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

Jason A. Fries, Phil Polgreen, MD, Alberto M. Segre
Presentation Outline

I. Public Health Motivation
II. Ad Parsing Pipeline
III. Demographic Results
IV. Future Work
I. Public Health Motivation
I. Public Health Motivation
Men Who Have Sex with Men (MSM) Health Disparities

\(~2\%\) of the population
Men Who Have Sex with Men (MSM) Health Disparities

~2% of the population
63% of new HIV cases

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Count (Percentage)</th>
<th>Age Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>11,400 (38%)</td>
<td>25-34</td>
</tr>
<tr>
<td>Black</td>
<td>10,600 (36%)</td>
<td>13-24</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>6,700 (22%)</td>
<td>25-34</td>
</tr>
</tbody>
</table>

— CDC “HIV Among Gay and Bisexual Men”, March 2013
STAMP OUT SYPHILIS AND GONORRHEA

7% MSM syphilis in 2000
STAMP OUT SYPHILIS AND GONORRHEA

72% MSM syphilis in 2011
STAMP OUT Syphilis and Gonorrhea

39% Increase
Drug resistant Gonorrhea

Cephalosporin-Resistant Gonorrhea in North America

JAMA Editorial
January 9, 2013

Increasing Minimal Inhibitory Concentrations (MICs) — an early warning of impending resistance.
### Patient Demographics/Risk Assessment Information

<table>
<thead>
<tr>
<th>Age/DOB:</th>
<th>Sex:</th>
<th>Race/Ethnicity:</th>
<th>Date of Test:</th>
</tr>
</thead>
<tbody>
<tr>
<td>County of Residence:</td>
<td>State:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Patient Risk Assessment Information

- **Check applicable answers and complete patient exposure information within past 12 months as required by CDC.**

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Had sex w/male?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex w/female?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex w/transgender?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex w/anon. partner?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex w/o condom?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex while intoxicated/high?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exchanged drugs/money for sex?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females-had sex w/known MSM?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had sex w/known IDU?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Been incarcerated?</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

- Injection drug use?  
- Shared injection equipment?  
- Injection/Non-Inject drug usage? (Note drugs: )  
- Was patient tested for HIV?  
- Patient’s HIV status?  
- Prior STD history?  
- Was patient counseled for HIV?  
- Met partners via internet?  
- Reason for exam?  
- Symptomatic  
- Asymptomatic  
- Contact to STD
New guy in town - Looking for a Sat.AM party host!
(castro/upper market) 35yr.

Thirty Five Six Two Italian/Austrian (white) Masc.
Safe play only…No BB and I am HIV Neg…You should be too.
YOU SHOULD BE ABLE TO HOST AND
HAVE CLOUDS TO BLOW
I OF COURSE WILL BRING ALONG PLENTY OF CASH

• do NOT contact me with unsolicited services or offers
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>New Guy in town - Looking for a Sat.AM <strong>parTy</strong> host!</td>
</tr>
<tr>
<td>2</td>
<td>Neighborhood <strong>District</strong> / Neighborhood <strong>District</strong> <strong>Age</strong> *</td>
</tr>
<tr>
<td>3</td>
<td>(castro / upper market) <strong>35yr.</strong></td>
</tr>
<tr>
<td>4</td>
<td><strong>Age</strong> *</td>
</tr>
<tr>
<td>5</td>
<td>Caucasian *</td>
</tr>
<tr>
<td>6</td>
<td>Caucasian *</td>
</tr>
<tr>
<td>7</td>
<td>Caucasian *</td>
</tr>
<tr>
<td>8</td>
<td>Thirty Five Six Two Italian/Austrian <strong>(white)</strong> Masc.</td>
</tr>
<tr>
<td>9</td>
<td>Safe play only...No BB and I am <strong>HIV Neg.</strong> You should be too.</td>
</tr>
<tr>
<td>10</td>
<td><strong>YOU SHOULD BE ABLE TO HOST AND HAVE CLOUDS TO BLOW</strong></td>
</tr>
<tr>
<td>11</td>
<td>I OF COURSE WILL BRING ALONG PLENTY OF CASH</td>
</tr>
<tr>
<td></td>
<td>New Guy in town - Looking for a Sat.AM <strong>party</strong> host!</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td><strong>Neighborhood District</strong> / <strong>Neighborhood District</strong></td>
</tr>
<tr>
<td></td>
<td>(castro) / upper market) <strong>Age</strong></td>
</tr>
<tr>
<td>2</td>
<td>35yr.</td>
</tr>
<tr>
<td>3</td>
<td><strong>Age</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Caucasian</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Age</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Caucasian</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Caucasian</strong></td>
</tr>
<tr>
<td>4</td>
<td>Thirty Five Six Two Italian/Austrian (white) Masc.</td>
</tr>
<tr>
<td>5</td>
<td>Safe play only...No BB and I am HIV Neg...You should be too.</td>
</tr>
<tr>
<td>6</td>
<td>YOU SHOULD BE ABLE TO HOST AND HAVE CLOUDS TO BLOW</td>
</tr>
<tr>
<td>7</td>
<td>I OF COURSE WILL BRING ALONG PLENTY OF CASH</td>
</tr>
</tbody>
</table>

**Geographic Location**  **Age**  **Race / Ethnicity**  **Sexual Behaviors**
I'm hiv+ and healthy poz or poz friendly (undetectable) poz guy looking to outlive this disease

CONDOM USE
looking for other buddies into uninhibited bareback fun
I have poppers and condoms
Prefer raw, will use a condom if you'd rather play safe
DRUG USE

Kickin back with my friend tina, some chronic 420, poppers, vodka/beer too...This is for now!

poppers and 420 friendly

(meth, marijuana, amyl nitrate, alcohol)

LOCATION

(San Francisco)
(Berkeley) (Glen Park)
(Philz Coffee)
(El Tapatio Nightclub Mission Street)
(Downtown / Civic / Van Ness)
AGE & RACE/ETHNICITY
Chill, handsome MWM Biz/
Family Man, DDF

I am 32 y/o black masculine

PARTNER PREFERENCE
ts tg m4t (transgender)

My preference is
CAUCASIAN/White (No Latino)

looking for hot 3some (group sex)
Public Health Intervention Tools

PARTNER NOTIFICATION

FORMS & SURVEYS
## Public Health Intervention Tools

**PARTNER NOTIFICATION FORMS & SURVEYS**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Race/Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Public Health Intervention Tools

#### PARTNER NOTIFICATION FORMS & SURVEYS

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Race/Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Had sex w/male?</td>
<td>yes □</td>
<td>no □</td>
</tr>
<tr>
<td>Had sex w/transgender?</td>
<td>yes □</td>
<td>no □</td>
</tr>
<tr>
<td>Had sex w/o condom</td>
<td>yes □</td>
<td>no □</td>
</tr>
<tr>
<td>Met partners via internet?</td>
<td>yes □</td>
<td>no □</td>
</tr>
<tr>
<td>Patient’s HIV status?</td>
<td>pos □</td>
<td>neg □</td>
</tr>
<tr>
<td>Non-injection drug usage (Note drugs:)</td>
<td>yes □</td>
<td>no □</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Sex</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Had sex w/male?</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Had sex w/transgender?</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Had sex w/o condom</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Met partners via internet?</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Patient’s HIV status?</td>
<td></td>
<td>pos</td>
</tr>
<tr>
<td>Non-injection drug usage (Note drugs:)</td>
<td></td>
<td>yes</td>
</tr>
</tbody>
</table>

Preferred partner race/ethnicity?
Where did you meet for sex?
II. Ad Parsing Pipeline
II. Ad Parsing Pipeline
...think of these as sentinel surveillance sites
Craigslist Data Set

2.5 years (2009 - 2012)

136 million ads (7.8B tokens)

32.5 million MSM ads (2.6B tokens)

55 GB of text

700 annotated ads
Surveillance Pipeline
Ad Information Extraction Subtasks

- Associate ads with geographic locations
- Identify MSM-targeted ads
- Extract author race/ethnicity and age information
- Label ads with high-risk sexual behavior search terms
Surveillance Pipeline

Ad Information Extraction Subtasks

- Associate ads with geographic locations
- Identify MSM-targeted ads
- Extract author race/ethnicity and age information
- Label ads with high-risk sexual behavior search terms

1. Relation Mining
2. Regular Expressions

1. Learning Terminology
2. Topic Modeling
3. Regular Expressions

Named Entity Recognition

Regular Expressions

Learning Terminology

Topic Modeling

Regular Expressions
Extracting Geographic Named Entities

“Geographic Entity Linking in Online Classified Ads” (in-review)

~ 75% Accuracy (for this work)
mean geocoding error ~15 miles
Age Extraction

Identifying age is actually pretty easy!

\( \text{\( \backslash d+ \)\( \backslash s* (yrs|yr|y/o|yo|years \ old) + \)\} } \)

The above regular expression results in a F1 score of 0.97
# How do Authors Discuss Race / Ethnicity on Craigslist?

<table>
<thead>
<tr>
<th>white</th>
<th>black</th>
<th>Japanese</th>
<th>costarican</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWM</td>
<td>blk</td>
<td>Korean</td>
<td>latinos</td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td>Chinese</td>
<td>mex</td>
</tr>
<tr>
<td>Austrian</td>
<td></td>
<td>Asian</td>
<td>hispanic</td>
</tr>
<tr>
<td>MWM</td>
<td></td>
<td></td>
<td>latin</td>
</tr>
<tr>
<td>Irish</td>
<td></td>
<td></td>
<td>dominican</td>
</tr>
<tr>
<td>WM</td>
<td></td>
<td></td>
<td>Puerto Rican</td>
</tr>
<tr>
<td>Caucasian</td>
<td></td>
<td>Asian</td>
<td>Hispanic/Latino</td>
</tr>
</tbody>
</table>

(just a small sample of the diversity of terms)
Race/Ethnicity Extraction

Race/ethnicity is more challenging

Two-stage classifier

CRF Labeler

13 lexical & syntactic features
label ∈ \{PARTNER, AUTHOR, NONE\}

Rule-based Classifier

Assign ad (author) to class using the final label set + original features.
CRF Labeling Features

1. Stemmed term
2. Race/Ethnicity Cluster
3. Digit
4. Craigslist Metadata
5. First Mention
6. In-list
7. Pluralization
8. Partner Preference
9. Punctuation
10. Slash
11. Left Verb Argument
12. Right Verb Argument
13. PRP + Being Verb

13 Lexical & Syntactic Features

(lexicons are built using training data and evaluated using cross-fold validation)
Classifier Output Example

First Mention: Predicts Author(Black) ✗
CRF-First: Predicts Author(Black) ✗
CRF-All: Predicts Author(Biracial) ✓
## Race/Ethnicity Performance

<table>
<thead>
<tr>
<th>Class</th>
<th>Ad N</th>
<th>F1 Score</th>
<th>vs. Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biracial</td>
<td>14</td>
<td>0.63</td>
<td>+ 159%</td>
</tr>
<tr>
<td>Hispanic / Latino</td>
<td>32</td>
<td>0.80</td>
<td>+ 3.1%</td>
</tr>
<tr>
<td>Black</td>
<td>27</td>
<td>0.85</td>
<td>+ 18.3%</td>
</tr>
<tr>
<td>Asian</td>
<td>19</td>
<td>0.91</td>
<td>+ 12.0%</td>
</tr>
<tr>
<td>Caucasian</td>
<td>136</td>
<td>0.92</td>
<td>+ 6.8%</td>
</tr>
<tr>
<td>Hawaiian / Pacific Islander</td>
<td>3</td>
<td>0.50</td>
<td>+ 85.8%</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>299</td>
<td>0.93</td>
<td>+ 2.2%</td>
</tr>
</tbody>
</table>

CRF-All improves on baseline (First Mention)
Calculating Prevalence Rates

Calculate a prevalence rate at a given geographic unit for each demographic variable of interest.

CBSA for Los Angeles, CA

\[
Prev = \frac{\text{Hispanic / Latino MSM Ads}}{\text{All MSM Ads}} = \frac{142,497}{1,610,742}
\]

\[
Weight = \frac{\text{All Ads}}{\text{All Ads}} = 5,593,782
\]

(see paper for equations)
III. Demographic Results
III. Demographic Results
Census Linear Regression

Metropolitan Core Based Statistical Areas (CBSA)

Asian
r-square 0.86
coefficient / 95% CI
0.65 [0.63, 0.68]

Hispanic/Latino
r-square 0.91
coefficient / 95% CI
0.81 [0.78, 0.84]

Black
r-square 0.71
coefficient / 95% CI
0.36 [0.34, 0.39]
Census Linear Regression

Metropolitan Core Based Statistical Areas (CBSA)

Biracial
r-square 0.43
coefficient / 95% CI
0.82 [0.72, 0.92]

Native Hawaiian / Pacific Islander
r-square 0.42
coefficient / 95% CI
0.4 [0.35, 0.45]

Caucasian
r-square 0.04
coefficient / 95% CI
-0.24 [-0.36, -0.13]
Caucasian Disclosure Rates

Less need to reveal race in very homogenous locations!

r-square 0.75
coefficient / 95% CI
0.15 [0.14, 0.16]
Craigslist Age Distributions
Census Hispanic / Latino
Online Health Surveillance

Internet surveillance can provide insight into communities that are difficult to observe through entirely offline means.

Algorithmic surveillance can provide near-real time supplemental information for formulating public health interventions.
IV. Future Work
IV. Future Work
Ultimate Goal: Forecast STD Outbreaks

Are sexual health behaviors predictive of STD outbreaks?

How can we mine behaviors in a semi-supervised fashion?
Related Publications

**Using Online Classified Ads to Identify the Geographic Footprints of Anonymous, Casual Sex-seeking Individuals.**
Jason A. Fries, Alberto M. Segre, Philip M. Polgreen
2012 ASE/IEEE Forth International Conference on Social Computing (September 2012)

**Using Craigslist Messages for Syphilis Surveillance**
Jason A. Fries, Alberto M. Segre, Linnea Polgreen, Philip M. Polgreen
International Meeting on Emerging Diseases and Surveillance (February 2011)

**The Use of Craigslist Posts for Risk Behavior and STI Surveillance**
Jason A. Fries, Alberto M. Segre, Linnea Polgreen, Philip M. Polgreen
9th Annual Conference of the International Society for Disease Surveillance (December 2010)
http://homepage.cs.uiowa.edu/~jfries/projects/craigslist.html

Interactive maps of race/ethnicity rates by county
Questions?