Mining the Demographics of Craigslist Casual Sex Ads to Inform Public Health Policy

Jason A. Fries
Department of Computer Science
University of Iowa
Iowa City, Iowa 52242
Email: jason-fries@uiowa.edu

Philip M. Polgreen, MD
Carver College of Medicine
College of Public Health
University of Iowa
Iowa City, Iowa 52242
Email: philip-polgreen@uiowa.edu

Alberto M. Segre
Department of Computer Science
University of Iowa
Iowa City, Iowa 52242
Email: alberto-segre@uiowa.edu

***Abstract*—Anonymous sexual encounters negotiated via the Internet present many challenges to public health officials addressing outbreaks of sexually transmitted infections. The anonymity and potential geographic scale of encounters weaken traditional tools like contact tracing and partner notification. These developments complicate interventions within the men who have sex with men (MSM) population, which has seen increasing health disparities in HIV and syphilis incidence rates over the last decade.**

This paper presents text-mining methods for conducting public health surveillance of the anonymous MSM populations using the online classified advertisement website Craigslist to negotiate casual sexual encounters. We analyze 2.5 years of Craigslist data (134 million ads) and present machine learning and rule-based approaches for efficiently mining race/ethnicity and age information from Craigslist text. Using previous work in geographic entity recognition, we link ads with specific locations and generate Craigslist MSM summary statistics for race/ethnicity and age cohorts in urban and rural geographic areas. This data is then compared to demographic information from the 2010 U.S. census to quantify how well it reflects the known, underlying population. We find significant correlations between Craigslist and census population statistics, suggesting our approach's utility for surveillance applications.

Keywords*—*Knowledge discovery, Natural language processing, Public healthcare, Supervised learning, Text mining.

I. Introduction

Anonymous sexual encounters negotiated via the Internet present many challenges to public health officials addressing outbreaks of sexually transmitted infections. The anonymity and potential geographic scale of encounters weaken traditional tools like contact tracing and partner notification. These developments complicate interventions within the *men who have sex with men* (MSM) population, which compared to the general population has seen widening health disparities over the last decade. MSM individuals comprise 75% of all reported primary and secondary syphilis cases as well as 63% of all new HIV infections in the U.S., disproportionately affecting young African American men [1]. These increases coincided with a decrease in safer sex practices [2]–[4].

Bull et al. found via a national survey that 43.3% of MSM (and 56.4% non-MSM individuals) negotiating sexual encounters via the Internet had traveled 100 or more miles to meet their partner [5]. A meta-analysis estimates that 40% of MSM meet their sex partners online [6] and the Internet has been identified as the largest venue where MSM meet sexual partners [7]. MSM who use the internet to seek sexual partners have reported higher rates of methamphetamine use and more sexual partners within the previous 6 months than those seeking partners found through offline means [8].

The online nature of these encounters suggests possible text mining applications for conducting public health surveillance. One such candidate is the online classified advertisement website Craigslist, the 10th most visited website in the U.S. as of September 2013 [9], which features a large community of MSM individuals. Since online classified ads are anonymous, authors frequently describe themselves in unstructured text, providing descriptive information about their sexual-health preferences as well as demographic details like race/ethnicity and age. Authors make safe or unsafe sexual encounter requests (*barebacking*), announce preferences for illegal drug use during encounters (e.g., methamphetamines, amyl nitrate), or reveal their HIV status and serosorting preferences. Moreover, all of this information is both timestamped and associated with a geographic location of often high spatial resolution.

These behavioral and demographic details speak, in part, to the types of questions asked by public health departments in partner notification surveys. Because ads are publicly visible and can be linked to specific locations, we can characterize geographic regions by the set of all MSM personal ads posted there and learn natural socio-geographic boundaries, similar to research using geocoded information on Twitter to characterize urban areas [10]. The ability to conduct public health surveillance in an automated fashion, efficiently collecting large-scale, location-specific demographic data about the anonymous populations using Craigslist would be useful to the public health community. Structural factors like race/ethnicity, age, gender, societal attitudes, etc. are identified as key features in creating and sustaining vulnerable populations, suggesting that such factors should be incorporated into the design of public health interventions [11], [12].

This paper examines the online classified advertisement website Craigslist, analyzing 2.5 years of data or 134 million ads. We present machine learning and rule-based approaches for efficiently mining age and race information from Craigslist text. Our race/ethnicity classifier reports combined F_1 scores (the weighted mean of precision and recall) from 0.63-0.93

for identifying author race/ethnicity mentions in text and 0.97 for extracting author age. Using previous work in geographic entity recognition, we link ads with specific locations and generate Craigslist MSM summary statistics for race/ethnicity and age cohorts in urban and rural geographic areas. These data are then compared to demographic information from the 2010 U.S. census to quantify how well this data reflects the known, underlying population. We found significant correlations between Craigslist and census population statistics, suggesting our approach's utility for surveillance applications.

II. Background

STI epidemiology utilizes the concept of core groups to define both the individuals engaging in high-risk behaviors (e.g., repeated sexual encounters, repeat infections, commercial sex work, etc.) and the geographic clustering of outbreaks associated with these groups. Models suggest that core groups are critical to maintaining transmission of disease within a population [13]. Unfortunately, the individuals comprising these groups are difficult to identify and locate, in part because membership within the core varies over time and because features that define a core group are not always easily observable. However, the spatial component or core area does appear to remain stable over the course of an outbreak, leading to intervention strategies that target *risk spaces* – the geographic locations linked to sexual encounters [14]. This suggests that knowledge of where individuals meet for sex and how those meeting places may change over time can provide useful information in designing targeted, geographically based interventions.

The link between Craigslist and *sexually transmitted infections* (STIs) has been explored by a number of researchers, with some research suggesting that the entry of Craigslist into local advertising markets can itself be linked to an increase in HIV/AIDS rates in the U.S. [15]. Among interviewed 2011 primary and secondary syphilis cases in Los Angeles, California (n=1,755), Craigslist was the second most common website used to meet sex partners [16]. Moskowitz and Seal explored the connection between ad posting frequency and MSM health outcomes, finding that men who frequently posted ads resulting in sexual encounters reported more negative health behaviors and STI rates [17]. Grov examined MSM ads posted in New York City, manually developing guidelines for annotating risk behaviors in ad text [18]. Fries et al. explored authorship attribution approaches for mining the geographic patterns of meeting locations of MSM individuals using Craigslist, as well as phrase/keyword-based behavioral surveillance methods [19]–[22]. They found that in California, self-reported HIV rates in ads were highly correlated with county-level HIV/AIDS prevalence, and that monitoring terminology associated with high-risk behaviors (e.g., unprotected sex requests, methamphetamine-use during sex, etc.) can, at the ecological level, be used to predict yearly, county-level syphilis incidence.

Natural language processing methods of extracting race/ethnicity from free text have largely been explored in the contexts of electronic medical records and social media. Johnson et al. looked at extracting patient race/ethnicity from medical record discharge summaries for comparison with form data gathered from a hospital admitting system [23]. Mislove et al. correlated Twitter user surnames with census data to infer probable race/ethnicity categories [24]. Other work has relied on the fact that race/ethnicity information is often included as structured metadata within a user's profile [25].

Ultimately the notion of race and ethnicity – in terms of categories and terminology – is itself a fraught issue, as noted by Bhopal [26]. Race/ethnicity in Craigslist ads is a self-disclosed variable which can take many textual forms. This introduces challenges in learning the terminology associated with discussing race/ethnicity on Craiglist, as well as mapping that terminology to specific race/ethnicity categories. Some research has suggested that the quality of self-reported race/ethnicity is superior than some sources of administrative data (e.g., VA outpatient clinic files), which exhibited less agreement with self-reported race/ethnicity in non-white groups [27].

III. Material and Methods

A. Craigslist Corpus

Craigslist is an online classified advertisement website that allows users to post free, anonymous classified ads in a variety of different categories (e.g., items for sale, job offerings, dating personals, etc.). Craigslist is organized around local geographic communities and is structured as a network of sub-websites (*sites*). Each site contains ads from its primary anchor city or state, as well as from smaller surrounding communities. There are between 1 and 28 sites per state, with each containing the same set of standardized, Craigslist-defined categories. All ads are publicly accessible via RSS (i.e., Really Simple Syndication – an open Internet standard for publishing content).

Craigslist ads are semi-structured (i.e., tagged with metadata), email-like text documents. They consist of a subject line, keyword-encoded metadata tags, and a body of text. When creating personal ads, authors must select a characterization of the type of relationship they are seeking and the type of person they wish to meet. This information is used by Craigslist to determine an ad's parent category and automatically generate an *encounter tag*, a 3-5 character tag encoding the gender of the author and their requested partner(s), e.g., `m4m` (men for men), `m4w` (men for women), etc. Authors can optionally provide their age and attach a *location tag* to ads, typically indicating a city or neighborhood. Once posted, an ad is available until it expires (7 days for high traffic sites, 45 days for all others) or is removed by the poster.

From July 1, 2009 until February 13, 2012, Craigslist data was downloaded using publicly available, Craigslist-provided RSS feeds using a general-purpose feed aggregator. The aggregator ran daily and retrieved feeds from 8 personal and 5 commercial categories in 412 sites across the United States. Commercial categories include legal services, appliances, pets, furniture, and parking and were chosen to approximate local, non-sexual Craigslist usage patterns. In total, 134 million ads were obtained via RSS. All Craigslist ad text was stripped of HTML markup, made lowercase, and sentence boundary detection done using the pre-trained Punkt sentence tokenizer from the Python module NLTK [28]. Sentences were then tokenized on whitespace and punctuation, with a rule-based system used to merge individual tokens into their final term

representation. Punctuation marks are retained as terms. Tokens were merged in cases where the token is a contraction or found in a manually created lexicon of emoticons, common abbreviations, and other classified ad vernacular, e.g.,"o.b.o" ("or best offer") is combined into a single term. Finally, all terms are stemmed using the NLTK Snowball stemmer.

TABLE I. CRAIGSLIST CORPORA SUMMARY

Name	Ads	Tokens	Selection Criteria
CRAIGSLIST	130.6M	7.8B	Daily sample of 8 personal ad and 5 commercial categories from 412 U.S. Craigslist sites.
MSM	28.6M	2.6B	CRAIGSLIST ad with encounter tag `t` $\in$ {`m4m, m4t, m4mm, mm4mm, mm4m`}
GOLD	700	42k	A uniform random sample of MSM ads from all sites (n=500), all California sites (n=100), and the *sfbay* site (n=100) which was then human-annotated with race and age information.
PHONE	303k	20M	MSM ads containing obfuscated telephone number (e.g., "867-5309" becomes "8sixseven5three oh nine").

MSM-targeted ads are identified by encounter tag, i.e., any ad with tag `t` $\in$ {`m4m, m4t, m4mm, mm4mm, mm4m`}, resulting in 32 million MSM-specific ads. Since authors can anonymously post multiple ads over time, we attempt to partially account for the resampling of individuals by collapsing posts into a single ad instance that are, with high probability, written by the same author. This is done by using a near duplicate detection approach based on locality sensitive hashing; specific technical details on using this approach are found in [19]. (This process also removes any spam ads that escaped Craigslists detection mechanisms.) 3.5M near-duplicate MSM ads were collapsed into single ad instances, resulting in a final set of 28.6M MSM and 102M non-MSM ads, collectively forming our CRAIGSLIST corpus.

Finally, in order to associate ads with specific geographic regions, we extract all geographic named entities (*toponyms*) from ad location tags and link them to canonical database representations - a task called *entity linking* or *normalization*. The algorithm outlined in [29] uses publicly available geographic shapefile data to automatically identify landmarks, roads, neighborhoods, cites, counties, and states mentioned in location tags. In an annotated testing set, this approach correctly linked 85% of all tags to their exact canonical representation, with an overall mean error of 5.7 miles. This linking allows us to compare Craigslist ad demographic attributes with known 2010 U.S. Census population data.

B. Annotation and Phone Number Corpora

Three disjoint datasets of annotated Craigslist MSM ads were created to train our race/ethnicity classifier and create a baseline corpus, GOLD, for all information extraction evaluations. Ads were selected randomly with uniform probability from the set of all sites (n=500), all California sites (n=100), and the *sfbay* site (n=100) and then annotated by the author JAF. Ads were annotated to identify author age and any mention of the race or ethnicity of an ad author or their preferred partner. Race/ethnicity categories follow 2000/2010 U.S. Census definitions: *Caucasian*, *Black*, *Asian*, *Hispanic/Latino* ethnicity, *Native Hawaiian/Pacific Islander*, and *Biracial*, (i.e., identifying as two or more races) [30]. No annotated ad text disclosed *American Indian/Native Alaskan* origins, so that population was not considered in our analysis.

To examine potential resampling (and oversampling) issues introduced by anonymous posting, we also created a second validation corpus for use in a sensitivity analysis of our results. This corpus, PHONE, utilizes the fact that many ad authors provide an obfuscated telephone number in ad text (e.g., "867-5309" becomes "8sixseven5three oh nine") to bypass Craigslist filters, which prohibit including phone numbers in personal ads. By matching phone numbers of this type across all ads, we can identify ad sets written by a single author.

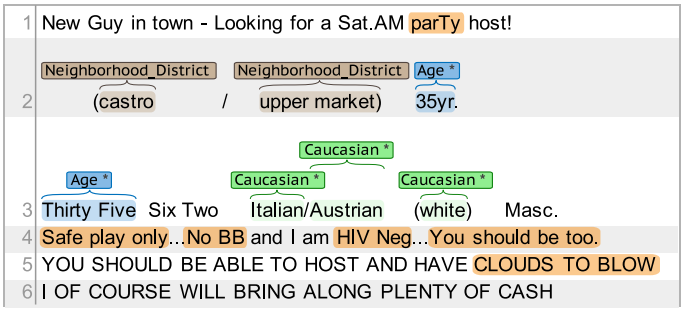

Fig. 1. Example annotations from an m4m *sfbay* ad. Race/ethnicity mentions (in green) are assigned to a racial group (e.g., Caucasian) and an attribute assigned to capture if the mention targets the author or their preferred partner. Orange highlighted text reflects the type of sexual health behaviors discussed in ads; preferences for condom-use during encounters, serosorting preferences, and possible illegal drug use during encounters. Here "parTy" and "CLOUDS TO BLOW" are slang for smoking crystal meth.

C. Extracting Age & Race/Ethnicity

Age information is typically provided as metadata in the ad subject line or the body of ad text. We search ads for all matches to the regular expression (a pattern used for string matching) `(\d+)\s*(yrs|yr|y/o|yo|years old)+` and select the first occurrence found as the author's age. Identifying mentions of author race/ethnicity is more challenging, requiring not only learning the terminology used to describe race/ethnicity, but also disambiguating word sense and adjective targets (e.g., "I'm white" vs. "I'm in a white t-shirt.") Extracting race/ethnicity labels can be viewed as the task of properly labeling the target of a modifying term, either the ad author, their potential partner, or unrelated entity.

We present two basic approaches for extracting race/ethnicity data from ads: (1) *First Mention*, a simple rule-based method; and (2) a hybrid method that extends *First Mention* using machine learning. The second approach follows information extraction work in the biomedical field, where identifying concepts in text can be viewed as a sequence labeling problem [31]. Each of these approaches is described in more detail below.

1) First Mention: We observed in our training data that in 80% of ads that disclose race, the author's race is mentioned first in absolute term offset. Using training data, we built a thesaurus of all labeled race/ethnicity vocabulary terms (e.g., "GWM","white","austrian", etc. maps to Caucasian) and implemented a simple rule-based heuristic which assigns author race based on the first race/ethnicity term found in ad text. No

attempt at disambiguation is attempted in this approach, which provides our baseline performance measure.

1	5 ' 9 ' 140 **blk** brn 30w 7c . . mix **blk** / **mexican**	
	NNNN N **N** N N N NN N **A** N **A**	
2	all race welcom . . . **latino** is a plus	
	N N N NNN **P** N N N	

Fig. 2. Excerpt of labeled output. Each term in a sequence is assigned a `label` ∈ { `AUTHOR (A), PARTNER (P), NONE (N)`} predicted based on 13 features, using conditional random fields. Given this labeling, both *First Mention* and *CRF-First* would incorrectly classify this ad's author as Black. *First Mention* fails to disambiguate the first usage (as hair color) of "blk" while *CRF-First* only considers the first race term labeled as `AUTHOR` . *CRF-All*, which considers all `AUTHOR` labels, would correctly predict Biracial.

2) Hybrid Method: This approach uses 11 Boolean and 2 nominal features (see below) to assign each term a `label` ∈ {`AUTHOR, PARTNER, NONE`} using linear-chain conditional random fields (CRFs). CRFs are undirected graphical models that, in the special case of a linear chain graph structure, can be used to efficiently label sequence data [32]. This machine generated label set is used by a rule-based classifier to then assign an ad to one of the 6 possible race/ethnicity categories (or Undisclosed if no `AUTHOR` tags are found or the labeled term isn't in our thesaurus). By constraining the machine learning step to detect all race/ethnicity mentions, independent of the class that mention belongs to, we help prevent overfitting in the less frequent categories in our training corpus.

For the rule-based classification, we consider two variations of the *First Mention* rule discussed above; (1) *CRF-First*; and (2) *CRF-All*. *CRF-First* uses the first `AUTHOR` labeled term in text to predict a race/ethnicity category, not just the first observed race/ethnicity vocabulary term. Second, *CRF-All* consider the set of all `AUTHOR` labels when assigning a category. If an ad contains `AUTHOR` labeled terms from more than one race/ethnicity terminology cluster, it is assigned to the Biracial class if those terms are separated by a slash (e.g., "white / black") and are not contained within a list. See Figure 2 for an example labeling and its resulting classification.

The performance of these methods is evaluated on the annotation corpus, averaged over 10 trials, with each trial using stratified, 10-fold cross validation. To build our race/ethnicity terminology lexicons, the n-1 folds of annotated training data are used to create the thesaurus used in labeling the documents of the nth fold. This cross-validation approach ensures that we capture the effects of lexical acquisition in our classifier. For *CRF-First* and *CRF-All*, stemmed token word windows of size 3-9 were tested with all features, with 6 performing best overall. Other features such as part-of-speech tags were tested, but did not result in a statistically significant improvement in performance and were not included in the final feature set below:

- *Stemmed Term (Nominal)*: The current term and a 6-term window of all surrounding words.
- *Race/Ethnicity Category (Nominal)*: Name of this term's parent race/ethnicity terminology cluster or None otherwise.
- *Digit + Unit of Measurement*: Term is a unit of measurement, e.g., "5'8'" "130lbs."
- *Metadata*: Term is part of age, location, or encounter tag metadata.
- *First Mention*: Term is the first labeled race/ethnicity term in ad text.
- *List*: Term co-occurs within a 5-term window of other race/ethnicity mentions, e.g., "totally into black, asian, latin or ethnic guys."
- *Pluralization*: Term belongs to a race/ethnicity category and ends in "s".
- *Partner Preference*: Qualifying terms within a 5-term window that express negation or partner preference, i.e., `term` ∈ {`no, not, into, none, only`}.
- *Punctuation*: Term is a punctuation mark.
- *Slash*: Term is in a race/ethnicity category and is separated from another race/ethnicity term by a slash character.
- *Left Verb Argument*: Term is within a 5-term left-window of a set of left/right-associative verbs: `v` ∈ {`looking, seeking, wanted`}. The verb "looking" is ignored if it occurs in the bigrams "good looking" or "nice looking."
- *Right Verb Argument*: Same as above but for the right-term window.
- *PRP + Being Verb*: Term is preceded by a syntactic pattern of the form `PersonalPronoun + BeingVerb` (e.g., "I am") where `BeingVerb` ∈ {`am,'m,:`}.

D. Calculating Demographic Rates

Using the methods described thus far, we extract race/ethnicity and age from all MSM ads. Demographic prevalence rates are calculated by collapsing all ads associated with a geographic region into a single bin L and checking for ads containing search terms associated with race/ethnicity and age attributes. Prevalence rate is then simply the percentage of MSM ads containing the search terms in question for a given location and time interval.

For our demographic analysis, L is defined as the set of all location tag toponyms contained within a U.S. county geographic boundary. Ads containing multiple toponyms, crossing multiple counties, are assigned fractional ad weights based on the number of county bins a location tag resolves to, given below by the function $geobins$. The $weight$ of any given set of ads A is calculated as:

$$w(A) = weight(A) = \sum_{ad \in A} \frac{1}{geobins(ad)} \tag{1}$$

Formally, prevalence is calculated as follows: given M, the set of all MSM ads for a given location and time interval, the prevalence of a terminology cluster T, at location L, at time window t_i is:

$$prev(T, L, t_i) = \sum_{tag \in L} \frac{w(\{ad \in M_{tag,t_i} : text(ad) \cap T \neq \emptyset\})}{w(\{ad \in M_{tag,t_i}\})} \tag{2}$$

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

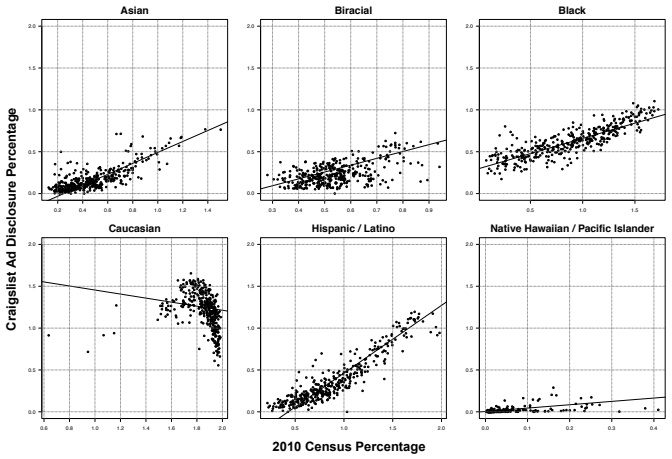

Fig. 3. Scatter plot of log-log regression results for 2010 census race/ethnicity percentages (x-axis, independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (y-axis, dependent variable). Plots are for all CBSA metropolitan geographic boundaries containing at least 1000 ads. 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.

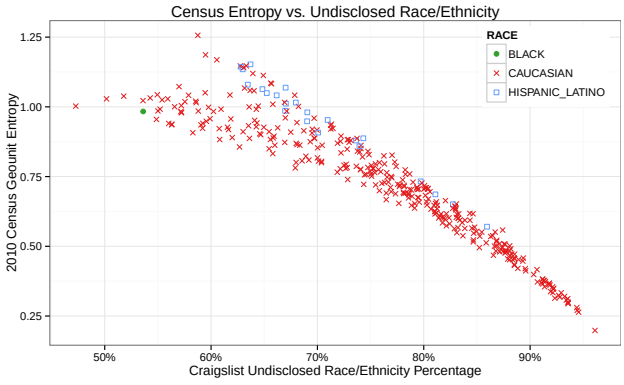

Fig. 4. Scatter plot of 2010 Census geographic Shannon entropy (y-axis), measured across race/ethnicity categories vs. the percentage of ads with undisclosed race/ethnicity (x-axis). The shape of each point indicates the majority race/ethnic group in that geographic boundary. Note how as a population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

TABLE II. AUTHOR RACE/ETHNICITY CLASSIFICATION PERFORMANCE MEASURES

Race/Ethnicity Class	Ad n	Algorithm	Recall μ (SD)	Precision μ (SD)	F_1 score μ (SD)	vs. Baseline F_1 score
Biracial	14	*First-Mention*	0.143 (0.00)	0.933 (0.13)	0.247 (0.01)	-
		CRF-First	0.093 (0.03) *	**1.000 (0.00) ***	0.168 (0.05) *	-31.9%
		CRF-All	**0.600 (0.03) ***	0.672 (0.05)*	**0.632 (0.02) ***	155.8%
Hispanic/Latino	34	*First-Mention*	**0.853 (0.00)**	0.706 (0.01)	0.772 (0.00)	-
		CRF-First	0.829 (0.01) *	0.788 (0.01) *	**0.808 (0.01) ***	4.6%
		CRF-All	0.724 (0.01) *	**0.886 (0.03) ***	0.796 (0.01) *	3.1%
Black	27	*First-Mention*	**1.000 (0.00)**	0.565 (0.01)	0.722 (0.01)	-
		CRF-First	0.970 (0.02) *	0.688 (0.02) *	0.805 (0.01) *	11.5%
		CRF-All	0.941 (0.03) *	**0.782 (0.02) ***	**0.854 (0.02) ***	18.3%
Asian	19	*First-Mention*	0.900 (0.02)	0.740 (0.00)	0.812 (0.01)	-
		CRF-First	**0.947 (0.00)***	0.896 (0.01) *	**0.921 (0.01) ***	13.4%
		CRF-All	0.879 (0.02) *	**0.943 (0.00) ***	0.910 (0.01) *	12.0%
Caucasian	136	*First-Mention*	0.868 (0.00)	0.851 (0.01)	0.859 (0.00)	-
		CRF-First	**0.909 (0.01) ***	0.920 (0.01) *	0.914 (0.01) *	6.4%
		CRF-All	0.900 (0.01) *	**0.935 (0.01) ***	**0.917 (0.01) ***	6.8%
Undisclosed	299	*First-Mention*	0.874 (0.00)	**0.952 (0.00)**	0.912 (0.00)	-
		CRF-First	**0.948 (0.01) ***	0.941 (0.00) *	**0.944 (0.00) ***	3.6%
		CRF-All	**0.948 (0.01) ***	0.916 (0.00) *	0.932 (0.00) *	2.2%
Hawaiian/Pacific-Islander	3	*First-Mention*	**0.333 (0.00)**	0.228 (0.04)	0.269 (0.03)	-
		CRF-First	**0.333 (0.00) ***	**1.000 (0.00) ***	**0.500 (0.00) ***	85.8%
		CRF-All	**0.333 (0.00) ***	**1.000 (0.00) ***	**0.500 (0.00) ***	85.8%

GOLD race/ethnicity classification performance measures. *First-Mention* is a simple rule-based heuristic which assigns author race using the first race/ethnicity term identified in ad text. *CRF-First* and *CRF-All* are hybrid machine learning/rule-based approaches that generate labels to identify author race. Both *CRF-First* and *CRF-All* use *First-Mention* as a baseline for percentage improvement in F-score, with * indicating a statistically significant difference ($p < 0.01$) using a paired t-test. Overall, *CRF-All* provides significant improvements over the baseline, especially in precision, This method performs poorly 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.")

TABLE III. CRAIGSLIST RACE/ETHNICITY VS. 2010 CENSUS

Race/Ethnicity Class	$\lceil$Ad $m\rceil$	Geounit Type	Geounit n	CRAIGSLIST Corpus			PHONE Corpus		
				R^2	Coef.	[95% CI]	R^2	Coef.	[95% CI]
Biracial	365,811	Metropolitan	365	0.43	0.82	[0.72, 0.92]	0.26	0.73	[0.60, 0.86]
	13,868	Micropolitan	523	0.53	0.60	[0.55, 0.64]	0.32	0.70	[0.62, 0.79]
	30,576	County	587	0.76	0.76	[0.73, 0.79]	0.47	0.72	[0.67, 0.77)
Hispanic/Latino	963,810	Metropolitan	365	0.91	0.81	[0.78, 0.84]	0.81	0.81	[0.77, 0.85]
	19,833	Micropolitan	523	0.56	0.54	[0.50, 0.58]	0.25	0.58	[0.50, 0.66]
	36,853	County	587	0.66	0.61	[0.58, 0.64]	0.54	0.71	[0.66, 0.75]
Black	992,014	Metropolitan	365	0.71	0.36	[0.34, 0.39]	0.54	0.46	[0.41, 0.5]
	38,985	Micropolitan	523	0.44	0.33	[0.30, 0.37]	0.24	0.51	[0.44, 0.59]
	41,675	County	587	0.45	0.24	[0.22, 0.26]	0.19	0.30	[0.26, 0.34]
Asian	359,114	Metropolitan	365	0.86	0.65	[0.63, 0.68]	0.63	0.57	[0.52, 0.61]
	11,472	Micropolitan	523	0.54	0.55	[0.51, 0.60]	0.36	0.58	[0.51, 0.64]
	38,535	County	587	0.66	0.55	[0.53, 0.58]	0.44	0.44	[0.41, 0.47]
Caucasian	4,869,614	Metropolitan	365	0.04	-0.24	[-0.35, -0.13]	0.04	-0.24	[-0.36, -0.11]
	215,777	Micropolitan	523	0.05	-0.38	[-0.52, -0.24]	0.01	-0.27	[-0.52, -0.02]
	218,246	County	587	0.04	-0.18	[-0.23, -0.12]	0.10	-0.46	[-0.55, -0.37]
Undisclosed (vs. Caucasian Census)	18,349,737	Metropolitan	365	0.34	0.23	[0.20, 0.27]	0.39	0.31	[0.27, 0.35]
	1,006,751	Micropolitan	523	0.19	0.21	[0.17, 0.24]	0.08	0.36	[0.26, 0.45]
	962,799	County	587	0.23	0.13	[0.11, 0.14]	0.07	0.19	[0.15, 0.24]
Native Hawaiian/Pacific Islander	20,289	Metropolitan	365	0.42	0.40	[0.35, 0.45]	0.21	0.40	[0.32, 0.47]
	1,646	Micropolitan	523	0.94	0.37	[0.37, 0.38]	0.90	0.47	[0.46, 0.48]
	4,237	County	587	0.96	0.45	[0.44, 0.45]	0.88	0.55	[0.54, 0.56]

CRAIGSLIST OLS log-log regression results for census race/ethnicity percentages (independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (dependent variable). Data points consist of the percentage population makeup, as determined by 2010 census data and measured Craigslist disclosures. Ad m (which can take fractional weights when crossing multiple geographic boundaries) is rounded up. Coefficients are in log units meaning, for example, a 1% increase in census population makeup for the Hispanic/Latino group results in a 0.81% C.I [0.78, 0.84] increase in Craigslist ads disclosing Hispanic/Latino origin.

B. Author Age

In the GOLD corpus, 90% (627/700) of ads contained author age information. Our regular expression correctly matched 95% (594/627) of all age tags with precision 0.99, recall 0.95 and F_1 score 0.97. Age correlations were strongest in metropolitan geographic areas, with the highest correlations

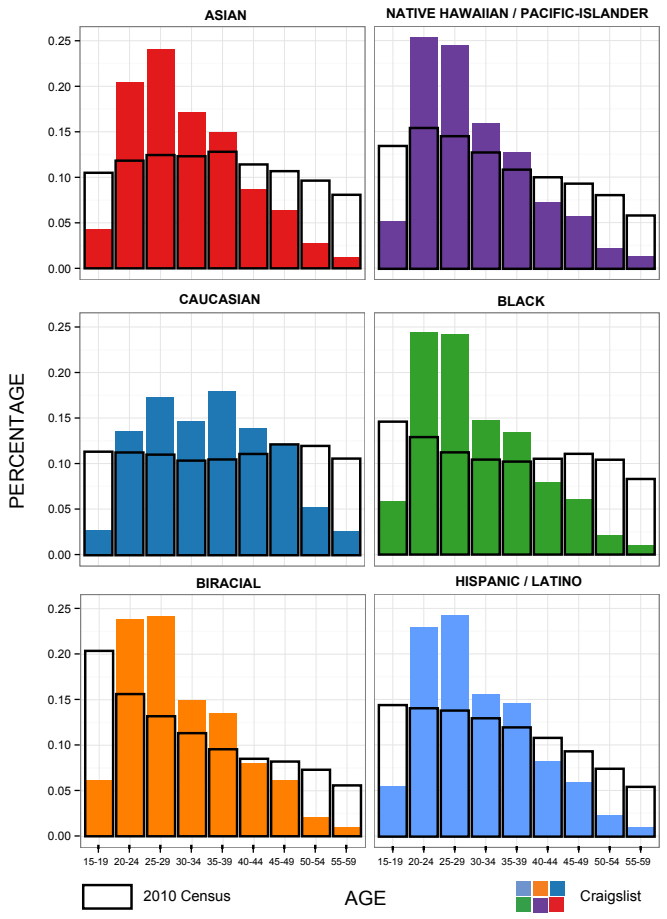

Fig. 5. Normalized histogram of aggregate, self-reported age by race for the CRAIGSLIST corpus. The black outline bars represent the 2010 census distribution of age groups vs. the solid colored bars, representing Craigslist age distributions. Caucasian ads trend older overall compared to other racial/ethnic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

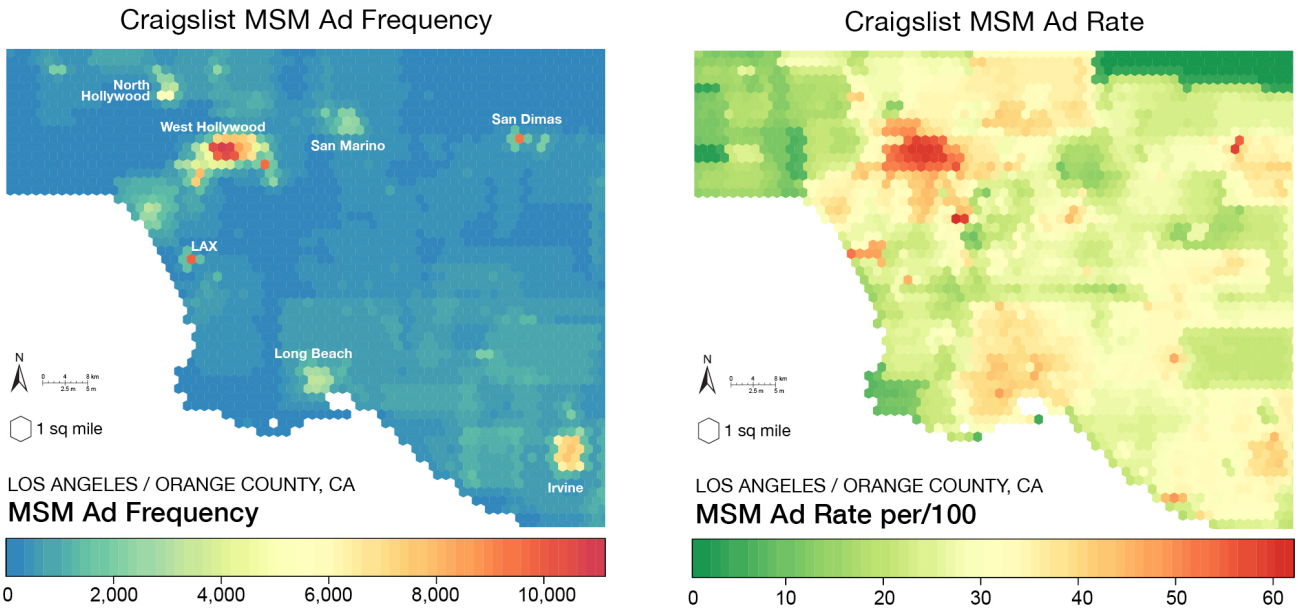

Fig. 6. Heat maps of California's Los Angeles/Orange County region, showing the total MSM ad mass for a given 1 square mile hexagon cell (left) and a normalized MSM ad rate for the same area (right). On the left, the most active MSM areas are labeled (e.g., West Hollywood, Los Angeles International Airport (LAX), etc.). The right map looks at the total percentage of Craigslist MSM activity, calculated as MSM ads divided by all CRAIGSLIST ads (personal ads + a sample of commercial activity categories like furniture, appliances, pets, etc.) for a given cell. 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.

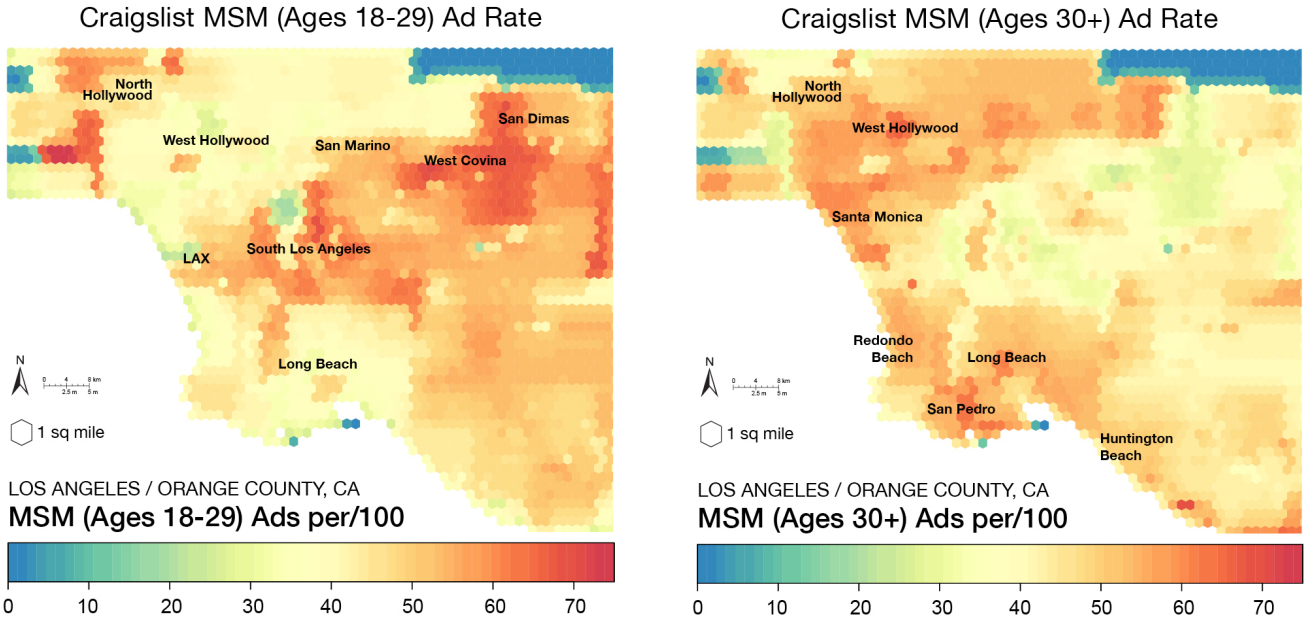

Fig. 7. Heat map of Craigslist self-disclosed author age in California's Los Angeles/Orange County region, 18-29 year-olds (left) and 30+ year-olds (right). The 18-29 age group makes up the majority of ads (n=854,144) in the central and northeastern parts of the map (e.g., South Los Angeles, West Covina, etc.). The 30+ age group (n=890,568) is more prevalent in a few northern and oceanside areas (e.g., Santa Monica, San Pedro, Huntington Beach).

overall. This difference may also speak to actual differences in Craigslist posting behavior across MSM individuals of different racial/ethnic backgrounds.

The similarities between the CRAIGSLIST and PHONE race/ethnicity regressions suggest that any resampling effects are effectively dealt with by our near-duplicate removal system or are otherwise small enough to have little impact on the overall analysis. PHONE is less informative in the context of age, partly due to the number of age classes (10+) and the resulting decrease in observations per class.

A. *Limitations*

There are many limitations to our information extraction methodology. First, Craigslist ads reflect an intention for an encounter, not necessarily that an encounter took place. However, ad content still provides insight into a community's practices

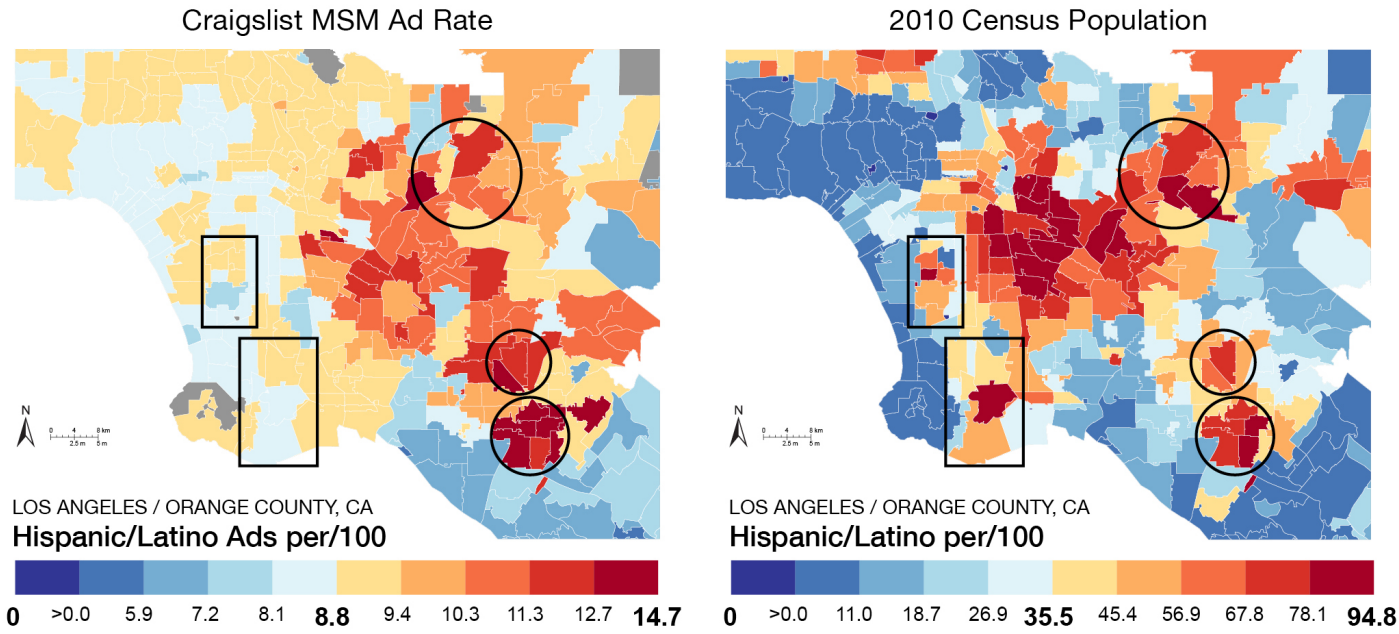

Fig. 8. Choropleth map of Hispanic/Latino author mentions in California's Los Angeles/Orange County region (left) compared to 2010 census data (right). For the Craigslist map, all toponyms are binned into their parent Zip Code Tabulation Area, weighted by percentage of overlap. Bins with $\leq$ 100 observations are greyed out. There is a visible correspondence between Craigslist meeting locations and the underlying population distributions, with some Craigslist location clusters closely follow the underlying population distribution, e.g., Santa Ana, Anaheim, and La Puente (circles). Other regions, e.g., Wilmington and Inglewood (rectangles), have meeting locations on the periphery of core population areas. This may correspond to an error in extracting geographic entities or it may reflect underlying travel and meeting preferences of MSM individuals in or around those areas.

respective of demographic information. Second, ad locations reflect where an ad author wishes to meet, not necessarily where that person lives. This means linking meeting location data with census data will always contain some element of error. Moreover, the geographic entity linking process itself is inexact; this can cause meeting location clusters to grow more diffuse as entities are misattributed to nearby locations.

Third, like most natural language processing applications, terminology acquisition and disambiguation create interesting challenges; our approach only uses terms identified in our training data to identify the racial/ethnic background of an author. Race/ethnicity categories like Hispanic/Latino have larger potential synonym sets than our annotated dictionaries capture (e.g., Cuban, Mexican, Chilean, etc.) and common, overloaded words like "black" and "white" highlight the importance of word sense disambiguation. Distant supervision approaches to word surface form clustering (e.g., Brown clustering [35], [36]) have been used successfully in named entity recognition applications and could be used to learn terminology and mappings to known racial/ethnic groups, improving the recall of our method.

Finally, our method is only evaluated within the context of Craigslist making it difficult to gauge performance in other settings; analyzing text from other online classified ad communities (e.g., Backpage) or MSM social networks like Grindr, Adam4Adam, etc. would be an interesting area of future exploration.

VI. Conclusion

The use of the Internet to seek out sexual partners has had a profound impact on the epidemiology of sexually transmitted infections. Sexual encounters arranged online are likely to be geographically broader, anonymous, and greater in number. Traditional approaches to control STIs involve contact tracing and partner notification, however Internet-mediated encounters can undermine these prevention and control measures. Demographic variables like race/ethnicity and age are essential for informing public health interventions, as interventions designed for one age or racial/ethnic group may not be appropriate for others. While some web sites collaborate with public health in providing risk reduction information, large-scale information about the communities themselves remains difficult to obtain. Mining information in an automatic fashion provides an inexpensive way to collect timely, large-scale, location-specific behavioral surveillance data that can be used to inform public health intervention strategies.

Once ads can be reliably linked with both location and demographic metadata a number of interesting applications become possible. Public health agencies could, for example, use emerging topic detection approaches to generate lists of "trending locations" (similar to systems used on Twitter or Foursquare for trending events or places) in order to identify MSM hotspots within cities. These queries could incorporate not only static demographic data (e.g., "where do 18-29 year-old black men meet for sexual encounters in Los Angeles, California?"), but also incorporate time to detect anomalies or other seasonal changes in meeting locations.

Methods for automatically learning the language used to discuss health behaviors is another interesting future research direction. The ability to build lexicons defining high-risk sexual behaviors could help answer more sophisticated surveillance questions. What is the geographic distribution

of unprotected sex requests or methamphetamine use during encounters in Los Angeles? What is the distribution of those behaviors in ads targeting West Hollywood? Typically these types of questions are answered through survey data. Manually identifying search terms associated with high-risk behaviors (e.g., "bareback" as slang for unprotected sex, "my friend tina" as code for methamphetamine", etc.) and geographically indexing ad text provides an inexpensive, automated way of of augmenting survey data. We leave these ideas for future exploration.

REFERENCES

[1] Centers for Disease Control and Prevention (CDC). (2014, Feb) HIV among African Americans. [Online]. Available: http://www.cdc.gov/hiv/pdf/risk_HIV_AfricanAmericans.pdf

[2] H. Jaffe, R. Valdiserri, and K. De Cock, "The reemerging HIV/AIDS epidemic in men who have sex with men," *JAMA: the journal of the American Medical Association*, vol. 298, no. 20, pp. 2412–2414, 2007.

[3] T. Sanchez, T. Finlayson, A. Drake, S. Behel, M. Cribbin, E. DiNenno, T. Hall, S. Kramer, A. Lansky, C. for Disease Control, and P. (US), *Human Immunodeficiency Virus (HIV) Risk, Prevention, and Testing Behaviors: United States, National HIV Behavioral Suveillance System: Men who Have Sex Men, November 2003 [to] April 2005*. US Department of Health and Human Services, 2006.

[4] D. Osmond, L. Pollack, J. Paul, and J. Catania, "Changes in prevalence of HIV infection and sexual risk behavior in men who have sex with men in San Francisco: 1997–2002," *Journal Information*, vol. 97, no. 9, 2007.

[5] M. McFarlane, S. Bull, and C. Rietmeijer, "The internet as a newly emerging risk environment for sexually transmitted diseases," *JAMA: the journal of the American Medical Association*, vol. 284, no. 4, pp. 443–446, 2000.

[6] A. Liau, G. Millett, and G. Marks, "Meta-analytic examination of online sex-seeking and sexual risk behavior among men who have sex with men," *Sexually transmitted diseases*, vol. 33, no. 9, p. 576, 2006.

[7] B. Rosser, W. West, and R. Weinmeyer, "Are gay communities dying or just in transition? Results from an international consultation examining possible structural change in gay communities," *AIDS care*, vol. 20, no. 5, pp. 588–595, 2008.

[8] E. Benotsch, S. Kalichman, and M. Cage, "Men who have met sex partners via the internet: Prevalence, predictors, and implications for HIV prevention," *Archives of Sexual Behavior*, vol. 31, no. 2, pp. 177–183, 2002.

[9] Alexa: The Web Information Company. (2013, September) Craigslist.org site info. [Online]. Available: http://www.alexa.com/siteinfo/craigslist.org

[10] S. Wakamiya, R. Lee, and K. Sumiya, "Crowd-sourced urban life monitoring: Urban area characterization based crowd behavioral patterns from Twitter," in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*. ACM, 2012, p. 26.

[11] K. M. Blankenship, S. J. Bray, and M. H. Merson, "Structural interventions in public health," *Aids*, vol. 14, pp. S11–S21, 2000.

[12] J. Kraut-Becher, M. Eisenberg, C. Voytek, T. Brown, D. S. Metzger, and S. Aral, "Examining racial disparities in HIV: Lessons from sexually transmitted infections research," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 47, pp. S20–S27, 2008.

[13] K. Bernstein, F. Curriero, J. Jennings, G. Olthoff, E. Erbelding, and J. Zenilman, "Defining core gonorrhea transmission utilizing spatial data," *American journal of epidemiology*, vol. 160, no. 1, pp. 51–58, 2004.

[14] C. Fichtenberg and J. Ellen, "Moving from core groups to risk spaces," *Sexually Transmitted Diseases*, vol. 30, no. 11, p. 825, 2003.

[15] J. Chan and A. Ghose, "Internet's dirty secret: Assessing the impact of technology shocks on the outbreaks of sexually transmitted diseases," 2012.

[16] California Department of Public Health. (2011, Dec) California syphilis elimination surveillance data. [Online]. Available: http://www.cdph.ca.gov/data/statistics/Documents/STD-Data-Syphilis-Elimination-Surveillance-Data.pdf

[17] D. A. Moskowitz and D. W. Seal, ""GWM looking for sex-SERIOUS ONLY": The interplay of sexual ad placement frequency and success on the sexual health of "men seeking men" on Craigslist," *Journal of Gay & Lesbian Social Services*, vol. 22, no. 4, pp. 399–412, 2010.

[18] C. Grov, "Risky sex-and drug-seeking in a probability sample of men-for-men online bulletin board postings," *AIDS and Behavior*, vol. 14, no. 6, pp. 1387–1392, 2010.

[19] J. A. Fries, A. M. Segre, and P. M. Polgreen, "Using online classified ads to identify the geographic footprints of anonymous, casual sex-seeking individuals," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 402–410.

[20] ——, "Towards linking anonymous authorship in casual sexual encounter ads," *Online Journal of Public Health Informatics*, vol. 5, no. 1, 2013.

[21] J. A. Fries, A. M. Segre, L. Polgreen, and P. M. Polgreen, "The use of Craigslist posts for risk behavior and STI surveillance," *International Society for Disease Surveillance Conference 2010*, p. 13, 2011.

[22] J. A. Fries, A. Ho, A. M. Segre, and P. M. Polgreen, "Using Craigslist messages for syphilis surveillance," in *International Meeting on Emerging Diseases and Surveillance (IMED)*, 2011.

[23] S. B. Johnson and C. Friedman, "Integrating data from natural language processing into a clinical information system." in *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1996, p. 537.

[24] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of Twitter users." *ICWSM*, vol. 11, p. 5th, 2011.

[25] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, "Tastes, ties, and time: A new social network dataset using Facebook.com," *Social Networks*, vol. 30, no. 4, pp. 330–342, 2008.

[26] R. Bhopal, "Glossary of terms relating to ethnicity and race: for reflection and debate," *Journal of Epidemiology and Community Health*, vol. 58, no. 6, pp. 441–445, 2004.

[27] U. Boehmer, N. R. Kressin, D. R. Berlowitz, C. L. Christiansen, L. E. Kazis, and J. A. Jones, "Self-reported vs administrative race/ethnicity data and study results," *American Journal of Public Health*, vol. 92, no. 9, pp. 1471–1472, 2002.

[28] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, 2009.

[29] J. A. Fries and A. M. Segre, "Geographic entity linking in online classified ads," 2014, in-review.

[30] United States Census Bureau. (2013, August) 2000 redistricting data (public law 94-171) summary file. [Online]. Available: http://www.census.gov/prod/cen2000/doc/pl94-171.pdf

[31] J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 524–527, 2010.

[32] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[33] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[34] A. E. Magurran, "Measuring biological diversity," 2004.

[35] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[36] S. Miller, J. Guinness, and A. Zamanian, "Name tagging with word clusters and discriminative training." in *HLT-NAACL*, vol. 4. Citeseer, 2004, pp. 337–342.