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

Sharad Goel
Yahoo! Research
111 West 40th Street
New York, NY 10018, US
goel@yahoo-inc.com

Jake M. Hofman
Yahoo! Research
111 West 40th Street
New York, NY 10018, US
hofman@yahoo-inc.com

M. Irmak Sirer
Northwestern University
2145 Sheridan Road
Evanston, IL 60208, US
irmak@northwestern.edu

Abstract

As the Web has become integrated into daily life, understanding how individuals spend their time online impacts domains ranging from public policy to marketing. It is difficult, however, to measure even simple aspects of browsing behavior via conventional methods—including surveys and site-level analytics—due to limitations of scale and scope. In part addressing these limitations, large-scale Web panel data are a relatively novel means for investigating patterns of Internet usage. In one of the largest studies of browsing behavior to date, we pair Web histories for 250,000 anonymized individuals with user-level demographics—including age, sex, race, education, and income—to investigate three topics. First, we examine how behavior changes as individuals spend more time online, showing that the heaviest users devote nearly twice as much of their time to social media relative to typical individuals. Second, we revisit the digital divide, finding that the frequency with which individuals turn to the Web for research, news, and healthcare is strongly related to educational background, but not as closely tied to gender and ethnicity. Finally, we demonstrate that browsing histories are a strong signal for inferring user attributes, including ethnicity and household income, a result that may be leveraged to improve ad targeting.

1 Introduction

Despite the rapid growth and importance of digital technologies over the past several decades, even relatively simple questions regarding browsing behavior remain unanswered. For example, how frequently do different demographic groups access online health resources? Do the heaviest Internet users—a relatively small number of whom account for the majority of all Web traffic—behave qualitatively differently than the rest of the population? Answers to such questions have implications ranging from assessing and responding to the digital divide to refining marketing and advertising strategies, yet addressing these issues has proven difficult via conventional methods.

To date, investigations of Web activity have typically relied on three types of data sources: surveys, toolbar tracking, and aggregated site-level analytics. Surveys are perhaps the most popular methodology for investigating Internet use, with the Pew Research Center regularly conducting such studies (Smith, 2010). While this approach is particularly well-suited for measuring attitudes and general usage trends, surveys typically involve small, non-representative samples of the population, and rely on users' incomplete—and sometimes inaccurate—statements of their own online behavior. In contrast, toolbar tracking data (e.g., as collected via user-installed toolbars distributed by Google, Microsoft, and Yahoo!) provide complete browsing histories for millions of users (Kumar and Tomkins, 2010). Though quite large in scale, these data often suffer from significant sample bias, and perhaps more importantly, it is difficult to disentangle an individual's browsing history from that of an entire household. Finally, site-specific analytics—as reported by Quantcast, for example—yield detailed data on specific site usage, but do not allow one to connect these statistics across sites or users. In short, though these approaches have certainly provided insight into how people use the Web, they are not tailored for accurately measuring detailed, individual-level behavior representative of the general population.

In contrast to such methods, our analysis of Web browsing behavior is based on complete activity logs for the approximately 250,000 users who participated in the Nielsen MegaPanel between June 2009 and May 2010. On shared computers, participants were required to log in to their own personal accounts, mitigating the intermingling of browsing sessions often present in toolbar data. Moreover, the panel is approximately representative of the general U.S. online population. Where applicable, pageviews were assigned to one of 85 high-level categories. In total, 9.2 million normalized domains were viewed by at least one panelist, and in aggregate, over three billion pageviews were recorded, with a median of 5,100 pageviews per user. In addition to these Web browsing histories, individual and household-level data were collected from each panelist, including age, sex, race, educational attainment, and household income. Leveraging these data, we are thus able to systematically investigate online browsing behavior at scale across a representative sample of the online population.

Though in this paper we focus on substantive scientific questions, we briefly note that the size of the dataset—billions of pageviews, comprising hundreds of gigabytes—presents several computational challenges. For example, even moderately complex aggregation tasks are impractical on single machines. As such, most of our work is conducted using the MapReduce parallel computation framework, which is particularly well-designed for such tasks.

To demonstrate the broad value of Web panel data for investigating browsing behavior, we explore three diverse topics. After reviewing related work in Section 2, we consider in Section 3 how Web usage changes as individuals spend more time online. We find that the heaviest 10% of users devote nearly twice as much of their time to social media relative to typical individuals, and spend a smaller fraction of their time on e-mail. In Section 4 we revisit the digital divide. Whereas this issue is generally framed in terms of those with and without access to the Internet, we instead focus on disparities in Web usage among those who are already online, finding that educational attainment is a particularly strong indicator of how often individuals turn to the Web for news, healthcare, and research. Finally, in Section 5, we evaluate the use of individual-level browsing histories to infer detailed demographic profiles, a task that is particularly relevant to ad networks that have limited information on user demographics but can often track browsing activity. Using an interpretable prediction framework, we show that browsing history is indeed a strong signal for inferring this otherwise difficult to obtain demographic information.

2 Related Work

Previous research on patterns of Internet use has generally focused on particular domains or services. For example, several survey-based studies have found significant race-dependent preferences on social networking sites, with White students preferring Facebook to MySpace (Hargittai, 2007; Watkins, 2009; Boyd, 2010). A recent and extensive analysis of activity on Facebook (Chang et al., 2010) revealed that the site's audience is converging to a distribution similar to that of the overall online population, although demographic groups tend to differ in their use of site features. Likewise, a demographic study of the Twitter population (Mislove et al., 2011) used self-reported name and location information to reveal that Twitter users are a highly biased sample of the U.S. population in terms of gender, race, and location. Further work highlights striking demographic differences in users' Web search intent and behavior (Weber and Castillo, 2010; Weber and Jaimes, 2011). For example, when searching for "wagner", U.S. women are most likely thinking of the 19th century German composer, while men are most likely interested in the paint sprayer company. Additionally, a detailed analysis of browsing sessions based on one week of Yahoo! toolbar data uncovered several topical and temporal patterns of user activity (Kumar and Tomkins, 2010). For example, sessions involving social networking sites tend to focus exclusively on this activity, whereas search-related sessions involve more diverse browsing behavior.

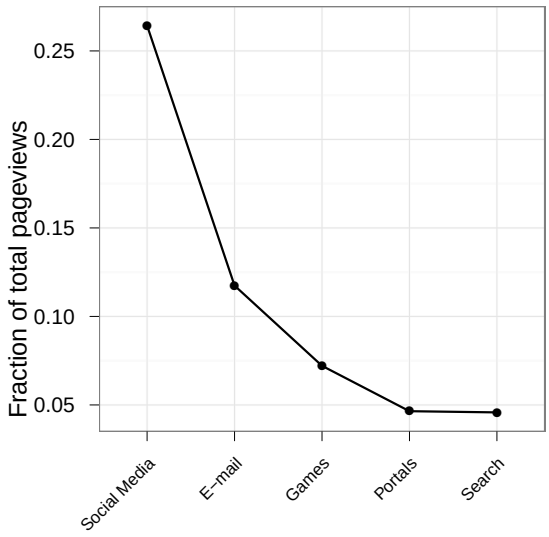

Figure 1: Percent of time spent on the top five most popular categories.

Finally, in addition to such descriptive work, several studies have focused on predicting demographic attributes from online activity. Recent work (Pennacchiotti and Popescu, 2011) inferred user demographics on Twitter, leveraging profile, content, and network information to predict political affiliation and race with reasonable accuracy. In analyzing the vulnerability of Web search query logs to privacy attacks, Jones et al. (2007) found that search queries are highly predictive of sex, age, location, and even individual identity. Similarly, using one week of activity for 189,000 users, Hu et al. (2007) were able to reliably infer limited demographic information from the content of visited sites. Due to the sparse sample of activity used, they first fit a model to predict site-level distributions over demographic dimensions, after which a smoothed, reduced dimensionality Bayesian framework was used to predict user age and sex from site visits. In an extension of this line of work, in Section 5 we use site visits alone to predict both these basic attributes as well as those that are typically harder to obtain, namely education, ethnicity, and household income.

3 Aggregate Usage Patterns

We begin by examining how users distribute their online time across different categories of websites, focusing on how these usage patterns relate to an individual's overall Internet activity level. For example, do the heaviest Web users behave qualitatively differently than those

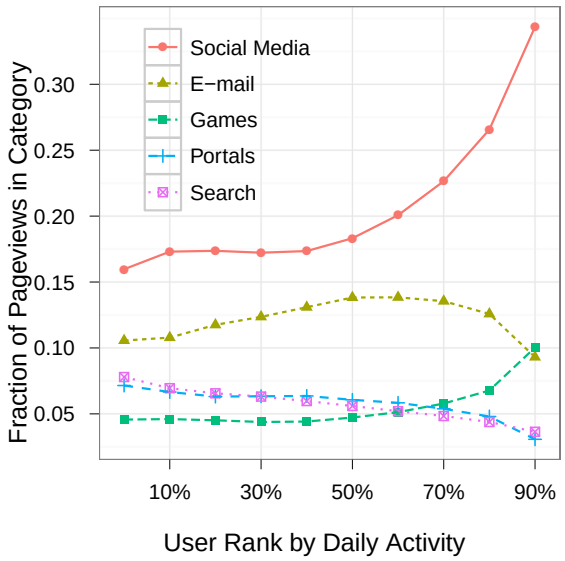

Figure 2: Variation in time spent on popular categories with overall activity.

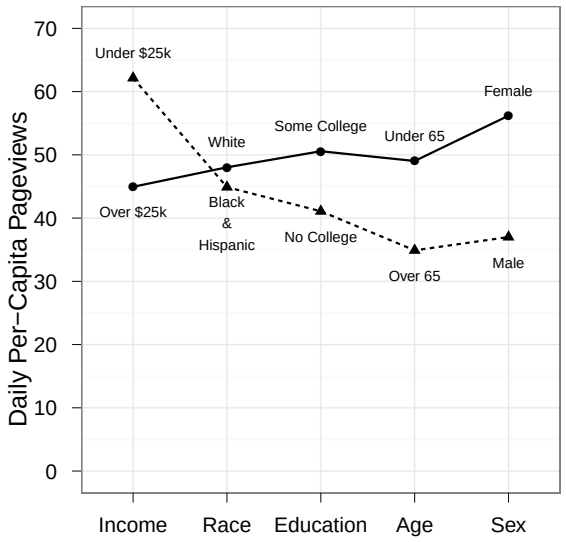

Figure 3: Average daily per-capita pageviews by demographic group for users 25 and older. The solid line represents majority groups and the dashed line represents minorities.

who use the Web infrequently, or do they simply visit the same types of sites more often? Given that the top 20% of users generate more than 60% of all pageviews, these heavy users' behavior is particularly consequential. We note that despite the simplicity of this question, it is difficult to reach an answer through conventional methods. Web panel data, by contrast, allow one to directly address such questions with high precision.

Figure 1 shows the fraction of time—as estimated via pageviews—allocated to the five most visited categories: social media (e.g., Facebook, MySpace & Twitter), e-mail (e.g., Yahoo! Mail, Windows Live & Gmail), games (e.g., Zynga & Pogo), portals (e.g., Yahoo! & AOL), and search (e.g., Google, Yahoo! & Bing). The distribution of views across categories is quite skewed, with the top five categories accounting for more than half of all Web activity, in agreement with previous work (Kumar and Tomkins, 2010). Even within these most popular categories, traffic is highly uneven: social media receives a dominant 25% of all pageviews, while portals and search receive less than 5% each.

Turning now to usage as a function of overall activity, we first bin users into deciles by daily pageviews; for each of these bins, we then compute how each group distributes their time across categories. The results, shown in Figure 2, reveal several trends, the most striking of which is a large increase in relative time on social media with total usage. For example, while the majority of users spend, on average, under 20% of their time on social media, the heaviest users allocate almost 35% of their time to this category. A similar, although less dramatic, effect is observed in online gaming. The opposite trend, however, is seen for e-mail, search, and portals, with the heaviest users spending less of their time on such sites.

In summary, while all groups spend the majority of their time in these popular categories, the heaviest users spend disproportionately more of their time on social media and less on e-mail relative to the overall population, suggesting there may be a tradeoff between the two activities. It is unclear, however, whether this finding foreshadows an overall shift towards social media, as today's heavy users may not be bellwethers for tomorrow's typical individual.

4 Reassessing the Digital Divide

In the 15 years since Hoffman and Novak highlighted the digital divide between those with and without Internet access (Hoffman and Novak, 1998), use of the Web has risen dramatically in the United States, increasing from roughly 30% to 80% of adults regularly going online.[1] Even with this tremendous growth in access, however, substantial inequalities persist across demographic groups. For example, in contrast to the 80% of adults in the general population who use the Internet, only about 70% of Blacks and 40% of people over 65 do so. Assessments and discussions of the digital divide typically focus on these disparities in basic access. Here we investigate a related but distinct ques-

[1]These and related data are available from the Pew Research Center at `http://pewinternet.org/Static-Pages/Trend-Data.aspx`.

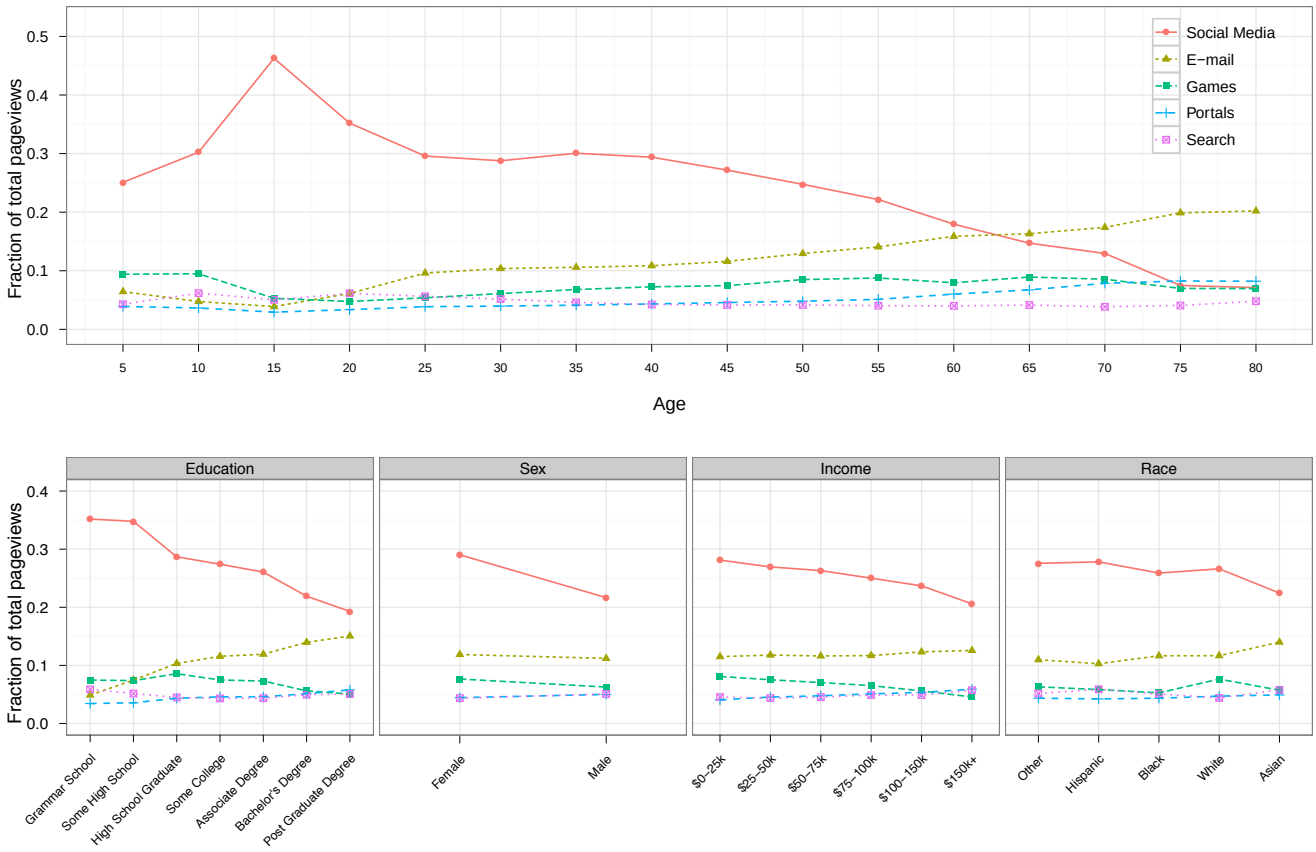

Figure 4: Fraction of time spent on the five most popular categories, split by demographic group. These percentages are normalized within group to adjust for variation in the total amount of time different groups spend online.

tion: Among those individuals who are already online, how do usage patterns vary across demographic groups?

Restricting to this population of active Internet users, we begin by estimating the amount of time different demographic groups spend online. Figure 3 shows pronounced differences in daily per-capita pageviews across groups when divided by income, education, age, and sex. (To account for skew in the distribution of time spent online, we take the geometric mean across users within each group.) Strikingly, though men and women have comparable *access* to the Internet, women spend considerably more *time* online than men, generating roughly 55 pageviews per day compared to less than 40 by men. Likewise, those individuals with at least some college education, as well as younger users, also spend much more time online relative to their less educated and older counterparts, exacerbating disparities in basic access between these groups. Younger and more educated individuals are not only more likely to *access* the Web at all, but are also more *active* compared to the general online population. Finally, while Blacks and Hispanics are less likely than Whites to go online, disparities in total usage are relatively small among those who use the Web.

Whereas above we looked at total time spent online by different demographic groups, we next consider how these groups allocate their time between categories of sites. Figure 4 shows usage across the five most popular categories, displaying the fraction of time each group spends on various activities. Similar to the overall distribution of time observed in Section 3, we see that all of the demographic groups we investigate spend the majority of their time visiting sites in these five categories. Despite these similarities, however, we also find significant variation in how different groups distribute their time amongst these popular categories. For example, while women spend nearly 30% of their time on social media sites, men spend approximately 20%. Social media usage, in fact, constitutes the most noticeable difference across demographic groups, with older, more educated, male, wealthier, and Asian Internet users spending a smaller fraction of their time on this category. Moreover, lower social media use by these groups is often accompanied by higher e-mail volume, similar to the overall trend noted earlier.

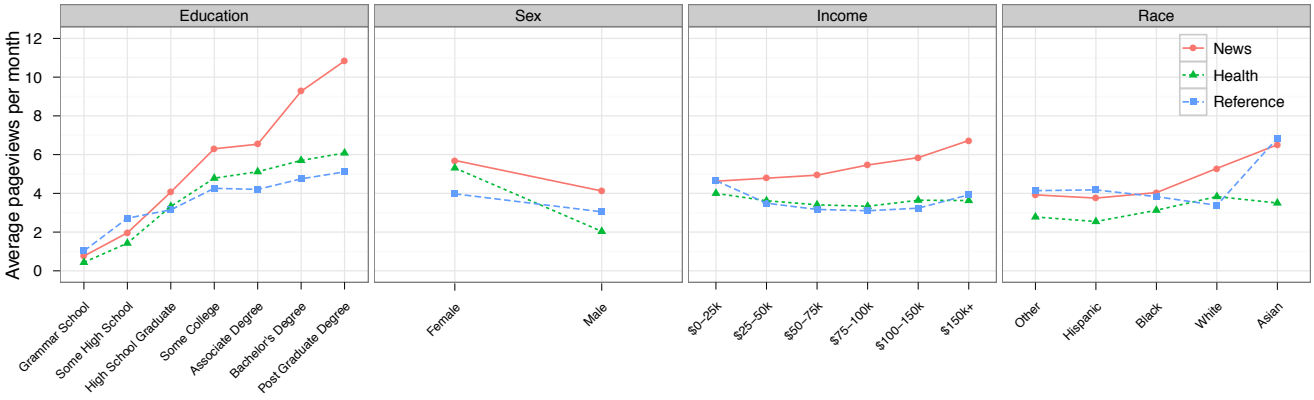

Figure 5: Use of news, health and reference sites by various demographic groups, as measured by monthly pageviews in each category.

Concern over the digital divide typically stems from the belief that Internet access improves quality of life along several core dimensions, including health and education, and that lack of access thus leaves certain subpopulations at a substantively important disadvantage. Moving away from the most popular categories, we now examine usage in three areas that directly relate to these particularly consequential outcomes: news, health (e.g., WebMD), and reference (e.g., Wikipedia). While our previous analysis measured site use in relative time (i.e., as a percentage of total activity), the value of visiting news, health and reference sites is likely minimally affected by how much total time one spends online. We thus report absolute use for these categories, with Figure 5 showing large usage differences in average monthly pageviews across demographic groups. Adults with a post-graduate degree, for example, spend more than three times as much time on health sites than those with only some high school education, and Asians spend more than 50% more time browsing online news than do other race groups. We emphasize that these results pertain to the subpopulation of users who already have Internet access, meaning that even when less educated and less wealthy groups gain access to the Web, they still use the Internet only relatively infrequently to access news, health and reference information.

As discussed above, there are stark differences in how various demographic groups use the Web, both in visits to the most popular categories of social media and e-mail, as well as in the potentially consequential categories of news, health, and reference. What drives these effects? Is it the case, for example, that differences among ethnic groups are largely accounted for by variation in educational attainment? To investigate this question, we estimate use of news, health, and reference sites as a function of several factors, including age, education, sex, income, race, and total Internet usage. Individual-level pageviews in each category are modeled via linear regression, where we include all available demographic features, their pairwise interactions, and quadratic terms. Specifically, for each category we have

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

where p_i is the number of pageviews for the i-th user in that category, x_{ij} specify user-level attributes, and the error terms ϵ_i are normally distributed. Though the observed variation is not accounted for by any single dimension, we indeed find that the effects of race largely disappear after controlling for other variables, while educational attainment continues to have the largest effect of the demographics included in our model. Figure 6 illustrates this point, showing model-adjusted estimates of usage as education and race are varied, with all other attributes held constant.[2] Moreover, in the case of health sites, we find that sex is an important determinant of usage, with women spending considerably more time in this category than men (Figure 7), a trend that has been suggested by past survey research (Fox, 2011).

Our examination of Internet usage reveals a nuanced picture of the digital divide. On the one hand, when older and less educated users gain access to the Web, their use remains relatively low, as measured both in total time spent online as well as in visits to the potentially consequential categories of news, health, and reference. These educational disparities are especially large, and persist even after controlling for other variables. On the other hand, differences between Whites and underrepresented minorities largely vanish among those who are already online. These results illustrate that the digital divide is more than a simple question

[2]In computing model estimates, we consider a prototypical individual who is 30 years old with a household income between $50,000 and $75,000.

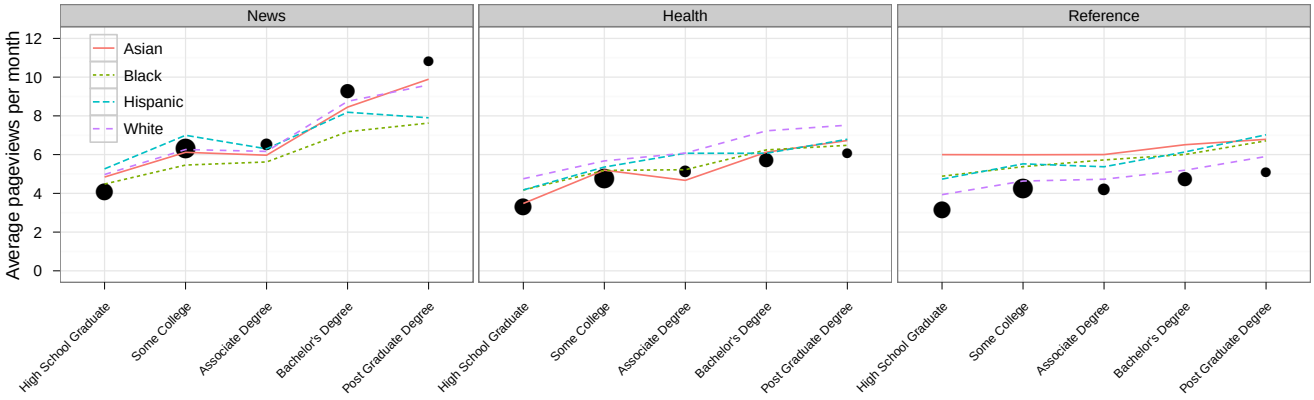

Figure 6: Observed and model-adjusted usage of news, health, and reference sites, by education and race. Circles show the observed group averages—with area proportional to the size of each group—and lines indicate the modeled values for a typical female user of the specified race.

of access, and highlight some of the issues that remain even among those with access to the Web.

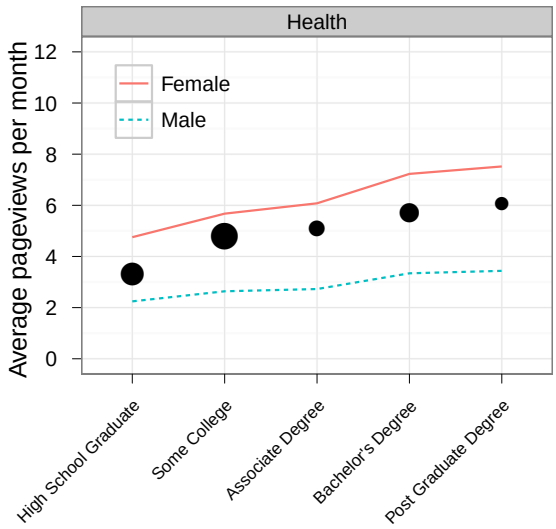

Figure 7: Observed and model-adjusted usage of news, health, and reference sites, by education and sex. Circles show the observed group averages—with area proportional to the size of each group—and lines indicate the modeled values for typical female and male users.

5 Inferring Demographics from Web Browsing Histories

Online advertisements are often targeted at particular demographic groups, usually defined in terms of age and sex. More sophisticated demographic targeting strategies, however, are hindered by the difficulty of obtaining comprehensive individual-level information. Here we explore the possibility of inferring detailed demographics from an individual's Web browsing history, which itself can be reconstructed by ad networks that track the movement of individuals across the Internet.[3]

Before discussing our methodology and results, we note that the feasibility of this demographic prediction task is unclear *a priori*. Our analysis of the digital divide shows that by and large different demographic groups spend the majority of their online time engaged in similar activities, namely social media and e-mail. Moreover, notwithstanding large group-level differences in certain categories, individual-level variation is generally quite large, diminishing discriminative power. On the other hand, however, there are many popular sites with extremely skewed audience demographics. For example, Fox News attracts millions of visitors each month, of which more than 90% are White, leaving open the possibility of accurately inferring demographics from browsing histories via visits to such informative sites.

For each of the five demographic dimensions—sex, age, race, education, and household income—we formulate the prediction task as binary classification (e.g., distinguishing between females and males, or Whites and non-Whites). Predictions are based on site-level visits to avoid loss of information in aggregating pageviews to categories. Web addresses for visited sites are normalized by retaining at most three domain levels and removing any remaining 'www' prefix. For example, `www.facebook.com` is transformed to `facebook.com`, and `us.mg0.mail.yahoo.com` is con-

[3]We point the reader to our proof-of-concept implementation at `http://bit.ly/surfpreds`, which predicts user demographics based on visits to the 10,000 most popular sites using the model discussed below.

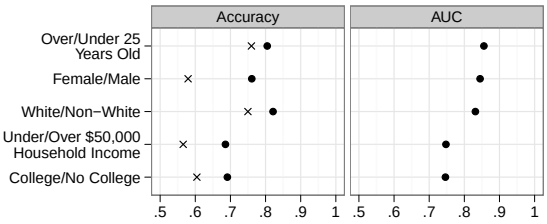

Figure 8: Summary of model accuracy (left) and AUC (right), indicated by solid circles, for all demographic attributes. Population skew is given by x's for comparison.

verted to `mail.yahoo.com`. Each user is represented by a sparse binary vector x_i indicating which of the top 100,000 most popular of these normalized domains they visited during the study period (i.e., $x_{is} = 1$ if the i-th user visited the s-th site); for household income, we consider all domains visited by the entire household.[4]

Given the scale of the data, and that the number of features is comparable to the number of examples, we use linear support vector machines (SVMs) (Schölkopf, Burges, and Smola, 1999) to infer demographic attributes from browsing history. Linear SVMs generate predictions of the form

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

where the predicted class is defined by the sign of $\hat{y}(x_i) \in \mathbb{R}$. To guard against overfitting, SVMs seek the weight vector w that maximally separates the positive and negative examples in the training set. Specifically, SVMs minimize the loss function

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

where $[z]_+ = (|z| + z)/2$ is the positive part of z, $y_i \in \{-1, 1\}$ encodes observed class in the data, and C is a tunable parameter that balances model fit against generalization error. Users are randomly divided into an 80% training set on which models are fit, a 10% validation set used to select the optimal parameter C for each demographic attribute, and a 10% held-out test set on which we evaluate and report final performance.

Figure 8 summarizes our results for all five demographic classification tasks. The left panel displays the accuracy of predictions, showing reasonable performance across all demographic dimensions, with slightly higher accuracies for age, sex, and race—80%, 76%, and 82%, respectively—than for education and income—70% and 68%. To help put these numbers in perspective, Figure 8 also includes the overall population skew for each demographic attribute, indicated by x's (e.g., 57% of the panel is female, while 76% is comprised of adults).

[4]Variants of these features, for example visitation frequency or tf-idf weights, result in similar performance.

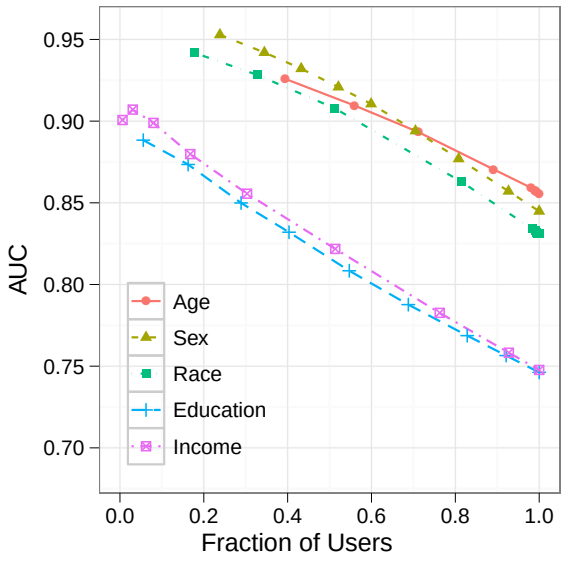

Figure 9: Change in model performance as the fraction of users on which predictions are made is varied. The rightmost points include predictions for all users, whereas points to the left are restricted to users farther from the decision boundary.

To adjust for the substantial demographic skew in the population, we also present AUC—or area under the ROC curve—in the right panel of Figure 8, a measure that effectively re-normalizes the majority and minority classes to have equal size. Intuitively, AUC is the probability that a model scores a randomly selected positive example higher than a randomly selected negative one (i.e., the probability that the model correctly distinguishes between a randomly selected female and male). Though an uninformative rule would correctly discriminate between such pairs 50% of the time, predictions based on browsing histories are relatively reliable, ranging from 74% to 85%. Thus, whether we measure performance in terms of accuracy or AUC, we find that browsing activity provides a strong signal for inferring individual-level demographic attributes.

The comprehensive nature of large-scale Web panel data enables us to achieve these results with less computational effort than previous approaches (Hu et al., 2007), and moreover, yields an interpretable predictive model. In particular, we can inspect the weight vector w to confirm that the most predictive (i.e., largest positively and negatively weighted) sites for each attribute are intuitively reasonable. We find, for example, that visiting the popular cosmetics company `lancome-usa.com` strongly indicates that a user is female, while visits to sports sites such as `espn.go.com` are highly predictive of being male. We note that though visits to such heavily weighted sites provide strong signals, many such sites are frequented by a rel-

atively small fraction of the population. Model performance thus benefits both from these highly informative features, as well as the many weak signals from visits to popular but less discriminating sites.

Finally, we investigate how model performance can be improved by focusing on the most "stereotypical" users. For example, those individuals who exclusively visit sports sites should be easier to classify as male than ones who visit both sports and cosmetics sites. In quantitative terms, by restricting predictions to users for whom the model is most confident, we expect gains in performance. Figure 9 illustrates this tradeoff between performance and coverage as we restrict to increasingly smaller sets of users that lie far from the decision boundary. For example, by restricting to the 50% percent of the population for which the model is most confident, we achieve AUC of approximately 0.9 for sex, race, and age, and roughly 0.8 for education and income—a roughly five percentage point increase compared to predictions over the entire population.

6 Conclusion

By exploring several diverse aspects of online browsing behavior, our work demonstrates the value of large-scale Web panel data for better understanding how people use the Internet, with consequences ranging from public policy to advertising. In particular, our data-intensive approach is well-suited for accurately measuring individual activity, a challenging task to accomplish via more conventional techniques. More broadly, this work highlights how large-scale data, together with the appropriate computational tools, can be leveraged to address otherwise difficult to study questions in the social sciences, illustrating the potential of the emerging field of computational social science.

Though this paper provides a novel perspective into online behavior, in many ways our findings raise as many questions as they answer. For example, in assessing the digital divide, we observe that while all demographic groups spend the majority of their time on the same popular activities (e.g., social media and e-mail), there are pronounced disparities in how frequently different groups access the particularly relevant categories of news, health, and reference. What are the effects, however, of these observed differences in behavior? While it is certainly reasonable that easier access to information on, for example, nutrition and contraception would lead to better health outcomes, rigorously establishing this relationship is not easy. In this regard, accurately measuring online behavior is an important, although preliminary, step in making informed decisions, leaving several directions for future work.

Acknowledgments

We thank Mainak Mazumdar and the Nielsen Company for providing the Web panel data, and Duncan Watts and David Pennock for helpful conversations.

References

Boyd, D. 2010. *White flight in networked publics? How race and class shaped American teen engagement with MySpace and Facebook*. Routledge. 203–222.

Chang, J.; Rosenn, I.; Backstrom, L.; and Marlow, C. 2010. ePluribus: Ethnicity on social networks. In *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI Press.

Fox, S. 2011. Health Topics. Pew Internet & American Life Project.

Hargittai, E. 2007. Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication* 13(1).

Hoffman, D., and Novak, T. 1998. Bridging the racial divide on the internet. *Science* 80(5362):390–391.

Hu, J.; Zeng, H.-J.; Li, H.; Niu, C.; and Chen, Z. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 151–160. ACM Press.

Jones, R.; Kumar, R.; Pang, B.; and Tomkins, A. 2007. I know what you did last summer: query logs and user privacy. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM)*, 909–914. ACM Press.

Kumar, R., and Tomkins, A. 2010. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 561–570. ACM.

Mislove, A.; Lehmann, S.; Ahn, Y.; Onnela, J.; and Rosenquist, J. 2011. Understanding the demographics of twitter users. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Schölkopf, B.; Burges, C.; and Smola, A. 1999. *Advances in Kernel Methods: Support Vector Learning*. The MIT press.

Smith, A. 2010. Home broadband adoption 2010: Summary of findings. Pew Internet & American Life Project.

Watkins, S. 2009. *The Young & Digital: What the Migration to Social Network Sites, Games, and Anytime, Anywhere Media Means for Our Future*. Columbia University Press.

Weber, I., and Castillo, C. 2010. The demographics of web search. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 523–530. ACM Press.

Weber, I., and Jaimes, A. 2011. Who uses web search for what? And how? In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*.