Text Analysis for Economics and Finance

Ruben Durante
ICREA-UPF, BGSE, IPEG, CEPR

DIW, October 2020

Supervised methods: measuring media bias

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Supervised methods: measuring media bias

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Supervised methods: measuring media bias

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

Supervised methods: measuring media bias

- How to measure partisan bias in the media?
- Mass media can slant the news to favor a particular point of view (*partisan bias*)
- Forms of partisan bias:
  - Unbalanced reporting of political events (language, citations, etc.)
  - More newstime devoted to like-minded politicians/experts
  - More emphasis on issues on which a party is perceived as stronger
  - More emphasis on bad performance or scandals of opposing party
- In any case, a very elusive object!
- Important to study other questions, e.g., what drives media bias?

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological score for major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, no letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological score for major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, no letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

A Measure of Media Bias (Groseclose-Milyo, 2005)

- Goal:
  - Estimate relative ideological score for major U.S. media outlets
- Empirical Idea:
  - Count the times a media outlet cites various think tanks and compare with the times members of Congress cite the same groups
  - Citations are the language, politicians the reference point
  - Only look at content (no editorials, no letters, etc.)
- Findings:
  - Strong liberal bias: all news outlets but two (Fox News, Washington Times) have scores to the left of the average member of Congress
  - Most leftist: CBS Evening News, NYT
  - Most centrist: PBS NewsHour, CNN's Newsnight, ABC's Good Morning America, USA Today

Data

- List of 200 of the most prominent US think tanks
- Search the *Congressional Record* for citations of these think tanks (1993-2002)
- Measure ideology of congressmen who cited think thanks using ADA score (0-conservative to 100-liberal)
- Omit cases where a politician/media outlet mentions a think tank just to criticize it or give it an ideological label
- *"The idea is that we only wanted cases where the legislator/journalist cited the think tank as if it were a disinterested expert on the topic"*

Data

- List of 200 of the most prominent US think tanks
- Search the *Congressional Record* for citations of these think tanks (1993-2002)
- Measure ideology of congressmen who cited think thanks using ADA score (0-conservative to 100-liberal)
- Omit cases where a politician/media outlet mentions a think tank just to criticize it or give it an ideological label
- *"The idea is that we only wanted cases where the legislator/journalist cited the think tank as if it were a disinterested expert on the topic"*

Methodology

- Dimensionality:
  - X is counts of citations to think tanks
  - Choose ex ante criterion for feature selection
- Training set: U.S. Congress
  - Outcome Y is ideology score based on roll-call voting
- Target: U.S. Media
  - Count references to think tanks in articles and compute $\hat{Y}$
- Dimension of data: $p = 50$ (keep 44, collapse others in 6 groups)
- Dimension of training data: $n = 535$

Methodology (cont.)

- Define a generative model for Y
- Let y_i be the ADA score of senator i
- Then:

$$Pr(x_{ijt} = 1) = \frac{exp(\alpha_j + \beta_j y_i)}{\sum_{j'} exp(\alpha_{j'} + \beta_{j'} y_i)}$$

- Assume same model for news media m but treat y_m as unknown
- Estimate α_j, β_j, and y via joint maximum likelihood

Think Tanks

TABLE I
THE 50 MOST-CITED THINK TANKS AND POLICY GROUPS
BY THE MEDIA IN OUR SAMPLE

	Think tank/policy group	Average score of legislators who cite the group	Number of citations by legislators	Number of citations by media outlets
1	Brookings Institution	53.3	320	1392
2	American Civil Liberties Union	49.8	273	1073
3	NAACP	75.4	134	559
4	Center for Strategic and International Studies	46.3	79	432
5	Amnesty International	57.4	394	419
6	Council on Foreign Relations	60.2	45	403
7	Sierra Club	68.7	376	393
8	American Enterprise Institute	36.6	154	382
9	RAND Corporation	60.4	352	350
10	National Rifle Association	45.9	143	336
11	American Association of Retired Persons	66.0	411	333
12	Carnegie Endowment for International Peace	51.9	26	328
13	Heritage Foundation	20.0	369	288
14	Common Cause	69.0	222	287
15	Center for Responsive Politics	66.9	75	264
16	Consumer Federation of America	81.7	224	256
17	Christian Coalition	22.6	141	220
18	Cato Institute	36.3	224	196
19	National Organization for Women	78.9	62	195
20	Institute for International Economics	48.8	61	194
21	Urban Institute	73.8	186	187
22	Family Research Council	20.3	133	160
23	Federation of American Scientists	67.5	36	139
24	Economic Policy Institute	80.3	130	138
25	Center on Budget and Policy Priorities	88.3	224	115
26	National Right to Life Committee	21.6	81	109
27	Electronic Privacy Information Center	57.4	19	107
28	International Institute for Strategic Studies	41.2	16	104
29	World Wildlife Fund	50.4	130	101
30	Cent. for Strategic and Budgetary Assessments	33.9	7	89

Politicians

TABLE II
AVERAGE ADJUSTED ADA SCORES OF LEGISLATORS

Legislator	Average score
Maxine Waters (D-CA)	99.6
Edward Kennedy (D-MA)	88.8
John Kerry (D-MA)	87.6
Average Democrat	84.3
Tom Daschle (D-SD)	80.9
Joe Lieberman (D-CT)	74.2
Constance Morella (R-MD)	68.2
Ernest Hollings (D-SC)	63.7
John Breaux (D-LA)	59.5
Christopher Shays (R-CT)	54.6
Arlen Specter (R-PA)	51.3
James Leach (R-IA)	50.3
Howell Heflin (D-AL)	49.7
Tom Campbell (R-CA)	48.6
Sam Nunn (D-GA)	48.0
Dave McCurdy (D-OK)	46.9
Olympia Snowe (R-ME)	43.0
Susan Collins (R-ME)	39.3
Charlie Stenholm (D-TX)	36.1
Rick Lazio (R-NY)	35.8
Tom Ridge (R-PA)	26.7
Nathan Deal (D-GA)	21.5
Joe Scarborough (R-FL)	17.7
Average Republican	16.1
John McCain (R-AZ)	12.7
Bill Frist (R-TN)	10.3
Tom DeLay (R-TX)	4.7

Results

TABLE III
RESULTS OF MAXIMUM LIKELIHOOD ESTIMATION

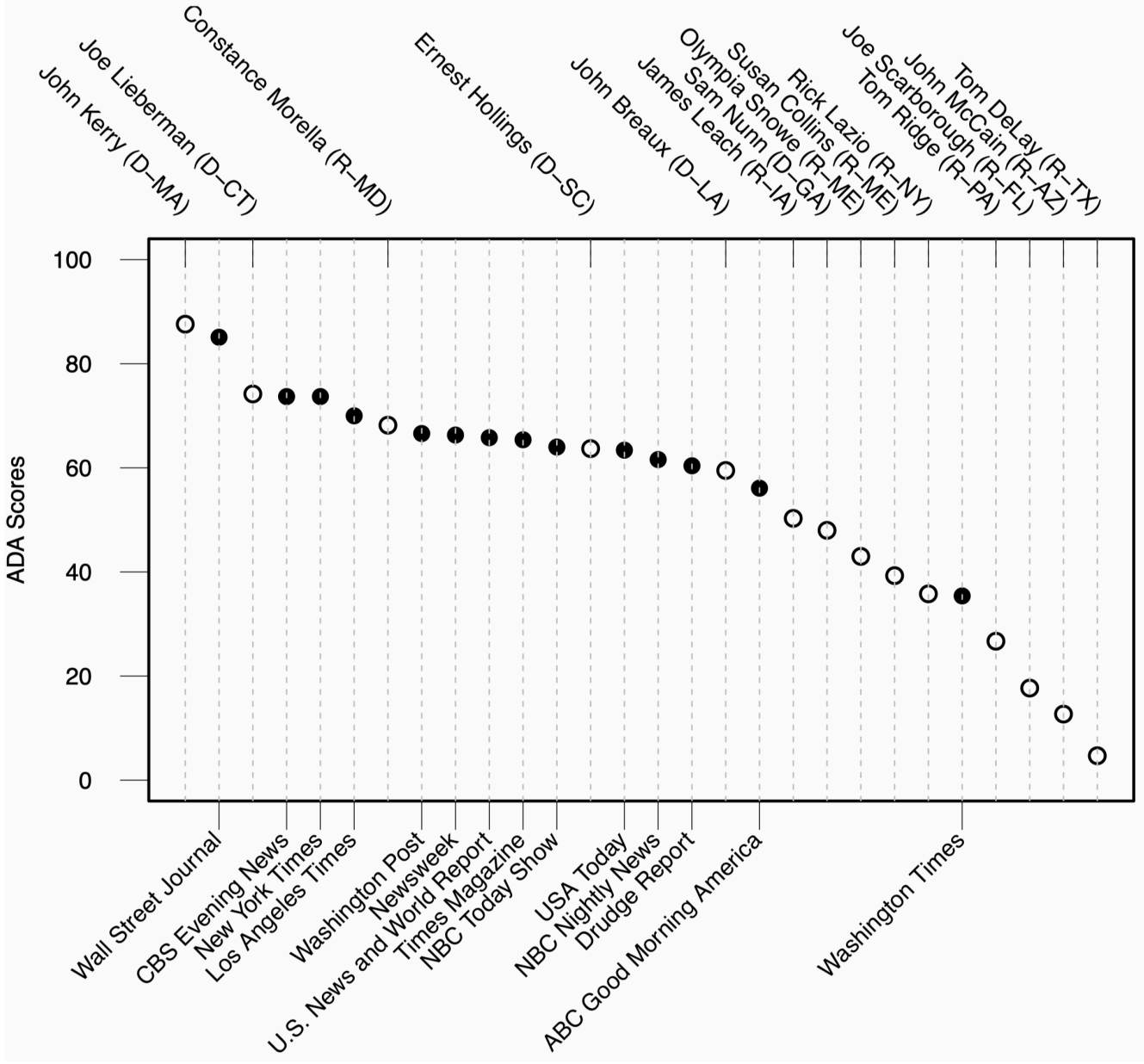
Media outlet	Period of observation	Estimated ADA score	Standard error
ABC Good Morning America	6/27/97– 6/26/03	56.1	3.2
ABC World News Tonight	1/1/94– 6/26/03	61.0	1.7
CBS Early Show	11/1/99– 6/26/03	66.6	4.0
CBS Evening News	1/1/90– 6/26/03	73.7	1.6
CNN NewsNight with Aaron Brown	11/9/01– 2/5/04	56.0	4.1
Drudge Report	3/26/02– 7/1/04	60.4	3.1
Fox News' Special Report with Brit Hume	6/1/98– 6/26/03	39.7	1.9
Los Angeles Times	6/28/02–12/29/02	70.0	2.2
NBC Nightly News	1/1/97– 6/26/03	61.6	1.8
NBC Today Show	6/27/97– 6/26/03	64.0	2.5
New York Times	7/1/01– 5/1/02	73.7	1.6
Newshour with Jim Lehrer	11/29/99– 6/26/03	55.8	2.3
Newsweek	6/27/95– 6/26/03	66.3	1.8
NPR Morning Edition	1/1/92– 6/26/03	66.3	1.0
Time Magazine	8/6/01– 6/26/03	65.4	4.8
U.S. News and World Report	6/27/95– 6/26/03	65.8	1.8
USA Today	1/1/02– 9/1/02	63.4	2.7
Wall Street Journal	1/1/02– 5/1/02	85.1	3.9
Washington Post	1/1/02– 5/1/02	66.6	2.5
Washington Times	1/1/02– 5/1/02	35.4	2.7
Average		62.6	

Results (cont.)

Rankings Based on Distance from Center

Rank	Media outlet	Estimated ADA score
1	*Newshour with Jim Lehrer*	55.8
2	*CNN NewsNight with Aaron Brown*	56.0
3	*ABC Good Morning America*	56.1
4	*Drudge Report*	60.4
5	*Fox News' Special Report with Brit Hume*	39.7
6	*ABC World News Tonight*	61.0
7	*NBC Nightly News*	61.6
8	*USA Today*	63.4
9	*NBC Today Show*	64.0
10	*Washington Times*	35.4
11	*Time Magazine*	65.4
12	*U.S. News and World Report*	65.8
13	*NPR Morning Edition*	66.3
14	*Newsweek*	66.3
15	*CBS Early Show*	66.6
16	*Washington Post*	66.6
17	*Los Angeles Times*	70.0
18	*CBS Evening News*	73.7
19	*New York Times*	73.7
20	*Wall Street Journal*	85.1

Results (cont.)



What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the bigrams and trigram most representative of *Republicans* and *Democrats*
  - Compute relative frequency of "Democratic" and "Republican" expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers' vs. owners' preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the bigrams and trigram most representative of *Republicans* and *Democrats*
  - Compute relative frequency of "Democratic" and "Republican" expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers' vs. owners' preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

What Drives Media Slant (Gentzkow-Shapiro, 2010)

- Goal:
  - Construct a measure of media slant based on the similarity of the language used by newspapers and politicians
- Methodology:
  - Consider official speeches by US congressmen
  - Identify the bigrams and trigram most representative of *Republicans* and *Democrats*
  - Compute relative frequency of "Democratic" and "Republican" expressions in the articles published in over 400 newspapers
  - Define the slant of each newspaper with respect to politicians
  - Test whether slant is driven by consumers' vs. owners' preferences
- Findings:
  - Slant highly correlated with political leaning of potential readers
  - Identity of media owner does not explain much

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency (instead of ADA)
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (MSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency (instead of ADA)
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (MSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Data

- Politicians:
  - All speeches from 2005 *Congressional Record*
  - Ideological score: vote share to Bush in 2004 presidential election in congressperson's constituency (instead of ADA)
- News content:
  - Headlines and text of all articles published on 433 US daily newspapers in 2005
  - Source: Newslibrary, ProQuest
  - Only news articles, no editorials
- Other:
  - Newspaper HQ location and relevant market (MSA)
  - Vote shares for Republicans and Democrats in relevant media market
  - Identity of newspaper's owner

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #1

- Pre-process the *Congressional Record* corpus: remove stop words, stemming (Porter)
- Consider all 2-word and 3-word phrases appearing in the corpus
- For each phrase p of length l, compute the total number of times it is used by Democrats and Republicans (f_{pld}, f_{plr})
- For each phrase compute the Pearson's χ^2 statistic:

$$\chi^2_{pl} = \frac{(f_{plr} f_{\sim pld} - f_{pld} f_{\sim plr})^2}{(f_{plr} + f_{pld})(f_{plr} + f_{\sim plr})(f_{pld} + f_{\sim pld})(f_{\sim plr} + f_{\sim pld})}$$

- χ^2: test statistic for the null hypothesis that the propensity to use phrase p is equal for Democrats and Republicans.
- Simple to compute, only requires f_{pld} and f_{plr}. Preferable to other naive statistics such as the ratio of uses by D or R

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.
- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)
- 2-word phrases that appeared more than 15,000 times in headlines (e.g., "third quarter" or "exchange rate")
- 3-word phrases that appeared less than 5 times in headlines
- 3-word phrases that appeared more than 1,000 times in headlines
- Any phrase that appeared in the full text of more than 400,000 documents
- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.
- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)
- 2-word phrases that appeared more than 15,000 times in headlines (e.g., "third quarter" or "exchange rate")
- 3-word phrases that appeared less than 5 times in headlines
- 3-word phrases that appeared more than 1,000 times in headlines
- Any phrase that appeared in the full text of more than 400,000 documents
- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Methodology: step #2

- Eliminate phrases that are not likely to be useful for diagnosing newspaper partisanship.
- 2-word phrases appearing less than 200 times in newspaper headlines (e.g., procedural phrases not used by the media)
- 2-word phrases that appeared more than 15,000 times in headlines (e.g., "third quarter" or "exchange rate")
- 3-word phrases that appeared less than 5 times in headlines
- 3-word phrases that appeared more than 1,000 times in headlines
- Any phrase that appeared in the full text of more than 400,000 documents
- Among the remaining ones, select the 500 phrases of each length with the highest values of χ^2_{pl}

Most representative Democratic phrases

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases

private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife

Three-Word Phrases

veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

Most representative Republican phrases

Panel B: Phrases Used More Often by Republicans

Two-Word Phrases

stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program

Three-Word Phrases

embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #3

- Use politicians' language and ideology to map phrases to ideology
- Re-index the phrases in the sample by $p \in [1, ..., 1000]$
- For each congressperson $c \in C$ we observe ideology y_c and phrase frequencies $\{f_{pc}\}_{p=1}^{1000}$
- Compute relative frequencies as: $\tilde{f}_{pc} = f_{pc} / \sum_{p=1}^{P} f_{pc}$
- For each phrase p, we regress $\tilde{f}_{pc}$ on y_c for the sample of congresspeople, obtaining an intercept and a slope parameter, a_p and b_p
- Hence we estimate 1,000 separate regressions, each with a sample the size of C
- Intuition: the larger b_p the more the use of a phrase is correlated with ideology (y_c)

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

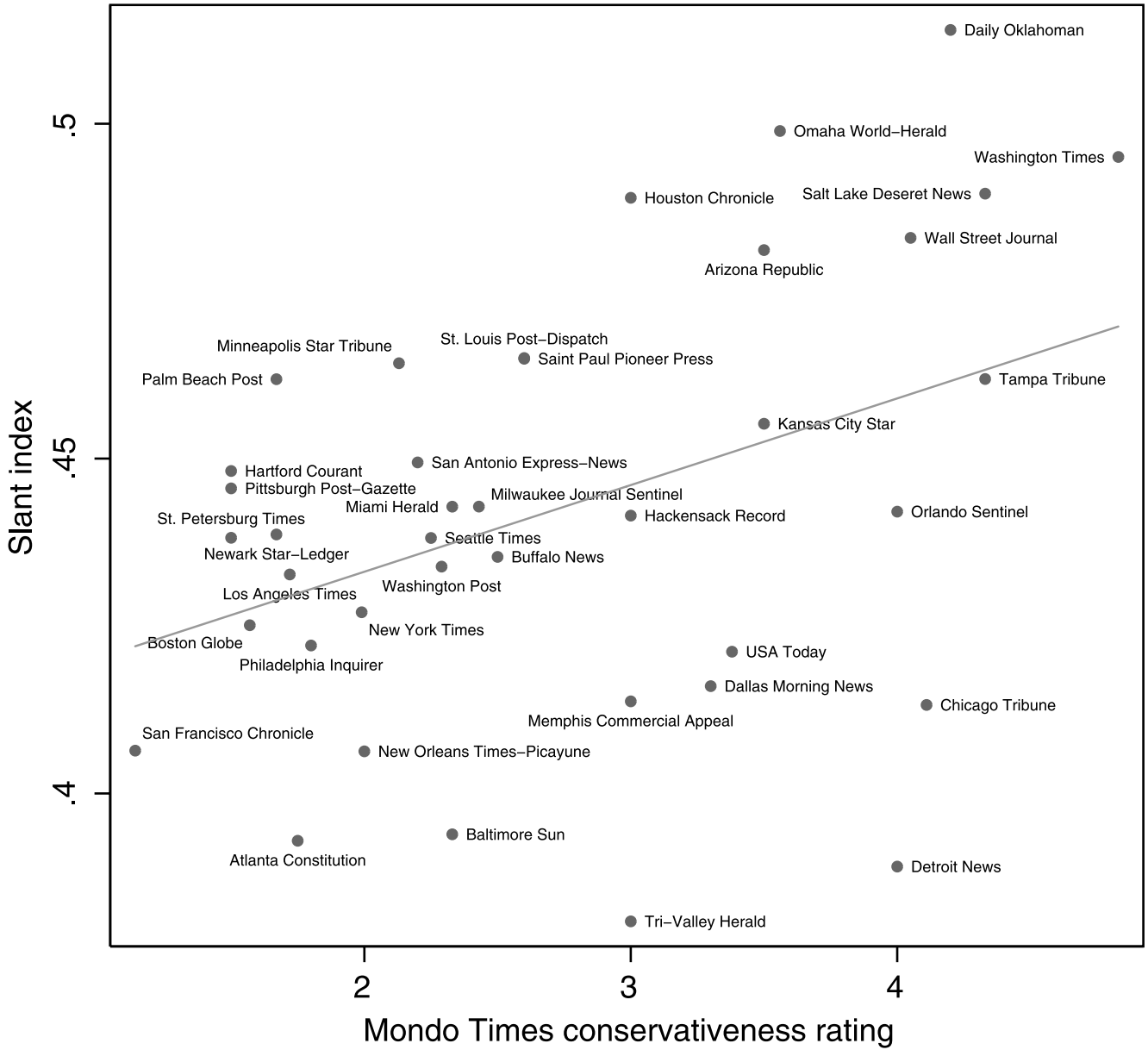
Methodology: step #4

- We want to assign each paper a measure of slant based on the frequency of phrases it uses and their ideological valence
- For each paper $n \in N$, we observe the relative frequency for each phrase $\tilde{f}_{pc}$, but not the ideology y_n, which we want to estimate
- For each newspaper n we regress $(\tilde{f}_{pn} - a_p)$ on b_p for the sample of phrases, obtaining the slope estimate:

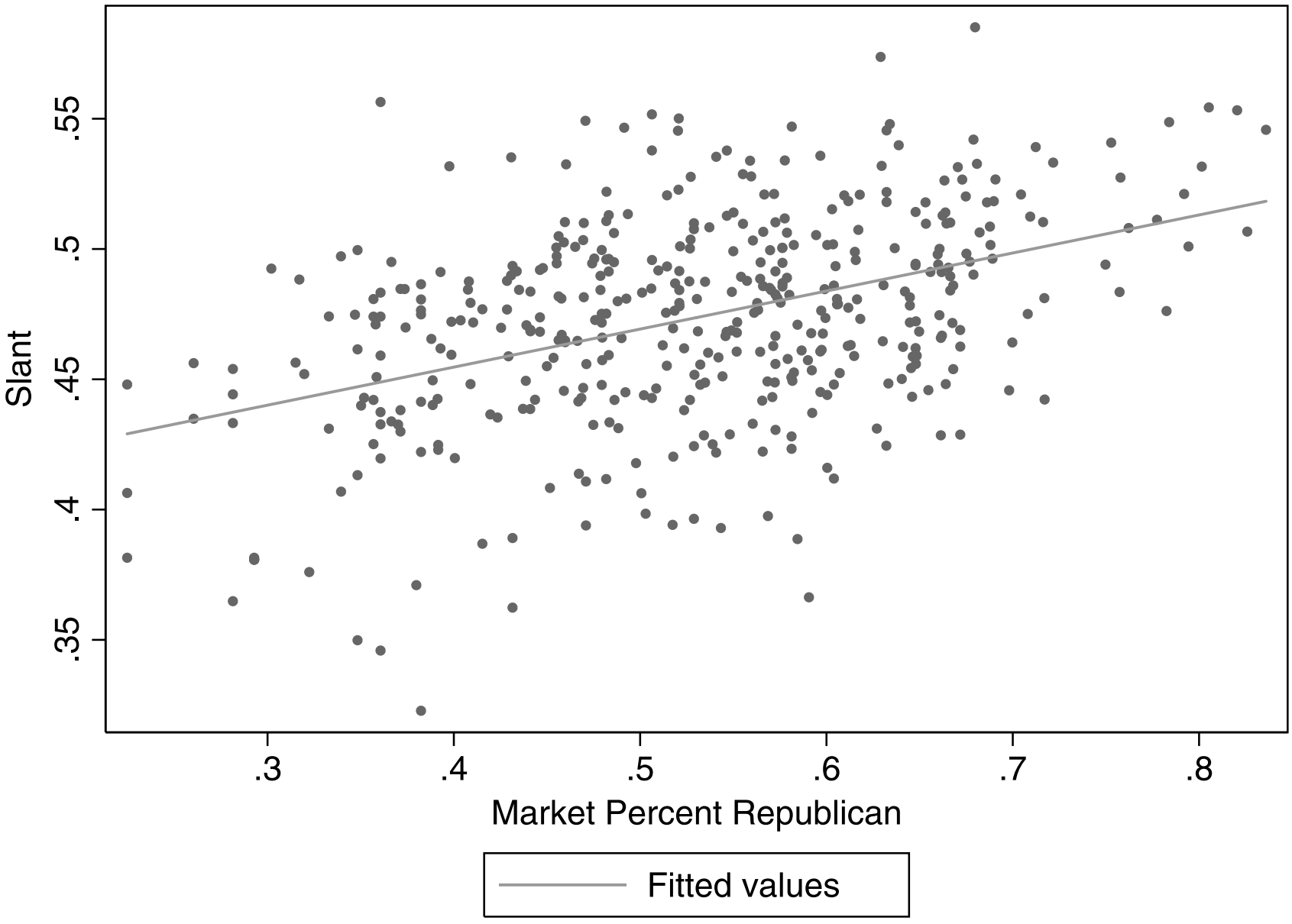
$$\hat{y}_n = \frac{\sum_{p=1}^{1000} b_p(\tilde{f}_{pn} - a_p)}{\sum_{p=1}^{1000} b_p^2}$$

- We estimate N separate regressions, each with a sample of 1,000
- Intuition: the higher the frequency $(\tilde{f}_{pn})$ of more ideological phrases (b_p), the higher the measure of slant $(\hat{y}_n)$

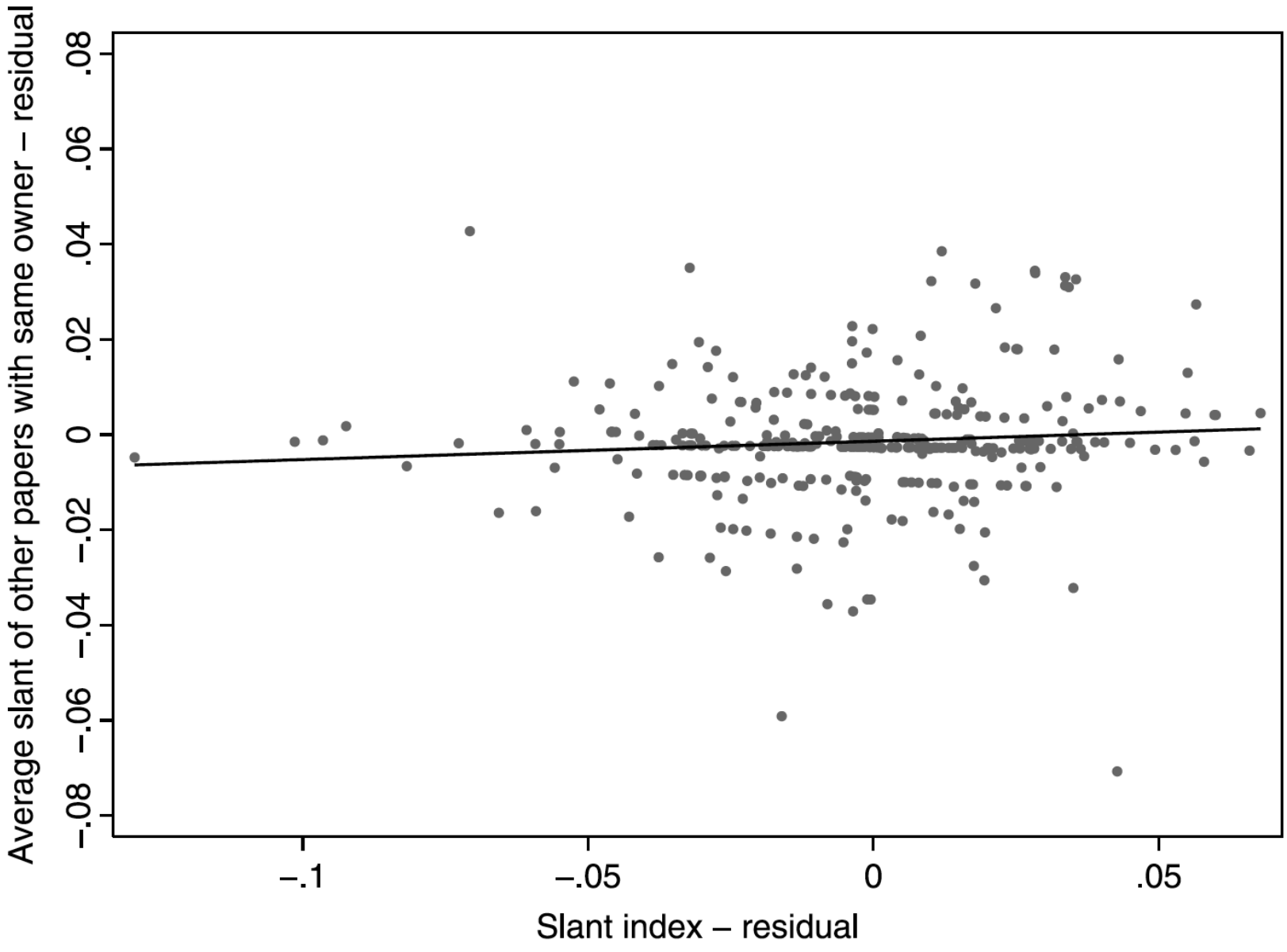
Validating the measure of slant



Using the measure: slant and readers' ideology



Using the measure: slant and owners' ideology



Using the measure: slant and owners' ideology

TABLE IV

ECONOMIC INTERPRETATION OF MODEL PARAMETERS[a]

Quantity	Estimate
Actual slant of average newspaper	0.4734
	(0.0020)
Profit-maximizing slant of average newspaper	0.4600
	(0.0047)
Percent loss in variable profit to average newspaper	0.1809
from moving 1 SD away from profit-maximizing slant	(0.1025)
Share of within-state variance in slant from consumer ideology	0.2226
	(0.0406)
Share of within-state variance in slant from owner ideology	0.0380
	(0.0458)

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Production of Information in an Online World (Cage et al., 19)

- Goal:
  - Study how much online media outlets copy content from each other in the news production process
- Methodology:
  - Consider all online news content produced by French media in 2013
  - Identify 25K news events with an event-detection algorithm
  - Identify first news item that breaks news about an event
  - Measure how much copying is used by subsequent news stories related to event with plagiarism detection algorithm
  - Measure effects of copying on readership/audience
- Findings:
  - Online copying in news production is widespread: 61.8% of content presents some form of copying
  - Producing original content is rewarded with larger viewership shares

Data

- Online News Content:
  - More than 2.5 million French news articles published online in 2013 (7K/day)
  - Transmedia approach: content from 86 media outlets including 1 news agency (AFP), 59 newspapers, 10 online-only media outlets, 7 radio stations, 9 TV channels
  - Source: French National Audiovisual Institute (public company)
- Viewership:
  - Daily audience measures for 58 out of the 86 outlets (AFP and some local newspapers not covered)
  - Number of shares of the article on Facebook and Twitter. Proxy for the number of views of an article

Methodology: step #1 event detection algorithm

- Consider headline and text of each article and compute its TF-IDF vector representation
- Compute the cosine similarity of each article-pair
- Iteratively aggregate articles into event-clusters if the cosine similarity is above a certain threshold (determined manually)
- "Close" an event if no article is aggregated to it within a 24-hour window
- Drop events that contain less than 2 distinct media outlets and less than 10 articles.
- Results in 25,215 news events, each lasting about 41 hours
- 33.4% of of the 2.5M articles are classified into an event
- Very large clusters are mostly "garbage clusters"

Methodology: step #2 plagiarism detection algorithm

- Consider all the articles within an event-cluster and order them by time. The first one is the news-breaking article.
- Define the reaction time of an article as the time elapsed since publication of the news-breaking article
- Compare the text of each article to that of all the preceding articles in the event-cluster
- If a portion of text of at least 100 characters in the article is identical to any previous portion that is already published, then that portion is defined as a copy
- For each article, calculate the originality rate, defined as:

$$\text{originality rate} = \frac{\#\text{ original characters}}{\#\text{ total characters}}$$

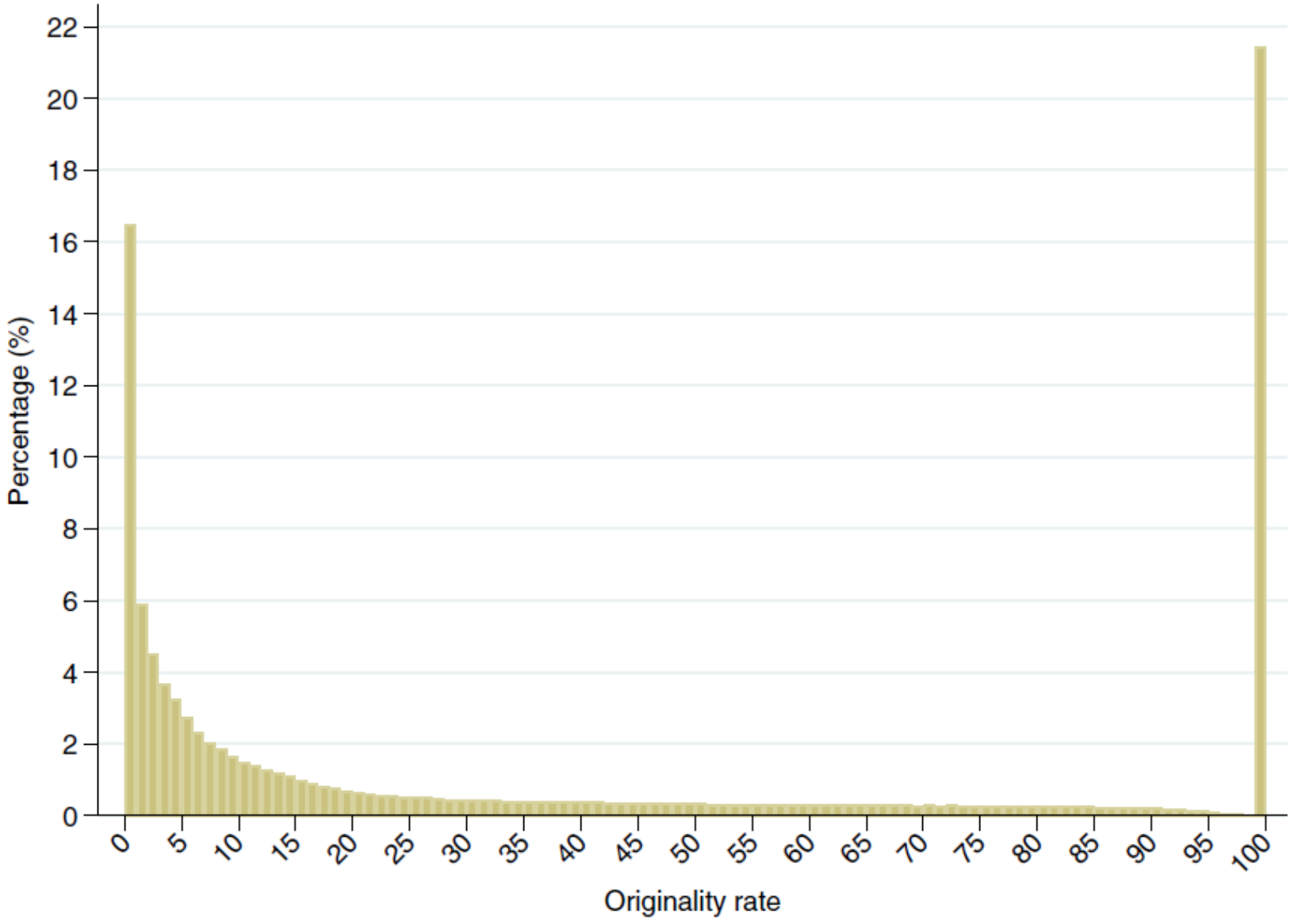
Summary statistics

TABLE 1
Summary statistics: articles (classified in events)

	Mean	Median	SD	Min	Max
Content					
Length (number of characters)	2,467	2,192	1,577	100	98,340
Original content (number of characters)	805	253	1,287	1	53,424
Non-original content (number of characters)	1,661	1,326	1,539	0	48,374
Originality (%)	36.5	14.5	39.8	0	100
Reactivity in hours	41.7	19.1	65.2	0	6,257
Audience					
Number of shares on Facebook	64	0	956	0	240,450
Number of shares on Facebook (winsorized)	37	0	136	0	1,017
Number of shares on Twitter	9	0	42	0	11,908
Number of shares on Twitter (winsorized)	7	0	19	0	126
Obs	851,864				

Notes: The table gives summary statistics. Year is 2013. Variables are values for the articles classified in events. The observations are at the article level. The "Number of shares on Facebook (winsorized)" variable is the version of the Facebook variable winsorized at the 99th percentile. Similarly, the "Number of shares on Twitter (winsorized)" variable is the version of the Twitter variable winsorized at the 99th percentile. Variables are described in more details in the text.

Originality rate distribution



Positive relationship between originality and reaction time

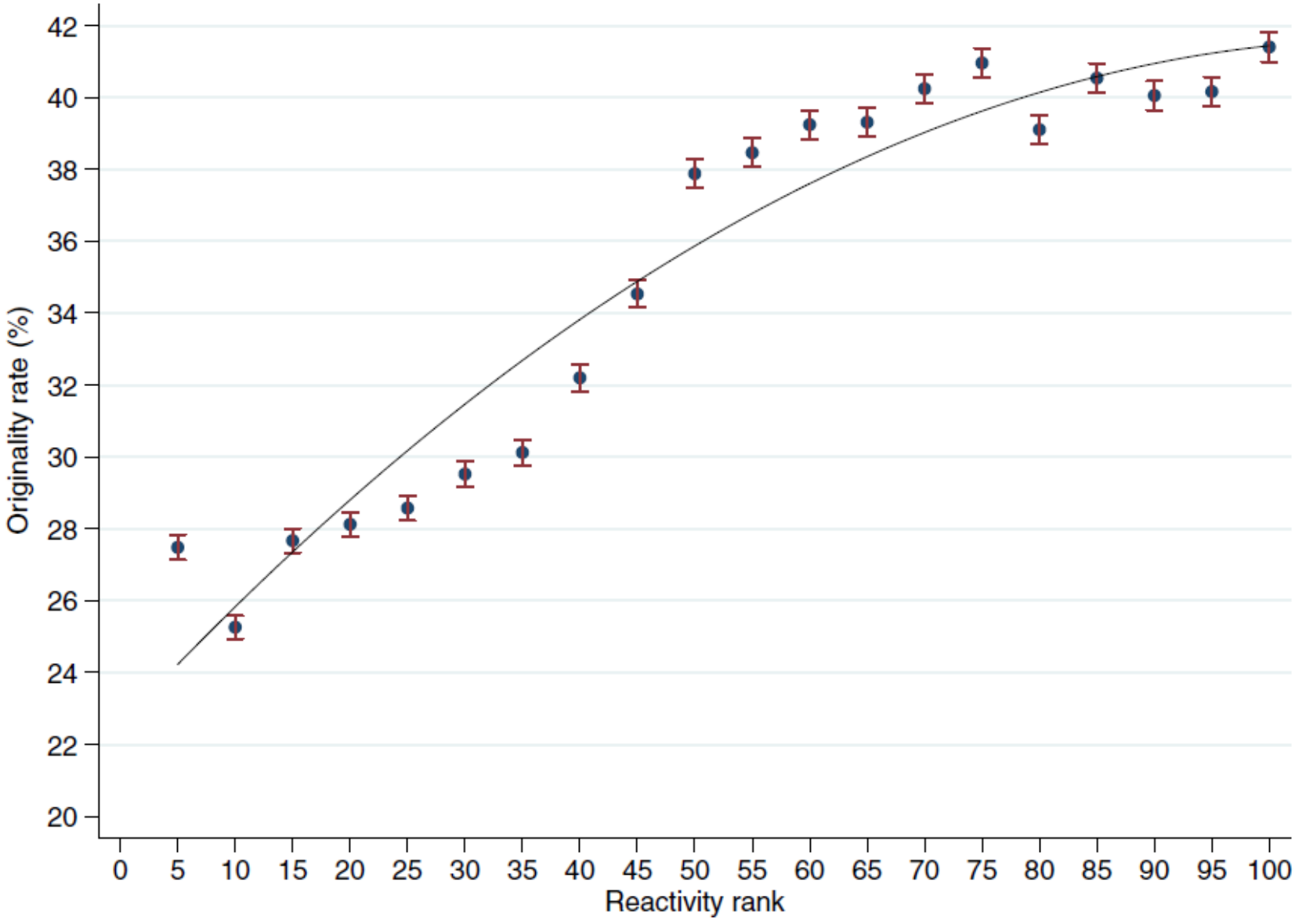


FIGURE 3

Correlation between originality and reaction time: average originality rate depending on the reactivity rank

Methodology: step #3 strategies to proxy article views

- "Naive" approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

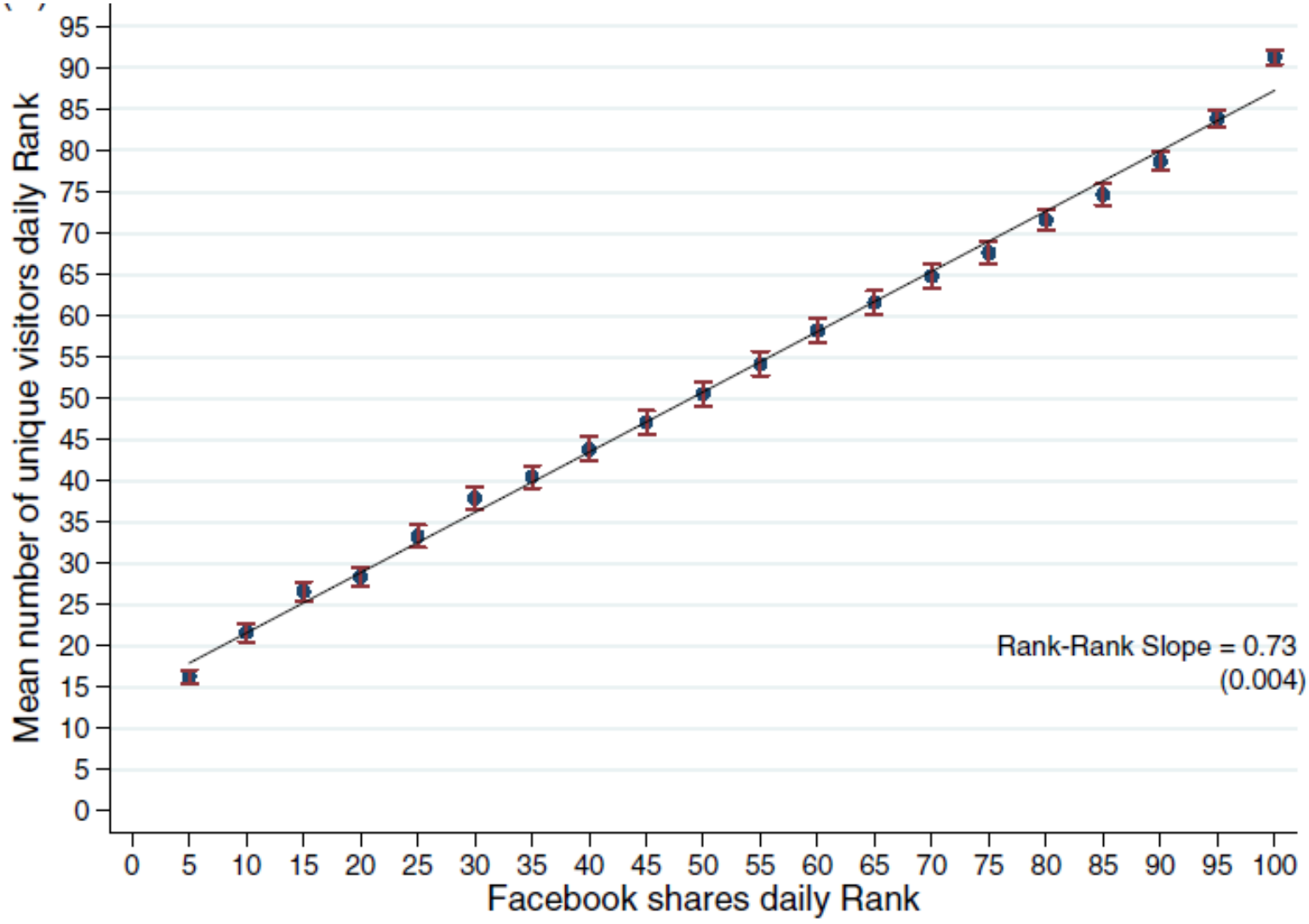
Methodology: step #3 strategies to proxy article views

- "Naive" approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

Methodology: step #3 strategies to proxy article views

- "Naive" approach:
  - On a given day and for a given outlet, assume all articles are equally popular
  - Number of views is naively defined as number of page views on website divided by number of articles
- Linear approach:
  - Number of article views is proportional to the relative number of shares of the article, within its outlet-day pair
- Social media approach:
  - Collect data from leading French newspaper *Le Monde* on article views from April to August 2017
  - Link the URL of the online article to Facebook (resp. Twitter) data in order to identify the number of shares on social media
  - Examine correlation between the number of views on *Le Monde* website and the number of social media shares
  - Use relationship to infer the number of views for other outlets

Social media shares and actual online views



Association between number of Unique visitors' and Facebook shares' Percentile Ranks

Positive relationship between popularity and originality

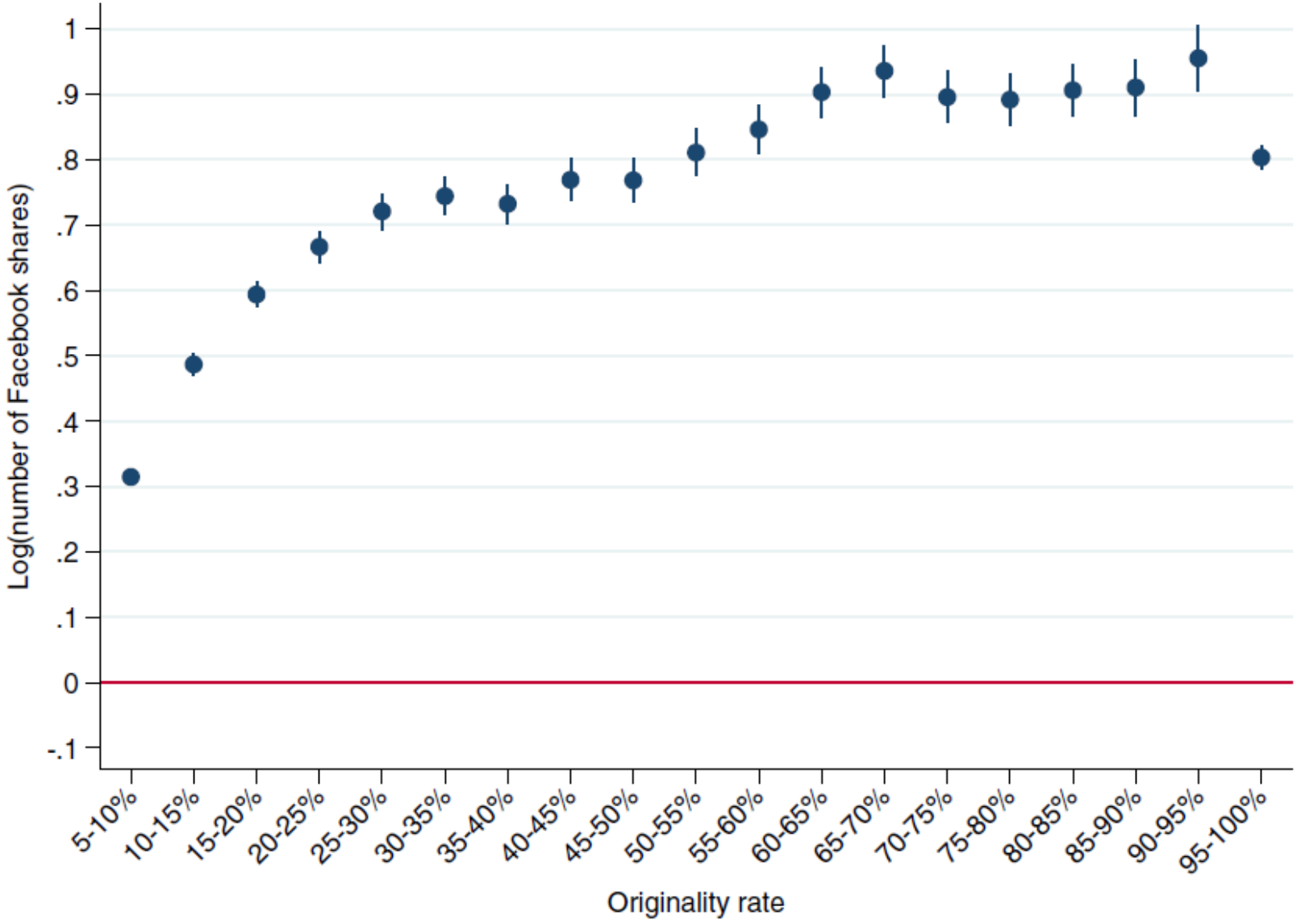


FIGURE 5

Facebook shares and originality rate

Share of original content in articles data

$$\frac{\sum_a \text{original content}_a \cdot \text{number of views}_a}{\sum_a \text{original content}_a \cdot \text{number of views}_a + \sum_a \text{non-original content}_a \cdot \text{number of views}_a}$$

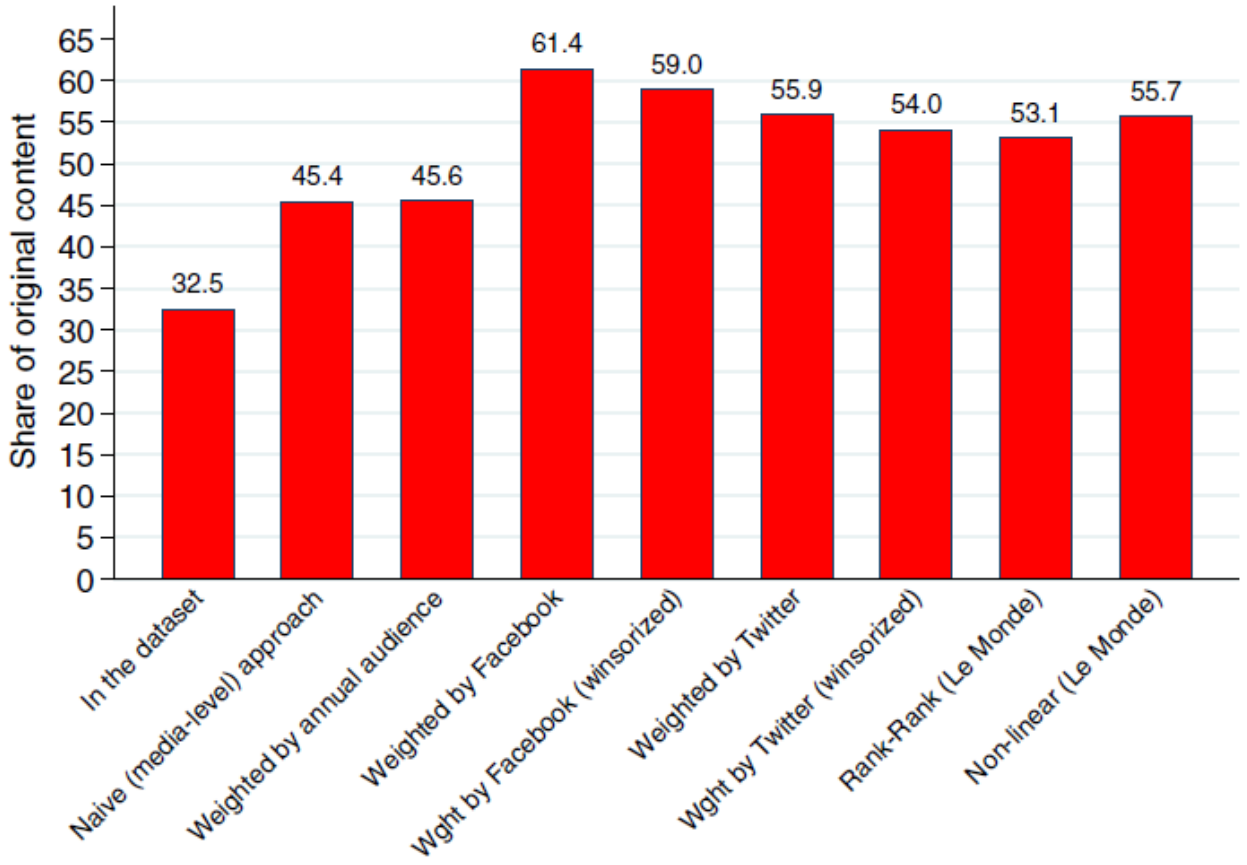


FIGURE 6

The audience-weighted share of original content