Harnessing Craigslist Personal Ads to Inform Federal HIV Prevention Funding

Varoon K. Bashyakarla
Yale University
185 Berry St., Suite 400
San Francisco, CA 94109
+13162580469
vbashyakarla@gmail.com

Claire Kelley
Yale University
24 Hillhouse Avenue
New Haven, CT 06511
+17756227632
claire1053@gmail.com

Allen Lin
Harvard University
35 Oxford Street
Cambridge, MA 02138
+15083384728
allenlin@g.harvard.edu

ABSTRACT
The United States has made substantial progress towards fighting HIV/AIDS, but the number of new HIV infections has remained steady over the past few years. Because men who have sex with men (MSM) bear the heaviest burden of the disease, existing research methods and prevention efforts often rely on the size of gay communities and their adoption of safe sex practices. However, the spread of HIV can also be curtailed by targeting prevention efforts at MSM who do not identify as gay. Here, data from every post available on Craigslist's "men seeking men" forum in every locale in the United States over the course of one week were collected. Among these 435,000 posts, over 40% of ads reference some form of discretion, effectively separating the ad's poster from the out gay community. In addition, states vary widely with respect to the proportion of ads seeking safe sex, and a small collection of primarily conservative states (Mississippi, Iowa, West Virginia, Wyoming, Montana, North Dakota, and South Dakota) were found to have fewer posts relative to their populations and markedly lower proportions of ads seeking safe sex compared to those of other states. These initial findings suggest that perhaps certain states that lack visible gay communities also lack HIV prevention resources in spite of the presence of MSM.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – Text Analysis; K.4.1 [**Computers and Society**] Public Policy Issues

General Terms
Measurement, Human Factors

Keywords
Text mining, Craigslist, HIV

1. INTRODUCTION
Since the outbreak of AIDS in the United States in 1981, HIV has infected over 1.8 million Americans and claimed over 650,000 lives in the United States [4, 5, 7]. In 2010, the White House announced a goal to reduce the number of new HIV infections by 25% by 2015 [11]. Recent prevention efforts have lowered HIV transmission rates, increased access to HIV screening and treatment, and lowered AIDS-related mortality [10]. However, the incidence of the virus has not declined, as the number of new HIV infections has remained steady at about 50,000 in recent years [6].

Men who have sex with men (MSM) are most seriously threatened by the risk of contracting HIV [2]. While the CDC estimates that MSM compose only 4% of the male population in the United States, MSM constituted 63% of all new infections and 78% of those among men in 2010 [2, 8]. In addition, domestic prevention efforts account for only 3% of the $29.7 billion in federal HIV/AIDS spending in the FY 2014 budget request [9]. With over 1 million Americans living with HIV, 16% of whom are unaware of their HIV+ status, the present challenge is two-fold [7]. First, limited HIV prevention resources must be allocated to locales in the United States commensurate with their relative need. And prior to any such distribution of resources, it is necessary to measure the needs of locales across the country in a manner that fairly represents their demand. This project concentrates on the latter goal.

2. DATA
The data consist of all ads posted and accessible on Craigslist during a one-week period (July 20-27, 2013) in the "men seeking men" personals section of the website. The following features were collected for each of the 434,956 ads scraped: a unique ad identifier, which suburbs the ad was cross-posted on (if any), part of the ad's URL designating which parent site the posting came from, the Craigslist city in which the ad was posted, the corresponding state, the age of the poster (if reported), any user-provided location data, what neighboring town/city the ad is from (if cross-posted), whether any photos were included in the ad, the title, the contents of the ad's message, the date/time on which the ad was originally posted, and subsequent timestamps at which the ad may have been updated. In total, these data span 419 cities in all 50 states, the territories, and D.C. In this timeframe, the Dallas/Fort Worth area had the largest number of posts (17,477), while the northeast South Dakota and southwest Texas regions tied for the fewest (3 each).

These data are germane to the topic at hand for a few reasons. Craigslist data is abundant, publicly accessible, and quite understudied. Additionally, the city-level aggregation available on Craigslist is well suited to studying HIV/AIDS, which is a largely urban disease. In 2011, the majority of new HIV diagnoses were made in cities with at least 500,000 people [3]. Furthermore, it's not unreasonable to suggest that these data provide a glimpse into parts of the MSM community and sexual practice more generally that mainstream research may not reach. Mining the content of user-created ads mitigates reporting bias; people are presumably more likely to be frank about their sexual practices in ads targeted at their potential sexual partners than they might be in reporting to a survey.

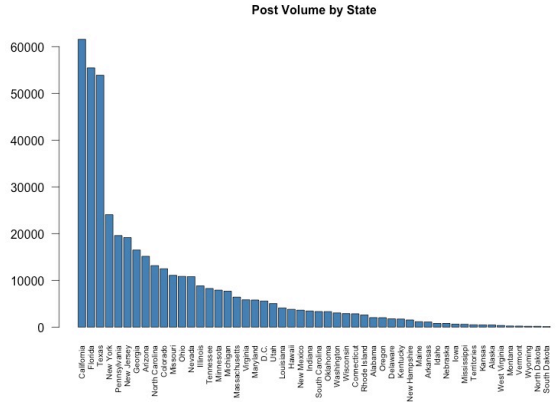

Figure 1: The distribution of post volume by state

3. POST ANALYSIS

3.1 Discretion and Qualitative Features

One key problem for HIV researchers is assessing need for prevention resources, which is often currently based on estimated gay populations, infection levels, and employment of safe-sex practices. However, the presence of MSM groups, which differ from gay-identified individuals, can improve understanding of where HIV resources are needed but have been difficult to quantify meaningfully. By incorporating Craigslist data from people who may not identify as gay, or who may not participate in gay events, we may infer a more robust understanding of these metrics. Out of the sample of about 435,000 cases, approximately 68,700 – or nearly 16% – explicitly mention the word "discreet" in the ad. Some ads posters use key words such as "anonymous" and "straight" to distance themselves from the out gay community. Incorporating other terms ("married," "not out," "str8," "straight-acting," "discrete," a common typo, etc.) brings the total to over 40% of ads that express some degree of differentiation from the gay-identified community. To the extent that these solicitations for sex are actually pursued, these data may help us understand the dynamics of MSM and HIV transmission among those who don't identify with the gay community, which traditional research methods can easily overlook.

To get a quick, visual overview at what posters were looking for, a word map of all contents of the ads was created. While this is not an analytical tool, it communicates a qualitative feature of the data that can inform and shape our analytical process. As the word map illustrates, posters tend to use words that seem to relate to more casual and sexually charged behaviors ripe for HIV prevention analysis.

Figure 2: A word map of the contents of all ads

3.2 HIV Implications

The HIV implications of these ads were studied next – in particular the mentions of HIV and safe sex. More than 30,000 of the ads directly mention HIV and 171,000 refer to safe sex (defined as those that mention "condom," "clean," "safe," or "ddf," a commonly-used acronym for "drug and disease free"). Manual inspection of many ads identified these words most commonly used to indicate a desire for safe and protected sex. About one-third of all ads are seeking safe sex, though this number varies greatly by state.

As the graph below shows, D.C. has the highest proportion of ads looking for safe sex, with almost half of the ads mentioning an associated key word. At the opposite end of the spectrum, less than one-fifth of ads from the U.S. territories seem to solicit safe sex. Wyoming, Montana, and North and South Dakota are the worst performing states, with less than one-third of ads looking for protected sex.

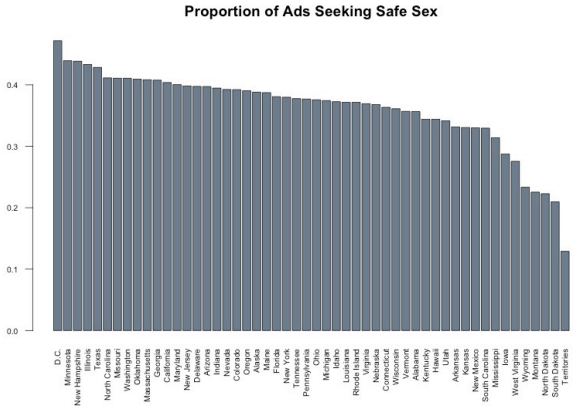

Figure 3: The distribution of proportion of ads seeking safe sex by region

To get a better impression of the data's geographic characteristics, a color-coded heat map of the states was produced with darker colors used to indicate lower proportions of ads seeking safe sex. The heat map shows geographically contiguous regions tend to cluster together with respect to the rate of safe sex solicitations.

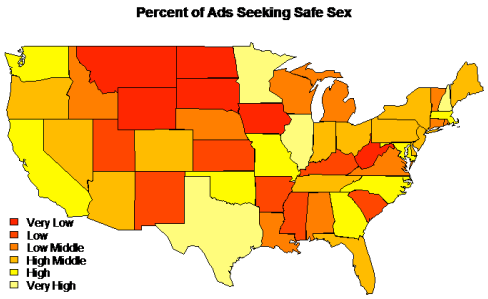

Figure 4: Heat map of contiguous states with respect to rate of safe sex solicitations

3.3 Political Factors

Next, to extend this analysis, it was decided to investigate the potential role of politics in HIV prevention resources. Because gay communities tend to thrive in liberal places, it is natural to wonder if forums like Craigslist cater more to conservative states. In these areas, men may be less inclined to solicit sex in person and instead elect to use the internet, which affords security and anonymity. The plot below graphs the proportion of ads seeking safe sex in each state against the number of posts per capita (using Census Bureau projections to estimate 2013 state-level populations) [12].

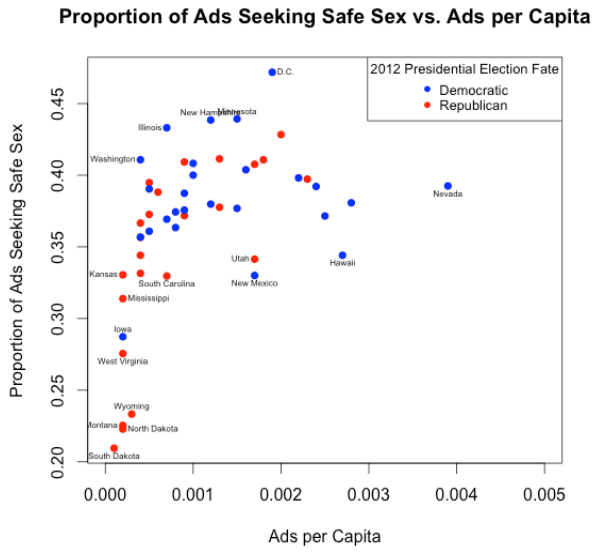

Figure 5: Proportion of ads seeking safe sex vs. Ads per capita by state, colored by political leaning

The data seem to suggest that the number of ads per capita for conservative states tends to be slightly lower than that of liberal states. Moreover, the positive association ($\rho \approx 0.35$) between the variables is striking. Additionally, the graph shows that certain states (Mississippi, Iowa, West Virginia, Wyoming, Montana, North Dakota, and South Dakota) all have relatively few MSM ads on Craigslist, and these ads solicit safe sex at rates markedly lower than that of most other states. In other words, this initial inspection seems to suggest that areas with fewer Craigslist posts per person also tend to have a lower percentage of ads seeking safe sex. It is natural to wonder, then, whether these areas that may lack visible gay communities also lack HIV prevention resources. Interestingly, 6 of these 7 states are conservative, and Mississippi and West Virginia were both among the group of states with the highest number of new HIV cases in 2011 [1].

4. DISCUSSION AND CONCLUSIONS

HIV prevention continues to pose a challenge for the United States, as the incidence of the virus has not abated over the past few years. MSM disproportionately bear the heaviest burden of HIV/AIDS in the United States. Moreover, in attempt to minimize the spread of the virus, researchers are often constrained by the presence of gay-identified populations. Anonymous data sources like Craigslist can complement existing research to elucidate the manner in which HIV prevention resources may be required among MSM who do not identify as gay or bisexual.

Initial results suggest that a substantial proportion of posts on Craigslist's "men seeking men" forum desire anonymity and vary widely with respect to desire for safe sex. A small collection of largely conservative states have a relatively low volume of posts relative to their populations, and ads in these states tend to seek safe sex at pronouncedly lower rates than that of other states.

5. FUTURE WORK

Future data mining and machine learning efforts with this rich data set can take many shapes: constructing a refined classifier to predict whether posts are seeking safe sex; monitoring racial preferences in the data and how they differ by region, political climate, and areas in which gay marriage has been legalized; and performing clustering to see which parts of the country are more similar to one another in terms of MSM activity and need for HIV prevention resources. Additionally, more data should be collected to test these initial findings, and current HIV prevention funding mechanisms should be studied to determine the extent to which Craigslist data may suggest allocating HIV prevention resources more strategically.

6. ACKNOWLEDGMENTS

We are grateful to Chris Peak whose documentary short, *Looking*, partly inspired this work and to Dropbox for funding travel to KDD to present this work at the Workshop on Data Science for Social Good.

7. REFERENCES

[1] Centers for Disease Control and Prevention, 2011. *HIV Surveillance Report*, Volume 23.

[2] Center for Disease Control and Prevention, December 2012. Estimated HIV incidence in the United States, 2007–2010. *HIV Surveillance Supplemental*, Report 2012; 17(No. 4).

[3] Center for Disease Control and Prevention, December 2013. *Today's HIV/AIDS Epidemic*.

[4] Center for Disease Control and Prevention, February 2013. *HIV Surveillance Report*, Vol. 23.

[5] Center for Disease Control and Prevention, June 1981. *Morbidity and Mortality Weekly Report*, Vol. 30, No. 21.

[6] Center for Disease Control and Prevention, November 2013. HIV in the United States: At a Glance.

[7] Center for Disease Control and Prevention, October 2013. *HIV Surveillance Supplemental Report*, Vol. 18, No. 5.

[8] D. W. Purcell, C.H. Johnson, A. Lansky, J. Prejean, R. Stein, P. Denning, Z. Gau, H. Weinstock, J. Su, and N. Crepaz, 2012. Estimating the population size of men who have sex with men in the United States to obtain HIV and syphilis rates. *Open AIDS Journal*, Volume 6 (Supp 1: M6), 98-107.

[9] Kaiser Family Foundation, 2014. U.S. Federal Funding for HIV/AIDS: The President's FY2014 Budget Request.

[10] National Institutes of Health, February 2013. Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents.

[11] The White House, July 2010. National HIV/AIDS Strategy for the United States. Washington, D.C.

[12] United States Census Bureau, 2013. State Totals: Vintage 2013.