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