Demographic Breakdown of Twitter Users: An analysis based on names

Hüseyin Oktay
University of Massachusetts
Amherst, MA
hoktay@cs.umass.edu

Aykut Fırat
Crimson Hexagon
Boston, MA
aykut@crimsonhexagon.com

Zeynep Ertem
Texas A&M University
College Station, TX
zeynepertem@tamu.edu

ABSTRACT

We propose an approach for age estimation using solely people's first names by extending an already existing method proposed by Chang et al. for ethnicity estimation. We demonstrate that proposed method is able to predict age of a person as well as the age breakdown of an entire population better than the natural alternatives. We then apply both the age and the ethnicity method to Twitter US users and perform the largest demographic analysis of the platform to the best of our knowledge. First, we closely replicate the findings about Twitter demographics in the most recent Pew Research report suggesting that name might be a useful indicator especially for aggregate analysis. Second, we demonstrate that our approach can overcome a methodological limitation in Pew Research study by estimating breakdown for all age groups including less than 18 years old age group. Third, we discover that Twitter US users has always been diverse, though some demographic groups are over-represented and some are under-represented with respect to the general internet users. We also find strong evidence that different demographic groups both in terms of age and ethnicity have different usage patterns on the platform in terms of their following relationships, topical conversations, and the time in the day to use the platform.

I INTRODUCTION

The demographic information of a population can be useful for social scientists, marketers, and policy makers. For example, of Twitter users tweeting about politics, the demographic information of Yelp reviewers commenting about a specific restaurant, or of citizens signing an e-petition about a government bill might be of interest to the respective stake holders. However, often times data to accurately obtain such information is not available for various reasons (practical, legal or ethical).

In this paper, we focus on two specific kinds of demographic information: (1) ethnicity and (2) age. We explore inferring such demographic variables using names. For ethnicity, we use an existing model proposed by Chang et al. [1] that uses last names of people in a population cross-referenced with census data to infer ethnicity.

For age, we extend this model to explicitly account for the ordinal dependency among age values. We estimate age distribution of a given population by only looking at first names cross-referenced with baby name frequency data from Social Security Administration (in the rest of the paper, we refer as SSD, social security data). We show that explicitly accounting for the ordinal dependency among age values has a better predictive power than natural alternatives including the already proposed method mentioned above.

Finally, we apply these two methods to Twitter US user-base, a microblogging platform where demographic information is almost always missing, whereas name information is often available. We estimate age and ethnicity breakdown of Twitter US users and show that our estimates based on names not only closely match the findings of most recent Pew Research report about Twitter demographics [2] but also include proportions for < 18 years olds that are overlooked in Pew Research report. Moreover, using these models, we also answer questions about Twitter US users:

- How diverse are Twitter US users in terms of age and ethnicity and how has diversity changed over time?
- How do users from different demographic groups use Twitter platform?
- How do different topics and different Twitter users attract users from different demographic groups?

We report that Twitter has always been diverse and certain demographic groups have been over-represented

and certain demographic groups have been underrepresented compared to the general internet population reported by Pew Research study. We also discover that different demographic groups use Twitter at different times during the day. Finally, we estimate that both different topics attract different demographic groups as contributors and different Twitter users attract different demographic groups as followers.

II METHODOLOGY

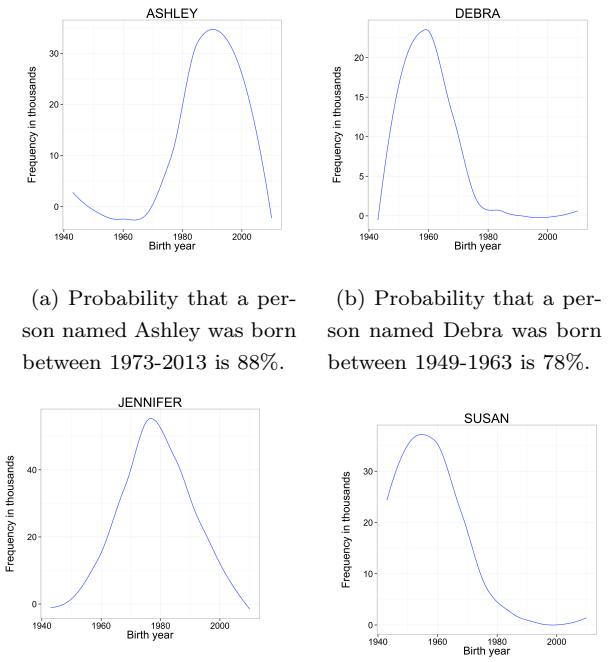
In this section, we describe two statistical models we use for ethnicity and age estimation using names. For ethnicity estimation, we use an existing model proposed by Chang et al. [1] that uses last name cross-referenced with census data. We refer the reader to that work for details about the model.

For age estimation, we use first names cross-referenced with baby name frequency data from SSD[1] and extend this earlier model to explicitly model age as an ordinal variable. Input to our model are a list of first names and parameters from the SSD (i.e., frequency of baby names in each birth year). Output from our model is population level breakdown of age categories.

The main advantage of the proposed model is that it requires no *labeled* datasets. The method uses unsupervised learning techniques to estimate age distribution for a given population. We note that we characterize such a model as *partially* supervised since parameters from the SSD can be interpreted as supervision to some extent. Since SSD provide aggregate summaries of baby name frequency statistics rather than individually labeled data instances, traditional supervised methods such as regression or multi-class classification are not directly comparable to our method.

Background The U.S. Social Security Administration releases data about the frequency of each baby name given in each year. Such data include the frequency of more than 150K different baby names for each year starting from 1881. Baby names show clear temporal trends. We show in Figure 1 several example first names and how their smoothed frequency values change over years. For example, among common names (names that belongs to at least 500K people), if a person's name is *Ashley*, then 90% of the time she is less than 40 years old. Similarly, if a person's name is *Debra*, then 78% of the time she is between 50-64 years old.

[1]http://www.ssa.gov/oact/babynames/limits.html



(a) Probability that a person named Ashley was born between 1973-2013 is 88%.

(b) Probability that a person named Debra was born between 1949-1963 is 78%.

(c) Probability that a person named Jennifer was born between 1964-1973 is 64%.

(d) Probability that a person named Susan was born between 1949-1963 is 59%.

Figure 1: Smoothed frequencies over birth years of example frequent first names show that baby names in US show a clear temporal trend.

A naive method to predict age at an individual level for a given name would be to assign it to the corresponding most frequent birth year in the SSD. To estimate aggregate density in a naive way, we can simply sum up frequencies corresponding to birth years for each name in a given population and then normalize the sum. Such a naive method performs inference independently for each user, and therefore assumes that users in the group are independent from each other. But what if the population of interest is mainly teenagers (e.g., followers of a young celebrity), then the independent population assumption might be violated.

We address this concern in the proposed Bayesian model (described in the next section) by allowing each name in a given population to inform each other name through joint inference. For example, if we have a population that is mainly teenagers, and a name in this population that has a uniform prior distribution

among almost all age groups. Then in joint inference these names inform each other. Therefore, the posterior probability of being a teenager increases for the name with a uniform prior.

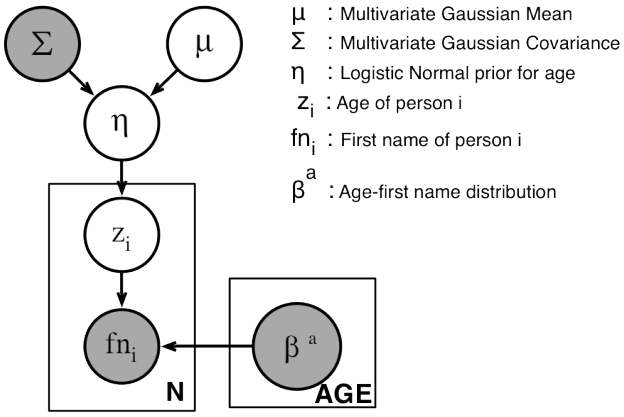


Figure 2: The proposed Bayesian graphical model to estimate density as well as predict individual age given a first name. Shaded variables indicate the inputs to the model, and unshaded variables are estimated or inferred from the model.

Model To account for joint inference, we use a generative Bayesian mixture model to estimate the age density of a given population. We model the individual age values as hidden variables, and we model the first names of corresponding individuals as observed variables. Also, parameters corresponding to first names for each age value are observed (through SSD). Then, we simply jointly infer the most likely values for hidden variables (individual age) given the observed variables (corresponding first names and corresponding frequencies in each age group).

The proposed model is closely related to mixture-models which allow populations to be a mixture of categories [1, 3]. Additionally, we explicitly account for the ordinal relations among the age values by using a logistic normal prior for age proportions with a predefined covariance matrix [4, 5] instead of using a Dirichlet prior.

Age values show an ordinal dependency that *teenagers* are smaller than *young adults* and *young adults* are smaller than *elderly*. We show the clear benefit of explicitly modeling age categories as ordinal variables with a logistic-normal prior in the experimental evaluation. Since logistic normal is not a conjugate-prior for multinomial distribution, we use variational inference.

We present the generative process of the proposed model more formally below. Let A be the number of different age categories, and N be the number of users in a given group.

1. For each age value $a \in 1...A$,
   (a) draw the distribution of first names, $\beta^f_{z_i} \sim Dirichlet(\nu)$
2. From a multivariate Gaussian with a mean μ and standard deviation σ, draw ν
3. Transform ν with the logistic function to get age proportions
4. For each person $n \in 1...N$,
   (a) Draw an age value z_i from $Multinomial(F(\nu))$
   (b) Draw the first name of the individual based on age value, $fn_i \sim Multinomial(\beta^{fn}_{z_i})$

The graphical representation of the corresponding model is shown in Figure 2. The parameters of the model are μ and Σ for multivariate Gaussian, and β_a for multinomial distributions of first names for each age category. We transform multivariate Gaussian to multinomial parameters using logistic-normal as follows:

$$F(\nu_i) = \eta_i = \frac{e^{\nu}_i}{\sum_j e^{\nu_j}}.$$

The shaded variables in Figure 2 correspond to the observed variables, and unshaded variables correspond to the hidden variables. We set $\beta^{fn}_{z_i}$ values from the SSD.

Age value x is an ordinal variable where $x < x+1 < x+2$. In a posterior probability sense, such ordinal dependency translates into having similar posterior probability values for age values that are close to each other, as shown in Figure 1. For example, if a person's name is *Ashley*, she has a high probability that she was born in 1990. By modeling the ordinal dependency among age values, we expect to also see high probabilities for values that are close to 1990 (e.g., 1991, 1989). On the other hand, a person named *Ashley* has a low probability that she was born in 1950, and therefore we expect low probability values for values close to 1950 (e.g., 1951, 1949).

To account for the ordinal dependency among the age values, we set the parameter for sigma as suggested by Agresti [5], as $\sigma_{ij} = \rho^{|i-j|}$ where $0 < \rho < 1$ and i and

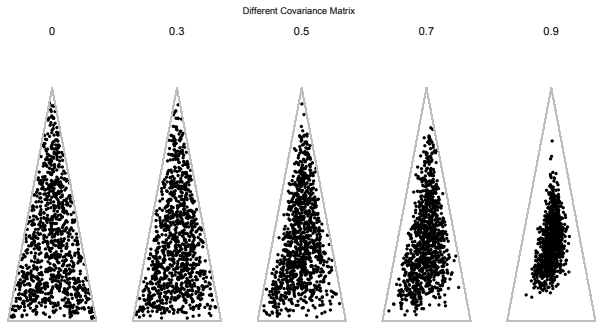


Figure 3: Possible values of class mixture proportions for a 3-valued variable as the covariance value changes for the logistic normal prior. When ρ is small, all possible combinations can be possible. Whereas as we increase ρ, the possible space is constrained. In the proposed model, this mixture proportion corresponds to the input vector for the multinomial.

j are indexes for consecutive possible values of the multinomial variable. In Figure 3 for a multinomial variable with 3 possible values, we show the possible probability vectors. In this figure, each corner represents a possible value for the multinomial, and each dot represents a probability vector with corresponding values for all possible values. The closer the dot is to the corner, the higher the probability for that particular value that corner represents. We vary the ρ value and show 1000 possible probability vectors in Figure 3. We observe that as the ρ value in the covariance matrix increases, the consecutive values in the probability vector for the multinomial variable become more and more dependent, mimicking the ordinal dependency. Therefore, we use logistic-normal prior with a fixed covariance matrix for age multinomials to capture the ordinal dependency.

Although logistic-normal seems like a better prior for modeling ordinal aspect of age, parameter estimation and inference are challenging since logistic-normal is not a conjugate prior for multinomial distribution. Therefore, we use variational methods instead of sampling based methods for parameter estimation and inference. We find close connections between the model we propose and *Correlated Topic Models* proposed by Blei and Lafferty [4]. In a topic model context, age categories in our model correspond to topics and first names in our model corresponds to words. We similarly use logistic-normal prior for multinomial as in correlated topic models. However, we explicitly encode correlation among topics as ordinal dependency among age values by setting the values for the covariance matrix instead of allowing any possible correlation. We also explicitly set topic-word distribution as the normalized age-first name frequency from SSD.

We revise the variational method for inference in the proposed method. Specifically, the main difference is that we have fewer variational and model parameters since Σ and β values are already fixed and we only estimate the remaining μ (see [4] for detail about variational parameters).

We must perform posterior inference to determine the values of the hidden variables given the observed variables. We are interested in estimating

$$p(\boldsymbol{\nu}, z_{1:N}|fn_{1:N}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta}^{fn}_{1:A}).$$

Given a model with $\beta_{1:A}, \mu, \Sigma$ and a list of first names $fn_{1:N}$, the variational algorithm optimizes the posterior with respect to the variational parameters, which are only λ and ζ since the model parameters corresponding to other variational parameters are fixed. We also use coordinate ascent, iteratively optimizing one variable while holding the other constant. For estimating μ, we use the variational expectation-maximization algorithm. In the E-step, we maximize the bound proposed in Blei and Lafferty [4] with respect to variational parameters. In the M-step, we maximize the bound with respect to model parameters. We run the variational EM algorithm until the relative change in the bound is less than 10^{-5}.

We calculate the expected value of each hidden variable under the approximate posterior and report results based on this expectation. For each first name, we get a vector π_i of length A (i.e., individual level prediction), where each entry is the probability of the person i being in that age value:

$$\boldsymbol{\pi}_i = p(z_{1:N}|fn_{1:N}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\beta}^{fn}_{1:A}).$$

We get population level statistics by calculating the column means of all the $\boldsymbol{\pi}_i$ values.

Experimental Evaluation

Data and Alternative Models We evaluate the proposed model by using voter registration data. We gather 154,016 voter registration records from Ohio in which we have the ground-truth data with first names and corresponding birth years. We compare performance of the following models:

Truth—ground truth data from voter registration.

Random—randomly assigning a person to a birth year, and then calculating the aggregate proportions.

Naive model — the model that aggregates SSD

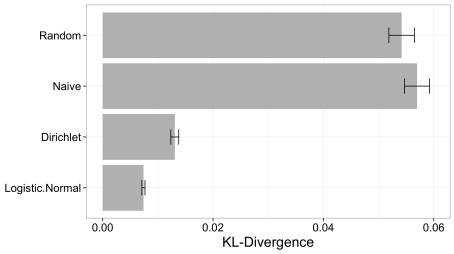


Figure 4: The model with logistic normal prior achieves the smallest kl-divergence as it both models age as an ordinal variable and performs a joint inference.

probabilities as described in Section II.

Dirichlet prior model with smoothing– a mixture model with joint inference, but that regards age categories as categorical variables (i.e,. not ordinal, the exact model proposed in Chang et al. [1]). We smooth the predictions of this model to make it more suitable for estimating an ordinal variable. (Raw predictions of this model performs the worst, but we think smoothing the predictions for this model is a fair comparison.)

Logistic normal prior model—the proposed model that explicitly accounts for the ordinal dependency among the age values.

Population-Level Evaluation We evaluate the proposed model on estimating the population-level statistics of a given population. We systematically sample populations with different mixtures of birth years from the ground truth data. We create 1000 different datasets each of which includes 10,000 first names, and we use the proposed model to estimate population-level statistics. We evaluate performance at an age category level, where we combine birth years into categories by simply summing the density values. For example, we estimate the density for the $18-29$ age category by summing all the birth year values corresponding to that category. We use the same four age categories used in Pew Research report as: $<30, 30-49, 50-64,$ *and* $65+$. For the evaluation metric, we use *kl-divergence*. The smaller the value, the better the performance.

In Figure 4 we plot average kl-divergence of predictions of each model for 1,000 test cases, and we observe that methods utilizing joint inference perform significantly better than the naive and the random model. We also report that *logistic-normal* prior has the lowest kl-divergence. These experimental results suggest that both jointly inferring posterior probability as well as explicitly modeling the ordinal dependency among age variables improves the estimation of population-level statistics.

III APPLICATION TO TWITTER

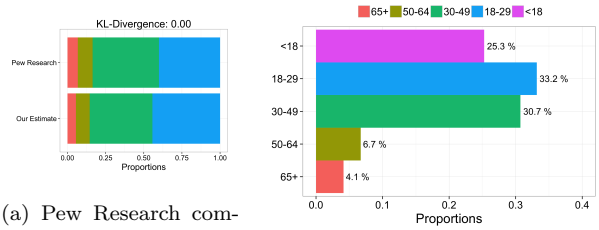
In this section, we apply the two statistical models to Twitter US user base to discover demographic information at the population level. Demographic information is almost always missing for Twitter users in their profile, however, often times their names are available.

Age Estimation In this section, we apply the proposed model described in Section II to Twitter US users. As the proposed method performs quite accurately on population level estimation, we perform population level analysis of the Twitter US user base. First, we compare our estimation with Pew Research report [2]. We then report on how age diversity of Twitter US users has changed over time with respect to the general Twitter US population, and a case study revealing age breakdown for populations identified by following a specific Twitter user and by participating in a specific conversation. Finally, we explore how assortative following relationship is among Twitter US users.

Comparison to Pew Research Report First, we calculate the age category breakdown of a Twitter user base snapshot from September 2013 (i.e., most recent report from Pew Research) and compare our findings to estimates in Pew Research Report about Twitter US demographics [2]. We define our categories as follows: *<18, 18-29, 30-49, 50-65, and 65+* as suggested in the Pew Research report. In Figure 5a, we compare our findings to the findings of Pew Research. We find that our findings very closely replicate that of Pew Research: most of the Twitter users are in age category 18-29, and users above 50 years old seem to be around 10% of all users.

We note that, the Pew Research study focused on Twitter users over 18 due to a methodological limitation (i.e., they could not perform phone surveys for users under 18). However, the proposed methodology can give estimates for any age category including under *18*. Figure 5b shows our estimate for all age categories, and we find evidence that Twitter US users under *18* might account for roughly the 25% of the existing Twitter user population.

Diversity on Twitter With the model described in Section II, we can estimate the relative age breakdown of Twitter users over time. We get a random



(a) Pew Research comparison

(b) Age breakdown for full data

Figure 5: Estimates of the proposed method are comparable to the findings in the Pew Research report. We also find that under 18 age group is the second most present age group among Twitter US users which is overlooked in Pew Research report.

sample of Twitter US users on the first day of each month starting from June 2011 until January 2014.

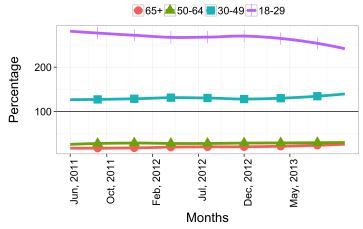


Figure 6: We compare different age groups in the US Twitter users to the general internet population as reported by the Pew Research. We observe that most present age group on Twitter has been over represented since the beginning.

In Figure 6, we compare the saturation of each age group above 18 years old to the age breakdown of the general internet users obtained from the Pew Research report [2]. In these plots, a line on 100% would indicate perfect representation of corresponding age category with respect to general internet population. A line above 100% line would indicate over-representation and a line under 100% line would indicate under-representation of that particular age group with respect to the general internet population.

We discover that Twitter US users have always been diverse. We also observe that age group 30-49 and age group 18-29 have been over-represented among Twitter US users and age group 18-29 seems to be in the decline. Conversely, age group 50-64 and 65+ have been under-represented on Twitter.

With the proposed method, we can also analyze which

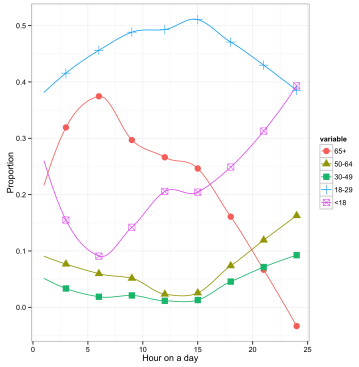
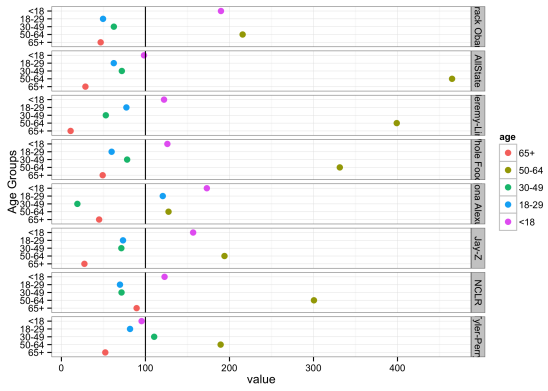


Figure 7: At what time during the day different age groups post Tweets in the US. Users over 65 and users under 18 use Twitter at completely different times during the day.

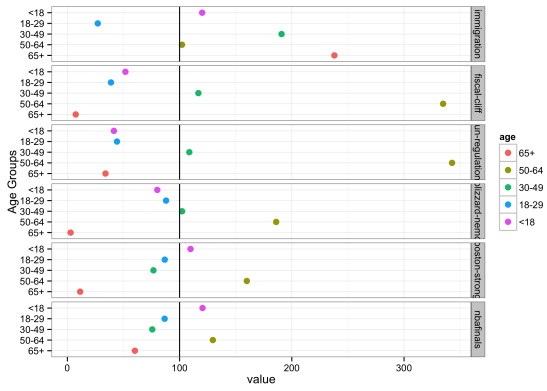
age group is actively using Twitter at what time during the day. We randomly sample 178,143 tweets posted at different hours from a one month interval. In Figure 7, we plot the daily activity of different age categories. We observe that Twitter US users under 18 and above 65 might be active on Twitter at completely different times of the day. Twitter users above 65 seems to be most active early in the morning. In contrast, Twitter users under 18 seem to be posting Tweets later in the day. For the largest age group of on Twitter, age category 18-29, Twitter usage peaks later in the afternoon.

Follower Analysis We also perform some example case studies on Twitter US users to show various potential applications of the model. First, we focus on age breakdown of followers of arguably popular Twitter US users, and we compare such breakdown figures with respect to the breakdown for general Twitter US users. By using the Twitter API, we get a 100,000 random sample of followers of *Barack Obama*, and we also get all followers of *Jay-Z*, a black singer in the US, *Jeremy Lin*, an Asian-American NBA player, *NCLR*, an NGO for supporting hispanic rights, *Tyler Perry*, a black actor in the US, *All-state*, an insurance company, *Whole Foods*, a grocery store chain in the US, *Ajiona Alexus*, a black teenager singer in the US. For each set of followers, we estimate the age breakdown and compare it with the age breakdown of general Twitter US population.

Figure 8a compares the age breakdown for each group of followers to the age breakdown of random Twitter US user sample. When an age group is well-represented with respect to the general Twitter population, the solid vertical lines intersect with 100%.



(a) Different Twitter users have different age groups as followers.



(b) Different topics have different age groups as participants.

Figure 8: We perform a case study on Twitter US users focusing on followers of a specific Twitter user and users engaging in specific conversation.

Dots to the left of this line indicate under-representation, and higher values indicate over-representation.[2] and dots to the right side of this line indicate over-representation of the corresponding age category with respect to general Twitter population.

For example, we observe that there are more than twice as many followers of *Barack Obama*, the President of the US, in the 50-64 age group as there are in the general US Twitter population. We also observe that there are more than four times many followers of *All state*, an insurance company, in the 50-64 age group as there are in the general Twitter US population. We see a similar trend for followers of *Whole Foods*, a national chain of grocery stores in the US, and also for followers of *NCLR*, an NGO organization for Hispanic rights located in Chicago.

On the other hand, for followers of *Ajiona Alexus*, a teenager celebrity, we observe that Twitter users under age 18 are over-represented compared to their corresponding proportion on Twitter. Main take away of these analysis is that different Twitter users attract users from different demographic groups as followers and this proposed model can accurately estimate such population level breakdown.

Topical Analysis Using the proposed model, we can also analyze the breakdown for different age groups engaging in conversation around specific topics on Twitter. In Figure 8b, we compare the age breakdown of users given a specific topic to the general age breakdown of Twitter US users. As explained above, 100% indicates perfect representation of each age group, smaller values indicate under-representation,

We observe that the age groups are represented differently in different conversation. For example, for the *immigration act* conversation on Twitter, measured by the relevant set of hashtags, we observe that Twitter users of age 65+ and 30-49 are over-represented (i.e., provided more posts than the expected from their proportion on Twitter). We also observe that users of age 18-29 are under-represented in the immigration act conversation on Twitter.

Users in age group 50-64 seem to be actively participating in conversations about *fiscal cliff, gun regulations, the blizzard nemo, boston-strong campaign* and also *nba-finals*. We also observe that users under age 18 are participating in conversations about *nba finals*.

Follow relationship analysis: We also analyze how different age groups interact with each other with respect to follow relationship on Twitter. We sample 100,000 random follow relationship from the Twitter graph using data from [6]. We then use our model to predict age of each user in the sample. For each relationship, we calculate an interaction matrix by simply taking the cross-product of age prediction vectors corresponding to users involving in follow relationship. We calculate an average interaction matrix by aggregating the matrices of each random edge we sample from Twitter graph. We normalize the average in-

[2]We gather data for each of these topics using the *Twitter Firehose API*. We filter tweets containing relevant keywords and hashtags. See the appendix for a detailed list of hashtags used for each topical analysis.

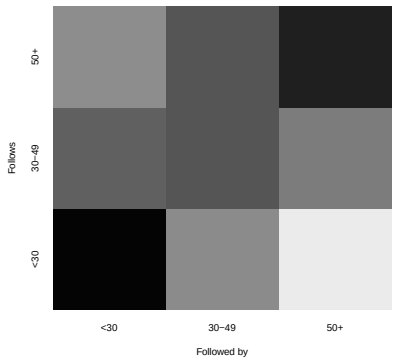


Figure 9: Following relationship by different age groups. Darker cells indicate that age group pairs follow each other (i.e., y follows x) more frequently than expected by chance. We observe that < 30 and 50+ age groups show assortative mixing in their relationship, however, $30 - 49$ age group shows a more diverse mixing.

teraction matrix, with the expected interaction we expect by chance.

Figure 9 shows the normalized interaction matrix as a heat map. In this matrix, age group in the y-axis follows age group on the x-axis. The darker the cell, the higher the interaction between corresponding pair of age groups. We observe that age groups < 30 and 50+ show an assortative mixing in their follow relationship (i.e., follow mostly users within their own age group), whereas, age group $30 - 49$ shows a more diverse mixing.

Ethnicity

Ethnicity on Twitter Overall In this section, we apply the bayesian model proposed by Chang et al. [1] to Twitter US users. First, we compare our estimation with estimation from Pew Research report [2]. We then report on how ethnicity diversity of Twitter US users has changed over time with respect to general Twitter US users, and finally a case study revealing ethnicity breakdown for populations identified by following a specific Twitter user and by participating in a specific conversation.

Comparison to Pew Research Report: First, we calculate the ethnicity breakdown for Twitter userbase snapshot from September 2013, and compare our findings with estimates in the most recent Pew Research report [2]. We find that our findings by considering the last name of users closely replicates the findings of Pew Research as shown in Figure 10.

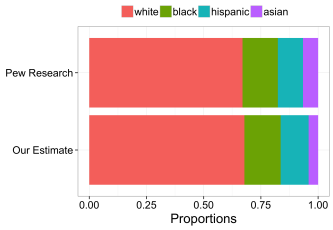


Figure 10: Pew Research comparison

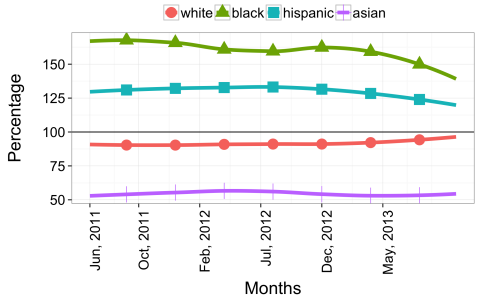


Figure 11: Ethnicity proportions on Twitter over time. Minorities are over-represented.

Diversity on Twitter: With such Bayesian model, we can estimate the ethnic breakdown among Twitter US users over time. Here we get a random sample of Twitter US users on the first day of each month between June 2011 and January 2014.

Figure 11 shows the ethnic proportions on Twitter over time with respect to ethnic breakdown of the general US internet population obtained from the Pew Research report. A line on 100% line indicates well-represented ethnicity group with respect to corresponding breakdown in the general US population; a line over 100% line indicates over-representation; and a line under 100% line indicates under representation.

We discover that Twitter US users have always been diverse in terms of ethnicity. We also report that black and hispanic ethnic groups have been over-represented among Twitter US users, inline with already existing anecdotal evidence from various blog posts. Also, we estimate that white ethnic group have been under-represented, though seems to be reaching to its fair share in general US population.

With the proposed method, we can also estimate which ethnic group is actively using Twitter at what time throughout the day. In Figure 12, we show that

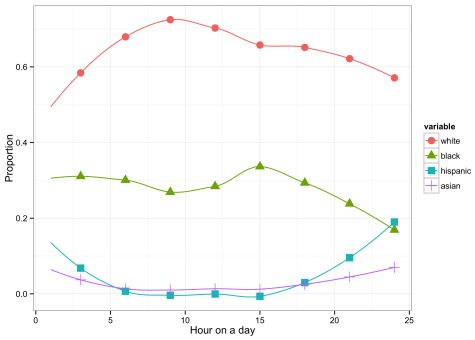


Figure 12: Usage for White Twitter users peaks during the morning whereas usage for Black users peaks during later afternoon. Usage for Hispanic and Asian users peaks around later in the evening.

usage for white users peaks early in the morning, whereas usage for black users peaks later in the afternoon. Usage for hispanic users peaks later at night.

Follower Analysis We focus on ethnic breakdown of followers of arguably popular Twitter users, and we compare such breakdown figures with general Twitter US ethnic breakdown. We focus on the same Twitter users as we reported for age estimation. Figure 13a show ethnic breakdown for followers of each Twitter with respect to general Twitter US ethnic breakdown. We discover that black users are over-represented among the followers of *Barack Obama*, whereas white Americans are slightly under-represented.

We also show that black origin users are over-represented among the followers of *Tyler Perry* (proportion of black users is 4 times more than the general Twitter users). Asian origin users are over-represented among the followers of *Jeremy Lin* and hispanic origin users are over represented among the followers of *NCLR*.

Again, we find that different Twitter users attract users from different demographic groups as followers, and by using names we can accurately estimate population ethnic breakdown.

Topical Analysis: We can also analyze the breakdown for different ethnic groups engaging in conversation around specific topics on Twitter. In Figure 13b, we compare the ethnic breakdown of Twitter users engaging in conversation about a specific topic to the general ethnic breakdown of Twitter US users. Similarly, 100% vertical line represent well-representation, smaller values represent under-representation and larger values represent over-representation.

We observe that different ethnic groups are represented differently in different conversations. For example, for the immigration topic on Twitter, we observe that contribution from asian and hispanic users to be over-represented. Interestingly, contribution to gun-regulation and fiscal-cliff conversation is over-represented by white users.

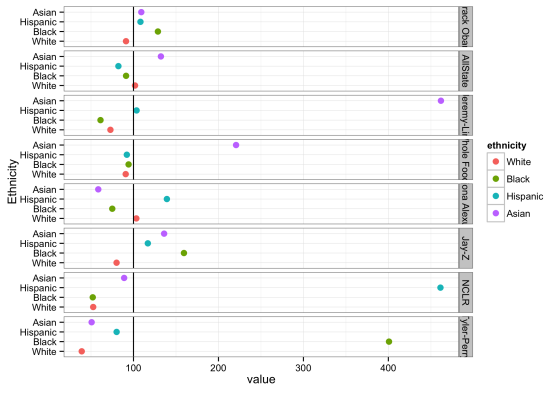
IV RELATED WORK

Many academic studies estimated the demographic information of social media and internet users. Sharad et al. [7] use web panel data including demographic information about internet users to uncover the different usage patterns among different demographic groups. Mislove et al. [8] use self reported location and name information in Twitter profile to understand demographics of users along geographic region, gender, and ethnicity. Rao et al. [9] use manually annotated data to predict demographic information about Twitter users focusing on age, gender, and regional origin. However, for age they were only able to predict two categories: below 30 and above 30. Another paper by Zamal et al. [10] focuses gender, age and political affiliation using information from neighboring users in the social graph. They collect a labeled data set by searching through tweets where users mention their birthday. Finally, Nguyen et al. [11] use a regression model and also a classification model based on text features. They train their model on manually labeled training instances to uncover different linguistic usage among different age groups.

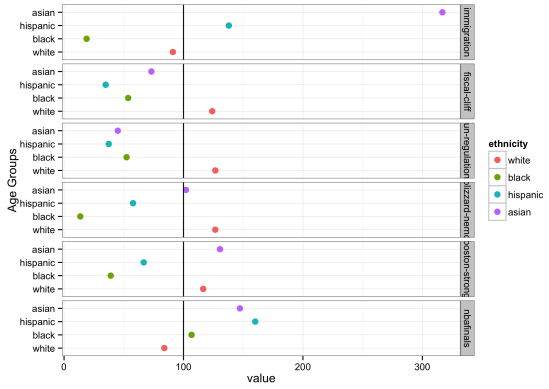
These studies use either reported age information or manually labeled training data sets to train *supervised* models such as regression or classification. In contrast, our proposed age model as well as the ethnicity model do not require labeled data.

O'Connor et al. [12] also propose a mixture model to understand demographic and linguistic variations among Twitter users using geo-tagged tweets cross-referenced with US Census statistics. However they do not include age in their approach. In a similar but a different context, Gallagher et al. [13] also use Social Security data about baby names as a prior in a mixture model to estimate the age of a person using features extracted from her image.

Related to our finding about the demographics of Twitter US users, Pew Research published a report most recently in December, 2013, reporting that users under 30 years old are the most active age group on Twitter, concluding that internet population is un-



(a) Ethnic breakdown for followers of different Twitter users are different from each other. More black origin Americans follow Tyler Perry; more asian/pacific islander Americans follow Jeremy Lin, and more hispanic origin Americans follow NCLR.



(b) Ethnic breakdown for users engaging in conversation about different topics. More asian origin Americans contribute to immigration discussion; more white Americans contribute to gun-regulations and fiscal-cliff discussions.

equally represented on Twitter. They focused on a snapshot of users in September, 2013 by performing phone surveys with users above 18 years old. Similarly, our analysis finds evidence to echo conclusions in this report, and extends the limitations of this study in two directions: First, we look at the temporal change in diversity on Twitter in the last two years. Second, we obtain estimates for all age categories, including under 18.

V CONCLUSION

In this paper, we perform the largest demographic analysis of Twitter US user-base in terms of both age and ethnicity by using names listed in user profiles. We propose a Bayesian generative model to estimating age using first name cross-referenced with Social Security data about baby names. We demonstrate that with the proposed model, we can accurately estimate the age breakdown at the population level. Since age categories are *ordinal* (i.e., not nominal), we show that explicitly modeling such ordinal dependency among age categories improves the predictive performance of the model. The proposed model makes extremely accurate population-level estimates, allowing us to answer interesting population-level questions.

We use an existing Bayesian model for ethnicity estimation using last names cross-referenced with census data. We closely replicate the results proposed in the Pew Research report about Twitter demographics suggesting that names in user profile might be a useful indicator for demographic estimation. Although, we note that our analysis is based on self-reported signal in user profile (i.e., namely name), and results should be interpreted with a grain and salt.

We also find evidence that Twitter US user base has always been diverse both in terms of age and ethnicity, and certain demographic groups have been over-represented whereas certain demographic groups are under-represented compared to the general internet user demographics in the US. Finally, we find that while some age groups are assortative in their following relationships, some age groups seem to be following users from diverse age groups.

These proposed models currently uses data specific to US which makes it only applicable to data about US users as currently implemented. However, extensions can be trivially created for other countries assuming that similar aggregate level statistics are available.

These models can be extended in multiple ways. First, we can combine these mixture models with a regular LDA to jointly infer topics that different demographic groups post about. For example, if different age groups post about different topics, than combining this model with a regular LDA can learn topics that different age groups talk in different frequency.

Second, we can use these models coupled with text created by corresponding users to implement sociolinguistic exploratory tools by compiling words from tweets to understand different linguistic usage among different demographics groups.

VI ACKNOWLEDGEMENTS

Discussions with David Arbour contributed to the paper. Helpful comments and patient editing were made by Cynthia Loiselle. Crimson Hexagon Forsight platform was very useful in getting the necessary data sets.

VII APPENDIX

Here we list the keywords we use to filter tweets related to given topic.

Immigration Act–*#cir OR #immigration OR #CIR OR #immyouth OR #DREAMact OR #cirasap OR #dwn OR #StopICE*

Fiscal Cliff–*("fiscal cliff" OR fiscalcliff)*

Gun Control–*(obama OR romney) AND "gun control"*

Blizzard Nemo–*("winter storm" OR blizzard OR Nemo OR winterstorm OR (snow AND storm) OR snowstorm OR snow*

Boston Strong–*BostonStrong OR "Boston Strong" OR OneFundBoston OR "One Fund" OR "Boston Marathon" OR BostonMarathon OR WeAreBoston OR BostonStrongest OR BelieveInBoston OR WeAreOneBoston OR PrayForBoston*

NBA Finals–*("San Antonio" OR spurs OR "tim duncan" OR "tony parker" OR miami OR "the heat" OR "dwayne wade" OR lebron OR MIA OR SA) AND (finals OR nbafinals OR win OR winning OR champion OR champions OR ring OR "win finals") AND -(eastern OR east OR west OR western OR indiana OR pacers OR "game 7" OR memphis OR grizzlies OR http OR "moving on")*

References

[1] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow, "epluribus: Ethnicity on social networks," in *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-10)*, (Washington DC), AAAI Press, May 2010.

[2] M. Duggan and J. Brenner, "The demographics of social media users," in *Posted on http://pewinternet.org/Reports/2013/Social-Media-Update.aspx*, December 2013.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), pp. 147–154, 2005.

[5] A. Agresti, *Analysis of Ordinal Categorical Data*. No. Chapter 11 in Probability and Statistics, Wiley, second ed., 2010.

[6] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *KDD*, pp. 1023–1031, 2012.

[7] S. Goel, J. M. Hofman, and M. I. Sirer, "Who does what on the web: A large-scale study of browsing behavior," in *International AAAI Conference on Weblogs and Social Media*, 2012.

[8] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the Demographics of Twitter Users," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, (Barcelona, Spain), July 2011.

[9] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, (New York, NY, USA), pp. 37–44, ACM, 2010.

[10] F. A. Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors.," in *ICWSM* (J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, eds.), The AAAI Press, 2012.

[11] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, ""how old do you think i am?": A study of language and age in twitter," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[12] B. O'Connor, J. Eisenstein, E. P. Xing, and N. A. Smith, "A mixture model of demographic lexical variation," *Proceedings of the NIPS Workshop on Machine Learning for Social Computing*, 2010.

[13] A. Gallagher and T. Chen, "Estimating age, gender and identity using first name priors," in *Proc. CVPR*, 2008.