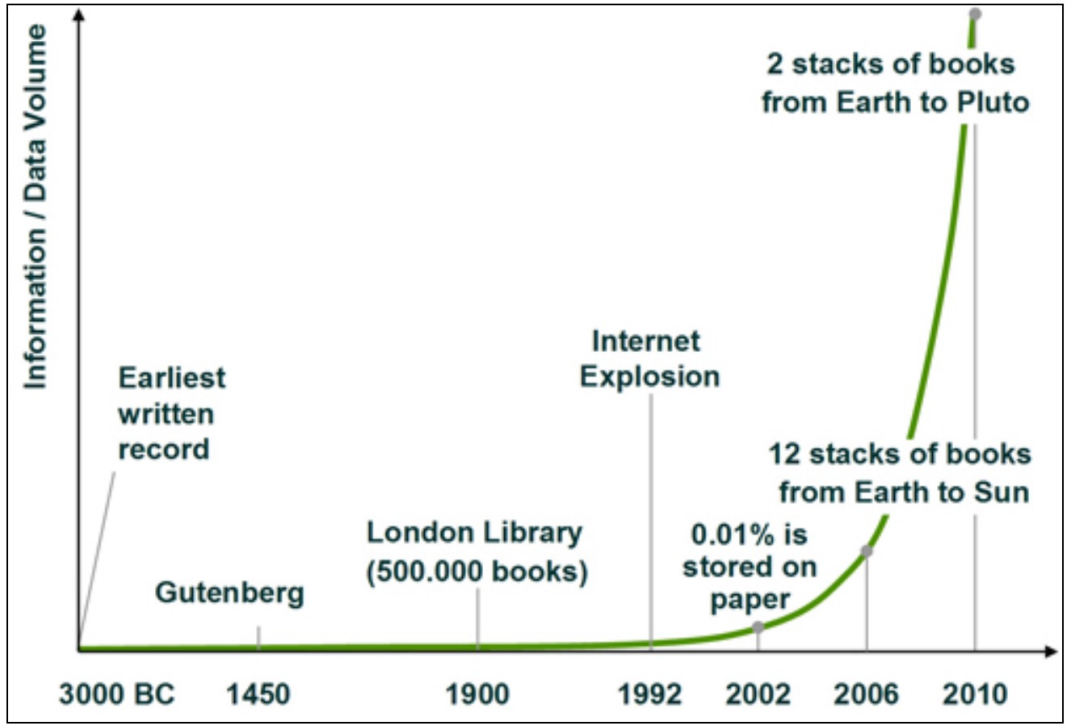
The Long Game: Preserving Human Digital Trace Data for Future Generations

Patrick Park

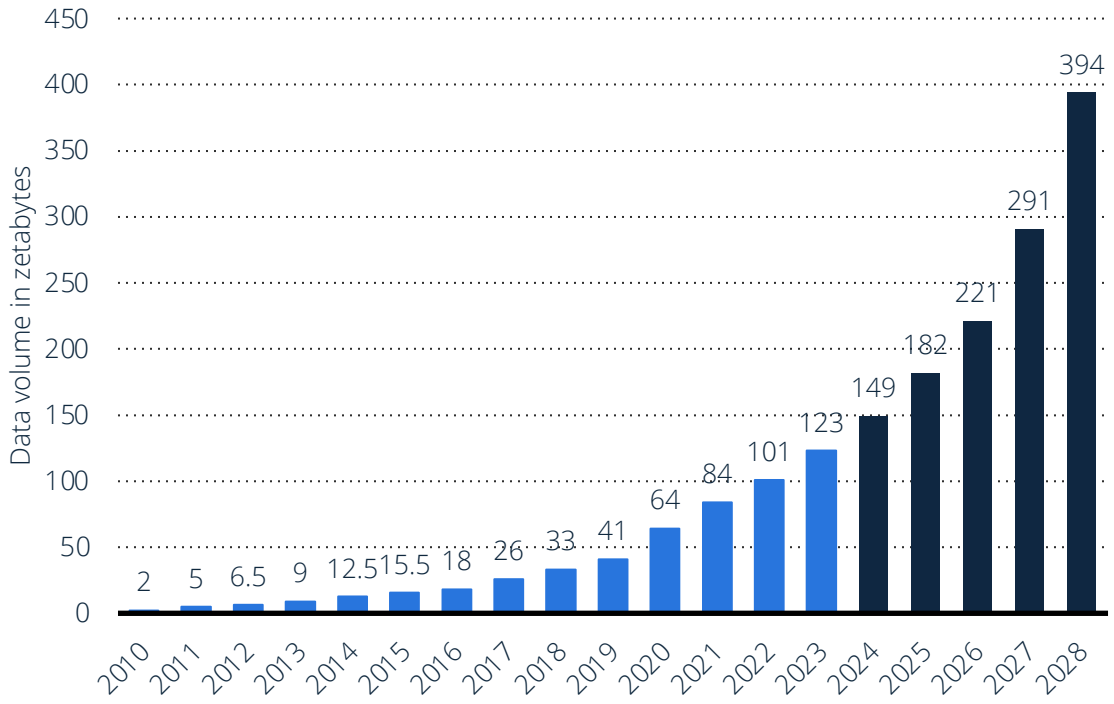
Software and Societal Systems Department

Carnegie Mellon University

Data Explosion



Mestl et al., 2009



Statistica, 2024

Academic Bias for Fresh Data

"… **the dataset is pretty old**; Twitter is now pretty different I'm guessing; and that may affect both results and how panelists assess the proposal." (HCI)

"… but (a) **it is over 10 years old** and prior to the current twitter management regime, …" (Sociology)

"Journal of Advanced Nursing (JAN) specifies in its authorship guideline that the period of **data collection should ideally be no more than 3 years before submission** of the manuscript (Dale and Logsdon, 2022)."

Justification for New Data

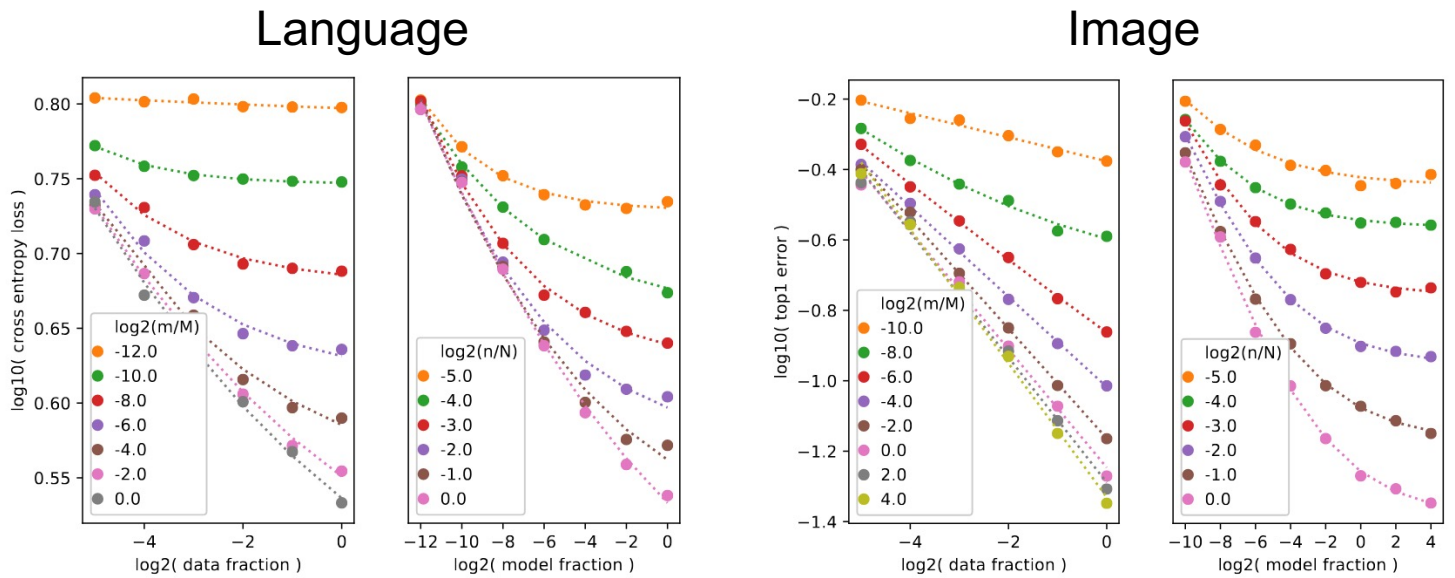
Researchers often need new data to solve current pressing problems

Academic incentives: New data also give a competitive advantage for publication

Exponential growth of new data can dramatically improve AI performance

(Scaling hypothesis)

AI Systems Are Hungry for New Data



(a) Wiki103 cross entropy vs. data and model size. (b) CIFAR10 top1 error vs. data and model size.

Figure 2: Error vs. data size (left part of each subfigure) and model size (right part) for Wiki103 and CIFAR10. Solid dots are measurements, dashed lines are best fit to saturating power-law.

Improving AI requires ever larger data (diminishing returns)

Error rates decrease with respect to model size and data size as approximately power-law functions (Rosenfeld et al. 2020)

The Problem: No Time for Digestion

Hard to keep up with state-of-the-art: Concentration of scholarly attention

Data-driven research: Studying the data, rather than addressing the real questions

Discoveries are prone to recency bias

Hence, theories and explanations based on those discoveries are also potentially biased

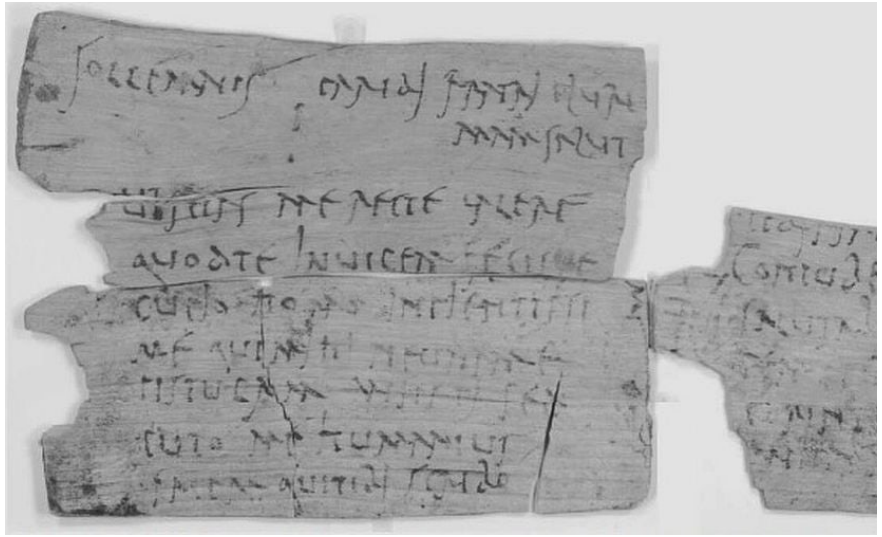
The Need for Long-Term Data Preservation

"In my office, I have a bookshelf that's eight feet high, 10 feet wide, and it contains pretty much all the main surviving Greek and Roman literary texts…

One bookshelf, it's a big bookshelf, but that's what we use to interpret this world."

– Gregory Aldrete, historian

The Need for Long-Term Data Preservation

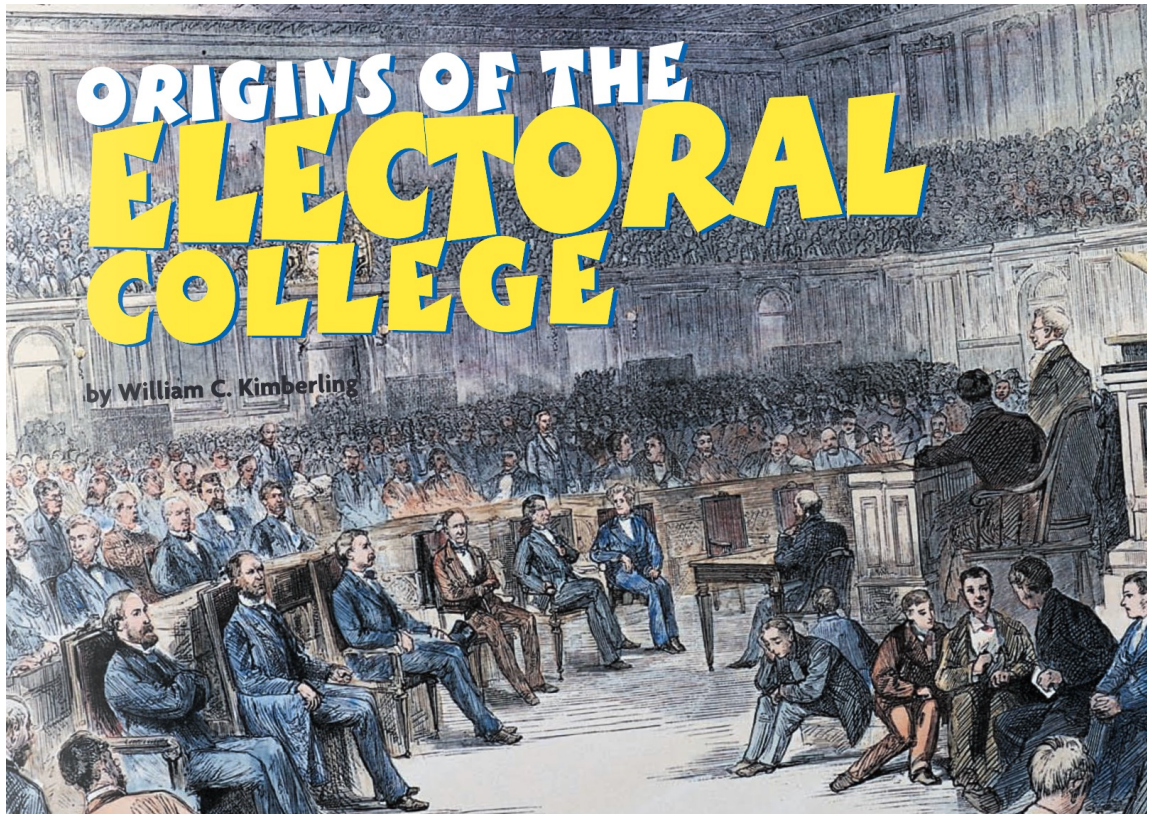


Given all the things we learned from a bookshelf amount of information, how would have history changed if 40M books passed down from the Roman Empire?

Why Old Data?

Important for understanding the present

- “Why did we need this in the first place?” (Institutional dementia)
- “Why do we still need it?”
- What additional social structures formed on the basis of this law in later history?



Why Old Data?

Scientific knowledge accumulation

1. Some social processes take time to unfold
2. Digesting information takes time
3. Unexpected new uses of data

Long-Term Social Processes

Moving to Opportunity

A randomized social experiment in the 1990s that gave very poor families the opportunity to move out of subsidized housing in poor neighborhoods to more affluent neighborhoods

Idea: Positive neighborhood effects on children's social mobility

Moving to Opportunity

Why Did the Moving to Opportunity Experiment Not Get Young People into Better Schools?

Xavier de Souza Briggs
Massachusetts Institute of Technology

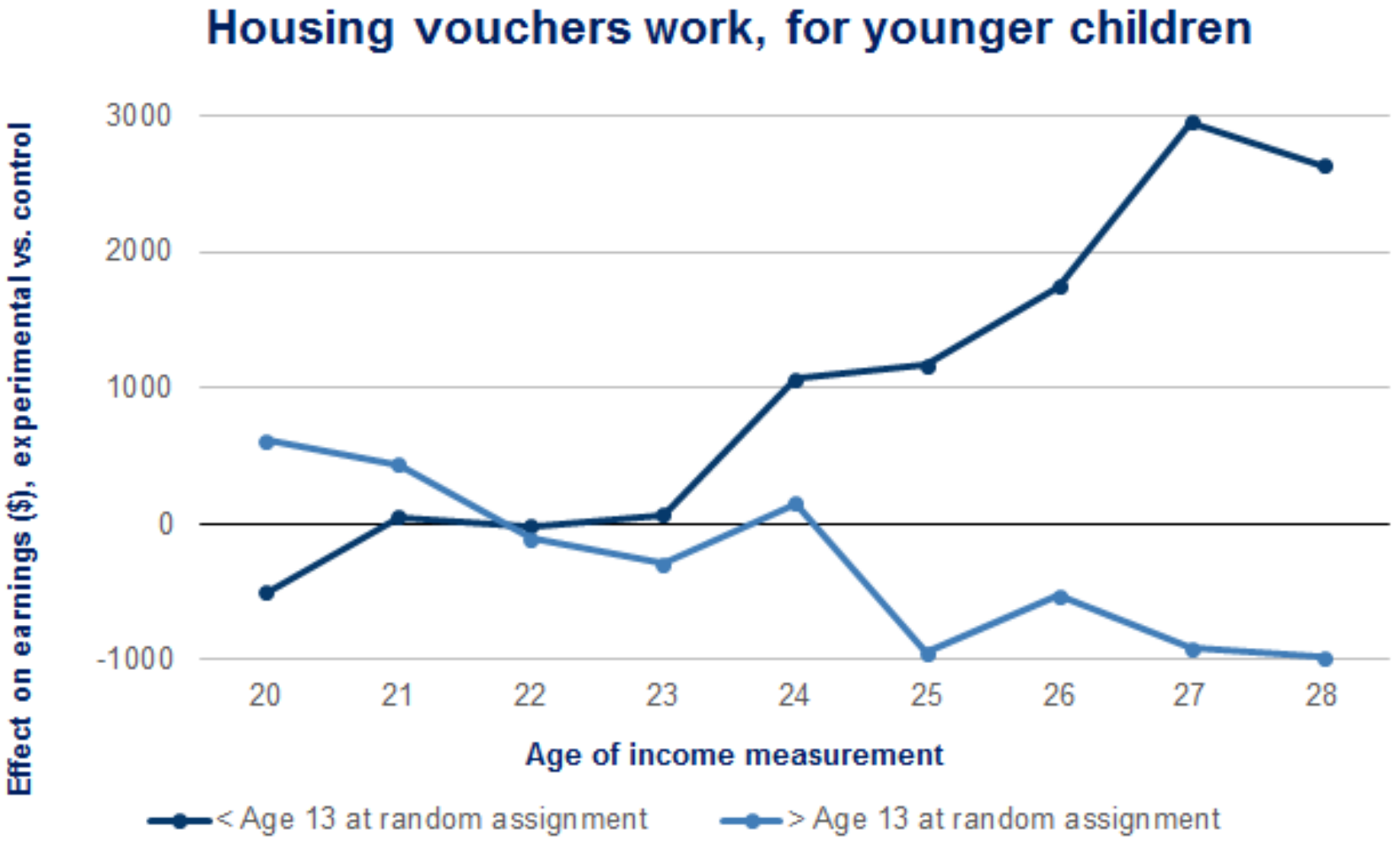
Kadija S. Ferryman
The Urban Institute

Susan J. Popkin
The Urban Institute

María Rendón
Harvard University

Early studies showed weak, insignificant effect on children's outcomes

Failed experiment



The real effects only started to emerge in the longer-term

Big effect on children who moved at younger ages

Source: Chetty et al., "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment," Figure 1: Impacts of experimental voucher by age of earnings measurement

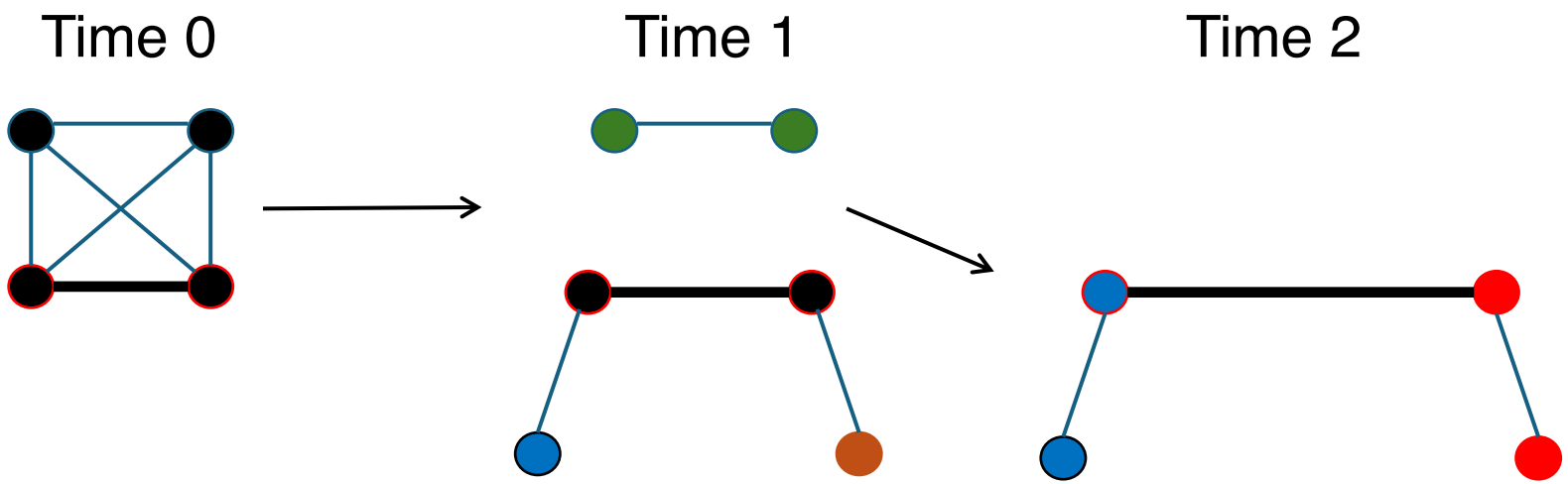
BROOKINGS

Moving to Opportunity

Possibility of other social experiments and policy interventions that may have been prematurely declared “failed”

Long-Term Evolution of Social Ties

Scholarship on how social relationships evolve over time is rare



Park, Xu, and Carley 2024

Evolution of Social Ties

Difficult to trace relationships in the long-term

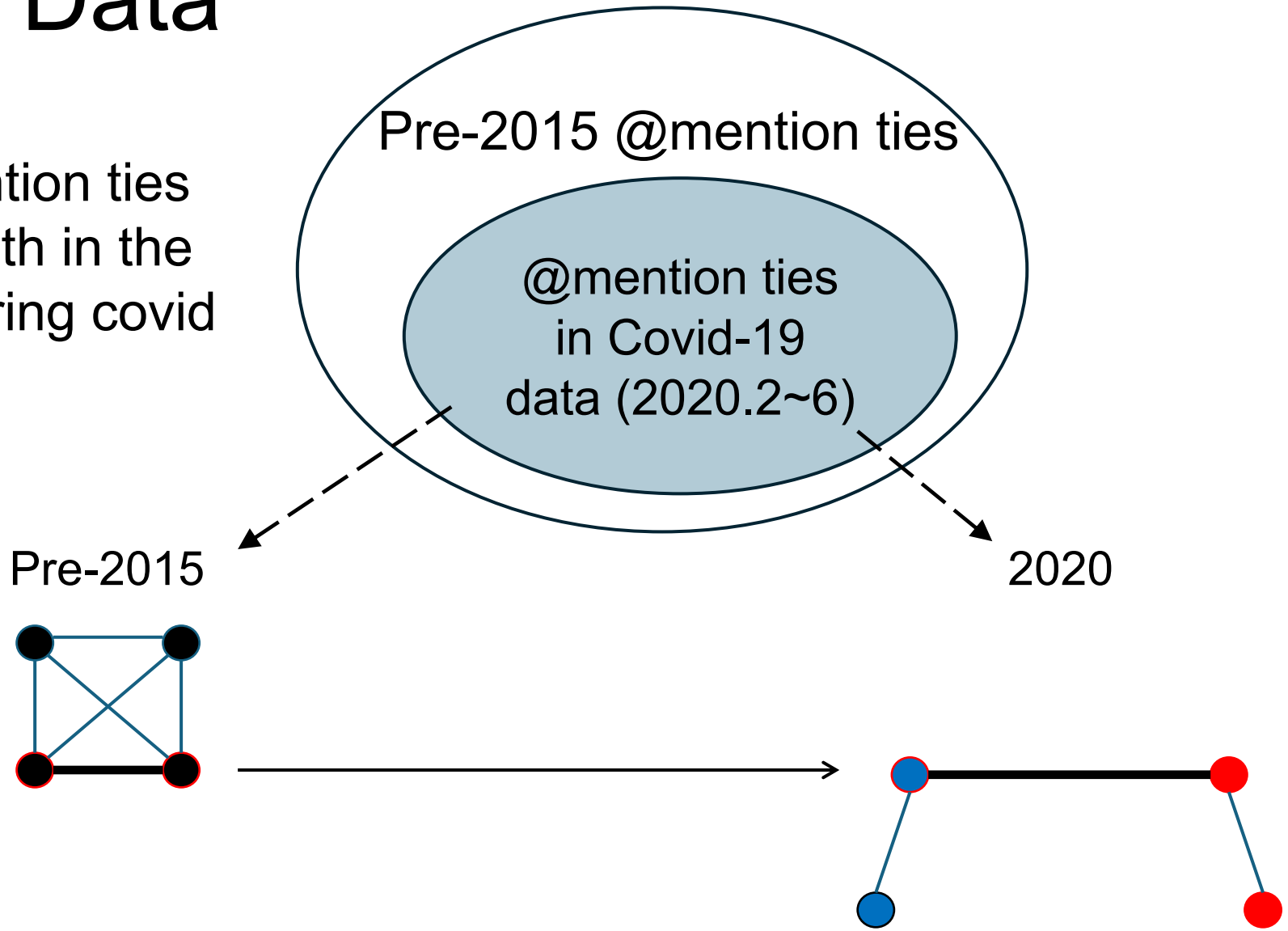
- Participant attrition
- Costly

Social media study required preserving 200TB of data for years

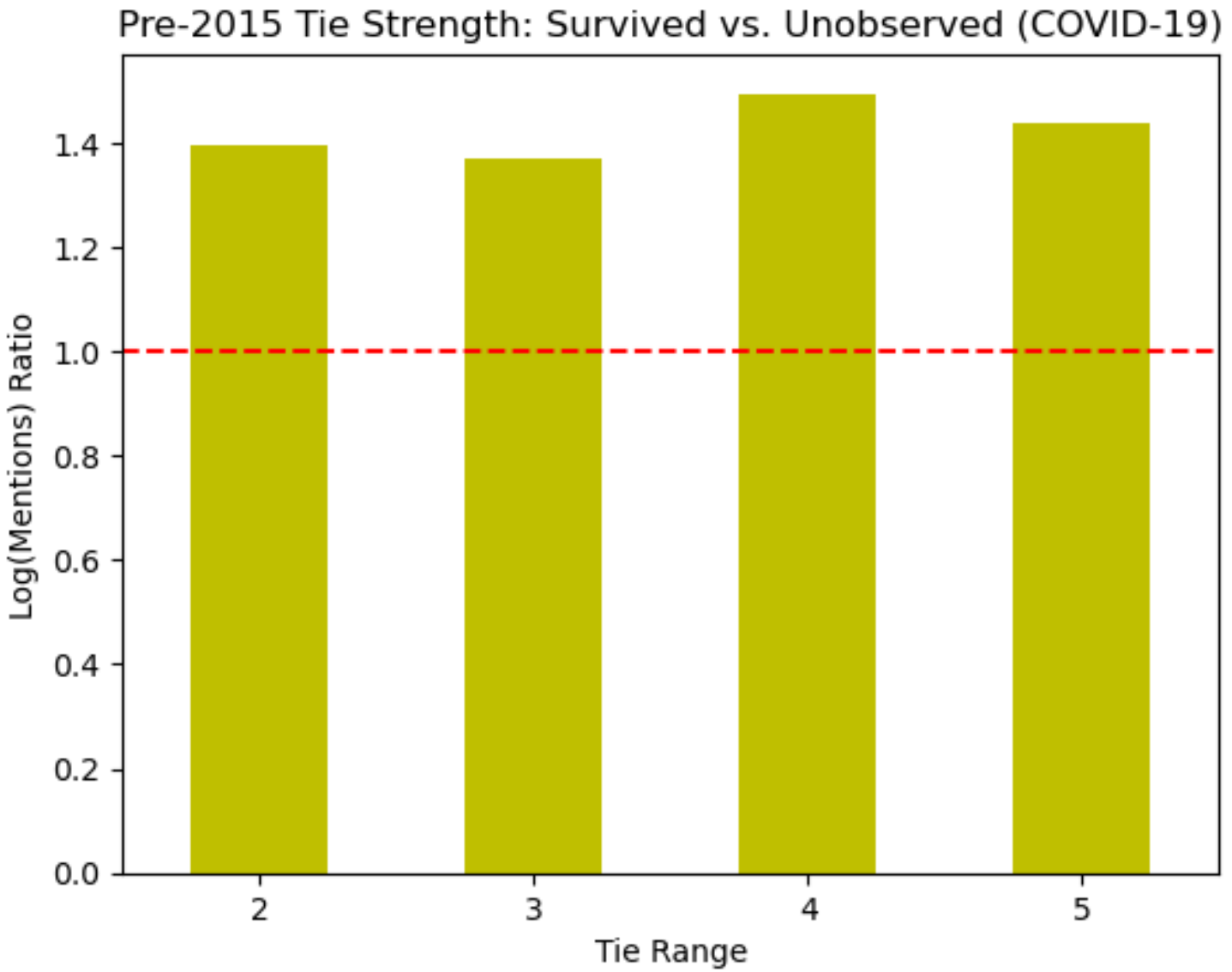
- High data storage cost in the cloud computing service
- Hard drives and cold storage solutions tend to fail → data loss

Twitter Data

430K @mention ties observed both in the past *and* during covid

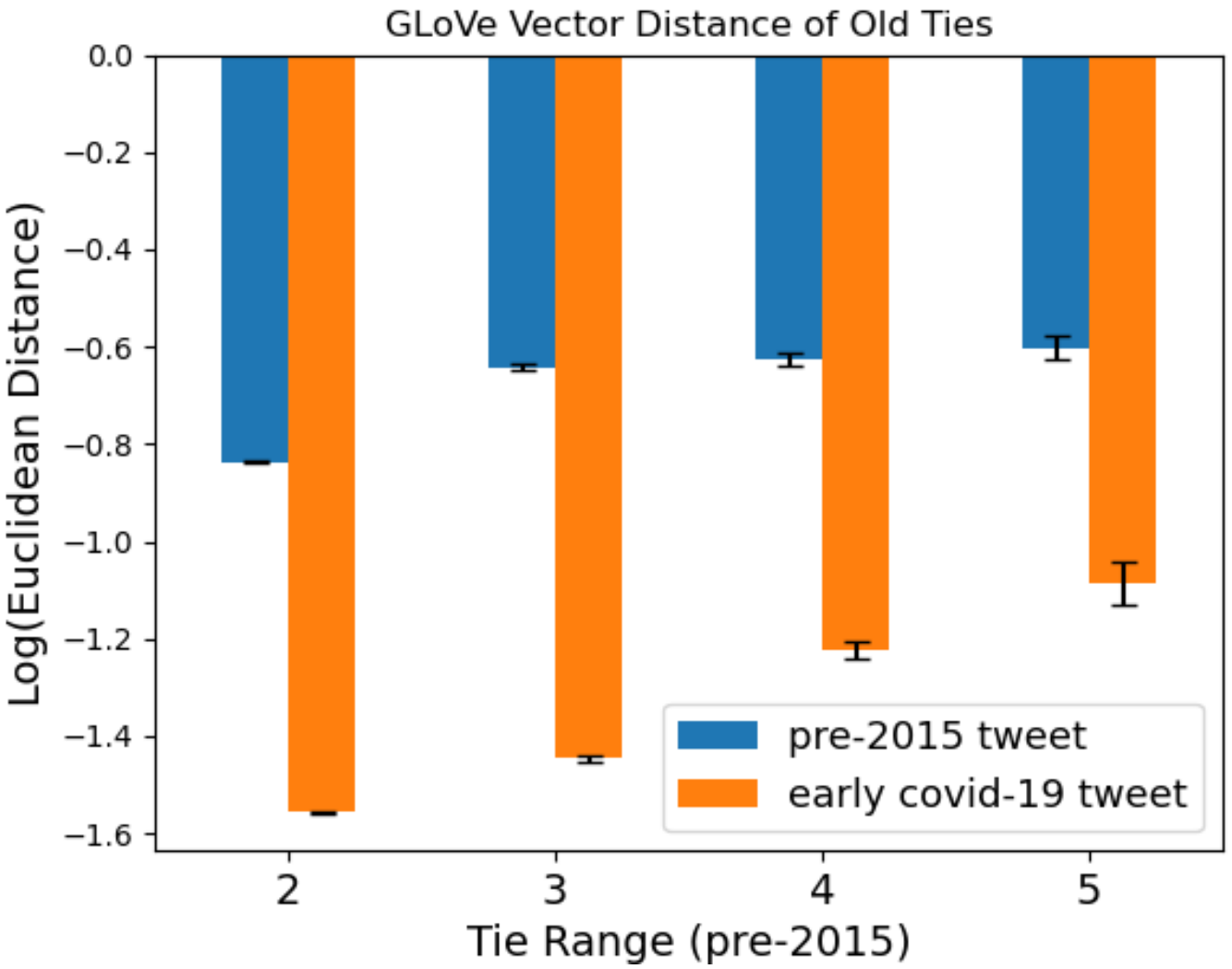


The pre-2015 ties observed in the Covid data (survived) were stronger than the rest (unobserved)

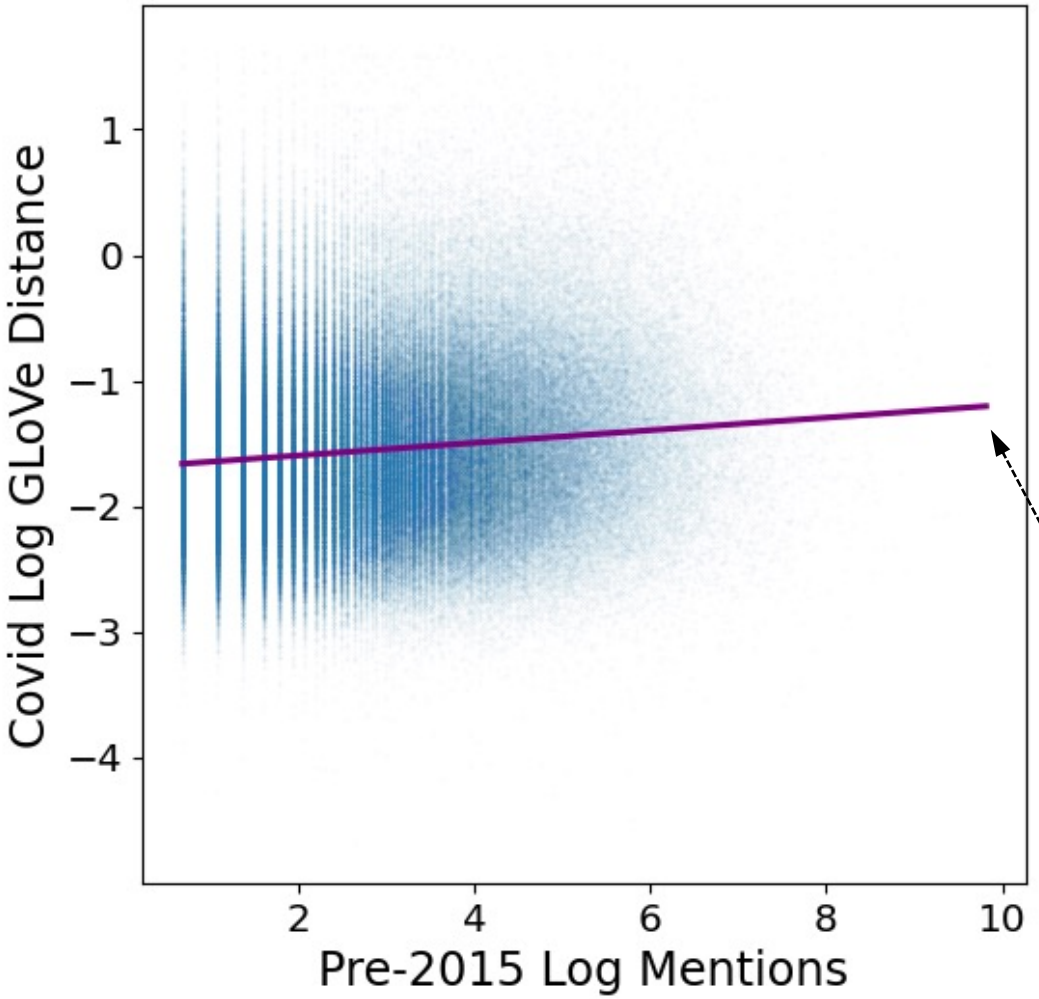


Old ties that survive have shorter cognitive distance in the present (orange vs. blue)

Embedded old ties (range=2) have shorter cognitive distance than the bridging ties

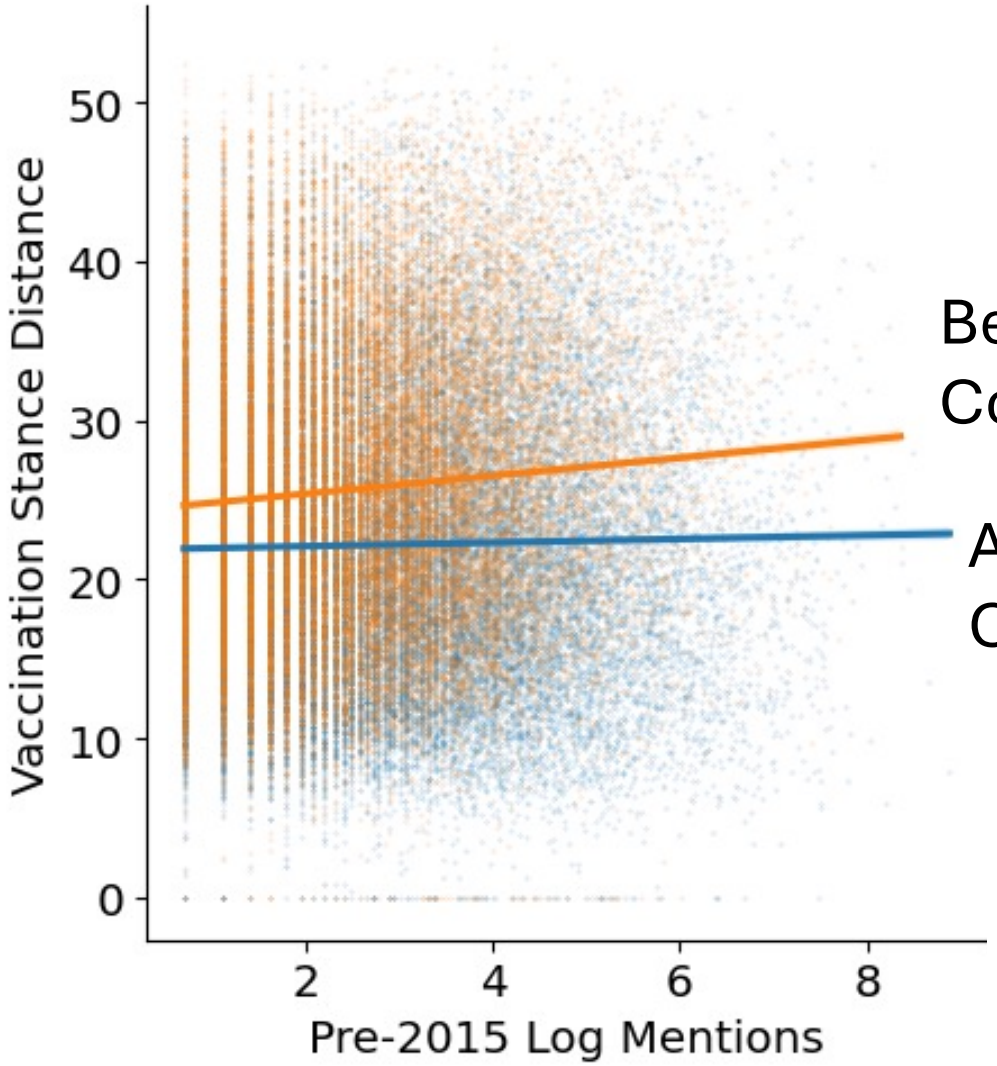


Cognitive Difference



Ties that used to be stronger are more cognitively distant about **Covid-19**

Vaccination Stance Distance



Below-Median
Covid Mention Frequency

Above-Median
Covid Mention Frequency

Homophily and Tie Survival

Table 2. Logistic Regression on Bidirected vs. Unidirected Ties in the Covid Data

Variable	**B**	**S.E.**	
Log Mention Frequency (pre-2015)	0.33	0.02	**
Same Occupation	-0.04	0.09	
Same Family Role	0.76	0.06	**
Same Political Orientation	-0.10	0.04	*
Same Cultural Interest	-0.41	0.06	**
Same Sports Interest	-3.02	0.01	**

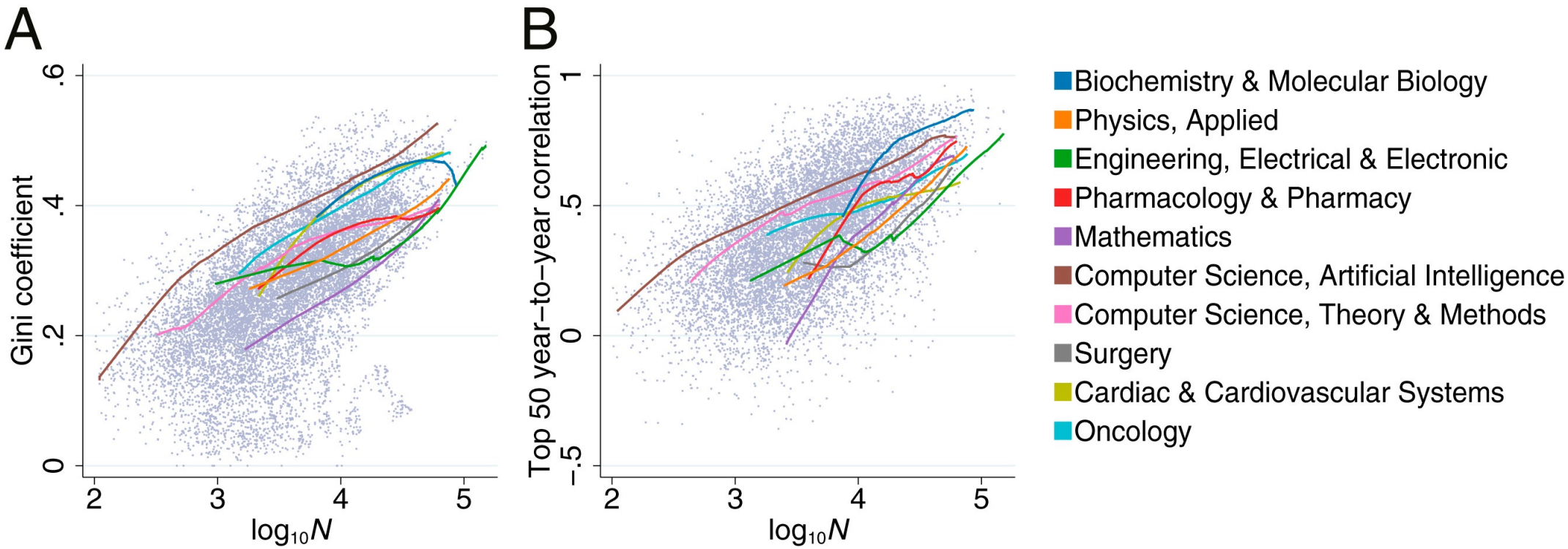
*<0.01, **< 0.001 N= 425597

$R^2 = 0.017$

Digesting Information Takes Time

Over-Concentration of Scholarly Attention

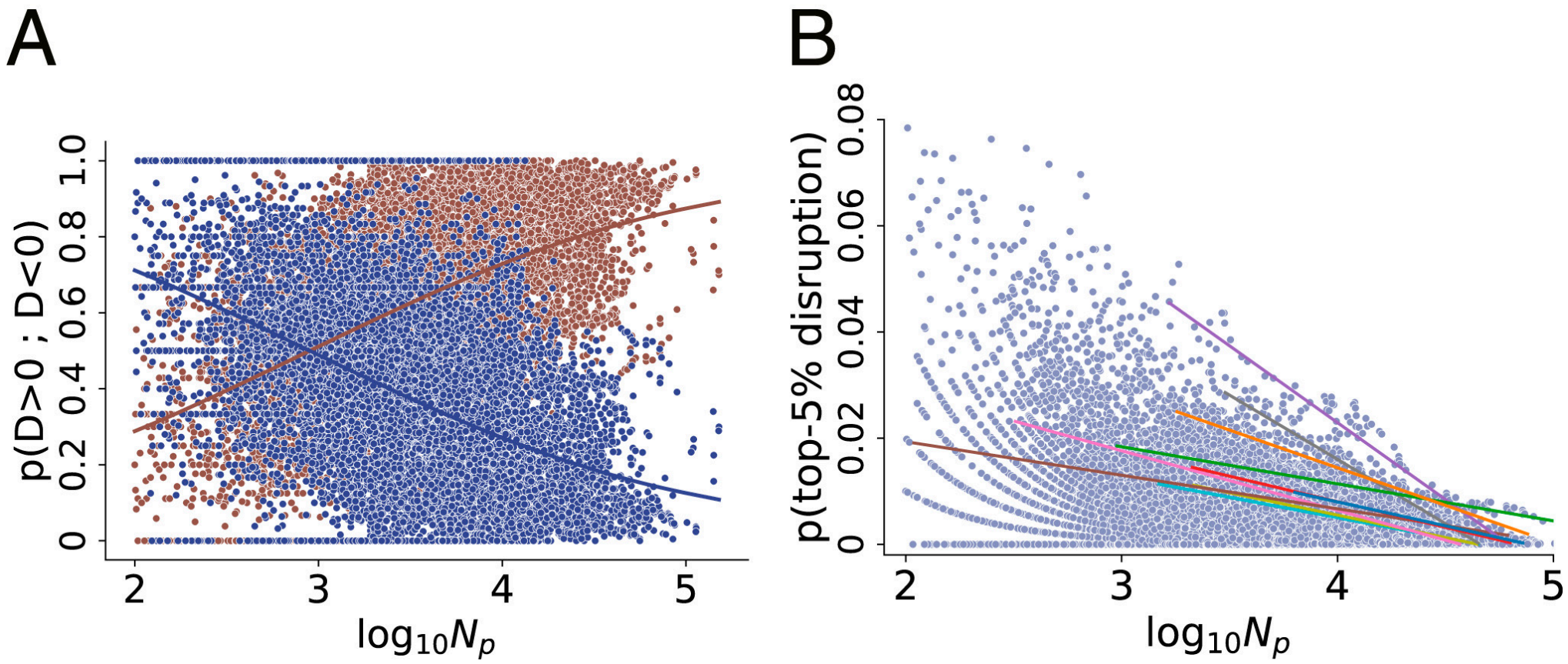
Top articles garner larger portions of citations as a discipline grows in size



Chu and Evans, 2021

Fewer Disruptive Ideas

Longer information digestion time needed for innovation



Chu and Evans, 2021

Novel Uses of Data

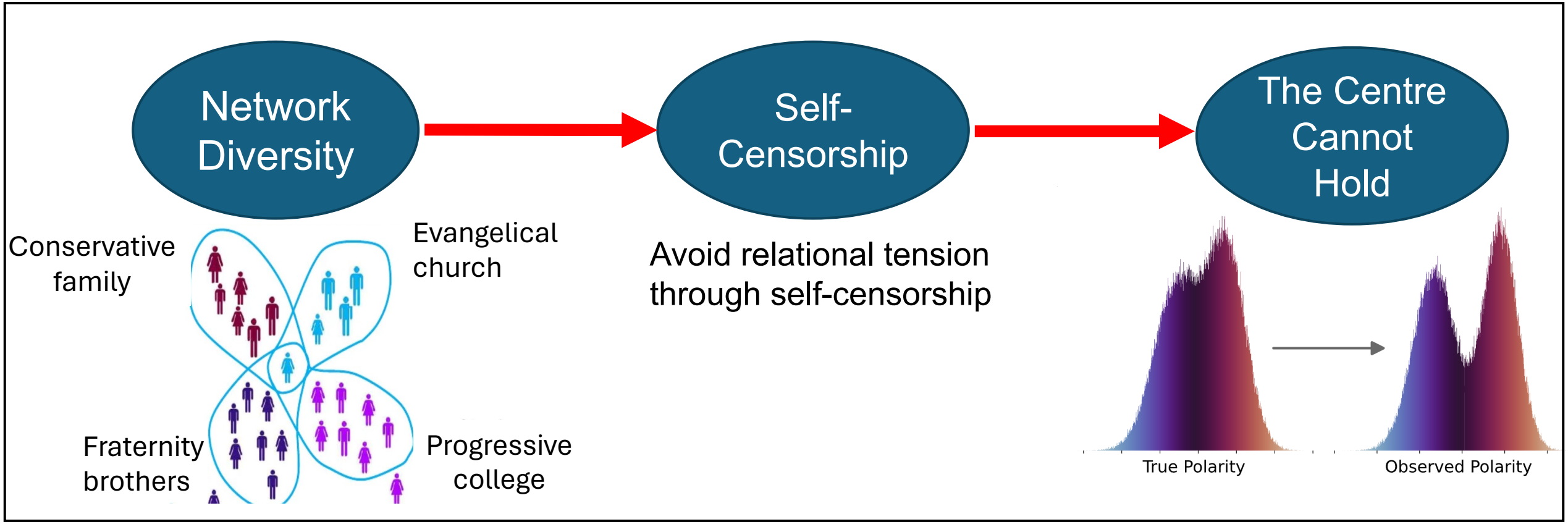
Paths to New Uses of Old Data

1. Exploring novel explanations of societal issues
2. New idea → new uses of old data
3. New findings → new explanations → rediscovery of old data
4. New discipline → new uses of old data
5. New combinations of old data
6. Data access to new users → innovative ideas

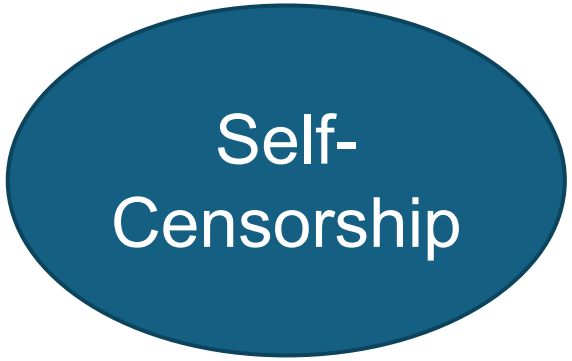
New Ideas Using Old Data

New Idea Sparks New Use of Old Data

Example: Explaining online political polarization



New Idea Sparks New Use of Old Data

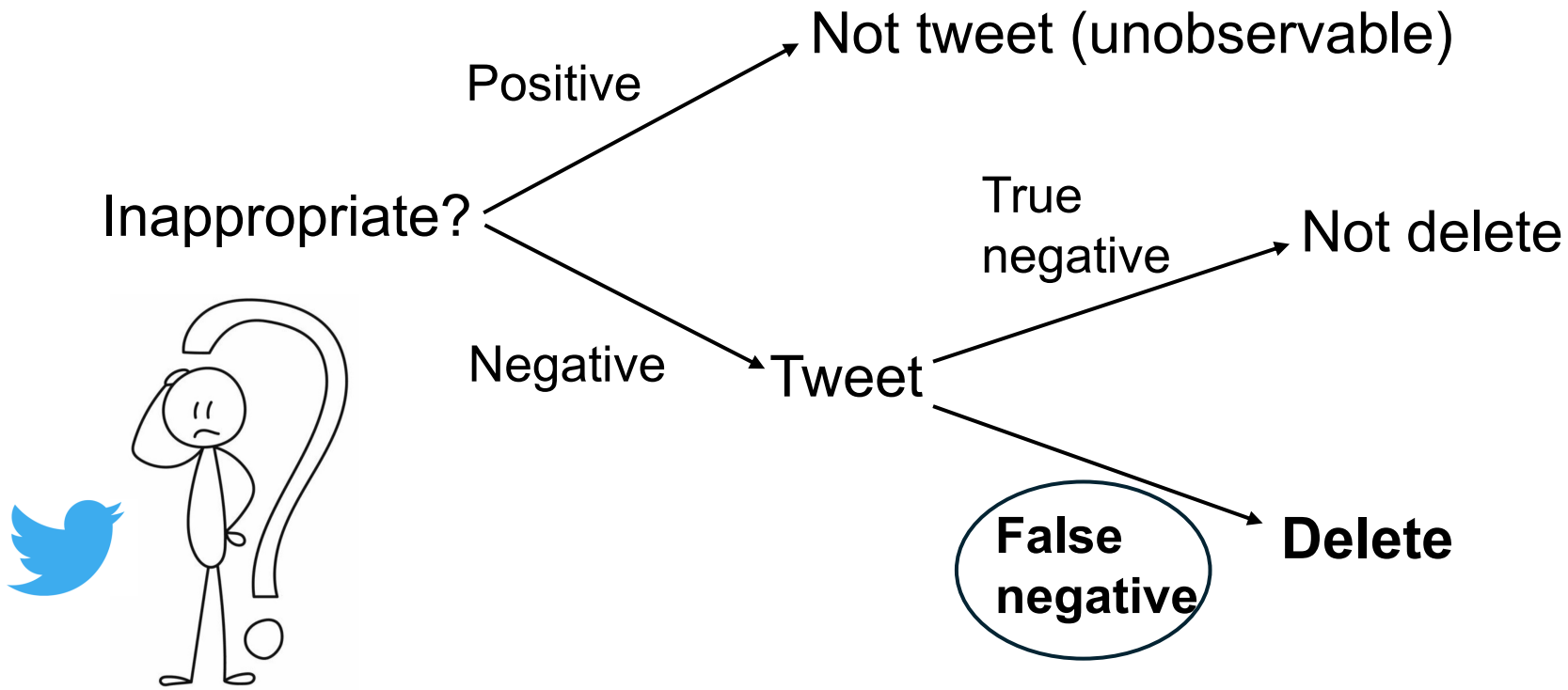


Problem: Self-censorship is hard to observe

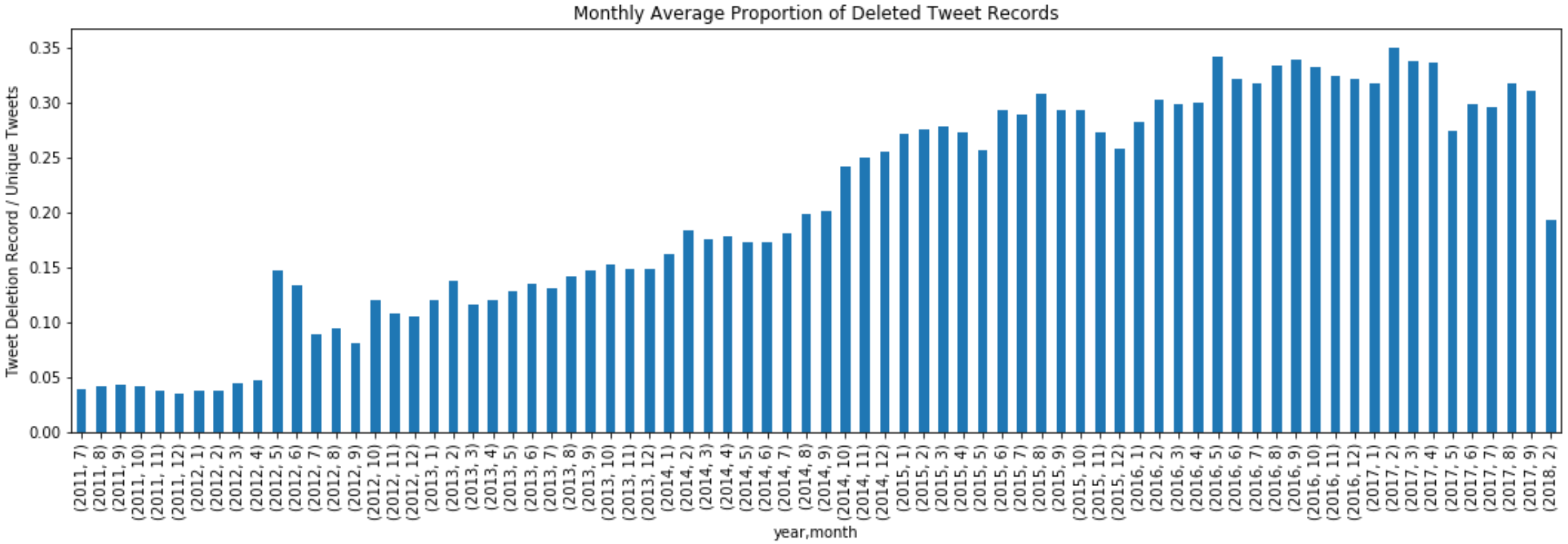
New Idea Sparks New Use of Old Data

Solution

- Self-censorship: Number of deleted tweets (logged)



New Idea Sparks New Use of Old Data

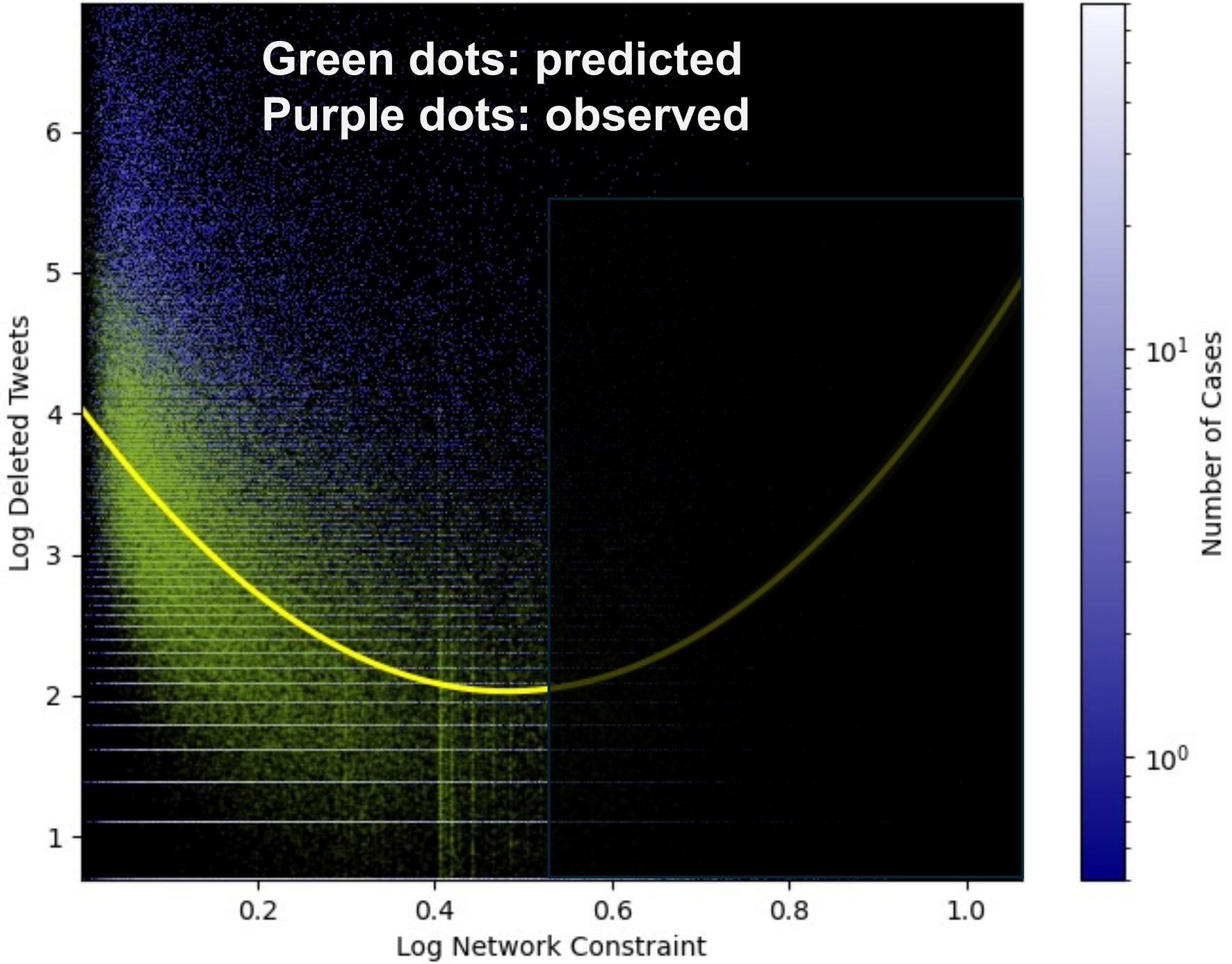


Tweet deletion has increased over the years (2011~2018)

Polarization also intensified during these years

26M US Twitter users

Users with
diverse networks
delete more tweets



New Findings, Rediscovery of Old Data

New Findings Open New Uses of Data

My own experience:

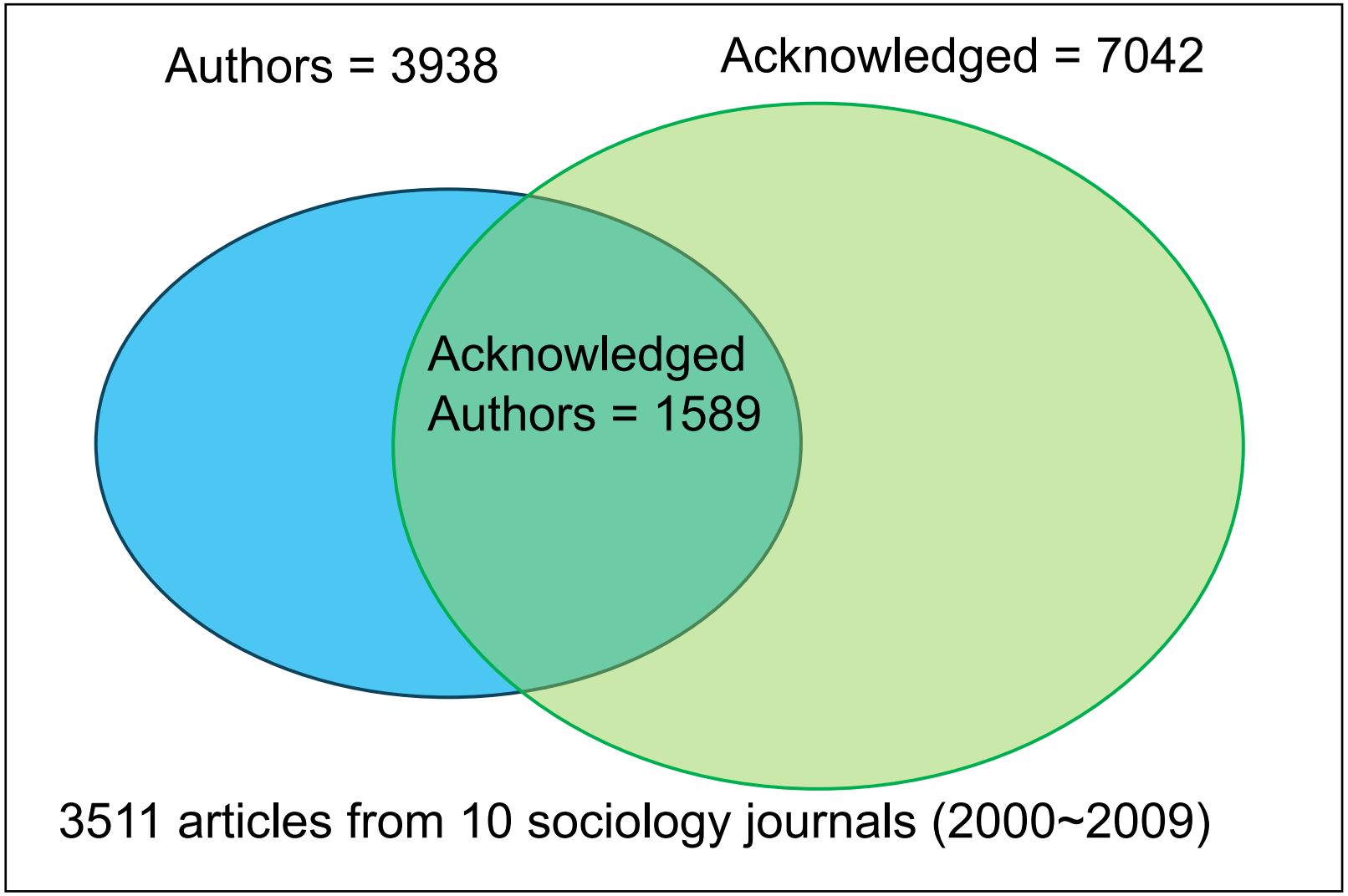
- Pet project during grad school
- Constructed a network of scholarly acknowledgements

A BEHAVIORAL MODEL OF RATIONAL CHOICE

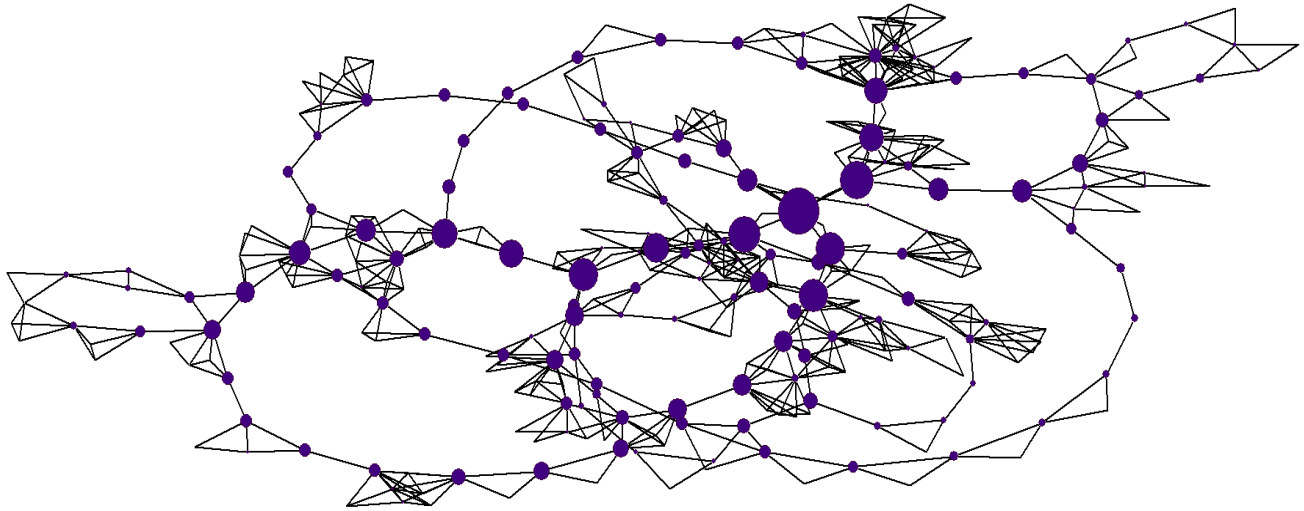
By HERBERT A. SIMON*

"The ideas embodied in this paper were initially developed in a series of discussions with **Herbert Bohnert, Norman Dalkey, Gerald Thompson, and Robert Wolfson** during the summer of 1952. These collaborators deserve a large share of the credit for whatever merit this approach to rational choice may possess."

New Findings Open New Uses of Data



New Findings Open New Uses of Data



Created some interesting visualizations

Ran some exploratory data analyses

Presented some findings at conferences

New Findings Open New Uses of Data

Main finding: Authors who have few coauthors tend to acknowledge more scholars

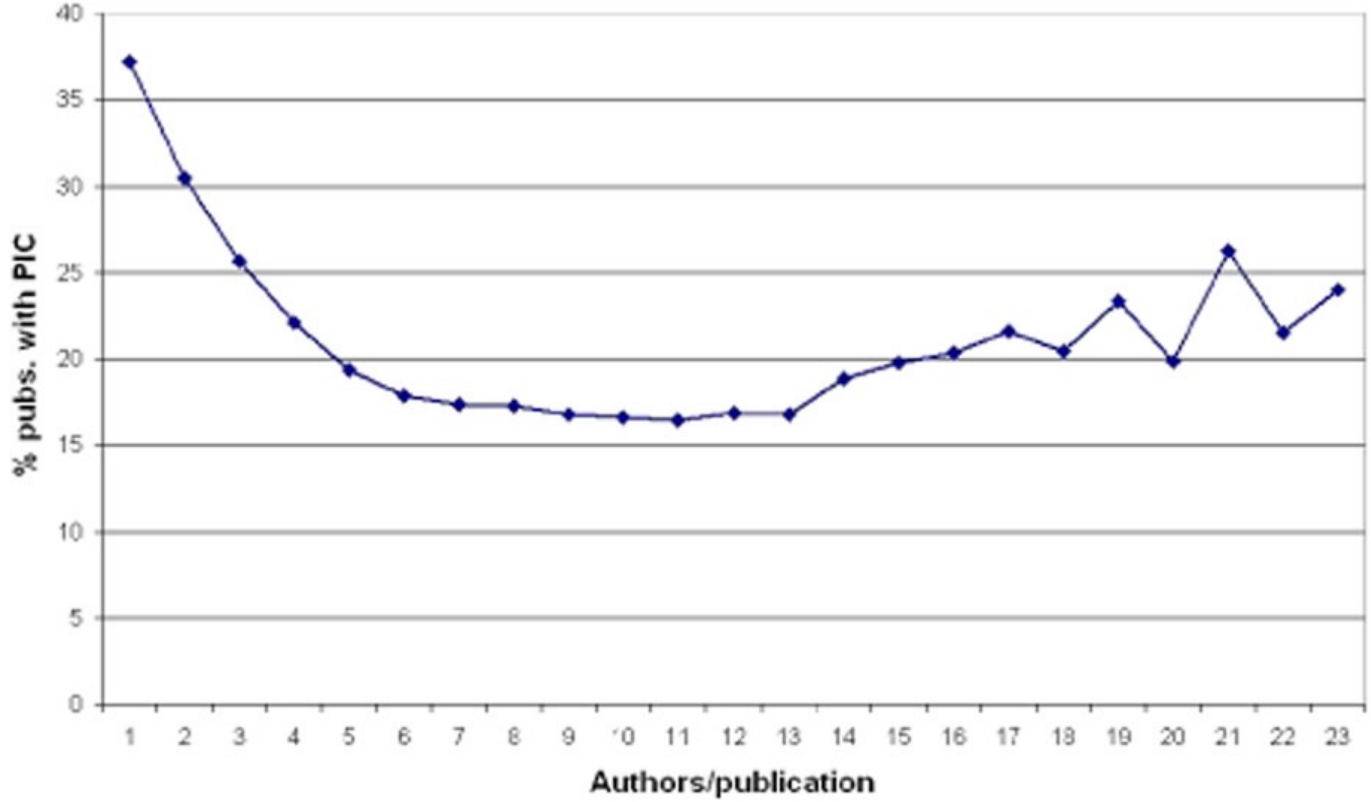
Table 2. OLS Regression for ego's average number of coauthors per publication

	Model 1			Model 2		
	B	S.E.		B	S.E.	
Intercept	1.46	0.03	***	0.79	0.06	***
Scholars Acknowledged by Ego (Log(outdegree))	-0.21	0.03	***	-0.13	0.03	***
Scholars Who Acknowledged Ego (Log(indegree))	-0.13	0.02	***	-0.02	0.03	
Mentor Index	-0.06	0.01	***	-0.12	0.01	***
Number of Publications	0.10	0.02	***	0.15	0.02	***
Years Since PhD				0.003	0.001	***
Female				0.02	0.05	
N	3938			1402		
Adjusted R^2	0.048			0.073		

(This was not news: already published using even better data)

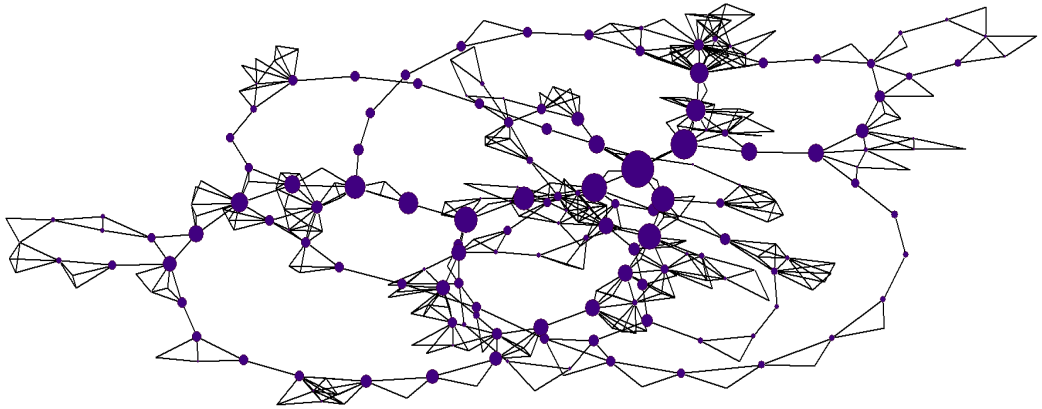
New Findings Open New Uses of Data

Publications with fewer authors are more likely to acknowledge scholars



Costas and van Leeuwen, 2012

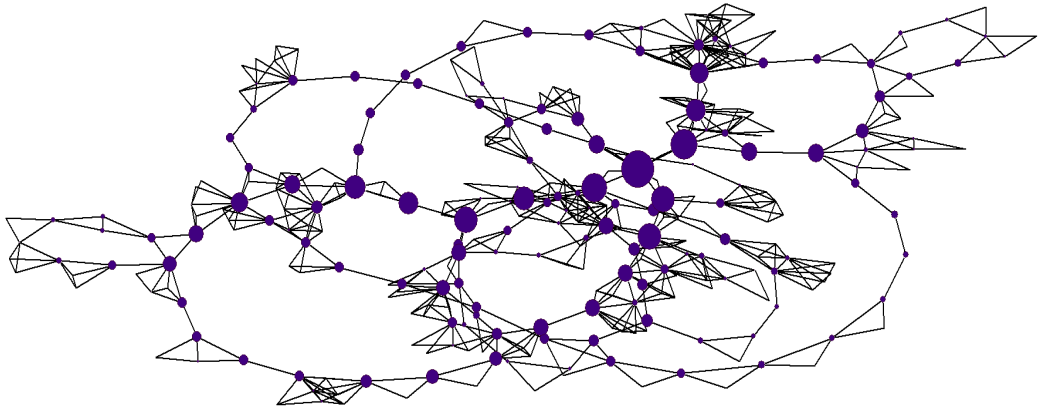
New Findings Open New Uses of Data



Acknowledgements, so what?

- What do acknowledgements represent?
- What questions can be addressed with them?
- Why should anybody care?

New Findings Open New Uses of Data



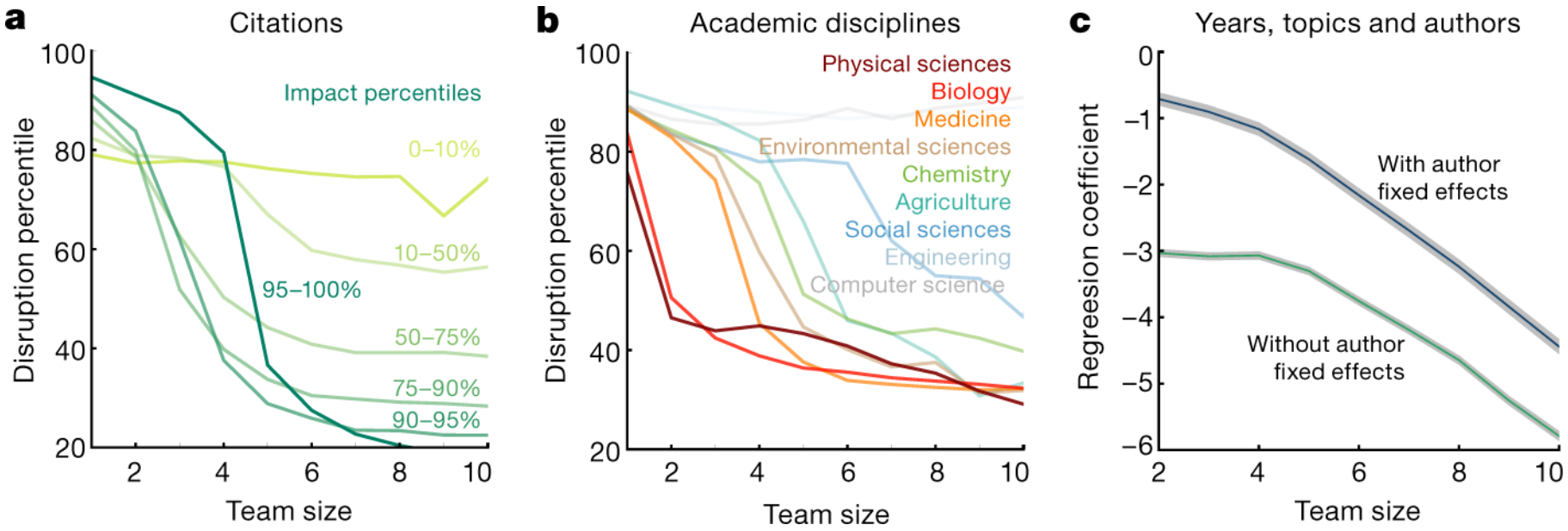
I had no convincing answers

So, pet project hibernated in backup drive for >10 years

New Findings Open New Uses of Data

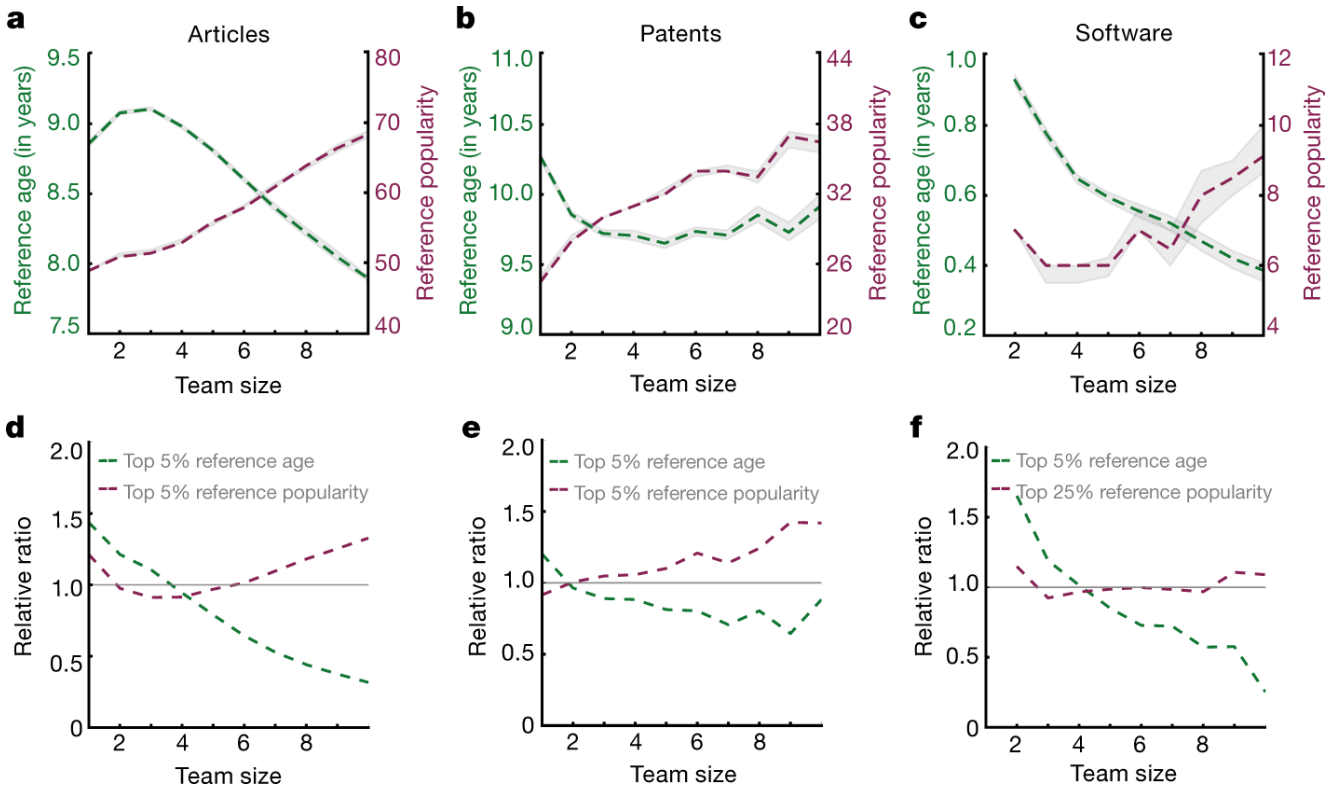
10+ years later:

- A study published in *Nature* showed that smaller academic teams (coauthors) tend to disrupt while larger teams develop



Wu et al., 2019

New Findings Open New Uses of Data



My question: Why are small teams more disruptive?

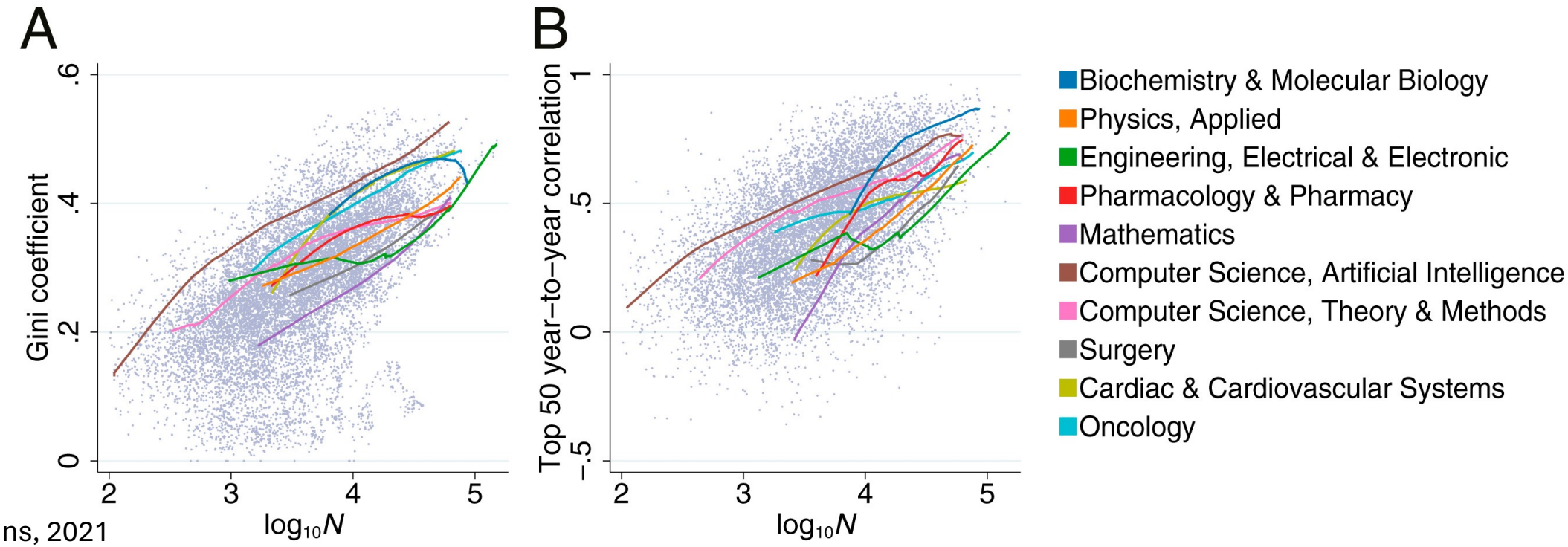
Paper found that:

Large teams search shallow

Small teams search deep

Concentration of Scholarly Attention

As a field grows, deep search becomes tenuous
- information overflow: too many articles to read
- people disproportionately pay attention to the top
- share of citations to top articles increases



Chu and Evans, 2021

New Findings Open New Uses of Data

If everyone looks to the small group of articles at the top,
Then where do disruptive ideas come from?

Hypothesis: They come from informal academic relationships

- Old graduate school friends
- Hallway chats with colleague
- Seminars, conferences
- Tennis court
- Mixers
- Farmers market

New Findings Open New Uses of Data

Large teams internalize intellectual resources in the form of coauthorship (strong ties)

Small teams draw intellectual resources and ideas from informal networks (weak ties)

So, maybe **small teams disrupt because the structure of informal networks consisting of weak ties facilitates broad circulation of diverse ideas across subfields and disciplines**

Acknowledgement: Measure of Informal Support

A BEHAVIORAL MODEL OF RATIONAL CHOICE

By HERBERT A. SIMON*

"The ideas embodied in this paper were initially developed in a series of discussions with Herbert Bohnert, Norman Dalkey, Gerald Thompson, and Robert Wolfson during the summer of 1952. These collaborators deserve a large share of the credit for whatever merit this approach to rational choice may posses."

Diffusion of Ideas through Informal Ties?

Compared to the structure of formal collaboration (i.e., coauthorship),

Is the structure of informal interaction network better for novel ideas to spread throughout different fields and subfields?

Informal Support Network in Science

Coauthorship Network

Acknowledgement Network

brittle

cohesive

Acknowledgement network did have a more cohesive structure

*Structural Cohesion: The extent to which many nodes in a network are reachable by other nodes through multiple paths

New Findings Open New Uses of Data

Scientific databases do not index acknowledgements

Because no one seriously thought of it

- Collaborations are encoded in coauthorships
- Intellectual indebtedness is coded in citations

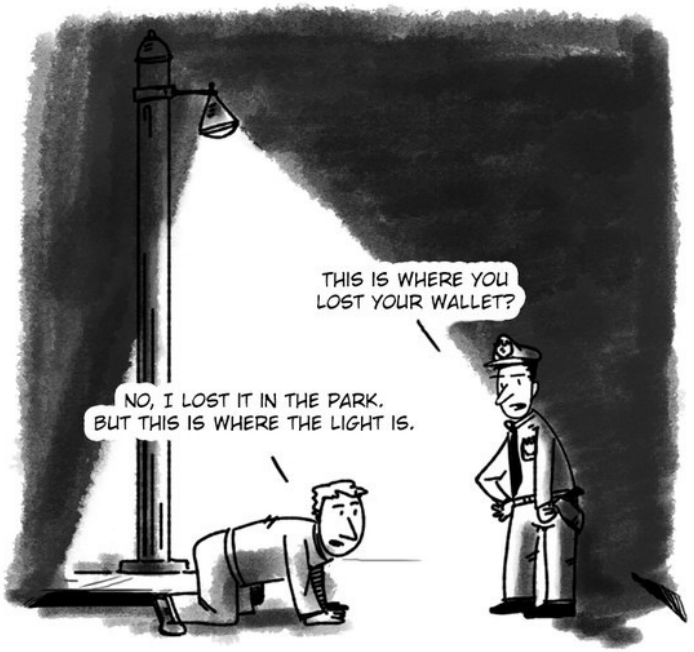
We search for answers where there is light

New answers require pointing theoretical light into the dark

But, disciplinary inertia and intellectual dogma can make this difficult

New Findings Open New Uses of Data

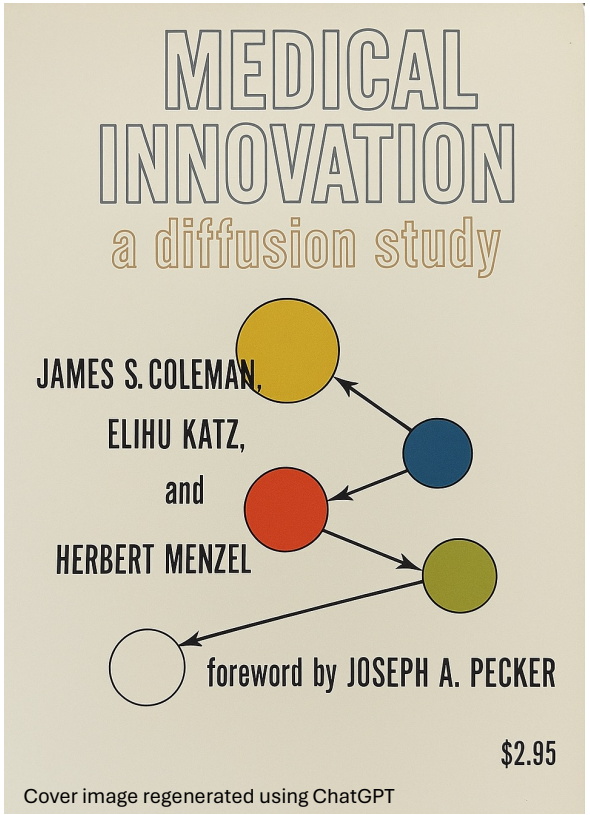
- New findings can stimulate new (applications of) theory
- And new ways to re-purpose old data



Challenges around Data Preservation

The Case of Medical Adoption Data

1966 → 1987



Cover image regenerated using ChatGPT

H1: Adoption through direct contacts

Social Contagion and Innovation: Cohesion versus Structural Equivalence[1]

Ronald S. Burt
Columbia University

Two classes of network models are used to reanalyze a sociological classic often cited as evidence of social contagion in the diffusion of technological innovation: *Medical Innovation*. Debate between the cohesion and structural equivalence models poses the following question for study: Did the physicians resolve the uncertainty of adopting the new drug through conversations with colleagues (cohesion) or through their perception of the action proper for an occupant of their position in the social structure of colleagues (structural equivalence)? The alternative models are defined, compared, and tested. Four conclusions are drawn: (*a*) Contagion was not the dominant factor driving tetracycline's diffusion. Where there is evidence of contagion, there is evidence of personal preferences at work.

AJS Volume 92 Number (May 1987):1287–1335

H2: Keeping up with the Joneses

The Case of Medical Adoption Data

New insights from 30-year old data (1950s → 1987)

New methods theory development and testing

The Case of Medical Adoption Data

New insights from 30-year old data (1950s → 1987)

New methods theory development and testing

Replication was costly and cumbersome

detailed comments on the manuscript. A copy of the data discussed here can be obtained by requesting Technical Report no. TR3, "The *Medical Innovation* Network Data," from Columbia University's Center for the Social Sciences. A copy of the microcomputer network-analysis software that includes the procedures used here to generate adoption norms can be obtained, with a program manual, by requesting Technical Report no. TR2, "STRUCTURE, Version 3.0" from the center. Enclose a check ($25 for no. TR2, $10 for no. TR3) made out to the Research Program in Structural Analysis with your request to help defray duplication and mailing costs. Data and software are sent on diskettes in DOS 360K format for an IBM microcomputer. Requests for reprints should be sent to Ronald S. Burt, Department of Sociology, Columbia University, New York, New York 10027.

Rediscovering The Value of Old Data

- **SBP-BRiMS 2025 Data Challenge**

Predict the adoption times of Midwestern physicians in the 1950s

Data: Medical Innovation Network Data

Throw your hat into the ring!