ePluribus : Ethnicity on Social Networks

Jonathan Chang and **Itamar Rosenn** and **Lars Backstrom** and **Cameron Marlow**
Facebook
1601 S California Ave
Palo Alto, California 94304
{jonchang, itamar, lars, cameron}@facebook.com

Abstract

We propose an approach to determine the ethnic breakdown of a population based solely on people's names and data provided by the U.S. Census Bureau. We demonstrate that our approach is able to predict the ethnicities of individuals as well as the ethnicity of an entire population better than natural alternatives. We apply our technique to the population of U.S. Facebook users and uncover the demographic characteristics of ethnicities and how they relate. We also discover that while Facebook has always been diverse, diversity has increased over time leading to a population that today looks very similar to the overall U.S. population. We also find that different ethnic groups relate to one another in an assortative manner, and that these groups have different profiles across demographics, beliefs, and usage of site features.

1. Introduction

The ethnicity[1] of a user base is an important demographic indicator that can be used for marketing, compliance, and analytics as well as a scientific tool for understanding social behavior and increasing diversity through outreach efforts. Unfortunately, ethnic information is often unavailable for practical, legal, or political reasons.

In this paper, we propose a technique that combines census information with user features such as surname to infer the overall ethnic breakdown of a population as well as the ethnicity of individual users. Our technique leverages the machinery of mixture modeling and we demonstrate that it achieves better predictive accuracy than other natural alternatives.

We then apply our technique to Facebook[2], a social networking site that does not explicitly collect ethnicity. Using the inferred ethnicity of the U.S. user base, we answer the questions:

- How diverse are Facebook users?
- How do users of different ethnicities use Facebook?
- What are the demographic characteristics of each ethnicity on Facebook?
- How do ethnicities interact on Facebook?

We report that Facebook ethnicities are assortative in their interactions, which is consistent with other studies. We also estimate that Facebook has always been diverse and that its diversity has increased significantly over the past year, to the point where the diversity of Facebook users in the U.S. nearly mirrors that of the overall population of the country.

2. Methodology

In this section, we describe a probabilistic, Bayesian approach to estimating the distribution of ethnicities of a population given only their names. Our model takes a list of users' names and applies both census name statistics and unsupervised learning techniques to estimate the ethnicities of members of the population.

We emphasize that our approach is only lightly supervised using census name statistics. Because representative ground truth is impossible to obtain, traditional supervised discriminative methods such as logistic regression and support vector machines cannot be directly applied to this problem.

2.1 Intuition

The U.S. Census Bureau's Genealogy Project publishes a data set containing the frequency of popular surnames along with a breakdown by race and ethnicity.[3] These data provide the rank in the population, the total count of people with the name, their proportion per 100,000 Americans, and the percent for various ethnicities: White, Black, Asian/Pacific Islander (API), American-Indian/Alaskan Native (AI/AN), two or more races and Hispanic.[4]

This data set allows us to predict a person's ethnicity based solely on their surname. Table 1 shows the top three surnames within the top 1,000 ordered by the percent in a given group. It shows that some ethnicities have distinctive surnames while others do not. For instance, 98.1% of individuals with the name Yoder are White, while the most predictive name for American Indian / Alaskan Native individuals only

[1] Throughout this paper we use the term "ethnicity" to refer to any racial or ethnic census category.

[2] http://www.facebook.com

[3] http://www.census.gov/genealogy/www/data/2000surnames/index.html

[4] While there are many preferences for describing people's ethnicity, we have chosen to use the terms used in the U.S. Census to be consistent with our data.

Name	Rank	Count	% in group
Caucasian			
Yoder	707	44245	98.1%
Krueger	863	36694	97.1%
Mueller	467	64305	97.0%
African American			
Washington	138	163036	89.9%
Jefferson	594	51361	75.2%
Booker	902	35101	65.6%
Asian / Pacific Islander			
Zhang	963	33202	98.2%
Huang	697	44715	96.8%
Choi	872	57786	96.4%
American Indian / Alaskan Native			
Lowery	752	41670	4.4%
Hunt	157	151986	3.9%
Sampson	844	37234	3.8%
Two or more races			
Khan	665	46713	15.6%
Singh	396	72642	15.3%
Ali	876	36079	13.4%
Hispanic			
Barajas	989	32147	96.0%
Orozco	690	45289	95.1%
Zavala	938	34068	95.1%

Table 1: Surnames highly indicative of each ethnicity. The subtable for each ethnicity shows the three surnames with the highest proportion of people in that ethnic group.

has 4.4% in that group. Because of the smaller sizes and less identifiable names of the American Indian / Alaskan Native and two or more race groups, we will only look at White, Black, Asian/Pacific Islander and Hispanic predictions in our analysis. A simple technique for finding the distribution of ethnicities of a population is based solely on the census numbers. We first count the number of people with each surname listed in the Census Genealogy data. For each of these names, we estimate the total number of each ethnicity by multiplying the number of people with the percentages above.

There is a rich history of authors using associations between surnames and ethnicities in curated sources such as census data to infer ethnicities (Ambekar et al. 2009; Buechley 1976; Coldman, Braun, and Gallagher 1988; Fiscella and Fremont 2006; Kali et al. ; Lauderdale and Kestenbaum 2000; Tucker 2005). One potential source of error in this estimate comes from our assumption that users are selected at random from the U.S. population. What if the population of interest is primarily White? Wouldn't a majority of the Smiths be White then, breaking our assumption? In order to address this concern, we refine our estimates using the model described in the following section. By allowing the population to be different from the curated source (census) population, and for each name to inform our interpretation of every other name, we demonstrate in Section 3. that our model more accurately estimates the total number of users of a given ethnicity.

2.2 Model

To address the concerns in the previous section, we take a Bayesian, probabilistic approach to estimating the ethnicities of members of a population. We imagine a stochastic process

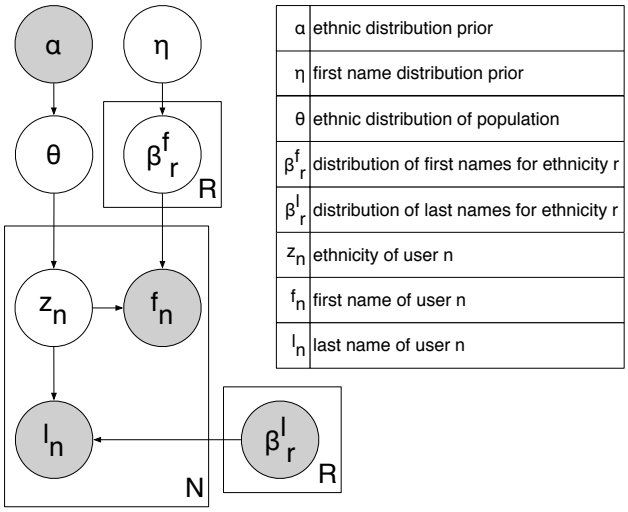


Figure 1: A graphical model representation of the model used to infer ethnicities. Shaded nodes are observed variables and unshaded nodes are unobserved. Plates indicate replication.

with individual ethnicities as hidden variables and individuals' names as observed variables. The problem of inferring ethnicity then becomes a problem of inferring the most likely value of these hidden variables given the names, a procedure we describe below.

Our model builds on the work in mixed-membership modeling (Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Lafferty 2004; Steyvers and Griffiths 2007) which allows populations to participate in multiple latent categories. Similar in spirit to our work is the work in (Tucker 2005) which, instead of defining a probabilistic model, iteratively adjusts a set of hard assignments. (Ambekar et al. 2009) uses the machinery of hidden Markov modeling to model names as a series of fragments.

In contrast, our model assumes that individuals are members of a population with unknown ethnic proportions. Individuals are drawn from the population and assigned names based on their ethnicity. Formally, let R denote the number of ethnicities and N denote the number of people in the population. Our model then uses the following generative process:

1. for each ethnicity $r \in \{1 \dots R\}$,
   (a) draw the distribution of first names for that ethnicity, $\boldsymbol{\beta}^f_r \sim \mathrm{Dir}(\boldsymbol{\eta})$;
2. draw the ethnic breakdown of the aggregate population, $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$;
3. for each person $n \in \{1 \dots N\}$,
   (a) draw the ethnicity of the individual from the aggregate population, $z_n \sim \mathrm{Mult}(\boldsymbol{\theta})$;
   (b) draw the surname of the individual based on the ethnicity, $l_n \sim \mathrm{Mult}(\beta^l_{z_n})$;
   (c) draw the first name of the individual based on the ethnicity, $f_n \sim \mathrm{Mult}(\beta^f_{z_n})$.

The parameters of the model are the multinomial distribution of names for each surname described in the previous section,

$\boldsymbol{\beta}^l_r, r \in \{1 \dots R\}$ and the Dirichlet hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$. We set the $\boldsymbol{\alpha}$ proportional to the ethnic breakdown of the population at large, and $\boldsymbol{\eta}$ so as to have a weak, symmetric Dirichlet prior on $\boldsymbol{\beta}^l_r$. This model is depicted graphically in Figure 1.

To determine the values of the hidden variables of the model — the individual ethnicities (z_n), the ethnicity of the aggregate population ($\boldsymbol{\theta}$), and the ethnic breakdown of first names ($\boldsymbol{\beta}^f$) — given the observed variables, we must perform *posterior inference*. That is, we must estimate

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}^f_{1:R}, z_{1:N} | f_{1:N}, l_{1:N}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta}^l_{1:R}).$$

Exact posterior inference is intractable for this model, so we turn to the approximate inference technique CVB0 described in (Asuncion et al. 2009).

In the sequel, we report results using the expected value of each hidden variable under the approximate posterior. For convenience, we define $\boldsymbol{\pi}_n$ as the estimated ethnic distribution of user n,

$$\boldsymbol{\pi}_n \triangleq p(z_n | f_{1:N}, l_{1:N}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\beta}^l_{1:R}),$$

where $\boldsymbol{\pi}_n$ is a vector of length R whose rth element expresses the probability that the user is of ethnicity r.

In Section 4., we analyze the relationships between ethnicities. For this, it is helpful to define the matrix of ethnic probabilities, Q_{n_1,n_2} for a pair of users n_1, n_2,

$$\boldsymbol{Q}_{n_1,n_2} \triangleq \boldsymbol{\pi}_{n_1} \boldsymbol{\pi}^\mathsf{T}_{n_2}, \tag{1}$$

where Q_{n_1,n_2} is an $R \times R$ matrix whose (r_1, r_2)th entry expresses the probability that user n_1 is of ethnicity r_1 and user n_2 is of ethnicity r_2.

We also aim to understand how some continuous variable associated with each user (such as age) correlates with ethnicity. We define the ethnicity-weighted mean of this variable as

$$\boldsymbol{\mu}_c \triangleq \frac{\sum_n \boldsymbol{\pi}_n c_n}{\sum_n \boldsymbol{\pi}_n}, \tag{2}$$

where division is element-wise. $\boldsymbol{\mu}_c$ is a vector of length R whose rth element expresses the mean value of the continuous variable c for users of ethnicity r.

When the variable c is categorical, taking on values $1 \dots K$, we define the weighted mean to be

$$\boldsymbol{M}_c \triangleq \frac{\sum_n \boldsymbol{\pi}_n \boldsymbol{c}^\mathsf{T}_n}{\sum_n \boldsymbol{\pi}_n}, \tag{3}$$

where division is row-wise. c_n is a categorical variable encoded as a length K indicator variable. $\boldsymbol{M}_c$ is an $R \times K$ whose (r, k)th element expresses the fraction of users who are of ethnicity r and whose value of c_n is k.

3. Validating the model

In order to validate the method described above, we collect ground-truth data from the social-networking website Myspace.[5] Our data set consists of 77954 users with self-reported names and ethnicities. We collect these profiles by searching

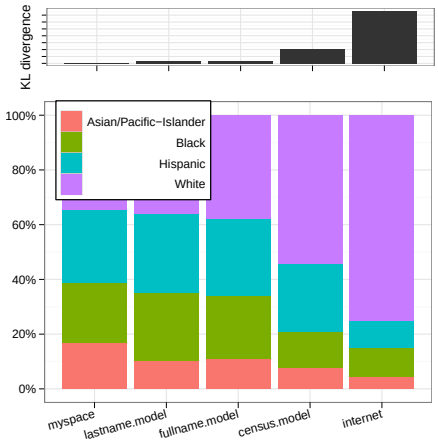


Figure 2: A comparison of different methods of estimating the ethnicity of Myspace users. The upper graph shows the KL-divergence between each model's estimate and the ground truth. The lower graph shows the predicted ethnic distribution for each method. The model we propose outperforms natural alternatives.

for users through the public search and specifying a given ethnicity. The list of results is biased by some internal update mechanism, so the percentages are unlikely to be representative of the Myspace population at large, but should be useful for evaluating the performance of our algorithm. We map the ethnic categories defined by Myspace onto Census-defined ethnicities.

Figure 2 compares various methods for estimating the ethnic breakdown of Myspace users. We emphasize that none of the estimation methods observe the ground truth. The columns of Figure 2 should be interpreted as follows:

myspace the ground-truth self-reported ethnicities of Myspace users;

lastname.model the proposed model of Section 2.2 *except* without observed first names;

firstname.model the proposed model of Section 2.2;

census.model the simple census-based model described in Section 2.1;

internet the estimated ethnic breakdown of Internet households on values from the National Telecommunications and Information Administration report on the Networked Nation.[6] We use the percent of households with Internet access as a proxy for the addressable Internet population of each ethnicity.

The lower graph of Figure 2 shows the predicted ethnic distribution for each method. The proposed model, with or without first names, is much closer to the truth ground than the alternatives. Taking simple sums of census surname data (as described in Section 2.1), or using the background ethnic breakdown of Internet households overestimates the proportion of White users, while underestimating Black and Asian/Pacific-Islanders. Because the user base is not representative of the census population, these methods fail to infer

[5] http://www.myspace.com

[6] http://www.ntia.doc.gov/reports/2008/NetworkedNation.html

Rank	White	Black	Asian / Pacific Islander	Hispanic
1	barb	latoya	rahul	luis
2	conor	latonya	syed	javier
3	peg	deandre	wei	jose
4	deb	lakeisha	minh	jorge
5	kurt	tameka	nguyen	hector
6	colleen	latrice	tuan	yesenia
7	meghan	jermaine	thanh	mayra
8	meaghan	lashonda	sandeep	julio
9	connor	jamaal	phuong	alejandro
10	brendan	lakisha	yi	cesar

Table 2: First names most associated with each ethnicity learned by the proposed model.

an accurate distribution over ethnicities. In contrast, our proposal, which takes this non-representativeness into account, is able to better model the ethnic breakdown of Myspace users.

We quantitatively measure the accuracy of each of these models using a measure of distributional similarity, KL-divergence, $\mathrm{KL}(q||p) = \sum_r q(r) \log \frac{q(r)}{p(r)}$ in the upper graph of Figure 2. A lower KL-divergence means that the estimated distribution is closer to the ground truth. As qualitatively observed in the previous paragraph, the **internet** and **census.model** estimation techniques diverge from the ground-truth, while the estimates of **lastname.model** and **fullname.model** are more accurate. In KL-divergence, **fullname.model** performs better than **lastname.model**. Because our data set is small, the amount of information that can be learned about each first name is small; we hypothesize that for larger data sets where more can be learned about each first name, the boost in accuracy of **fullname.model** over **lastname.model** will be larger.

In Figure 3, we visualize the ability of the proposed model to predict the accuracy of individual ethnicities in addition to the ethnic breakdown of the aggregate population. Each column represents a ground truth ethnicity; the stacks indicate the number of people predicted by the model to be of each ethnicity for that ground truth ethnicity. Ideally, the first bar would be entirely red, the second entirely green, and so on. Using the census model described in Section 2.1 will underestimate Asian/Pacific-Islanders, Blacks, and Hispanics. Using our model with just last names improves upon this estimate, increasing the number of people correctly identified in each of these categories. Using both first and last names further improves estimates, largely by making better distinctions between White and Black.

The model is able to learn first names associated with each ethnicity and leverage this information to improve its estimates of individuals' ethnicity. To illustrate this, we apply the model to the larger user base of U.S. Facebook users (described in the following section) to generate Table 2, which shows the first names most indicative of each ethnicity, i.e., those names with the highest smoothed posterior probability for each ethnicity.

4. Application to Facebook users

In this section we apply the approach described in Section 2.2 to the set of U.S. Facebook users. Using our proposed model, we report on how the diversity of Facebook has changed over time, how members of different ethnicities interact, and the demographic characteristics of each ethnicity.

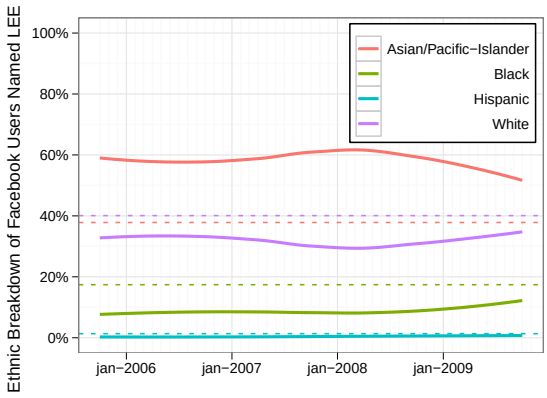


Figure 4: Ethnic breakdown of U.S. Facebook users with the surname LEE over time (solid lines). Dashed lines show the proportion of each ethnicity in the census.

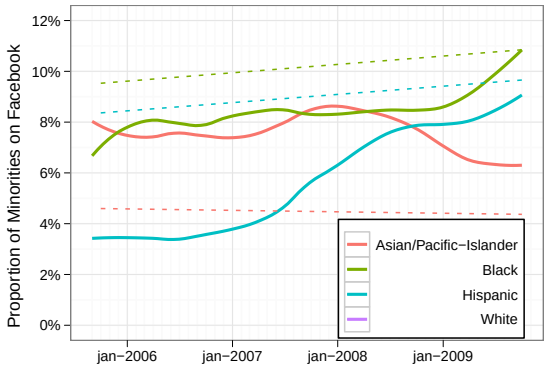
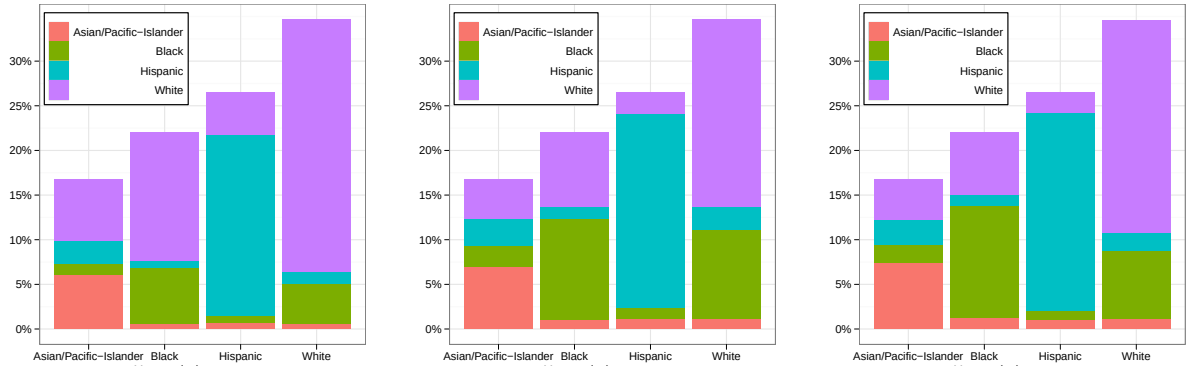


Figure 5: Ethnic breakdown of U.S. Facebook users over time (solid lines). Dashed lines show the proportion of each ethnicity in the addressable Internet population. Although Asian/Pacific-Islanders are still overrepresented, other minorities have nearly reached parity with the addressable Internet population.

Diversity over time The technique of Section 2.2 allows us to obtain a picture of the relative makeup of Facebook's racial subpopulations within the United States. Because the Facebook population is changing over time, as is the ethnic diversity of addressable Internet users, we compare these groups over time.

To illustrate how our model changes along with the changing population of Facebook users, Figure 4 shows the model's changing estimate of the distribution for the surname LEE. The dashed lines show the ethnic breakdown of people named LEE given by the Census Bureau tables. The disparity between the solid and dashed lines shows the possible bias when estimating ethnicity without the adjustments made by our model. For instance, the Census numbers would underestimate the number of Asian/Pacific Islanders on Facebook and overestimate the number of Black users on Facebook.

Figure 5 shows our model's prediction of the ethnic breakdown of the U.S. Facebook population over time (solid lines).



(a) census.model (b) lastname.model (c) fullname.model

Figure 3: A comparison of three methods for estimating the ethnicity of individual Myspace users. Each graph breaks down the model's predictions of ethnicity by the ground truth ethnicity (column).

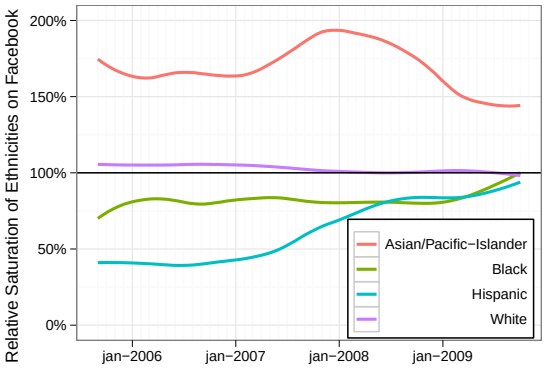


Figure 6: Relative saturation of ethnicities on Facebook. As the lines converge towards 100% (center), the makeup of U.S. Facebook converges towards that of the addressable Internet population.

The dashed lines show the breakdown of U.S. Internet households. (Because White users are a large majority, we have left them out of this plot.) Although Asian/Pacific-Islanders are still overrepresented, other minorities have nearly reached parity with the addressable Internet population.

Another approach to visualizing this data is to look at the relative saturation of each race. This is the fraction of users on Facebook compared to the fraction we would expect from the U.S. Internet population at that time. For instance, if Facebook had 100M users, and Asian Americans made up 4.4% of the U.S. Internet population, we would expect to find 4.4M Asian users on Facebook. If instead we observe 5M then the relative saturation would be roughly 114%.

Figure 6 shows Facebook saturation by ethnic and racial groups. Since 2005, Asian/Pacific Islanders have been much more likely to be on Facebook than Whites, which remains true to this day. While Hispanics were once only 40 percent as likely as Whites to be on the site, this likelihood has been steadily climbing since early 2007, and is currently at 80 percent. The graph also shows that Black users are now about equally as likely to be on the site as White users.

Interactions between ethnicities Predicting the ethnicity of each individual in a population allows us to understand how different ethnicities relate to one another, as described in Equation 1. Here, we analyze two kinds of Facebook relationships: romantic and friendship. We do so by examining the number of pairs of people in each relationship, broken down by ethnicity.

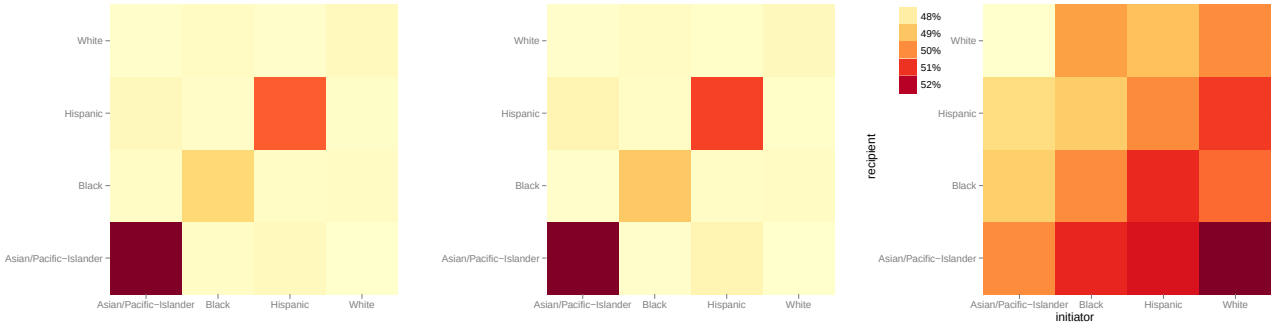
Let $\mathscr{R}$ denote the set of pairs engaged in the relationship of interest. Then let $\hat{\boldsymbol{Q}} \triangleq \frac{1}{|\mathscr{R}|} \sum_{(r_1,r_2)\in\mathscr{R}} \boldsymbol{Q}_{r_1,r_2}$ be a $R \times R$ matrix whose entries denote the estimated proportion of relationship being of a particular pair of ethnicities. Because the entries of this matrix will be biased towards highly frequent ethnicities, it is helpful to divide through by the expected value of each matrix entry if relationships were formed without regard to ethnicity, $\boldsymbol{Q}^* \triangleq \boldsymbol{\theta}^*\boldsymbol{\theta}^{*\mathsf{T}}, \boldsymbol{\theta}^*_r = |\{(r_1, r_2)|r_1 = r\}|$. The entries of the normalized matrix characterizes the dependence of the relationship on that ethnicity pair.

The normalized ethnic breakdown of romantic relationships is shown in Figure 7(a). Dark tiles indicate ethnic pairs who relate more frequently than expected by chance, while lighter tiles indicate ethnic pairs who relate less frequently. The dark tiles along the diagonal indicate that minority ethnicities are assortative in their relationship preferences. The lighter tiles on the off-diagonal indicate that ethnicities have a dispreference towards romantic relationships with other ethnicities.

The normalized ethnic breakdown of friendships using both normalizations is shown in Figure 7(b). The trends in romantic relationships are echoed in the friendships.

Figure 7(c) shows, for each pair of ethnicities, the fraction of friendships initiated by a user of the column ethnicity. Darker tiles indicate a prevalence towards initiation. Thus Asian/Pacific-Islanders are less likely to initiate friendships with those outside their ethnicity, while Whites are most likely to do so. The inclination or disinclination to seek friendships with those outside one's own ethnicity helps explain the insularity pattern observed in Figure 7(a) and (b).

The previous figures are suggestive that individuals' ethnicity can be predicted through their social ties. We explore this possibility by estimating each user's ethnicity using the



(a) Romantic (b) Friendship (c) Friendship initiations

Figure 7: Relationships by ethnicity. Darker tiles indicate ethnic pairs who relate more frequently than expected by chance, while lighter tiles indicate ethnic pairs who relate less frequently. (a) shows relationships between users in romantic relationships and (b) shows friendships. The darker tiles along the diagonal indicates that minorities are assortative in their relationship preferences. (c) shows the probability that a relationship between two given races is initiated by one.

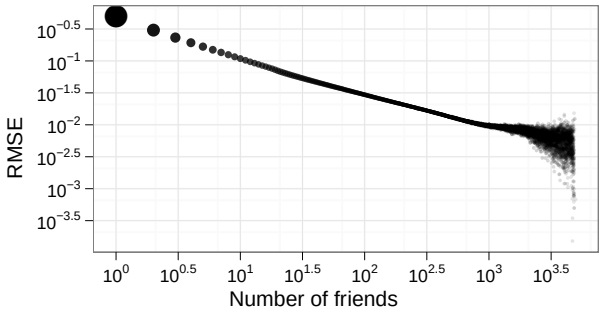


Figure 8: Root mean squared error when predicting users' ethnicity based on the ethnicity of their friends. Errors are plotted against the number of friends the user has. Points are sized according to the number of users having said number of friends.

average ethnicity of their friends, as imputed by our model. Because we do not have ground truth, we compare the ethnicity of each user as estimated by their friends versus their ethnicity as estimated by their name. The root mean squared difference between these two ethnic distributions is plotted in Figure 8, grouped by the number of friends the user has. The size of each point denotes the numer of users having that number of friends. As the number of friends increases, the estimate of a user's ethnicity based on their friends improves until it achieves an error of 0.01 when the number of friends reaches 1000. For users having more than 1000 friends it becomes more difficult to tease out trends because of smaller sample sizes but it appears there is no longer a steady improvement in estimates with increases in friend count. If a user becomes too promiscuous in their friending activity, thereby gaining more friends, those additional friends provide too little information with which to discriminate that user's ethnicity.

Ethnic demographics Predicting the ethnicity of each individual in a population also enables us to understand the different demographic characteristics of each ethnicity, as described in Equation 2 and Equation 3. We compute the mean value of a variable for each ethnicity to analyze how usage, language, political affiliation, gender, and geography depend on ethnicity. As in Section 4., it will be useful to divide each variable by its mean value among the entire U.S. Facebook population to better gauge how the variable depends on ethnicity.

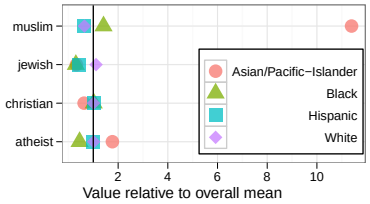


Figure 9: The religion of U.S. Facebook users by ethnicity. Numbers are relative to the total number of U.S. Facebook users with that locale.

Figure 9 shows how users' religion depend on ethnicity, for a few major religions. Asian/Pacific Islanders are much more likely than average to be Muslim; Blacks also self-identify as Muslim more frequently than one would expect by chance. In contrast, both Asian/Pacific-Islanders, Blacks, and Hispanics are less likely to identify as Jewish.

Figure 10 shows how a user's locale depends on ethnicity. Asian/Pacific Islanders are the most distinct, heavily favoring Chinese, Vietnamese, Korean, and Japanese whereas Hispanics favor Spanish and Portuguese. Blacks are less likely to have a locale other than French or English.

Figure 11 shows users' self-reported political affiliation by ethnicity. Hispanics and Whites under-report their political preference and consequently appear less frequently than expected (to the left of the black line) in almost all categories. Whites are more frequent in the Libertarian, Conservative, and Very Conservative categories, while Asian/Pacific-Islanders are more frequent in the Moderate, Liberal, and Very Liberal categories. Asian/Pacific-Islanders also self-identify as Apathetic much more frequently than other ethnicities.

Figure 10: The locale of U.S. Facebook users by ethnicity. Numbers are relative to the total number of U.S. Facebook users with that locale.

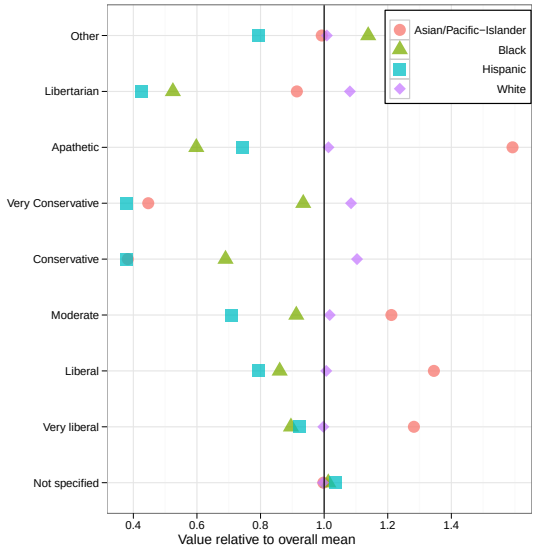


Figure 11: The political affiliation of U.S. Facebook users by ethnicity. Numbers are relative to the total number of U.S. Facebook users with that affiliation.

Figure 12 shows how users of different ethnicities use the site. Asian/Pacific-Islanders are the most engaged, with an unexpectedly high number of wall, video, note, gift, comment, and group sharing actions. Blacks and Hispanics share less than the site average; photos are a notable exception for Hispanics, while Blacks tend to update their status more often and send more private messages.

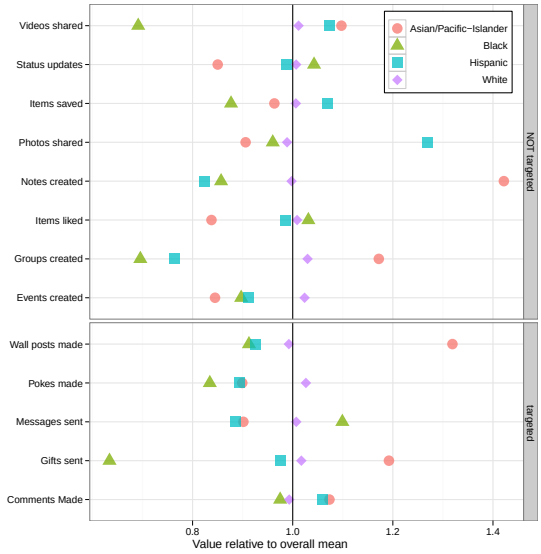


Figure 12: Usage characteristics of U.S. Facebook users by ethnicity. Numbers are relative to the total number of U.S. Facebook users with that affiliation.

Figure 13 shows the geographical distribution of different ethnicities. Each point is a spatially-binned collection of users, sized and colored by the relative proportion of users of that ethnicity. Larger/redder points denote areas where that ethnicity is more prevalent than one would expect by chance. For Blacks, highly prevalent areas are concentrated in the South, while for Hispanics, these areas are in the Southwest, California, and Florida. Asian/Pacific-Islanders, on the other hand, are more prevalent in coastal urban centers, especially the San Francisco Bay Area.

5. Discussion and Related Work

There has been much discussion on the issue of race and class in the context of the Internet, and social media in particular. While most of the early dialog focused on access (e.g., the "Digital Divide"), researchers have more recently shifted their focus to the differentiation in scope and use of the Internet for varying purposes (DiMaggio et al. 2004). Recently, some research has suggested that online social network membership is becoming increasingly assortative (Boyd 2009), and that usage of online social media is becoming differentiated by socio-economic status (Hargittai and Walejko 2008).

We propose an approach to determine the ethnic breakdown of a population based solely on people's names and data provided by the U.S. Census Bureau. We demonstrate that our approach is able to predict the ethnicities of individuals as well as the ethnicity of an entire population better than natural alternatives. We apply our technique to the population of U.S. Facebook users and discover that while Facebook has always been diverse, this diversity has increased over time leading to a population that today looks very similar to the overall U.S. population.

Caveats The observations made in this paper are based on a fairly noisy feature, namely people's names, and while the results are significant the interpretation should come with a

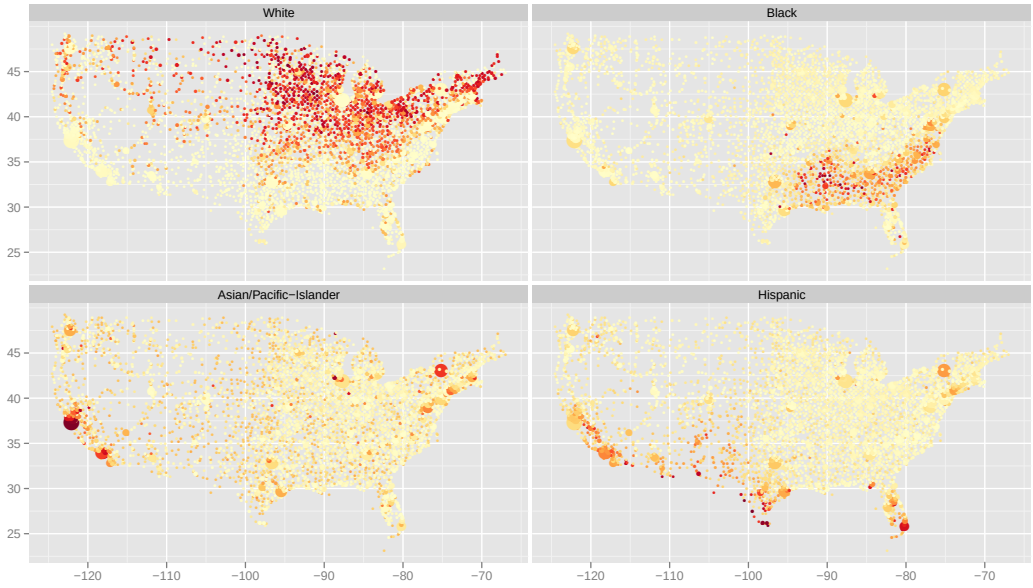


Figure 13: The geographical distribution of U.S. Facebook users by ethnicity. Each point is a spatially-binned collection of users, sized and colored by the relative proportion of users of that ethnicity.

few caveats. First, we have empirically evaluated our model in Section 3., but have not yet theoretically modeled error throughout our calculations Second, we have left out a significant portion of the population through smaller ethnic groups. These groups should be included with appropriate interpretation for a complete analysis. Finally, and most importantly, while ethnicity is an important factor in understanding user behavior, it is often only a proxy for other variables, such as socioeconomic status, or education. A complete analysis should control for all such factors to understand which inferred features have the most significant impact.

Future Work The approach taken in this paper suggests a framework for understanding user behavior in terms of demographic features determined through unsupervised modeling. A number of extensions could extend the observations made in this paper. First, the findings of assortative mixing are possibly the most important piece of behavioral analysis. While we have only presented a snapshot in time, this analysis could very easily be extended over a period of time to understand how relationships evolve as Facebook grows and friendships grow over time. Second, relationships can also be an important source of information for modeling ethnicity, and the results of Figure 8 are suggestive that this information can improve predictions dramatically. Third, the data provided by the census goes beyond surnames, and includes information about location, professions and other features disclosed by social network users. These data could be easily incorporated to improve the predictive power, as shown in Figure 13.

References

Ambekar, A.; Ward, C.; Mohammed, J.; Male, S.; and Skiena, S. 2009. Name-ethnicity classification from open sources. In *KDD*.

Asuncion, A.; Welling, M.; Smyth, P.; and Teh, Y. W. 2009. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Boyd, D. 2009. The Not-So-Hidden Politics of Class Online.

Buechley, R. W. 1976. Generally useful ethnic search system, GUESS. In *Annual Meeting of the American Names Society*.

Coldman, A. J.; Braun, T.; and Gallagher, R. P. 1988. The classification of ethnic status using name information. *Journal of Epidemiology and Community Health* 42(4):390–395.

DiMaggio, P.; Hargittai, E.; Celeste, C.; and Shafer, S. 2004. *From unequal access to differentiated use: A literature review and agenda for research on digital inequality*. New York, NY: Russell Sage Foundation. 355–400.

Erosheva, E.; Fienberg, S.; and Lafferty, J. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*.

Fiscella, K., and Fremont, A. M. 2006. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research* 41(4):1482–1500.

Hargittai, E., and Walejko, G. 2008. The participation divide: Content creation and sharing in the digital age. *Information Communication and Society* 11(2):239.

Kali, J. C.; Bethel, J.; Burke, J.; Morganstein, D.; and Westat, S. H. Using names to check accuracy of race and gender coding in naep. *ASA Section on Survey Research Methods*.

Lauderdale, D. S., and Kestenbaum, B. 2000. Asian american ethnic identification by surname. *Population Research and Policy Review* 19:283–300.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*.

Tucker, D. K. 2005. The cultural-ethnic language group technique as used in the Dictionary of American Family Names (DAFN). *Onomastica Canadiana* 87(2).