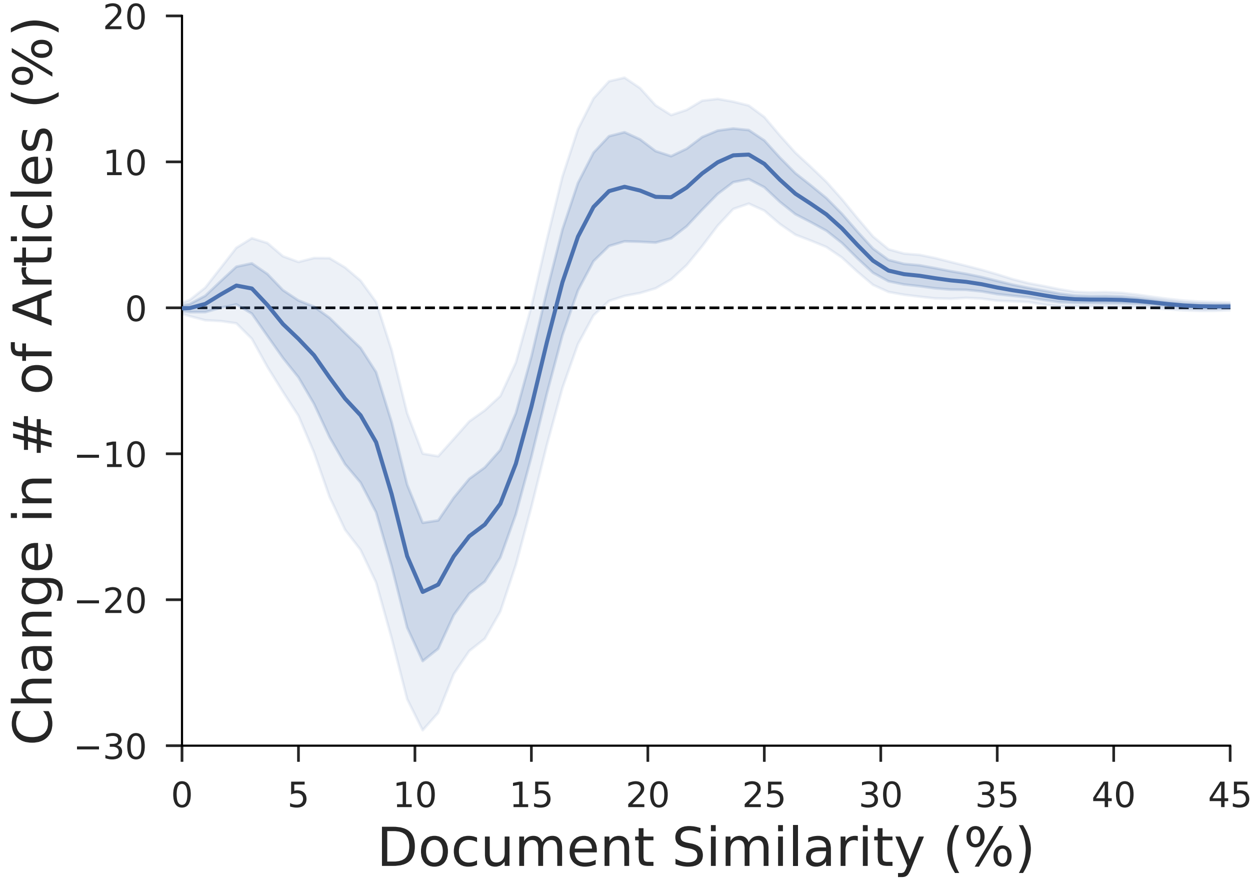
Additional filtration required

- pare down category tree with PageRank (cycles)
- train SVM classifier on hand-categorized pages

Observational similarity



Distributional changes



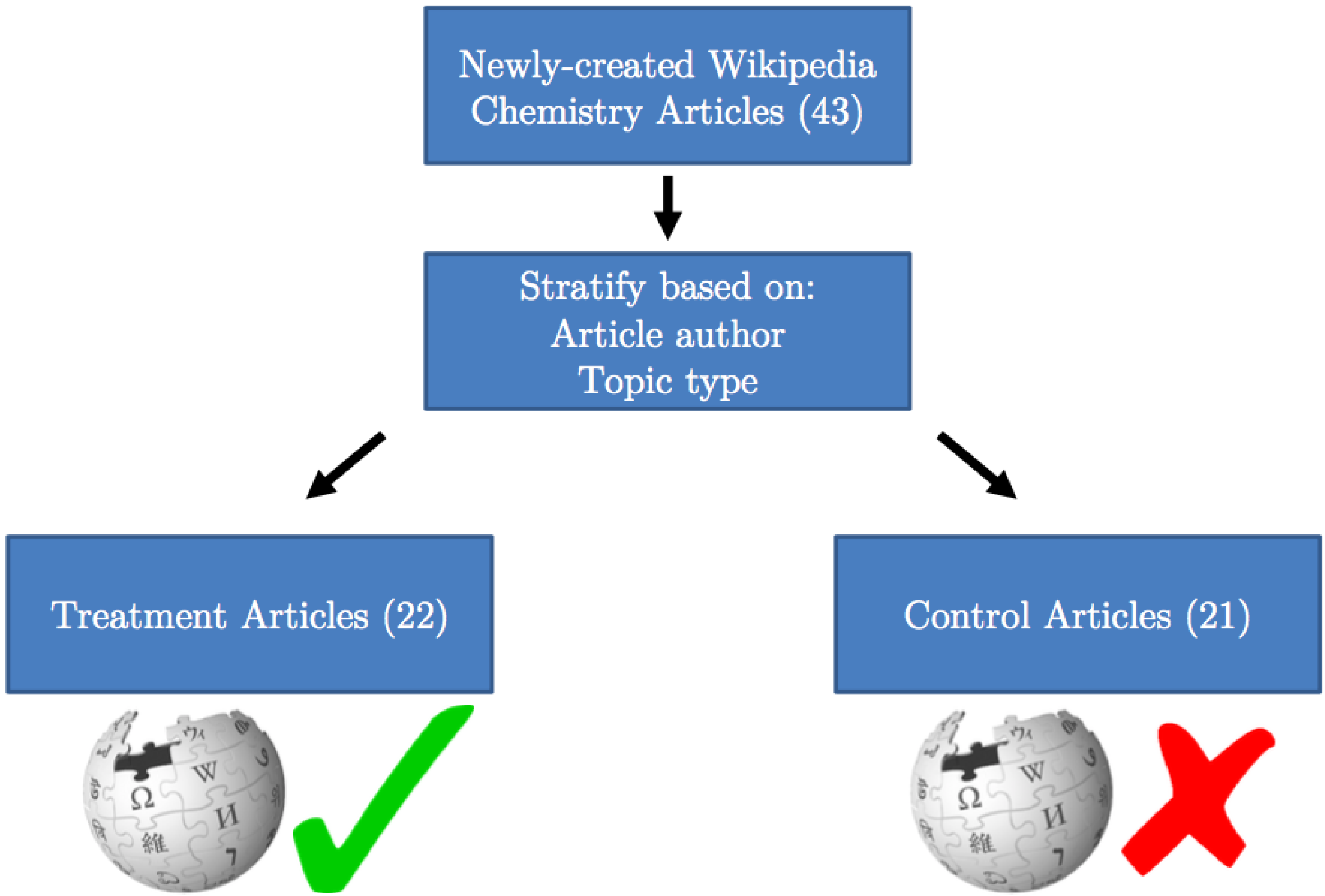
Experiment

To get issues of causality, we take a randomized control trial (RCT) approach

Identify 43 topics in chemistry that could use Wikipedia entries, but don't currently have them (look at graduate syllabi)

Contract out writing of summaries on these topics to chemistry grad students and publish a random subsample of them

Experimental Design



Sample Treatment Article

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Elimination reaction of free radicals

From Wikipedia, the free encyclopedia

Free radicals can undergo elimination reactions to form olefins, a reaction known as elimination reaction of free radicals.[1] Such reactions are usually not major pathways for radical mediated reactions.

Radicals can undergo disproportionation reaction through radical elimination mechanism (Figure "Radical dis... er sa...

Ra... mple, wh... ted via rad... me... nilar typ... de...

Average views per article *per month*: 4,400 (!)

Radical elimination reactions are found in enzyme-catalyzed pathways. In the dehydrogenation reaction of acyl-CoA to form enoyl-CoA, FAD accepts two protons and two electrons to form FADH2 under the catalysis of acyl-CoA dehydrogenase.[3] The mechanism involves formation of acyl-CoA β-radical that undergo elimination to form the enoyl-CoA product (Figure "Radical elimination reaction in acyl-CoA dehydrogenase-catalyzed reaction").

Radical disproportionation via radical elimination mechanism

Depolymerization of polystyrene via radical elimination mechanism

Radical elimination reaction in acyl-CoA dehydrogenase-catalyzed reaction.

References [edit]

1. ^ Ansylen, E. V.; Dougherty, D. A. Modern Physical Organic Chemistry. p. 586, ISBN 978-1-891389-31-3
2. ^ Grassie, N.; Kerr, W. W. Trans. Faraday Soc., 1957, 53, 234-239
3. ^ Thorpe, C.; Kim, J. J.; FASEB J., 1995, 9, 718-725

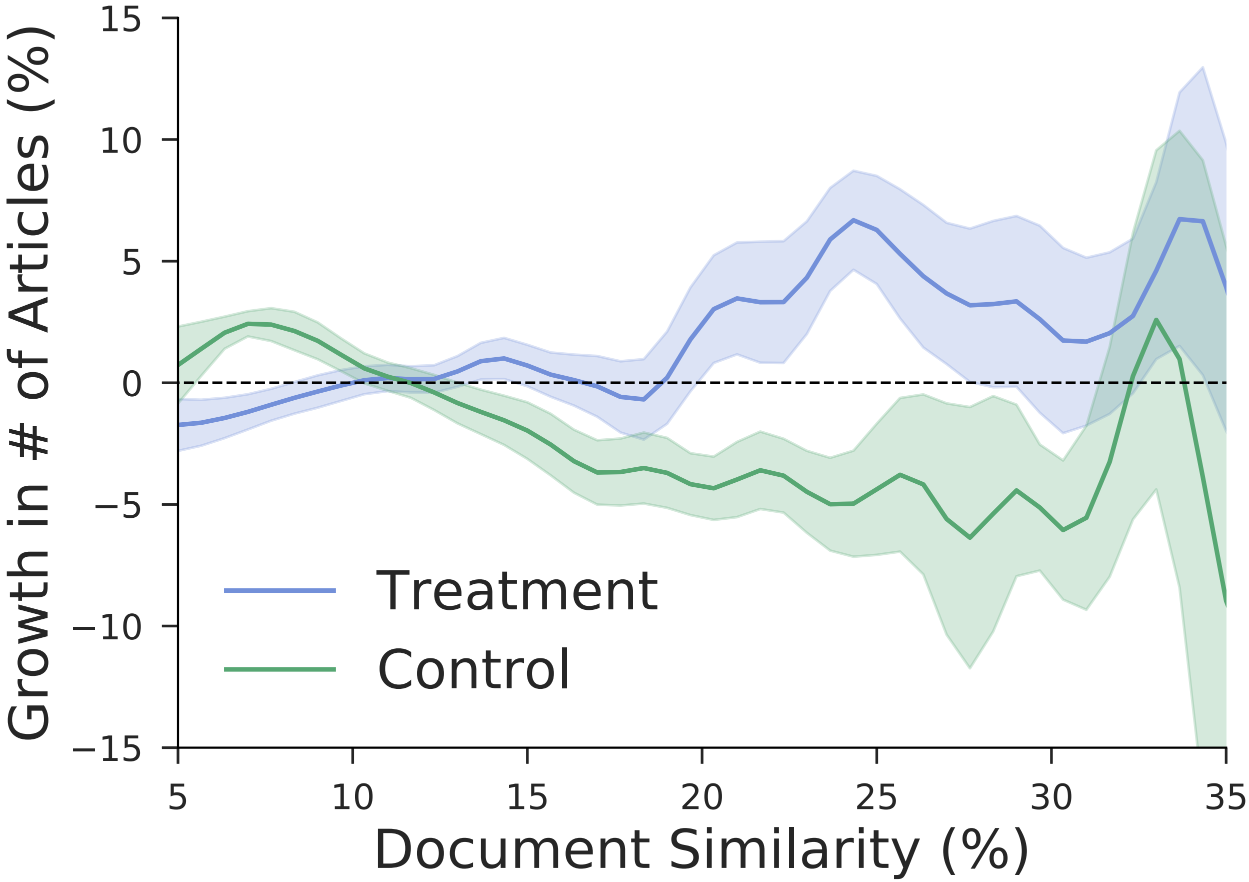
Experimental Balance

	Treatment (mean)	Control (mean)	T-test (p-value)	KS-test (p-value)
# words	241	243	0.47	0.16
# links	11.1	10.9	0.82	0.99
# figures	1.9	1.9	0.98	1.00
# academic references	3.0	2.4	0.26	0.99
# google hits (millions)	1.9	4.3	0.32	0.08*

Observational articles: average starting length is **226 words**

Basline results

Results are bootstrapped at the Wikipedia article level



Nature of effect

Published Wikipedia articles show a distinct pattern not present in non-published ones

Control shows a "negative" trend in distributional shift

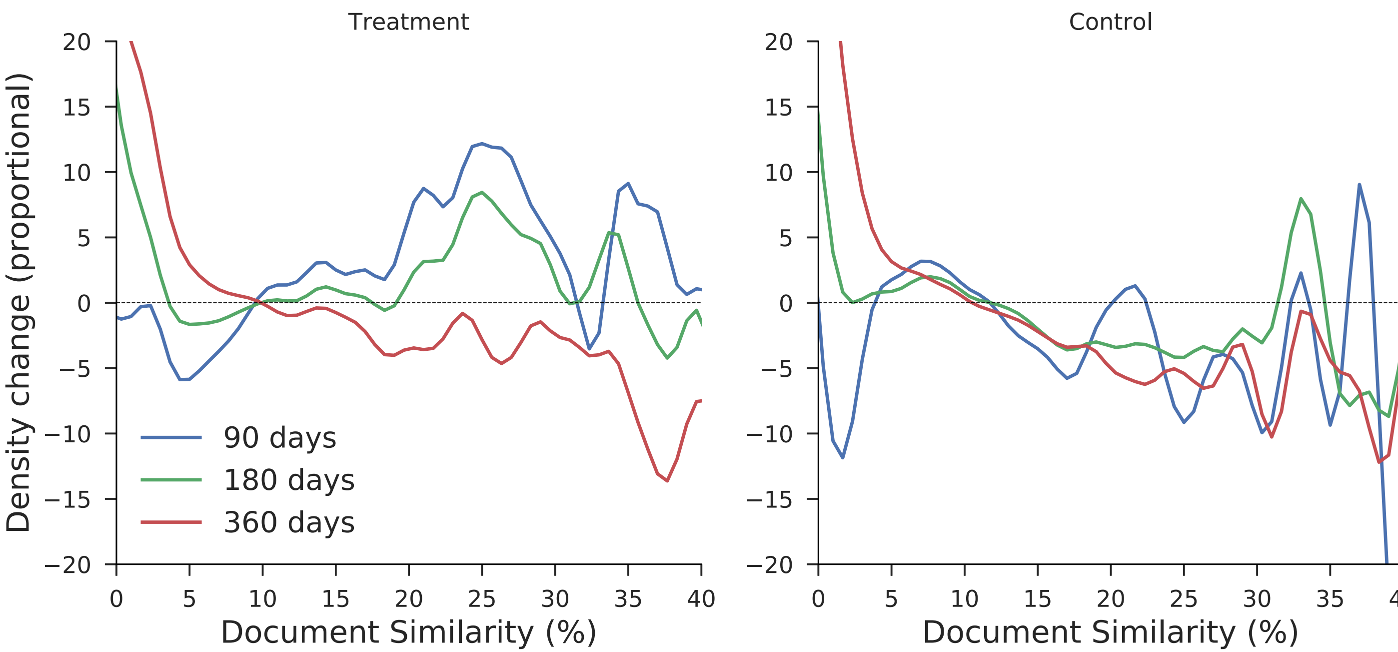
- science is constantly advancing and Wikipedia tries to keeps up

What is the observational analogue to the control articles?

- we do simulations of natural drifts in word usage frequency patterns and find find effects similar to our control

Observation windows

We look at a pre and post windows of length 90, 180, and 360 days. Also have a 90 day post-delay for publication



Why the fast decay?

Chemistry has a very rapid publish cycle (lucky them!). Articles are usually accepted within 2 months of submissions

After article submission, self-editing nature of Wikipedia takes over. This limits the *observability* of the effect over time

- Because of this we use only the text of the orginal submission ("intent to treat")

Regression design

Diff-in-diff on treatment vs non-treatment and before vs after window

$$\text{Similarity}_{ws} \sim 1 + \text{Treat}_w + \text{After}_s + \text{Treat}_w \times \text{After}_s$$

Because our unit of observation is a Wikipedia-science article pair, standard errors may be correlated

- This is particularly problemat given the number of Wikipedia articles (43)

We use the dyadic clustering method of Cameron and Miller (2015) for standard errors and boostrapping at Wikipedia level

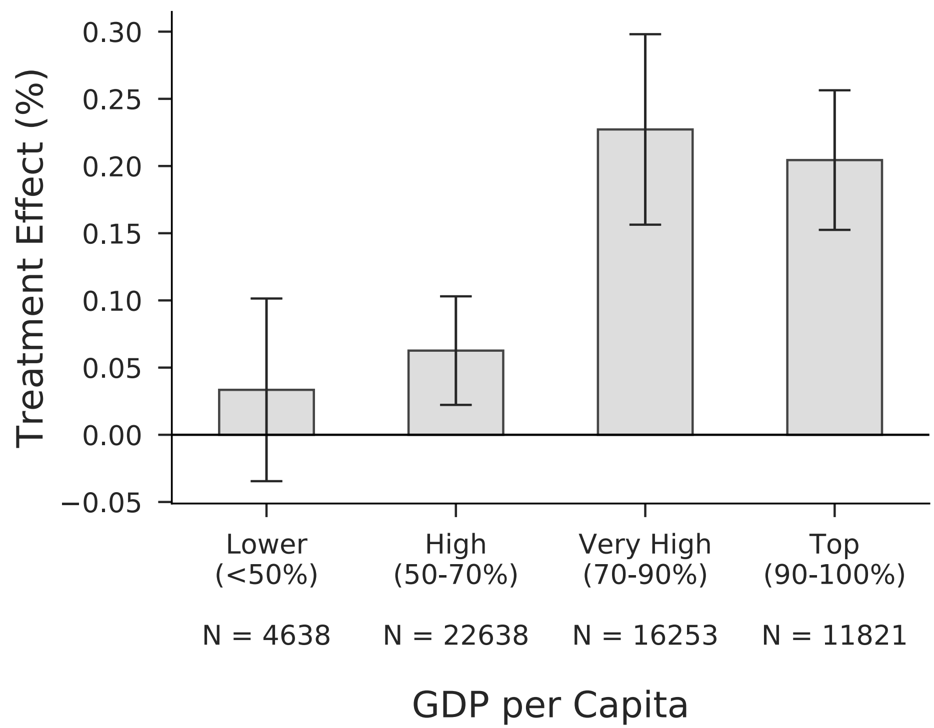
Regression results

Below are estimates for OLS and various quantiles

	Similarity (OLS)	**Similarity (q=25%)**	**Similarity (q=50%)**	**Similarity (q=75%)**
Intercept	11.2404*** (0.3778)	8.0502*** (0.3456)	10.4781*** (0.3719)	13.6000*** (0.3934)
Treated	−0.1367 (0.5859)	−0.2383 (0.3982)	−0.4068 (0.4865)	−0.4743 (0.6423)
After	−0.0768*** (0.0192)	−0.0499** (0.0201)	−0.0715*** (0.0253)	−0.1103*** (0.0399)
Treated x After	0.1181*** (0.0358)	0.0804*** (0.0263)	0.1041** (0.0412)	0.1815*** (0.0604)

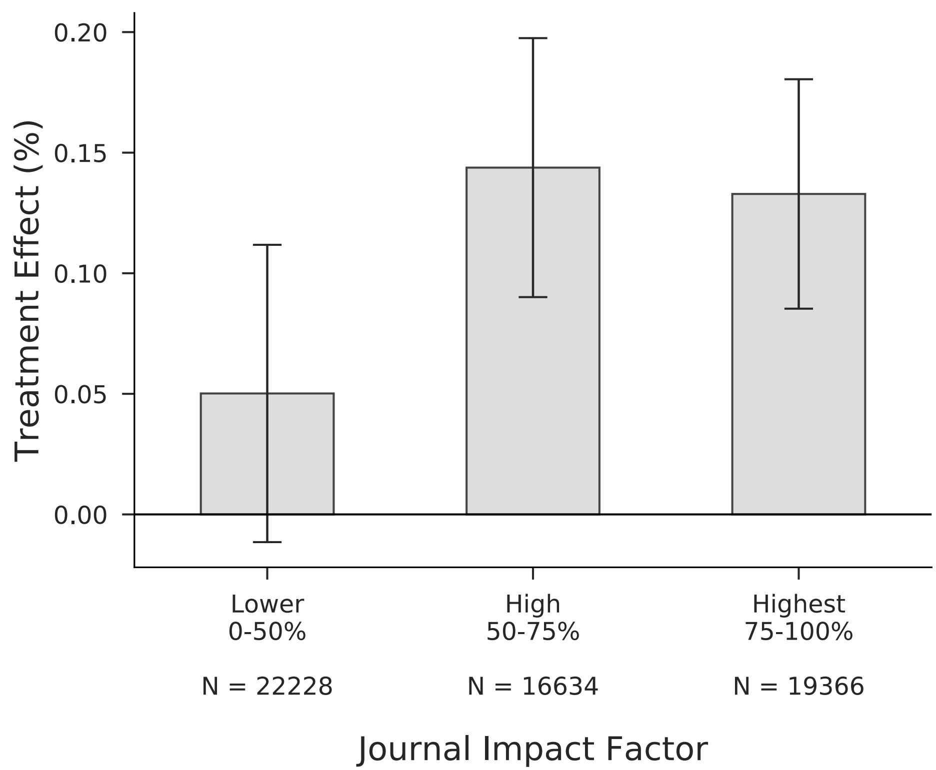
Country effects

Using modal institution of science authors, we can include GDP per capita of the country as a covariate



Journal quality

Similarly, we can utilize the ranking of the journal of publication



Page views

We can see page view counts for all our articles, with controls naturally given zero views

Suprisingly, we see no effect of page views on similarity

- could be that people are glimpsing at Google preview
- presumably only a small fraction of people are paper writers, hence a noisy measure

Citations

We can also look at the effect on citations of articles that are mentioned in a treated Wikipedia entry. Effect seems to be stronger for those with already high citations.

	Citations Growth (%)
Intercept	−15.7676 (10.8078)
Cites Pre	0.2224 (0.2693)
Treated	56.0418** (27.3931)
Log Views	−7.6745* (4.4475)

Article sections

Typical chemistry article layout (very common)

- introduction → methods → results → conclusion

Where does the effect seem to be concentrated?

- introduction, results, and conclusion have similar effects
- methods section shows no effect

Wikipedia may not determine which experiments are done, but could affect how people interpret them and understand them within existing literature

Econometrics

We also performed a similar experiment on econometrics

- there was no effect!

What might be beind this?

- publications lags are much longer (we've waited two years since publications)
- economics has a strong working paper culture (thanks RePeC!) while chemistry is legally restricted

Interpretation

What is the ultimate impact of these effects? One might worry that researchers are focusing on certain topics at the expense of others (with no change in total research)

- if one changes their action in response to new information, this makes them (weakly) better off?

Could be a public downside if there is some inefficiency in research efforts

- diffusion $\rightarrow$ escape competition vs dilution of returns
- choice between incremental and radical innovation

Pratical effect

Hard to disentangle **intensive** and **extensive** margin of effect of Wikipedia

Suppose this intervention affected 1% of chemistry articles

- then each article changed by 10% (10% x 1% = 0.1%)
- total of 600 articles affected (30 per treatment entry)

Note that changes are in importance weighted words

- simulations show that 10% similarity change ≈ 10% random words changed

Conclusion

Observational component has shown the existence a baseline dynamic between Wikipedia and science

Randomized trial data clarifies issuess of causality

- Wikipedia not only reflects the current state of science -- it helps shape it

Wikipedia could be a cost-effective way to promote knowledge diffusion

Future work

Could eventually look at relationship with other text sources such as patents (link to productivity)

Look at other public knowledge repositories such as Github or StackExchange

How important is cross-field knowledge diffusion? Do fields have something like an input-output matrix or a hierarchy?

Theory

How can we incorporate Wikipedia effect into existing growth models?

Standard Jones (1995) framework looks like

$$\dot{A} = A^{\phi} R^{\lambda} = A^{\phi} (s_R L)^{\lambda}$$

A - technology, ϕ - feedback, R/s_R - researchers/share, L - population

On a balanced growth path, this leads to

$$g \equiv \frac{\dot{A}}{A} = \frac{(s_R L)^{\lambda}}{A^{1-\phi}} = \frac{\lambda n}{1 - \phi}$$

Multi-field

Critical parameter is ϕ, which determines how existing knowledge affects the generation of new knowledge

Effect may be not only within fields but across fields, so consider multiple interacting fields

$$\dot{A}_i = \left[\prod_j A_j^{\delta_{ij}}\right]^{\phi_i} (s_i L)^{\lambda_i}$$

The matrix δ determines the strength of between-field interactions, ϕ vector determines overall effects

Knowledge Growth

Can express the growth rate as combination of pure effects and interactions

$$\frac{\dot{A}_i}{A_i} = \left[\prod_j \left(\frac{A_j}{A_i}\right)^{\delta_{ij}}\right]^{\phi_i} \times \frac{(s_i L)^{\lambda_i}}{A_i^{1-\phi_i}}$$

On a balanced growth path, growth rates satisfy

$$g_i = \phi_i \sum_j \delta_{ij} g_j + \lambda_i n$$

Whenever $\phi_i < 1$ and $\sum_j \delta_{ij} = 1$ (WLOG), this is a contraction mapping.

Diffusion Matrix

We can express the solution using linear algebra

$$
\begin{aligned}
g &= \delta_\phi g + \lambda n \\
\Rightarrow \quad g &= \left[I - \delta_\phi\right]^{-1} \lambda n
\end{aligned}
$$

where $(\delta_\phi)_{ij} \equiv \phi_i \delta_{ij}$

Proposition: Whenever $\delta = I$ or $\phi_i = \phi$ and $\lambda_i = \lambda$ for all i, the resulting growth rates are separeble

$$
g_i = \frac{\lambda_i n}{1 - \phi_i}
$$

Asymmetries

We require systematic differences across fields for δ_ϕ to be important. Some classes of matrices

$$\text{Symmetric} \qquad \delta_\phi = \begin{bmatrix} \delta & 1-\delta \\ 1-\delta & \delta \end{bmatrix}$$

$$\text{Hierarchical} \qquad \delta_\phi = \begin{bmatrix} \delta & 1-\delta \\ 0 & 1 \end{bmatrix}$$

Hierarchical

Growth rate in the hierarchical case

$$g = \frac{\lambda n}{1 - \phi} \begin{bmatrix} 1 - \frac{\phi_1 - \phi_1 \delta}{1 - \phi_1 \delta} \frac{\phi_1 - \phi_2}{1 - \phi_2} \\ 0 \end{bmatrix}$$

Thus with output $y = \alpha a$, we will have

$$\frac{\partial g_y}{\partial \delta} > 0 \quad \Leftrightarrow \quad \frac{\partial g_1}{\partial \delta} > 0 \quad \Leftrightarrow \quad \phi_1 > \phi_2$$

Might think that Wikipedia effect is unambiguously positive, but could be a matter of time allocation

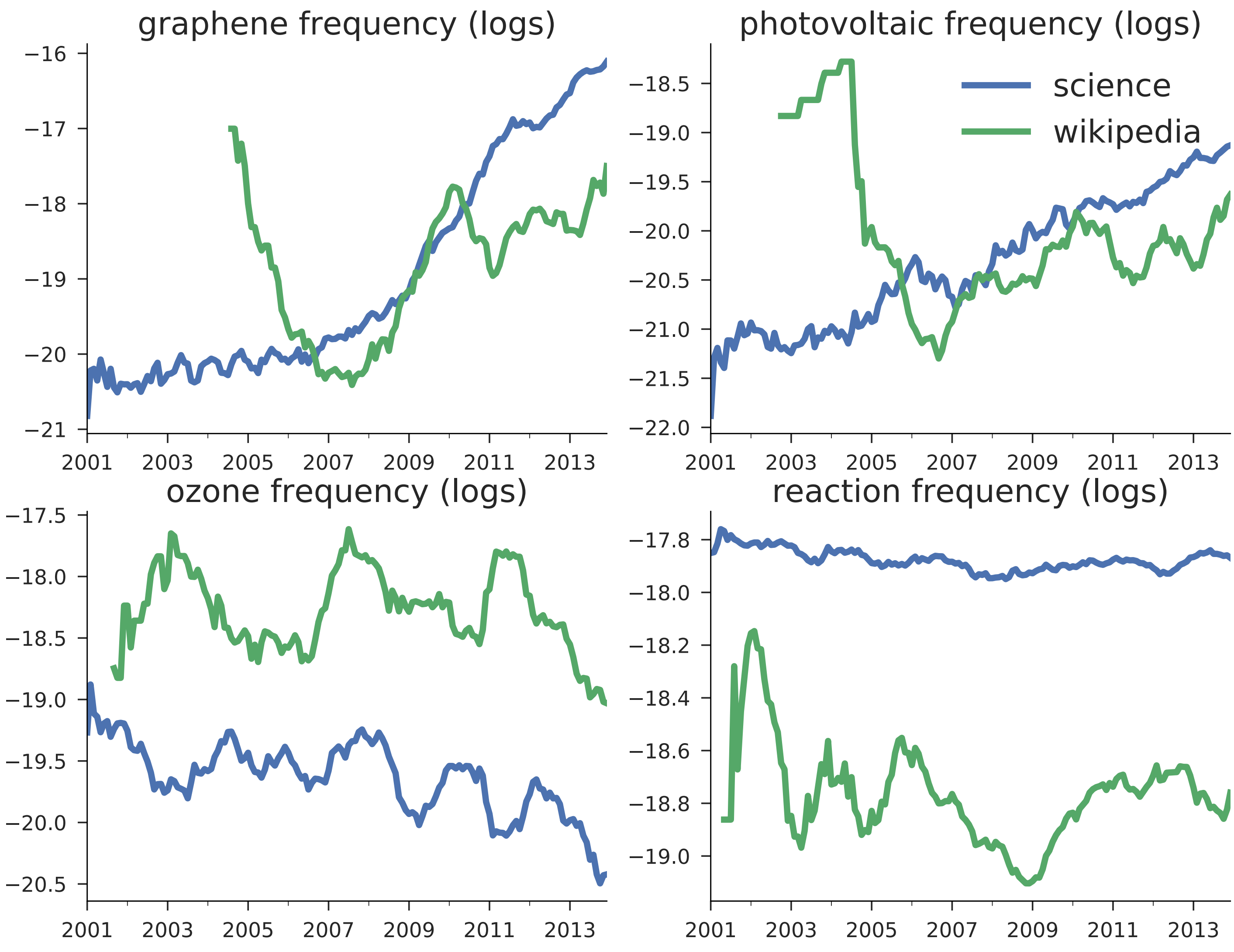
Word Frequency

Final output of Wikipedia data is stream of **new** words being added through edits

Aggregate this to monthly word frequency vectors over 1.3M words appearing 3 or more times in corpus.

Analogous monthly series of frequency vectors over words appearing in published scientific articles (chemistry)

Word Frequency (wikigrams)



Wikipedia → Science

Logs on logs, assuming positive ($R^2 = 0.9167$)

Science(t+1)	coef	std err	p-value
Wikipedia(t)	0.0776	0.0008	0.0000
Science(t)	0.9167	0.0006	0.0000

Binary outcome model ($R^2 = 0.2733$)

Science(t+1) > 0	coef	std err	p-value
Intercept	0.1927	0.0000	0.0000
Wikipedia(t) > 0	0.2261	0.0001	0.0000
Science(t) > 0	0.4208	0.0001	0.0000

Adoption Dynamics

Might be worried about pretrends in literature frequency. Controlling for levels and changes at t takes care of adoption curve dynamics.

Wikipedia → Science: diffs on diffs (logs) ($R^2 = 0.2258$)

ΔScience(t+1)	**coef**	**std err**	**p-value**
Wikipedia(t)	0.0397	0.0008	0.0000
ΔScience(t)	-0.4407	0.0014	0.0000
Science(t)	-0.0436	0.0006	0.0000

Science → Wikipedia

Diffs on diffs, assuming positive (logs) ($R^2 = 0.2061$)

ΔWikipedia(t+1)	coef	std err	p-value
Science(t)	0.0904	0.0007	0.0000
ΔWikipedia(t)	-0.2588	0.0016	0.0000
Wikipedia(t)	-0.1844	0.0011	0.0000

Binary outcome model ($R^2 = 0.4346$)

Wikipedia(t+1) > 0	coef	std err	p-value
Intercept	0.0623	0.0003	0.0000
Science(t) > 0	0.1443	0.0004	0.0000
Wikipedia(t) > 0	0.6052	0.0005	0.0000

Emails!

Consider the effect of a large, exogenous change in the vocabulary used in both Wikipedia and science