

Creative Destruction

The structural
consequences
of scientific
curation

Peter McMahan

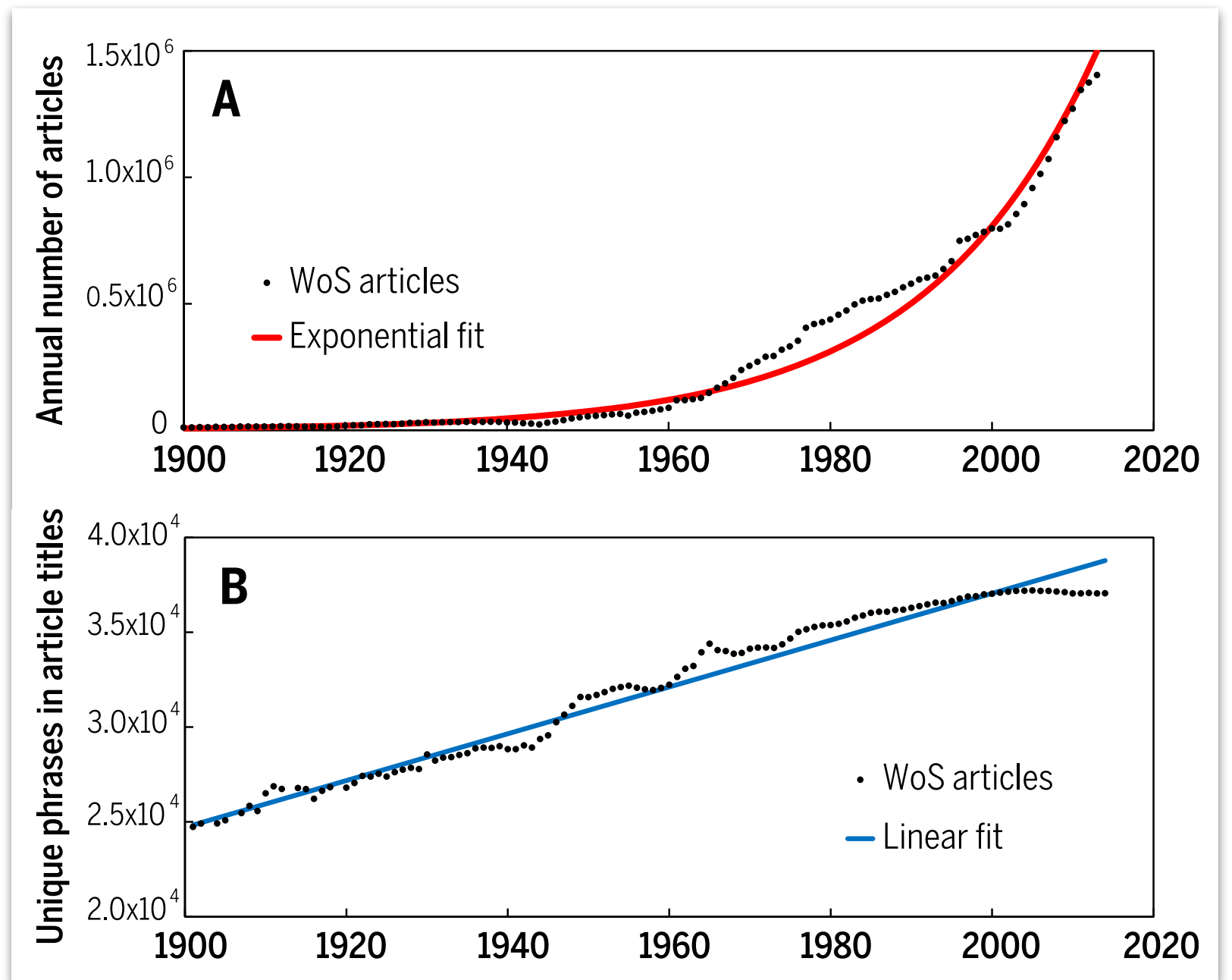
peter.mcmahan@mcgill.ca

Daniel A. McFarland

mcfarland@stanford.edu

Knowledge Growth

Keeping up with scientific output



Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, et al. "Science of Science." *Science* 359, no. 6379 (March 2, 2018).

Knowledge Synthesis

What do reviews *do*?

Curated *summaries* of the important concepts, innovations, and debates within a scholarly area.

“inform interested readers who have limited knowledge of [a] topic, whether students new to the field or seasoned researchers from other domains.”

(Freeman and Jeanloz 2015)

Summarization as translation

Synthesis of knowledge is not neutral; “packing down” is a *creative* act.

∴ ***Black boxes***

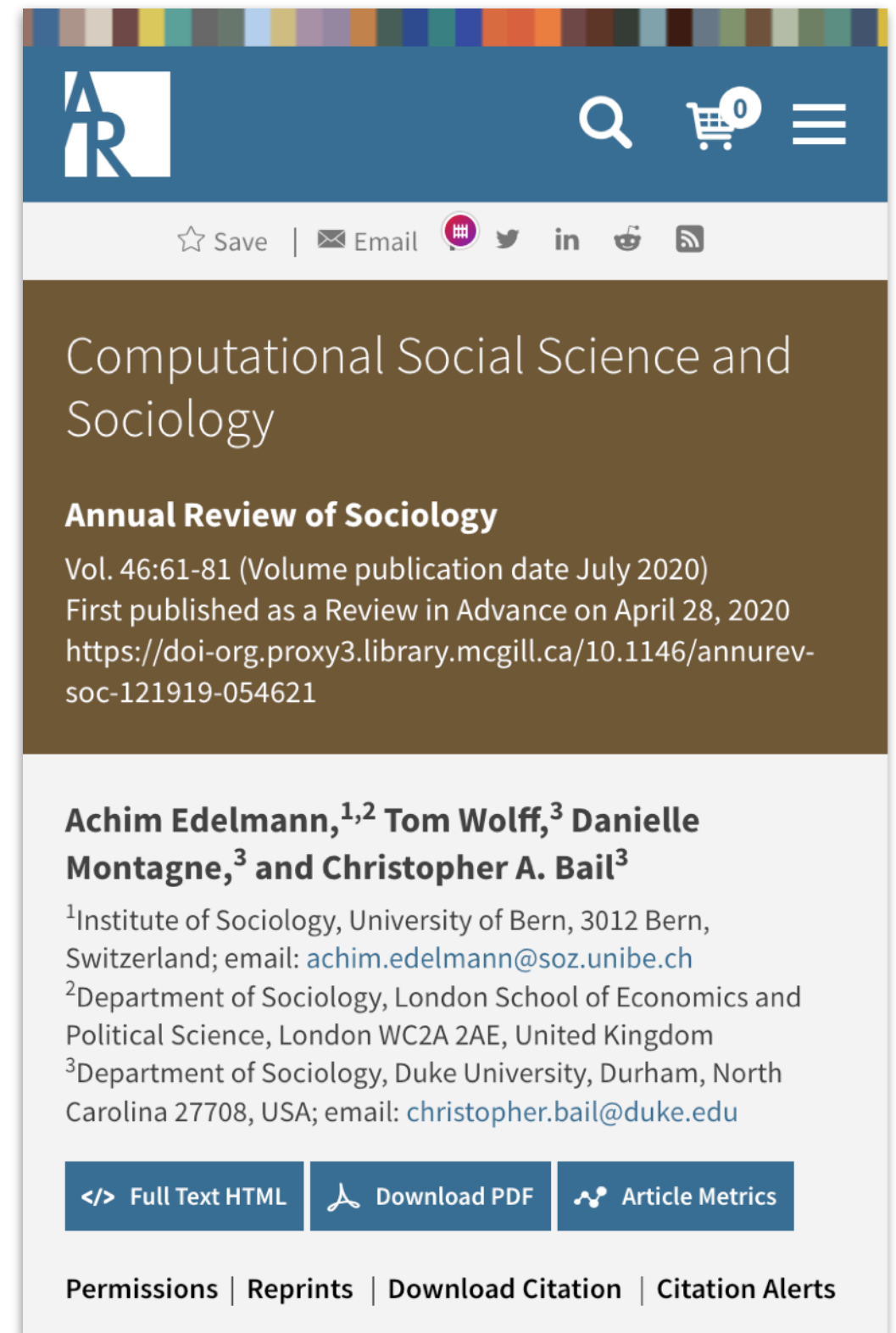
Whitley

∴ ***Boundary objects***

Latour; Star and Griesemer

∴ ***Exemplars***

Bourdieu; Kuhn; Frickel and Gross



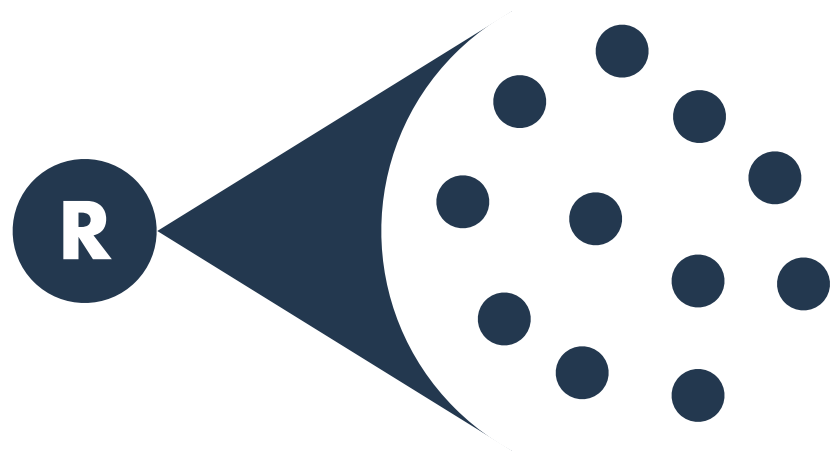
Levels of analysis

How do review articles shape scientific knowledge?

(Two levels of analysis)

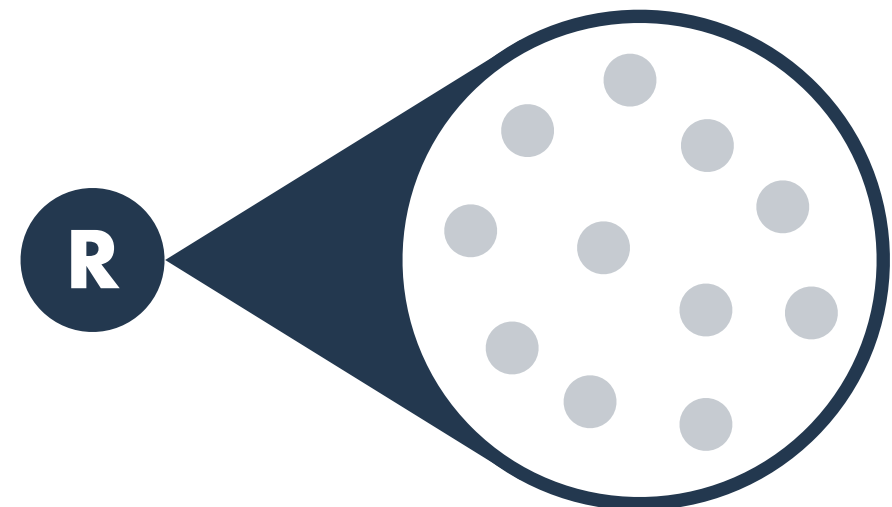
Article

How do reviews affect the **popularity** of the **articles** they cite?



Subfield

How do reviews affect the **structure** of the **subfield** they cite?



Citation Data

Articles & Citations

Web of Science (WoS)

Articles published 1990–2016

Review articles

54 *Annual Review* journals

Omits many ‘review-like’ publications

Subsample

Top 50 journals cited by each *Annual Review* journal

About 5.9 million publications in about 1,200 journals

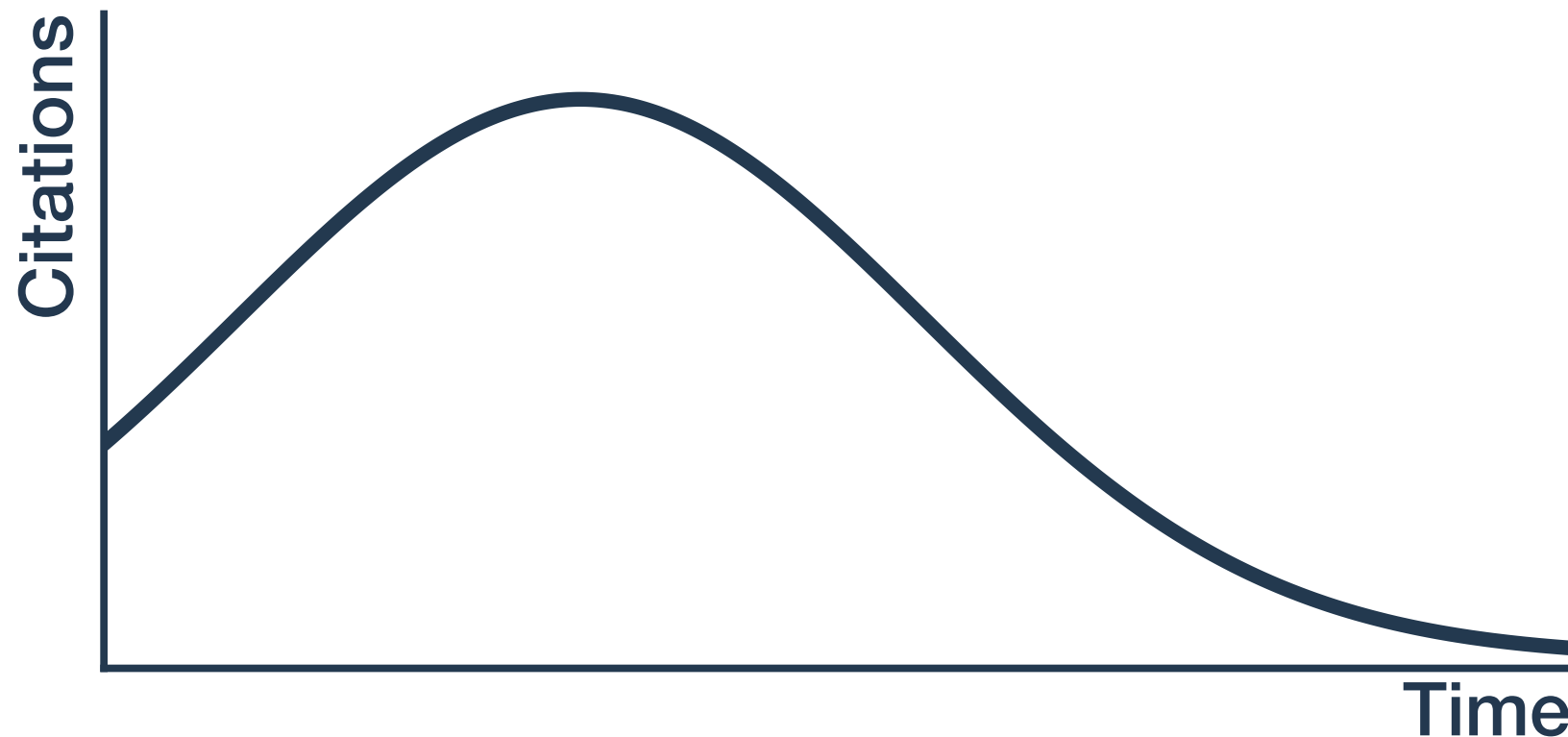
Promote or poach?

How do reviews affect the **popularity** of the **articles** they cite?

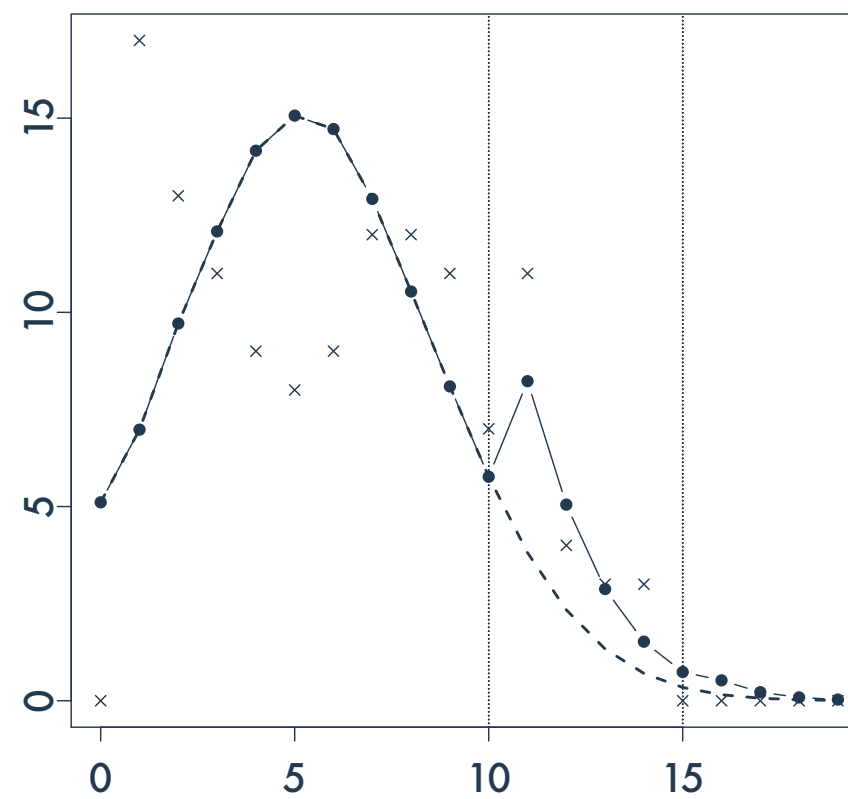
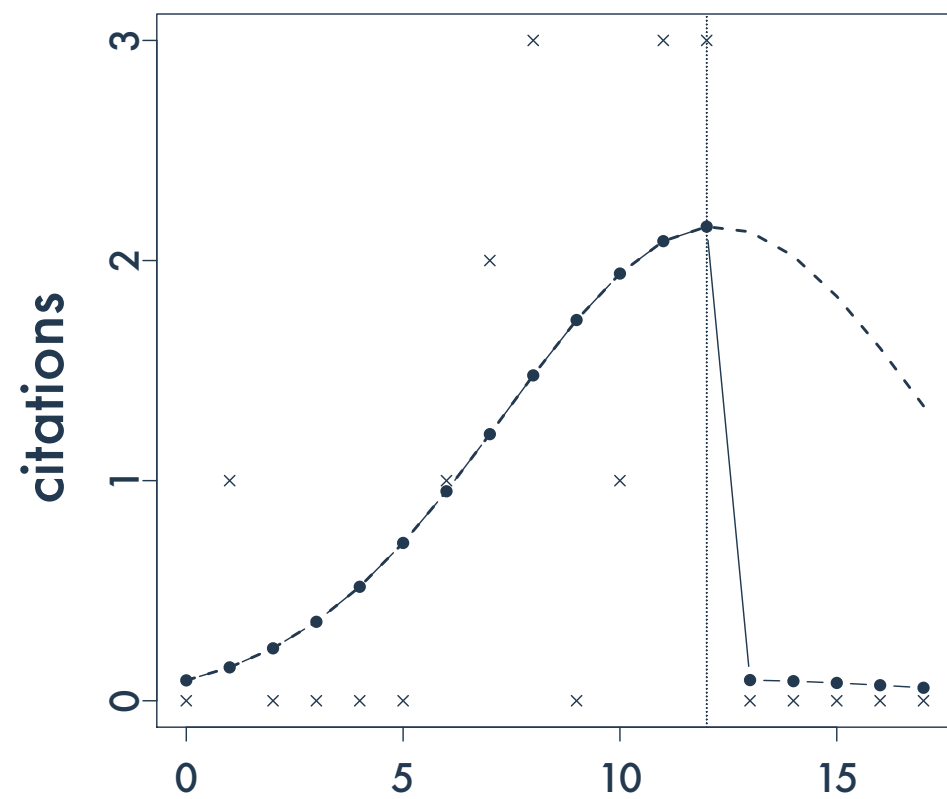
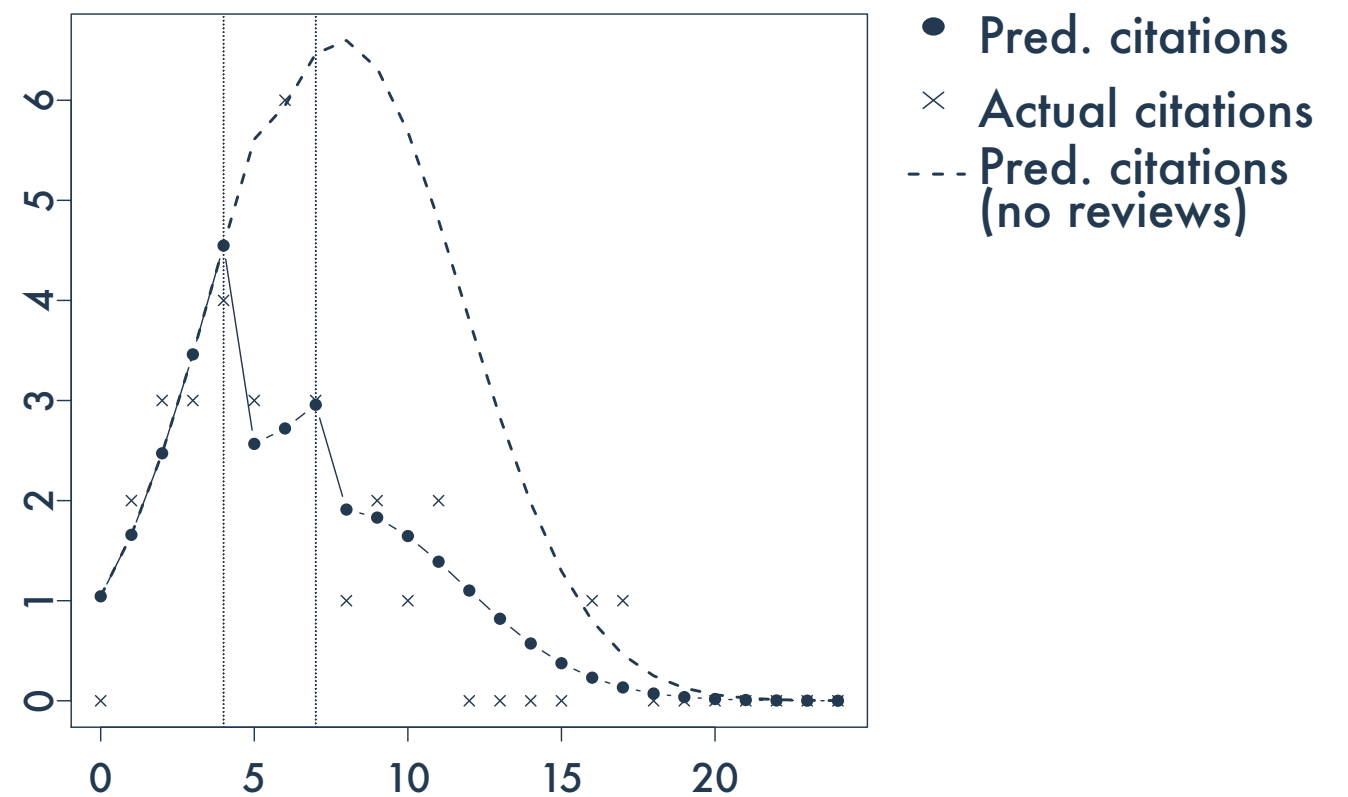
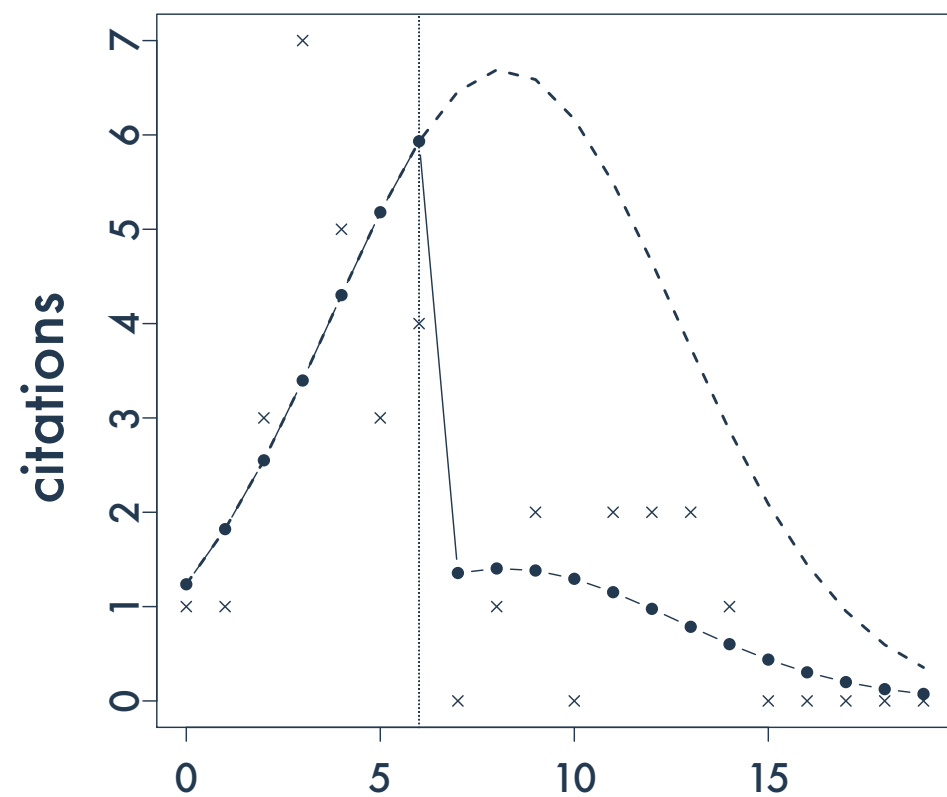
Predicting articles' "citation lifecycle"

Multilevel model accounting for time since publication, discipline, journal, and article-level effects

Look for discontinuities in citation curve associated with inclusion in a review



Promote or poach?

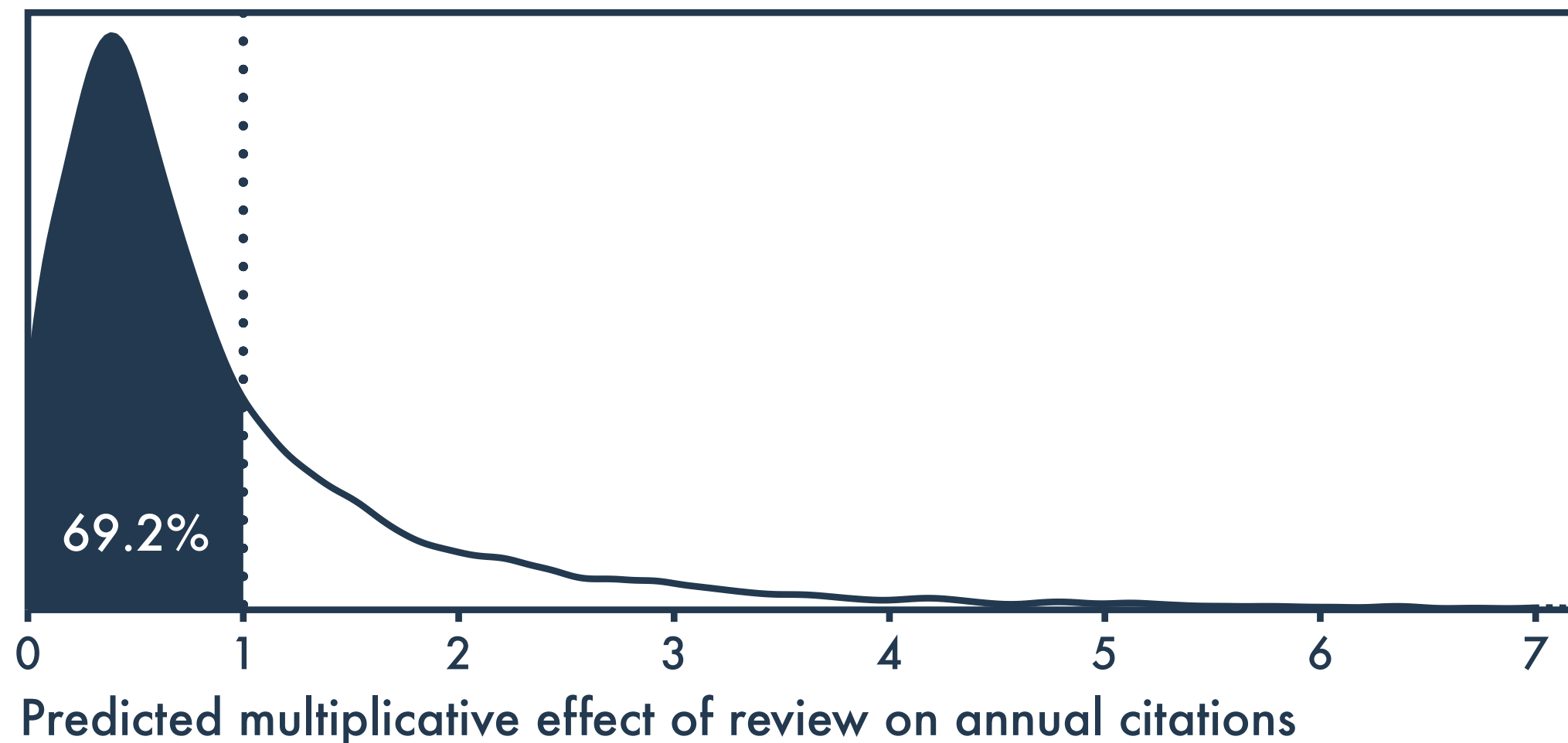


years since publication

Promote or poach?

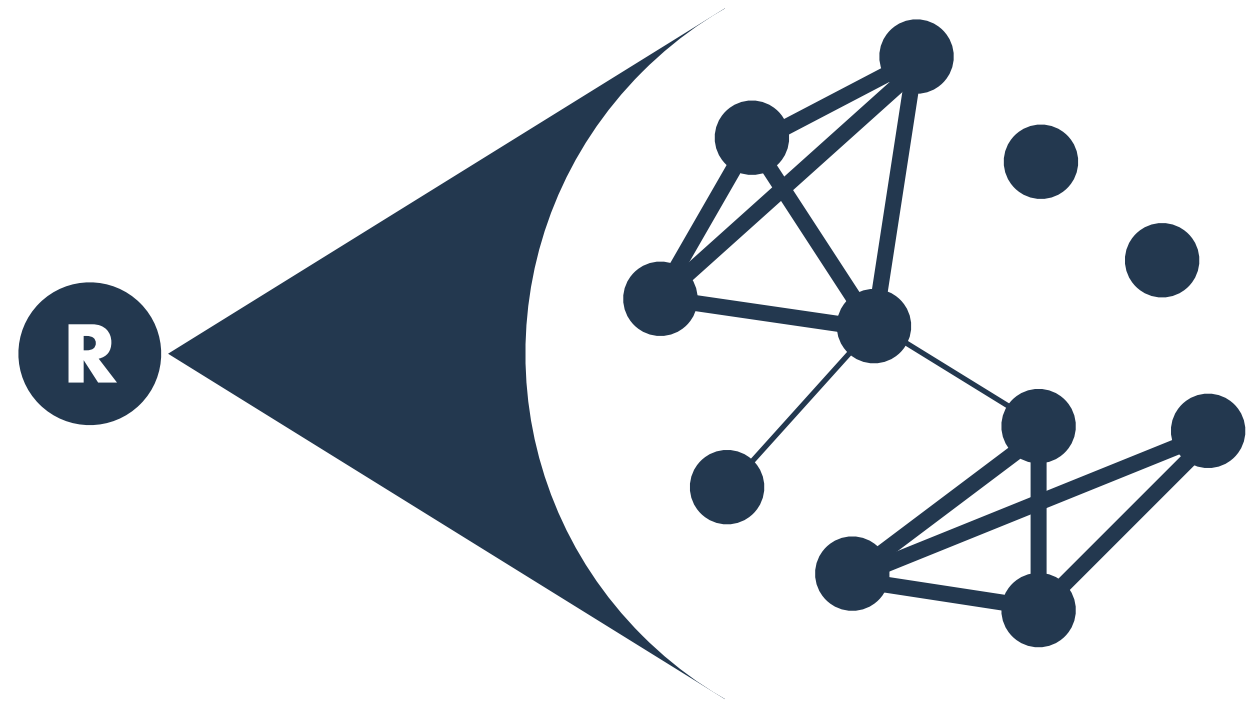
The median reviewed article can expect **38.3% fewer citations** in every subsequent year.

However, a handful of cited articles will receive a substantial **boost** to their expected future citations.



Subfield structure

How do reviews affect the **structure** of the **subfield** they cite?

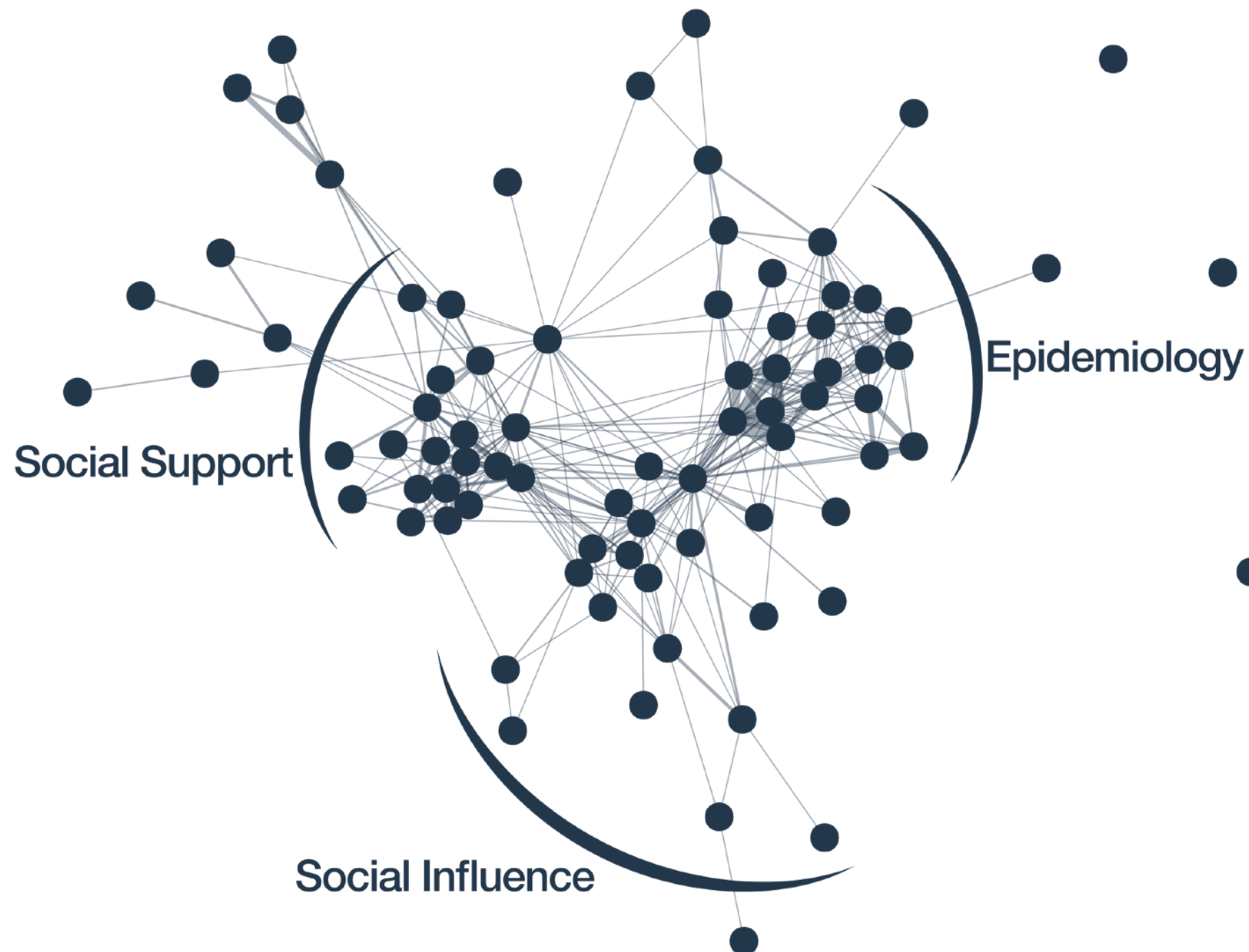


Cocitation measures the shared attention (positive *and* negative) that articles receive.

Cocitation focuses on the way research is *used* in scholarly discourse.

Cocitation Networks

Smith, Kirsten P., and Nicholas A. Christakis.
2008. "Social Networks and Health." Annual
Review of Sociology 34 (1): 405–29.



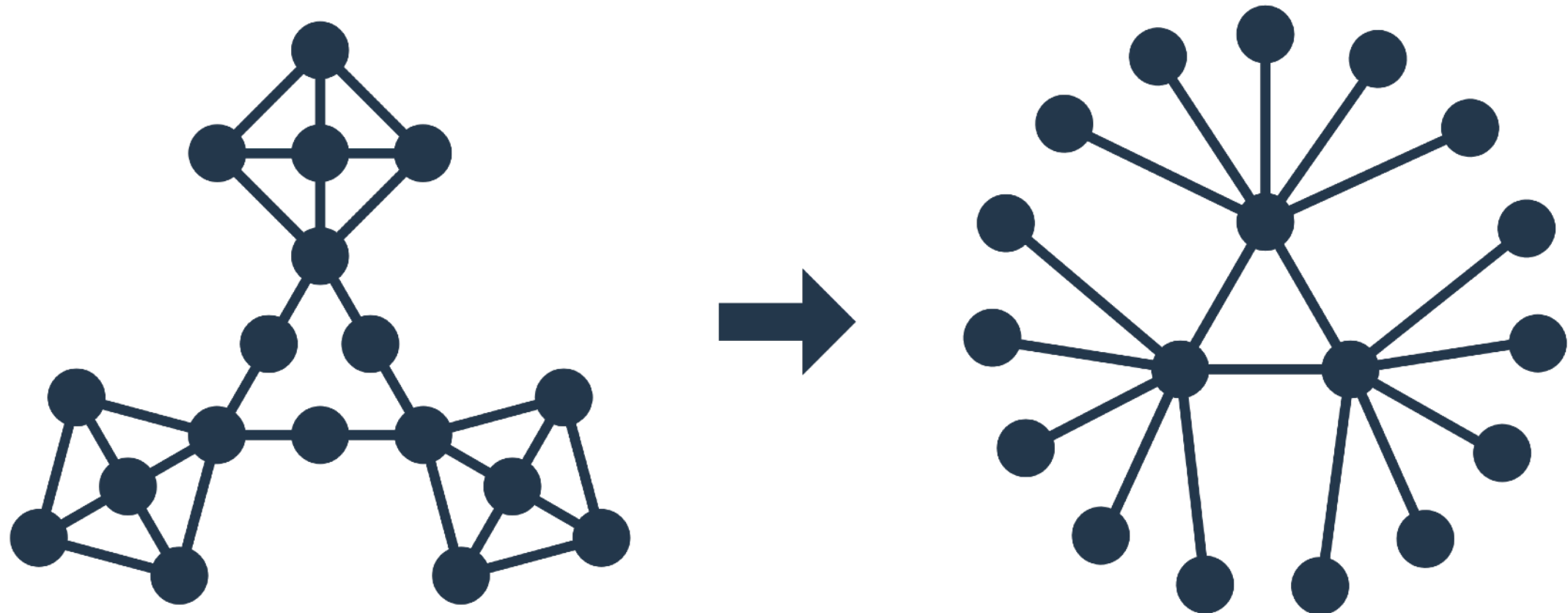
Structural change

Structural effects of review

We compare cocitation structure over seven years *before* and *after* the publication of a review

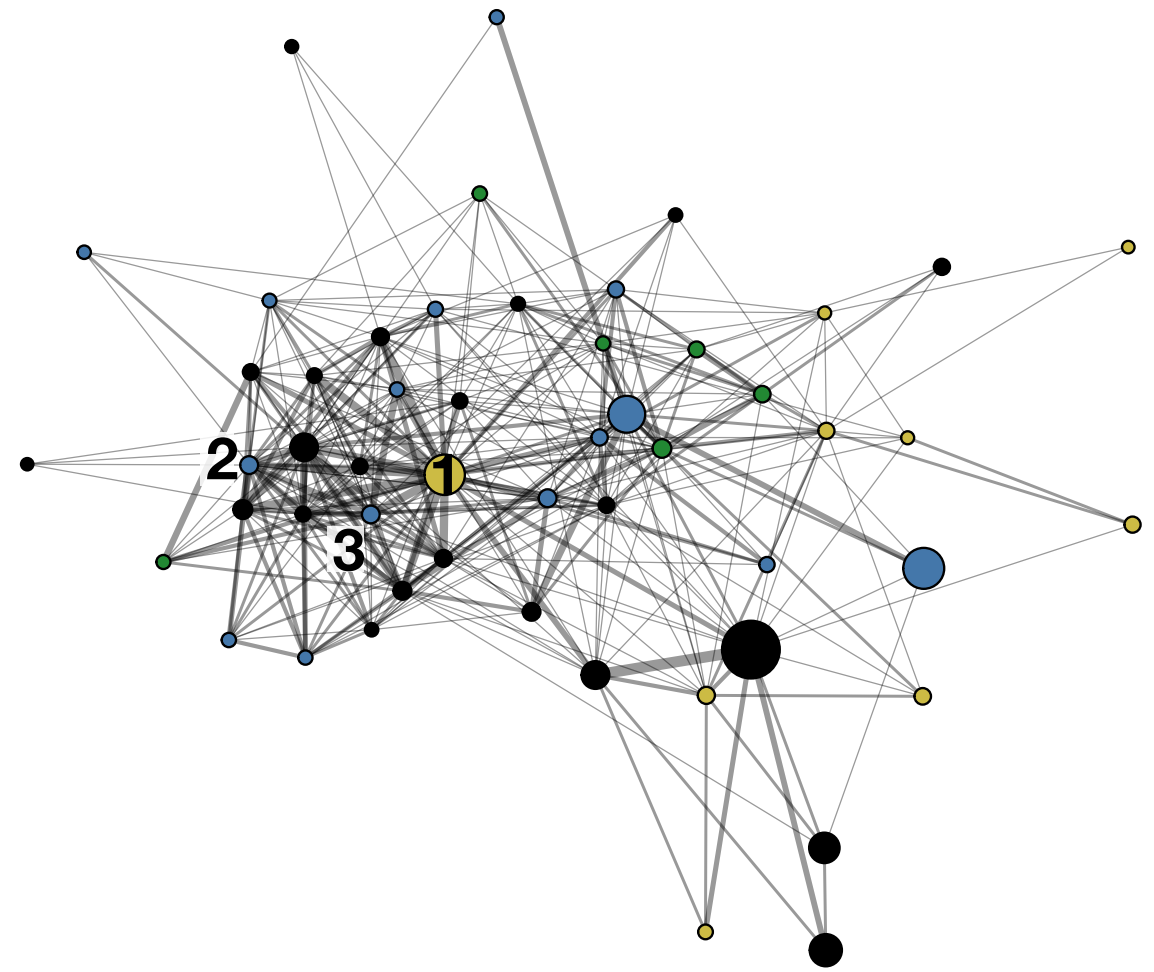
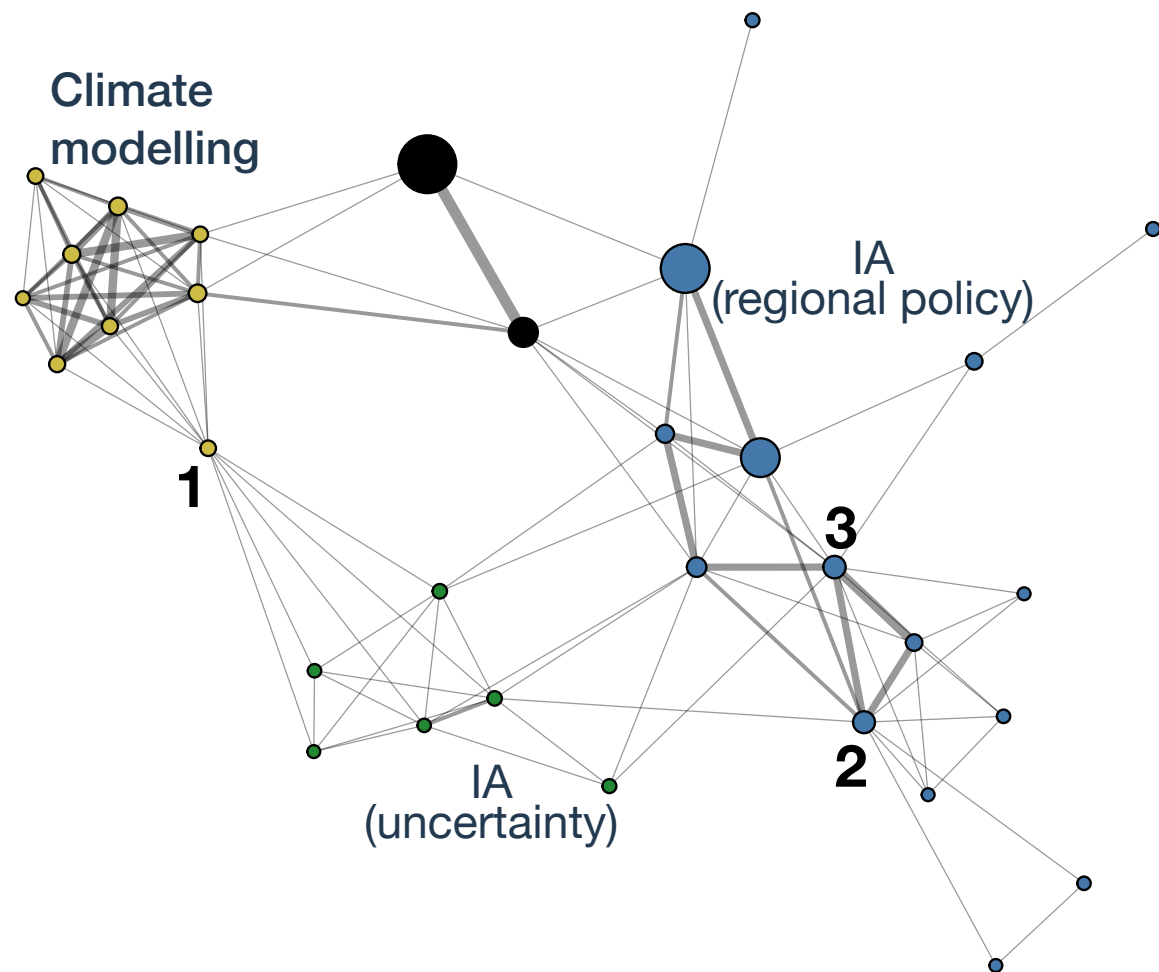
Matched sample controls for characteristics of emerging scientific subfields

Reviewed specialties have significantly *shorter* path lengths and *fewer* cohesive sub-clusters



Restructuring knowledge

Integrated assessment models of global climate change
(Parson and Fisher-Vanden 1997)



Summary

Questions answered:

Review articles draw attention away from the articles they cite

most cited work sees significant decrease in future citations

Reviews create centralized research communities

scholarly attention turns away from intra-group relations, toward exemplary hubs

Reviews promote bridging research

scientific perception refocusses on research that links distinct branches

Crystallization of research specialties

Moment of legitimization for research topic


independent research programs cohere into unified whole

Necessary form of knowledge curation

Need to simplify findings in an ever-expanding body of knowledge

Details are lost in order to incorporate into the larger body of scientific knowledge

Thank you & Questions



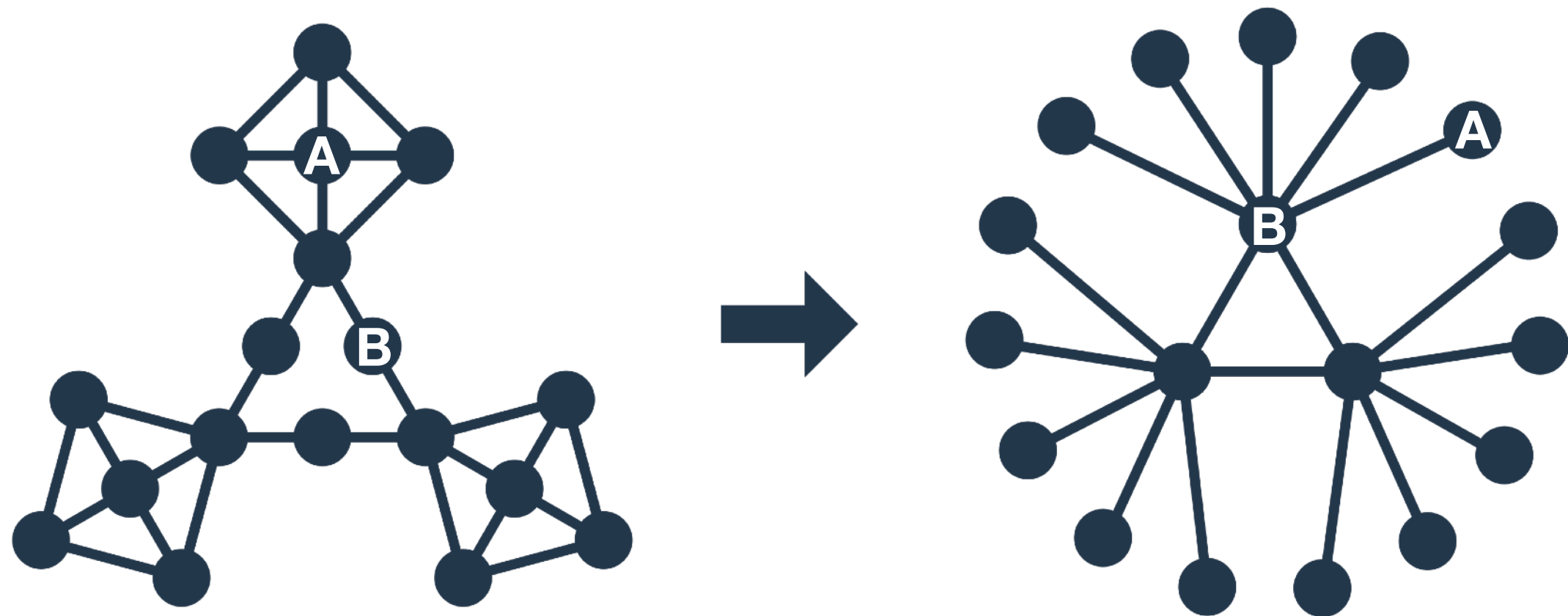
Peter McMahan

peter.mcmahan@mcgill.ca

Slides available at

<http://petermcmahan.com>

Centering exemplars



Which articles do reviews elevate as *exemplars* within a scientific specialization?

Future conversation is *not* centered around already-central research (**A**)

Highly cited and central work will likely be pushed toward the periphery after a review is published

Exemplars chosen from *bridging articles* (**B**)

Research that connects contrasting communities experiences the biggest boost in centrality

Article-level model

$$\mathbf{C}_{ijt} \sim \text{NBinom}(\lambda_{ijt}, d)$$

$$\log(\lambda_{ijt}) = \beta_{0ij} + \beta_{1i}\mathbf{t} + \beta_{2i}\mathbf{t}^2 + \beta_{3ij}\mathbf{Reviews}_{ij(t-1)} + \beta_{4j}\mathbf{Reviewability}_{ij(t-1)}$$

$$\beta_{0ij} = \gamma_{00j} + \gamma_{01}\mathbf{NumCites}_i + \gamma_{02}\mathbf{WoS}_i + \gamma_{02}\mathbf{Pages}_i + \eta_{0i}$$

$$\beta_{1i} = \gamma_{10} + \eta_{1i}$$

$$\beta_{2i} = \gamma_{20} + \eta_{2i}$$

$$\beta_{3ij} = \gamma_{30j} + \gamma_{31}\mathbf{NumCites}_i + \gamma_{32}\mathbf{WoS}_i + \gamma_{32}\mathbf{Pages}_i + \eta_{3i}$$

$$\beta_{4j} = \gamma_{40j}$$

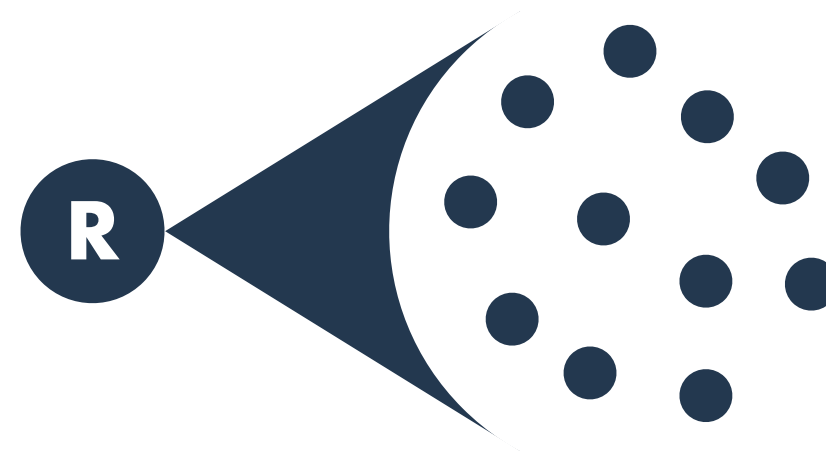
$$\gamma_{00j} = a_{00} + v_{0j}$$

$$\gamma_{30j} = a_{30} + a_{31}\mathbf{IF}_j + a_{32}\mathbf{Sub1}_j + \cdots + a_{37}\mathbf{Sub6}_j$$

$$\gamma_{40j} = a_{40} + v_{4j}$$

Article-level model

How do reviews affect
the **popularity** of the
articles they cite?



Multilevel Model



Article-level model

Population-level effects	Estimate	Std Error	(Z value)
(Intercept)	-1.48926	0.05147	(-28.94)
t	-3.47962	0.01326	(-262.48)
t^2	-2.18744	0.00645	(-339.02)
Reviews	-0.1192	0.03639	(-3.28)
Reviewability	0.2862	0.01189	(24.07)
Impact Factor	0.28647	0.01747	(16.4)
Pages	-0.08607	0.0125	(-6.89)
Citations	0.38025	0.00394	(96.53)
Prop. WoS	1.22897	0.01415	(86.85)
Subj: Nat. Sci.	0.05144	0.04757	(1.08)
Subj: Eng. & Tech	-0.02725	0.05142	(-0.53)
Subj: Med. & Health	0.1886	0.05101	(3.7)
Subj: Agr. Sci.	-0.1962	0.07352	(-2.67)
Subj: Soc. Sci.	-0.10433	0.05527	(-1.89)
Subj: Humanities	-0.60872	0.10442	(-5.83)
Reviews × Impact Factor	0.02643	0.00633	(4.18)
Reviews × Pages	-0.01779	0.00945	(-1.88)
Reviews × Citations	0.0425	0.00726	(5.86)
Reviews × Prop. WoS	-0.52196	0.04075	(-12.81)
Reviews × Subj: Nat. Sci.	-0.00867	0.02895	(-0.3)
Reviews × Subj: Eng. & Tech	-0.06128	0.03219	(-1.9)
Reviews × Subj: Med. & Health	0.11775	0.02759	(4.27)
Reviews × Subj: Agr. Sci.	-0.10006	0.05467	(-1.83)
Reviews × Subj: Soc. Sci.	0.10427	0.0354	(2.95)
Reviews × Subj: Humanities	-0.05329	0.08834	(-0.6)

Multilevel negative-binomial model estimates. Coefficients reflect expected effects on total yearly citations received. All covariates except *Reviews* standardized to have zero mean and unit standard deviation.

Group-level effects	Level	Std Dev
(Intercept)	Article	10.0341
t	Article	22.7016
t^2	Article	5.3351
Reviews	Article	2.1927
(Intercept)	Journal	0.2743
Reviewability	Journal	0.1265

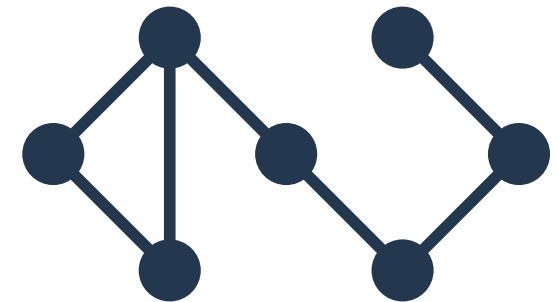
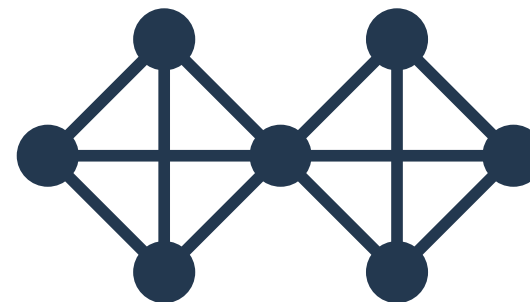
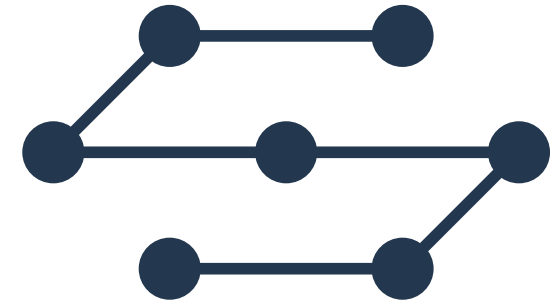
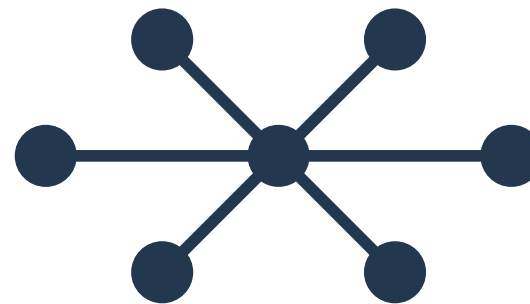
Cocitation Structure

Number of “steps” to
traverse a community
(six degrees of separation)

Average path length

← SHORTER

LONGER →



Tendency
toward tightly
connected
clusters of
research

Clustering

↑ LESS

↓ MORE

Structural model

Results describing structural changes to reference clusters. The left two columns of results show estimates using a weighted, 5% sample of the full data, and the remaining columns show the same results estimated on the propensity–score matched sample. The final two columns use co-citation networks that exclude articles citing the root article. Values in parentheses represent 95% credible intervals on all estimates.

$$\begin{pmatrix} \mathbf{AvgPathLength}_i^\Delta \\ \mathbf{Clustering}_i^\Delta \end{pmatrix} \sim \text{MVNorm} \left(\begin{pmatrix} \mu_{1i} \\ \mu_{2i} \end{pmatrix}, \Sigma \right)$$

$$\begin{aligned} \mu_{ji} = & \beta_{j0} + \beta_{j1} \mathbf{Review}_i + \\ & \beta_{j2} (\# \mathbf{Vertices}_i) + \beta_{j3} (\# \mathbf{Vertices}_i)^2 + \\ & \beta_{j4} (\mathbf{Density}_i^0) + \beta_{j5} (\mathbf{Density}_i^0)^2 + \\ & \beta_{j6} \mathbf{AvgPathLength}_i^0 + \beta_{j7} \mathbf{Clustering}_i^0 \end{aligned}$$

Structural model

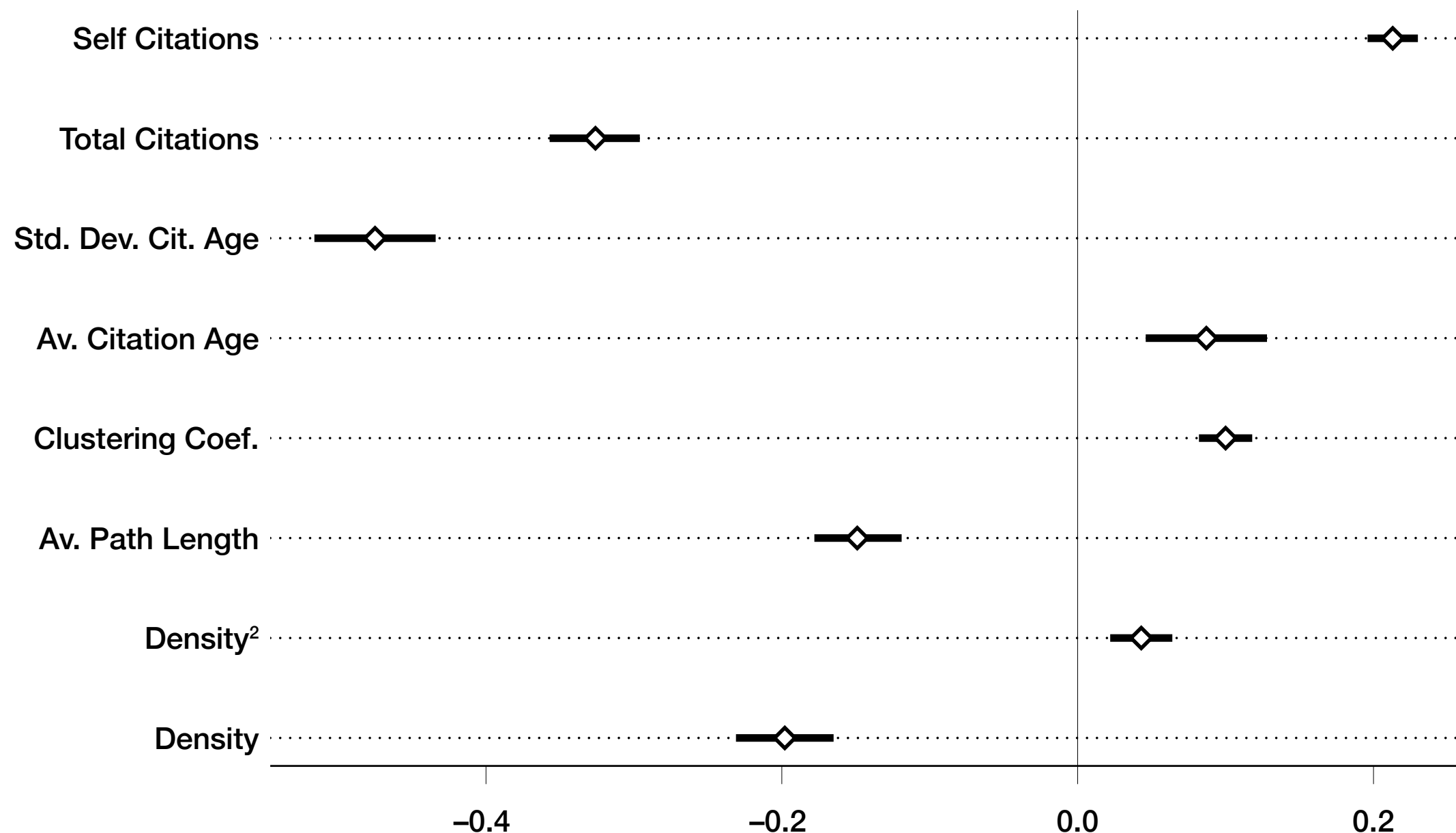
	5% sample		Matched sample	
	AvgPathLength ^Δ	Clustering ^Δ	AvgPathLength ^Δ	Clustering ^Δ
(Intercept)	-0.008 (-0.010, -0.005)	-0.032 (-0.035, -0.029)	-0.002 (-0.027, 0.023)	0.012 (-0.008, 0.031)
Review	-0.145 (-0.164, -0.125)	-0.208 (-0.228, -0.188)	-0.109 (-0.136, -0.082)	-0.070 (-0.090, -0.049)
#Vertices	-0.135 (-0.138, -0.132)	-0.026 (-0.029, -0.023)	-0.207 (-0.222, -0.192)	0.010 (-0.001, 0.021)
#Vertices²	0.008 (0.007, 0.009)	0.006 (0.005, 0.008)	0.044 (0.031, 0.058)	0.005 (-0.006, 0.016)
Density⁰	-0.105 (-0.107, -0.102)	-0.083 (-0.085, -0.081)	-0.107 (-0.125, -0.089)	-0.034 (-0.048, -0.020)
(Density⁰)²	0.002 (-0.000, 0.003)	0.027 (0.025, 0.028)	0.012 (0.001, 0.023)	0.017 (0.009, 0.026)
AvgPathLength⁰	-0.638 (-0.640, -0.636)	0.001 (-0.001, 0.003)	-0.625 (-0.639, -0.610)	0.012 (0.001, 0.022)
Clustering⁰	0.033 (0.031, 0.035)	-0.620 (-0.622, -0.617)	0.026 (0.010, 0.043)	-0.817 (-0.829, -0.804)
Res. Std. Dev.	0.778 (0.776, 0.779)	0.803 (0.802, 0.804)	0.789 (0.780, 0.799)	0.597 (0.590, 0.604)
Res. Cor.	0.070 (0.062, 0.072)		0.186 (0.169, 0.202)	

Results describing structural changes to reference clusters. The left two columns of results show estimates using a weighted, 5% sample of the full data, and the remaining columns show the same results estimated on the propensity–score matched sample. The final two columns use co-citation networks that exclude articles citing the root article. Values in parentheses represent 95% credible intervals on all estimates.

Matching model

$\text{Review}_i \sim \text{Bernoulli}(\text{logit}^{-1}(\mu_i))$

$$\begin{aligned}\mu_i = & \beta_0 + \beta_1(\#\mathbf{Vertices}_i) + \beta_2(\#\mathbf{Vertices}_i)^2 + \beta_3(\mathbf{Density}_i^0) + \beta_4(\mathbf{Density}_i^0)^2 + \\ & \beta_5\mathbf{AvgPathLength}_i^0 + \beta_6\mathbf{Clustering}_i^0 + \beta_7\mathbf{AvgCiteAge}_i + \\ & \beta_8\mathbf{StdDevCiteAge}_i + \beta_9\mathbf{TotalCites}_i + \beta_{10}\mathbf{SelfCites}_i\end{aligned}$$



Matching model

$\text{Review}_i \sim \text{Bernoulli}(\text{logit}^{-1}(\mu_i))$

$$\mu_i = \beta_0 + \beta_1(\#\mathbf{Vertices}_i) + \beta_2(\#\mathbf{Vertices}_i)^2 + \beta_3(\mathbf{Density}_i^0) + \beta_4(\mathbf{Density}_i^0)^2 + \beta_5\mathbf{AvgPathLength}_i^0 + \beta_6\mathbf{Clustering}_i^0 + \beta_7\mathbf{AvgCiteAge}_i + \beta_8\mathbf{StdDevCiteAge}_i + \beta_9\mathbf{TotalCites}_i + \beta_{10}\mathbf{SelfCites}_i$$

Coefficient	Estimate	95% cred. int.	
(Intercept)	-5.293	(-5.336	-5.250)
$\#\mathbf{Vertices}$	1.230	(1.186	1.274)
$\#\mathbf{Vertices}^2$	-0.097	(-0.108	-0.086)
$\mathbf{Density}^0$	-0.198	(-0.231	-0.165)
$(\mathbf{Density}^0)^2$	0.043	(0.022	0.064)
$\mathbf{AvgPathLength}^0$	-0.149	(-0.178	-0.119)
$\mathbf{Clustering}^0$	0.100	(0.082	0.118)
$\mathbf{AvgCiteAge}$	0.087	(0.046	0.128)
$\mathbf{StdDevCiteAge}$	-0.475	(-0.516	-0.434)
$\mathbf{TotalCites}$	-0.326	(-0.357	-0.296)
$\mathbf{SelfCites}$	0.213	(0.196	0.230)

Exemplars model

$$\mathbf{EVCentrality}_{ij}^{\Delta} \sim \text{Normal}(\mu_{ij}, \sigma_2)$$

$$\begin{aligned} \mu_{ij} = & \beta_{1j} \mathbf{Citations}_{ij}^0 + \beta_{2j} \mathbf{EVCentrality}_{ij}^0 \\ & + \beta_{3j} \mathbf{Transitivity}_{ij}^0 \\ & + \beta_{4j} (\mathbf{EVCentrality}_{ij}^0 \times \mathbf{Transitivity}_{ij}^0) \\ & + \beta_{5j} \mathbf{SelfCite}_{ij} + \beta_{6j} \mathbf{ReviewCited}_{ij} \end{aligned}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} \mathbf{Review}_j$$

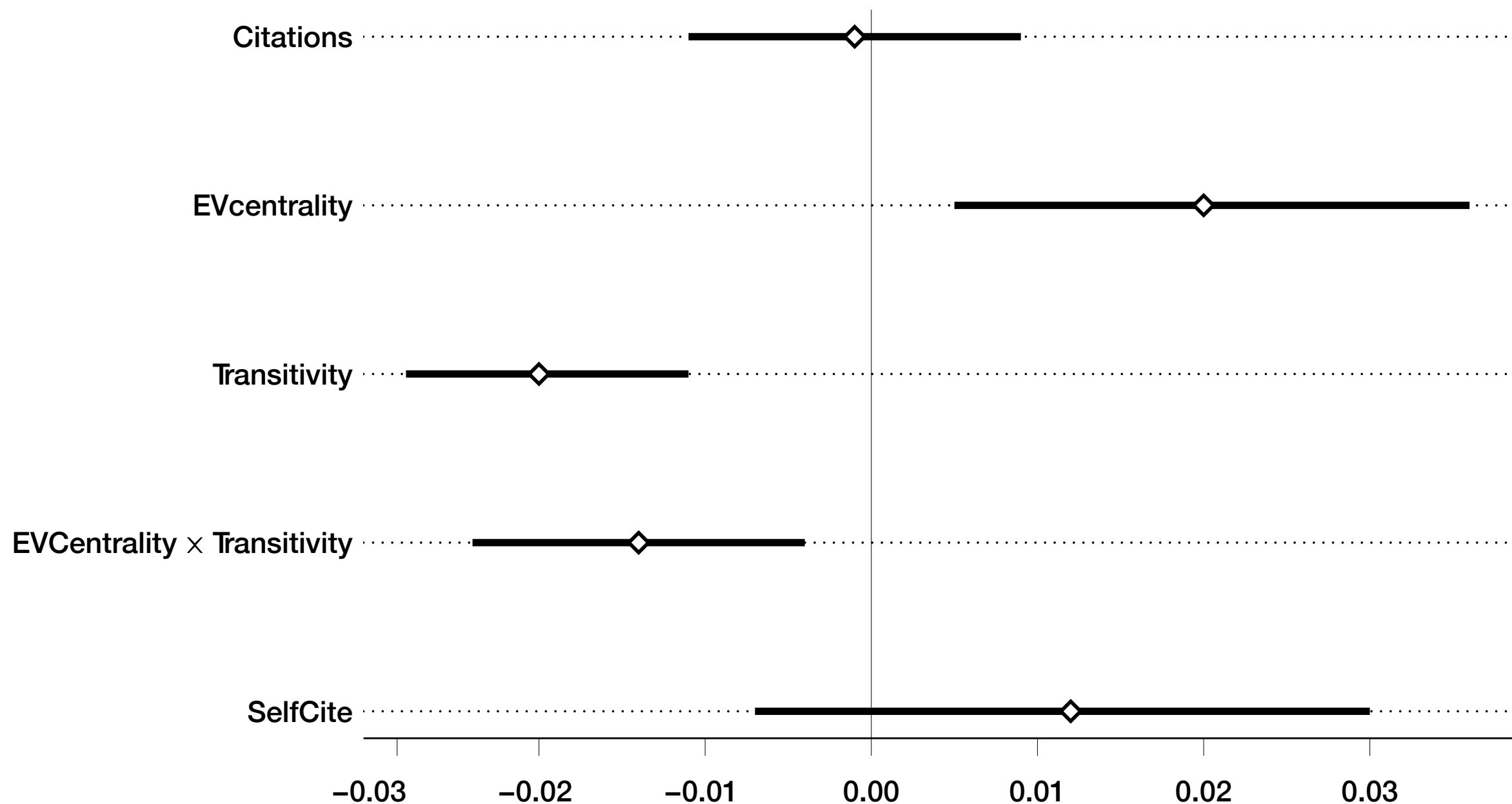
	EVCentrality ^Δ
Citations ⁰	-0.0612 (-0.068, -0.054)
EVCentrality ⁰	-0.504 (-0.515, -0.493)
Transitivity ⁰	-0.083 (-0.089, -0.077)
EVCentrality ⁰ × Transitivity ⁰	-0.101 (-0.108, -0.094)
SelfCite	0.037 (0.022, 0.051)
ReviewCited	0.107 (0.066, 0.147)
Review × Citations ⁰	-0.001 (-0.011, 0.009)
Review × EVCentrality ⁰	0.020 (0.005, 0.036)
Review × Transitivity ⁰	-0.020 (-0.028, -0.011)
Review × EVCentrality ⁰ × Transitivity ⁰	-0.014 (-0.024, -0.004)
Review × SelfCite	0.012 (-0.007, 0.030)
Review × ReviewCited	-0.024 (-0.073, 0.025)
Observations	589,735
Adjusted R ²	.260

Exemplars model

$$\mathbf{EVCentrality}_{ij}^{\Delta} \sim \text{Normal}(\mu_{ij}, \sigma_2)$$

$$\begin{aligned} \mu_{ij} = & \beta_{1j} \mathbf{Citations}_{ij}^0 + \beta_{2j} \mathbf{EVCentrality}_{ij}^0 + \beta_{3j} \mathbf{Transitivity}_{ij}^0 \\ & + \beta_{4j} (\mathbf{EVCentrality}_{ij}^0 \times \mathbf{Transitivity}_{ij}^0) + \beta_{5j} \mathbf{SelfCite}_{ij} + \beta_{6j} \mathbf{ReviewCited}_{ij} \end{aligned}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} \mathbf{Review}_j$$



Data description

Table 1. Summary Statistics for the Full Sample, by Journal Subject Area and Total

	Journals	Articles	Articles reviewed > 0 times	Articles reviewed > 1 time	Mean citations/ year (median, IQR)	Citations referenced (median, IQR)	Prop. citations to WoS (median, IQR)
Nat. Sci.	636	3,938,866	365,226	105,538	.85 (.33, 1.91)	29 (18, 44)	.56 (.29, .76)
Eng. & Tech.	173	1,005,023	34,489	5,552	.58 (.25, 1.32)	23 (13, 34)	.5 (.22, .72)
Health Sci.	264	1,727,804	133,197	30,751	1 (.33, 2.33)	30 (15, 44)	.62 (.31, .82)
Agr. Sci.	64	258,336	8,741	972	.62 (.29, 1.29)	27 (18, 38)	.44 (.19, .66)
Soc. Sci.	277	310,736	33,061	8,054	.62 (.25, 1.50)	34 (19, 53)	.27 (.10, .49)
Humanities	39	45,759	2,437	375	.44 (.17, 1.00)	31 (14, 52)	.27 (.07, .62)
Total	1,155	5,901,565	509,325	140,044	.85 (.33, 1.96)	29 (17, 44)	.55 (.26, .77)

Note: Many journals are listed in multiple subjects, so total counts will be less than the sum of the subject counts. In addition to the number of journals and number of articles, the number of articles cited at least once and at least twice by a review article are shown. Median and inter-quartile range across articles are shown for the mean annual number of citations received, the number of works cited, and the proportion of those citations that reference an article in the WoS database.

Who is an exemplar

