Predicting the Demographics of Twitter Users from Website Traffic Data

Aron Culotta and **Nirmal Kumar Ravi**
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
aculotta@iit.edu, nravi@hawk.iit.edu

Jennifer Cutler
Stuart School of Business
Illinois Institute of Technology
Chicago, IL 60616
jcutler2@stuart.iit.edu

Abstract

Understanding the demographics of users of online social networks has important applications for health, marketing, and public messaging. In this paper, we predict the demographics of Twitter users based on whom they follow. Whereas most prior approaches rely on a supervised learning approach, in which individual users are labeled with demographics, we instead create a distantly labeled dataset by collecting audience measurement data for 1,500 websites (e.g., 50% of visitors to gizmodo.com are estimated to have a bachelor's degree). We then fit a regression model to predict these demographics using information about the followers of each website on Twitter. The resulting average held-out correlation is .77 across six different variables (gender, age, ethnicity, education, income, and child status). We additionally validate the model on a smaller set of Twitter users labeled individually for ethnicity and gender, finding performance that is surprisingly competitive with a fully supervised approach.

1 Introduction

Social media are increasingly being used to make inferences about the real world, with application to politics (O'Connor et al. 2010), health (Dredze 2012), and marketing (Gopinath, Thomas, and Krishnamurthi 2014). Understanding the demographic makeup of a sample of social media users is critical to further progress in this area, as it allows researchers to overcome the considerable selection bias in this uncontrolled data. Additionally, this capability will help public messaging campaigns ensure that the target demographic is being reached.

A common approach to demographic inference is supervised classification — from a training set of annotated users, a model is fit to predict user attributes from the content of their writings (Argamon et al. 2005; Schler et al. 2006; Rao et al. 2010; Pennacchiotti and Popescu 2011; Burger et al. 2011; Rao et al. 2011; Al Zamal, Liu, and Ruths 2012). This approach has a number of limitations: collecting human annotations is costly and error-prone; many demographic variables of interest cannot easily be labeled by inspecting a profile (e.g., income, education level); by restricting learning to a small set of labeled profiles, the generalizability of the classifier is limited. Additionally, most past work has focused on text as the primary source of evidence, making limited use of network evidence.

In this paper, we use regression to predict six demographic variables (gender, age, ethnicity, education, income, and child status) of a set of Twitter users based solely on whom they follow. Rather than using a standard supervised approach, we construct a distantly labeled dataset consisting of web traffic demographic data from Quantcast.com. By pairing web traffic data for a site with the followers of that site on Twitter.com, we fit a regression model between a set of Twitter users and their expected demographic profile.

With this data, we explore several questions:

RQ1. Can the demographics of a set of Twitter users be inferred from network information alone? We find across six demographic variables an average held-out correlation of .77 between the web traffic demographics of a website and that predicted by a regression model based on the site's Twitter followers. We can learn, for example, that high-income users are likely to follow The Economist and young users are likely to follow PlayStation.

RQ2. Can a regression model be extended to classify individual users? Using a hand-labeled validation set of users annotated with gender and ethnicity, we find that the regression model is competitive with a fully-supervised approach.

RQ3. How much follower information is needed for inference? We find that the identities of only 10 followed accounts per user, chosen at random, is sufficient to achieve 90% of the accuracy obtained using 200 followed accounts.

In the remainder of the paper, we will first review related work, then describe the data collected from Twitter and QuantCast and the feature representation used for the task; next, we will present regression and classification results; finally, we will conclude and outline directions for future work.[1]

[1] Code is available here: https://github.com/tapilab/aaai-2015-demographics.

2 Related Work

Predicting attributes of social media users is a growing area of interest, with recent work focusing on age (Schler et al. 2006; Rosenthal and McKeown 2011; Nguyen, Smith, and Ros 2011; Al Zamal, Liu, and Ruths 2012), sex (Rao et al. 2010; Burger et al. 2011; Liu and Ruths 2013), race/ethnicity (Pennacchiotti and Popescu 2011; Rao et al. 2011), and personality (Argamon et al. 2005; Schwartz et al. 2013). Other work predicts demographics from web browsing histories (Goel, Hofman, and Sirer 2012).

The majority of these approaches rely on hand-annotated training data, require explicit self-identification by the user, or are limited to very coarse attribute values (e.g., above or below 25-years-old).

A related lightly supervised approach includes Chang et al. (2010), who infer user-level ethnicity using name/ethnicity distributions provided by the Census; however, that approach uses evidence from first and last names, which are often not available, and thus are more appropriate for population-level estimates. Rao et al. (2011) extend this approach to also include evidence from other linguistic features to infer gender and ethnicity of Facebook users; they evaluate on the fine-grained ethnicity classes of Nigeria and use very limited training data. More recently, Mohammady and Culotta (2014) trained an ethnicity model for Twitter using county-level supervision.

There have been several studies predicting population-level statistics from social media. Eisenstein, Smith, and Xing (2011) use geolocated tweets to predict zip-code statistics of race/ethnicity, income, and other variables using Census data; Schwartz et al. (2013) similarly predict county health statistics from Twitter. However, none of this prior work attempts to predict or evaluate at the user level.

The primary methodological novelties of the present work are its use of web traffic data as a form of weak supervision and its use of follower information as the primary source of evidence. Additionally, this work considers a larger set of demographic variables than prior work, and predicts a much more fine-grained set of categories (e.g., six different age brackets instead of two or three used previously).

3 Data

3.1 Quantcast

Quantcast.com is an audience measurement company that tracks the demographics of visitors to millions of websites. This is accomplished by using cookies to track the browsing activity of a large panel of respondents (Kamerer 2013).

We sampled 1,532 websites from Quantcast and downloaded statistics for six demographic variables:

- **Gender:** Male, Female
- **Age:** 18-24, 25-34, 35-44, 45-54, 55-64, 65+
- **Income:** $0-50k, $50-100k, $100-150k, $150k+
- **Education:** No College, College, Grad School
- **Children:** Kids, No Kids
- **Ethnicity:** Caucasian, Hispanic, African American, Asian

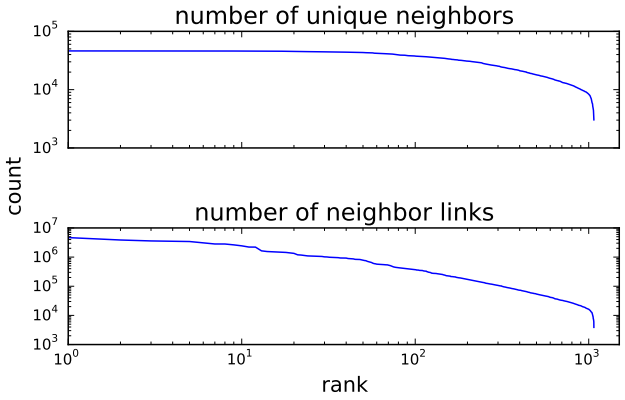

Figure 1: Rank-order frequency plots of the number of neighbors per account and the number of links to all neighbors per account.

Each variable represents the estimated percentage of visitors to a website with a given demographic.

3.2 Twitter

For each website collected in the previous step, we executed a script to search for its Twitter account, then manually verified it; 1,066 accounts from the original set of 1,532 were found. An assumption of this work is that the demographic profiles of followers of a website on Twitter are correlated with the demographic profiles of visitors to that website. While there are undoubtedly biases introduced here (e.g., Twitter users may skew younger than the web traffic panel), in aggregate these differences should have limited impact on the final model.

For each account, we queried the Twitter REST API to sample 120 of its followers, using `followers/ids` request. This sample is not necessarily uniform. The Twitter API documentation states that "At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues."

For each of these followers, we then collected up to 5,000 of the accounts they follow, called *friends*, using the `friends/ids` API request. Thus, for each of the original accounts from Quantcast, we have up to $(120 * 5K = 600K)$ additional accounts that are two hops from the original account (the friend of a follower). We refer to these discovered accounts as *neighbors* of the original Quantcast account.

Of course, many of these accounts will be duplicates, as two different followers will follow many of the same accounts (i.e., *triadic closure*) — indeed, our core assumption is that the number of such duplicates represents the strength of the similarity between the neighbors.

For each of the original accounts, we compute the fraction of its followers that are friends with each of its neighbors and store this in a *neighbor vector*. For example, suppose a Quantcast account A has two followers B and C; B follows D and E; and C follows D and F. Then the neighbor vector

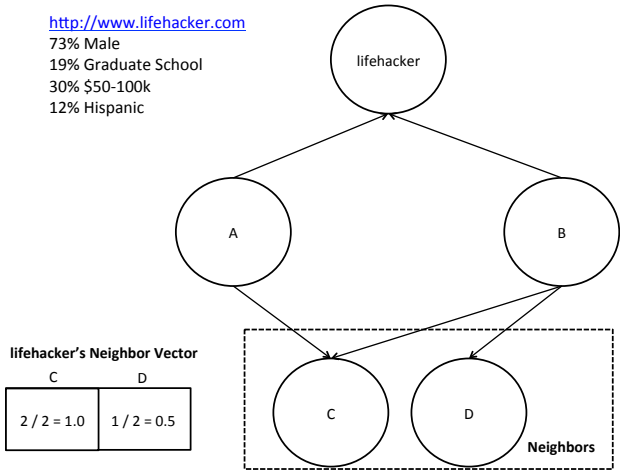

Figure 2: Data model. We collect QuantCast demographic data for each website, then construct a **Neighbor Vector** from the Twitter connections of that website, based on the proportion of the company's followers that are friends with each neighbor.

for A is $\{(D, 1), (E, .5), (F, .5)\}$. This suggests that A and D are closer neighbors than A and E. Figure 2 depicts this representation.

The resulting dataset consists of 1.7M unique neighbors of the original 1,532 accounts. To reduce dimensionality, we removed neighbors with fewer than 100 followers, leaving 46,622 unique neighbors with a total of 178M incoming links. Figure 1 plots the number of unique neighbors per account as well as the number of neighbor links per account.

4 Analysis

4.1 Regression

For each Quantcast site, we pair its demographic variables with its neighbor vector to construct a regression problem. Thus, we attempt to predict the demographic profile of the followers of a Twitter account based on the friends of those followers.

Due to the high dimensionality and small number of examples, we use elastic net regularization, which combines both L1 and L2 penalties. Furthermore, since each output variable consists of dependent categories (e.g., age brackets), we use a multi-task variant of elastic net to ensure that the same features are selected by the L1 regularizer for each category. We use the implementation of `MultiTaskElasticNet` in `scikit-learn` (Pedregosa and others 2011).[2]

Recall that standard linear regression selects coefficients β to minimize the squared error on a list of training instances $\{\mathbf{x}_i, y_i\}_{i=1}^N$, for feature vector $\mathbf{x}_i$ and expected output y_i.

[2] After tuning on a validation set for one task, we fix alpha=$1e-5$ and l1_ratio=0.5.

$$\beta^* \leftarrow \text{argmin}_\beta \sum_{i=1}^{N} (y_i - \beta^T \mathbf{x}_i)^2$$

Lasso imposes an L1 regularizer on β, while ridge regression imposes an L2 regularizer on β. Elastic net combines both penalties:

$$\beta^* \leftarrow \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T \mathbf{x}_i)^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$$

where λ_1 and λ_2 control the strength of L1 and L2 regularizers, respectively. The L1 regularizer encourages sparsity (i.e., many 0 values in β), while the L2 regularizer prevents β values from becoming too large.

Multi-task elastic net extends elastic net to groups of related regression problems (Obozinski, Taskar, and Jordan 2006). E.g., in our case, we would like to account for the fact that the regressions for "No College", "College", and "Grad School" are related; thus, we would like the sparse solutions to be similar across tasks (that is, L1 should select the same features for each task).

Let $\beta^{(j)}$ be the coefficients for task j, and let $\beta_k = (\beta_k^{(1)} \ldots \beta_k^{(M)})^T$ be the vector of coefficients formed by concatenating the coefficients for the kth feature across all M tasks. Then multi-task elastic net objective enforces that similar features are selected across tasks:

$$\beta^* \leftarrow \text{argmin}_\beta \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} (y_i^{(j)} - \beta^{(j)T} \mathbf{x}_i^{(j)})^2 + \lambda_1 \sum_{k=1}^{p} ||\beta_k||_1 + \lambda_2 ||\beta||_2^2$$

where N_j is the number of instances for task j and p is the number of features.

Regression Results We perform five-fold cross-validation and report the held-out correlation coefficient (r) between the predicted and true demographic variables. Figure 3 displays the resulting scatter plots for each of the 19 categories for 6 demographic variables.

We can see that overall the correlation is very strong: .77 on average, ranging from .55 for the 35-44 age bracket to .89 for Male and African American. All of these correlation coefficients are significant using a two-tailed t-test ($p < 0.01$), with a Bonferroni adjustment for the 19 comparisons. These results indicate that the neighbor vector provides a reliable signal of the demographics of a group of Twitter users.

To further examine these results, Table 1 displays the accounts with the 5 largest coefficients per class according to the fit regression model. These contain many results that match common stereotypes: sports accounts are correlated with men, video game accounts are correlated with younger people, financial news accounts are correlated with greater income, and parenting magazines are correlated with people who have children. There also appear to be some geographic effects, as California-related accounts are highly weighted for both Hispanic and Asian categories. There seems to

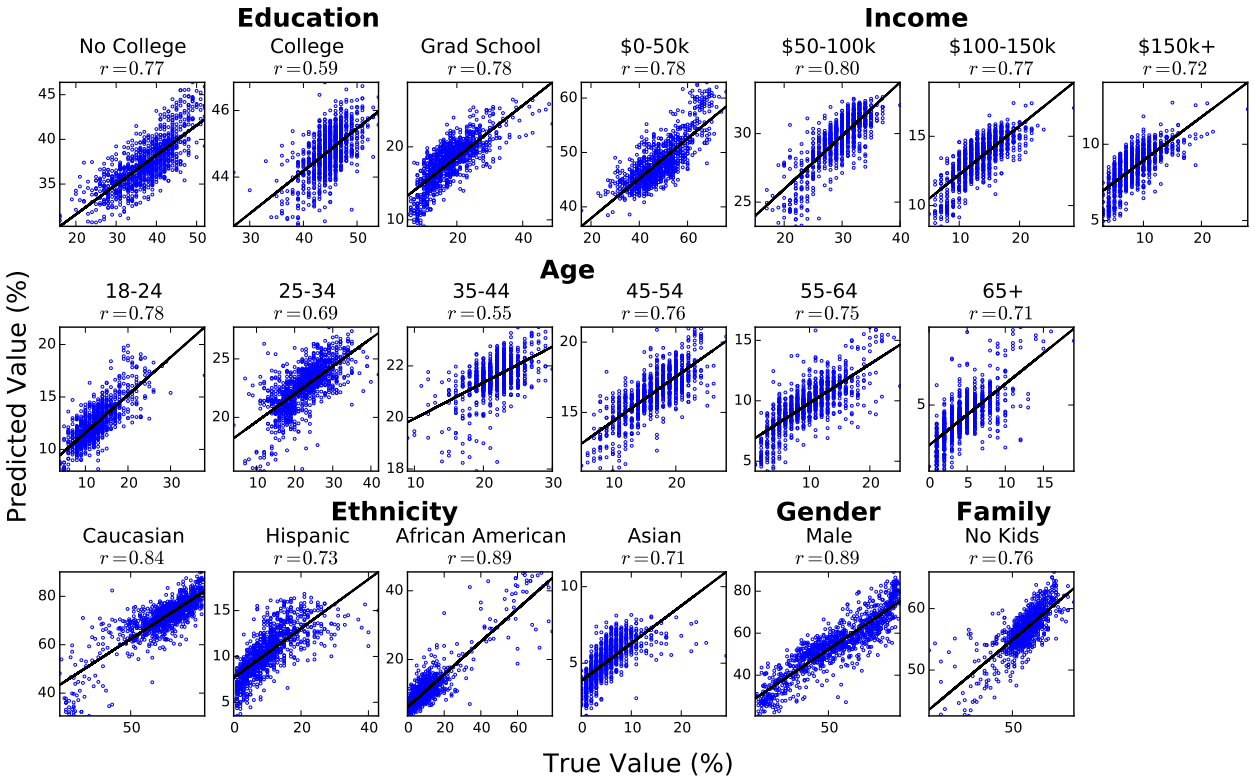

Figure 3: Scatter plots of the true demographic variables from Quantcast versus those predicted from the neighbor vector, along with the held-out correlation coefficient (r).

Model	Average Correlation
multi-task elastic net	0.772
elastic net	0.769
ridge	0.671

Table 2: Average held-out correlation across all demographic variables for three competing regression models.

be good city-level resolution — Los Angeles accounts (*latimes*, *Lakers*) are more strongly correlated with Hispanics, whereas San Francisco accounts (*SFGate*, *SFist*, *SFWeekly*) are more strongly correlated with Asians.

Finally, we compare multi-task elastic net with the single-task variant of elastic net and ridge regression (with regularization parameters tuned as before). Table 2 shows a large improvement from ridge to elastic net, and a more modest improvement when using the multi-task formulation.

4.2 Classification

The regression results suggest that the neighbor vector of an account is highly predictive of the demographics of its followers. In this section, we provide additional validation using data manually annotated at the user level.

Many of the demographic variables are difficult to label at the individual level — e.g., income or education level is rarely explicitly mentioned in either a profile or tweet. Indeed, an advantage of the approach here is that aggregate statistics are more readily available for many demographics of interest that are difficult to label at the individual level. For validation purposes, we focus on two variables that can fairly reliably be labeled for individuals: ethnicity and gender.

These were collected as follows: First, we used the Twitter Streaming API to obtain a random sample of users, filtered to the United States (using time zone and the place country code from the profile). From six days' worth of data (December 6-12, 2013), we sampled 1,000 profiles at random and categorized them by analyzing the profile, tweets, and profile image for each user. We categorized 770 Twitter profiles into one of four ethnicities (Asian, African American, Hispanic, Caucasian). Those for which ethnicity could not be determined were discarded (230/1,000; 23%).[3] The category frequency is Asian (22), African American (263), Hispanic (158), Caucasian (327). To estimate inter-annotator agreement, a second annotator sampled and categorized 120 users. Among users for which both annotators selected one of the four categories, 74/76 labels agreed (97%). There was some disagreement over when the category could be determined: for 21/120 labels (17.5%), one annotator indicated the category could not be determined, while the other se-

[3]This introduces some bias towards accounts with identifiable ethnicity; we leave an investigation of this for future work.

Category	Value	Top Accounts
Gender	Male	AdamSchefter, SportsCenter, espn, WIRED, mortreport
	Female	TheEllenShow, Oprah, MarthaStewart, Pinterest, FoodNetwork
Age	18-24	PlayStation, IGN, RockstarGames, Ubisoft, steam_games
	25-34	azizansari, louisck, lenadunham, mindykaling, WIRED
	35-44	TMZ, Oprah, BarackObama, andersoncooper, cnnbrk
	45-54	cnnbrk, FoxNews, AP, CNN, ABC
	55-64	FoxNews, cnnbrk, AP, WSJ, WhiteHouse
	65+	FoxNews, cnnbrk, WSJ, AP, DRUDGE_REPORT
Income	$0-50k	YouTube, PlayStation, IGN, RockstarGames, KevinHart4real
	$50-100k	cnnbrk, espn, SportsCenter, AP, WSJ
	$100-150k	WSJ, TheEconomist, nytimes, washingtonpost, Forbes
	$150k+	WSJ, TheEconomist, nytimes, Forbes, BloombergNews
Education	No College	YouTube, PlayStation, RockstarGames, katyperry, KevinHart4real
	College	ConanOBrien, louisck, danieltosh, azizansari, WIRED
	Grad School	NewYorker, nytimes, TheEconomist, WSJ, washingtonpost
Children	No Kids	NewYorker, StephenAtHome, nytimes, TheEconomist, WIRED
	Has Kids	parentsmagazine, parenting, TheEllenShow, thepioneerwoman, HuffPostParents
Ethnicity	Caucasian	FoxNews, jimmyfallon, TheEllenShow, blakeshelton, cnnbrk
	Hispanic	latimes, Lakers, SFGate, kobebryant, SFist
	African American	KevinHart4real, Drake, iamdiddy, Tip, kendricklamar
	Asian	SFGate, SFist, TechCrunch, WIRED, SFWeekly

Table 1: Accounts with the highest estimated coefficients for each category.

lected a category. Gender annotation was done automatically by comparing the first name provided in the user profile with the U.S. Census list of names by gender.[4] Ambiguous names were removed.

For each user, we collected up to 200 of their friends using the Twitter API. We removed accounts that restricted access to friend information; we also removed the Asian users due to the small sample size, leaving a total of 615 users. For classification, each user is represented by the identity of their friends (up to 200). Only those friend accounts contained in the 46,622 accounts used for the regression experiments were retained. Figure 4 shows the number of friends per user for each dataset.

As a baseline, we trained a logistic regression classifier with L2 regularization, using a binary representation of each user's friends. To repurpose the regression model to perform classification, we must modify the coefficients returned by regression. We first compute the z-score of each coefficient with respect to the other coefficients for that category value. E.g., all coefficients for the *Male* class are adjusted to have mean 0 and unit variance. This makes the coefficients comparable across labels. Furthermore, we set to 0 any negative coefficient. To classify each user, we then compute the sum of coefficients for each friend, and select the class with maximum value.

Classification Results Figure 5 displays the macro-F1 value for ethnicity (three classes) and gender (two classes). The regression model is fit using only the Quantcast data, while the classification model uses three-fold cross-validation using the labeled user accounts. We compare the

[4]http://www.census.gov/genealogy/www/freqnames.html

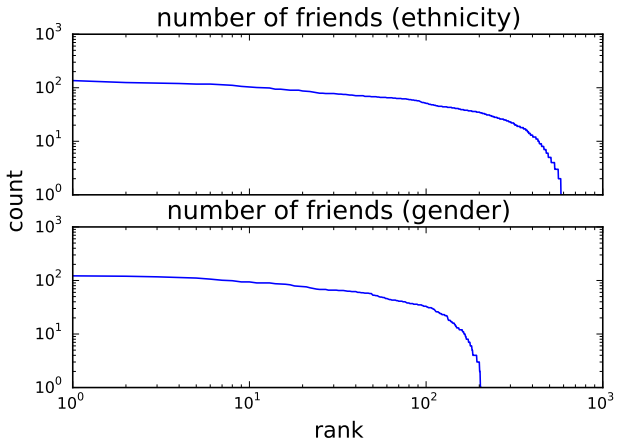

Figure 4: Rank-order frequency plots of the number of friends per user in each of the labeled datasets (ethnicity, gender). These friends are restricted to one of the 46,622 accounts used in the regression experiments.

regression approach with logistic regression using an increasingly larger number of labeled examples.

For gender, the regression model outperforms the classification approach, which is surprising given that the regression model does not have any hand-labeled profiles for training. For ethnicity, the regression approach outperforms classification until over half of the labeled data is used to fit the classification approach, after which the classification approach dominates. In general, the accuracy of the two approaches is comparable.

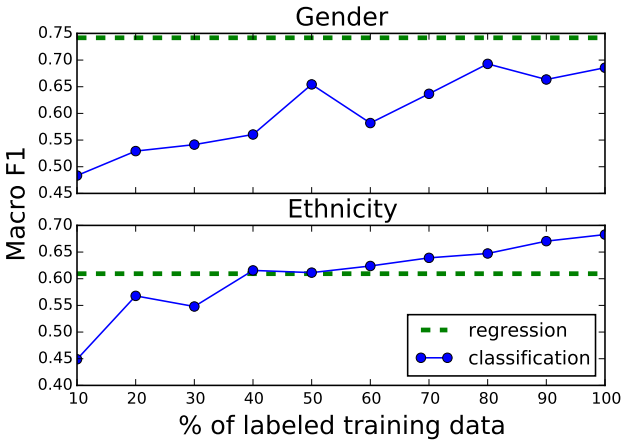

Figure 5: Classification results comparing a standard logistic regression classifier (`classification`), trained using cross-validation, versus the proposed approach (`regression`), which is fit solely on statistics from Quantcast, with no individually labeled data.

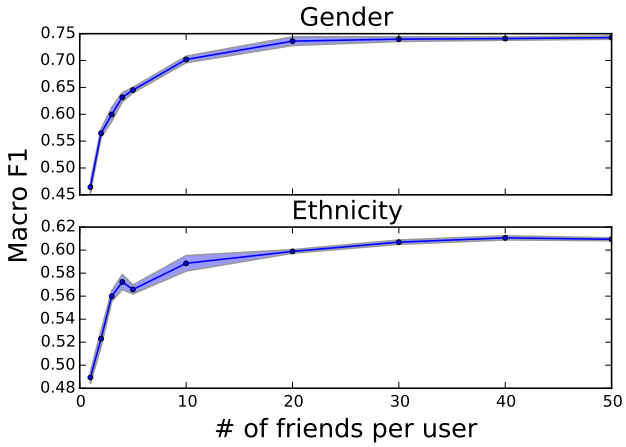

Figure 6: Macro F1 as the number of friends per user increases (with standard errors).

Sensitivity to number of friends Finally, we investigate how much information we need about a user before we can make an accurate prediction of their demographics. Whereas the previous results considered up to 200 friends of each user, we consider smaller numbers of friends to determine how the number of friends collected affects accuracy. Figure 6 displays the macro-F1 value for trials in which the number of friends per user is one of $\{1, 2, 3, 4, 5, 10, 20, 30, 40, 50\}$ (values greater than 50 did not significantly increase accuracy). The friends are sampled at random, and the results are averaged over five trials. We can see that accuracy plateaus quickly: for both tasks, the F1 score using only 10 friends is within 5% of the score using all 200 friends.

This result has implications for scalability — Twitter API rate limits make it difficult to collect the complete social graph for a set of users. Additionally, this has important privacy implications; revealing even a small amount of social information also reveals a considerable amount of demographic information. Twitter users concerned about privacy may wish to disable the setting that makes friend identity information public.

5 Conclusion

In this paper, we have shown that Twitter follower information provides a strong source of information for performing demographic inference. Furthermore, pairing web traffic demographic data with Twitter data provides a simple and effective way to train a demographic inference model without any annotation of individual profiles. We have validated the approach both in aggregate (by comparing with Quantcast data) and at the individual level (by comparing with hand-labeled annotations), finding high accuracy in both cases. Somewhat surprisingly, the approach outperforms a fully-supervised approach for gender classification, and is competitive for ethnicity classification.

In the future, we will test the generalizability of this approach to new groups of Twitter users. For example, we can collect users by city or county and compare the predictions with the Census demographics from that geographic location. Additionally, we will investigate ways to combine labeled and unlabeled data using semi-supervised learning (Quadrianto et al. 2009; Ganchev et al. 2010; Mann and McCallum 2010).

References

Al Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Argamon, S.; Dhawle, S.; Koppel, M.; and Pennebaker, J. W. 2005. Lexical predictors of personality type. In *In proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 1301–1309. Stroudsburg, PA, USA: Association for Computational Linguistics.

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*.

Dredze, M. 2012. How social media will change public health. *IEEE Intelligent Systems* 27(4):81–84.

Eisenstein, J.; Smith, N. A.; and Xing, E. P. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 1365–1374. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ganchev, K.; Graca, J.; Gillenwater, J.; and Taskar, B. 2010. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.* 11:2001–2049.

Goel, S.; Hofman, J. M.; and Sirer, M. I. 2012. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*.

Gopinath, S.; Thomas, J. S.; and Krishnamurthi, L. 2014. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Science*. Published online in Articles in Advance 10 Jan 2014.

Kamerer, D. 2013. Estimating online audiences: Understanding the limitations of competitive intelligence services. *First Monday* 18(5).

Liu, W., and Ruths, D. 2013. What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium on Analyzing Microtext*.

Mann, G. S., and McCallum, A. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J. Mach. Learn. Res.* 11:955–984.

Mohammady, E., and Culotta, A. 2014. Using county demographics to infer attributes of twitter users. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

Nguyen, D.; Smith, N. A.; and Ros, C. P. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Scie nces, and Humanities*, LaTeCH '11, 115–123. Stroudsburg, PA, USA: Association for Computational Linguistics.

Obozinski, G.; Taskar, B.; and Jordan, M. 2006. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11:122–129.

Pedregosa, F., et al. 2011. Scikit-learn: Machine learning in Python. *Machine Learning Research* 12:2825–2830.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. In Adamic, L. A.; Baeza-Yates, R. A.; and Counts, S., eds., *ICWSM*. The AAAI Press.

Quadrianto, N.; Smola, A. J.; Caetano, T. S.; and Le, Q. V. 2009. Estimating labels from label proportions. *J. Mach. Learn. Res.* 10:2349–2374.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, 37–44. New York, NY, USA: ACM.

Rao, D.; Paul, M. J.; Fink, C.; Yarowsky, D.; Oates, T.; and Coppersmith, G. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM*.

Rosenthal, S., and McKeown, K. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 763–772. Stroudsburg, PA, USA: Association for Computational Linguistics.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. W. 2006. Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 06–03.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*.