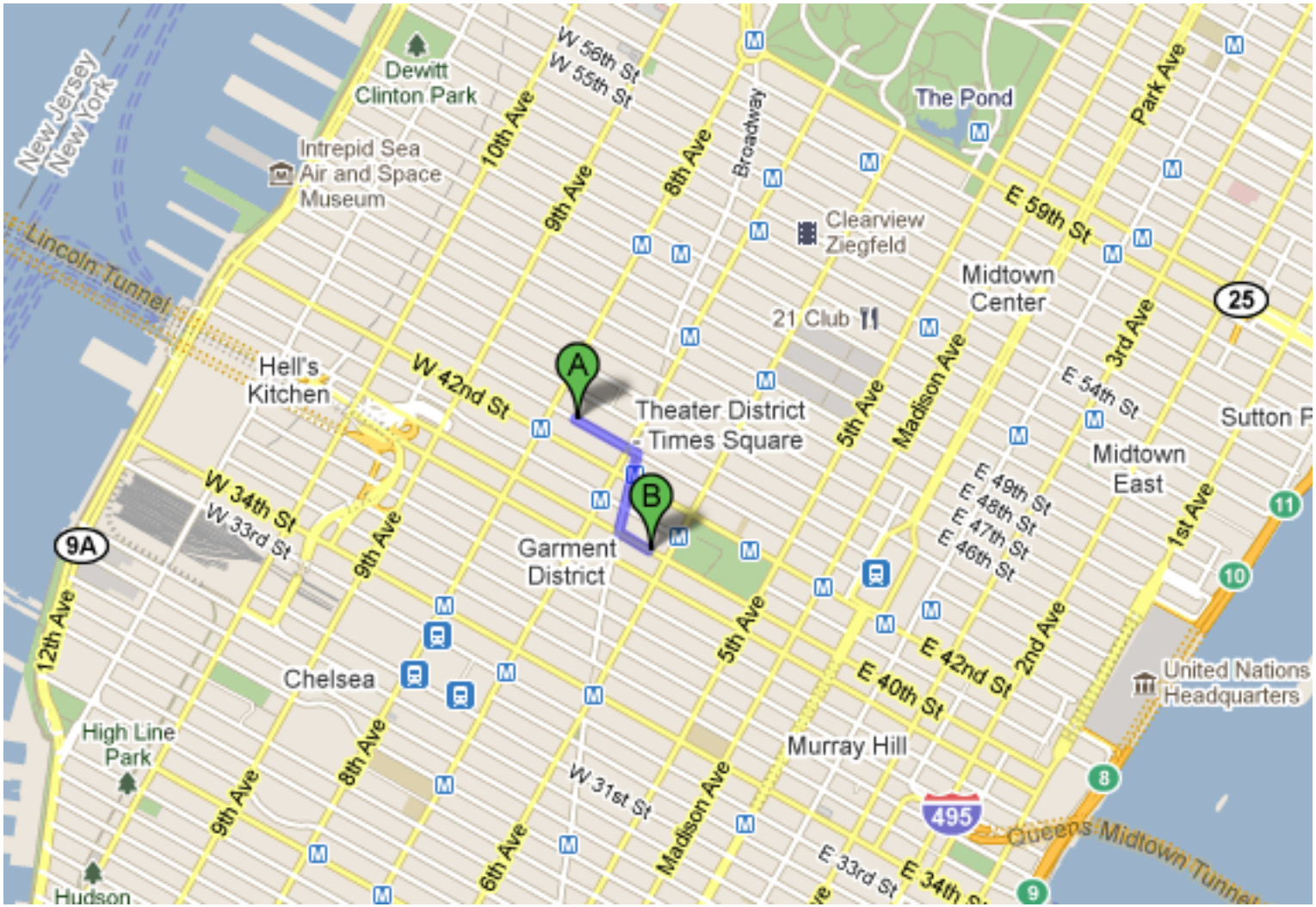
Learning from Web Activity

Jake Hofman

Joint work with Irmak Sirer & Sharad Goel
Yahoo! Research

June 27, 2011


New Jersey
New York
Lincoln Tunnel
Dewitt
Clinton Park
Intrepid Sea
Air and Space
Museum
W 56th St
W 55th St
10th Ave
9th Ave
8th Ave
Broadway
The Pond
Park Ave
E 59th St
Clearview
Ziegfeld
Midtown
Center
21 Club
25
3rd Ave
E 54th St
Hell's
Kitchen
W 42nd St
A
Theater District
Times Square
5th Ave
Madison Ave
Sutton P
Midtown
East
E 49th St
E 48th St
E 47th St
E 46th St
1st Ave
11
10
W 34th St
W 33rd St
9A
9th Ave
B
Garment
District
12th Ave
Chelsea
5th Ave
E 42nd St
2nd Ave
E 40th St
United Nations
Headquarters
High Line
Park
Murray Hill
8th Ave
W 31st St
Madison Ave
495
Queens Midtown Tunnel
8
9th Ave
6th Ave
E 33rd St
E 34th St
9
Hudson

"... applies **social sciences** *techniques, including algorithmic game theory, economics,* **network analysis**, *psychology, ethnography, and mechanism design, to* **online situations**."

`http://research.yahoo.com`

SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

"... a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale ..."

http://sciencemag.org/content/323/5915/721

SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

"... shares with other nascent interdisciplinary fields (e.g., sustainability science) the need to develop a paradigm for training new scholars ..."

http://sciencemag.org/content/323/5915/721

The ~~clean~~ real story

*"**We have a habit** in writing articles published in scientific journals to make the work as finished as possible, **to cover all the tracks**, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on. So there **isn't any place to publish**, in a dignified manner, **what you actually did** in order to get to do the work ..."*

-Richard Feynman
Nobel Lecture[1], 1965

[1] `http://bit.ly/feynmannobel`

Outline

- The clean story
- The real story
- Lessons learned

The clean story

(covering our tracks)

Motivation

Science 17 April 1998:
Vol. 280 no. 5362 pp. 390-391
DOI: 10.1126/science.280.5362.390

< Prev | Table of Contents | Next >

POLICY

INFORMATION ACCESS

Bridging the Racial Divide on the Internet

Donna L. Hoffman and Thomas P. Novak

+ Author Affiliations

The Internet is expected to do no less than transform society (1); its use has been increasing exponentially since 1994 (2). But are all members of our society equally likely to have access to the Internet and thus participate in the rewards of this transformation? Here we present findings both obvious and surprising from a recent survey of Internet access and discuss their implications for social science research and public policy.

Previous work is largely survey-based and focuses and group-level differences in online access

Motivation

"As of January 1997, we estimate that 5.2 million African Americans and 40.8 million whites have ever used the Web, and that 1.4 million African Americans and 20.3 million whites used the Web in the past week."

-Hoffman & Novak (1998)

Motivation

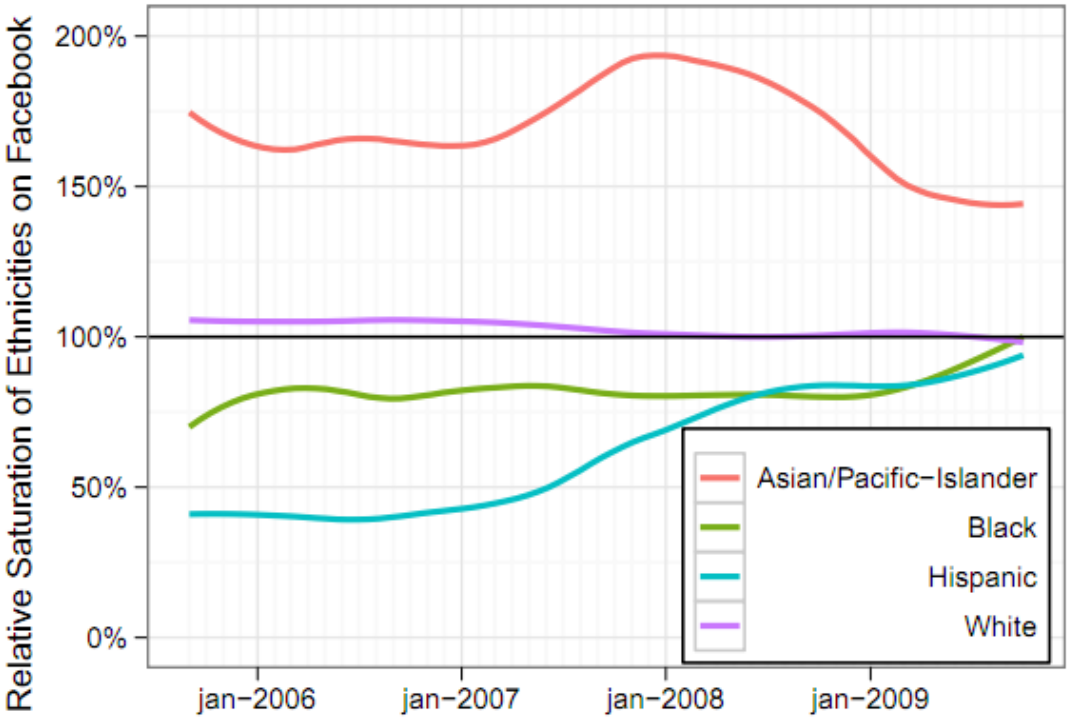


Figure 6: Relative saturation of ethnicities on Facebook. As the lines converge towards 100% (center), the makeup of U.S. Facebook converges towards that of the addressable Internet population.

Chang, et. al., *ICWSM* (2010)

Motivation

Focus on activity instead of access

How diverse is the Web?

To what extent do online experiences vary across demographic groups?

Data

nielsen MegaPanel

- Representative sample of 265,000 individuals in the US, paid via the Nielsen MegaPanel[2]
- Log of anonymized, complete browsing activity from June 2009 through May 2010 (URLs viewed, timestamps, etc.)
- Detailed individual and household demographic information (age, education, income, race, sex, etc.)

[2] Special thanks to Mainak Mazumdar

Data

- Transform all demographic attributes to binary variables
  e.g., Age $\rightarrow$ Over/Under 25, Race $\rightarrow$ White/Non-White, Sex $\rightarrow$ Female/Male

Data

- Transform all demographic attributes to binary variables
  e.g., Age → Over/Under 25, Race → White/Non-White, Sex → Female/Male
- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`, `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`

Data

- Transform all demographic attributes to binary variables
  e.g., Age → Over/Under 25, Race → White/Non-White, Sex → Female/Male
- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`, `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- Restrict to top 100k (out of 9M+ total) most popular sites (by unique visitors)

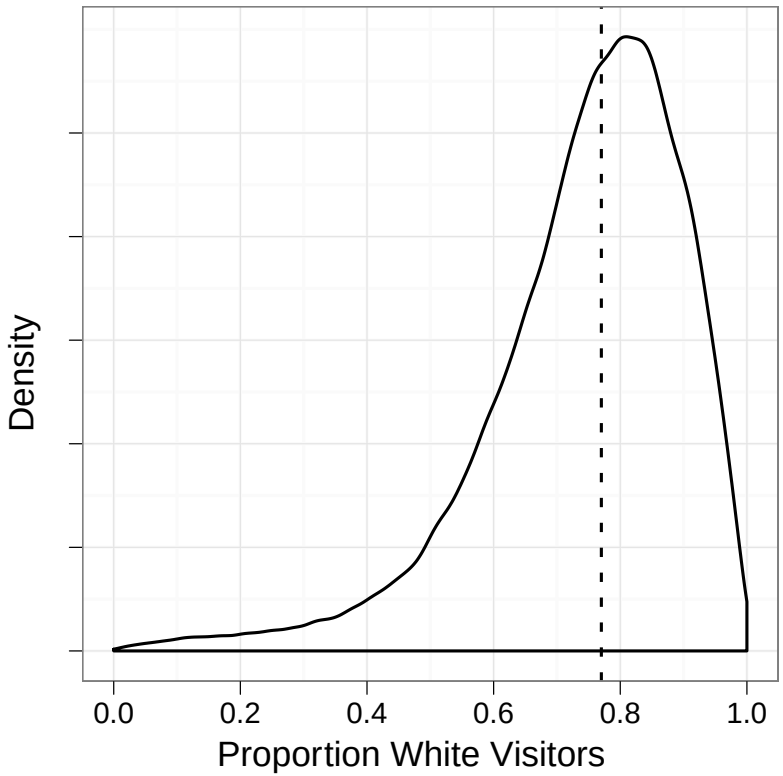
Data

- Transform all demographic attributes to binary variables
  e.g., Age → Over/Under 25, Race → White/Non-White, Sex → Female/Male
- Normalize pageviews to at most three domain levels, sans `www`
  e.g. `www.yahoo.com` → `yahoo.com`, `us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- Restrict to top 100k (out of 9M+ total) most popular sites (by unique visitors)
- Aggregate activity at the site, group, and user levels

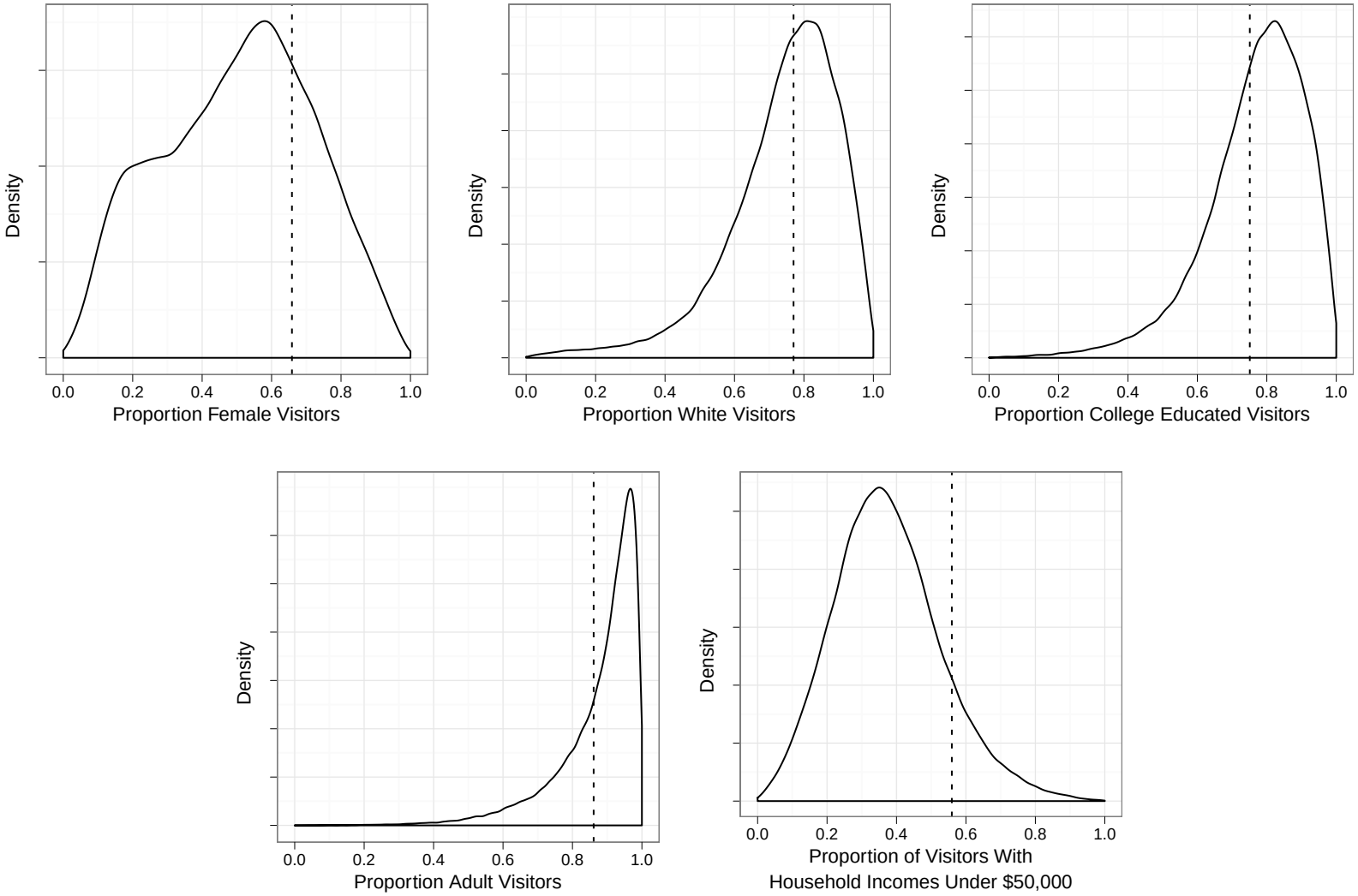
Site-level skew

How diverse are site audiences?

- For each site and attribute, calculate the skew in visitors (e.g., 93% of pageviews on `foxnews.com` are by White users)
- For each attribute, plot the distribution of visitor skew across all sites



Site-level skew



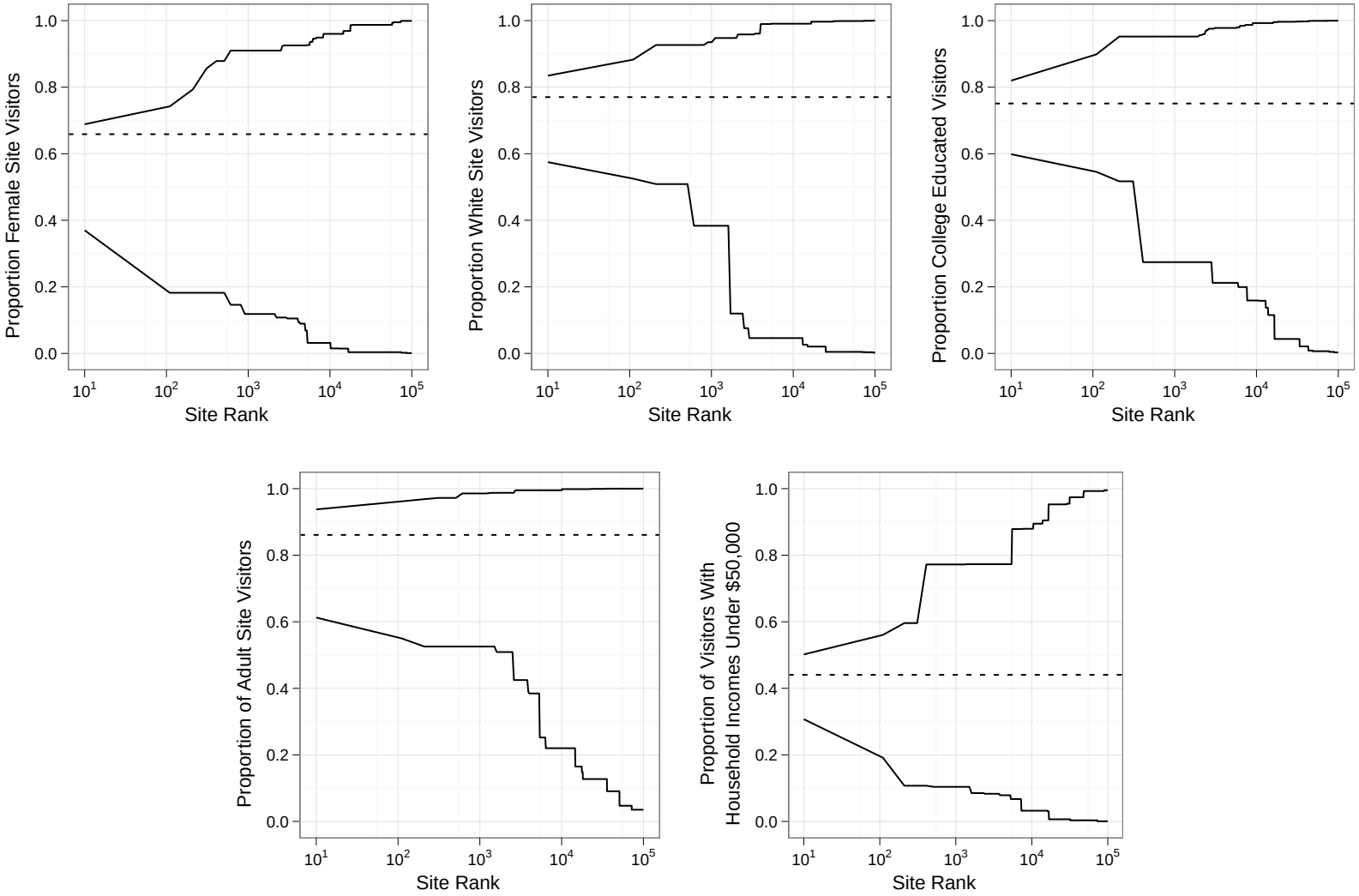
Site-level skew

Many sites have skew close the overall mean, but there also popular, highly-skewed sites

	Greater Than 90%	Less Than 10%
Female	youravon.com collectionsetc.com	coveritlive.com needlive.com
White	foxnews.com wunderground.com	blackplanet.com mediatakeout.com
College Educated	news.google.com nytimes.com	slumz.boxden.com sythe.com
Over 25 Years Old	mail.yahoo.com apps.facebook.com	nanowrimo.org cbox.ws
Household Income Under $50,000	scarleteen.com boards.adultswim.com	opentable.com marketwatch.com

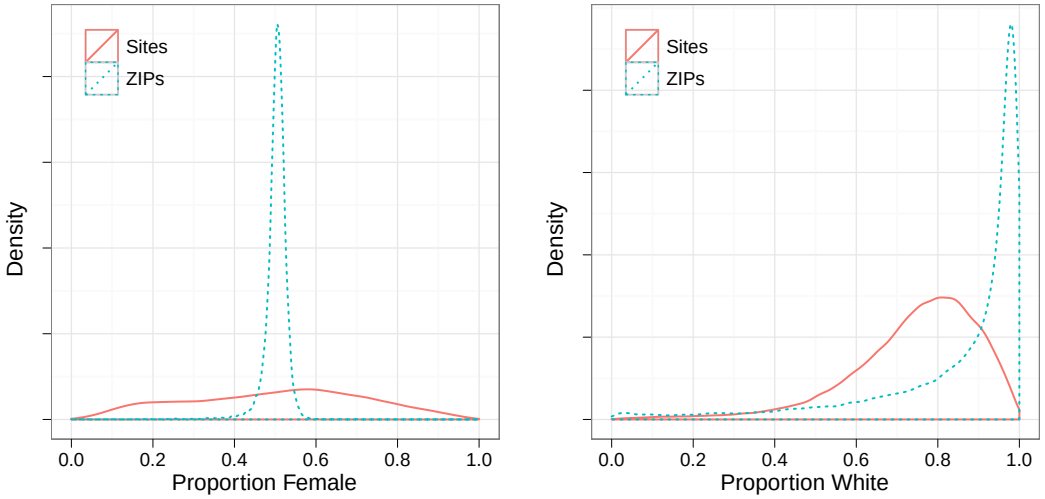
Site-level skew

This skew persists even when we restrict attention to the top 10k or 1k sites



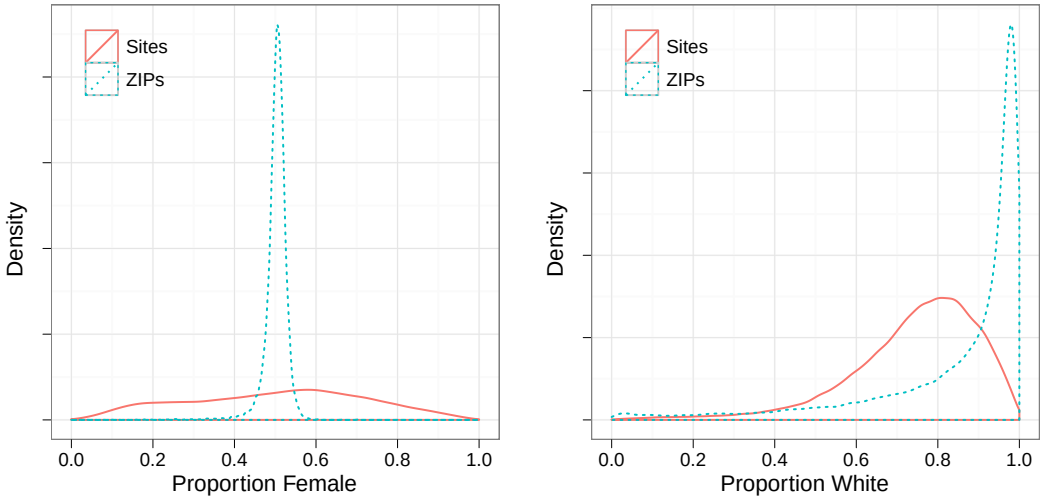
Sites vs. ZIPs

How do diversity of the online and offline worlds compare?



Sites vs. ZIPs

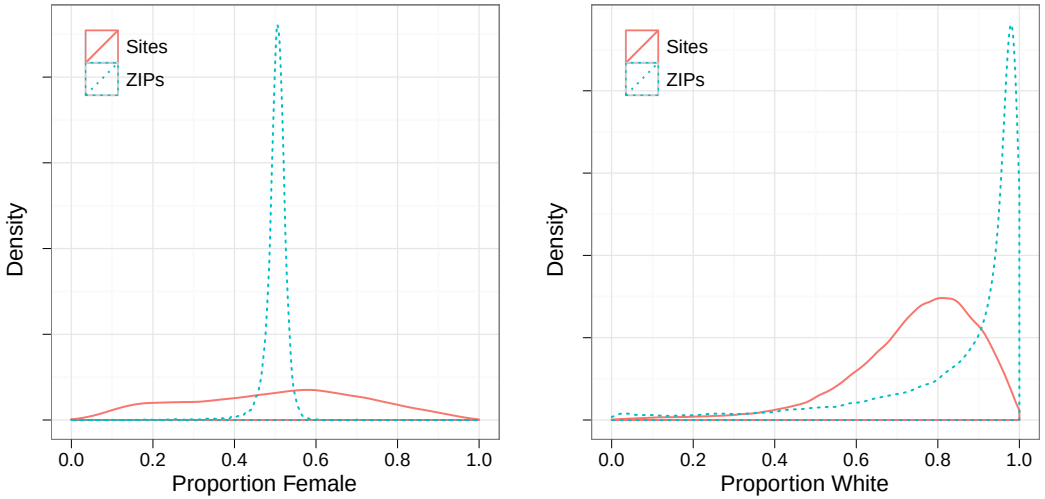
How do diversity of the online and offline worlds compare?



As expected, neighborhoods are more gender-balanced than sites

Sites vs. ZIPs

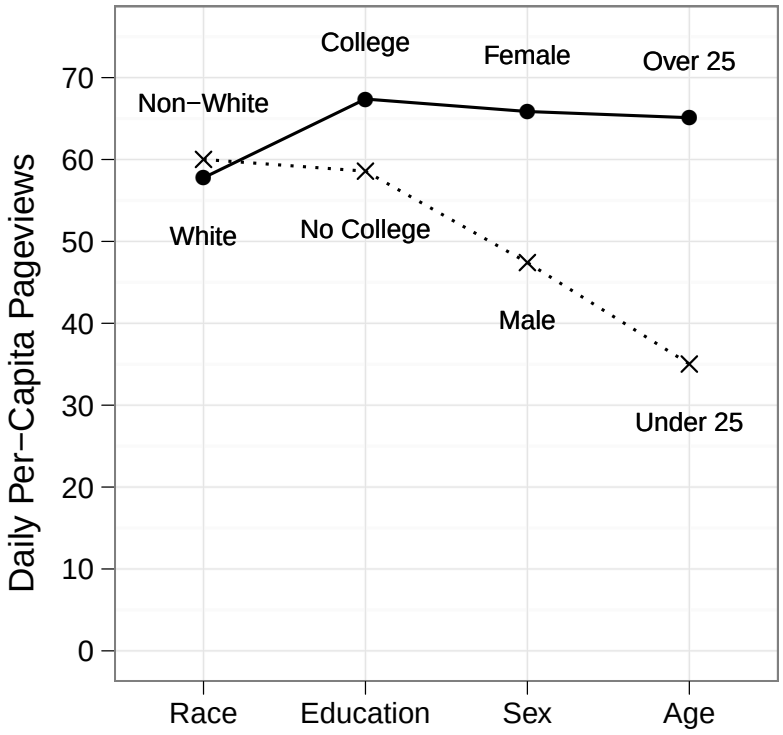
How do diversity of the online and offline worlds compare?



But sites typically have more racially diverse audiences than neighborhoods have residents

Group-level activity

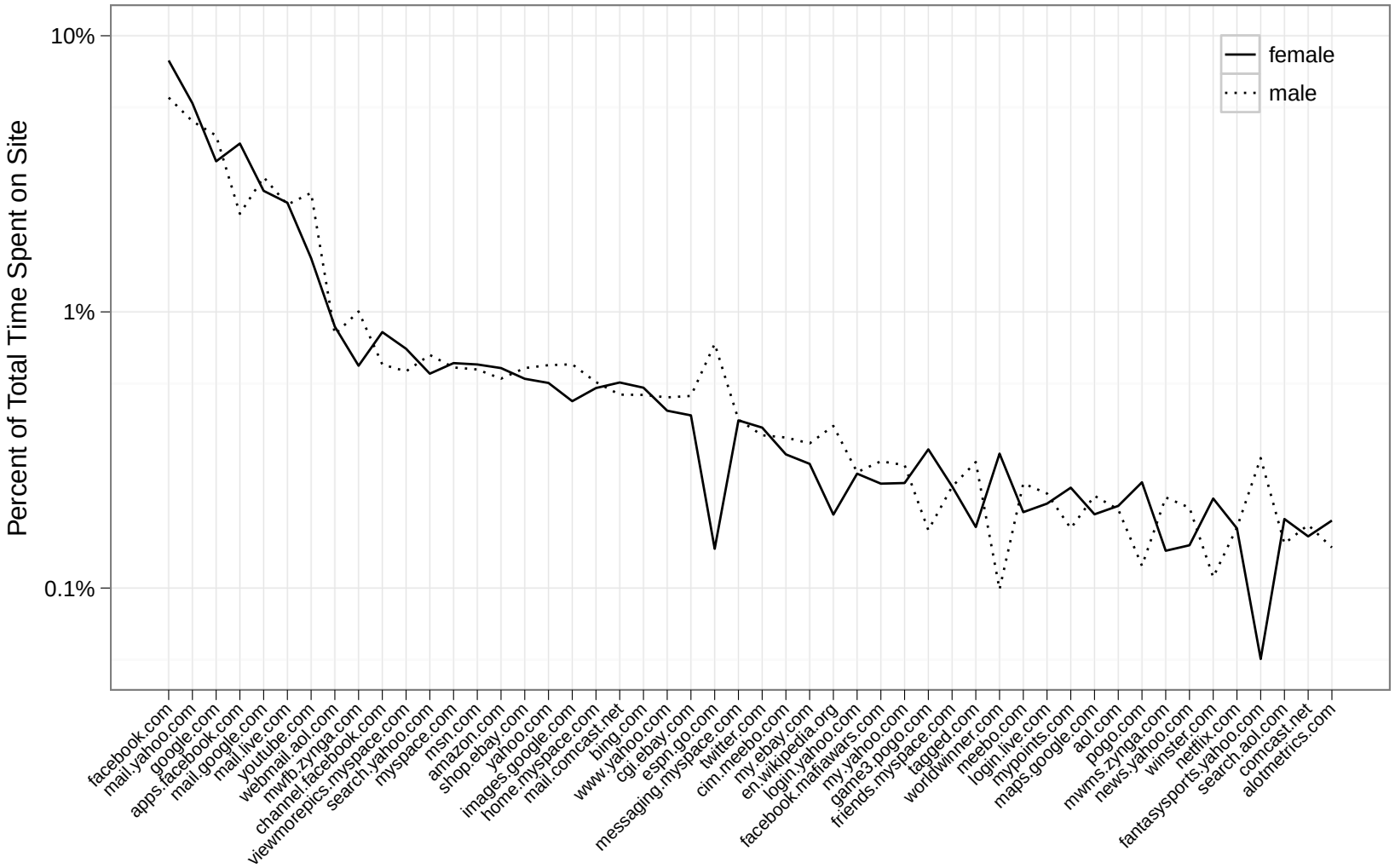
How does browsing activity vary at the group level?



Large differences exist even at the aggregate level
(e.g. women on average generate 40% more pageviews than men)

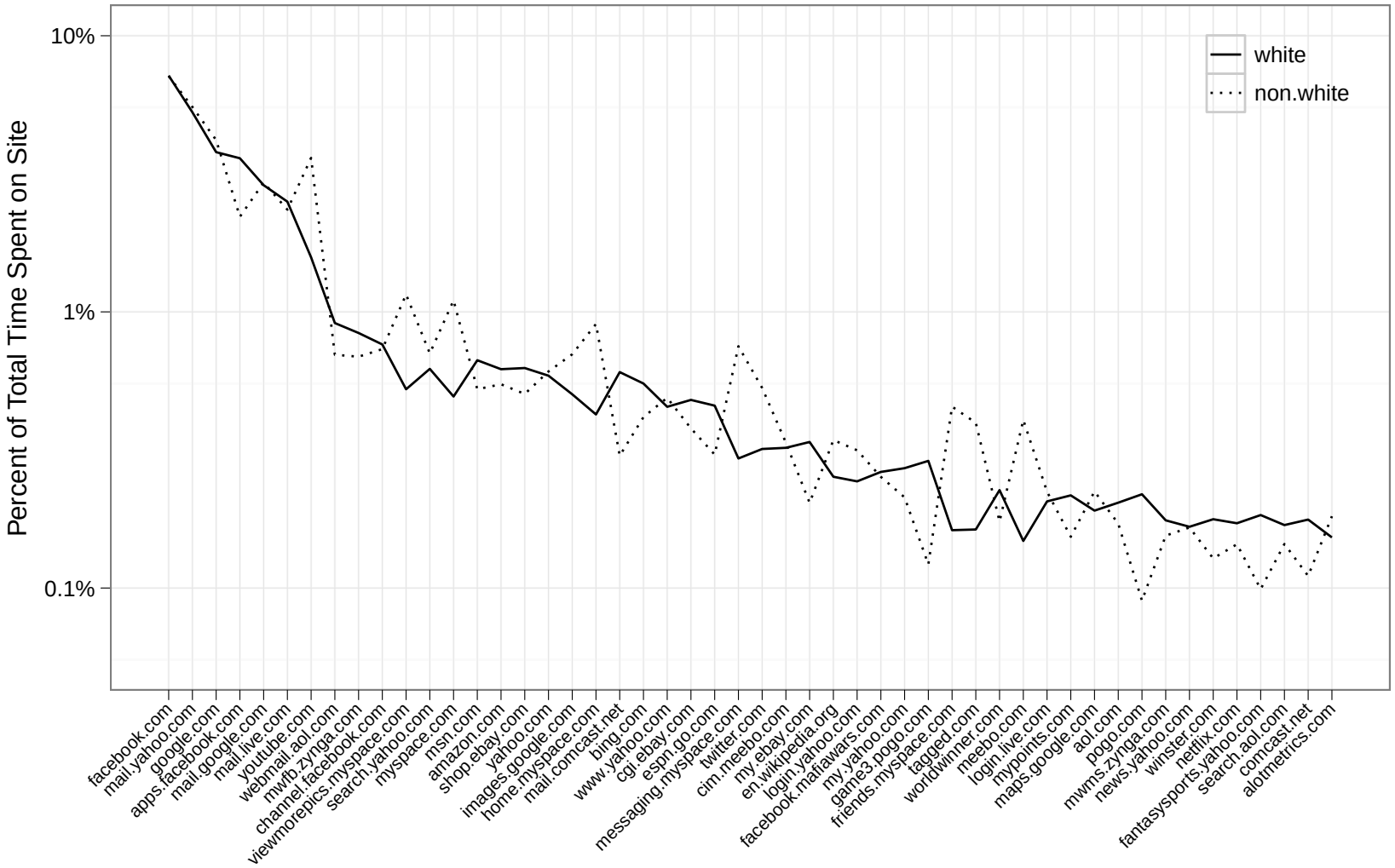
Group-level activity

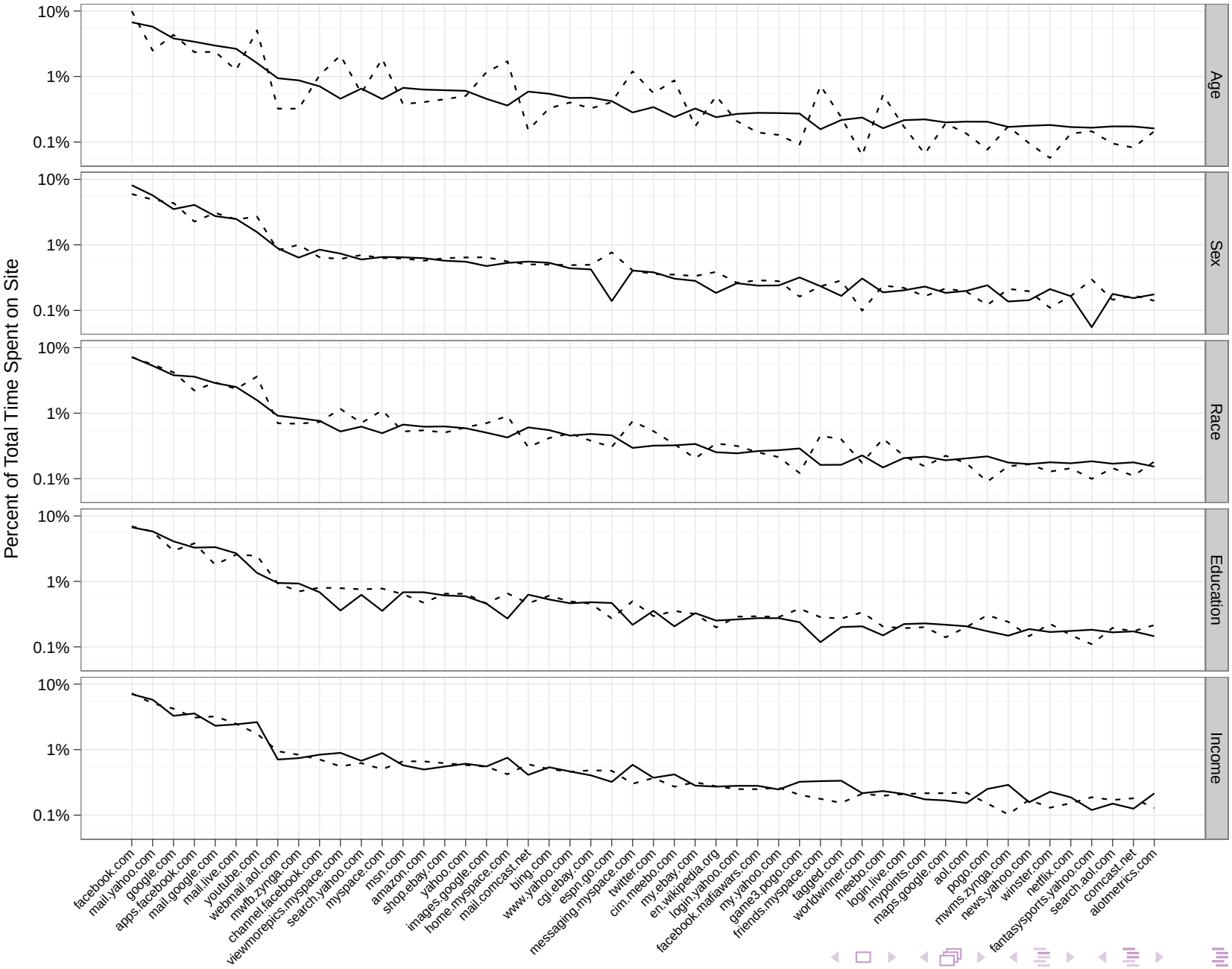
All groups spend more than a third of their time on a handful of email, search, and social networking sites



Group-level activity

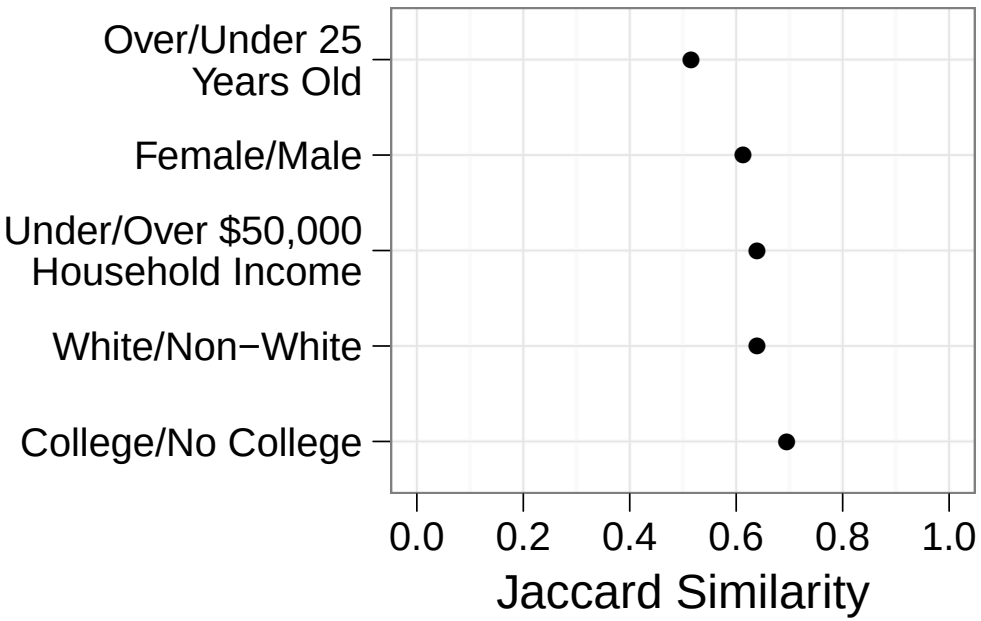
But different groups distribute their time differently, both on universally popular and on more niche sites




Percent of Total Time Spent on Site
10%
1%
0.1%
Age
Sex
Race
Education
Income
facebook.com
mail.yahoo.com
google.com
apps.facebook.com
mail.google.com
mail.live.com
youtube.com
webmail.aol.com
mwfb.zynga.com
channel.facebook.com
viewmorepics.myspace.com
search.yahoo.com
myspace.com
msn.com
amazon.com
shop.ebay.com
yahoo.com
images.google.com
home.myspace.com
mail.comcast.net
bing.com
www.yahoo.com
cgi.ebay.com
espn.go.com
messaging.myspace.com
twitter.com
cim.meebo.com
my.ebay.com
en.wikipedia.org
login.yahoo.com
facebook.mafiawars.com
my.yahoo.com
game3.pogo.com
friends.myspace.com
tagged.com
worldwinner.com
meebo.com
login.live.com
mypoints.com
maps.google.com
aol.com
pogo.com
mwms.zynga.com
news.yahoo.com
winster.com
netflix.com
fantasysports.yahoo.com
search.aol.com
comcast.net
alotmetrics.com

Group-level activity

There is both reasonable overlap and variation amongst the most popular sites within groups.



Individual-level prediction

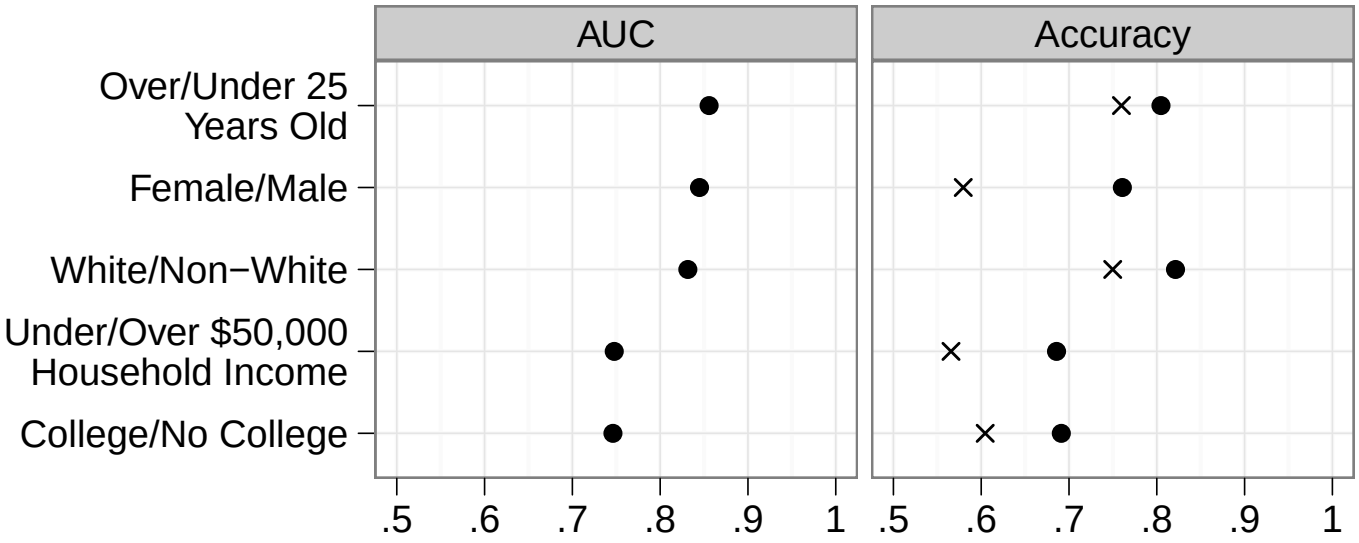
How well can one predict an individual's demographics from their browsing activity?

- Represent each user by the set of sites visited
- Fit linear models[3] to predict majority/minority for each attribute on 80% of users
- Tune model parameters using a 10% validation set
- Evaluate final performance on held-out 10% test set

[3] Using SVM-perf

Individual-level prediction

Reasonable (∼70-85%) accuracy and AUC across all attributes



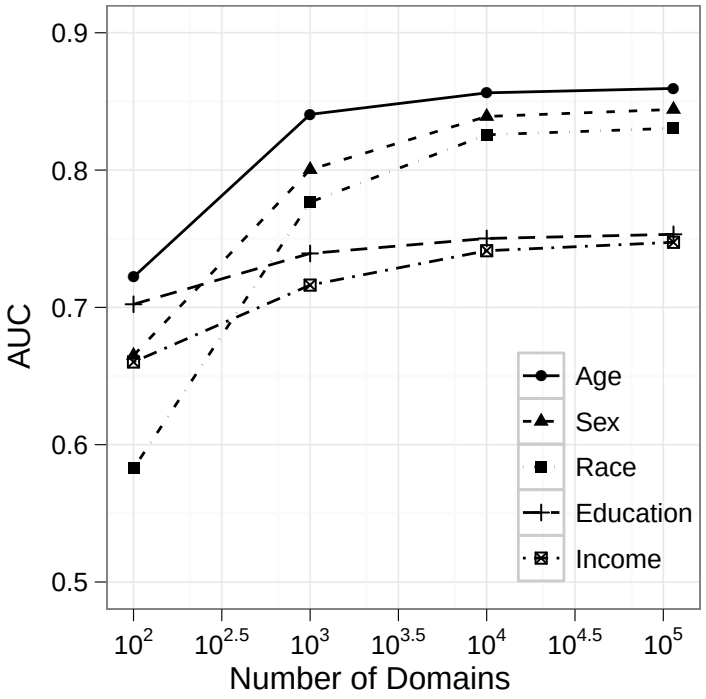
Individual-level prediction

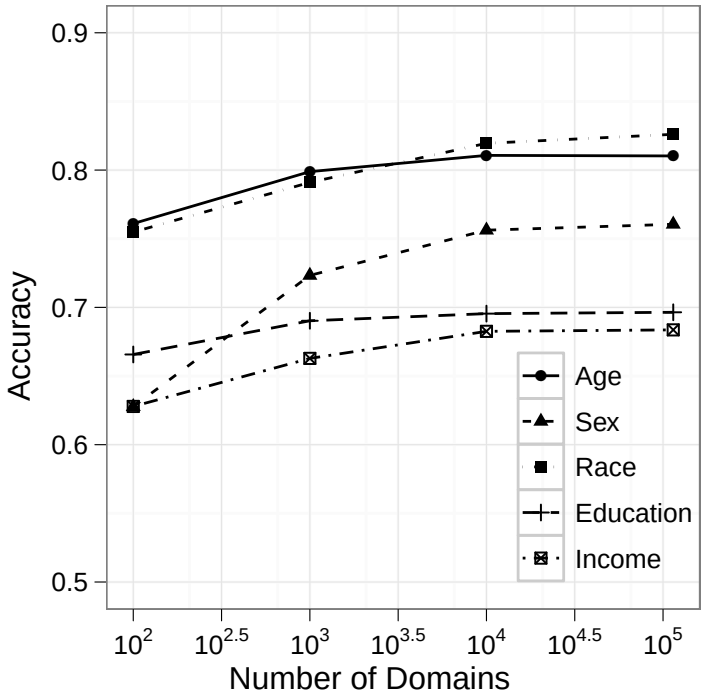
Highly-weighted sites under the fitted models

	Large positive weight	Large negative weight
Female	winster.com lancome-usa.com	sports.yahoo.com espn.go.com
White	marlboro.com cmt.com	mediatakeout.com bet.com
College Educated	news.yahoo.com linkedin.com	youtube.com myspace.com
Over 25 Years Old	evite.com classmates.com	addictinggames.com youtube.com
Household Income Under $50,000	eharmony.com tracfone.com	rownine.com matrixdirect.com

Individual-level prediction

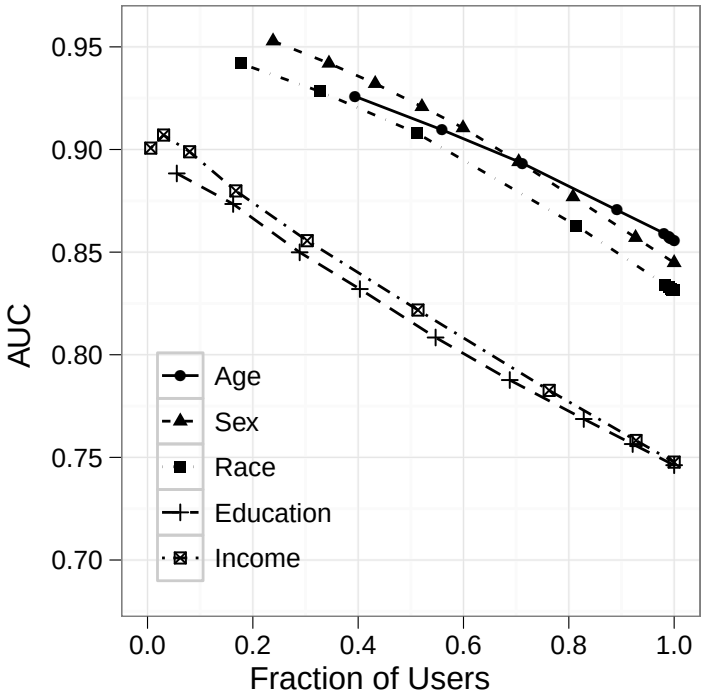
Similar performance even when restricted to top 1k sites

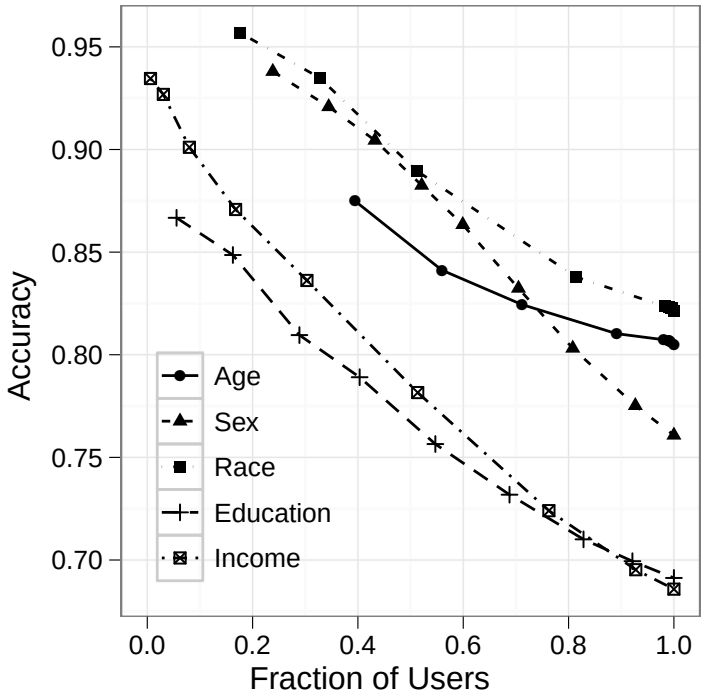




Individual-level prediction

Substantially better performance when restricted to "stereotypical" users (∼80-90%)

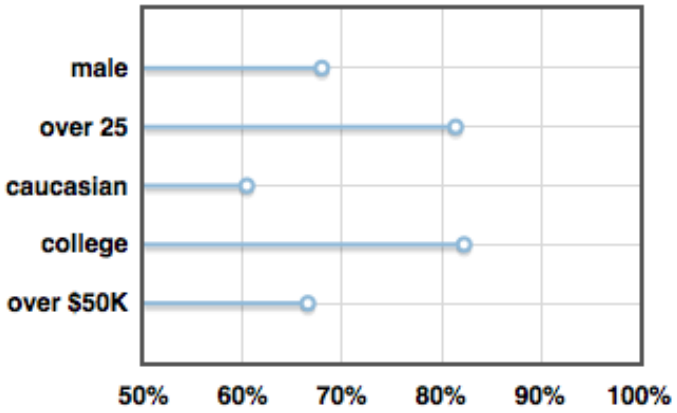




Individual-level prediction

Proof of concept browser demo

From the 28 sites we found in your browser history, it appears that you're a **caucasian male** who is **over 25** years old with a **college** education earning **over $50K** per year.



http://bit.ly/surfpreds

The real story

(what we actually did)

The real story

0. Got several hundred GBs of MegaPanel data from Nielsen[4], looked at a small sample

```
# ls -alh nielsen_megapanel.tar
-rw-r--r--  100G Jul 17 13:00 nielsen_megapanel.tar
```

[4] Special thanks to Mainak Mazumdar

The real story

1. Discussed (many) possible projects

- Infer the number of individuals using the same browser or behind the same ip?
- Determine number of actual uniques advertisers are receiving?
- Predict user demographics from a few minutes of browsing activity for ad-targeting?

The real story

2. Modeled real-valued age as a function of site visits

- Worked on this for an embarassingly long time
- Tried various options for cleaning and normalizing data (100GB $\rightarrow$ $\sim$5GB)
- Investigated several methods for feature selection (e.g., naive Bayes, mutual information, *popularity*)

Hadoop + Pig (+ awk)

```
-- define streaming command for normalizing urls
DEFINE add_star_domains `awk 'n=split($2,a,"."); (n==3) {print $1"\t*."a[2]"."a[3]"\t"$3}'`;

-- flatten user histories
user_pageviews = FOREACH users GENERATE
        uid,
        FLATTEN(history) AS (did, pageviews, weight);

-- join user pageviews against top domains
user_pageviews = JOIN user_pageviews BY did, top_domains BY did USING 'replicated';
user_pageviews = FOREACH user_pageviews GENERATE
        user_pageviews::uid AS uid,
        top_domains::domain AS domain,
        user_pageviews::pageviews AS pageviews;

-- stream through awk to extract two-level domains (e.g. *.yahoo.com)
user_pageviews = STREAM user_pageviews THROUGH add_star_domains AS (uid:long, domain:chararray, pageviews:int);

-- regroup and count pageviews by normalized domains
-- (userid, normalized_domain, num_pageviews)
user_pageviews = GROUP user_pageviews BY (uid, domain) PARALLEL 10;
user_pageviews = FOREACH user_pageviews GENERATE
        group.uid AS uid,
        group.domain AS domain,
        SUM(user_pageviews.pageviews) AS pageviews;
```

The real story

3. Settled for classification of binary outcomes (e.g. adult/non-adult)

- 265,000 users (examples) and 100,000 sites (features)
- Logistic regression in R, e.g.

```
model <- glm(is.adult ~ ., data=nielsen, family=binomial)
```

???

The real story

3. Settled for classification of binary outcomes (e.g. adult/non-adult)

$$\hat{y}(x_i) = w \cdot x_i + b$$

$$\text{Support Vector Machine}: \quad L(y, \hat{y}) = C \sum_i [1 - y_i \hat{y}(x_i)]_+ + ||w||^2$$

The real story

4. Investigated *why* classification worked reasonably well

- Generated descriptive statistics across all attributes at the site and group levels
- Compared site statistics to ZIP code data from the US Census
- Compared time distribution across groups

The real story

5. Realized that we now had one of the most comprehensive studies available on demographic diversity of the web

Conclusion

(lessons learned)

Lessons learned

Data jeopardy

Regardless of scale, it's difficult to find the right questions to ask of the data

Lessons learned

Rapid iteration

The ability to iterate quickly, asking and answering many questions, is crucial

Lessons learned

Data cleaning

Cleaning and normalizing data is a substantial amount of the work

Lessons learned

Modeling

Simple methods (e.g., linear models) work surprisingly well, especially with lots of data

Lessons learned

It's easy to cover your tracks—things are often much more complicated than they appear

Thanks. Questions?

http://messymatters.com/webdemo

http://jakehofman.com
hofman@yahoo-inc.com