We use a corpus-wide time window for all demographic analyses (7/1/2009 - 2/13/2012). We also calculate a $usage$ parameter for each geographic bin L, defined as the sum of all ads A associated with that location at time window t_i. This includes ads from the set of all monitored categories C (both commercial and MSM/non-MSM personal ads) and is used to weight regressions across geographic locations by measuring the degree to which the local community uses Craigslist services.

$$usage(L, t_i) = \sum_{category \in C} \sum_{tag \in L} w(\{A_{category,tag,t_i}\}) \quad (3)$$

These functions provide the input for all our analyses, which use a weighted, log-log transformed, ordinary least squares (OLS) regression to compare the relationship between disclosed race/ethnicity and age in Craigslist ads and the underlying population. For the census regressions, each county forms an observation, weighted by that location's usage score. The percentage of Craigslist ads for each race/ethnicity or age group is the dependent variable and the percentage of that same group in the 2010 census data is the independent variable. All statistical analyses were done using R version 2.14.1 [33].

IV. Results

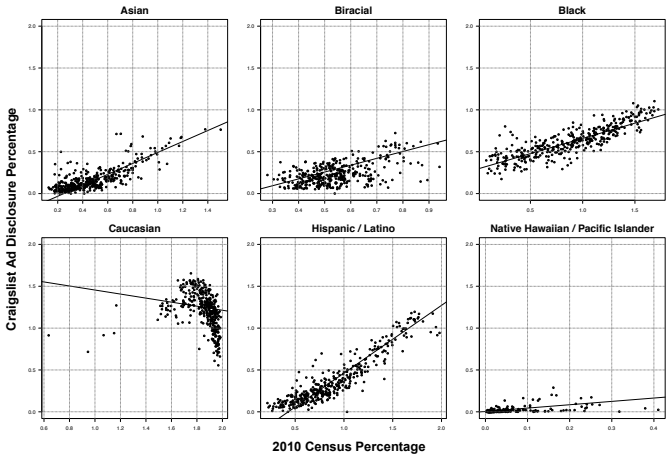

Fig. 3. Scatter plot of log-log regression results for 2010 census race/ethnicity percentages (x-axis, independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (y-axis, dependent variable). Plots are for all CBSA metropolitan geographic boundaries containing at least 1000 ads. For the Caucasian plot (lower left corner) the percentage of race disclosures begins to drop precipitously in the interval 64%-100% (1.8-1.2), suggesting that as the population grows more homogeneously Caucasian there is less need to mention race in ads.

A. Author Race/Ethnicity

In the GOLD corpus, the CRF labeling classifier for the first stage of *CRF-All* and *CRF-First* had the following per-label category F_1 scores: AUTHOR 0.86 (SD 0.01); PARTNER 0.78 (SD 0.02); and NONE 0.99 (SD 0.0). Detailed performance measures of the final output of *CRF-First* and *CRF-All* compared to the baseline *First-Mention* algorithm are found in Table II. *CRF-All* performed best overall, with statistically significant improvements between 2.2% to 156% in F_1 score over the baseline ($p < 0.05$ using a two-sided t-test) and scored the highest precision values for every category except Biracial and Undisclosed. *CRF-All* performed best at identifying Caucasian and Asian ad authors, with F_1 scores of 0.92 and 0.91 respectively. Biracial and Hawaiian/Pacific-Islander classes were the worst performing category overall, with F_1 scores of 0.63 and 0.50.

Table III includes counts and weighted OLS results comparing the race/ethnicity distributions for county, micropolitan, and metropolitan areas in CRAIGSLIST vs. 2010 census data. Most ads, 71%, did not disclose race/ethnicity information. Caucasian formed the majority-identified category with 18.5%, followed by Black 3.7%, Hispanic/Latino 3.6%, Biracial 1.4%, Asian 1.4%, and Hawaiian/Pacific-Islander 0.1%. All reported census regressions were statistically significant at $p < 0.05$. Figure 3 shows scatter plots for the metropolitan CBSA component of this analysis. Only Caucasian had a negative coefficient value, with disclosure rates decreasing as the percentage of Caucasians increased in a given geographic boundary. Rate of non-disclosed race/ethnicity is examined more closely in Figure 4, which shows a scatter plot of 2010 Census geographic Shannon entropy compared to the percentage of ads with undisclosed race/ethnicity. Shannon entropy is a measure of type diversity; it increases as members are more evenly distributed across categories (i.e., evenness) [34]. Note how as a population grows more homogenous and less evenly distributed across types (i.e., lower entropy), the rate of undisclosed race/ethnicity ads increases.

Overall, almost all categories were significantly correlated with known subpopulation makeup, with metropolitan areas tending to have the highest R^2 values. Metropolitan Hispanic/Latino ads were the most correlated with an R^2=0.91 (coefficient 0.81, 95% CI [0.78, 0.84]), followed by Asian, Black, Biracial and Hawaiian/Pacific-Islander ads. The least correlated were Caucasian and Undisclosed (using Caucasian census values) ads. Comparing the CRAIGSLIST analysis with PHONE, we find regression coefficients and R^2 values are very similar in both ad sets and statistically significant in all classes.

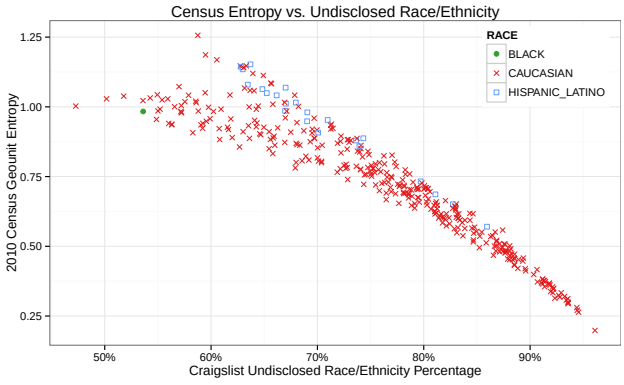

Fig. 4. Scatter plot of 2010 Census geographic Shannon entropy (y-axis), measured across race/ethnicity categories vs. the percentage of ads with undisclosed race/ethnicity (x-axis). The shape of each point indicates the majority race/ethnic group in that geographic boundary. Note how as a population grows more homogenous and less evenly distributed across types (i.e., lower entropy), the rate of undisclosed race/ethnicity ads increases.