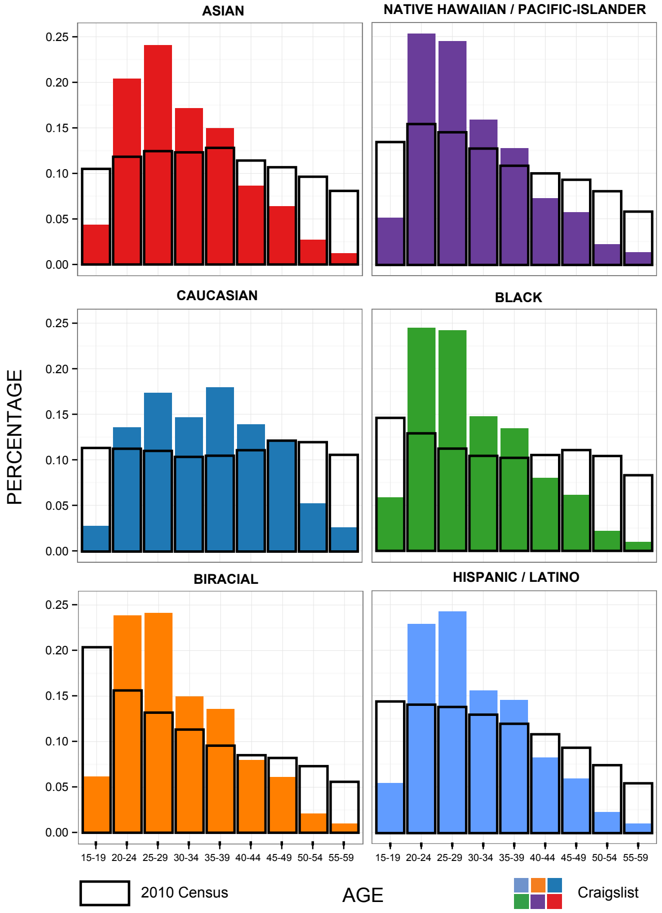

Fig. 5. Normalized histogram of aggregate, self-reported age by race for the CRAIGSLIST corpus. The black outline bars represent the 2010 census distribution of age groups vs. the solid colored bars, representing Craigslist age distributions. Caucasian ads trend older overall compared to other racial/ethnic groups, with ages 40-69 comprising 31% of Caucasian ads compared to 15%-17% for other racial/ethnic groups.

found in the 20-24, 25-29, and 45-49 age groups. The strongest correlation is found in the 20-24 age group; R^2=0.64, coefficient 0.88, 95% CI [0.81, 0.95]. Micropolitan and county boundaries were only weakly correlated. Age regression results for PHONE in metropolitan areas were also statistically significant, but with larger differences in coefficient values and lower overall R^2 values.

Figure 5 shows aggregated normalized distributions of ads disclosing age and race/ethnicity. Using Pearsons chi-square test for independence, we found that a significant relationship exists between age and reported race in Craigslist ads ($\chi^2 = 407,334$, $df = 90$, $p < 0.01$). Caucasian ads trend older overall, compared to other racial/ethnic groups. 28% of all Caucasian ads disclose ages between 15-29 while other groups range from 40% to 44%. The percentage of ads in the 30-39 range are similar across all groups (25%-29%). Ages 40-69 comprised 31% of Caucasian ads compared to 15%-17% for other racial/ethnic groups.

C. Maps

We generate several maps to illustrate the spatial distribution of MSM activity and demographic attributes in the Los Angeles / Orange County area of California, using ads from the *losangeles*, *orangecounty*, *santabarbara*, and *santamaria* sites (n=1,954,679). Since ad toponyms capture varying degrees of spatial resolution, all toponyms are normalized by binning ads into either a 1 sq. mile cell or their parent Zip Code Tabulation Area (ZCTA) geographic boundary (depending on the map), weighted by percentage of geographic overlap. Cell maps are smoothed using a Gaussian filter with σ=0.5. Figure 6 shows MSM activity percentages; this measure normalizes MSM activity rates across a community's use of Craigslist for non-sexual purposes. Red areas reflect regions where the majority of Craigslist ads in our corpus are MSM-related. Figure 7 shows the differences in spatial distributions of the 18-29 and 30-44 age groups. Figure 8 shows a ZCTA choropleth map of Hispanic/Latino authored ads. The distribution of authors within these Craigslist clusters visibly corresponds with the underlying population distribution of Hispanic/Latino individuals in census data. Several clusters appear more concentrated on the edges of population regions, with lower author race/ethnicity rates reported in the cluster center or core; other regions had a more direct correspondence between disclosures and population density. In Los Angeles similar clustering behavior was observed in Caucasian, Black, and Asian populations.

V. DISCUSSION

Our *CRF-All* method performs well at extracting race/ethnicity information in ads, providing a 2.2 - 156% improvement in mean F_1 score over a simple heuristic baseline method. *CRF-All* performs poorly only in classifying Hawaiian/Pacific-Islanders due to low sample size and in identifying Biracial ads, largely because of terminology overlap with other classes (e.g., "I am a hispanic/white male."), particularly Hispanic/Latino ads. Age is a relatively simple variable to extract from ads, since it is included as a numeric metadata tag in virtually every personal ad.

Overall we found the percentage of race/ethnicity and age disclosures in ads do reflect the population makeup of the locations provided in location tags at the county and CBSA level. Exploration of ZCTA-level spatial binning in well-populated cities like Los Angeles suggest there are opportunities for even higher spatial resolution in some circumstances. While Caucasian authorship initially seems uncorrelated with the underlying population, the absence of strong correlations more likely reflects the fact that as a population grows more homogenous, there is less need to explicitly mention race in ads. This is also suggested by the fact that the percentage of Undisclosed ads increase as the population becomes more Caucasian.

Using *CRF-All*, we show that the majority of the MSM population using Craigslist that disclose race are Caucasian, with most authors being in the age range of 20-49. Caucasian authors tend be older than other racial and ethnic groups using Craigslist, with 31% of all Caucasian ads reporting ages between 40-69 years old compared to 15%-17% for other racial/ethnic groups. This difference likely reflects the intrinsic difference in distribution of Caucasian individuals across geographic regions, as well as the more uniform (rectangular) age distribution in the Caucasian population compared to minority groups like Hispanic/Latinos, which trend younger