

Media Diversity and the Concentration of Online Attention

Matthew Hindman

*Doctoral Fellow, National Center for Digital Government
John F. Kennedy School of Government, Harvard University*

How Open is the Web?

- It is commonly assumed that the Web is far more open and diverse than traditional media
 - A generation ago, media limited to three television networks, handful of radio stations, and a single hometown paper
 - Now anyone can put up a Web page
- Reality much more complicated:
 - Millions and millions of sites online...
 - ...and yet everyone chooses the same ones.
- 50% of traffic goes to the top 0.05% of Web sites

The Link Structure of the Web

Consider an online community where links are distributed as follows:

- 1 site has 1,000,000 links
- 10 sites have at least 10,000 links
- 100 sites have at least 100 links
- 1,000 sites have at least 1 link

This is a ***Power Law*** distribution

- Probability that a site has K links proportional to K^{-a}

Over the entire Web inbound links follow a power law, with $a = 2.1$ for inbound links

(Barabasi 1999, Kumar 1999, Lawrence 1998)

Why Do We Care?

The number of inbound links pointing to a site is a good proxy for its visibility on the open Web

- Two ways to find new information online:
 - Surfing away from known sites
 - Search tools (Google, Yahoo! directories, etc.)
- Both methods funnel traffic to sites with large numbers of inbound links
- Number of inbound links and number of visitors to a site are highly correlated
 - Correlation Coefficient = .702

Methodology

Four Steps:

1. Create lists of 200 highly-ranked sites in a variety of political categories from Google and the Yahoo! Directory
 - “seed sites”
2. Build robots to crawl outward from these sites, following every link in turn, 3 links deep
 - Approx. 250,000 pages per community (3,000,000 total)
 - Should capture 75%+ of search behavior



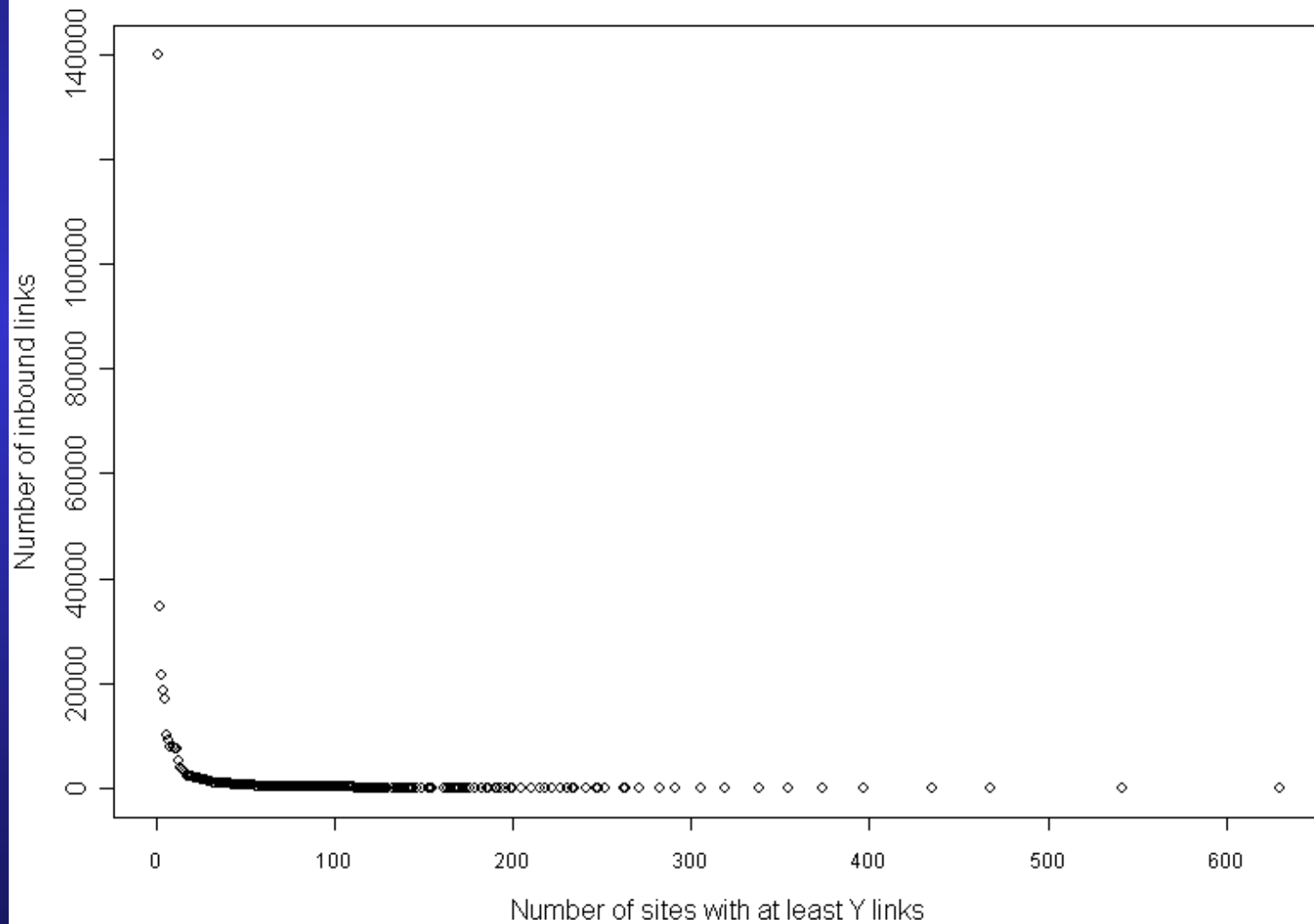
Methodology (cont.)

3. Analyze pages with automated methods, to see whether new pages are relevant to the given category
 - Support Vector Machine (SVM) classifiers:
 - Automatically learn to differentiate between two types of pages based on the words and word-pairs they contain
4. Look at the distribution of inlinks within these positive pages

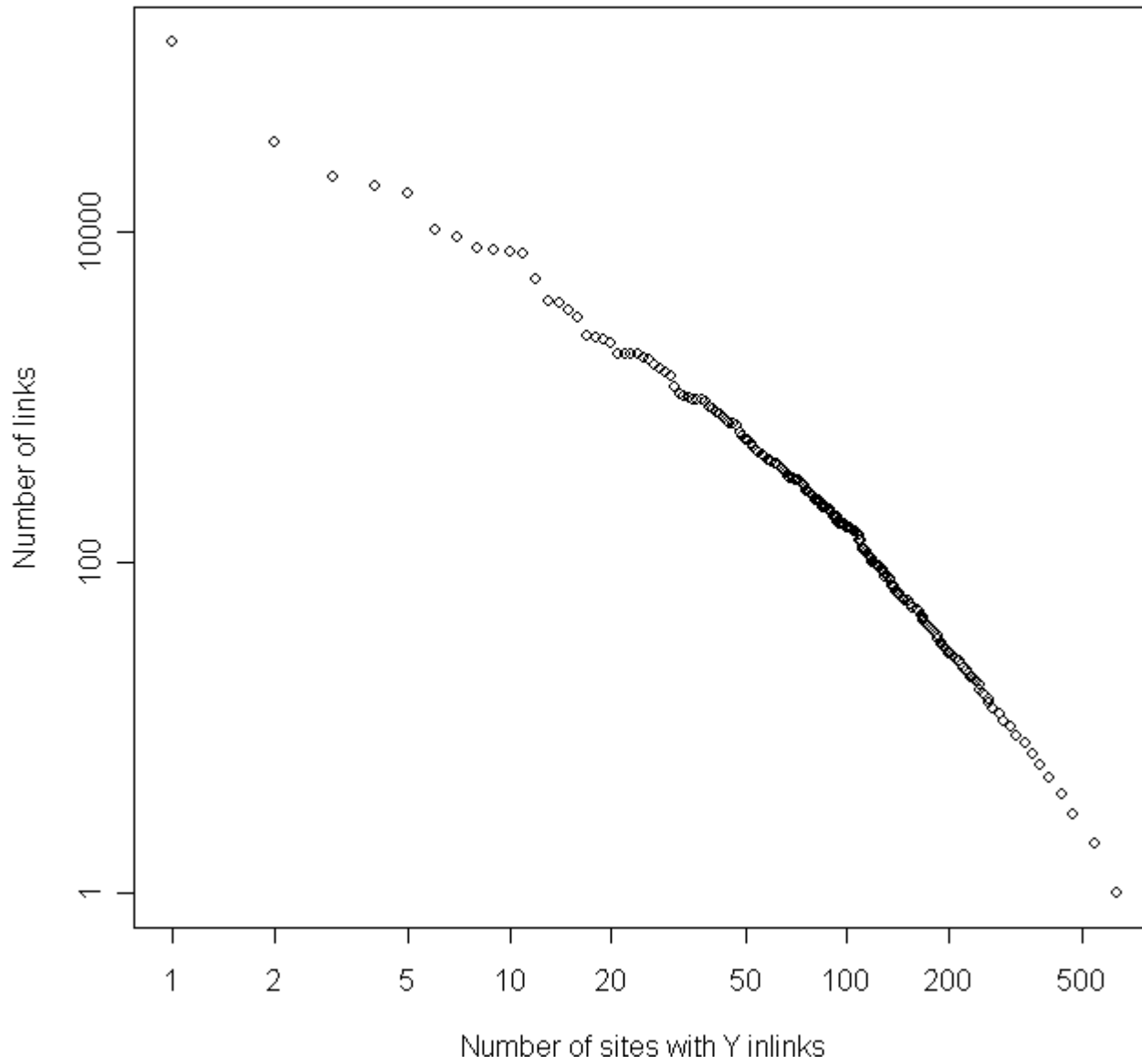
Link Concentration

	Sites	% links, Top Site	% links to Top 10 sites	% links to top 50
Abortion (Yahoo)	706	15.4	43.2	79.5
Abortion (Google)	1,015	31.1	70.6	88.8
Death Penalty (Yahoo)	725	13.9	63.5	94.1
Death Penalty (Google)	781	15.9	53.5	88.5
Gun Control (Yahoo)	1,059	28.7	66.7	88.1
Gun Control (Google)	630	39.2	76.8	95.9
President (Yahoo)	1,163	53.0	83.2	94.9
President (Google)	1,070	21.9	65.3	90.9
U.S. Congress (Yahoo)	528	25.9	74.3	94.8
U.S. Congress (Google)	1,350	22.0	51.4	82.3
General Politics (Yahoo)	1,027	6.5	36.4	70.3
General Politics (Google)	3,243	13.0	44.0	74.0

Gun Control Sites



Gun Control Sites (Google seed set)



Power-Law Fit

	Slope ($-\alpha$)	Y-Intercept (logged)	R^2
Abortion (Yahoo)	-1.54	11.8	.90
Abortion (Google)	-1.48	11.8	.97
Death Penalty (Yahoo)	-1.68	12.0	.97
Death Penalty (Google)	-1.95	13.9	.95
Gun Control (Yahoo)	-1.45	11.6	.96
Gun Control (Google)	-1.80	13.1	.97
President (Yahoo)	-1.65	13.0	.99
President (Google)	-1.70	13.2	.97
U.S. Congress (Yahoo)	-1.90	13.2	.97
U.S. Congress (Google)	-1.53	12.9	.95
General Politics (Yahoo)	-1.25	10.5	.96
General Politics (Google)	-1.45	13.5	.98

Web v. Traditional Media

Look at national market share across radio stations, print media, and Web sites.

Data:

- Radio: Average listenership for all 1280 commercial stations in top 50 US markets (*Arbitron*)
- Print: Circulation figures for all US newspapers and all US magazines (*Audit Bureau of Circulation*)
- Web: Additional data on aggregate link distribution, top news sites, Weblogs
- TV: Primetime ratings of 171 shows (*AC Nielsen*)

Choosing Metrics

- *The Gini Coefficient*
 - Commonly used to measure income equality
 - How does the average player compare with the top dogs
- *Herfindahl-Hirschman Index (HHI)*
 - Highlights power of the largest outlets
 - In this case stations or sites, not parent companies

Web v. Traditional Media (II)

<i>Media Type</i>	<i>Gini</i>	<i>Gini, top 20</i>	<i>HHI</i>
Television—Primetime Ratings	.35	.09	93
Radio—Stations in top 50 markets	.53	.12	19
Print—All US Newspapers	.69	.25	73
Print—All US Magazines	.70	.37	123
WWW—All sites, links	.96	.45	323
WWW—All news sites, traffic	n/a	.31	n/a
WWW—All Weblogs, traffic	.89	.42	286
WWW—Abortion sites, links	.94	.55	1754
WWW—Gun control sites, links	.96	.57	1705
WWW—Presidency sites, links	.98	.63	3207

Economics of Information

“What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it”

—Herbert Simon, 1971

- Users “satisfice” with the first site that is good enough
 - Repeated by thousands of users for year after year, creates these power law structures
 - Preferential attachment—“Rich Get Richer”

New Media, New Limits

- Challenges to diversity on the Web haven't disappeared, they're just different:
 - Most costs of creating content unchanged
 - Distribution is the only major cost reduction
 - Cognitive limitations
 - Vastness of cyberspace
 - Economics: more choices often mean poorer decisions
 - Architecture of the medium
 - More links = better rankings, more traffic
 - Can't eliminate links without unweaving the Web

Geographic barriers which *created* diversity are erased

Zen and the Measurement of Online Diversity

- If someone puts a Web site online...
 - ...and no one visits it or can even find it...
 - *Does it count towards online diversity?*
- Assumption has often been that making lots of information sources easy to retrieve would increase their use
 - The *Field of Dreams* model: “If you build it, citizens will come”
- Most Web content continues the broadcast model

Thank you

Matthew Hindman

Doctoral Fellow, NCDG, Harvard
mhindman@princeton.edu

Power Laws and Public Policy

Q: What can be done about these power laws?

A1: Not much...

- Emerge spontaneously almost everywhere
- Only exceptions:
 - Niches with both horizontal and vertical knowledge (*Universities, public companies*)
 - Markets which have to be local (*wedding photographers*)
- Difficult to imagine effective regulation

A2: ...Nothing *should* be done

- Most users are shockingly unsophisticated
- Power laws make the Web navigable

Power Laws and Public Policy (II)

Implications for regulating other media:

- Counting the number of sites available online gives the wrong answer on diversity
- The right answer requires more and different data:
 - Link data: Good, important, but only a start
 - Audience data: Nielsen//Netratings sample
 - Search engine data: What sites are people looking for and what are they finding?
 - Experimental data: What resources can the average Web user reach if they have to?

A Tale of Two Mediascapes

Two hypothetical newspaper readership nationwide:

Condition A:

- No one gets to chose their newspaper
- Equal readership across hundreds of publications

Condition B:

- Everyone can chose from any of 1000 newspapers
- Top 5 newspapers have 80% of the market share

Which condition is more diverse?

