Structure and Dynamics of Online Information Ecosystems

Concentration, Segregation, and Governance

Burak Özturan

Dissertation defense · Northeastern University

Committee: David Lazer (Chair) · David Rothschild · Nicholas Beauchamp · John Wihbey

The common question

- A small fraction of users produces most of what circulates online, and audiences for that content are sorted into groups that do not resemble one another.
- Both patterns are documented. Neither has been examined at the level that matters for political representation: the individual citizen with a known relationship to the political system.
- The obstacle has been data. Platform samples carry no reliable individual-level attributes; surveys cannot observe years of behavior at the level of the individual link.

All three chapters use one instrument: a panel of US registered voters matched to their Twitter/X activity.

One panel, three questions

	Question	Object of measurement
Chapter 1	Who produces it?	A property of people
Chapter 2	How do audiences sort?	A property of sources
Chapter 3	What flows through it?	A property of content

- 1,643,182 registered voters matched to Twitter accounts: age, gender, party registration, and location observed, not inferred.
- Bots, brands, and non-US accounts excluded by construction.
- Sharing observed July 2018 to May 2023.

A second thread running through all three

In each chapter, a measurement convention the field adopted without much scrutiny turns out to be doing real work in the conclusions.

Ch 1

Observation-window length reproduces the entire factor-of-three disagreement in published concentration estimates, within one dataset.

Ch 2

Measuring partisanship, inside news, by visit, at outlet level, with Independents discarded, is not a neutral set of defaults.

Ch 3

A decline estimated on one dataset is corroborated on a second with a different sampling frame.

Each chapter's headline result would change, and in two cases reverse, under the conventions it argues against.

Chapter 1

Concentration Without Representation

with David Lazer

The measurement problem

Published estimates of online sharing concentration vary by a factor of three at comparable thresholds, and the variation has gone unexplained.

- Studies differ in platform, content definition, population, and observation window.
- No study has varied these systematically within a single dataset, so it has been impossible to say which choice drives the disagreement.

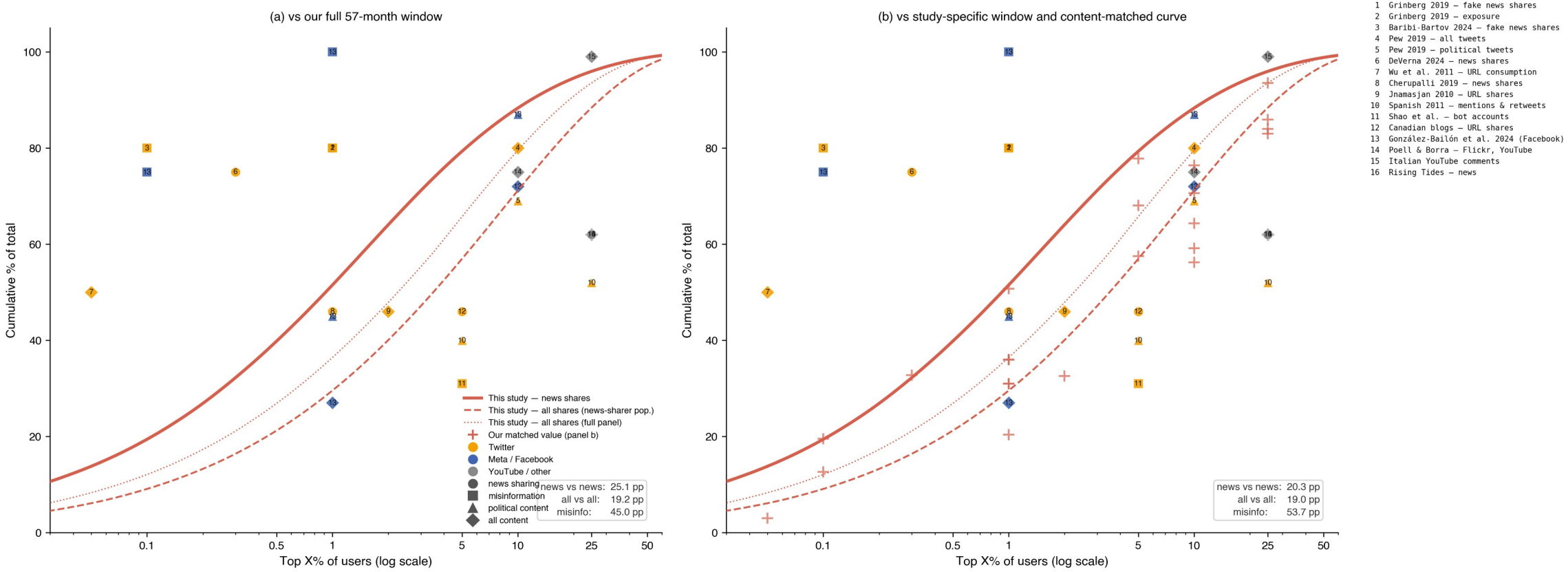
972,830
panelists shared at least one URL

483,695
shared at least one news link

41M
news shares across 57 months

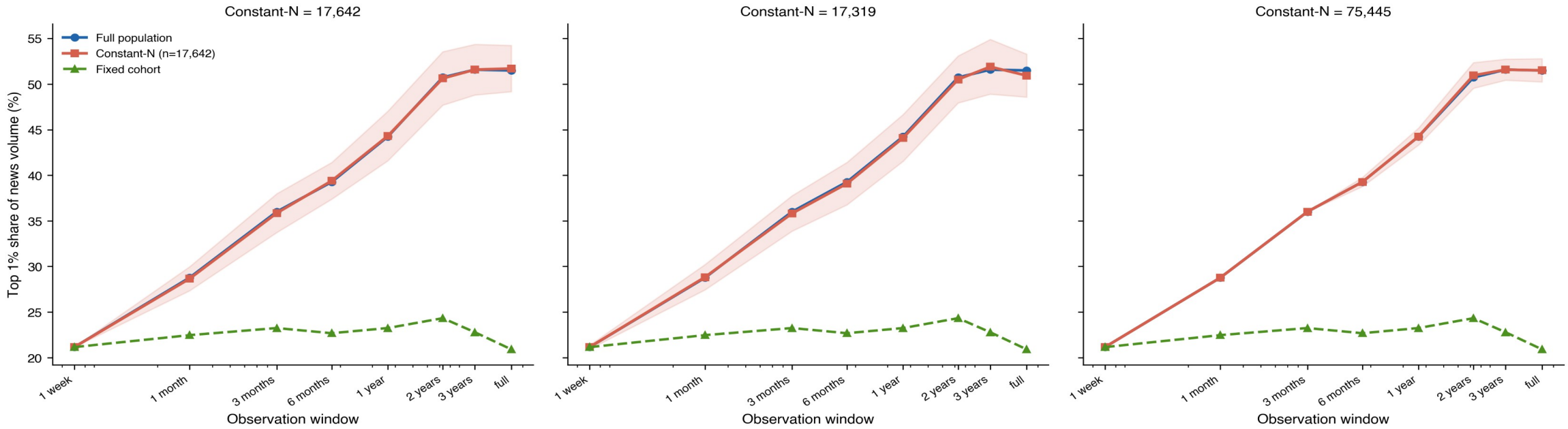
Selection operates before intensity does: the 489,135 who never shared news are younger (36.8 vs. 40.6 years) and more Republican (34.7% vs. 30.7%) than those who did.

Window length alone reproduces the whole literature



Top-1% share rises from 17% at a one-week window to 50% at 57 months, population and content held fixed. Matching each study's window cuts the mean absolute gap against our curve from 26.5 to 10.0 percentage points.

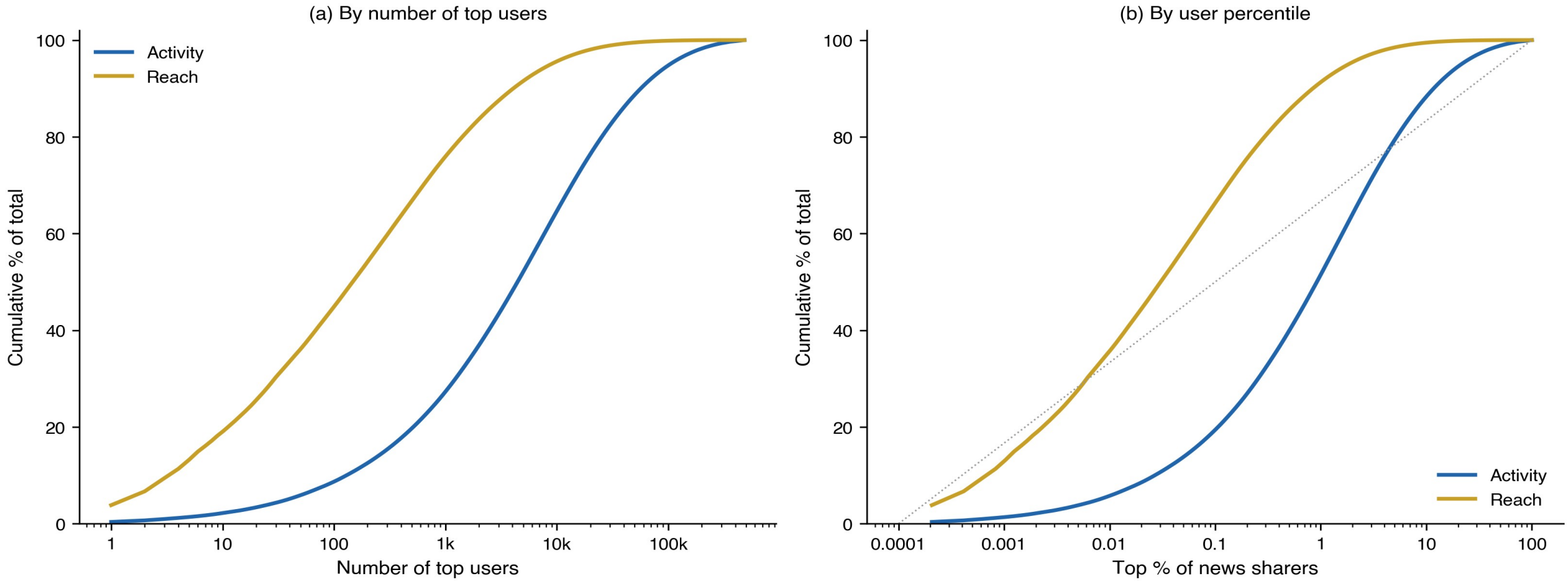
The mechanism is compositional, not behavioral



- Constant-N: holding sample size fixed does not flatten the curve, so it is not a sample-size artifact.
- Fixed cohort: the same users followed across windows stay flat near 21%, against a population estimate reaching 50% — a 29-point gap.

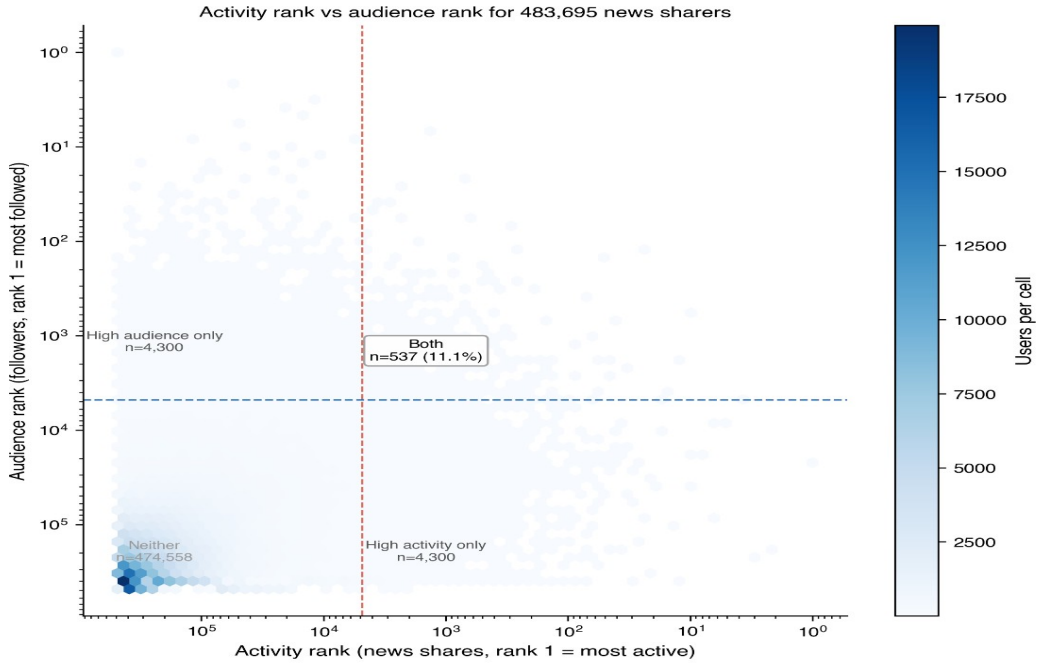
Entrants accumulate in the denominator faster than they add to the numerator: at the full window they are 32.5% of the active pool and 4.6% of volume.

Activity and reach are both concentrated



Top 1% = 51.5% of shares (Gini 0.911), but 91.3% of reach. The top 0.1% produce 19.4% of volume and 66.3% of potential exposure. The reach curve lies above activity across the whole distribution.

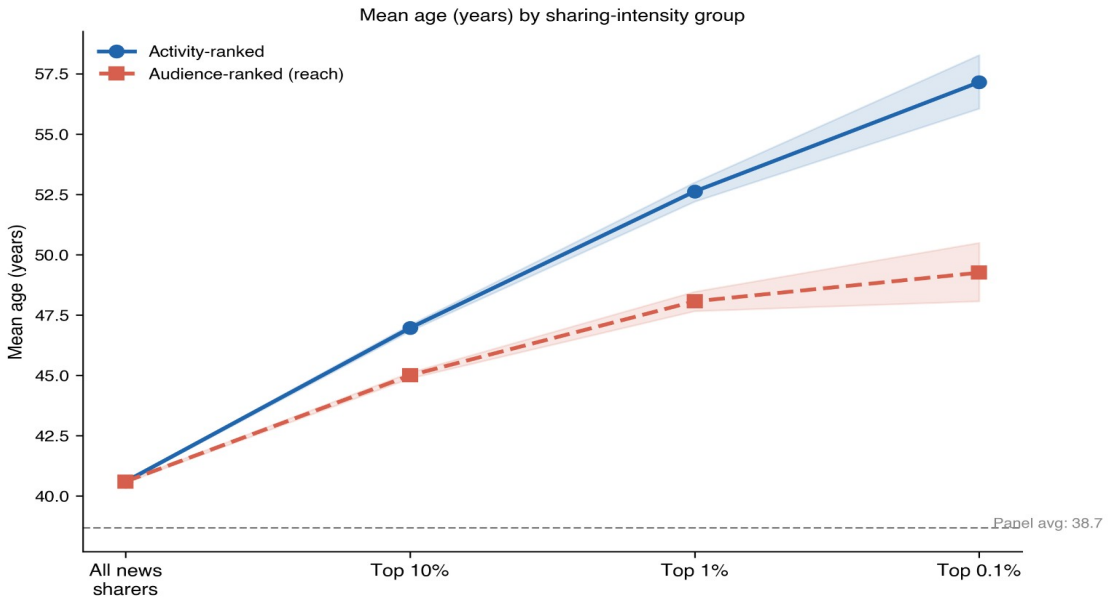
But they identify different people

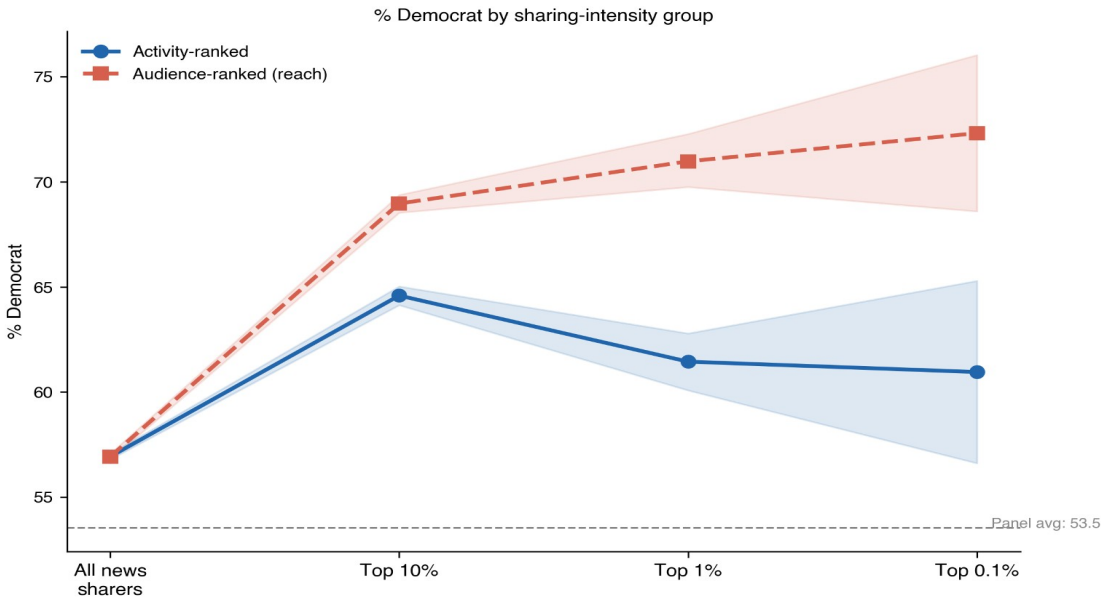


Only 11.1% overlap at the top 1% — eleven times chance, but nine in ten members of each elite do not belong to the other.

Activity elite median: 1,484 followers. Reach elite median: 26,062 — a seventeenfold difference. Spearman ρ = 0.29 overall, near zero within deciles.

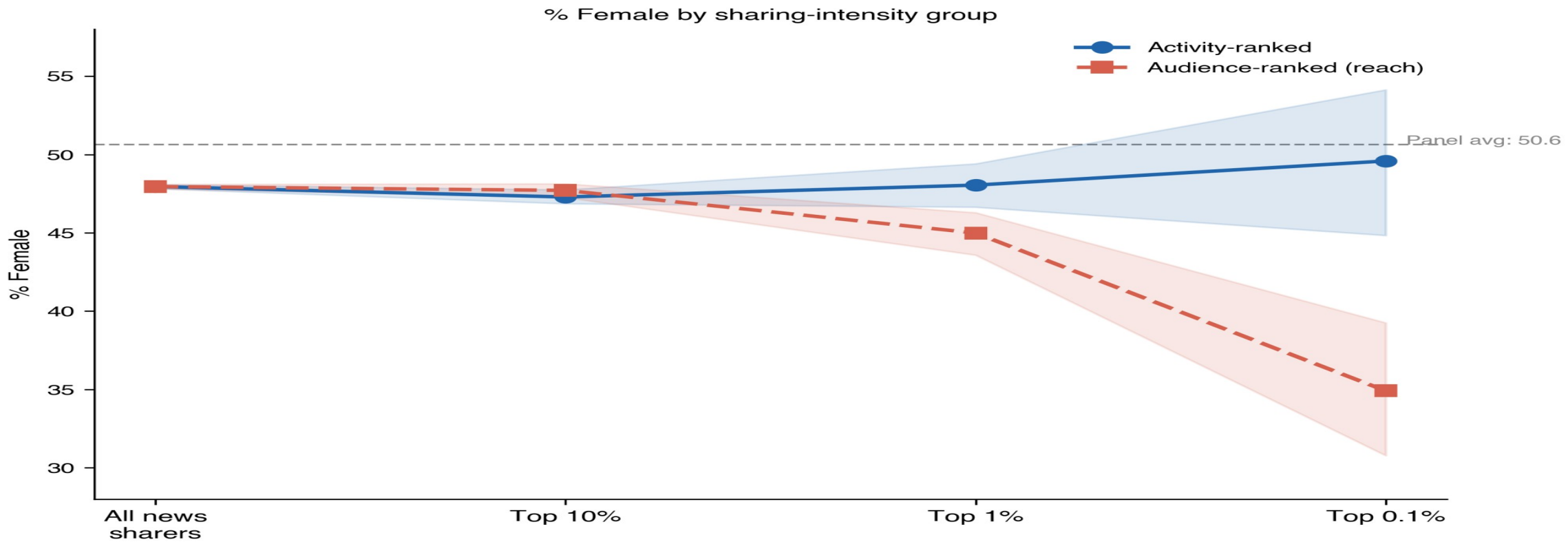
Two elites, two demographic profiles





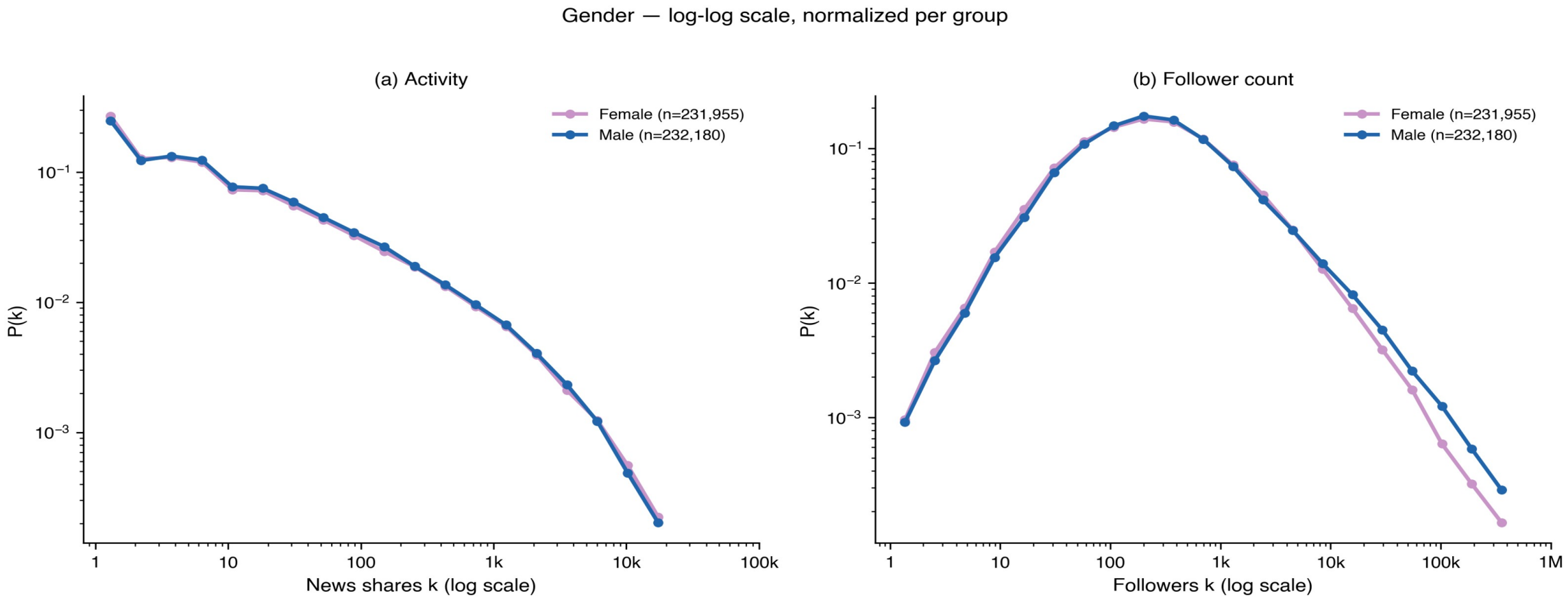
- Age rises from 40.6 to 57.1 years under activity ranking, but only to 49.2 under reach.
- Democratic share rises more steeply under reach (72%) than activity (61.5%).

The gender gap only one measure can see



Flat under activity. Falls from 48.4% to 34.5% under reach. A design measuring only volume would correctly conclude men and women participate equally — and would be wrong about who is heard.

Where the gap is produced



Activity distributions coincide exactly. Follower distributions separate above ~1,000 followers. 99th percentile: 16,414 (men) vs. 11,839 (women).

Chapter 1: what it establishes

- Concentration estimates are not comparable across studies unless the observation window is reported and matched.
- Activity and reach are separate quantities that identify separate populations and imply separate interventions. Rate limits act on one; algorithmic demotion acts on the other.
- Both elites are older and more Democratic than the panel; only the reach elite is disproportionately male.
- Elite position is durable at the head (55.0% of the top 1% persist; 74.9% of the top 0.1% stay within the top 1%) and near-memoryless in the middle (rank correlation ≤ 0.12).

The people who share the most are largely not the people whose shares travel furthest.

Chapter 2

Demographic Segregation in Online Information Ecosystems

The literature narrows twice over

1. It measures partisanship

The outcome varies a lot across studies. Which social divide gets measured almost never varies.

2. It measures news

News is a small share of online activity, with a partisan composition unlike media diets as a whole.

The restrictions compound: the one divide most likely to dominate, inside the one content category where it is most likely to dominate — and usually with Independents discarded.

Measure: excess Theil H against a permutation null, person-weighted, URL level, minimum two distinct visitors per source, 59 months. Every dimension reduced to a common two-group footing by averaging over pairwise contrasts.

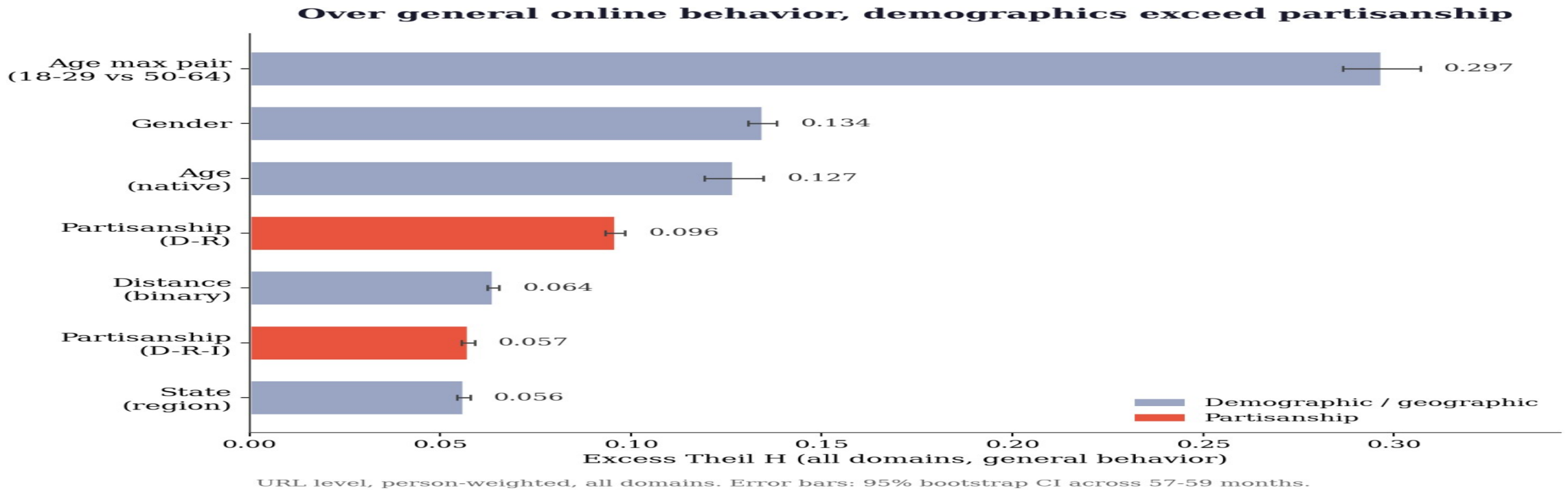
Why category counts have to be equalised

Excess Theil H shrinks mechanically as a variable is split into more categories, whatever the underlying sorting: more groups means fewer users per group at any source, so the observed index and its null converge.

Dimension	Adjustment
Age	Averaged over the six contrasts among four brackets
Partisanship	Averaged over three contrasts among D, R and I — counts Independents instead of discarding them
Gender	Already binary, no adjustment
Geography	Opposite problem: state collapsed to four Census regions, distance split at its median

Both adjustments work against the result: they raise partisanship and lower age relative to their native constructions. The comparison is conservative.

Over general sharing, demographics lead



Both demographic dimensions sit above every partisan and geographic measure. Partisanship is the lowest dimension measured, geography included.

Each bar is that dimension's native construction (native D/R/I partisanship 0.057; native age 0.127; age's most extreme pair 0.297). On the common two-group footing used for every cross-dimension claim in this chapter, the values are gender 0.134, age 0.134, partisanship 0.048.

The ordering does not depend on how you count

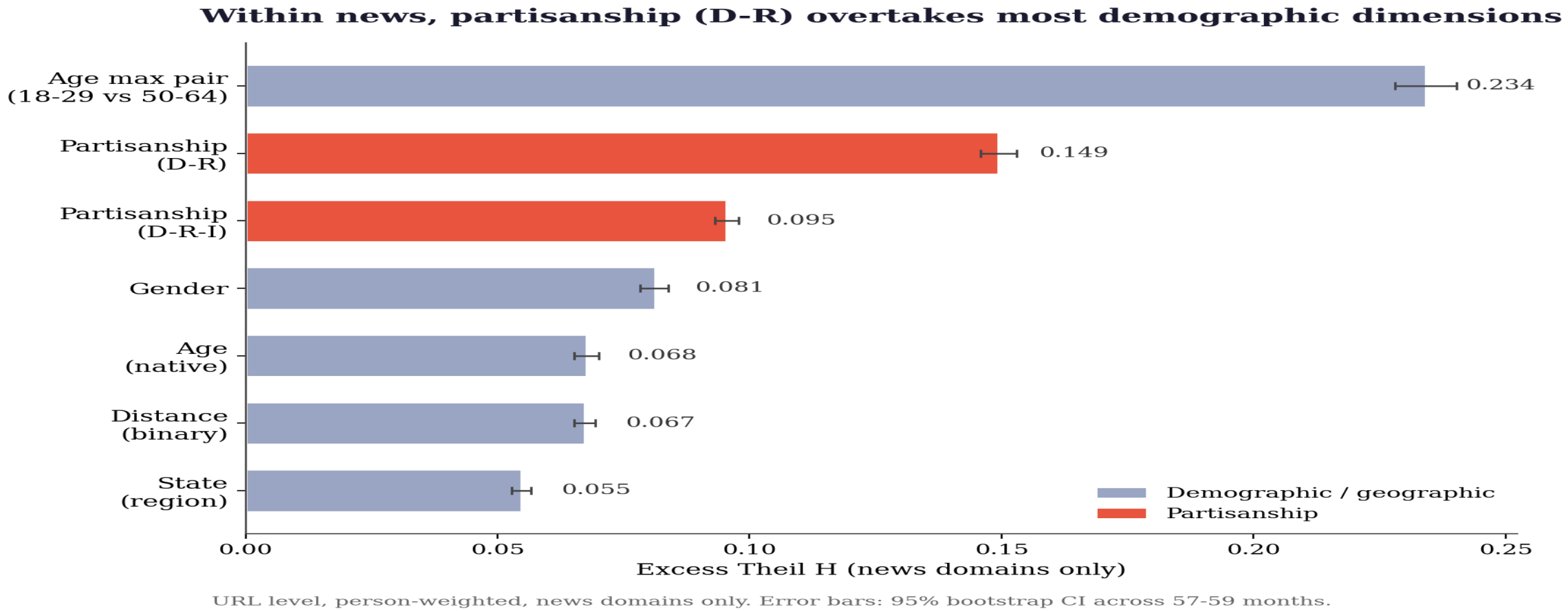
All domains	Gender	Age	Partisanship (D/R/I)
URL, person-weighted	0.134	0.134	0.048
URL, volume-weighted	0.128	0.196	0.078
Domain, person-weighted	0.048	0.085	0.022

- Bootstrap intervals over months do not overlap under volume weighting (partisanship 0.075–0.081; gender 0.124–0.133; age 0.190–0.202).

One comparison is sensitive.

Gender exceeds the Democrat–Republican contrast under person weighting (0.134 vs. 0.096) and falls below it under volume weighting (0.128 vs. 0.145). Reported, not hidden.

Within news, one contrast overtakes



Native constructions again: native D/R/I partisanship 0.095 exceeds native gender 0.081 here, and native age reads 0.068. Pairwise-averaged — the fair footing — the same three are partisanship 0.079, gender 0.081, age 0.114. The Democrat–Republican contrast is the only measure that crosses between the two bands, and only within news.

The exception is real, and it is narrow

- Restricted to Democrats and Republicans, partisan segregation rises sharply inside news: 0.096 → 0.149 at the URL level, 0.028 → 0.121 at the domain level. Replicates under Gentzkow–Shapiro isolation, which shares no functional form with Theil H (0.102 → 0.159), and rises to 0.229 at a twenty-visitor threshold.
- But all three pairs move the same way, and D–R is larger than the other two combined (D–R +0.053; R–I +0.026; D–I +0.013). Independents act as an undifferentiated middle, not a third pole.
- Averaged over the three contrasts, partisanship rises only 0.048 → 0.079: indistinguishable from gender (0.081, paired difference 0.003, [−0.001, 0.006]) and still short of age (0.114, difference 0.035, [0.032, 0.038]).

What rises within news is the distance between two committed partisan camps, not the organising power of party across the electorate.

News against non-news is the decisive contrast

URL level	Non-news	News	Direction
Partisanship (D–R)	0.092	**0.149**	▲ rises
Gender	0.140	**0.081**	▼ falls
Age	0.129	**0.114**	▼ falls

- Non-news supplies roughly three quarters of distinct URL sources, so the pooled “all domains” baseline is dominated by the non-news half and understates how partisan news is on its own.

One honest qualification.

Scored as a single native three-way index rather than pairwise-averaged, partisanship reaches 0.095 within news and exceeds native gender (0.081) in 54 of 59 months. Age is unaffected by the choice either way.

Gender is not partisanship under another name

$\phi = 0.095$

They are barely correlated. Party explains under 1% of the variance in gender.

$0.134 > 0.096$

Magnitude rules it out. Gender exceeds the D–R contrast over general sharing. A proxy cannot be larger than the thing it proxies for.

Opposite signs

Within news partisanship rises while gender falls by nearly half (0.134 → 0.081); outside news both reverse (gender 0.140, partisanship 0.092).

Gender organises the large non-political share of sharing, where party is close to irrelevant. Partisanship organises the small political share, where gender is close to irrelevant.

Most female-leaning domains: beauty and shopping, animal welfare, influencer commerce. Most male-leaning: sports, gaming, technical services.

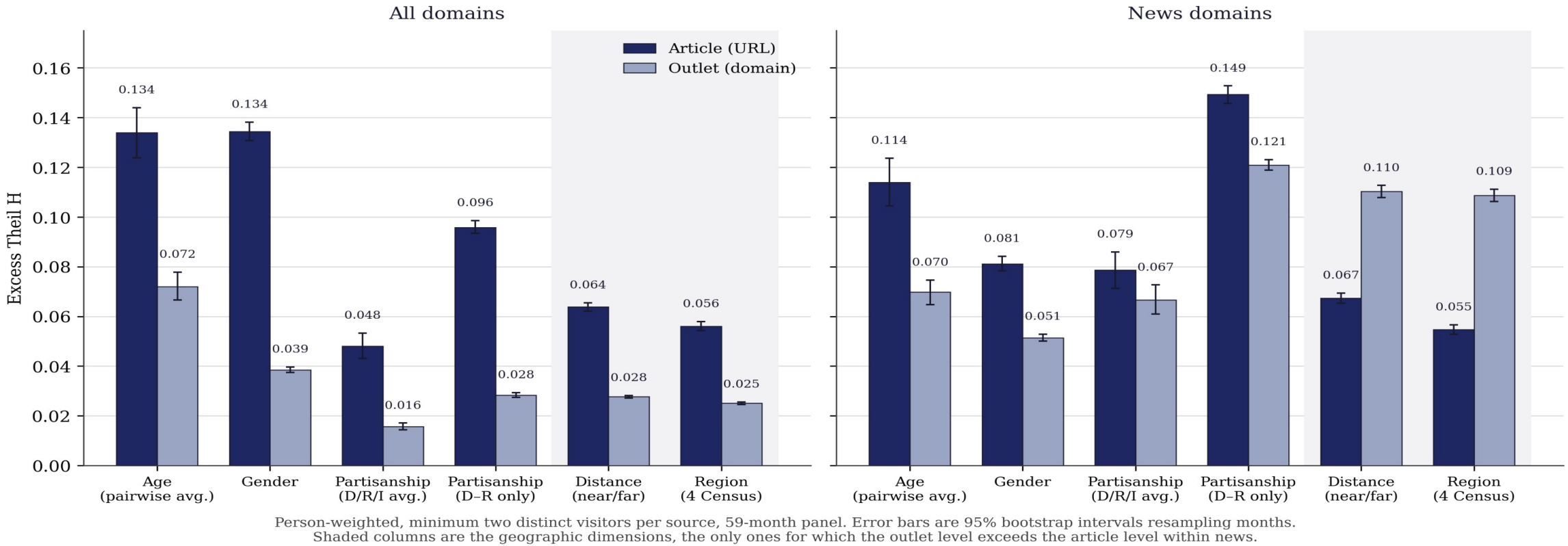
Sorting lives in articles, not outlets

All domains	URL	Domain	Ratio
Gender	0.134	0.039	3.4×
Age	0.134	0.072	1.9×
Partisanship (D/R/I)	0.048	0.016	3.0×
Partisanship (D–R)	0.096	0.028	3.4×

- The ideological-segregation literature established this for partisanship. It is not particular to partisanship.
- Age is the dimension most visible at the outlet level: age groups differ in which publications they draw on at all. Gender shows the opposite — same outlets, different stories.

Ordering survives inside news: gender 0.081→0.051, age 0.114→0.070, D/R/I 0.079→0.067, D–R 0.149→0.121.

Geography is the exception that clarifies the rule



Person-weighted, minimum two distinct visitors per source, 59-month panel. Error bars are 95% bootstrap intervals resampling months.
Shaded columns are the geographic dimensions, the only ones for which the outlet level exceeds the article level within news.

Over all domains geography behaves like everything else (distance 0.064 URL vs. 0.028 domain; region 0.056 vs. 0.025). Within news it inverts: the article level barely moves (0.067; 0.055) while the outlet level rises roughly fourfold, to 0.110 and 0.109 — the only bars where outlet exceeds article.

Why the exception matters

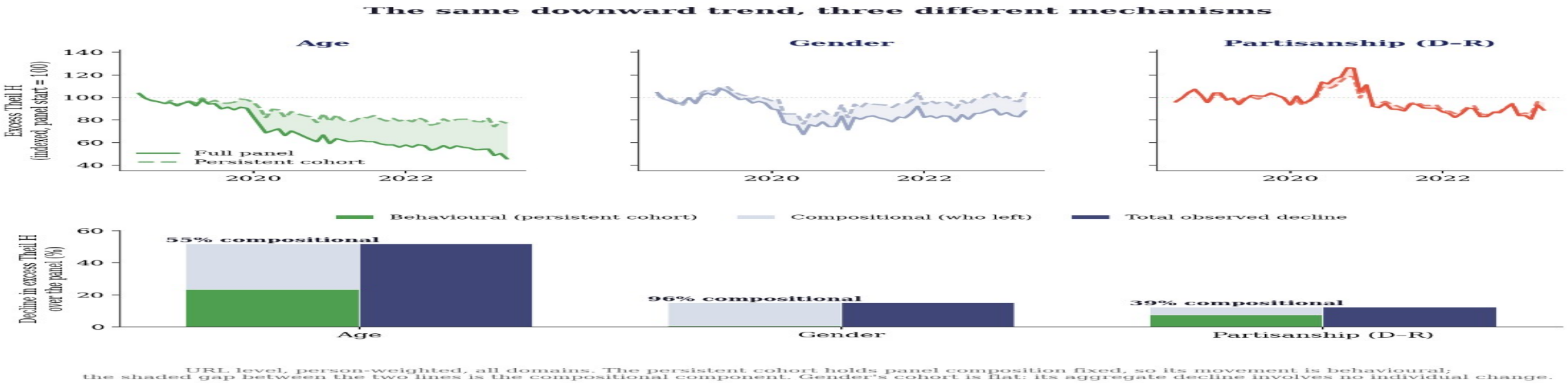
The likeliest explanation is local news. Geographically bounded audiences are a property of publications, not of stories.

- It confirms the outlet-versus-article distinction tracks how content is actually organised, rather than being an artefact of how the index behaves at different levels of aggregation.
- It shows the article-level concentration found for age, gender and partisanship is a property of those divides, not a mechanical feature of URL data — geography, measured identically, runs the other way.

Verified three ways.

k-means on raw coordinates (varies the partition, keeps the index); mean pairwise great-circle distance against a random-draw null (keeps neither): news domains draw audiences living 184 km closer together than chance, against 12 km for non-news. All three agree.

One shared trend, three different causes



URL level, person-weighted, all domains. The persistent cohort holds panel composition fixed, so its movement is behavioural; the shaded gap between the two lines is the compositional component. Gender's cohort is flat: its aggregate decline involves no individual change.

URL level	Full panel	Persistent cohort	Compositional
Age	−52%	−23%	~55%
Gender	−15%	−0.6%	96%
Partisanship	−12%	−7.5%	~39%

The same trend line, two opposite stories

- The panel loses roughly 70% of monthly active users over five years, and loses them unevenly. An aggregate trend cannot tell “users grew more alike” from “the distinctive ones left.”
- A cohort active in both the first and last six months fixes composition by construction, so movement inside it is behavioural.
- Stayers vs. leavers agrees, dimension by dimension. Eventual leavers were already more separated at the outset by age (0.349 vs. 0.303) and gender (0.154 vs. 0.126). For partisanship it runs the other way (stayers 0.094 vs. 0.082) — attrition was not selecting on partisan distinctiveness.

Gender's decline involves essentially no individual changing. Partisanship's is mostly people changing. Given only the two trend lines — 15% and 12% — a reader would tell one story, and it would be wrong for one of them.

Chapter 2: what it establishes

- Partisanship is the lowest of every divide measured over general sharing, at roughly a third of either demographic dimension — geography included.
- The reversal prior work reports belongs to committed partisans inside news. Admitting Independents removes nearly half the news effect.
- Segregation is built at the article level for every social divide, not just partisan ones. Geography is the exception, and it is the exception local news predicts.
- Parallel aggregate declines rest on unrelated causes: composition for gender, behaviour for partisanship, a mixture for age.

An assessment of cross-cutting exposure restricted to party is not a partial answer to that question so much as a full answer to a narrower one.

Chapter 3

Declining Information Quality Under New Platform Governance

with Katya Ognyanova and David Lazer

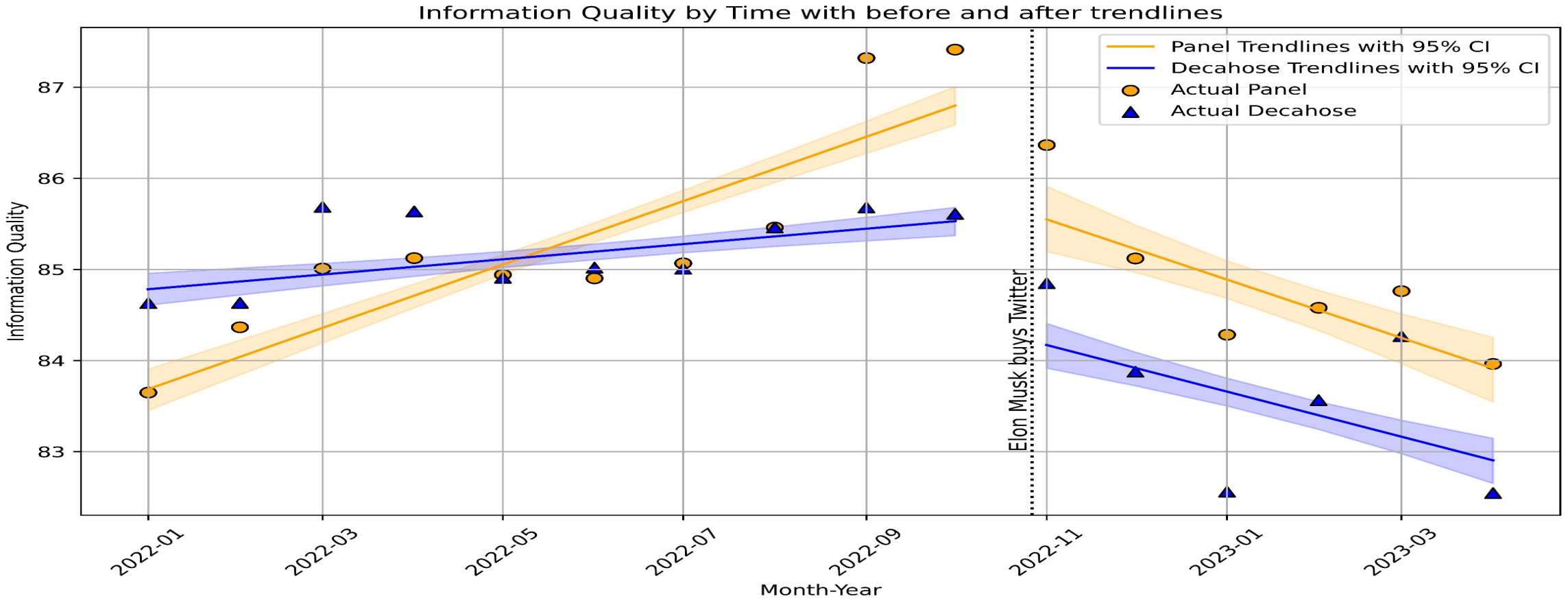
The event and the design

On 27 October 2022 Twitter changed ownership, followed by executive departures, large-scale layoffs, and revisions to content policy including reinstatement of suspended accounts and dissolution of the Trust and Safety Council.

- Measure. Monthly average NewsGuard reliability score (0–100) of domains linked in shared tweets. Original tweets and retweets only; satire and user-generated platforms excluded.
- Design. Interrupted time series, Jan 2022 to Apr 2023, OLS with Newey-West standard errors, bootstrapped CIs.
- Two datasets. The voter-file panel (stable, US, no bots) and the Decahose (10% of global tweets, includes bots and organizations). 18.5 billion tweets.

A decline appearing in both is hard to attribute to the composition of either.

Quality declined in both datasets



Improving before the transition in both. Immediate drop of 1.45 points (Panel) and 1.39 (Decahose), then a sustained downward trend.

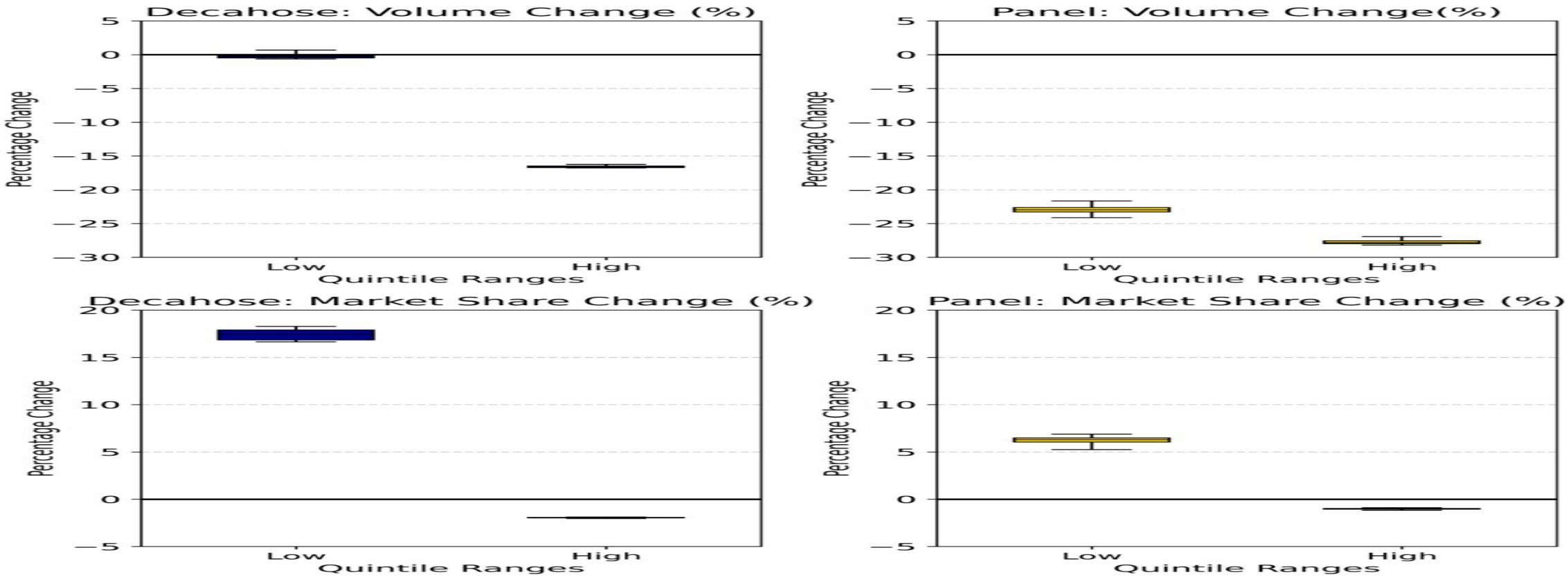
The ITS estimates

	Twitter Panel	Decahose
Trend before (β_1)	+0.34/mo ***	+0.08/mo *
Immediate change (β_2)	−1.45 **	−1.39 ***
Trend after ($\beta_1+\beta_3$)	−0.37/mo	−0.27/mo

- Not just a one-time drop, but a continuing trend.
- Robustness: VIF < 5; intervention date shifted ±1–3 months; ARIMA specification; alternative credibility scores.

** $p \le .05$, ** $p \le .01$, *** $p \le .001$. Newey-West standard errors, lag-1.*

Driven by market share, not volume collapse



Decahose: low-quality volume essentially flat (−0.28%) against a sharp high-quality contraction (−16.58%). Panel: high-credibility volume falls further (−27.84%) than low (−22.96%). The 40–59 band gains +29.4% (Decahose) and +30.1% (Panel) market share.

Chapter 3: what it establishes

- Information quality reversed direction at the ownership transition, in two datasets with different sampling frames.
- The mechanism is compositional. High-quality content shrank in both absolute and relative terms while mid- and low-credibility sources became more visible.
- Effect sizes are small on a 0–100 scale and the ecosystem remains above NewsGuard's high-credibility threshold. High-quality sources still carry ~90% of shared content. The direction and persistence are what matter.
- Limits. No exposure data, so we cannot say users encountered more low-quality content. Domain-level rather than URL-level scoring. Access ended May 2023.

Taken together

Three findings that recur

Age

The most consistent signal. 40.6 to 57.1 years across the sharing distribution (Ch 1); the largest or tied-largest source of audience segregation in every content regime, including news (Ch 2).

Party

Weaker than its prominence implies. Lowest of every divide over general sharing, geography included (Ch 2). And the partisan composition of the elite flips with content class: 61% Democratic for news, 64% Republican for misinformation in the same panel.

Gender

Operates in tails and in non-political content. Invisible under activity, fourteen points under reach (Ch 1); tied-largest over general sharing, halves within news (Ch 2).

Each would be missed by a statistic computed at the median.

What the two Twitter-panel chapters say jointly

- Chapter 1: the users who drive circulation are demographically unlike the public, and the direction of that distortion depends on whether we count volume or audience.
- Chapter 2: the audiences receiving that content sort more strongly by who people are than by which party they support.
- Demographic structure sits on both sides of the same information environment.

One informative disagreement.

The October 2022 transition moved information quality (Ch 3) but produced no discontinuity in any segregation series (Ch 2).

Governance acts more directly on what is available than on how people sort themselves across it.

Contribution

Substantive

The online information environment is unequal in production, sorted in reception along lines the field has not been measuring, and contingent in quality on the institution hosting it.

Methodological

Report the full concentration curve and the observation window. Report activity and reach separately. Measure divides other than party, and count Independents. Measure content other than news. Do not read outlet-level data as story-level sorting. Do not read a trend line as behaviour without asking whether the population held still.

Instrumental

Voter-file linkage makes demographic questions about online behavior answerable without inference from behavior itself.

Limitations

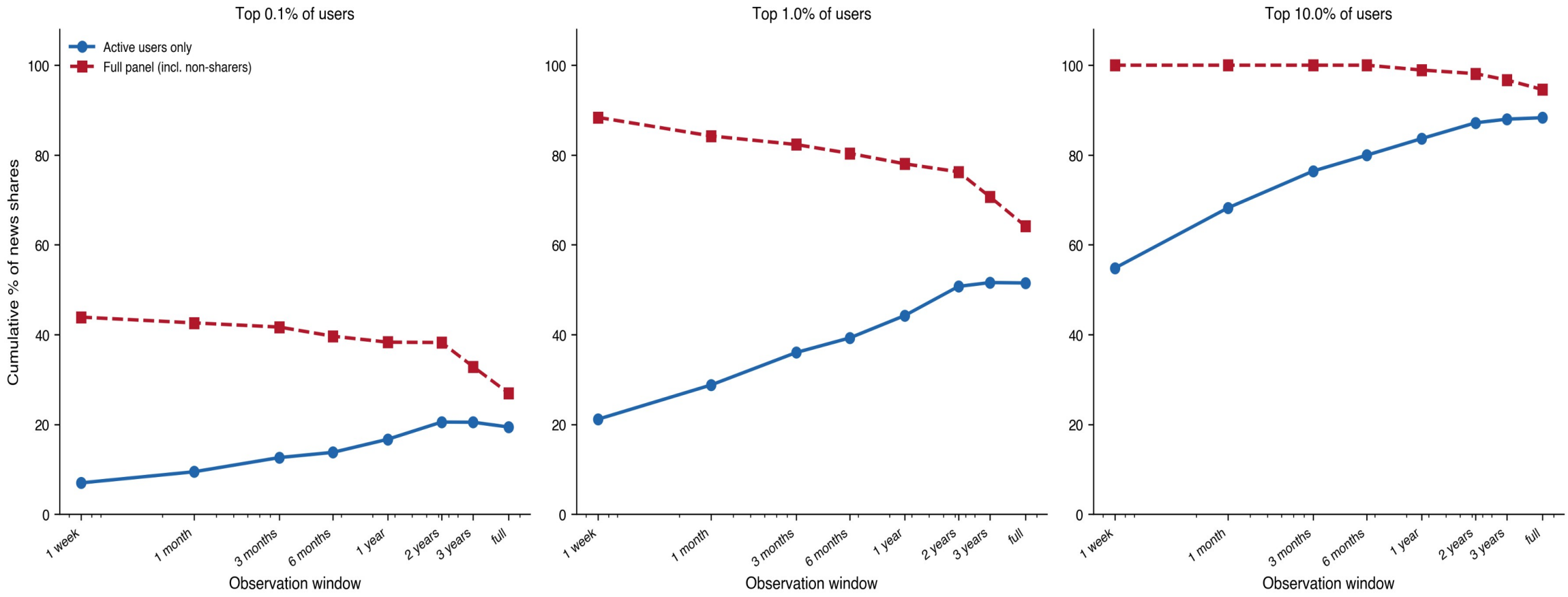
- Circulation, not exposure. We observe what users shared, which is downstream of their choices, their tie structure, and ranking systems. We cannot separate the three.
- One platform, changing across the window. Ownership, moderation, and user base all shifted. Temporal results describe this platform in these years.
- Panel construction. Linkage favors rare names and inherits voter-file biases, overrepresenting White users.
- Category-count corrections discard resolution. The collapses that make cross-dimension comparison possible are reported as full sweeps in the appendices so the trade-off is inspectable rather than hidden.
- Descriptive. No chapter identifies a causal effect in the experimental sense, and none speaks to attitudinal consequences.

Thank you

Questions?

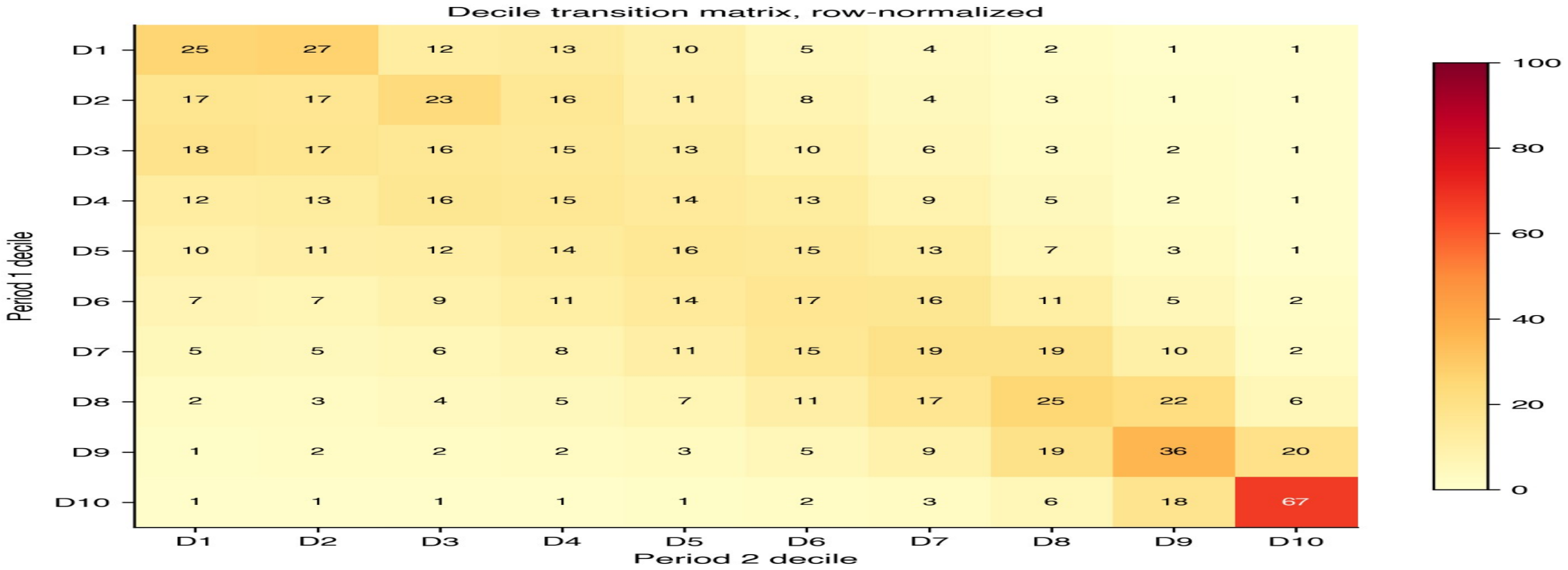
Backup slides

Backup: denominators move in opposite directions



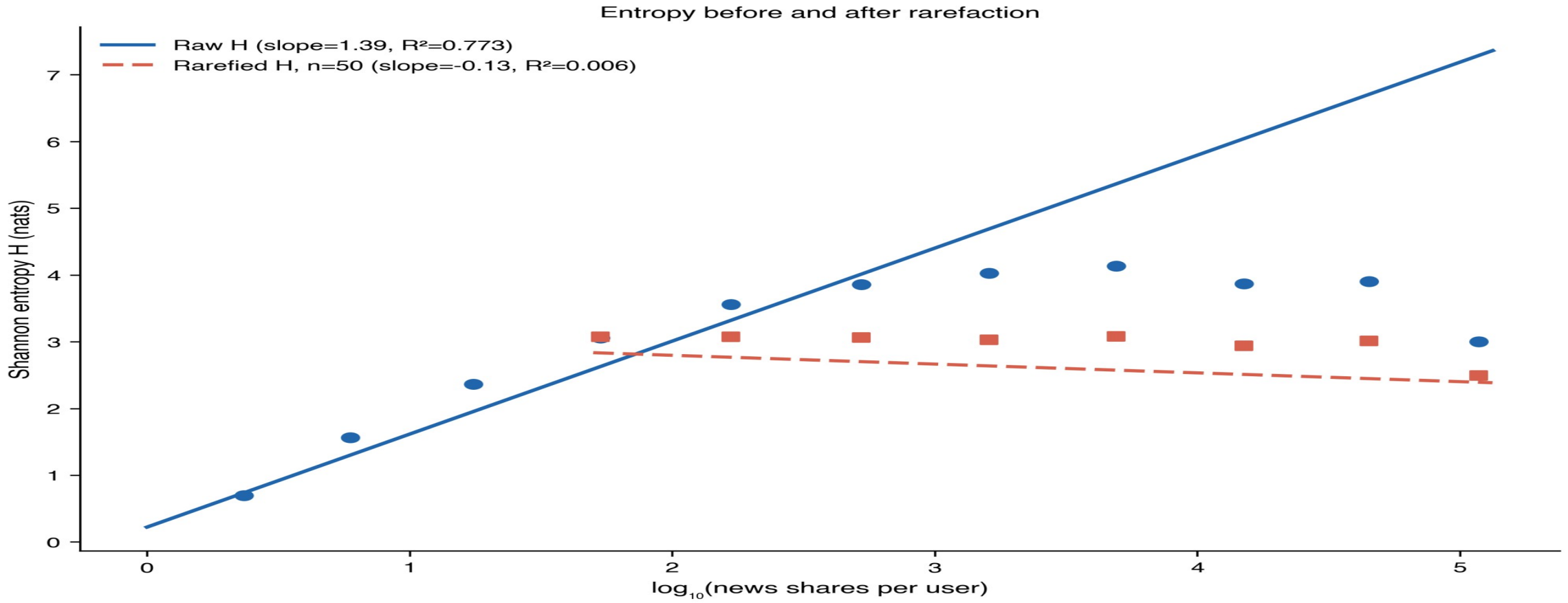
Conditioning on active users, concentration rises with window length. Over all registered voters it falls. Same data, opposite direction. At the full window: 51.5% vs. 64.1% at the top 1%.

Backup: persistence at the head, volatility in the middle



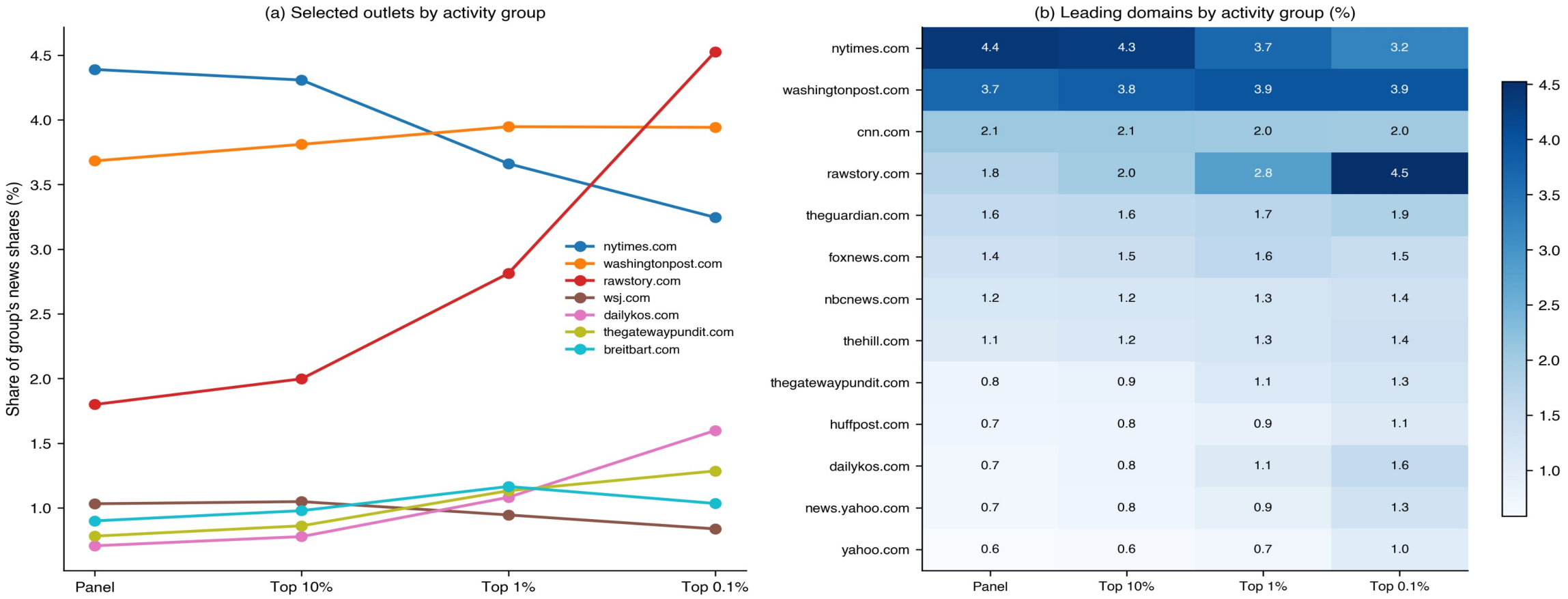
55.0% of the top 1% persist; 49.3% of the top 0.1% hold that rank and 74.9% stay within the top 1%. Above 3,000 shares persistence is 71.3% against 50.7% below. Diagonal dominance confined to the top decile — this reconciles stochastic-popularity findings with supersharer-persistence findings.

Backup: heavy sharers are not narrower, just louder



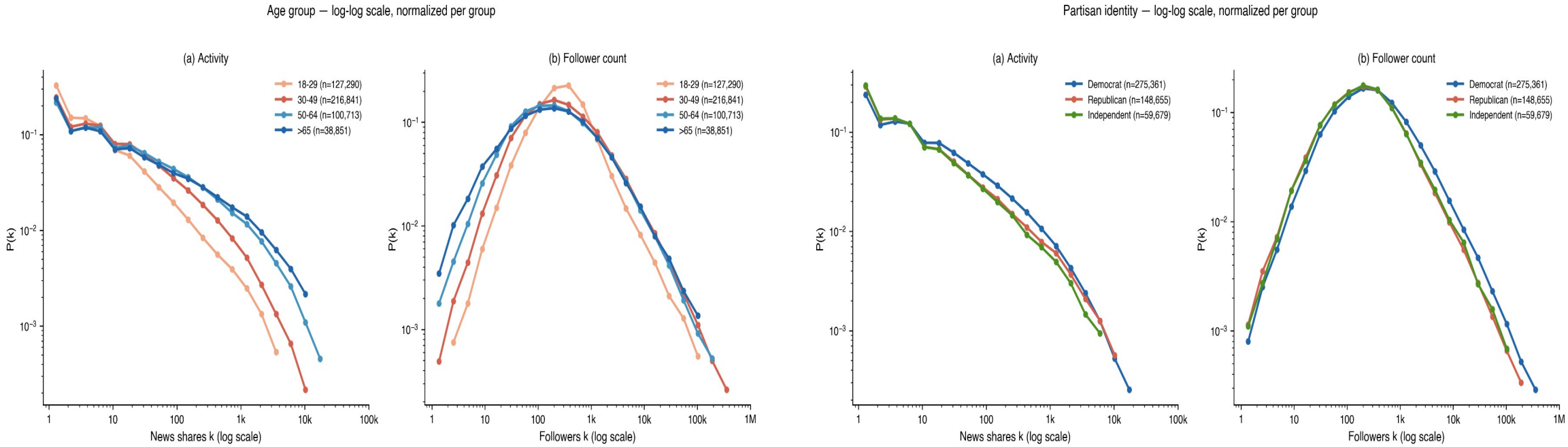
Raw entropy rises 1.39 nats per tenfold increase in shares (R^2 = 0.773), but that is a finite-sample artifact. Rarefying kills it (slope −0.132, R^2 = 0.006).

Backup: what the elite shares



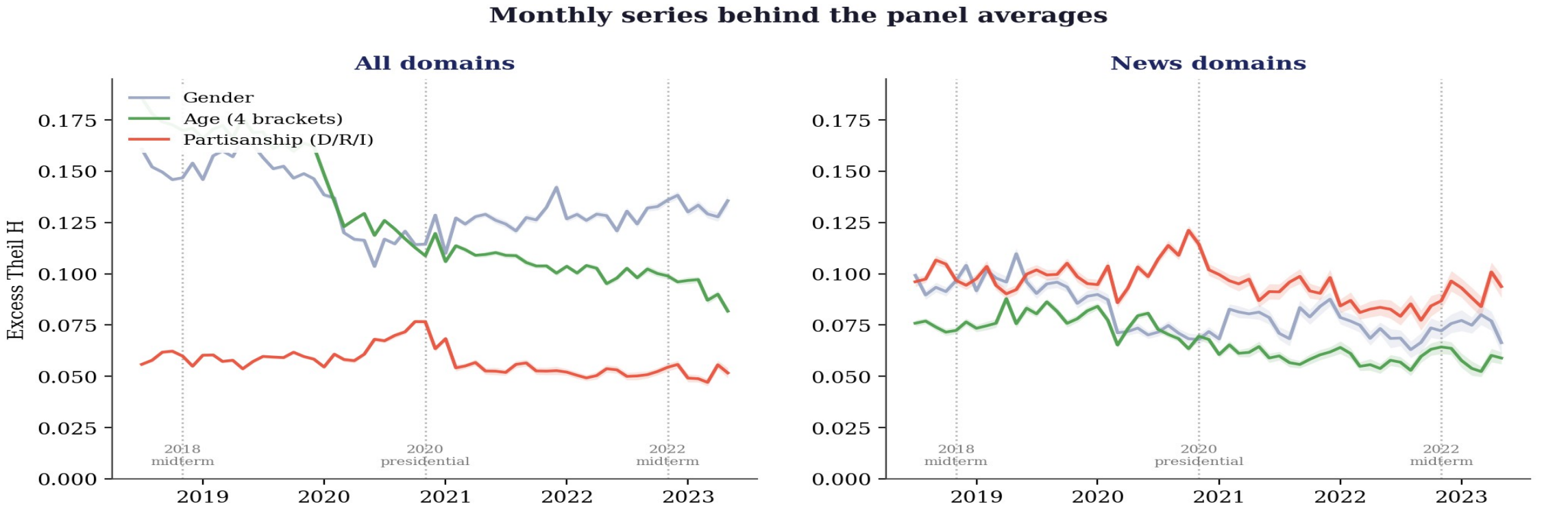
Raw Story rises from 1.8% to 4.5% and overtakes the NYT at the top 0.1%. The shift is a property of who shares, not of narrower individual diets.

Backup: age and partisan degree distributions



Age predicts sharing volume reliably but not follower count. Democrats hold a modest audience advantage visible only at the highest follower counts.

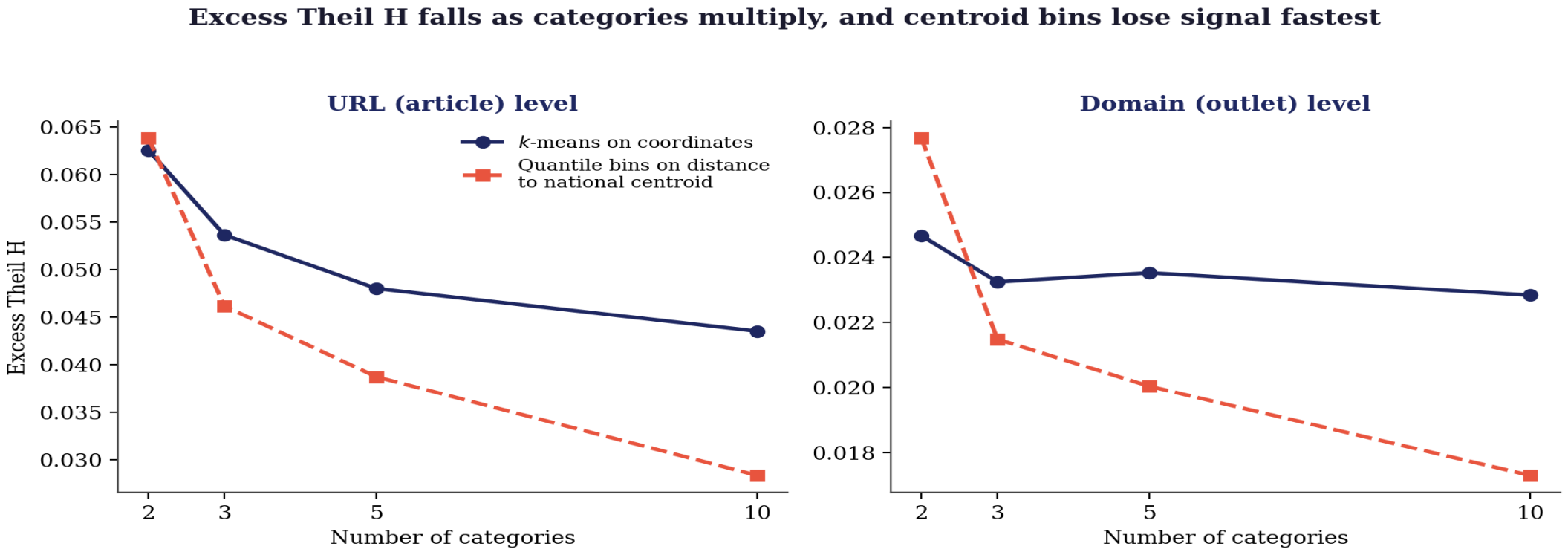
Backup: the monthly segregation series



URL level, person-weighted, minimum two distinct visitors per source. Age and partisanship are each their single native index (four brackets, three parties); gender is native and binary. Dotted lines mark US election months. Shaded bands are ±1.96 standard deviations of each month's permutation null, an approximate month-by-month uncertainty range.

The ranking holds in every month but November 2020, at the peak of a presidential election, when partisan segregation within news runs some 28% above its panel mean. No discontinuity at the October 2022 ownership change.

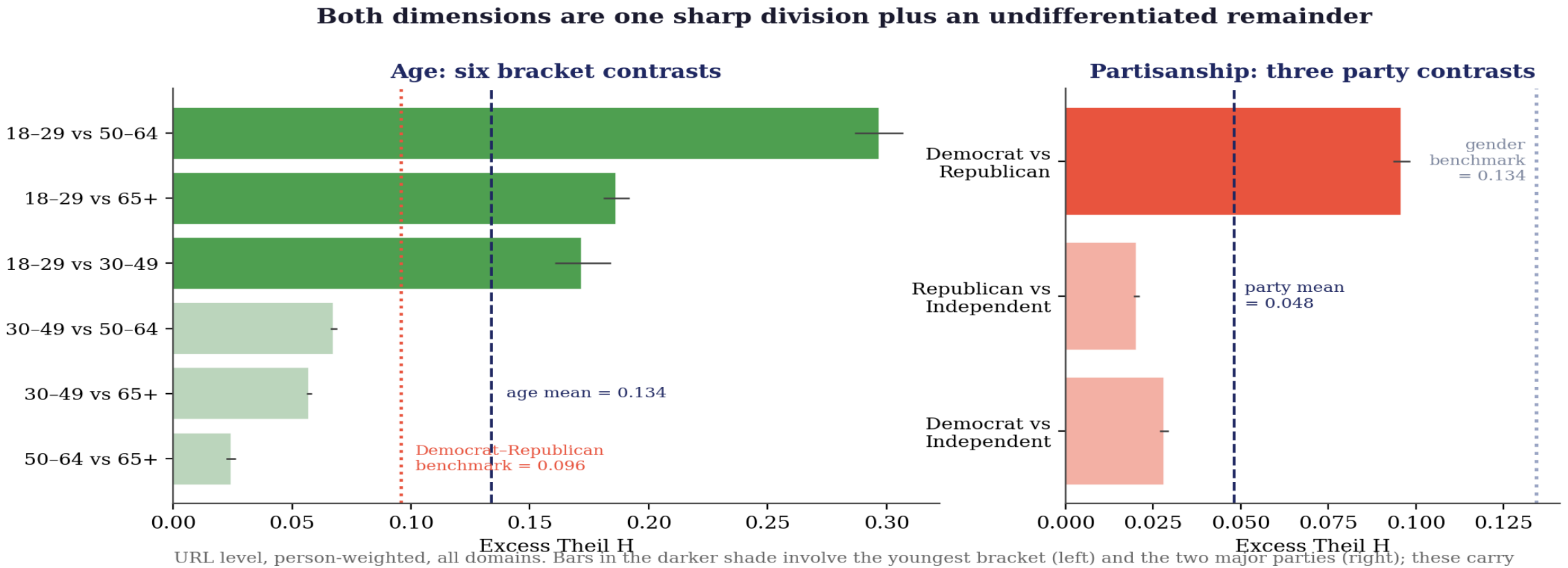
Backup: why category count must be corrected



All domains, person-weighted, minimum two distinct visitors per source. Both measures agree at two categories; they diverge as resolution rises because people equidistant from the national centroid on opposite coasts share a quantile bin, whereas *k*-means groups by genuine mutual proximity.

Excess Theil H falls mechanically as categories multiply, holding content and filter fixed. This is why age, gender, partisanship and geography are all reduced to a common two-group footing.

Backup: person vs. visit weighting

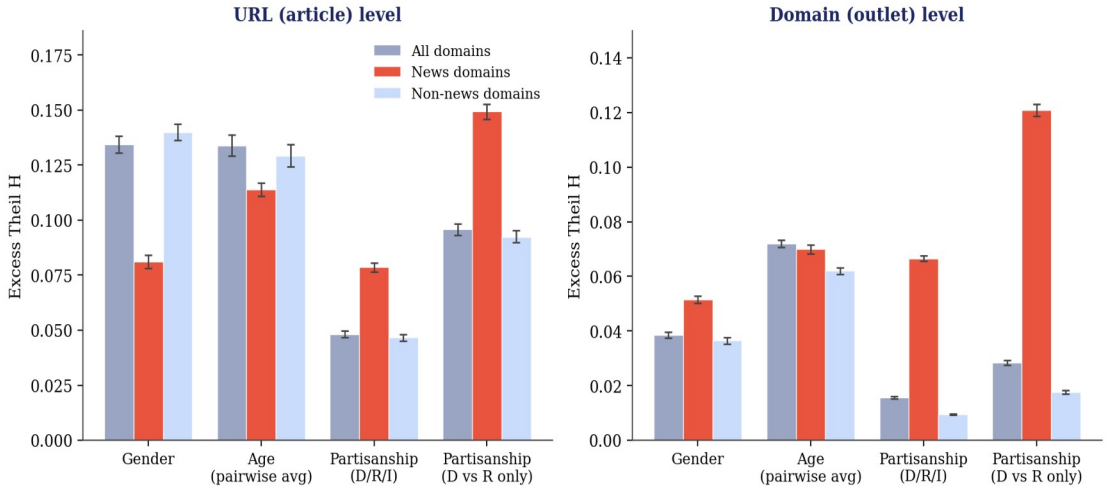


URL level, person-weighted, all domains. Bars in the darker shade involve the youngest bracket (left) and the two major parties (right); these carry the dimension's aggregate strength. Dotted lines mark the cross-dimension benchmark: every age contrast involving the youngest bracket exceeds the Democrat–Republican value, and no party contrast except Democrat–Republican reaches gender. Error bars are 95% bootstrap intervals across the 59 months.

Visit weighting raises every dimension, sometimes by close to half, without reordering which is largest.

Backup: geography verified three ways

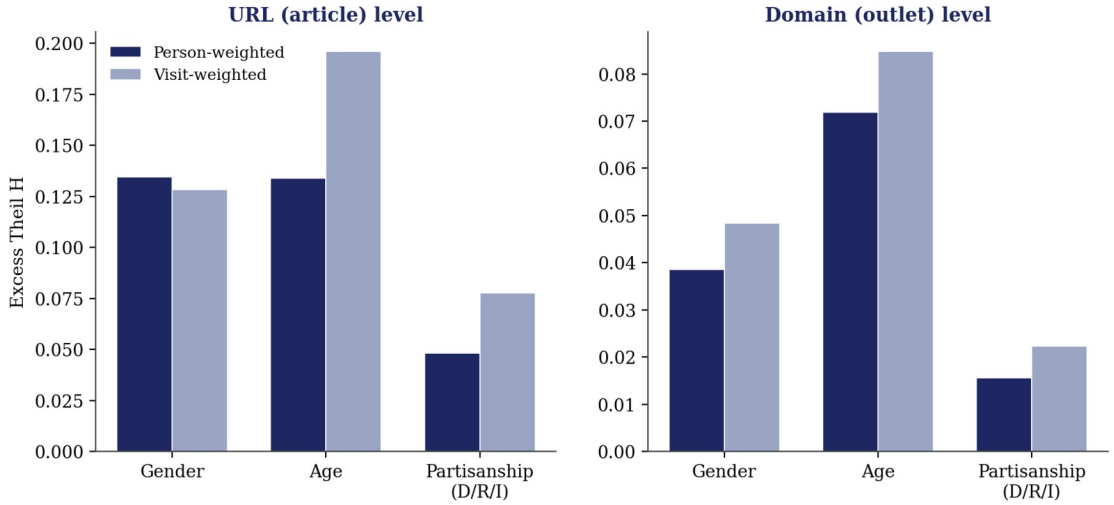
Partisan segregation rises in news only for the Democrat-Republican contrast; demographic segregation falls



Person-weighted, minimum two distinct visitors per source. Error bars: 95% bootstrap CI across 59 months

k-means on raw coordinates

Weighting by visits rather than people raises every dimension, but does not reorder them

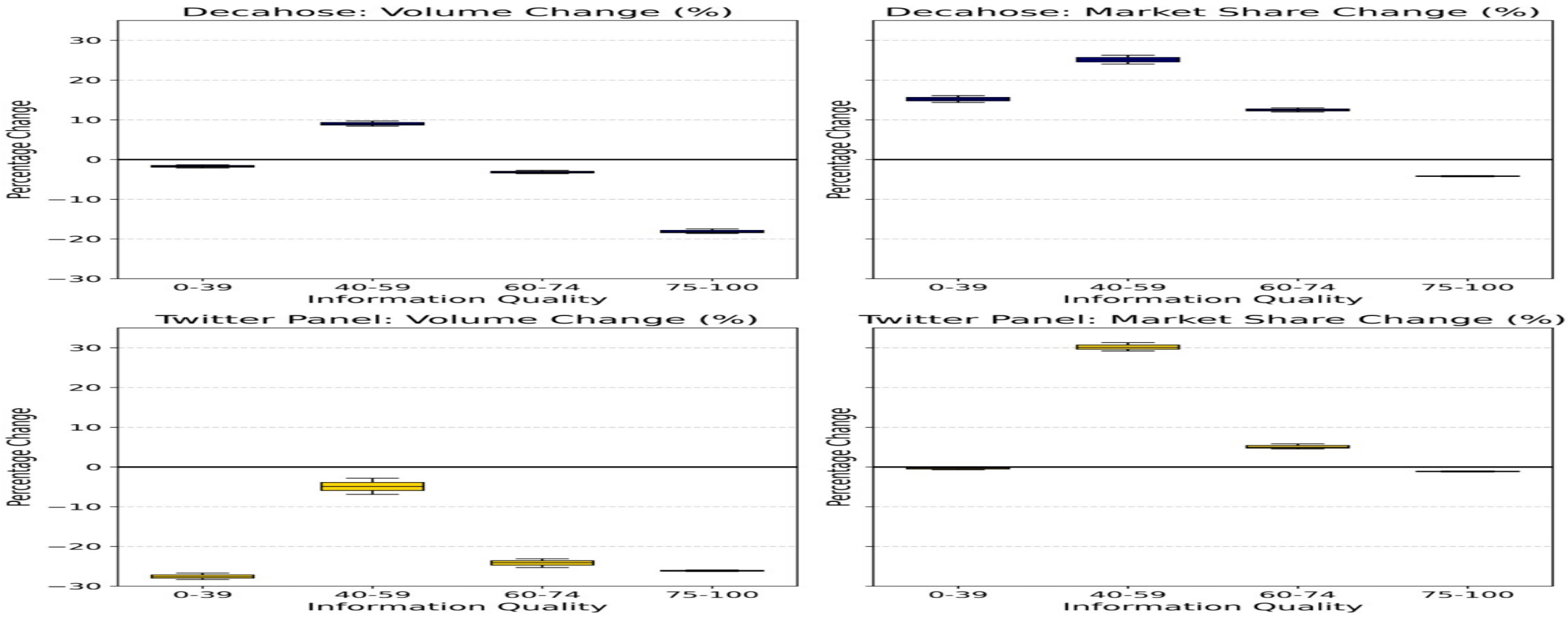


All domains, minimum two distinct visitors per source. Age and partisanship are averaged over their pairwise group contrasts.

Pairwise kilometres, no index

The clustering check varies the partition while keeping the index; the distance check keeps neither. At four clusters, matched to the four Census regions, data-driven and administrative partitions return overlapping intervals in every cell. All three agree.

Backup: quality shift by credibility band



Volume and market share change across all four NewsGuard bands, for Decahose (top) and Panel (bottom). Largest Decahose volume decline is the 75–100 group (−17.18%); largest market-share gain is 40–59.

Backup: why the panel is the right instrument

- Attributes observed, not inferred. Age, gender, party registration, and address come from the voter file, avoiding the selection that behavior-inferred partisanship introduces.
- A defined population. Registered voters have a known relationship to democratic participation, which makes the representativeness question meaningful.
- Clean by construction. Bots, brands, and non-US accounts are excluded because every user is a matched voter.
- Known biases. Linkage favors rare names; the panel overrepresents White users and underrepresents Hispanic and Asian users.