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` ∈ {`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` ∈ {`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.

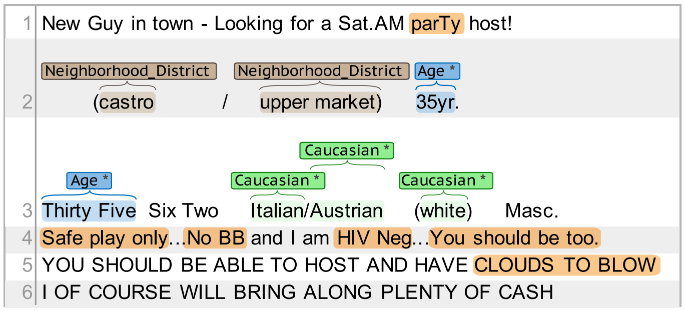
1 New Guy in town - Looking for a Sat.AM parTy host!
2 (castro / upper market) 35yr.
3 Thirty Five Six Two Italian/Austrian (white) Masc.
4 Safe play only...No BB and I am HIV Neg...You should be too.
5 YOU SHOULD BE ABLE TO HOST AND HAVE CLOUDS TO BLOW
6 I OF COURSE WILL BRING ALONG PLENTY OF CASH

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