Who Does What on the Web: A Large-Scale Study of Browsing Behavior

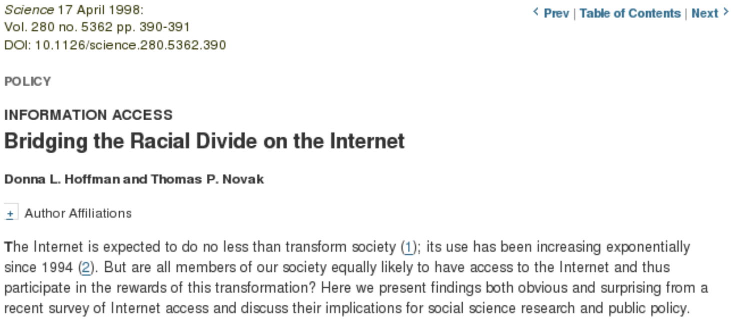
Sharad Goel [1] Jake M. Hofman [1] M. Irmak Sirer [2]

Microsoft Research

Northwestern University

June 5, 2012

Motivation



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

POLICY

INFORMATION ACCESS

Bridging the Racial Divide on the Internet

Donna L. Hoffman and Thomas P. Novak

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

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[1]
- 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.)

[1] Special thanks to Mainak Mazumdar

Data

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

Data

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

- 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

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

- 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)

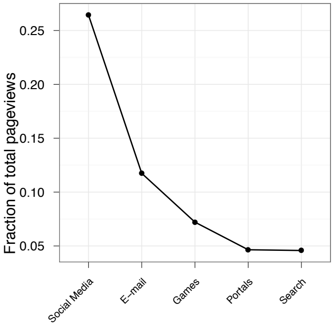
Data

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

- 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 page category, demographic group, and individual user levels

Aggregate usage patterns

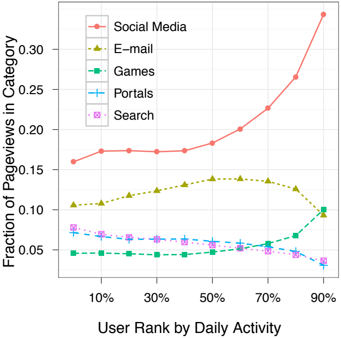
How do users distribute their time across different categories?



All groups spend the majority of their time in the top five most popular categories

Aggregate usage patterns

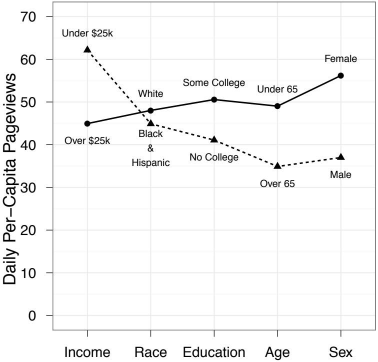
How do users distribute their time across different categories?



Highly active users devote nearly twice as much of their time to social media relative to typical individuals

Group-level activity

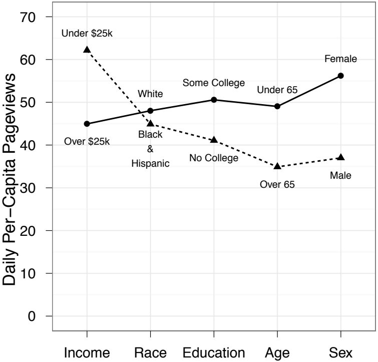
How does browsing activity vary at the group level?



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

Group-level activity

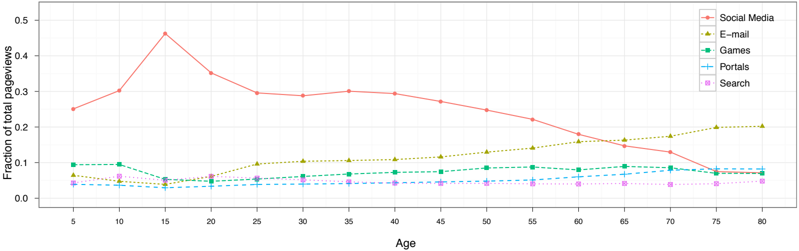
How does browsing activity vary at the group level?

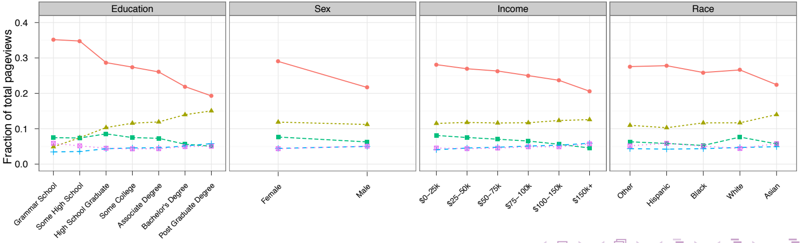


Younger and more educated individuals are both more likely to access the Web and more active once they do

Group-level activity

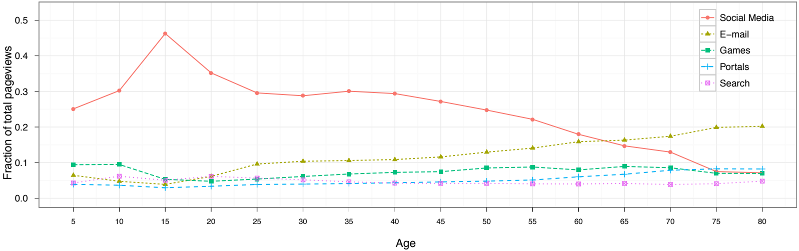
All demographic groups spend the majority of their time in the same categories

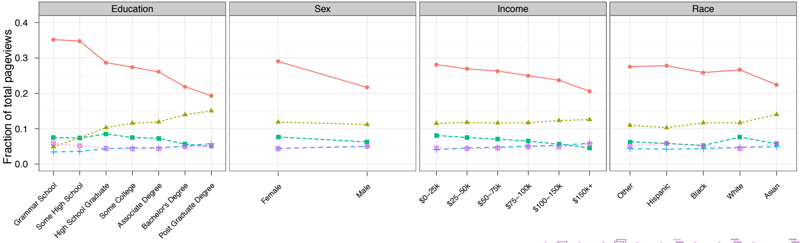




Group-level activity

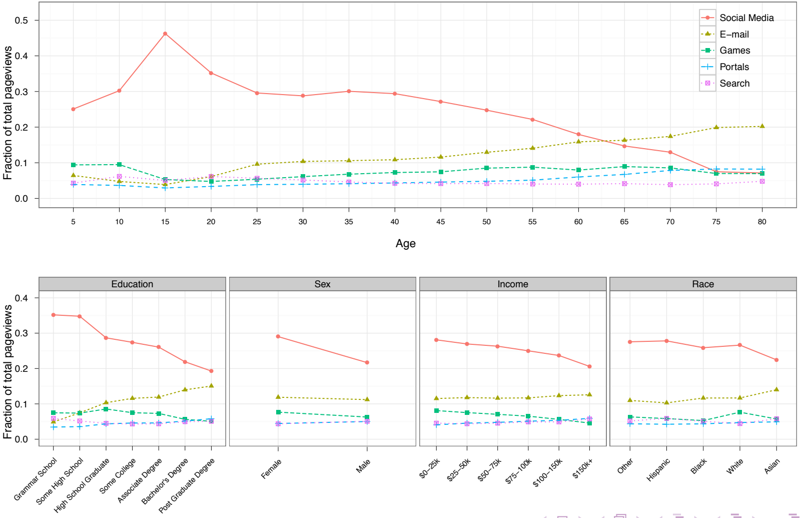
Older, more educated, male, wealthier, and Asian Internet users spend a smaller fraction of their time on social media





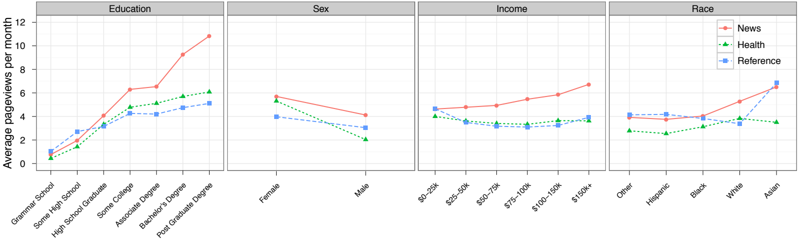
Group-level activity

Lower social media use by these groups is often accompanied by higher e-mail volume



Revisiting the digital divide

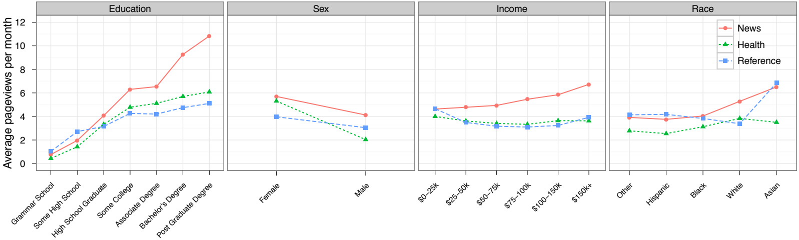
How does usage of news, health, and reference vary with demographics?



Post-graduates spend three times as much time on health sites than adults with only some high school education

Revisiting the digital divide

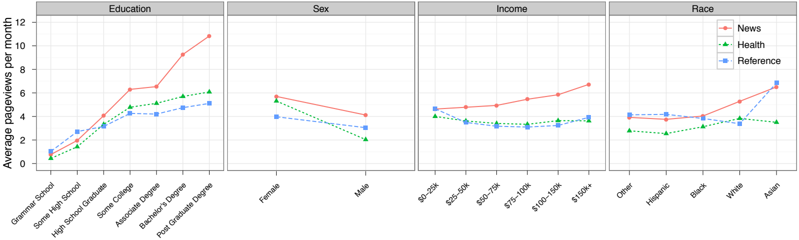
How does usage of news, health, and reference vary with demographics?



Asians spend more than 50% more time browsing online news than do other race groups

Revisiting the digital divide

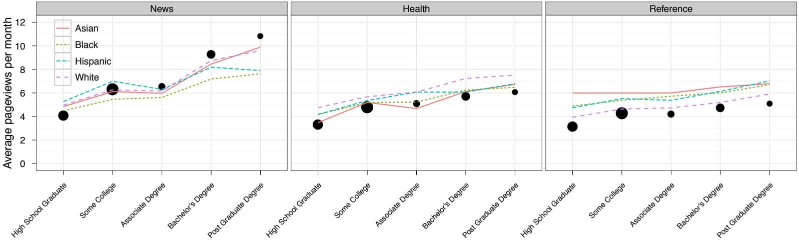
How does usage of news, health, and reference vary with demographics?



Even when less educated and less wealthy groups gain access to the Web, they utilize these resources relatively infrequently

Revisiting the digital divide

How does usage of news, health, and reference vary with demographics?

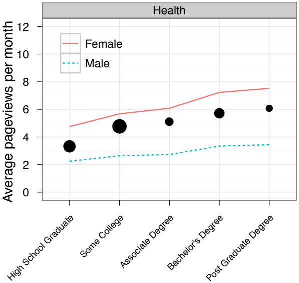


Controlling for other variables, effects of race largely disappear, while education continues to have large effect

$$p_i = \sum_j \alpha_j x_{ij} + \sum_j \sum_k \beta_{jk} x_{ij} x_{ik} + \sum_j \gamma_j x_{ij}^2 + \epsilon_i$$

Revisiting the digital divide

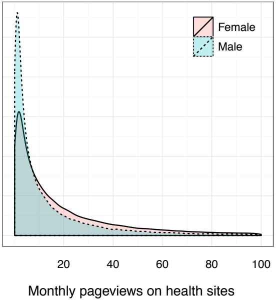
How does usage of news, health, and reference vary with demographics?



However, women spend considerably more time on health sites compared to men

Revisiting the digital divide

How does usage of news, health, and reference vary with demographics?



However, women spend considerably more time on health sites compared to men, although means can be misleading

Individual-level prediction

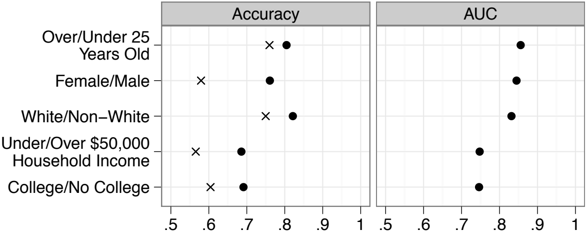
How well can one predict an individual's demographics from their browsing activity?

- Represent each user by the set of sites visited
- Fit linear models[2] 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

[2] http://bit.ly/svmperf

Individual-level prediction

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



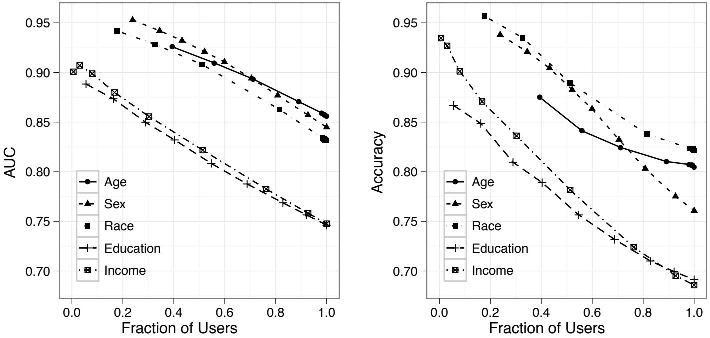
Individual-level prediction

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

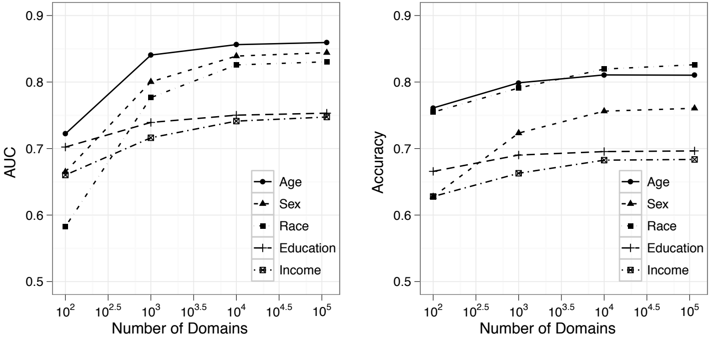
Individual-level prediction

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

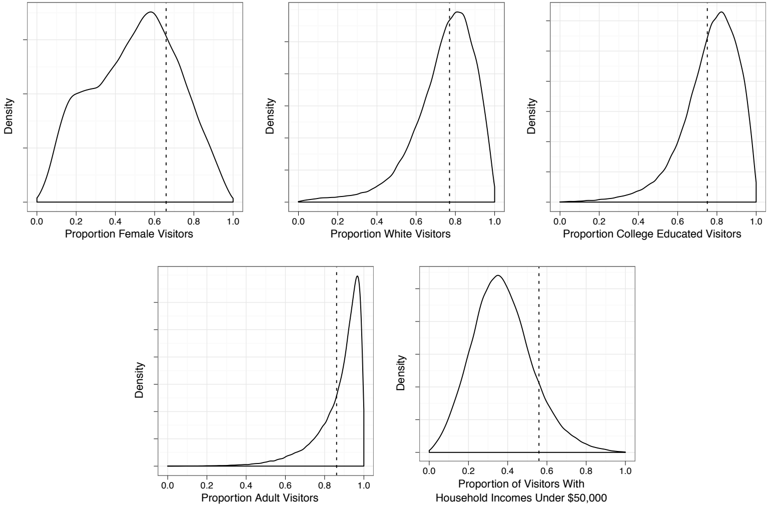


Individual-level prediction

Similar performance even when restricted to top 1k sites



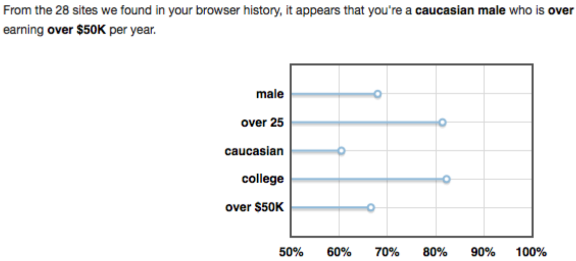
Site-level skew



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

Individual-level prediction

Proof of concept browser demo[3]



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`

[3] Requires Firefox 3.6 or older

Summary

- All demographic groups spend the majority of their time in the same categories
- Highly active users spend disproportionately more of their time on social media and less on e-mail relative to the overall population
- Access to research, news, and healthcare is strongly related to education, not as closely to ethnicity
- User demographics can be inferred from browsing activity with reasonable accuracy

Thanks. Questions?

sharadg@microsoft.com

jmh@microsoft.com

irmak@northwestern.edu