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