

---

Theses and Dissertations

---

Summer 2015

# Modeling words for online sexual behavior surveillance and clinical text information extraction

Jason Alan Fries  
*University of Iowa*

Copyright 2015 Jason Alan Fries

This dissertation is available at Iowa Research Online: <https://ir.uiowa.edu/etd/2076>

---

## Recommended Citation

Fries, Jason Alan. "Modeling words for online sexual behavior surveillance and clinical text information extraction." PhD (Doctor of Philosophy) thesis, University of Iowa, 2015.  
<https://ir.uiowa.edu/etd/2076>.

---

Follow this and additional works at: <https://ir.uiowa.edu/etd>



Part of the [Computer Sciences Commons](#)

MODELING WORDS FOR ONLINE SEXUAL BEHAVIOR SURVEILLANCE  
AND CLINICAL TEXT INFORMATION EXTRACTION

by

Jason Alan Fries

A thesis submitted in partial fulfillment of the  
requirements for the Doctor of Philosophy  
degree in Computer Science  
in the Graduate College of  
The University of Iowa

August 2015

Thesis Supervisor: Professor Alberto Maria Segre

Graduate College  
The University of Iowa  
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

PH.D. THESIS

---

This is to certify that the Ph.D. thesis of

Jason Alan Fries

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the August 2015 graduation.

Thesis committee: \_\_\_\_\_  
Alberto M. Segre, Thesis Supervisor

\_\_\_\_\_  
Philip M. Polgreen

\_\_\_\_\_  
Ted Herman

\_\_\_\_\_  
Padmini Srinivasan

\_\_\_\_\_  
Nick Street

## ACKNOWLEDGEMENTS

First, I would like to thank Professor Alberto Segre and Dr. Phil Polgreen for all of their advice and assistance over the years; both spent untold hours securing data access and funding resources for my research. Without their help this thesis would not exist. On the same note, my eternal gratitude to Professor Ted Herman, who nominated me for an undergraduate research project and started me on the unexpected road to graduate school. Many thanks to Professors Sriram Pemmaraju and Padmini Srinivasan for providing their expertise at crucial moments. My gratitude to Professor Nick Street for his pointed and invaluable critiques along the road to my defense, as well as his eagle-eyed, detail-oriented ability to catch errors involving hyphenated adjectives.

Thank you to all the past and present members of the CompEpi research group for your feedback and insights. Our research conference adventures will always stay with me: ice skating in front of the Vienna Rathaus with Don, Chris, and Geoff; catching a terrifying late night cab in Rome with Jim, David, and Thomas; triumphantly winning an empty manure bag at the Deadwood pub quiz with Mauricio, Valerie, Lucio, and Pat. These colleagues and friends made graduate school bearable during stressful times.

For all the people and creatures of Deweyville, defacto city-state of Iowa City, I doubt I will ever stumble across a neighborhood and community as fun or as welcoming. I never felt much attachment to the apartment of my childhood, with its low-slung view of the interstate and a mouldering Happy Chef. I'm glad I traded those walls for a better view and a place I can actually call home.

And finally, my lifelong thanks to Katherine Massoth for providing a shoulder to lean on in frantic times, making coffee and cookie runs at a moment's notice, and generally being supportive of the long hours I spent silently staring into my computer screen.

## ABSTRACT

How do we model the meaning of words? In domains like information retrieval, words have classically been modeled as discrete entities using 1-of- $n$  encoding, a representation that elides most of a word’s syntactic and semantic structure. Recent research, however, has begun exploring more robust representations called *word embeddings*. Embeddings model words as a parameterized function mapping into an  $n$ -dimensional continuous space and implicitly encode a number of interesting semantic and syntactic properties. This dissertation examines two application areas where existing, state-of-the-art terminology modeling improves the task of information extraction (IE) – the process of transforming unstructured data into structured form. We show that a large amount of word meaning can be learned directly from very large document collections.

First, we explore the feasibility of mining sexual health behavior data directly from the unstructured text of online “hookup” requests. The Internet has fundamentally changed how individuals locate sexual partners. The rise of dating websites, location-aware smartphone apps like Grindr and Tinder that facilitate casual sexual encounters (“hookups”), as well as changing trends in sexual health practices all speak to the shifting cultural dynamics surrounding sex in the digital age. These shifts also coincide with an increase in the incidence rate of sexually transmitted infections (STIs) in subpopulations such as young adults, racial and ethnic minorities, and men who have sex with men (MSM). The reasons for these increases and their possible connections to Internet cultural dynamics are not completely understood. What is apparent, however, is that sexual encounters negotiated online complicate

many traditional public health intervention strategies such as contact tracing and partner notification. These circumstances underline the need to examine online sexual communities using computational tools and techniques – as is done with other social networks – to provide new insight and direction for public health surveillance and intervention programs.

One of the central challenges in this task is constructing lexical resources that reflect how people actually discuss and negotiate sex online. Using a 2.5-year collection of over 130 million Craigslist ads (a large venue for MSM casual sexual encounters), we discuss computational methods for automatically learning terminology characterizing risk behaviors in the MSM community. These approaches range from keyword-based dictionaries and topic modeling to semi-supervised methods using word embeddings for query expansion and sequence labeling. These methods allow us to gather information similar (in part) to the types of questions asked in public health risk assessment surveys, but automatically aggregated directly from communities of interest, in near real-time, and at geographically high-resolution. We then address the methodological limitations of this work, as well as the fundamental validation challenges posed by the lack of large-scale sexual behavior survey data and limited availability of STI surveillance data.

Finally, leveraging work on terminology modeling in Craigslist, we present new research exploring representation learning using 7 years of University of Iowa Hospitals and Clinics (UIHC) clinical notes. Using medication names as an example, we show that modeling a low-dimensional representation of a medication’s neighboring words, i.e., a word embedding, encodes a large amount of non-obvious semantic information. Embeddings, for example, implicitly capture a large degree of the hierarchical structure of drug families as

well as encode relational attributes of words, such as generic and brand names of medications. These representations – learned in a completely unsupervised fashion – can then be used as features in other machine learning tasks. We show that incorporating clinical word embeddings in a benchmark classification task of medication labeling leads to a 5.4% increase in F1-score over a baseline of random initialization and a 1.9% over just using non-UIHC training data. This research suggests clinical word embeddings could be shared for use in other institutions and other IE tasks.



## PUBLIC ABSTRACT

How do we model the meaning of words? In domains like information retrieval, words have classically been represented as discrete entities. Recent research, however, has begun exploring more robust, continuous space representations of words which implicitly encode rich syntactic and semantic information. This dissertation examines two application areas where existing, state-of-the-art terminology modeling improves the task of information extraction (IE) – the process of transforming unstructured data into structured form. We show that a large amount of word meaning can be learned directly from very large document collections.

First, we explore the feasibility of mining sexual health behavior data directly from the unstructured text of online “hookup” requests. One of the central challenges in this task is constructing lexical resources that reflect how people actually discuss and negotiate sex online. Using a 2.5-year collection of 130 million Craigslist ads, we discuss text mining and machine learning methods for modeling subpopulation risk behaviors, such as illegal drug use, unprotected sex, and HIV status. These methods allow us to reframe many public health risk assessment survey questions as text classification problems, automatically aggregating surveillance data directly from communities of interest, in near real-time, and at high geographic resolution. We then address the methodological limitations of this work, as well as the fundamental validation challenges posed by the lack of large-scale sexual behavior survey data.

Finally, leveraging work on terminology modeling in Craigslist, we present new re-

search exploring representation learning using 7 years of University of Iowa Hospitals and Clinics (UIHC) clinical notes. Using medication names as an example, we show that modeling a low-dimensional representation of a medication’s neighboring words, i.e., a word embedding, encodes a large amount of non-obvious semantic information. Embeddings, for example, implicitly capture a large degree of the hierarchical structure of drug families as well as encode relational attributes of words, such as generic and brand names of medications. These representations – learned in a completely unsupervised fashion – can then be used as features in other machine learning tasks. We show that incorporating clinical word embeddings in a benchmark classification task of medication labeling leads to a 5.4% increase in F1-score over a baseline of random initialization and a 1.9% over just using non-UIHC training data. This research suggests clinical word embeddings could be shared for use in other institutions and other IE tasks.

## TABLE OF CONTENTS

LIST OF TABLES . . . . .	xii
LIST OF FIGURES . . . . .	xiv
CHAPTER	
1 INTRODUCTION . . . . .	1
1.1 Community Sexual Behavior Surveillance . . . . .	1
1.1.1 Motivation . . . . .	1
1.1.2 Background . . . . .	4
1.2 Modeling Terminology and Relationships in Clinical Text . . . . .	6
2 BACKGROUND . . . . .	8
2.1 Information Extraction (IE) . . . . .	8
2.1.1 Pattern-based Approaches . . . . .	8
2.1.2 Supervised Learning . . . . .	10
2.1.3 Unsupervised or Weakly-supervised Learning . . . . .	11
2.2 Applied IE . . . . .	13
2.2.1 Clinical Text . . . . .	13
2.2.1.1 General Approaches . . . . .	15
2.2.1.2 Medication Extraction . . . . .	17
2.3 Distributional Semantic Models . . . . .	19
2.3.1 Matrix Factorization Approaches . . . . .	21
2.3.2 Hierarchical clustering . . . . .	22
2.3.3 Neural Language Models (NLM) . . . . .	23
2.3.4 Word Embeddings . . . . .	24
2.3.5 Recurrent Neural Networks . . . . .	26
2.3.6 Evaluating Distributional Semantic Models . . . . .	28
3 DATA SETS: DESCRIPTIONS AND SUMMARY STATISTICS . . . . .	29
3.1 Sexual Behavior Data . . . . .	29
3.1.1 Craigslist . . . . .	29
3.2 Clinical Text . . . . .	37
3.2.1 i2b2 Medication Extraction . . . . .	37
3.2.2 University of Iowa Hospital and Clinics . . . . .	39
4 BUILDING A SURVEILLANCE PIPELINE . . . . .	40

4.1	Surveillance Pipeline NLP Subtasks . . . . .	40
4.2	Geographic Named Entity Normalization . . . . .	41
4.2.1	Materials/Methods . . . . .	43
4.2.1.1	Annotation Corpora & Toponym Knowledge Base . . . . .	43
4.2.1.2	Named Entity Recognition and Linking . . . . .	46
4.2.1.3	Evaluation Measures . . . . .	53
4.2.2	Results . . . . .	55
4.2.3	Discussion . . . . .	59
4.3	Authorship Attribution . . . . .	61
4.3.1	Methods . . . . .	63
4.3.1.1	Near-Duplicate Detection . . . . .	63
4.3.1.2	Constructing Footprint Travel Graphs . . . . .	68
4.3.2	Results . . . . .	69
4.3.3	Discussion . . . . .	73
4.4	Demographics: Extracting Age & Race/Ethnicity . . . . .	75
4.4.1	Methods . . . . .	76
4.4.1.1	Extracting Age & Race/Ethnicity . . . . .	76
4.4.1.2	Calculating Demographic Rates . . . . .	80
4.4.2	Results . . . . .	81
4.5	Conclusion . . . . .	85
5	COMMUNITY SEXUAL BEHAVIOR SURVEILLANCE . . . . .	88
5.1	Modeling Sexual Behaviors in Text . . . . .	88
5.1.1	Dictionary/Search Term-based Methods . . . . .	89
5.1.2	Word/Phrase Learned Representations . . . . .	90
5.1.3	Topic Modeling . . . . .	96
5.2	Geo-temporal Validation . . . . .	98
5.3	Links between Sexual Behaviors and Disease . . . . .	100
5.3.1	Materials/Methods . . . . .	100
5.3.1.1	CDC County-level Disease Data . . . . .	100
5.3.1.2	Computing Topic Prevalence Rates . . . . .	101
5.3.2	Results . . . . .	101
5.3.2.1	State-level Disease Correlations . . . . .	101
5.3.2.2	Self-reported HIV Status . . . . .	102
5.3.2.3	Syphilis Forecasting in California . . . . .	103
5.3.3	Discussion . . . . .	103
6	SEMANTIC MODELING OF MEDICATIONS IN CLINICAL TEXT . . . . .	107
6.1	Introduction . . . . .	107
6.2	Methods & Materials . . . . .	109
6.2.1	Corpora and Preprocessing . . . . .	109
6.2.2	Medication Data Set Creation . . . . .	111

6.2.3	Training Word Embeddings . . . . .	112
6.2.4	Evaluating the Quality of Embeddings . . . . .	112
6.2.5	Sequence Labeling . . . . .	114
6.3	Results . . . . .	117
6.3.1	Medication AHFS-type Classifier . . . . .	117
6.3.2	Relation Encoding . . . . .	121
6.3.3	Sequence Labeling Performance . . . . .	122
6.4	Discussion . . . . .	124
6.4.1	Limitations . . . . .	126
7	CONCLUSION . . . . .	132
7.1	Future Work . . . . .	135
	REFERENCES . . . . .	137

## LIST OF TABLES

Table

3.1	Craigslist Corpora Summary . . . . .	30
3.2	Annotation Corpus Toponym Distributions . . . . .	34
3.3	i2b2 Annotation Entity Summary . . . . .	38
4.1	GeoExact Toponym Linking Performance . . . . .	55
4.2	Entity Linking Miles Error . . . . .	57
4.3	Ad Cluster Sizes (U.S. and California) . . . . .	70
4.4	Graph Summary Statistics . . . . .	73
4.5	Author Race/Ethnicity Classification Performance Measures . . . . .	82
4.6	Craigslist Race/Ethnicity vs. 2010 Census . . . . .	83
5.1	Craigslist Race/Ethnicity Terms: 10-nearest Neighbors . . . . .	91
5.2	Craigslist Drug Slang Terms: 10-nearest Neighbors . . . . .	92
5.3	Drug Euphemisms “Snow”: 10-nearest Neighbors . . . . .	94
5.4	Negated Sentences: 10-nearest Neighbors . . . . .	95
6.1	UIHC Clinical Document Collections . . . . .	111
6.2	Sample Semantic Medication Analogy Questions . . . . .	114
6.3	Clinical Text Medication Terms: 10-nearest Neighbors (UIHC-ALL) . . . . .	117
6.4	Medication Classifier ( $\geq 5$ Instances-per-class) . . . . .	118
6.5	Medication Classifier: AHFS Tier 1 ( $\geq 10$ Instances-per-class) . . . . .	119
6.6	Antibacterial Classifier ( $\geq 5$ Instances-per-class) . . . . .	120

6.7	Semantic Evaluation: Medication Brand Name Prediction (Progress Notes) . . .	121
6.8	i2b2 Medication Labeling: All Tags . . . . .	122
6.9	i2b2 Labeling: Drug, Dose, Mode, Frequency . . . . .	123
6.10	i2b2 Labeling: Drug . . . . .	123

## LIST OF FIGURES

Figure		
2.1	Simple Elman-type Recurrent Neural Network. $\mathbf{U}$ is the input $\times$ hidden layer weight matrix. $\mathbf{V}$ is the context layer $\times$ hidden layer matrix, and $\mathbf{W}$ is the output weight matrix. Dotted lines indicate recurrent edge weights. . . . .	26
3.1	Log-linear scatterplot of mean daily post rates for MSM ads (solid black line) calculated over 723 days of Craigslist posts. Representative sites at 5-number summary percentiles are listed on the x-axis. Red dotted lines indicate $\pm 1$ SD for MSM ads. 50% of sites (205/412) have daily MSM post rates between 7 and 62 ads per day. The top 5 MSM sites, <i>losangeles</i> , <i>newyork</i> , <i>sfbay</i> , <i>chicago</i> , <i>dallas</i> , have daily rates of 1716, 1654, 1648, 1648, 1286 respectively. Mean daily rates for used furniture ads (gray points) are provided for comparison; furniture ads are on average higher than MSM ad rates (mean difference of 27 ads); 19% of sites have mean MSM rates higher than furniture ads. Nationwide, the mean rate of MSM ads was 38,193 (SD 5,998) posts per day. . . . .	32
3.2	Example GOLD annotations from an m4m <i>sfbay</i> ad. Race/ethnicity mentions (in green) are assigned to a racial group (e.g., Caucasian) and tagged as referencing 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. . . . .	33
4.1	Sample <i>sfbay</i> Craigslist ad subject lines. Each subject line may include an optional location tag, enclosed in parentheses (shown here in boldface), of up to 40 characters in length. Tags contain no formatting constraints beyond length and may refer to arbitrary geographic entities (cities, buildings, streets, etc.) or no entities at all. . . . .	45
4.2	Classified ad toponym entity labeling pipeline. The above flowchart uses the <i>iowacity</i> site location tag “IC/CR/NL” as an example to illustrate the entity linking pipeline. Location tags are first normalized and segmented into entity spans, with site specific lexicons used to generate candidate entity sets (steps 1 and 2). The final set of entities is then disambiguated using a combination U.S. Census data and a set of rule-based scoring functions. . . . .	48



4.3	Log-log scatterplot of the frequency of all location entity tags vs. the corresponding area of that tag in square miles. Note the high geographic resolution of entity locations, with 97% of all tags correspond to areas smaller than 640 sq mi – the median area of a U.S. county using 2010 census data. Overall tags reflect the high spatial resolution of ad entities: 71% of all tags refer to areas smaller than the city of San Francisco, California (47 sq mi); 50% of all tags refer to a geographic boundary with an area less than 15 sq mi; and 10% of tags refer to areas smaller than the median size of a U.S. neighborhood (0.65 sq mi). Vertical dotted lines indicate the size of named locations to provide a relative sense of scale. . . . .	58
4.4	True positive rates for hash fingerprint feature selection (using unigram and bigram feature sets) and evaluated at Hamming distances $k = [1..10]$ . Note that unigram features perform poorly overall, while bigrams perform well at distances $1 \leq k \leq 7$ . . . . .	64
4.5	Histogram of intra-cluster Jaccard similarities, before removing spurious near-duplicate clusters. 48% of all matches consist of exact duplicate ads (the peak on the far right of the plot). The majority of detected near-duplicates are true matches, with only 12% of all detected clusters falling below our 0.6 similarity threshold. . . . .	66
4.6	Performance measures for partial cluster detection using the near-duplicate identification approach to linking anonymous authorship in Craigslist ads. In general, our near-duplicate approach detects a small subset of ads that are written by the same author, but suffers from poor recall when identifying all of an author’s ads. Recall performance suffers as the number of ads written by a single author increases. . . . .	69
4.7	Empirical Cumulative Distribution Function (left) and histogram (right) of hop distances, calculated as the haversine distance (in miles) between location tags of successive ads within a given ad cluster. 86% of all hop distances involve distances under 50 miles while 4% involve distances over 250 miles. Note as the haversine distance is the minimal great-circle distance between 2 points on a sphere, these values underestimate the actual number of miles traveled. . . . .	71
4.8	(Top, red) Average hourly entropy of location tags counts, binned by U.S. state, and (bottom, blue) average posting counts by hour for all U.S. clusters. Black dashed lines divide days and grey dotted lines correspond to noon. Higher entropy corresponds to more tag variability across bins, indicating a greater willingness to travel. Observe how the periodicity of tag entropy largely corresponds to posting frequency, with slightly more location variability observed Friday and Saturday evenings and Saturday and Sunday mornings. . . . .	72

4.9	Graph of all authorship-linked ad posts crossing county lines (left) and the same graphs corresponding adjacency matrix with Cuthill-McKee [40] ordering of nodes (right). The adjacent matrix shows clustering between major counties in California (e.g, Los Angeles County, Orange County, etc. . . . .	74
4.10	Visualization of the Craigslist <i>sfbay</i> subgraph (V=591, E=7916), showing all edges with weights > 25. Node color represents outdegree and edge color represents weight (blue=low and red=high). Observe how geographic clustering is clearly visible with respect to the spatial distribution of city locations. . . . .	75
4.11	Excerpt of labeled output. Each term in a sequence is assigned a <code>label</code> $\in$ { <code>AUTHOR</code> (A), <code>PARTNER</code> (P), <code>NONE</code> (N)} predicted based on 13 features, using conditional random fields. Given this labeling, both <i>First Mention</i> and <i>CRF-First</i> would incorrectly classify this ad’s author as Black. <i>First Mention</i> fails to disambiguate the first usage (as hair color) of “blk” while <i>CRF-First</i> only considers the first race term labeled as <code>AUTHOR</code> . <i>CRF-All</i> , which considers all <code>AUTHOR</code> labels, would correctly predict Biracial. . . . .	77
4.12	Scatter plot of log-log regression results for 2010 census race/ethnicity percentages (x-axis, independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (y-axis, dependent variable). Plots are for all CBSA metropolitan geographic boundaries containing at least 1000 ads. For the Caucasian plot (lower left corner) the percentage of race disclosures begins to drop precipitously in the interval 64%-100% (1.8-1.2), suggesting that as the population grows more homogeneously Caucasian there is less need to mention race in ads. . . . .	84
4.13	Scatter plot of 2010 Census geographic Shannon entropy (y-axis), measured across race/ethnicity categories vs. the percentage of ads with undisclosed race/ethnicity (x-axis). The shape of each point indicates the majority racial/ethnic group in that geographic boundary. Note how as a population grows more homogenous and less evenly distributed across types (i.e., lower entropy), the rate of undisclosed race/ethnicity ads increases. . . . .	85
4.14	Percentage of MSM ads out of all sampled Craigslist ads. In some states such as Louisiana, MSM ads are the majority of sampled ads, while in others such as Washington and Oregon, most ads are commercial ads (i.e., an ad from the categories { <code>legal services</code> , <code>appliances</code> , <code>pets</code> , <code>furniture</code> , and <code>parking</code> }).	86
4.15	Hispanic/Latino Craigslist county-level percentages as determined by CRF-All (left) vs. 2010 Census data (right). . . . .	86
5.1	Word embeddings <i>k</i> -nearest-neighbors for “bb” (left) and “favors” (right), where “bb” is slang for unprotected sex and “favors” is slang for illicit drugs. . . . .	92

5.2	2D stochastic neighbor embedding of Craigslist word embeddings. Point labels are determined through annotations in GOLD and geographic entities identified by TopoLinker. Note how much of the space contains geographic terms. Geographic sub-clusters frequently correspond to specific usage classes (e.g., university acronyms) or represent site-specific clustering. Stochastic neighbor embedding generated using Barnes-Hut T-SNE [147]. . . . .	93
5.3	Some example topics generated from Craigslist MSM ads using LDA ( $k=125$ ). Word size corresponds to per-topic probability. . . . .	97
5.4	Spatial distribution of “Party and Play” (PNP) topic (left) and “Married/Bi/Straight” (SEXUAL IDENTITY) topic (right). PNP is higher prevalence on the west coast while SEXUAL IDENTITY is generally higher in the south and east coast, with clusters in Idaho, Nevada, Utah, and Wyoming. . . . .	97
5.5	Two time series models for predicting NOAA-reported snowfall using daily, standardized Craigslist snow-related ad rates. On the left: fitted values plot of predicted snowfall for Tuolumne County California; on the right: Chicago, Illinois (Cook, DuPage, Kane, Lake, and Will counties). The top models estimate snowfall using only NOAA data while the bottom models incorporate Craigslist ad content and do a better job overall of predicting snowfall. The significant temporal association between Craigslist ads and snowfall as well as the reduction in error when using ads as a predictor of snowfall infers the usefulness of Craigslist data as a predictor of phenomena in the physical world. . . . .	98
5.6	U.S. State-level analysis of 2011 Caucasian PNP (party and play) behavior rates vs. 2011 Caucasian PS syphilis (left) and HIV (right) incidence rates using a weighted log-log OLS regression. Disease incidence rates are on the x-axis and PNP rates are on the y-axis. All results are statistically significant ( $p < 0.05$ ). HIV incidence vs. PNP rates: adjusted r-squared 0.38, coefficient 0.43, 95% CI [0.26,0.59]; and PS syphilis incidence vs. PNP rates: adjusted r-squared 0.55, coefficient 0.60, 95% CI [0.41,0.78]. 2009 and 2010 HIV incidence were similarly correlated with PNP rates, as was 2010 PS syphilis. . . . .	103
5.7	Log-log scale scatter plots of the OLS regression model of California MSM HIV+ (left) and BB (right) ontologies vs. combined HIV/AIDS prevalence rates. In these plots the size of the circle corresponds to the population of the county data point. Note the visible clusters between high population and low population counties. . . . .	104

6.1	T-SNE [147] visualization of medication embeddings, clustered using the AHFS Tier 1 classification system ( $\geq 10$ instances). The number indicates cluster id and color indicates one of the top 7 clusters in terms of set size. Note how <i>Anti-infectives</i> (red) and <i>Central Nervous System Agents</i> (green) are comprised of multiple sub-clusters. These in fact frequently correspond to the next hierarchical tier of medications (see Figures 6.2 and 6.3). . . . .	128
6.2	Medication clustering for AHFS Tier 2 <i>Central Nervous System Agents</i> (green colored points in Figure 6.1). Note the same sub-clustering observed at Tier 1. Numbered points with very close overlap typically correspond to brand name/generic name versions of medication mentions. . . . .	129
6.3	AHFS Tier 3 clustering for <i>Psychotherapeutic Agents</i> and <i>Anxiolytics, Sedatives, and Hypnotics</i> (orange and dark purple respectively in Figure 6.2). Note how much of hierarchical nature of the AHFS medication classification system is implicitly captured by the word embedding training process. . . . .	130
6.4	Generic/Brand name relation encoding accuracy. Performance gains using larger embedding spaces begin diminishing after 300. Progress notes performed best in capturing generic/brand name relationships, even though that corpus is roughly 27% of the size of the set of all notes. Note that multiple training epochs greatly improved the performance of the discharge summaries corpus, making it competitive to the larger corpora, even though it is substantially smaller (4% the size of all notes). . . . .	131

## CHAPTER 1 INTRODUCTION

### 1.1 Community Sexual Behavior Surveillance

#### 1.1.1 Motivation

The Internet has fundamentally changed how individuals locate sexual partners. The rise of dating websites, location-aware smartphone apps like Grindr and Tinder that facilitate casual sexual encounters (“hookups”), as well as changing trends in sexual health practices all speak to the shifting cultural dynamics surrounding sex in the digital age. These shifts coincide with an increase in the incidence rate of sexually transmitted infections (STIs), especially in certain populations such as young adults, racial and ethnic minorities, and men who have sex with men (MSM). The reasons for these increases and their possible connections to Internet cultural dynamics are not completely understood. What is apparent, however, is that sexual encounters negotiated online complicate many traditional public health intervention strategies like contact tracing and partner notification. The relative ease of locating partners, partner anonymity, and the possible geographic breadth of encounters weaken the efficacy of these traditional public health tools. These circumstances underline the need to examine online sexual communities using computational tools, as is done with other social networks, and explore how such data can be incorporated into intervention and surveillance programs.

These developments complicate interventions within the *men who have sex with men* (MSM) population, which has experienced 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 [30]. These increases coincided with a decrease in safer sex practices [63, 105, 124]. In a national survey, Bull et al. found 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 [88]. A meta-analysis estimates that 40% of MSM meet their sex partners online [77] and the Internet has been identified as the largest venue where MSM meet sexual partners [118]. Moreover, 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 [16].

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 [5], 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 (i.e., *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 [149]. 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 to negotiate sexual encounters 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 [21, 69].

This dissertation explores, in part, the feasibility of extracting sexual health behavior information directly from the unstructured text found in online “hookup” requests. One of the central challenges in this task is constructing lexical resources that reflect how people actually discuss and negotiate sex online. Using a 2.5-year collection of over 130 million Craigslist ads (a large venue for MSM casual sexual encounters), I discuss computational methods for automatically learning terminology characterizing risk behaviors in the MSM community, e.g., illegal drug use, unprotected sex. These approaches range from keyword-based dictionaries and topic modeling to semi-supervised methods using learned word representations (word embeddings) for query expansion and semantic indexing. In addition to these contributions, I also present machine learning and rule-based approaches for efficiently linking ads to geography locations, linking anonymous authorship across ads, and mining age and race information from Craigslist text. Together, these methods allow us to gather

information similar (in part) to the types of questions asked in risk assessment surveys, but automatically aggregated directly from communities of interest, in near real-time, and at high geographic resolution.

Finally I discuss the limitations of this work. These include technical limitations imposed by the data set itself (user anonymity, lack of exact geographic location, etc.) as well as more fundamental validation challenges posed by the limited availability of STI surveillance data and the relative lack of large-scale sexual behavior survey data. Despite these limitations, public health interventions will increasingly need to occur both on and offline, suggesting that automated tools for monitoring sexual behaviors in social media should be incorporated into public health intervention design.

### 1.1.2 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 [18]. 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 [47]. 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.

Additionally, some researchers theorize that STI epidemics have different transmission dynamics depending upon the spatial origin of an outbreak: epidemics originating from within core areas may be more difficult to bring under control than those originating from non-core areas [56]. This suggests that the ability to monitor sexual traffic between geographic clusters may provide additional insight into the progression of an epidemic. While the role of air travel in the spread of diseases such as influenza has been explored by many researchers [57,66] the impact that local travel has on the dynamics of STIs is less discussed in the literature. Large metropolitan areas, such as Los Angeles, involve some implicit level of local travel, but the degree to which sexual networks connect neighboring cities, or even geographic locations within cities themselves (beyond known clusters of socio-demographic disease correlates), is less understood. Travel across state and county lines all have potential impact on public health department surveillance and intervention strategies, which are ultimately bound by geographic and political boundaries.

The link between Craigslist and 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. [31]. 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 [27]. 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 [101]. Grov examined MSM ads posted in New York City, manually developing guidelines for annotating risk behaviors in ad text [58].

## 1.2 Modeling Terminology and Relationships in Clinical Text

The second component of this dissertation looks at semi-supervised methods for modeling medication terminology and relations directly from large collections of unlabeled clinical text, the unstructured text component of the electronic medical record (EMR). Biomedical domain experts invest considerable effort into building curated lexicons and ontologies of medical concepts, which provide a common language for encoding and exchanging data for billing and other purposes. These resources are frequently leveraged for clinical information extraction (IE) tasks like medication extraction – the process of identifying medication names and their corresponding attributes in text. Comprehensive IE is identified as a key requirement of clinical question answering systems and many significant research challenges remain as open problems. We present work examining how a core clinical IE task, sequence labeling, is improved by incorporating representations of words and phrases learned directly from large data sets.

This work examines a large collection of clinical text documents from the University of Iowa Hospitals and Clinics (UIHC), spanning 7 years and comprising over 8 million unique patient visits. We apply recent advances in generating word embeddings to model representations of medication names and phrases in clinical text. Embeddings model words as a parameterized function mapping into an  $n$ -dimensional vector space and encode a number of interesting semantic and syntactic properties of words. Using medication names as

an example, I show that a low-dimensional representation of a medication's context, i.e., its neighboring words, encodes a surprising amount of non-obvious semantic information. For example, embeddings implicitly capture a large degree of the hierarchical structure of drug families. Using UIHC embeddings as features in a KNN classifier, we can predict multiple classification levels for medications, achieving  $F_1$ -scores between 0.79 - 0.86. We also show that embeddings can encode relational properties such as generic and brand names of medications.

Finally, we demonstrate that these embeddings – learned in a completely unsupervised fashion – can then be used as features in other machine learning tasks. Incorporating clinical word embeddings in a benchmark classification task of medication labeling results in a 5.4% increase in  $F_1$ -score over a baseline of random initialization and a 1.9% over just using non-UIHC training data. This research suggests clinical word embeddings can function as empirically generated lexicons, modeling meaningful semantics and contextual features in an automated fashion. These representations could be shared for use in other institutions and for other IE tasks.

## CHAPTER 2 BACKGROUND

### 2.1 Information Extraction (IE)

*Information extraction* (IE) is the process of converting *unstructured data*, information without a pre-defined relational structure, into structured, logical form. Identifying relations in unstructured data such as text or multimedia allows systems to better reason over that data, enabling complex information retrieval tasks that require some degree of logical inference. Automatic summarization (generating condensed representations of documents), question answering tasks (responding to natural language queries with the correct answer rather than a list of relevant documents), and information retrieval in general all benefit from comprehensive IE. Unfortunately, comprehensive IE draws upon virtually all foundational tasks in *natural language processing* (NLP), including many challenging open problems: term tokenization, sentence boundary detection, part-of-speech tagging, shallow parsing (chunking), dependency parsing (deep parsing), named entity recognition, semantic role labeling, discourse modeling, etc. This section provides a brief historical overview of IE methodology and subtasks, specifically focusing on applications and open problems in the clinical text domain.

#### 2.1.1 Pattern-based Approaches

Early IE approaches were primarily rule-based or pattern matching systems. Riloff et al. looked at systems for tagging noun phrases and answer keys in text [115]. Riloff later expanded upon this with a pattern extraction system that used syntactic rule sets

to automatically identify patterns using a pre-classified corpus of relevant and irrelevant documents [114]. Brin proposed Dual Iterative Pattern Relation Extraction (DIPRE) which used a seed set of patterns to identify the relation of books and authorship in web documents. These relation patterns are populated by tuples of arguments expanding the pattern set using documents collected from the web [25]. Agichtein and Gravano built upon DIPRE to develop Snowball, a more generalized system designed to quickly assist the manual construction of seed patterns with a more sophisticated definition of relations [2]. Ravichandran and Hovy proposed a question answering system that used question/answer pairs to bootstrap additional patterns [111].

Most these approaches share a common design pattern which requires a small set of manually constructed regular expression patterns which are then bootstrapped to discover more patterns or rules. Learned patterns are then incorporated into the system to improve recall at some (possibly large) cost to precision. These approaches introduced the problem of *semantic drift*, as labeling errors compound with each learned instance and patterns “drift” from their intended meaning. While many approaches have been suggested for minimizing semantic drift (e.g., type checking relation arguments, combining multiple classifier approaches to minimize error, etc.) these approaches still present additional limitations when dealing with web-scale data sources or when used in out-of-domain applications [108, 117].

The overall weakness of these approaches is a reliance on identifying sets of surface form representations of concepts and using heuristics to match mentions to those sets. The brittleness of pattern-based systems, combined with the high cost of potentially enumerating seed patterns for any target concepts of interest, led to the use of more generalized machine

learning systems for IE.

### 2.1.2 Supervised Learning

Some of the brittleness associated with exclusively pattern-based IE methods are mitigated through the use of more adaptable machine learning algorithms. The current predominant IE paradigm is supervised learning. Here a small corpus of relevant linguistic features or other concepts of interest are labeled by hand and used to train a classifier. IE is a higher-level NLP task and is usually addressed via a pipeline of many individual NLP modules. IE can be viewed as a sequence labeling problem, where the classification task is to assign a tag to a given word (e.g., POS-tagging, Inside-Outside-Begin (IOB) chunking, etc.). The pipeline structure means the labels generated in a prior step can be used as features in upstream classification tasks. In this way, syntactic features of text like POS tags can be leveraged for high-level tagging, such as named entity recognition or semantic role labeling. State-of-the-art IE systems typically employ a combination of machine learning and manually engineered heuristics.

The choice of classification algorithm largely depends on the task, but usually it is a combination of support vector machines (SVMs) and a graphical model approach, e.g., Hidden Markov Models, Maximum Entropy Markov Models, or (increasingly) some variation of Conditional Random Fields (CRFs). Sarawagi found that semi-CRFs – CRFs that label segments of a sequence rather than individual elements – generally outperform standard CRFs at little additional computation cost [125].

### 2.1.3 Unsupervised or Weakly-supervised Learning

Much research has been devoted to robust ways of extracting relations in an unsupervised or *weakly-supervised* fashion, i.e., using existing structured data to automatically annotate unlabeled data. Mintz et al. presented a method requiring only weakly annotated training data (i.e., using an existing knowledge base to heuristically label instances and train an extractor), using Freebase data to extract sentences containing target entities and train a relation classifier [98]. Open Information Extraction (OIE) is an approach geared specifically towards mining web-scale corpora with arbitrarily large numbers of relations. This approach relies on using a small, unlabeled corpus of data to learn candidate relations, which are then used to extract relations from larger text collections. OIE bears conceptual similarity to early pattern-based approaches, but makes use of better machine learning and distributed computing resources. Banko et al. present the system *TextRunner* which uses a single, untagged corpus as input and generates a set of indexed relations [10]. Rather than using linguistic parsing tools to preprocess all data, *TextRunner* only uses high-cost parsers to train the preliminary relation extractor. This extractor builds sets of candidate binary relations  $t = (e_i, r_{i,j}, e_j)$  and labels them as positive given certain constraints on the syntactic structure shared by entities  $e_i$  and  $e_j$ . These positive instances are then used to extract relations from the corpus-at-large without utilizing a parser, ultimately outputting a probability-scored set of candidate relations. Here probability is a measure of redundancy as proposed by Downey et al., where normalized relations are created by removing non-essential verb/noun modifiers and the resulting counts of each relation are used to estimate the probability that an observed tuple is a correct instance of a relation [42]. Turney pro-

posed latent relational analysis (LRA) where a  $|V^n| \times |D|$  matrix was used to learn latent binary relations between  $n$ -tuples of words  $V^n$  in a set of documents  $D$  [140]. LRA proposes an  $n$ -tuple  $\times$  document matrix for detecting latent relations between argument sets. Riedel et al. recently presented a matrix factorization approach for extracting relations using *universal schemas* – the union of existing relational database schemas with OIE-style schemas acquired directly from language – to learn implication rules between relations [113, 156]. Representing schema as a *relation*  $\times$  *entity tuple* matrix, this approach uses matrix factorization techniques commonly found in recommender systems to predict the truth value of unseen matrix entries. In this way, they address data sparseness issues while simultaneously modeling potentially asymmetric implicature (entailment rules) between relations. One intriguing argument in their work is that methods should focus on predicting source data in lieu of directly modeling semantic equivalence. Clustering surface forms, for example, is a common way of modeling semantic equivalence, but one that often results in surface form clusters sharing only some vague notion of semantic equivalence and ultimately unable to provide a reliable mechanism for implicature.

Completely unsupervised approaches in IE are more rare and tend to focus on leveraging surface form clustering for use in other tasks. Poon and Domingos presented an unsupervised ontology induction approach that uses lambda calculus to represent ISA and IS-PART relations, parsing and transforming sentence dependency trees into quasi-logical forms whose subexpressions are defined using lambda forms [110]. These approaches are heavily dependent on the accuracy (or existence) of knowledge bases and upstream linguistic parsers like part-of-speech tagging, generating sentence dependency trees, named entity



recognition, etc.

## 2.2 Applied IE

### 2.2.1 Clinical Text

*Clinical text* is defined as the unstructured textual component of the electronic medical record (EMR). This text usually takes the form of notes associated with specific points of care during a patient’s stay in a healthcare facility. There are several different types of notes, often varying across facilities, but some common examples in hospitals include *progress notes* (short notes covering some time interval) and *discharge summaries* (a summary of a patient’s stay). Collectively these documents capture a patient’s *clinical narrative* – “a first person ‘story’ written by a clinician that describes a specific clinical event or situation” [87]. Clinical narrative is capable of capturing important aspects of patient care that are difficult to express in the structured charting or template mechanisms that make up most EMR systems.

Comprehensive information extraction is a key requirement for clinical question answering systems which entails comprehensive semantic parsing of the clinical narrative [4]. However clinical text possesses many properties that distinguish it from other domain text and requires special consideration when developing IE systems. Unlike text from biomedical domains for example, which usually consists of formal, copy-edited documents found in journals, books, etc., clinical text favors brevity and is often short and ungrammatical. Clinical text also makes use of frequent and overloaded abbreviations, acronyms or other institution-specific shorthand [91]. Liu et al. explored word sense disambiguation in biomedical doc-

uments and observed that 33.1% of abbreviations extracted from UMLS concepts map to multiple full forms [78,79]. Moreover, different types of notes are written for different purposes which impact their overall clarity. Discharge summaries usually include well formed summary data about a patient and are written to be read. Nursing progress notes, on the other hand, may function as an unstructured text container primarily documenting observations at arbitrary points in time. In some sense, clinical text orthographically resembles a hybrid of biomedical and microblog text (e.g., Twitter).

Performance in clinical text has typically lagged state-of-the-art NLP approaches largely due to the lack of accessible, standardized research corpora. Since clinical documents contain protected health information (PHI) and are subject to federal HIPAA regulations, they require special processing and permission to distribute. While there have been several task-specific datasets (e.g., medication extraction, de-identification, etc.) made available to the research community, general NLP tasks benefit from rich formal datasets such as the Brown corpus [51] as well as sophisticated linguistically annotated resources like Treebank and PropBank [67,86]. Clinical text has lacked such resources until very recently. Albright et al. created a corpus of multi-layered annotated clinical text, capturing many levels of linguistically standardized tagging and semantic labeling. Developed as part of a larger Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ), this document collection consists of 127k tokens of clinical text sampled from the Mayo Clinic medical record system. The performance of various supervised learning tasks like POS tagging to named recognition, improved by as much as 10% when trained on this clinical text corpus vs. non-clinical corpora [4].

Some empirical work suggests that clinical text contains many different *sublanguages*, a linguistic theory developed by Harris defining a specialized and constrained form of natural language [91]. A sublanguage imposes constraints on the types of sentences that can be constructed in a language and is usually observed in highly technical domains. Similar sublanguages have been observed across healthcare institutions [41]. Since sentence construction is governed by external knowledge (i.e., some sentences express concepts that are clearly incorrect), knowledge of sublanguage rule constraints can aid parsing and information extraction tasks. Some work suggests that in clinical text, where documents may reflect narrow clinical scopes, term statistics or semantic models trained on one note type may perform poorly on other note types [107].

### **2.2.1.1 General Approaches**

One of the pioneering systems of medical text processing is the Linguistic String Project (LSP) [120,121]. LSP used a system of template grammars and clinical SNOMED-CT vocabulary dictionaries to represent common phrases found in clinical documents and has been used on a variety of clinical text. Medical Language Extraction and Encoding (MedLEE) made use of structured vocabularies and medical concept dictionaries in addition to syntactic and language modeling components [52]. Originally designed for extracting data from radiological reports, MedLEE was extended to other domains. Both these systems made use of manually crafted grammar and syntax rules rather than machine learning approaches.

MetaMap is a widely used NLP pipeline for mapping concepts in biomedical text (typically research articles and other scholarly text) onto the ULMS Metathesaurus [8].

Originally designed for improving information retrieval from MEDLINE/PubMed citations, MetaMap has evolved over the past decade into a modern, generalized architecture for labeling biomedical text with ULMS concepts [9]. It now includes sophisticated modules for tokenization, shallow parsing, phrase variant generation, word sense disambiguation, etc. The ONYX system is designed to extract concepts and other relations from speech transcriptions for use in decision support and other clinical applications [35]. SymText and MPLUS (M+) presented a medical language Bayesian Network for extracting semantic relations from radiology reports [36].

Generalized open source NLP tools are often incorporated as underlying components in larger clinical text frameworks. Health Information Text Extraction (HITEx) is a system released as part of i2b2's Interpreting Biology for the Bedside project [102]. Similar in functionality to MetaMap, HITEx is based on the GATE natural language processing framework [39]. The Mayo Clinic Text Analysis and Knowledge Extraction System (cTAKES) is based on the Apache Foundation's OpenNLP toolkit (consisting primarily of maximum entropy classifiers) and the Unstructured Information Management Architecture (UIMA) framework, which is designed for the analysis of unstructured data [126]. The default classification models in cTAKES are trained using a collection of annotated Mayo Clinic CT documents and the framework is frequently used as baseline evaluation measure for clinical IE systems.

Beyond foundational NLP tasks, like POS-tagging, tokenization, shallow parsing, etc., most frameworks include specialized modules that address specific issues identified as useful in clinical text IE. The negation of clinical concepts is commonly addressed using the

NegEx algorithm, a negation system based on regular expressions, which is incorporated as a component of cTAKES and MetaMap [32]. Word Sense Disambiguation – using context to assign an unambiguous concept to an ambiguous mention – is implemented in cTAKES, commonly handled using ULMS concepts to aid in disambiguation [3, 54]. Several pre-processing tasks appear to improve performance. Spell checking has been examined by several researchers, either using conventional spell check software or specially trained UMLS-based systems [96, 138]. Segmenting documents on known section headers also appears to improve performance.

Many other specialized medical NLP systems exist in the literature, but they are often designed for non-clinical documents (e.g., IBM BioTeKS) or tailored for very specific applications (e.g., the Medical Knowledge Analysis Tool, which extracts cancer characteristics from pathology reports) [82]. Many tools are based on core frameworks like MetaMap or cTAKES, with extensions for important clinical IE tasks like medication extraction, medical concept recognition, etc.

#### **2.2.1.2 Medication Extraction**

Due to the challenges in de-identifying patient records, most clinical text data sets are small and released for a targeted IE challenge, e.g., de-identification, extracting temporal relations, etc. Informatics for Integrating Biology and the Bedside (i2b2) has held several yearly competitions involving a small release of annotated clinical text documents (usually discharge summaries). These annotations are in a non-standard (i.e., not Penn Treebank) format, but provide clear guidelines for annotators.

Medication extraction in clinical text involves identifying drug names in text and all of their modifying attributes, such as dosage and mode of administration. The i2b2 2009 medication challenge required labeling discharge summary text with 6 different medication attribute types and linking those attributes to a specific medication. The full i2b2 tag set is described in depth in §3.2.1. In the original competition, the best participating system used a CRF classifier to first label mentions as their i2b2 tag type [106]. Linkages between mentions were then treated as a binary classification task on medication and attribute tuples, resolved using a SVM. They report 88%  $F_1$  scores for drug identification and similar scores for dosage, mode, and frequency extraction. Duration and reason performed comparatively poorly with scores around 44%, likely due to the complex ways in which both of those attributes are discussed. Other researchers have explored hybrid rule-based and machine learning approaches [75, 130] or ways of incorporating MetaMap output into classification tasks [9]. Li et al. view medications and their attributes as a sequential, multi-layer classification problem [75]. This approach uses multiple CRF-based layers to generate independent label sets and takes the union of all individual label outputs, resulting in similar performance.

The medication extraction system MedEx identified medications and attributes by specifying a sequential semantic grammar and parsing using a top-down chart parser [155]. They report high performance ( $F_1$  scores of 93-96%) for name, frequency and other constrained lexical fields, but duration and other temporal fields still perform poorly (57-74%). More recent contributions focus not only on identifying medications in text, but normalizing those mentions into a canonical format like the UMLS/RxNorm concept unique identifier

(RxCUI). The Medication Extraction and Normalization (MedXN) tool reports  $F_1$  scores of 97.5% for names and 90% for most attributes using only tokenization modules from cTAKES, engineered regular expression patterns, and rule-based linkages [133]. Their performance exceeded MedEx for medication and attributes and MetaMap for names. However, the system requires expertise in engineering patterns for actual deployment (all evaluations were done on Mayo Clinic CT data). Thus all the methods outlined in this section primarily rely on hand-engineered lexical features encoding syntactic or semantic knowledge and none of the approaches incorporate much external domain knowledge beyond the use of lexicons.

### 2.3 Distributional Semantic Models

Language modeling – estimating the joint probability of a sequence of words – is a foundational component of NLP and information retrieval systems. Language models have classically been implemented using  $n$ -gram statistics, which estimate the joint probability of an  $m$ -long word sequence by a combination of preceding  $n-1$  contexts:

$$P(w_1^m) = \prod_{i=1}^m P(w_i | w_1^{i-1}) \approx \prod_{i=1}^m P(w_i | w_{i-n+1}^{i-1}) \quad (2.1)$$

The restriction to  $n$ -grams (i.e.,  $n$  consecutive tokens) may itself be generalized to skip-grams (tokens separated by arbitrary gaps) and statistical approaches to smoothing to address sparsity in empirical count data. Smoothing techniques vary from simple discounting methods to more more complicated back-off models [34]. These approaches all have the benefit of low computational complexity and the ability to easily scale to web-scale data sets. Language models are usually evaluated using *perplexity*, a measure of cross entropy [93]. Intuitively, perplexity captures how well a model can predict a given word in some context,

with a lower perplexity score indicating a superior model.

One disadvantage of  $n$ -gram models is that they represent terms as discrete entities, eliding syntactic and semantic properties of words and decreasing generalizability when dealing with out-of-vocabulary words. This has led to research into ways of modeling individual words in a continuous feature space and using those representations to generate joint probability estimates. One interesting development towards these ends is the increasing ability to model words as a function of the distribution of contexts in which they occur. These models are variously named *distributional semantic models*, *continuous word representations*, or *word embeddings*. Rather than viewing words as discrete units, distributional approaches model words in a continuous vector space where similar words, in a semantic or syntactic sense, are clustered within the learned parameter space.

This approach is predicated on the *Distributional Hypothesis* which proposes that word semantics can be modeled as a function of the similarity between different distributions of linguistic contexts [61, 97]. More simply, this hypothesis states that semantically-related words are used in similar contexts or neatly stated by Firth as “[y]ou shall know a word by the company it keeps” [49, 61]. A large body of research has empirically motivated this hypothesis, showing that characterizing words by the distributional properties of their contexts, such as a vector of raw co-occurrence counts, does approximate some aspect of word meaning. One advantage of representing words in a continuous (instead of the 1-of- $n$  vector model common in information retrieval) space is that it provides a mechanism for computing the semantics of unknown words with respect to known words, improving the generalizability of natural language processing systems.



### 2.3.1 Matrix Factorization Approaches

While co-occurrence captures word semantics, raw co-occurrence vectors are very high-dimensional, sparse, and perform poorly in most semantic tasks. Count vectors are usually preprocessed and de-noised using dimensionality reduction techniques like principle component analyses (PCA), singular value decomposition (SVD), random indexing [122], or manual reweighting schemes. Semantic word classes can be built using these richer representations of linguistic context, and there is a long history of methods for generating continuous word representations discussed in the research literature [12, 141]. Latent Semantic Analysis (LSA) forms initial work on generating continuous space word representations. Originally, this was a decomposition of the **term x document** matrix (consisting of term frequency counts) where the resulting low-rank approximation corresponds to some latent concept space. Many approaches operate on a  $|V| \times |C|$  sparse matrix capturing vocabulary  $V$  counts (or a strength of association measure, like positive pointwise mutual information) in some set of contexts  $C$  [74]. When  $C =$  a single document, the matrix is a standard *term*  $\times$  *document* matrix, the representation commonly used in LSA. When  $C = V$  the matrix forms what Lund et al. called *semantic memory* or *word space* and is used in hyperspace analogue to language (HAL) to model concepts [81]. One of the key insights into modeling words is a better understanding of how the definition of  $C$ , starting with document and moving to small context windows around a word, impacts the interpretability of word classes.

### 2.3.2 Hierarchical clustering

A very simple distributional semantic model is to create *word classes* by clustering word surface forms on some definition of context (e.g.,  $n$ -grams, skip-grams, etc.). Individual words are then defined by a word cluster id instead of the typical 1-of- $n$  vector encoding. This approach has the considerable advantage that it can be run on unlabeled training data, enriching the semantics of known words to include out-of-vocabulary words learned from unlabeled data. For machine learning tasks, words are clustered in an offline step to generate semantic word classes and word features are then defined using cluster membership id instead of individual terms, improving classifier performance.

Brown clustering, an agglomerative, hierarchical method using bigram context, is commonly used for defining word classes and has been successfully used in many text-mining applications [116,139]. Hierarchical clustering has the advantage of not only creating many layers of semantically-related terms, but generating a meaningful, generalized dataset that can then be distributed for use in other machine learning applications. Unfortunately, the general complexity of Brown clustering (and most non-special cases of agglomerative clustering) is  $O(|V|^3)$ , where  $V$  is vocabulary size, imposing practical constraints on the size of data sets it can be used on<sup>1</sup>. Agglomerative clustering also has the disadvantage that cluster membership definitions are fixed (at arbitrary levels of resolution), coarsening the ability to modify clusters after they are generated.

---

<sup>1</sup>Liang proposed an  $O(|V|k^2 + |T|)$  approach for agglomerative clustering where  $k$  is the number of word classes and  $T$  is text length [76].

### 2.3.3 Neural Language Models (NLM)

An alternate approach to classic language models Neural Language Models (NLMs), which use a neural network to model the probability of a word given some context. Unlike  $n$ -gram language models, NLMs estimate this joint probability using a learned real-valued feature vector for every word in its model. The original feedforward NLM, as proposed by Bengio et al. [15], intuitively consists of a projection of discrete word indices into a continuous space and a probability estimator operating over that space [128]. This consists of a projection layer  $P$  feeding input into a fully connected multi-layer perceptron with  $\mathbf{H}$  hidden units. For a vocabulary  $\mathbf{V}$ , input into the projection layer consists of a 1-of- $n$  encoding of all preceding  $n - 1$  words in an input context  $h_j = w_{t-n+1}, \dots, w_{t-2}, w_{t-1}$ . The linear projection matrix  $\mathbf{V} \times \mathbf{P}$  maps this discrete representation into a continuous one, where the  $i$ -th row corresponds to the  $i$ -th word's continuous representation. This projection matrix is shared across all contexts. Output is then the posterior probability of all words in the vocabulary for  $w_t$  given context  $h_j$ , using softmax normalization to force all output values to the range of  $[0,1]$ .

The network is trained via a standard back-propagation approach, learning both the projection layer (i.e., word embeddings) and all internal multi-layer perceptron weights simultaneously, minimizing the following error function based on predicting the next word in a training sentence:

$$E = \sum_{i=1}^{|\mathbf{V}|} t_i \log p_i + \beta \left( \sum_{jl} m_{jl}^2 + \sum_{ij} v_{ij}^2 \right) \quad (2.2)$$

Here  $t_i$  is the correct output probability (1.0 for the correct training word and 0.0 for

all others) and  $p_i$  is the softmax normalization operating over output layer nodes  $o_i$ :

$$p_i = e^{o_i} / \sum_{r=1}^N e^{o_r} \quad (2.3)$$

which corresponds to the cross-entropy between output and target probability distributions.

The second part of the error function is a regularization term based on hidden layer and output weights  $m_{jl}$  and  $v_{ij}$  where  $\beta$  is determined empirically. Probability estimates are then a function of smoothed word representations of any  $n - 1$  context and are thus able to better generalize to unseen  $n$ -gram contexts.

NLMs show much improved performance over standard  $n$ -gram models and can be trained by a variety of neural network structures, e.g., feedforward networks, convolutional networks, recurrent neural networks, etc. Each choice of network structure entails different runtime complexities and representational trade-offs. Simple NLMs coarsely model temporal or sequential data by specifying the context  $n$  as a pre-determined window size. Sequences are then fed as input as a time-delayed buffer often called a taped delay line. While this is conceptually easy, this approach impacts generalization, as the same inputs effect different network weights based on the position in the buffer. Simple NLMs also lack a concrete form of short-term memory beyond words moving to a new position in the tape delay.

### 2.3.4 Word Embeddings

One of the more interesting aspects of NLMs, aside from their superior performance in language modeling, is the discovery that the resulting parameter matrix of word representations (word embeddings) does a surprisingly good job of modeling word semantics. NLMs use a supervised learning approach, learning an  $n$ -dimensional parameterization of a word

that best predicts its own neighboring words, within some context window  $c$ . The advantage of this training objective is that it comes with no annotation costs; training is done using sequence information derived from an input text. In practice, the non-linearity of NLMs isn't actually necessary to generate high quality word representations and is discarded to speed up training. Currently, the best performing word embeddings for many tasks are generated using word2vec [94]. Here, using their Skip-gram model, words are sampled from  $c$  during training, giving less weight to more distant words. Formally the Skip-gram model's objective function maximizes the average log probability

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \quad (2.4)$$

where  $p(w_{t+j} | w_t)$  is defined using the softmax function:

$$p(w_O | w_I) = \frac{\exp(v'_{w_O} \top v_{w_I})}{\sum_{w=1}^W \exp(v'_w \top v_{w_I})} \quad (2.5)$$

where  $v_w$  and  $v'_w$  are input and output vector representations of word  $w$  and  $W$  is the size of the entire input vocabulary. Since the softmax function is  $\mathcal{O}(W)$  the full softmax is approximated using hierarchical softmax [100], reducing the computation to  $\mathcal{O}(\log W)$ . To account for the highly skewed frequencies of observed words (which can bias learned representations of rare words), word2vec subsamples each word using a heuristically derived probability function and threshold.

Training embeddings using predictive or reconstructive objective functions are shown to result in embeddings that dramatically outperform count-based semantic models [11]. Evidence suggests that incorporating these embeddings into existing linear or shallow classifiers improves performance in sequence labeling classification tasks [139]. More recently,

there have been advances in fully end-to-end systems, i.e., supervised-learning systems that require less (if any) manually engineered features as input. Collobert et al. [37] presented a system that used a lookup table of word embeddings as the only feature input with competitive performance in tasks like POS-tagging, NER, and chunking. Recently Zhang and LeCun presented a convolutional neural network for text classification that only uses characters as input [157].

### 2.3.5 Recurrent Neural Networks

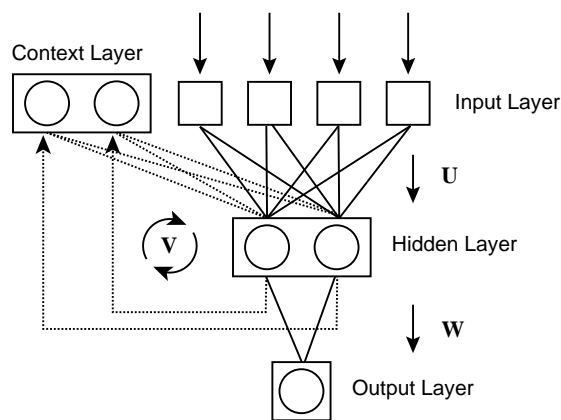


Figure 2.1: Simple Elman-type Recurrent Neural Network.  $\mathbf{U}$  is the input  $\times$  hidden layer weight matrix.  $\mathbf{V}$  is the context layer  $\times$  hidden layer matrix, and  $\mathbf{W}$  is the output weight matrix. Dotted lines indicate recurrent edge weights.

One of the key challenges in sequence modeling is learning long-distance dependencies. Standard neural networks have no conception of “memory” in terms of previously seen states, making such dependencies difficult to model. One algorithm that speaks to this limitation is an Elman network [43], a form of Recurrent Neural Network (RNNs). RNNs are very similar

to classical feedforward neural networks, except that they incorporate an additional hidden layer that captures some element of sequence history. RNNs can naturally incorporate dense, real valued feature vectors and have been used for other embedding-based sequence labeling tasks [90]. In the canonical RNN design, the output of the hidden layer at  $t - 1$  is retained in a special context layer, which is fed back into the hidden layer at  $t$  (see Figure 2.1). Errors can then be backpropagated using a technique called *Backpropagation Through Time* (BPTT) [119]. BPTT essentially maintains a copy of weights at a given time step, evaluates error by unfolding the state layer across time, and then refolding time to collapse weights to a single state layer update. This mechanism enables RNNs to learn a summary representation of past states during training, which is hypothesized to help model longer term dependencies in sequence data [129].

Formally, RNNs are defined as:

$$h(t) = f(\mathbf{U}x(t) + \mathbf{V}h(t - 1)) \quad (2.6)$$

$$y(t) = g(\mathbf{W}h(t)) \quad (2.7)$$

where  $\mathbf{U}$  is the input-to-hidden layer matrix,  $\mathbf{V}$  is the hidden layer-to-context layer matrix, and  $\mathbf{W}$  is the output layer matrix.  $f(x)$  and  $g(x)$  are the hidden and output layer activation functions, in this work sigmoid and softmax respectively. The output layer emits a probability distribution in the dimension of the candidate label set size which is then used to classify a token.

Other neural network structures have been successfully used in certain language tasks. Convolutional networks, which create tiles of overlapping input sequences, are commonly used in image recognition tasks, but have also been successfully used to induce dependency

structure at the sentence level [23]. Other network structures or activation functions (e.g., recursive auto-encoders, k-max pooling layers, etc.) and have been used to optimize performance in tasks like paraphrase detection [131].

### 2.3.6 Evaluating Distributional Semantic Models

The question of what defines a high-quality word embedding is itself challenging; several evaluation data sets have been proposed as candidate benchmarks [13,48,92]. Broadly, evaluations take the form of detecting synonyms, identifying members of concept sets, or inferring unseen relations in test word tuples. Some evidence suggests that embeddings generated via LSA or other approaches using a bag-of-words model don't capture contextual regularities as effectively as NLM embeddings [95].

Many evaluations focus on synonymy, such as the TOEFL synonym detection task or the WordSim 353 data set. The concept categorization task, where nouns are clustered and compared to gold standard categories, is another proposed measure [7]. More complicated evaluations explore the relational aspects of word spaces, such as predicting unseen words in analogy questions. A NLM trained on a 6B token corpus, evaluated on 8,869 semantic relation pairs (e.g. *city-in-state*, *man-woman*, *currency*, etc.) and 10,675 syntactic relations (e.g., *opposites*, *past tense*, *plural nouns*, etc.), reported 65.6% overall accuracy in predicting the correct missing word [92]. Banko et al. created a BLESS, an comprehensive evaluation dataset that tested performance across multiple different relation types extracted from the ukWaC and Wackypedia corpora [13].



## CHAPTER 3 DATA SETS: DESCRIPTIONS AND SUMMARY STATISTICS

### 3.1 Sexual Behavior Data

#### 3.1.1 Craigslist

Craigslist is an online classified advertisement website that allows users to post free, anonymous classified advertisements (*ads*) in a variety of different categories (e.g., items for sale, job offerings, dating personals, etc.). As of September 2013, it is the 10th most visited website in the United States [5]. Craigslist predominately facilitates the exchange of traditional goods (e.g., cars, musical instruments, furniture), but via the personals category, it also supports a large community of casual-sex-seeking individuals, predominantly but not exclusively MSM. These personal ads announce each poster's availability and preference for sexual encounters. 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).

#### 1. Craigslist

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 relation-

Table 3.1: Craigslist Corpora Summary

Name	Ads	Tokens	Selection Criteria
CRAIGSLIST	130.6M (134.5M)	-	Daily sample of 8 personal ad and 5 commercial categories from 412 U.S. Craigslist sites.
MSM	28.6M (32.5M)	2.1B	CRAIGSLIST ad with encounter tag $t \in \{m4m, m4t, m4mm, mm4mm, mm4m\}$ with near-duplicate ads removed
GOLD	700	42k	A uniform random sample of MSM ads from all sites ( $n=500$ ), all California sites ( $n=100$ ), and the <i>sfbay</i> site ( $n=100$ ) which was then human-annotated with race and age information.
PHONE	303k	20M	MSM ads containing an obfuscated telephone number (e.g., “867-5309” becomes “8sixseven5three oh nine”).
LOCTAGS	-	-	Site-wise frequency distributions of all CRAIGSLIST location tags, consisting of 88K unique location tags.

ship 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. Only ad counts (and subject tags) were retained for commercial categories. All personal ad categories were retrieved. In total, 134.5M ads were obtained (130.6M with near-duplicates removed), forming the CRAIGSLIST corpus. A list of all subcorpora derived from CRAIGSLIST is provided in Table 3.1 and described in detail below.

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 [20]. Sentences were then tokenized on whitespace and punctuation with a rule-based system used to merge individual tokens into their final term 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.

## 2. MSM

The MSM corpus is a subset of CRAIGSLIST containing all ads targeting the MSM individuals. The target population is identified Craigslist encounter tag, defining MSM as the set of all ads containing a tag  $t \in \{m4m, m4t, m4mm, mm4mm, mm4m\}$ . Filtering by tag results in 32.5 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 that are, with high probability, written by the same author into a single ad instance. This is done using the near-duplicate detection method described in §4.3. This results in a final collection size of 28.6M ads, currently the largest data set of online sexual encounter ads examined in the literature.

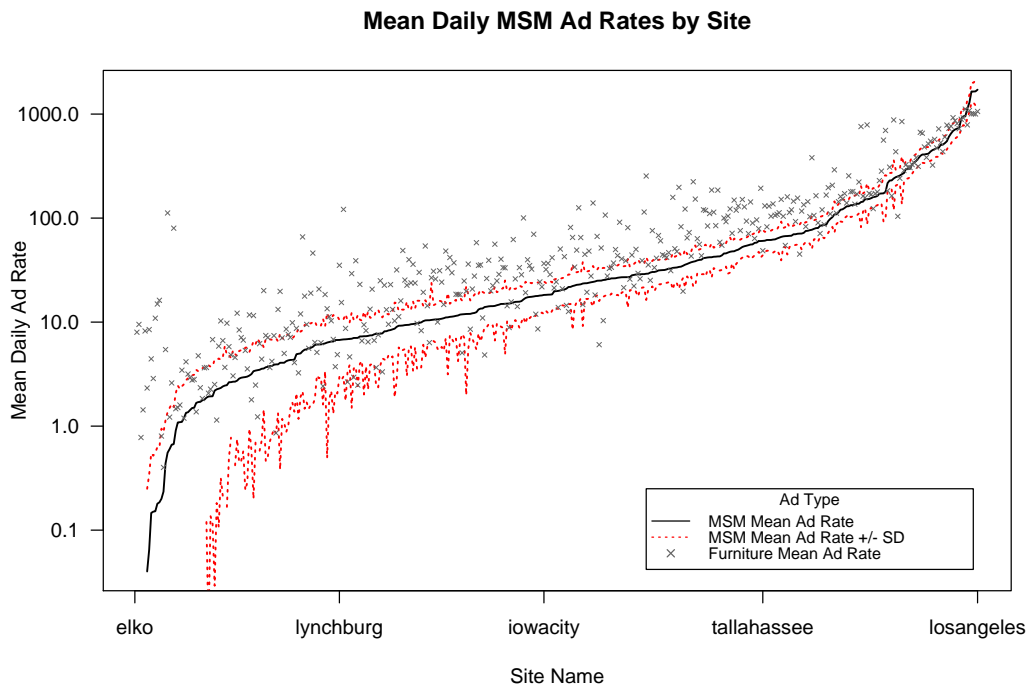


Figure 3.1: Log-linear scatterplot of mean daily post rates for MSM ads (solid black line) calculated over 723 days of Craigslist posts. Representative sites at 5-number summary percentiles are listed on the x-axis. Red dotted lines indicate  $\pm 1$  SD for MSM ads. 50% of sites (205/412) have daily MSM post rates between 7 and 62 ads per day. The top 5 MSM sites, *losangeles*, *newyork*, *sfbay*, *chicago*, *dallas*, have daily rates of 1716, 1654, 1648, 1648, 1286 respectively. Mean daily rates for used furniture ads (gray points) are provided for comparison; furniture ads are on average higher than MSM ad rates (mean difference of 27 ads); 19% of sites have mean MSM rates higher than furniture ads. Nationwide, the mean rate of MSM ads was 38,193 (SD 5,998) posts per day.

Figure 3.1 presents posting rate statistics for MSM ads.

### 3. GOLD

Three disjoint samples of annotated CRAIGSLIST ads were created as a gold standard (i.e., GOLD) to evaluate performance of our information extraction methodologies. Ads from this corpus 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.

1	New Guy in town - Looking for a Sat.AM parTy host!
2	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Neighborhood_District</span>  <span style="border: 1px solid black; padding: 2px;">(castro)</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Neighborhood_District</span>  <span style="border: 1px solid black; padding: 2px;">/ upper market)</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Age *</span>  <span style="border: 1px solid black; padding: 2px;">35yr.</span> </div> </div>
3	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Age *</span>  <span style="border: 1px solid black; padding: 2px;">Thirty Five</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Six Two</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Italian/Austrian</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">(white)</span> </div> <div style="text-align: center;"> <span style="border: 1px solid black; padding: 2px;">Masc.</span> </div> </div>
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

Figure 3.2: Example GOLD annotations from an m4m *sfbay* ad. Race/ethnicity mentions (in green) are assigned to a racial group (e.g., Caucasian) and tagged as referencing 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.

All annotation was done using an open source text annotation tool [134]. These 700 ads were labeled with geographic named entities, author/partner age and race/ethnicity, as well as a set of sexual behaviors. Annotated entities are described in detail below and a sample annotation is shown in Figure 3.2.

## 1. LOCATIONS

A. **GPE**: State, County, Area-code, Informal-region, Indian-reservation, CDP, City, Public-transit, Zipcode, Neighborhood/District, HWY/Street

B. **Place**: Hotel, Park, Airport, Address, University, Other

C. **Unknown**

## 2. Age

3. **Race**: Caucasian, Black, Asian, Hispanic/Latino, American-Indian, Native-Hawaiian/Pacific-Islander, Biracial

4. **SexAct**: Oral, Anal, Masturbation
5. **Encounter Preference**
6. **Drug**: Marijuana, Alcohol, Amyl-nitrate, Meth, Cocaine
7. **HIV/AIDS**
8. **Phone Number**

## Geographic Entities

Table 3.2: Annotation Corpus Toponym Distributions

Entity Category	Entity Type	In-tag $n$	In-body $n$	N	Percentage
GPE	<b>City/CDP</b>	542	105	647	60.9%
	<b>Neighborhood/District</b>	150	31	181	17.0%
	<b>HWY/Street</b>	42	26	68	6.4%
	<b>County</b>	23	2	25	2.4%
	Informal-region	19	4	23	2.2%
	<b>State</b>	12	5	17	1.6%
	Area-code	3	2	5	0.5%
	Public-transit	3	1	4	0.4%
	<b>Zipcode</b>	2	1	3	0.3%
	Unknown	0	1	1	0.1%
Place	University	20	8	28	2.6%
	Other	16	29	45	4.2%
	Hotel	5	6	11	1.0%
	Address	1	1	2	0.2%
	Park	1	2	3	0.3%
TOTALS		839	224	1063	

All three datasets were annotated to identify geographic named entities (*toponyms*), with types consisting of 2 categories: *geopolitical entities* (GPEs) such as counties, cities,

neighborhoods, etc., and *places* which consist of point locations or institutions, such as local businesses, hotels, universities, and airports. Entities were identified in both the location tag and ad body text for a total of 1063 toponyms.

Table 3.2 contains summary statistics on the annotation corpus toponym types. In total, 839 entities were found in location tags, with 89% (625/700) of tags referencing at least one toponym. A tag may contain no entities (75/700), a single mention (469/700), or multiple mentions (156/700). Entities found in the body of an ad were more likely to reference *place* entity types than location tags alone (odds ratio 4.7, p-value <0.05, 95% CI [2.9, 7.5] using two-sided Fisher’s exact test). All 771 GPEs found in location tags were manually linked to canonical geographic knowledge base (KB) representations using U.S. census data. This set of city, neighborhood, road, county, state and ZIP code entities form the validation set used in our entity linking evaluation. Some annotated entities were missing a canonical representation in our KB ( $n=11$ ). In these circumstances, the entity was assigned to the closest geographically bounding entity of the same type. For example, because the neighborhood “China Basin” in San Francisco is not present in our database, its geographic area is folded into the larger neighborhood of “South of Market”, which is then used as the true entity.

Most entity mentions in the GOLD corpus correspond to City/CDP types (64.6%), followed by Neighborhoods, and HWY/Streets. Collectively these entity types account for 87.5% of all mentions in annotated location tags. The remaining annotated entities were divided between coarser resolution mentions (e.g., state, county, area codes) (7.4%) and high-resolution mentions corresponding to business names, buildings, etc. (5.1%).

## Age and Race/Ethnicity Mentions

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) [142]. No annotated ad text disclosed *American Indian/Native Alaskan* origins, so that population was not considered in any analysis.

## 4. PHONE

The PHONE corpus consists of MSM ads where authors’ provide an obfuscated phone number in ad text. For example, “867-5309” becomes “8sixseven5three oh nine.” Since such obfuscation typically just consists of replacing digits with word equivalents in a sequence, phone numbers can be normalized by matching candidate phone numbers using a regular expression and only retaining those numbers that match valid phone number lengths. Linking ads by matching phone numbers provides a mechanism for linking anonymous ads across time and forms the evaluation data set for examining our near-duplicate detection system.

## 5. LOCTAGS

The LOCTAGS corpus consists of site-wise frequency distributions of all location tags. Tag text is normalized using the same process described above for the CRAIGSLIST corpus. All location tags are filter by removing any tag with a raw frequency count under 100. The final set of normalized, thresholded tags consists of 88K unique location tags.



## 3.2 Clinical Text

### 3.2.1 i2b2 Medication Extraction


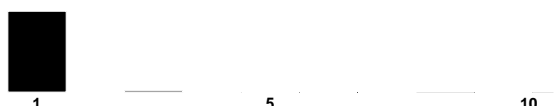

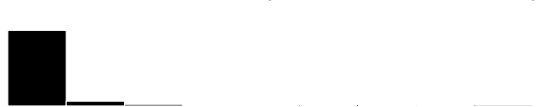

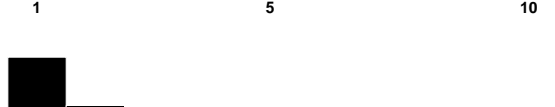
*Informatics for Integrating Biology and the Bedside* (i2b2) has released several clinical text collections for specific yearly information extraction challenge tasks. This work uses the 2009 data set focusing on the extraction of medications and medication-related information from CT. This data set consists of 1250 discharge summaries with 251 documents (33k lines, median line length of 9 words/tokens) annotated to include words and phrases covering 6 types of medications mentions [145]. These entities are described in detail in the i2b2 annotation guidelines [146] and are described briefly below:

1. **Medication/Drug:** Drug/medication mentions “for which the patient is the experimenter”, e.g., “lasix”, “vitamins A.” Annotations exclude mentions of food, water, tobacco, alcohol, and illicit drugs.
2. **Dosage:** The amount of drug administered, e.g., 100 mg.
3. **Mode:** The method by which a drug is administered, e.g., intravenous, topical, orally.
4. **Frequency:** Terms or phrases indicating how often should this medication be administered.
5. **Duration:** Terms or phrases indicating how long this medication should be administered.
6. **Reason:** The medical reason for prescribing a medication.

The i2b2 annotations further indicate whether a mention occurs in a narrative or list section of a discharge summary, which this dissertation does not consider. Note that discharge

summaries are not split into sentences but are instead split into lines of some max width. Table 3.3 provides some summary information on annotated entities themselves, including median entity token length, entity counts, etc. The histograms in the right column show the distribution of entity word counts. Note that drug, dose, mode, and frequency mentions predominately consist of single words while duration and reason mentions include multiple words. This shows that mentions of duration and reason contain a much higher percentage of phrases than the other tag categories.

Table 3.3: i2b2 Annotation Entity Summary

Entity Type	Entity N	Median	Max	Entity Word Count	
				Word Count Distribution	
Drug	8,939	1	14		
Dose	4,456	2	10		
Mode	3,386	1	5		
Frequency	4,039	1	41		
Duration	553	3	13		
Reason	1,635	2	7		

### 3.2.2 University of Iowa Hospital and Clinics

The *University of Iowa Hospital and Clinics Corpus* (UIHC) corpus is a very large collection of clinical text data collected from the UIHC's Epic-based electronic medical record system. These notes, linked by patient identifier, span 2007 - 2014 and include over 79 different clinical document types (e.g., discharge summaries, clinical notes, progress notes, etc.) Notes are not de-identified, dash records are linked by patient identifier across time and include protected health information (PHI) such as attending healthcare worker, dates of admission, etc. The entire corpus consists of 880M sentences, 6.6B unigram tokens, 34M unique word types (900k when thresholded at a minimum occurrence of 5). For purposes of this dissertation, we examine 4 corpora in detail: *discharge summaries*, *progress notes*, *clinic notes*, and the set of *all notes*.

## CHAPTER 4 BUILDING A SURVEILLANCE PIPELINE

### 4.1 Surveillance Pipeline NLP Subtasks

In order to utilize Craigslist ads for surveillance purposes a number of important NLP subtasks must be first addressed:

1. **Geographic Entity Normalization:** Craigslist ads make frequent mention of geographic entities that can be used to link ads to specific locations. However these mentions occur in unstructured text, meaning they must first be identified and then linked to a canonical knowledge base representation of geographic entities (*normalization*). We present a geographic entity (*toponym*) recognition and linking system that can link geographic locations in ad text to their corresponding U.S. census entity.
2. **Authorship Attribution:** Unlike most social networks such as Twitter or Facebook, Craigslist is an anonymous, broadcast form of social media; ads are not linked to any notion of identity. This poses sampling challenges when conducting behavioral surveillance, since we risk oversampling individuals who frequently post ads to Craigslist. To partially account for this, we present an authorship attribution system that can cluster ads likely written by the same author. We leverage ads that include identifying information (i.e., phone numbers) to validate the performance of this system.
3. **Extract Demographic Information:** From a public health perspective, knowing how behaviors stratify across age and demographic variables such as race/ethnicity can provide insight into designing intervention programs. While variables such as age

are relatively easy to extract using regular expressions, mentions of race/ethnicity are considerably more challenging. This is due to both the breadth of terminology used to describe race as well as the ambiguity between mentions of the author’s race and the description of their desired partner. This work reframes the problem of extracting author and partner race/ethnicity as a sequence labeling task, using a machine learning approach to assign tags to words which are used to properly disambiguate and extract mentions.

This chapter proposes solutions to all of these issues. While each individual subtask can be improved upon, we show this system provides sufficient accuracy for extracting high-fidelity geotemporal data from the MSM sexual community operating on Craigslist.

## 4.2 Geographic Named Entity Normalization

This section presents an automatic toponym recognition and linking algorithm, *Classified Ad Toponym Linker (TopoLinker)*, that uses location tag text to geocode ads to a canonical knowledge base (KB) of entities – a task called *entity linking*, *entity disambiguation*, or *normalization*. Identifying and linking toponyms in online ads pose challenges beyond those seen in conventional geocoding applications due to the highly informal nature of location tags. Unlike geographic queries used in mapping contexts, location tags don’t necessarily conform to a hierarchical left/right ordering of geographic entities or contain common search patterns (e.g., directions from point A to B). Tags reflect a high degree of implicit spatial information and are often underspecified or otherwise ambiguous; street names may drop prefixes or suffixes; places may be referred to using colloquial names or unusual spelling

variations; and geographic boundary definitions may themselves may be ambiguous, as is the case with neighborhoods. Tags may include multiple separate entity mentions (as authors provide a set of possible locations where they could travel for a transaction) or may contain no location mentions at all (e.g., “anywhere”, “your place”). The disambiguation task is further complicated by the lack of sentence context clues in location tags to help differentiate named entities from non-entity text. Leveraging the content of ads for disambiguation also poses challenges, as ad text itself is often short and highly uniform in content, further diminishing context clues for disambiguation. Our *TopoLinker* entity linking method includes some solutions to these challenges and provides a substantial increase in performance over a naive method that simply uses the global site location to geocode ads (e.g., associating all ads in the Craigslist site *sfbay* with San Francisco, California.).

In our hand-annotated validation set, *TopoLinker* correctly linked 85% of all tags to their exact canonical geographic entity. Overall, *TopoLinker* resulted in a 77% reduction in mean error over the naive site location method, measured as distance between predicted entity and true entity. Running *TopoLinker* on a document collection of 106 million Craigslist location tags, we find that ads reference a variety of high-resolution spatial data, with 97% of all ads corresponding to areas smaller than the median area of a U.S. county and 10% tags refer to areas smaller than the median size of a U.S. neighborhood.

## 4.2.1 Materials/Methods

### 4.2.1.1 Annotation Corpora & Toponym Knowledge Base

#### Location Tag Corpus

Online classified ads are typically *semi-structured* (i.e., tagged with metadata), email-like text documents. Craigslist ads consist of a subject line, keyword-encoded metadata tags, and a body of unstructured text. The majority of ads include a tag corresponding to the geographic location or region where the author wishes to negotiate a transaction (see Figure 4.1 for example subject lines and location tags). Location tags are user-generated, meaning they are a noisy data stream with many textual variations or *surface forms* of the same underlying named entity. The tag set associated with the canonical entity “San Francisco, CA”, for example, includes “sf”, “sanfrancisco”, as well as implicit tags like “downtown”. Tags may reflect multiple locations (“cole valley / ashbury hts”), refer to informal geographic regions (“bay area”, “the valley”), or contain non-entity text (“couch , sofa”, “my place”, “on sale”, etc.). Location tags are analogous to “check-ins” found in location-based networks (LBNs) such as FourSquare, but with a highly variable spatial resolution and encoding a number of superordinate-subordinate relationships. For example, some tags correspond to metropolitan-sized areas (e.g., “L.A.”) while others are venue-specific (e.g., “Dolores Park”, “Marriott Marquis”, etc.) and capture the spatial resolution typically seen in LBNs.

We examine two data sets this section: (1) a sample of 700 location tags that was human-annotated to identify geographic place names, i.e., GOLD and (2) the site-wise distribution of normalized location tag counts, i.e., LOCTAGS. Both of these data sets are

described in detail in §3.1.1.

In order to generate a list of probable low information stop words [123], each location tag term is also assigned an *idf* (inverse document frequency) [65] weight. This weight is calculated by creating a collection of 414 pseudo-documents consisting of the set of all location tags for a given site, which allows the *idf* calculation to assign more weight to terms that occur across fewer sites (i.e., terms that likely correspond to local regional features). The bottom 1000 ranked IDF terms are further filtered to remove variations of the term “downtown” and any term matching a city name with a population  $\geq 10,000$ ; this filtered list forms our final stop word set.

Location tag metadata is not unique to Craigslist and most classified ad services provide some mechanism of indicating the geographic scope of an ad. Table 4.1 includes some representative examples of the surface forms seen in the LOCTAGS dataset as well as several examples of location tags from Backpage ads, manually selected from the *losangeles* site in November 2013, to illustrate the similarities in tag format.

### **Toponym Knowledge Database**

When selecting a canonical knowledge base (KB) resource, the Freebase website (a knowledge base consisting of crowdsourced structured data) or Wikipedia are common choices in named entity recognition and linking tasks. However, our linking system needs to execute geospatial queries, such as distance from geographic boundaries, in a generalized way as well as maintain entity identifiers across multiple U.S. Census data sets. While Freebase encodes some geographic relations (*contains*, *partially contains*, *adjoins*, etc.),



Schwinn Bantam Girls - \$200 (**santa clara**)  
 Women's roadbike!!! - \$80  
 Sony DSC-H200 - \$130 (**san jose south**)  
 CHEM 32 - General, Organic, & Biochemistry (**ingleside / SFSU / CCSF**)  
 85/65/14 NEW SET OF 4 TIRES ON SALE \$159 (**STOCKTON-(TIRES WHEELS BRAKES ALIGNMENT)**)  
 I'm new here - w4m 23 (**downtown / civic / van ness**)  
 Spin the Bottle/7 Minutes in Heaven - w4m (**Happy Hour**)  
 You were selling dog leashes with your old grumpy dog - w4m (**Ferry Building**)  
 Black guy coming from Starbucks - w4m 23 (**hayward / castro valley**)  
 Fireman on stevens creek - w4m (**bascom/stevens creek**)

Figure 4.1: Sample *sfbay* Craigslist ad subject lines. Each subject line may include an optional location tag, enclosed in parentheses (shown here in boldface), of up to 40 characters in length. Tags contain no formatting constraints beyond length and may refer to arbitrary geographic entities (cities, buildings, streets, etc.) or no entities at all.

it lacks a mechanism for easily importing new geographic data respective of these relations.

Geospatial databases like PostGIS provide a consistent query interface across any geographic data and easy mechanisms for importing new spatial data. As such, we built our canonical geographic representations with a PostGIS database of common location entities found in the United States [62]. Using public 2009 & 2010 TIGER/Line geographic boundary files, canonical entities were created for all landmarks, roads, cities/towns/census-designated places (CDP), core based statistical areas (CBSA), counties, and other major geographic locations [144]. Because census data does not define any boundary type directly corresponding to neighborhoods – frequently used in ad location tags but itself a nebulous geographic delineation – we used the online real estate database company Zillow's publicly available shapefiles to define neighborhood entities [158]. The tool used to download all geographic datasets and create the PostGIS database used in this paper is freely available online [53].

For every Craigslist site, we manually generated a canonical set of county and CBSA

boundaries comprising that site’s *coverage footprint* – the geographic region serviced by the site in question. These coverage areas are defined as the set of geographic boundaries covered by the convex hull of the metro/micropolitan statistical area of the site’s primary city or, for areas that do not directly correspond to cities, the set of cities and towns contained within that region (e.g., Inland Empire, California). The list of such cities is extracted from the Wikipedia page for each region, which Craigslist provides a link to on each site’s homepage.

#### 4.2.1.2 Named Entity Recognition and Linking

We present two methods for associating ads with geographic locations: a baseline method named *Centroid* and *TopoLinker*, the entity linking algorithm that is the primary contribution of this section. Both methods were evaluated against ground truth, represented by the GOLD annotation corpus.

##### Method I: Centroid Algorithm

All Craigslist ads have an implicit location by virtue of that ad’s posting site; ads posted on *sfbay*, for example, likely correspond to the city of San Francisco, California. This suggests a naive geocoding method in which every ad is simply assigned to its parent-site’s coverage footprint. Since location tags may contain multiple entity mentions, tags are first segmented using a rule-based tokenizer with each mention or *span* identified by splitting the entire tag on a set of manually defined delimiters, where `delim`  $\in$  `{\|&;:}`. For example, the location tag “downtown/civic/van ness” is split into 3 spans: [ `downtown`, `civic`, `van ness` ] and each span in this sequence is then assigned to the coverage footprint. This method does not provide a mechanism for exact entity linking, as it fails to identify fine-

grained entity types such as roads or neighborhoods, but it does provide a convenient baseline measure for evaluating how well an algorithm can link an entity to its true location, gauging performance in terms of distance error between true and predicted locations.

## Method II: TopoLinker Algorithm

The *TopoLinker* algorithm presents a more sophisticated approach to identifying and linking entities in location tags. Broadly speaking, *TopoLinker* works by generating probable synonym sets for both the original location tag and the set of all possible candidate locations in a given Craigslist site. This tag and entity aliasing creates a large, varied set of candidate entity matches, which are then filtered to a final predicted labeling using geographic constraints (e.g., containment, distance), city population counts, and other implicit information such as the parent ad site. The *TopoLinker* pipeline (see Figure 4.2) proceeds in three stages: 1) generate site lexicons and toponym entity aliases; 2) match location tags to generate candidate named entity sets; and 3) disambiguate candidates and link surface forms to canonical named entities. Each stage of *TopoLinker* is explained in detail below:

### Stage 1: Lexicon & Entity Alias Generation

For each site, a candidate named entity lexicon is created consisting of all KB roads, neighborhoods, landmarks, counties, and state entities found within the site’s coverage footprint and all cities within 200 miles of the footprint boundary (parameter selected empirically). Craigslist location tags make frequent use of informal abbreviations or other surface form transformations of entity names. Many of these variations can be addressed by introducing a series of alias transformation rules that generate surface form synonym sets, which

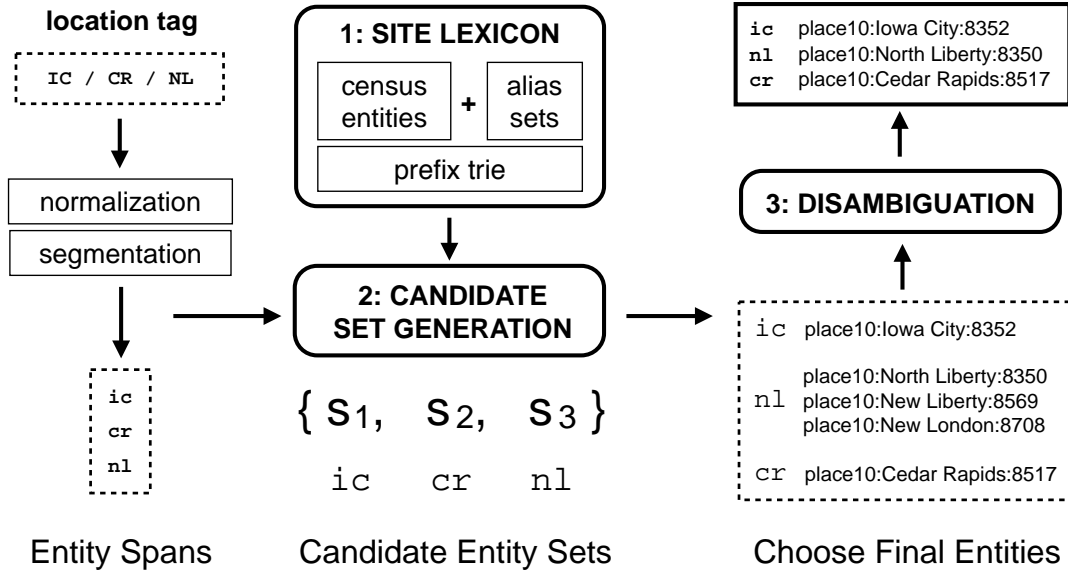


Figure 4.2: Classified ad toponym entity labeling pipeline. The above flowchart uses the *iowacity* site location tag “IC/CR/NL” as an example to illustrate the entity linking pipeline. Location tags are first normalized and segmented into entity spans, with site specific lexicons used to generate candidate entity sets (steps 1 and 2). The final set of entities is then disambiguated using a combination U.S. Census data and a set of rule-based scoring functions.

are then used to match candidate entities. The full set of alias transformation rules and their target modified entity types are explained below:

1. **Acronym** (*Neighborhoods, Cities*) Use the first letter of each whitespace tokenized term in the entity name, delimited by a character from the set  $\{ \_ - / \}$ . Example:  
Iowa City  $\rightarrow \{ ic, i/c \}$
2. **SegmentDropAcronym** (*Neighborhoods, Cities*) Any entity name containing the suffix  $s \in \{bay, ville, town, burg, wood\}$  is aliased using the entity’s first letter + suffix to create acronyms and concatenated variations, delimited by a character from the set  $\{ \_ ' - \}$ . Example: Provincetown  $\rightarrow \{ p'town, p-town, ptown,$

pt, p/t }

3. **Syllable** (*Neighborhoods, Cities*) Using a hyphenation dictionary [72], any entity name with 2 or more tokens, alias using the `first syllable + remaining terms`. Example: `Strawberry Plains` → `straw plains`
4. **PrefixDrop** (*Roads*) Strip any cardinal direction prefix term. Example: `N Dubuque St` → `dubuque st`
5. **SuffixDrop** (*Roads, Neighborhoods, Counties*) Strip any official USPS street suffix term [143]. Example: `N Dubuque St` → `n dubuque`
6. **Root** (*Roads*) Strip both cardinal direction prefixes and USPS street suffixes to create root entity name. Example: `N Dubuque St` → `dubuque`
7. **Road** (*Roads*) Create common variations for highway and interstate names. Example: `US Hwy 1` → `{us 1, us1, us-1}`, `I-35` → `{i35, i 35}`
8. **Concatenation** (*Neighborhoods, Cities*) Remove all whitespace and punctuation. Example: `Los Angeles` → `losangeles`, `O'Fallon` → `ofallon`
9. **Dictionary** (*Cities, States*) A set of common entity aliases created through manual inspection of location tag frequency lists, as well as the full set of international airport codes which are often used as shorthand for city names. Example: `Minneapolis` → `mpls`, `Los Angeles` → `lax`

All surface form aliases are then loaded into a trie data structure for efficient string prefix matching, which is then used to label whitespace-tokenized location tags with exact string-to-entity matches.

## Stage 2: Candidate Entity Set Generation

Location tags are segmented into candidate entity spans using the same rule-based tokenizer described under the *Centroid* algorithm. Each location tag span is further expanded by creating a set of aliased, permuted spans, generated by collapsing and expanding USPS official street suffixes and cardinal compass directions. For example, the full permuted span set for the location tag “rich/rose/ft.bend” is [ {rich}, {rose}, {fort bend, ft bend, ft bnd} ]. Here “rich” and “rose” have no aliases and remain unchanged, while with “ft.bend” both “ft.” (fort) and “bend” are postal suffixes, resulting in an expanded span set.

For each span, our KB prefix trie is queried to generate a list of candidate entity matches, grouped by entity type (e.g., neighborhood, city, etc.) and length of string match (measured in token length). All string matches for the target span and its aliases form the candidate entity set. To reduce the overall size of the solution space, each candidate set is then filtered to include only the set of longest string matches for any given span, discarding any span sub-string entity matches.

## Stage 3: Disambiguation & Linking

Each location tag is labeled with  $n$  candidate entity sets (1 set per identified span) which then must be disambiguated to select the final, per-span predicted entity. The number of detected spans in a location tag provides a convenient way of partitioning the disambiguation problem; the majority of annotated tags (66%) have 1 entity mention, with the rest having 0 or multiple entities. For tags with single entities, we primarily have string similarity and an estimate of the overall distribution of entity types to guide disambiguation. For

tags with multiple entities, we would like to leverage the co-occurrence of multiple entities to help disambiguate candidates, based on the insight that ad authors are generally describing a small bounded region, comprised of like-types, where they are willing to meet. We present two scoring functions for ranking and selecting the best candidate entity for a given tag span.

### Single Entity Linking

Given a single detected entity span and candidate set  $C_1$  of entities  $e$ , we choose a predicted entity using:

$$best(C_1) := \underset{e \in C_1}{\operatorname{argmin}} w(e) + lev(span, name(e)) \quad (4.1)$$

where  $lev$  is the Levenshtein distance [73] between the span string and the candidate’s canonical name and  $w$  is a weighting reflecting geographic precedence by entity type. Assigning a penalty to each type captures the degree to which we favor choosing certain candidate toponym types over others. We use weights in the range 0 – 6 reflecting the following total ordering of entity types: [place, neighborhoods, point/landmark, roads, county, state, zipcode].

Because single entity spans can be highly ambiguous, we build in several manual tests for restricting candidate entity types in spans that would otherwise be missed due to geographic precedence; any span containing county-equivalent terms (e.g., “county”, “parish”, “borough”) is restricted to county entities; any 2 letter acronym matching a state abbreviation to states; and any span containing digits, ordinal suffix, or street suffixes is restricted to roads. Finally, if a span is present in our stop word list or no matches are otherwise discovered, we return a `None` link.

## Multiple Entity Linking

Given  $n$  detected entity spans, where each span is assigned a set of candidate entity sets  $S_n = \{s_1, s_2, \dots, s_n\}$ , let  $C_n$  be defined as all combinations over  $S_n$ , with the best  $n$  labeling defined as:

$$n\text{-best}(C_n) := \operatorname{argmin}_{d \in C_n} \sum_{i=1}^n \sum_{j=1}^n d_{ij} \quad (4.2)$$

where  $d$  is a pairwise distance matrix of all  $n$  entities in a given combination. Naively, this approach explodes our solution space as we evaluate every possible candidate entity set. Road entity types are particularly challenging in this regard, due to the large number of shared root words in naming conventions (e.g., “Main St”, “Main Ave”, etc.). In practice however, some simple candidate filtering largely mitigates this issue; 99% of all observed location tags in GOLD have 3 or fewer entity mentions and filtering each set of per-span candidates to the best scored entity type substantially reduces the possible solution space size. Each span candidate set is filtered to include only the best ranked entity type (using geographic precedence weights) and the resulting subset permuted to generate  $C_n$ .

As with the single entity linker, we incorporate several rules for restricting span candidate sets to certain entity types to reduce sizes. We use the same restriction tests for counties, states, stop words, etc. as in single entities, as well as incorporate a road intersection test to identify possible road entities. Finally, if no candidate entity types in a given sequence are of type city or higher (in geographic precedence), then an implicit *anchor city* entity is added to the existing candidate set. This anchor city is the highest ranked city entity, chosen from the set of all cities within a site footprint, weighted by  $1/d * \text{population}$  where  $d$  is distance from the footprint centroid. If any entities are not contained within that



city, they are removed from the candidate set.

### Acronym Disambiguation

Any span of length  $\leq 3$  is considered an acronym and needs special consideration in terms of scoring since acronyms break disambiguation using string similarity. Both scoring functions deal with the special case of acronym spans by ranking each candidate using raw population count data for any city found in a given site’s coverage footprint, essentially using population as a proxy estimate for the likelihood of a given acronym corresponding to a target city. City entities are ranked first and if no viable candidates are found, then neighborhood entities are ranked using distance from the site centroid. In each case, the set of candidate entities is reduced to the single top ranked item, which is scored as 0 for purposes of all further entity ranking calculations.

Finally, after evaluating linking performance parameters, *TopoLinker* was run on the entire LOCTAGS location tag corpus, generating summary statistics on entity spatial resolution and frequency (see Figure 4.3).

#### 4.2.1.3 Evaluation Measures

We use 3 different metrics to evaluate entity linking performance. All metrics are calculated using the annotated corpus GOLD. Duplicate occurrences of entities within a single location tag are considered separate mentions. For the equations below,  $T$  denotes the complete set of true entities and  $P$  denotes the final set of predicted entities for all GOLD tags. For a given location tag,  $t_{tag}$  denotes the true entity set,  $c_{tag}$  denotes the candidate set, and  $p_{tag}$  denotes the final predicted entity set.

1. **Candidate Recall (CndR):** To evaluate how well our initial candidate matching stage performs, we define a measure that simply looks to see if the correct entity is present in the set of all possible candidates. This measure reflects the best possible linking performance, given perfect disambiguation, using input candidate set  $C$ .

$$candidate-recall(C) := \left( \sum_{tag \in Gold} |c_{tag} \cap t_{tag}| \right) / |T| \quad (4.3)$$

2. **GeoExact:** Predicted entities are evaluated in terms of standard precision and recall, where a predicted entity label  $p \in p_{tag}$  is correct iff  $p \in t_{tag}$ . Note that this measure disregards token span matching and only evaluates the final set of predicted entities. Incorrectly predicted entity labels penalize precision and missed true entity labels penalize recall.

$$precision(P) := \left( \sum_{tag \in Gold} |p_{tag} \cap t_{tag}| \right) / |P| \quad (4.4)$$

$$recall(P) := \left( \sum_{tag \in Gold} |p_{tag} \cap t_{tag}| \right) / |T| \quad (4.5)$$

3. **Mean Distance Error:** Formally, for each tag we generate a predicted label mapping  $M_{tag} = p_{tag} \rightarrow t_{tag}$ . Pair  $(p, p_t) \in M_{tag}$  is constructed using overlapping true and predicted entity token spans in location tags. Conflicted mappings (i.e., more than one label entity per true entity) are decided using length of token overlap or in cases of ties the label that occurs first in token ordering. *Mean Distance Error* is then measured in terms of great-circle distance, in miles, between the geographic centroids of predicted entity  $p$  and the true entity  $p_t$ . Here  $M_{tag}$  includes null labelings  $(\emptyset, p_t)$  to account for

false negatives in  $P$  labelings. Null label errors are measured as the distance from the parent site centroid to  $p_t$ .

$$dist-error(P) := \frac{1}{|T|} \sum_{tag \in Gold} \sum_{(p, p_t) \in M_{tag}} distance(p, p_t) \quad (4.6)$$

## 4.2.2 Results

### Normalization Performance

Table 4.1: GeoExact Toponym Linking Performance

Entity Type	N	Without Aliasing				With Aliasing			
		CndR	Recall	Precision	F <sub>1</sub>	CndR	Recall	Precision	F <sub>1</sub>
City	542	0.77	0.76	0.92	0.83	0.93	0.90	0.94	0.92
Neighborhood	150	0.73	0.65	0.95	0.77	0.86	0.74	0.93	0.83
HWY/Street	42	0.12	0.10	0.50	0.16	0.83	0.64	0.56	0.60
Zipcode	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
County	23	0.26	0.26	1.00	0.41	0.78	0.30	1.00	0.47
State	12	0.08	0.00	1.00	0.00	1.00	1.00	0.86	0.92
None	74	1.00	0.93	0.33	0.49	1.00	0.95	0.53	0.68
TOTAL	771	0.73	0.70	0.76	0.73	0.91	0.85	0.85	0.85

Generating candidate entity sets using only exact string names (i.e., without alias rules) results in candidate recall of 0.73; incorporating aliasing results in 0.91, a 25% increase. Aliasing increases the size of candidate sets from a median baseline size of 1 to a median of 15. Filtering candidate sets to include only the set of longest string matches results in 1% drop in candidate recall, but a 40% reduction in median set size, from 15 to 9 entities per entity span. For the  $n$ -best entity scoring function, where the solution space consists of permuted sequences across  $n$  candidate sets, candidate sequences are further filtered to include only

the highest ranked entity type (using geographic precedence), reducing the median solution set size by 90% (from 10 to 1) and the maximum set size by several orders of magnitude (from 15,177,600 to 1,053).

Table 4.1 contains results for the GeoExact measure, which evaluates *TopoLinker's* exact entity linking performance. CndR (*Candidate Recall*) measures the performance of *TopoLinker's* candidate generation step and reflects the best possible linking performance, assuming perfect disambiguation. Performance without aliasing rules resulted in lower recall (2% to 54%) in all classes. *TopoLinker* (with aliasing enabled) correctly linked 85% of entities in the GOLD annotation corpus to their canonical representation. Incorporating alias generation provides clear benefits, especially in the HWY/Street category, improving candidate recall scores by 0.18 (25%) and linking recall by 0.15 (21%) overall. Performance was best in the majority entity class, City, with a combined  $F_1$  score of 0.92 and worst in the HWY/Street with an  $F_1$  score of 0.60.

Per-entity type candidate recall shows the various benefits provided by entity name aliasing, with increases seen in every category. Especially large increases were seen in HWY/Street and State entities (86% and 92% improvement respectively) – both categories where surface form mentions rarely match the canonical U.S. Census form.

Table 4.2 reports an alternate way of comparing this methods using the mean distance error measure. Overall, *TopoLinker* reduced mean error (in miles) by 77% when compared to the naive *Centroid* approach, which only uses site location information to link entities. Across all entity types, *TopoLinker* shows a 24% - 100% reduction in mean error, measured as the distribution of distances (in miles) between the centroids of all machine-labeled entities and

Table 4.2: Entity Linking Miles Error

Entity Type	<i>Centroid Method</i>		<i>TopoLinker</i>		vs. Centroid Method
	Mean Err. (SD)	Min/Max	Mean Err. (SD)	Min/Max	
City	27.3 (30.0)	[1.8 - 317.1]	4.0 (22.7)	[0.0 - 317.1]	-85.5%
Neighborhood	13.9 (8.7)	[0.5 - 47.2]	8.5 (27.5)	[0.0 - 218.1]	-39.0%
HWY/Street	16.0 (7.2)	[1.1 - 32.8]	10.9 (35.3)	[0.0 - 228.9]	-32.0%
Zipcode	42.8 (34.9)	[8.0 - 77.7]	0.0 (0.0)	[0.0]	-100.0%
County	29.8 (27.3)	[2.5 - 102.6]	22.6 (33.5)	[0.0 - 123.0]	-24.0%
State	97.0 (83.9)	[9.3 - 282.1]	0.0 (0.0)	[0.0]	-100.0%
TOTAL	25.3 (30.0)	[0.5 - 317.1]	5.7 (25.0)	[0.0 - 317.1]	-77.4%

their corresponding true location, with a total mean error of 5.7 (S.D. 25) miles. Average distance errors by entity type for *TopoLinker* ranged from 4.0 - 22.6 miles compared to *Centroid* where they ranged from 13.9 - 97.0 miles.

In Table 4.2 *Mean Error* is the mean distance (in miles) between the centroids of all machine-labeled entities and their corresponding true location, with lower scores indicating better performance. *Centroid* links all toponym mentions to their parent site centroid; *TopoLinker* links all toponyms to their knowledge base representation. Note that mean error is lower for the *TopoLinker* system across all toponym types, with 24%-100% reduction in error across all entity types. *Topolinker* errors are more geographically dispersed than *Centroid*, which is expected, since candidates may be chosen from any location inside the site coverage footprint.

### Craigslist Entity Summary

Using *TopoLinker* to extract entities from the entire LOCTAGS corpus links on average 87.1% (SD 11.6%) of all ads in a given site to 1 or more geographic locations, for a total of 87M

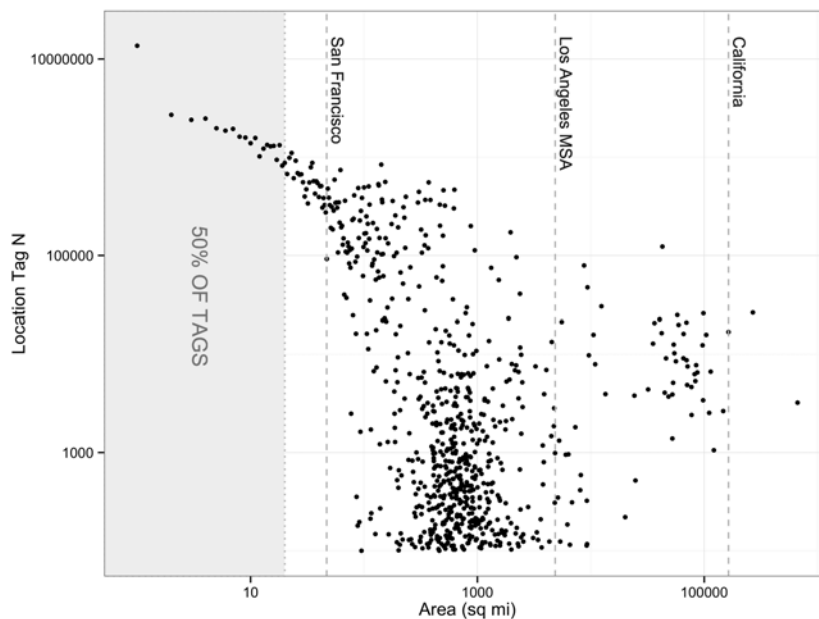


Figure 4.3: Log-log scatterplot of the frequency of all location entity tags vs. the corresponding area of that tag in square miles. Note the high geographic resolution of entity locations, with 97% of all tags correspond to areas smaller than 640 sq mi – the median area of a U.S. county using 2010 census data. Overall tags reflect the high spatial resolution of ad entities: 71% of all tags refer to areas smaller than the city of San Francisco, California (47 sq mi); 50% of all tags refer to a geographic boundary with an area less than 15 sq mi; and 10% of tags refer to areas smaller than the median size of a U.S. neighborhood (0.65 sq mi). Vertical dotted lines indicate the size of named locations to provide a relative sense of scale.

geocoded ads. In total, 20,601 unique named toponym entities were identified, the majority being City/CDP entities, followed by HWY/Street and Neighborhood entities. Figure 4.3 shows a log-log scatterplot of the relationship between tag frequency and corresponding geographic area, measured in square miles. Almost all tag entities, 97.6%, correspond to areas smaller than 640 sq mi – the median area of a U.S. county using 2010 census data. The majority of entities correspond to geographic regions with much higher spatial resolution; 67.8% of all tags refer to areas smaller than the city of San Francisco, California (47 sq mi) and 15.8% of tags refer to areas  $\leq 1$  sq mi, approximately the area of a typical U.S.

neighborhood (median area of 0.65 sq mi).

### 4.2.3 Discussion

#### **TopoLinker Performance**

The *TopoLinker* method can correctly link 85% of toponym entities in our annotation corpus, with a mean error of 5.7 miles for incorrect matches. *TopoLinker* brings significant improvements to the task of correctly extracting geographic information from ads, reducing mean error (in distance) by 77% when compared to a naive approach using only site location information. Moreover, we provide an open source tool for geocoding ads that uses only publicly available data. *TopoLinker* typically fails in circumstances of slang or informal names for locations (e.g., “sac” for Sacramento); further work on terminology acquisition could lead to better linking performance. *TopoLinker* also performs poorly with HWY/Street entities – the third most common occurring entity in the Craigslist corpus. Roads are the most ambiguous entity mentions in our corpus, as indicated by the high candidate recall gains seen when using alias rules. Unlike cities, where population provides a reasonable proxy for frequency, roads lack an immediately obvious mechanism for disambiguating candidates. HWY/Street precision suffers in our system because many road names (or their aliases) also match other entities in a given region. Moreover, beyond removing candidates not contained within a co-occurring or inferred city entity, there is little way for our system to identify a false positive from tag data alone.

Analyzing the entire corpus shows that the overwhelming majority of location tags, 98%, correspond to locations smaller than the median size of a U.S. county. While we sam-

pled ads for only a small fraction of the total number of Craigslist categories (13 out of approximately 130), the spatial resolution is likely similar across most categories. This suggests that the spatial resolution of ads can be used for at least county-level surveillance tasks. Moreover many ads, especially those targeting larger cities, contain even higher geographic resolution data in the form of neighborhoods, streets, etc. This fact is likely explained by both the sheer number of potential meeting locations within metropolitan areas, as well as the need to provide more specific information in a city like Los Angeles, where city or county-level tags are too ambiguous to be very informative. Overall, this suggests that geographic information retrieval and information extraction could be conducted at the city level.

There are many sources of error in our entity linking pipeline. Some linking errors are the result of how certain geographic boundaries are defined in our knowledge base. For example, some neighborhoods are missing from the Zillow data set (e.g., “Sherman Oaks” in Los Angeles) or have multiple geographic entities folded into a single boundary. Combining additional sources of geographic data such as OpenStreetMap, Flickr geotagged shapefiles, etc. could improve linking performance, especially in cases like neighborhoods, where no official boundary definition exists. Entity geographic precedence conflicts can also occur, as many cities and neighborhoods in a given region share the exact same canonical name. Disambiguation in these contexts could be improved by looking for clues in the text of the ad body, which our method currently does not do.

Finally, the major cause of error in our system is the diversity of entity surface forms. While our generative alias rules cover many common cases and dictionary methods can address certain predictable surface form synonyms (e.g., “NYC”, “philly”) the long tail



of spelling variations and other unusual abbreviations present problems. Entity mentions like “el sob” for El Sobrante, California or colloquial shorthand like “san berdoo” for San Bernardino are not handled, nor are informal names for non-city locations like “Miracle Mile” or “Vegas Strip”. Learning common entity synonyms could be done by incorporating external data sources such as Wikipedia disambiguation pages, while less common synonyms could be identified using surface form clustering (e.g., Brown clustering [26], etc.).

### 4.3 Authorship Attribution

This work examines a corpus of anonymous personal ads seeking sexual encounters from the classifieds website Craigslist and presents a way of linking multiple ads posted across time to a single author. We make two observations about anonymous Craigslist personal ads: 1) many users appear to post ads at regular intervals, making little or no textual changes to the body of their message; and 2) changes are often confined to the subject line and contain updated location information. This re-posting phenomenon likely stems from the fact that Craigslist, as of March 2012, had no mechanism for users to permanently delete or remove an ad from their account history. Every user’s account management page maintains a log of all posting activity, including ad text, date of post, and ad category. This log, when coupled with Craigslist’s 2-click mechanism for quickly reposting expired ads, creates an incentive for some users to repost ads, either maintaining the exact same content, or – more commonly – making small changes to the text.

Under the presumption that highly similar ads (measured using cosine similarity), when not spam, originate from the same author, we can use efficient near-duplicate detection

techniques to cluster ads within some threshold similarity. Linking ads in this way allows us to preserve the anonymity of authors while still extracting useful information on the frequency with which authors post ads, as well as the geographic regions in which they seek encounters. Naively, while this process detects many clusters of similar ads, the lack of a true corpus of authorship-linked ads makes it difficult to validate and tune the parameters of our system. Fortunately, 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 create a corpus of ad clusters known to be written by a single author. This authorship corpus can then be used to evaluate and tune our existing near-duplicate detection system, and in the future identify features for more robust authorship attribution techniques.

By examining each of these clusters and identifying differences in user-supplied location tags, we can then reconstruct an approximation of an anonymous individual's movement *footprint* over time, as well as estimate the rate at which ad authors seek sexual encounters. Using the spatial information present in these ad clusters, we can then construct an estimate of the geographic region where an individual is willing to meet for a casual sexual encounter and calculate associated travel distances within that region. This can provide insight not only into the spatial clustering of behaviors defining core groups, but the degree to which geographic clusters themselves are interconnected.

### 4.3.1 Methods

#### 4.3.1.1 Near-Duplicate Detection

The *near-duplicate detection problem* is the task of efficiently identifying textual content or documents in a collection that, using a pairwise measurement of similarity, fall within a similarity threshold value in the range  $[0, 1]$ . Detecting near-duplicates has many applications in web-crawling [84], plagiarism detection [6], and document clustering [64]. Our task is similar to that of plagiarism detection, in that we wish to detect repeated subsequences of text across many different Craigslist ads. The idea that various word/phrase choices, text formatting (e.g., consistently capitalizing words for dramatic effect), and other author-specific linguistic features or *stylometry* can be used to identify individuals in online contexts was explored in [1]. We do not consider any of these more advanced approaches to identifying online identity in anonymous settings, but instead focus on identifying nearly-identical ads using  $n$ -gram document features (i.e., sequences of  $n$  consecutive words).

Since exactly solving common similarity functions such as the Jaccard index are computationally expensive and therefore intractable for large document collections [153], a number of locality-sensitive hashing techniques have been developed for efficiently identifying approximate sets with similar features. Charikar’s simhash [33] is a probabilistic dimension reduction technique by which a set of input features is hashed into a  $f$ -bit long *fingerprint* such that similar input items hash to similar values with very high probability.

## Fingerprint Generation

To calibrate our selection of document features and evaluate the detection accuracy of our parameter choices, we conducted several preliminary experiments. All Craigslist ads from the San Francisco Bay Area site were tokenized into case-sensitive unigrams and bigrams (i.e., single words and adjacent word pairs) and a 64-bit simhash fingerprint created for each resulting document term vector. We then calculated the all-pairs Hamming distance [60],  $k$ , for each ad. 50 pairs of hashes at Hamming distances  $k = [1...10]$  were randomly selected from each  $n$ -gram fingerprint set. A human annotator then evaluated each pair as a true or false positive near-duplicate match. The guidelines for making decisions were as follows: pairs were flagged as false positives when they contained little to no sentence or phrasal commonalities; true positives required that ads share substantial matching textual content, orthographic features (e.g., capitalization, slight differences in sentence order, and/or minor textual changes). Based on these results (see Figure 4.4) we choose a Hamming distance of 6-bits for our detection threshold.

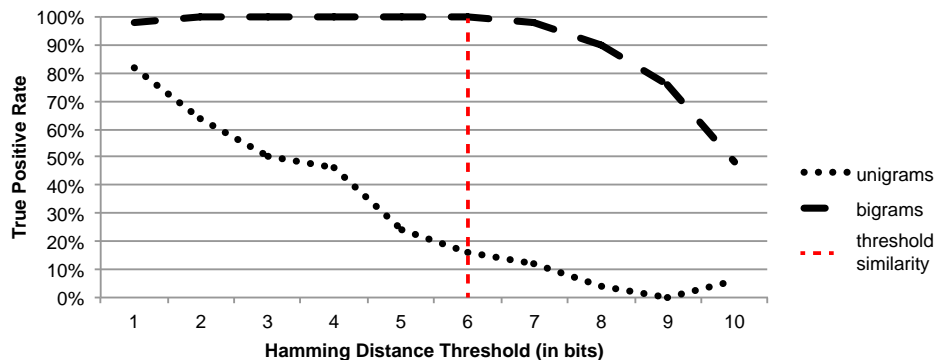


Figure 4.4: True positive rates for hash fingerprint feature selection (using unigram and bigram feature sets) and evaluated at Hamming distances  $k = [1...10]$ . Note that unigram features perform poorly overall, while bigrams perform well at distances  $1 \leq k \leq 7$ .

To efficiently compute the pairwise Hamming distance of all 33M ads, we utilized the method described in Manku et al. [85]. The key observation to their approach is that the top  $n$ -bits of each fingerprint, when sorted, function approximately as a counter, allowing queries to quickly reduce the search space necessary for finding hashes within some distance threshold. All 64-bit fingerprints are split into 4 tables of 16-bit blocks, with each table's blocks permuted such that each has a different leading block of bits. Each ad's fingerprint is then queried against these 4 tables and all fingerprints within  $k=6$  bits of the query are returned and greedily assigned to a cluster. Ads with no matches are considered singleton messages. Ads that are previously assigned to a cluster are not queried.

### Removing Spurious Ad Clusters

The end result of the query process consists of the set of all singleton messages and  $n$  matched ad clusters. Because these ad clusters contain both a level of spurious matches (as a result of the fingerprinting process) and a degree of multi-post spam messages, additional filtering is required to identify clusters corresponding to the behavior of real users over some interval of time.

First, to filter out spurious hash matches, for each cluster we calculated the pairwise inter-cluster Jaccard index for the corresponding ad term vectors. Any cluster in which the median similarity value falls below a threshold  $r$  (we used 0.60) is discarded (see Figure 4.5 for the distribution of inter-cluster similarities). Likewise any single document contained within a cluster that falls below this threshold is also removed.

The remaining clusters are further processed to address ads reposted multiple times

within a small time window to the same Craigslist site, which seem to typically result from users making minor changes or corrections to the text content of an ad and reposting. Ads within a cluster that are posted within 4 hours of each other are collapsed into 1 posting instance. Clusters consisting solely of near-duplicates within the same 24-hour interval are also considered singleton ads.

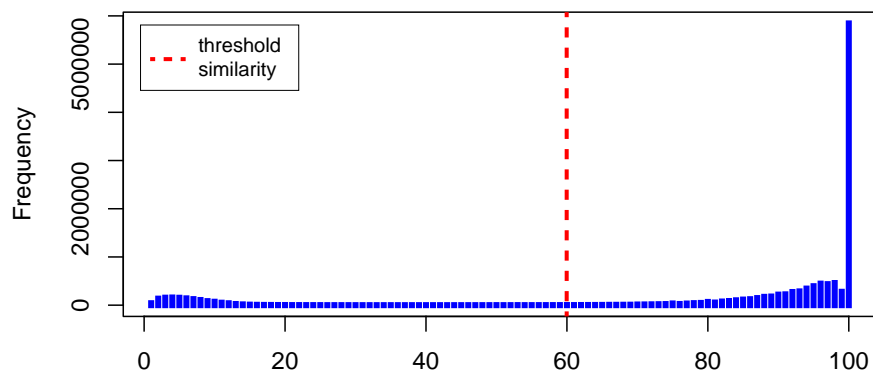


Figure 4.5: Histogram of intra-cluster Jaccard similarities, before removing spurious near-duplicate clusters. 48% of all matches consist of exact duplicate ads (the peak on the far right of the plot). The majority of detected near-duplicates are true matches, with only 12% of all detected clusters falling below our 0.6 similarity threshold.

## Removing Spam

Identifying near-duplicate ads on Craigslist also identifies spam ad clusters. These spam clusters can consist of web advertisements, chain letters, or requests unrelated to negotiating a direct-contact encounter (e.g., ads asking to chat over webcams, posting nationally to trade erotic pictures, etc.). All of these spam types exhibit wide geographic dispersion, with multiple ads posted within a narrow window of time to numerous, geographically dis-

tant Craigslist sites, making it highly unlikely they correspond to realistic travel intentions. While Craigslist maintains internal mechanisms for detecting both similar content posted across multiple sites and duplicate ads posted within the same 48-hour window, manual inspection of detected clusters reveals that spam still exists in our corpus.

We removed any cluster containing references to external URLs (e.g., “guyflings.com”, “caliguys.com”). and any cluster using inline HTML table markup to obfuscate text for bypassing Craigslist’s internal spam filters (e.g., “Go<tr>o<tr>gle<tr> fo<tr>r i<tr>t”). To filter out any remaining spam clusters, we identified several non-lexical features of spam clusters and used a machine learning approach to classify them. Certain Craigslist spam ad clusters exhibit features similar to those observed in botnets [154], where spam can be characterized, in part, by rapid bursts of posting activity across a widely dispersed geographic area. Since our corpus covers the entire United States, we can extract a geographic distribution for all clusters and identify clusters with large, implausible geographic and temporal footprints. All feature are calculated as a vector of values, aggregated into 7-day intervals, over the duration (*lifespan*) of each ad cluster. Cluster Features as described below:

- *Maximum Weekly Burst Rate*: the maximum number of ads posted within any single 7-day interval.
- *Mean and Maximum Weekly Geographic Dispersion*: for all ads ordered in time, we calculate the sum of all hop distances within a single 7-day interval, and record the mean and maximum value of the resulting vector.
- *Geotemporal Stability*: the percentage of weeks for each cluster in which there was 0 dispersion, i.e., either no ad was posted or the ad contained no change in location.

The entire ad cluster corpus was split by ad cluster size (i.e., the number of ads comprising a single cluster) into 10 bins and clusters randomly selected from each bin. In total, 525 ad clusters were manually inspected and flagged as spam or valid. Guidelines for flagging spam involved ordering all posts within a cluster by time and examining the hop distances between each posted ad location; clusters involving multiple long distance hops within a short window of time were flagged as spam. We used WEKA [59] to test a number of classifiers, of which an AdaBoostM1 classifier evaluated with a DecisionStump as the base classifier performed best, with a weighted average true positive score of 93.5% and ROC area value of 0.96.

#### 4.3.1.2 Constructing Footprint Travel Graphs

We built 2 directed travel graphs for the state of California, representing the change in location tags within clusters over time. The first graph contains all entity-mapped location tags, while the second aggregates all tags into their county of origin. Edges are created using pairs of ads within clusters, where all ads are ordered in time and each edge represents a location tag change between each ad comprising the pair. For example, if the first ad in a cluster contained the location tag “glendale” (a city in Los Angeles) and the next detected ad within that same cluster contained the location tag “long beach” (a city also in Los Angeles, but 30 miles south) a directed edge (*Glendale, Long Beach*) is created connecting both locations. Edge weights were assigned in units of  $w$  ads, with  $w$  taking on fractional values in circumstances where a single location tag contained multiple locations (e.g., “burbank/north hollywood/glendale”). In these cases, edges were created for each pair of tags across ads.



For the county graph, edges are only created when changes to location tags cross county lines. Given these edges, we then calculated the haversine (or great circle) distance for all paths to derive estimates for the “roaming” geographic distance covered by each cluster’s footprint. Table 4.4 contains graph summary statistics for both footprint graphs.

### 4.3.2 Results

#### Near-Duplicate Identification Summary

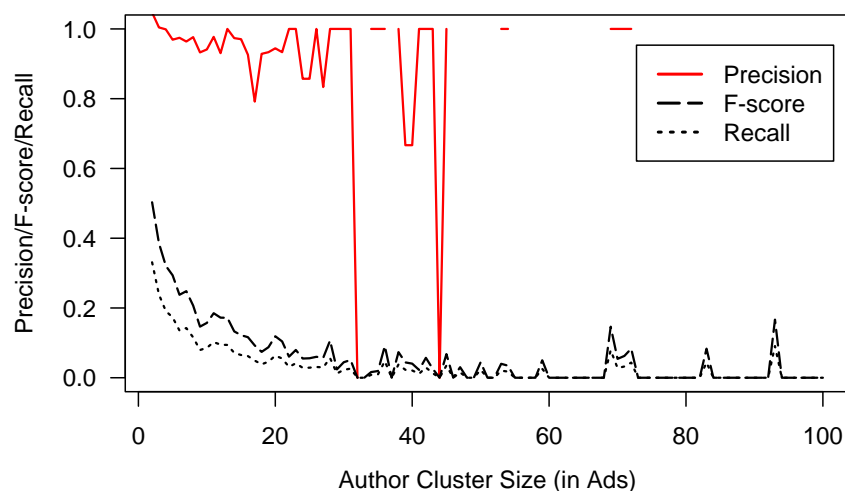


Figure 4.6: Performance measures for partial cluster detection using the near-duplicate identification approach to linking anonymous authorship in Craigslist ads. In general, our near-duplicate approach detects a small subset of ads that are written by the same author, but suffers from poor recall when identifying all of an author’s ads. Recall performance suffers as the number of ads written by a single author increases.

652,014 ads contained phone numbers, producing a total of 46,079 authorship-linked ad clusters. For detecting all ads within a cluster, precision ranged from 0.05 to 0.0 and recall from 0.02 to 0.0 for all cluster sizes. For detecting partial clusters, see Figure 4.6.

Table 4.3: Ad Cluster Sizes (U.S. and California)

	U.S.		California	
	Cluster N	Ad N	Cluster N	Ad N
Spam	66.8K	165.2K	17.5K	53.1K
Collapsed to Singleton	36.2K	-	6.0K	-
Valid Clusters	1.1M	3.7M	219.3K	723.9K
≥ 4 ads per Cluster	249.2K	1.7M	49.6K	344.2K
Cross-site Clusters	10.5.K	84,215	3.3K	29.5K
Cross-state Clusters	7.3K	57,126	2.0K	18.5K

Table 4.3 contains summary statistics for the final filtered set of ad clusters for both the entire United States and the state of California. These remaining filtered clusters represent likely (though still anonymous) posting behavior of a single Craigslist user, where each ad can now be viewed as a check-in to a given location at some point in time. These check-ins define the geographic region within which an ad’s author is willing to travel for a casual sexual encounter. The textual changes authors make to their ads over time reveal useful information and can be used to gain insight into changes in behavior. By examining the changes in each cluster’s set of user-supplied location tags, for example, we can calculate how the geographic region changes over time and derive an estimate of how far that author is willing to travel for an encounter. Users also update biographic details, their age for example, which can provide a weak form of validation. Over the entire set of ads reporting age (14% of all clusters in California), the mean age change was 0.25 years (S.D. 1.4 years). Of the 1,410 ad clusters spanning longer than one year (0.6% of the total clusters), 67% of the clusters contain changes incrementing their reported age by 1 or 2 years, with a mean change of 2.3 years (S.D. 2.6) over the 2-year corpus.

## Footprint Distances and Location Entropy

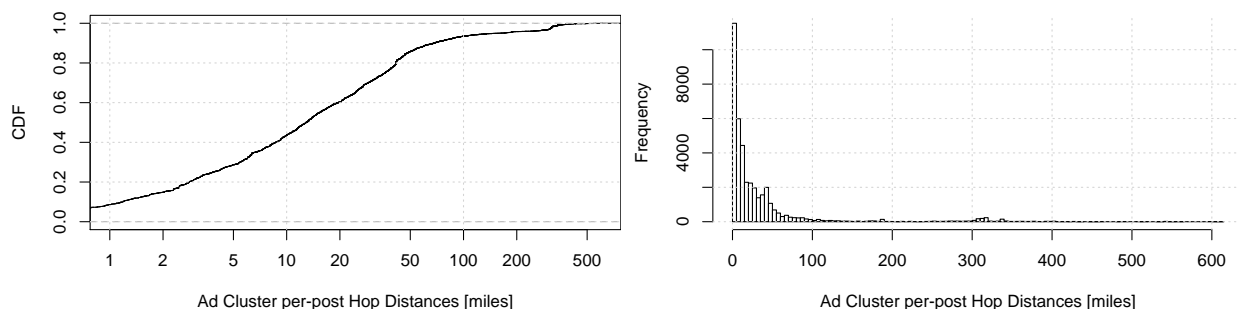


Figure 4.7: Empirical Cumulative Distribution Function (left) and histogram (right) of hop distances, calculated as the haversine distance (in miles) between location tags of successive ads within a given ad cluster. 86% of all hop distances involve distances under 50 miles while 4% involve distances over 250 miles. Note as the haversine distance is the minimal great-circle distance between 2 points on a sphere, these values underestimate the actual number of miles traveled.

Figure 4.7 shows the distribution of per-hop roaming distances for all detected California ad clusters. The majority of all cluster paths involve footprint distances of no more than 50 miles, with a small percentage (4%) indicating long-distance travel. The most common of these paths was between Los Angeles and San Francisco, as well as paths to outlining mountainous counties such as El Dorado and Placer County (the counties adjoining Lake Tahoe, a well-known vacation destination).

Figure 4.8 shows the hourly average Shannon entropy of ad location tags and average post counts for the duration of our corpus. A higher entropy value corresponds to more observed variability in location tags for that hour. All location tag counts are binned by U.S. state of origin. Note how entropy periodicity mirrors posting frequency, with two distinct morning and evening posting peaks. While Tuesday-Friday involve more evening posting

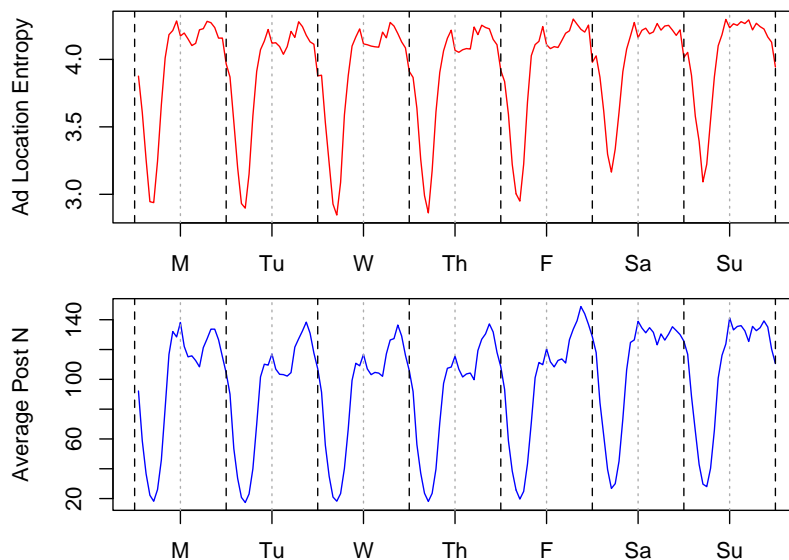


Figure 4.8: (Top, red) Average hourly entropy of location tags counts, binned by U.S. state, and (bottom, blue) average posting counts by hour for all U.S. clusters. Black dashed lines divide days and grey dotted lines correspond to noon. Higher entropy corresponds to more tag variability across bins, indicating a greater willingness to travel. Observe how the periodicity of tag entropy largely corresponds to posting frequency, with slightly more location variability observed Friday and Saturday evenings and Saturday and Sunday mornings.

on average, the corresponding entropy of location tags are essentially equivalent between peaks. Weekend (Saturday and Sunday) mornings show a slightly higher entropy overall, revealing that those up in the early morning hours express more willingness to travel than early morning/late night posts made during the work week.

## Travel Networks

Figure 4.10 shows the San Francisco Bay Area as a directed graph of changes in location tags, which can be viewed as a measure of flow between locations, in terms of an individual's willingness to travel to that location for a sexual encounter. Observe how the layout respects the actual spatial distribution of cities in the Bay Area and how most flow occurs between downtown San Francisco and San Jose – the result of both legitimate travel

Table 4.4: Graph Summary Statistics

California Graphs	All Location Tags	County-level
Nodes	4,464	55
Edges	43,023	792
Avg. Clustering Coefficient	0.022	0.566
Avg. Degree	0.301	14.4
Diameter	15	5
Avg. Path Length	5.7	2
Min Edge Weight	1	1
Max Edge Weight	154	445,978

between these two locations and the semantic ambiguity of the tag “downtown” which may also refer to downtown San Jose. Figure 4.9 shows the county-level, directed graph of travel between counties in California. Note how some counties are characterized by asymmetric inward/outward flow. This can be partly explained by the bias inherent in how people report location tags – authors are more likely to report tags about locations they intend to travel to – as well as the asymmetric desirability of certain long-distance travel locations (i.e., vacation spots).

### 4.3.3 Discussion

We have shown that a near-duplicate identification system can successfully and efficiently cluster Craigslist ads with similar content. Moreover, even in the noisy, unstructured data environment of anonymous personal ads, we show that it is still possible to extract meaningful signal. By leveraging the information contained within these clusters, we are able to construct graph representations of the geographic footprints of sexual encounter requests at varying levels of spatial resolution. We also report data capturing an author’s

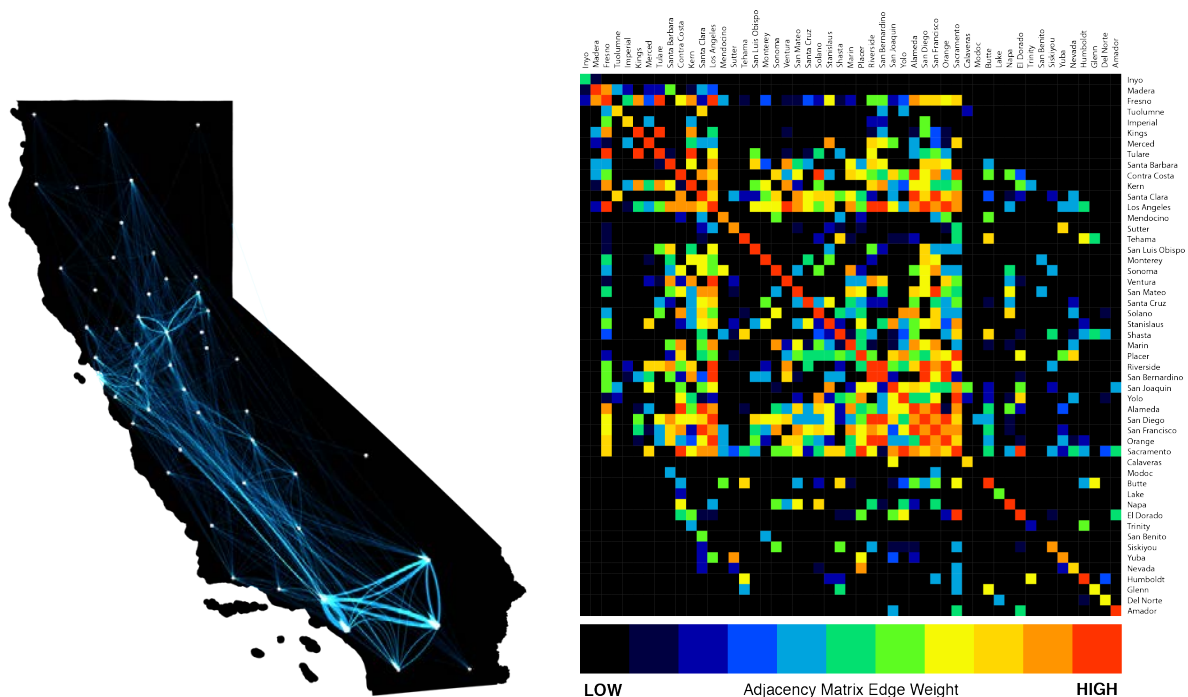


Figure 4.9: Graph of all authorship-linked ad posts crossing county lines (left) and the same graphs corresponding adjacency matrix with Cuthill-McKee [40] ordering of nodes (right). The adjacent matrix shows clustering between major counties in California (e.g, Los Angeles County, Orange County, etc).

willingness to travel within these footprints in miles, the first such results that did not originate from survey data. We find that the majority of encounter requests occur within a 50-mile area around a given location, with a small percentage of travel involving distances of 100 or more miles.

We find that near-duplicate detection alone is insufficient to detect all ads within a cluster. This follows the intuition that an author’s total set of ads is itself comprised of multiple self-similar subsets. While a near-duplicate detection approach can correctly identify subsets of ads linked to a single author, this process alone cannot attribute multiple clusters to a single author. However, we do find that the process can, with high precision and low recall, detect a significant subset of ads associated with a single author, allowing us

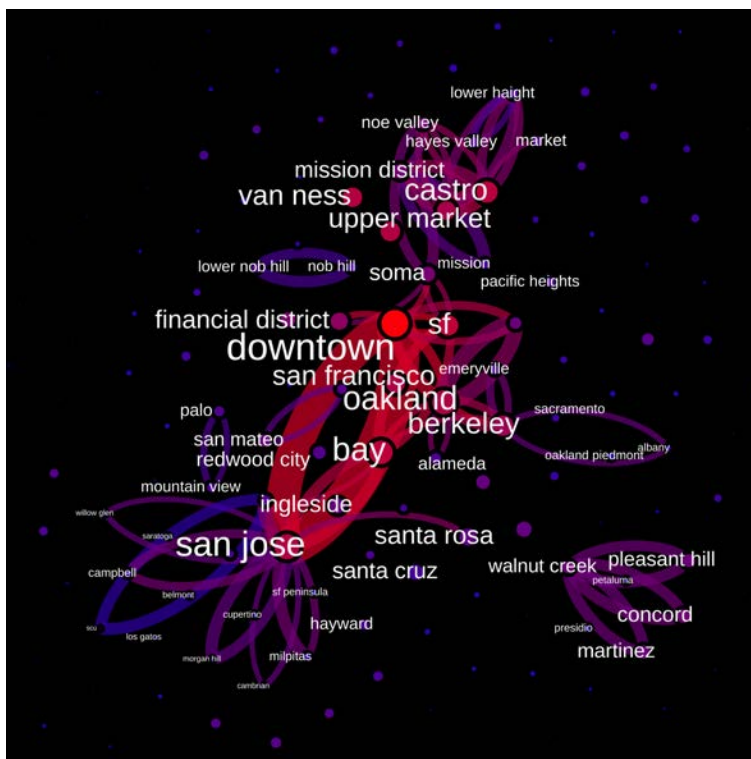


Figure 4.10: Visualization of the Craigslist *sfbay* subgraph ( $V=591$ ,  $E=7916$ ), showing all edges with weights  $> 25$ . Node color represents outdegree and edge color represents weight (blue=low and red=high). Observe how geographic clustering is clearly visible with respect to the spatial distribution of city locations.

to measure some anonymous individuals Craigslist posting behavior over time.

#### 4.4 Demographics: Extracting Age & Race/Ethnicity

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 [21, 69].

This section presents 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 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.

#### 4.4.1 Methods

##### 4.4.1.1 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 [106]. Each of these approaches is described in more detail below.

### First Mention (FM)

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 attempt at disambiguation is attempted in this approach, which provides our baseline performance measure.

1	5	'	9	'	140	<b>blk</b>	brn	30w	7c	.	.	mix	<b>blk</b>	/	<b>mexican</b>
	NNNN	N	<b>N</b>	N	N	N	NN	N	<b>A</b>	N	<b>A</b>				
2	all	race	welcom	.	.	.	<b>latino</b>	is	a	plus					
	N	N	N	NNN	<b>P</b>	N	N	N							

Figure 4.11: Excerpt of labeled output. Each term in a sequence is assigned a `label`  $\in$  { AUTHOR (A), PARTNER (P), NONE (N)} predicted based on 13 features, using conditional random fields. Given this labeling, both *First Mention* and *CRF-First* would incorrectly classify this ad’s author as Black. *First Mention* fails to disambiguate the first usage (as hair color) of “blk” while *CRF-First* only considers the first race term labeled as AUTHOR. *CRF-All*, which considers all AUTHOR labels, would correctly predict Biracial.

## Hybrid Method

This approach uses 11 Boolean and 2 nominal features (see below) to assign each term a `label`  $\in$  `{AUTHOR, PARTNER, NONE}` using linear-chain conditional random fields (CRFs). CRFs are undirected graphical models that, in the special case of a linear chain graph structure, can be used to efficiently label sequence data [70]. This machine generated label set is used by a rule-based classifier to then assign an ad to one of the 6 possible race/ethnicity categories (or Undisclosed if no `AUTHOR` tags are found or the labeled term isn't in our thesaurus). By constraining the machine learning step to detect all race/ethnicity mentions, independent of the class that mention belongs to, we help prevent overfitting in the less frequent categories in our training corpus.

For the rule-based classification, we consider two variations of the *First Mention* rule discussed above; (1) *CRF-First*; and (2) *CRF-All*. *CRF-First* uses the first `AUTHOR` labeled term in text to predict a race/ethnicity category, not just the first observed race/ethnicity vocabulary term. Second, *CRF-All* consider the set of all `AUTHOR` labels when assigning a category. If an ad contains `AUTHOR` labeled terms from more than one race/ethnicity terminology cluster, it is assigned to the Biracial class if those terms are separated by a slash (e.g., “white / black”) and are not contained within a list. See Figure 4.11 for an example labeling and its resulting classification.

The performance of these methods is evaluated on the annotation corpus, averaged over 10 trials, with each trial using stratified, 10-fold cross validation. To build our race/ethnicity terminology lexicons, the  $n-1$  folds of annotated training data are used to create the thesaurus used in labeling the documents of the  $n$ th fold. This cross-validation

approach ensures that we capture the effects of lexical acquisition in our classifier. For *CRF-First* and *CRF-All*, stemmed token word windows of size 3-9 were tested with all features, with 6 performing best overall. Other features such as part-of-speech tags were tested, but did not result in a statistically significant improvement in performance and were not included in the final feature set below:

- *Stemmed Term (Nominal)*: The current term and a 6-term window of all surrounding words.
- *Race/Ethnicity Category (Nominal)*: Name of this term’s parent race/ethnicity terminology cluster or None otherwise.
- *Digit + Unit of Measurement*: Term is a unit of measurement, e.g., “5’8” “130lbs.”
- *Metadata*: Term is part of age, location, or encounter tag metadata.
- *First Mention*: Term is the first labeled race/ethnicity term in ad text.
- *List*: Term co-occurs within a 5-term window of other race/ethnicity mentions, e.g., “totally into black, asian, latin or ethnic guys.”
- *Pluralization*: Term belongs to a race/ethnicity category and ends in “s”.
- *Partner Preference*: Qualifying terms within a 5-term window that express negation or partner preference, i.e.,  $\text{term} \in \{\text{no}, \text{not}, \text{into}, \text{none}, \text{only}\}$ .
- *Punctuation*: Term is a punctuation mark.
- *Slash*: Term is in a race/ethnicity category and is separated from another race/ethnicity term by a slash character.
- *Left Verb Argument*: Term is within a 5-term left-window of a set of left/right-

associative verbs:  $v \in \{\text{looking, seeking, wanted}\}$ . The verb “looking” is ignored if it occurs in the bigrams “good looking” or “nice looking.”

- *Right Verb Argument*: Same as above but for the right-term window.
- *PRP + Being Verb*: Term is preceded by a syntactic pattern of the form `PersonalPronoun + BeingVerb` (e.g., “I am”) where `BeingVerb`  $\in \{\text{am, 'm, :}\}$ .

#### 4.4.1.2 Calculating Demographic Rates

Using the methods described thus far, we extract race/ethnicity and age from all MSM ads. Demographic prevalence rates are calculated by collapsing all ads associated with a geographic region into a single bin  $L$  and checking for ads containing search terms associated with race/ethnicity and age attributes. Prevalence rate is then simply the percentage of MSM ads containing the search terms in question for a given location and time interval.

For our demographic analysis,  $L$  is defined as the set of all location tag toponyms contained within a U.S. county geographic boundary. Ads containing multiple toponyms, crossing multiple counties, are assigned fractional ad weights based on the number of county bins a location tag resolves to, given below by the function *geobins*. The *weight* of any given set of ads  $A$  is calculated as:

$$w(A) = \text{weight}(A) = \sum_{ad \in A} \frac{1}{\text{geobins}(ad)} \quad (4.7)$$

Formally, prevalence is calculated as follows: given  $M$ , the set of all MSM ads for a given location and time interval, the prevalence of a terminology cluster  $T$ , at location  $L$ , at

time window  $t_i$  is:

$$prev(T, L, t_i) = \sum_{tag \in L} \frac{w(\{ad \in M_{tag, t_i} : text(ad) \cap T \neq \emptyset\})}{w(\{ad \in M_{tag, t_i}\})} \quad (4.8)$$

We use a corpus-wide time window for all demographic analyses (7/1/2009 - 2/13/2012).

We also calculate a *usage* parameter for each geographic bin  $L$ , defined as the sum of all ads  $A$  associated with that location at time window  $t_i$ . This includes ads from the set of all monitored categories  $C$  (both commercial and MSM/non-MSM personal ads) and is used to weight regressions across geographic locations by measuring the degree to which the local community uses Craigslist services.

$$usage(L, t_i) = \sum_{category \in C} \sum_{tag \in L} w(\{A_{category, tag, t_i}\}) \quad (4.9)$$

These functions provide the input for all our analyses, which use a weighted, log-log transformed, ordinary least squares (OLS) regression to compare the relationship between disclosed race/ethnicity and age in Craigslist ads and the underlying population. For the census regressions, each county forms an observation, weighted by that location’s usage score. The percentage of Craigslist ads for each race/ethnicity or age group is the dependent variable and the percentage of that same group in the 2010 census data is the independent variable. All statistical analyses were done using R version 2.14.1 [136].

#### 4.4.2 Results

In the GOLD corpus, the CRF labeling classifier for the first stage of *CRF-All* and *CRF-First* had the following per-label category  $F_1$  scores: **AUTHOR** 0.86 (SD 0.01); **PARTNER** 0.78 (SD 0.02); and **NONE** 0.99 (SD 0.0). Detailed performance measures of the final output

Table 4.5: Author Race/Ethnicity Classification Performance Measures

Race/Ethnicity	Ad $n$	Method	Recall	Precision	$F_1$	$\pm\Delta F_1$
Biracial	14	<i>FM</i>	0.14	0.93	0.25	-
		<i>CRF-First</i>	0.09	<b>1.00</b>	0.17	-31.9%
		<i>CRF-All</i>	<b>0.60</b>	0.67	<b>0.63</b>	155.8%
Hispanic/Latino	34	<i>FM</i>	<b>0.85</b>	0.71	0.77	-
		<i>CRF-First</i>	0.83	0.79	<b>0.81</b>	4.6%
		<i>CRF-All</i>	0.72	<b>0.89</b>	0.80	3.1%
Black	27	<i>FM</i>	<b>1.00</b>	0.57	0.72	-
		<i>CRF-First</i>	0.97	0.69	0.81	11.5%
		<i>CRF-All</i>	0.94	<b>0.78</b>	<b>0.85</b>	18.3%
Asian	19	<i>FM</i>	0.90	0.74	0.81	-
		<i>CRF-First</i>	<b>0.95</b>	0.90	<b>0.92</b>	13.4%
		<i>CRF-All</i>	0.88	<b>0.94</b>	0.91	12.0%
Caucasian	136	<i>FM</i>	0.87	0.85	0.86	-
		<i>CRF-First</i>	<b>0.91</b>	0.92	0.91	6.4%
		<i>CRF-All</i>	0.90	<b>0.94</b>	<b>0.92</b>	6.8%
Undisclosed	299	<i>FM</i>	0.88	<b>0.95</b>	0.91	-
		<i>CRF-First</i>	<b>0.95</b>	0.94	<b>0.94</b>	3.6%
		<i>CRF-All</i>	<b>0.95</b>	0.92	0.93	2.2%
Hawaiian/Pacific-Islander	3	<i>FM</i>	<b>0.33</b>	0.23	0.27	-
		<i>CRF-First</i>	<b>0.33</b>	<b>1.00</b>	<b>0.50</b>	85.8%
		<i>CRF-All</i>	<b>0.33</b>	<b>1.00</b>	<b>0.50</b>	85.8%

of *CRF-First* and *CRF-All* compared to the baseline *First-Mention* algorithm are found in Table 4.5. *CRF-All* performed best overall, with statistically significant improvements between 2.2% to 156% in  $F_1$  score over the baseline ( $p < 0.05$  using a two-sided t-test) and scored the highest precision values for every category except Biracial and Undisclosed. *CRF-All* performed best at identifying Caucasian and Asian ad authors, with  $F_1$  scores of 0.92 and 0.91 respectively. Biracial and Hawaiian/Pacific-Islander classes were the worst performing category overall, with  $F_1$  scores of 0.63 and 0.50.

Table 4.6 includes counts and weighted OLS results comparing the race/ethnicity

Table 4.6: Craigslist Race/Ethnicity vs. 2010 Census

Race/Ethnicity	[Ad $m$ ]	Geotype	CRAIGSLIST Corpus			PHONE Corpus		
			$R^2$	$\beta$	[95% CI]	$R^2$	$\beta$	[95% CI]
Biracial	366K	Metro	0.43	0.82	[0.72, 0.92]	0.26	0.73	[0.60, 0.86]
	14K	Micro	0.53	0.60	[0.55, 0.64]	0.32	0.70	[0.62, 0.79]
	31K	County	0.76	0.76	[0.73, 0.79]	0.47	0.72	[0.67, 0.77]
Hispanic/Latino	964K	Metro	0.91	0.81	[0.78, 0.84]	0.81	0.81	[0.77, 0.85]
	20K	Micro	0.56	0.54	[0.50, 0.58]	0.25	0.58	[0.50, 0.66]
	37K	County	0.66	0.61	[0.58, 0.64]	0.54	0.71	[0.66, 0.75]
Black	992K	Metro	0.71	0.36	[0.34, 0.39]	0.54	0.46	[0.41, 0.5]
	39K	Micro	0.44	0.33	[0.30, 0.37]	0.24	0.51	[0.44, 0.59]
	42K	County	0.45	0.24	[0.22, 0.26]	0.19	0.30	[0.26, 0.34]
Asian	359K	Metro	0.86	0.65	[0.63, 0.68]	0.63	0.57	[0.52, 0.61]
	11K	Micro	0.54	0.55	[0.51, 0.60]	0.36	0.58	[0.51, 0.64]
	39K	County	0.66	0.55	[0.53, 0.58]	0.44	0.44	[0.41, 0.47]
Caucasian	4.9M	Metro	0.04	-0.24	[-0.35, -0.13]	0.04	-0.24	[-0.36, -0.11]
	216K	Micro	0.05	-0.38	[-0.52, -0.24]	0.01	-0.27	[-0.52, -0.02]
	218K	County	0.04	-0.18	[-0.23, -0.12]	0.10	-0.46	[-0.55, -0.37]
Undisclosed vs. Caucasian	18.4M	Metro	0.34	0.23	[0.20, 0.27]	0.39	0.31	[0.27, 0.35]
	1M	Micro	0.19	0.21	[0.17, 0.24]	0.08	0.36	[0.26, 0.45]
	963K	County	0.23	0.13	[0.11, 0.14]	0.07	0.19	[0.15, 0.24]
Native	20K	Metro	0.42	0.40	[0.35, 0.45]	0.21	0.40	[0.32, 0.47]
Hawaiian/ Pacific Islander	2K	Micro	0.94	0.37	[0.37, 0.38]	0.90	0.47	[0.46, 0.48]
	4K	County	0.96	0.45	[0.44, 0.45]	0.88	0.55	[0.54, 0.56]

distributions for county, micropolitan, and metropolitan areas in CRAIGSLIST vs. 2010 census data. Most ads, 71%, did not disclose race/ethnicity information. Caucasian formed the majority-identified category with 18.5%, followed by Black 3.7%, Hispanic/Latino 3.6%, Biracial 1.4%, Asian 1.4%, and Hawaiian/Pacific-Islander 0.1%. All reported census regressions were statistically significant at  $p < 0.05$ . Figure 4.12 shows scatter plots for the metropolitan CBSA component of this analysis. Only Caucasian had a negative coefficient value, with disclosure rates decreasing as the percentage of Caucasians increased in a given

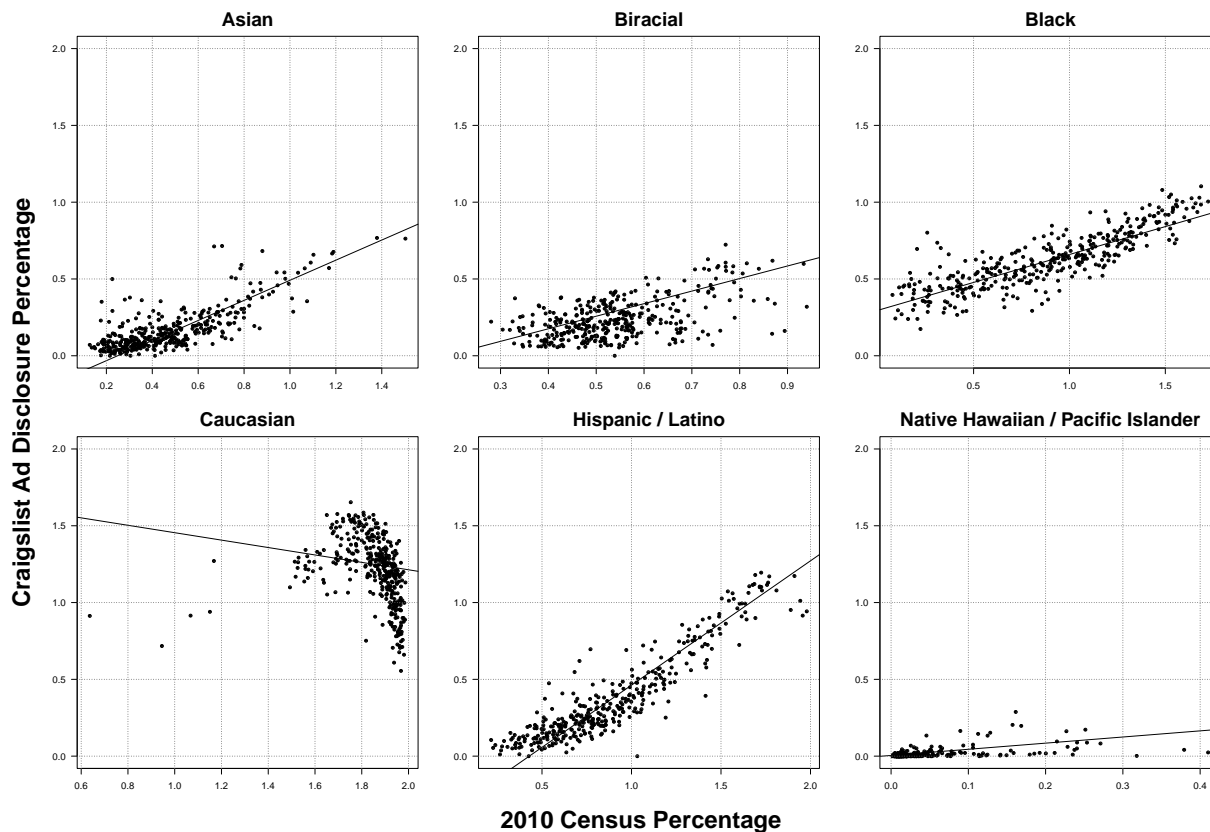


Figure 4.12: Scatter plot of log-log regression results for 2010 census race/ethnicity percentages (x-axis, independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (y-axis, dependent variable). Plots are for all CBSA metropolitan geographic boundaries containing at least 1000 ads. For the Caucasian plot (lower left corner) the percentage of race disclosures begins to drop precipitously in the interval 64%-100% (1.8-1.2), suggesting that as the population grows more homogeneously Caucasian there is less need to mention race in ads.

geographic boundary. Rate of non-disclosed race/ethnicity is examined more closely in Figure 4.13, which shows a scatter plot of 2010 Census geographic Shannon entropy compared to the percentage of ads with undisclosed race/ethnicity. Shannon entropy is a measure of type diversity; it increases as members are more evenly distributed across categories (i.e., evenness) [83]. Note how as a population grows more homogenous and less evenly distributed across types (i.e., lower entropy), the rate of undisclosed race/ethnicity ads increases.



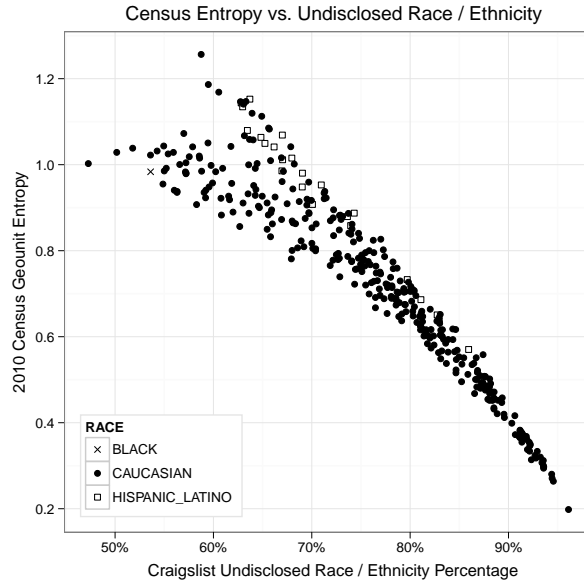


Figure 4.13: Scatter plot of 2010 Census geographic Shannon entropy (y-axis), measured across race/ethnicity categories vs. the percentage of ads with undisclosed race/ethnicity (x-axis). The shape of each point indicates the majority racial/ethnic group in that geographic boundary. Note how as a population grows more homogenous and less evenly distributed across types (i.e., lower entropy), the rate of undisclosed race/ethnicity ads increases.

Overall, almost all categories were significantly correlated with known subpopulation makeup, with metropolitan areas tending to have the highest  $R^2$  values. Metropolitan Hispanic/Latino ads were the most correlated with an  $R^2=0.91$  (coefficient 0.81, 95% CI [0.78, 0.84]), followed by Asian, Black, Biracial and Hawaiian/Pacific-Islander ads. The least correlated were Caucasian and Undisclosed (using Caucasian census values) ads. Comparing the CRAIGSLIST analysis with PHONE, we find regression coefficients and  $R^2$  values are very similar in both ad sets and statistically significant in all classes.

## 4.5 Conclusion

Utilizing the unstructured data of Craigslist ads for public health surveillance requires solving several key NLP subtasks. This section has presented systems for addressing several

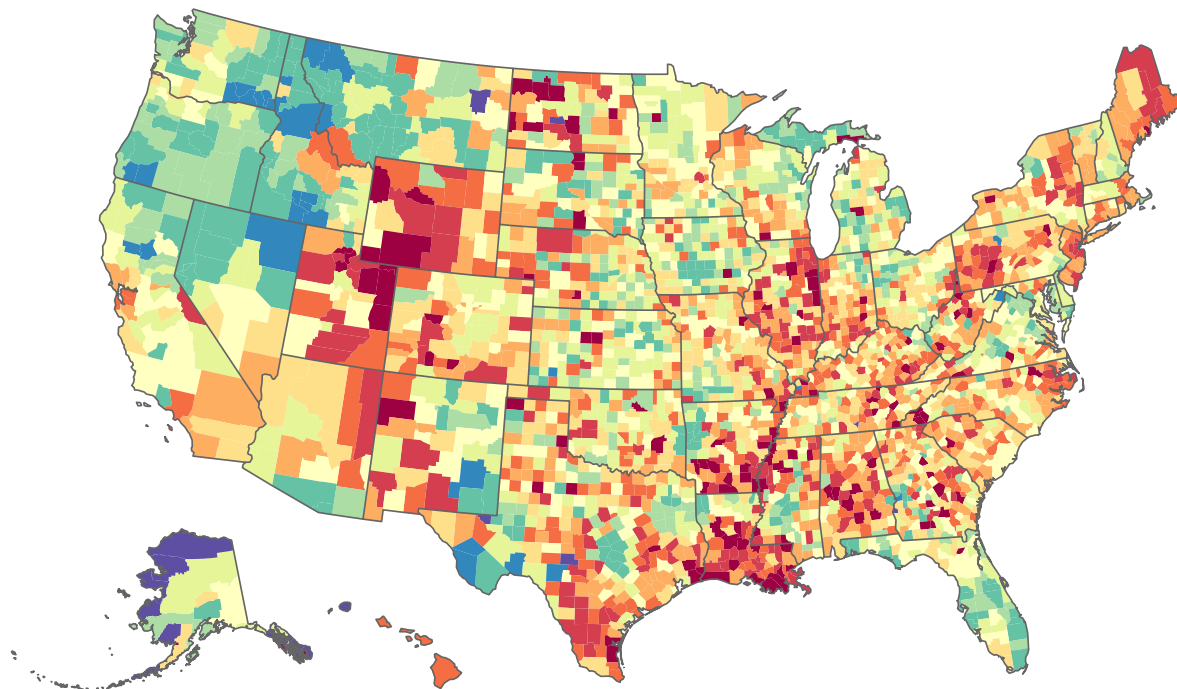


Figure 4.14: Percentage of MSM ads out of all sampled Craigslist ads. In some states such as Louisiana, MSM ads are the majority of sampled ads, while in others such as Washington and Oregon, most ads are commercial ads (i.e., an ad from the categories {legal services, appliances, pets, furniture, and parking}).

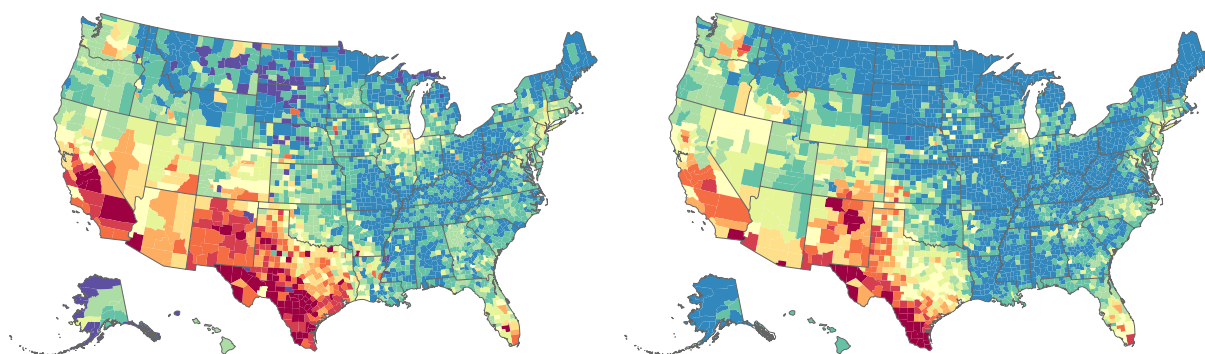


Figure 4.15: Hispanic/Latino Craigslist county-level percentages as determined by CRF-All (left) vs. 2010 Census data (right).

of those tasks: 1) geocoding ads; 2) identifying subsets of ads written by the same author; and 3) efficiently extracting demographic data from ad text. All of these systems are crucial components of our proposed behavioral surveillance pipeline. Moreover, by decomposing the larger surveillance task into these systems, we provide a structure for future improvement, as each individual component can be improved upon independently.

## CHAPTER 5 COMMUNITY SEXUAL BEHAVIOR SURVEILLANCE

### 5.1 Modeling Sexual Behaviors in Text

The final component of a sexual health surveillance pipeline is identifying relevant behavioral information in the text of Craigslist ads. However, unlike the tasks described in the previous section, which lend themselves to concrete objectives and gold standard training data sets (correctly normalizing geographic entities, clustering authors, etc.) characterizing concepts or topics in text is considerably more challenging from a validation perspective. What defines a “good” representation of a sexual behavior? Do we wish to identify text that correlates with known high risk sexual behaviors, creating more interpretable results for public health analysts? Or do we wish to view disease forecasting as our objective and accept a more opaque model? In this section, we present three approaches for generating sexual behavior topics.

Unlike census data, there is no large scale sexual behavior data against which to validate our results. Instead we present several proxy validation methods, showing that Craigslist ads reflect other quantifiable real-world phenomena. We first show that the geographic scope and scale of Craigslist is suited to supplement surveillance efforts and that personal ads capture a spectrum of relevant information. To validate the demographic information extracted from Craigslist we compare it to published data from traditional census surveys. Secondly, to demonstrate how the MSM ad content captures real world temporal and geographic variation, we examine the frequency at which posters mention snowfall in

ads to construct a time series model for predicting actual snowfall. This same methodology is then extended to capture ad author’s self-disclosed HIV/AIDS status, which is correlated to county-level HIV/AIDS prevalence rates, demonstrating how Craigslist data can be used as a proxy for HIV prevalence. Finally, we analyze how ad content also encodes a variety of known high-risk behaviors, which can then be used to predict syphilis rates at an ecological level.

### 5.1.1 Dictionary/Search Term-based Methods

Through manual inspection of ads, we constructed 4 behavior search term clusters (or *concepts*) corresponding to known high-risk sexual behaviors in the MSM community: self-disclosed HIV+ status (HIV+), unprotected sex requests or “barebacking” (BB), amyl nitrate use (POPPERS), and “party and play” requests (PNP) – a slang term describing sexual encounters involving the use of illegal drugs. Term clusters were built through manual inspection of ads and by consulting the Wikipedia pages for “Party and Play” and “Poppers” [150, 151]. We also defined sets of left/right negation terms which are used to filter out false positive matches (“no pnp”) or identify negated matches (“no rubbers”) within a window size of 5 terms. Listed below are each concept’s search term set for positive matches:

**HIV+:** (poz | poz4poz | positive | poz-friendly | hiv positive | hiv pos | hiv poz | pos play | pos bottom | pos top)

**POPPERS:** (popper | poppers | popper-friendly)

**BB:** (bb | bareback | barebacking | raw | bare back | bare ride | bare play)

<NEGATION> + (condom | condoms | rubber | rubbers | safe | protection)

**PNP:** (pnp | party-friendly | partying | meth | tina | blow clouds | cialis  
| viagra | party favors)

**Negation:** *left:* (not into | not | no) *right* (free)

## 5.1.2 Word/Phrase Learned Representations

### Word Embeddings

In addition to terminology identified manually in terminology clusters above, we also utilize the set of annotated behaviors identified in the GOLD corpus. These terms and phrases are then expanded using word embeddings generated with word2vec [94] using the Skip-gram model. A formal description of this model is provided in §2.3.4. All MSM ads are preprocessed, normalized to lowercase, and all digits replaced with the character N. We then merge common bigrams into single tokens using Pointwise Mutual Information [94] as described by Mikolov et al.

$$score(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i) * P(w_j)} \quad (5.1)$$

where probability is simply the raw corpus term frequency, thresholded at a minimum occurrence of 10 to filter out low frequency bigrams. The final embedding model is trained using 300-dimensional vectors and a window size of 2.

Table 5.1 shows race/ethnicity terms (in bold) and their 10 nearest neighbors. Note that while there are some obvious errors (*af-am* under caucasian) terminology largely clusters with racial/ethnic identity groups. Note how *hispanic*, a term with often inconsistent usage when applied to American Hispanics, is a member of caucasian and latino categories. Table 5.2 illustrates Craigslist drug slang learned using KNN in word embedding space. Cocaine:

*yayo, ski, ski slopes.* MDMA: *thizz, molly.* Marijuana: *420, herb, pot, ganja, mary jane, oscar.* Methamphetamine: *tina/teena, clouds, shards.* Amyl Nitrate: *poppers.* Neighboring terms for frequent drugs (marijuana and amyl nitrate) largely consist of synonyms, while more ambiguous words like *pills* neighbor a variety of prescription medications.

Figure 5.1 shows wordcloud [46] versions of the 10-nearest-neighbors for “bb” (unprotected sex) and “favors” (illicit drugs) where size corresponds to relative frequency. Note how in all examples, most terms are at least semantically related and often true synonyms of the original query word. Finally, Figure 5.2 shows a 2D visualization of all words used on Craigslist, with specific entity types highlighted using GOLD entity definitions.

Table 5.1: Craigslist Race/Ethnicity Terms: 10-nearest Neighbors

<b>caucasian</b>	<b>black</b>	<b>latino</b>	<b>pacific_islander</b>	<b>chinese</b>	<b>biracial</b>
caucasian	balck	latin	islander	korean	mixed
w_hite	blaack	hispanic	filipino	vietnamese	bi-racial
caucsian	blacl	lation	pac_islander	japanese	mixed_race
cacausian	blsck	mexican	malaysian	viet	multi_racial
causasian	blacck	lationo	pinoy	malaysian	euro-asian
cuacasian	bkack	laino	chinese_filipino	thai	lightskinned
caucassian	blaxk	hipanic	guamanian	chines	multi-ethnic
wihte	blkack	laitno	tahitian	indonesian	lightskined
anglo	blak	hispanice	chinese_japanese	okinawan	black&dominican
hispaniic	b;ack	latio	filipno	taiwanese	domican
af-am	blavk	hisp	maori	filipino	latinno
southasian	blk	latiino	flipino	pinoy	korean_american
cacasian	blakc	mex	hawaiian	chinese_japanese	half-black

Table 5.2: Craigslist Drug Slang Terms: 10-nearest Neighbors

pot	pills	poppers	clouds	tina	shrooms	hydrocodone
herb	viagras	popers	slopes	molly	ghb	percocet
420	ghb	poers	ski_slopes	mary_jane	benzos	klonopin
weed	pill	pppers	ganj	maryjane	zannies	ritalin
booze	tablets	poopers	cloudz	yayo	roxies	vicoden
four-twenty	xanax	popprs	brews	teena	addies	norcos
alcohol	viagra	pprs	blunts	mary_j	yayo	xanax
marijuana	blue_pills	poprs	party_goods	gina	hydrocodone	vicodin
ganja	cialis	poppes	tina_clouds	go-fast	norcos	benzos
fourtwent	tabs	ppprs	oscars	ghb	vicodin	cialas
maryjane	adderall	popps	shards	miss_tina	adderall	oxycodone
herbs	viagara	popperz	ncaa_games	thizz	viagara	vics

Figure 5.1: Word embeddings  $k$ -nearest-neighbors for “bb” (left) and “favors” (right), where “bb” is slang for unprotected sex and “favors” is slang for illicit drugs.

## Sentence Embeddings

The word2vec approach of modeling context can be extended to model sentences, where one vector per sentence is trained to predict all words within a sentence instead of a fixed context window [71]. Since every sentence in a corpus is assigned an embedding, there are potentially multiple embeddings per unique sentence. Ideally, we would like to model some aspect of paraphrase, where, for example, representations for “no raw”, “condoms a must”, and “don’t do bb”, would be near each other in embedding space. However, in



## Craigslist Word Embeddings

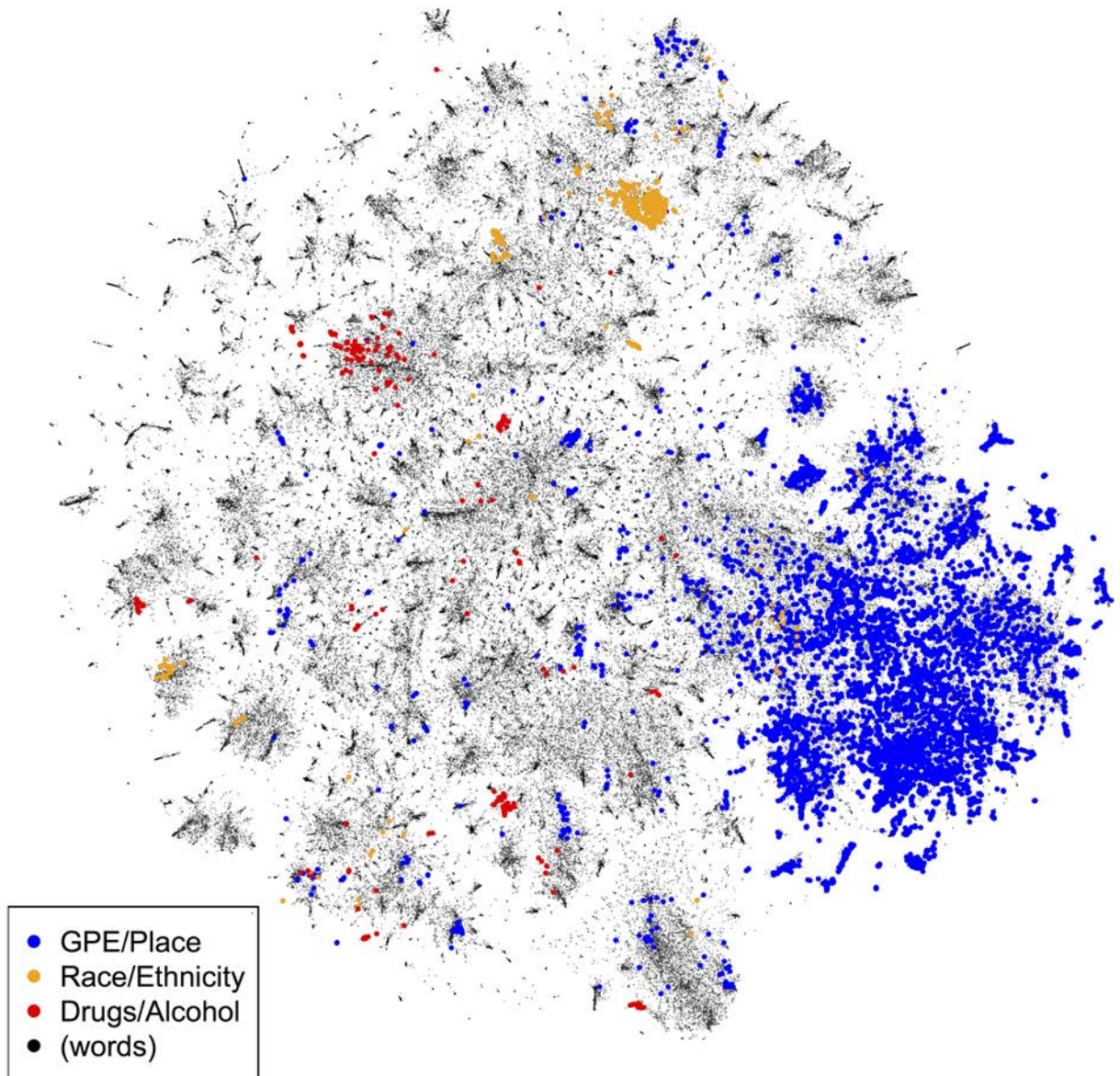


Figure 5.2: 2D stochastic neighbor embedding of Craigslist word embeddings. Point labels are determined through annotations in GOLD and geographic entities identified by TopoLinker. Note how much of the space contains geographic terms. Geographic sub-clusters frequently correspond to specific usage classes (e.g., university acronyms) or represent site-specific clustering. Stochastic neighbor embedding generated using Barnes-Hut T-SNE [147].

practice this is quite challenging. To illustrate how well some sentence embeddings encode meaning, we look at two simple examples of binary classification problems: disambiguating drug mentions and negation detection.

Table 5.3: Drug Euphemisms “Snow”: 10-nearest Neighbors

<b>love to play in the snow</b> (refers to <b>cocaine</b> )	<b>i like the snow and snowy days</b> (refers to <b>weather</b> )
love snow	snowy weekend
i love snow	its supposed to snow today
have snow	rainy the last N days
you have the snow	<i>cloudy snow</i>
love the snow	cloudy rain day in ames
want snow	this cold and dreary winter weekend or week
<i>fuck the snow</i>	this is real cooler day in amarillo after a snowy winter
in the snow	i am real it is snowy today
i have the snow and the ass	put snowy
<i>lots of snow today</i>	sunny days in subject

The words “snow”, ”ski”, or ”slopes” are all used as euphemisms for cocaine as well as weather and outdoor activities; simple keyword-based prevalence estimates collapses both meanings. Table 5.3 contains the 10-nearest neighbors for euphemistic sentences discussing cocaine (snow) and those discussing weather, where the true meaning of each sentence was determined manually by reading the sentence’s parent ad. In absence of ad context, it is difficult even for a human to disambiguate euphemistic drug references from actual weather chatter, although possessive mentions of snow (e.g., “have snow”, “you have the snow”) have awkward phrasing under the predominant meaning of snow. We see that while there are a few incorrect entries (italicized) sentence embeddings do manage to divide drug vs. weather

usages. Sentence representations are modeled using individual word embeddings and these results show how certain words can dominate sentence meaning: “play” and “snow” dominate drug references while “snow” and “days” are closer to discussion of actual temporal weather phenomena. The phrase “cloudy snow” is a drug reference, but “cloudy” is used most frequently to describe weather, while “clouds” refers to meth or marijuana.

Table 5.4: Negated Sentences: 10-nearest Neighbors

<b>i am into bareback</b>	<b>i am not into bareback only play safe</b>
looking for bareback play	i am not into bareback only play safe
looking for bareback play	safe play only no bareback
bareback	not into bareback
bareback	<i>into bareback only</i>
love bareback	<i>must be into bareback</i>
bareback	<i>bareback only</i>
bareback only	<i>bareback only</i>
bareback top	not into bareback
bareback only	no bareback fucking safe sex only
bareback only	<i>just be into fucking bareback</i>

Table 5.4 shows the 10-nearest neighbors for a sentence requesting unprotected sex and a negated form of the same meaning. Note that while some aspects of negation are captured in the righthand-side of the table, only 50% are semantically correct. Moreover, no synonyms for bareback are included in the top results. This is partly due to the fact that every sentence, unique or not, gets a separate representation introducing considerable computation costs and redundancy in the top matched neighbors. One could conceivably implement a KNN classifier to determine negation, but this elides that fact that the naive approach

of extending word embeddings fails to capture important sentence semantics. This partly illustrates the need for more sophisticated methods for generating sentence or document level representations [131].

### 5.1.3 Topic Modeling

While dictionary and semi-supervised methods of lexical acquisition perform well in many tasks, they still require a significant amount of human effort to construct labeled data. Labeling data requires some level of domain expertise and while fields like medicine have considerable resources and experts to devote to maintaining rich ontologies and lexicons, social media typically must rely on crowdsourced assets like Wikipedia or the Amazon Turk. A more subtle issue is human-curated dictionaries and ontologies reflect a set of assumptions on the annotator, ones that may or may not reflect the latent structure of the data itself. This fact, coupled with the ever-evolving nature of social media, make it important to explore more automated approaches of modeling terminology.

One alternative approach is to use an unsupervised method that models topics or concepts of discussion directly from text data, such as Latent Dirichet Allocation (LDA) [22]. Broadly speaking, LDA assumes documents are a mixture of several latent conceptually-related categories called *topics*. These topics are modeled as a distribution of words  $d$  over  $k$  topics where  $k$  is specified as a hyperparameter. There are additional concentration hyperparameters  $\alpha$  and  $\beta$  which control the distribution of topics per document and words per topic respectively. The optimal value of  $k$ ,  $\alpha$ , and  $\beta$  are typically determined experimentally or otherwise guided by domain knowledge.

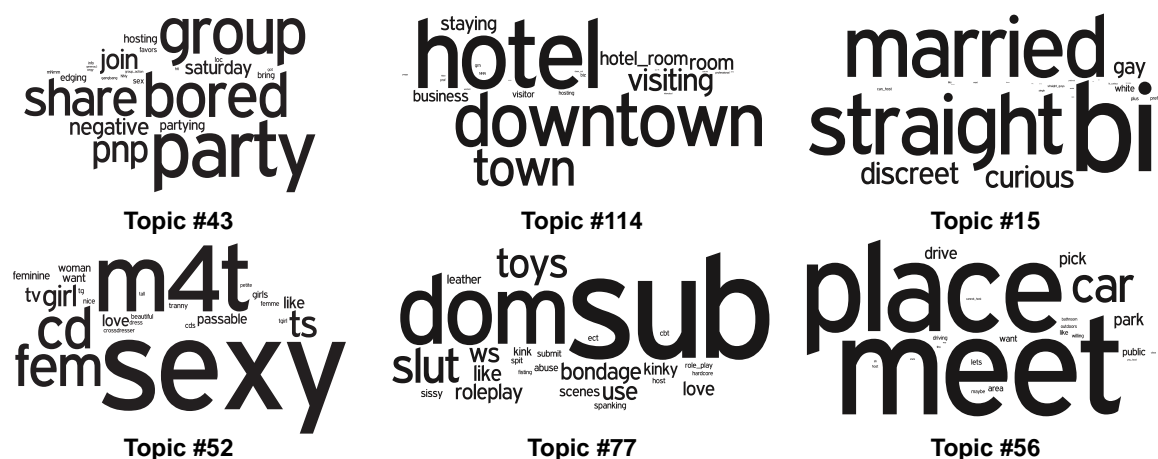


Figure 5.3: Some example topics generated from Craigslist MSM ads using LDA ( $k=125$ ). Word size corresponds to per-topic probability.

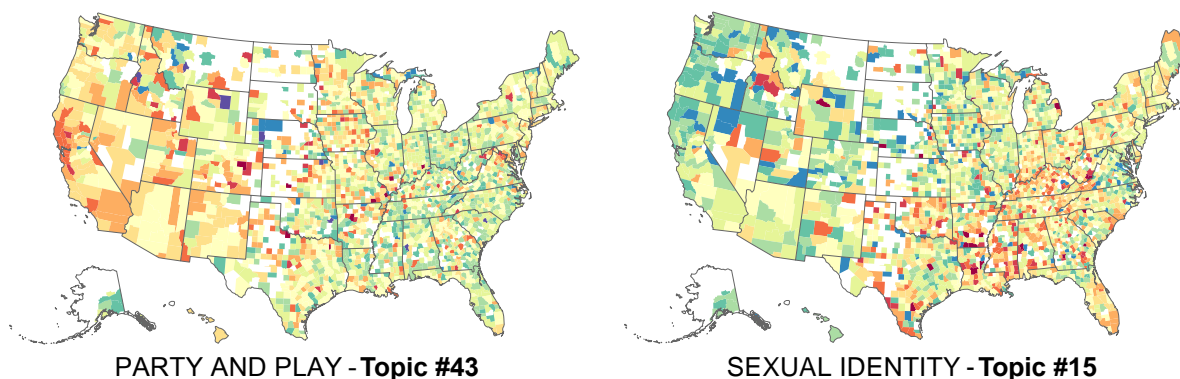


Figure 5.4: Spatial distribution of “Party and Play” (PNP) topic (left) and “Married/Bi/Straight” (SEXUAL IDENTITY) topic (right). PNP is higher prevalence on the west coast while SEXUAL IDENTITY is generally higher in the south and east coast, with clusters in Idaho, Nevada, Utah, and Wyoming.

For Craigslist MSM ads, we filter all document to include only the top 5K most frequent words and remove all geographic entities identified by TopoLinker. The advantage of this is that we generate generalized, country-wide topics; the disadvantage is that we discard a high degree of localized, regional behavioral data. Both perspectives have value from a public health perspective, but this work only examines country-wide topics. We

experimented with 25 to 500 topics and symmetric/asymmetric priors and found through manual inspection that  $k=125$  with asymmetric priors ( $\alpha$  auto-configured using gensim [112]) generated the most interpretable topics. Figure 5.3 illustrates some representative topics generated by LDA. Observe that a topic corresponding to party and play (upper left-hand corner) is discovered automatically using LDA. Figure 5.4 shows the spatial distribution of some of the same topics.

## 5.2 Geo-temporal Validation

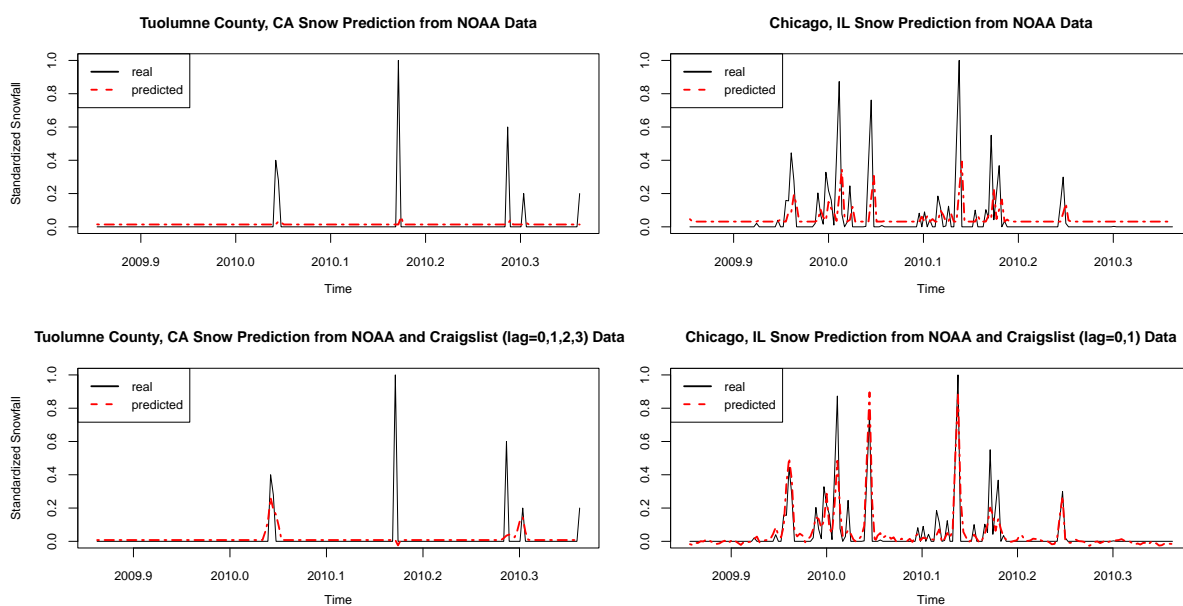


Figure 5.5: Two time series models for predicting NOAA-reported snowfall using daily, standardized Craigslist snow-related ad rates. On the left: fitted values plot of predicted snowfall for Tuolumne County California; on the right: Chicago, Illinois (Cook, DuPage, Kane, Lake, and Will counties). The top models estimate snowfall using only NOAA data while the bottom models incorporate Craigslist ad content and do a better job overall of predicting snowfall. The significant temporal association between Craigslist ads and snowfall as well as the reduction in error when using ads as a predictor of snowfall infers the usefulness of Craigslist data as a predictor of phenomena in the physical world.

It is impossible to directly validate the accuracy of the information within Craigslist ads. Thus, we use an indirect approach to establish that Craigslist personal ads do actually contain up-to-date, real-world temporal information and to validate the accuracy of our geocoding framework. Specifically, we exploited the observation that Craigslist posters sometimes refer to local weather conditions, as they could be relevant to their availability (e.g., snow complicating plans to have Internet mediated encounters). We validated the accuracy of this time and location sensitive snow-related information within Craigslist ads against actual weather data. Using measurement data from the National Oceanic and Atmospheric Administration’s (NOAA) “U.S. Daily Snowfall and Snow Depth Data” website [103], we constructed daily snow fall time series for Chicago, Illinois (Cook, DuPage, Kane, Lake, and Will counties) and Tuolumne County along the Sierra Nevada mountains in California. From October 1st, 2009 to March 31st, 2010, we computed daily snowfall rates as the sum of reported values from all weather stations occurring within each county.

For this same interval, we built a time series capturing snow-related discussions in MSM Craigslist ads. For each county, we calculated a daily snow-related ad ratio, defined as the number of ads matching a set of regular expressions corresponding to snow-related ad content divided by the count of all ads in all monitored categories for that same day. To investigate the temporal association of snow-related discussions and actual snowfall, we computed a cross-correlation function (CCF). Because cross-correlation between time series can yield spurious results due to the effects of common temporal patterns, we employed a prewhitening process [38]. Using the temporal association suggested by the CCF, we formulated a seasonal autoregressive integrated moving-average (ARIMA) time series regression

model with autocorrelated errors. Finally, we evaluated how well the Craigslist time series data performed in predicting daily reported NOAA snowfall rates in terms of overall reduction in observed error when including Craigslist as a predictor.

Figure 5.5 shows the time series models for Chicago, IL and Tuolumne County, CA. For Chicago, we found a large and significant association between observed snowfall and snowfall mentioned concurrently and the previous day in Craigslist, indicating people discuss snowfall on Craigslist the day of and the day prior to the actual snow event. Incorporating the 2 Craigslist variables into our model resulted in an observed reduction in error of 75.6% over a model using no Craigslist data. For Tuolumne County, CA we found a large and significant association with concurrently mentioned snowfall1, 2, and 3 days previously. Incorporating these variables into our model resulted in a 9.2% reduction in observed error. Note, snow is much less common in California and is restricted to more mountainous regions.

### **5.3 Links between Sexual Behaviors and Disease**

#### **5.3.1 Materials/Methods**

##### **5.3.1.1 CDC County-level Disease Data**

All disease data from this work is provided by the Centers for Disease Control's (CDC) National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) Atlas [50]. The Atlas provides state and county-level disease data for primary/secondary syphilis 2008-2013 and HIV incidence/ prevalence 2010-2012.



### 5.3.1.2 Computing Topic Prevalence Rates

Using the dictionary-based term clusters described in §5.1.1, we define the prevalence rate for a single term as the relative frequency:

$$prev(w, county) = \frac{freq(w, county)}{\sum_{\hat{w} \in V} freq(\hat{w}, county)} \quad (5.2)$$

where the denominator consists of all ads geolocated to a given county, for a given time period under investigation. Term clusters are simply a summation across all terms in the cluster.

Each cluster’s prevalence rate is computed for every county and state for the United States. To show that behavioral information itself can be extracted from ads and correlated with disease rates, 2010 BB and PNP rates are also correlated with California syphilis incidence rates. We used a weighted least squares, log-log-transformed ordinary least squares model on county-level, 2010 primary and secondary syphilis rates for males. Our dependent variable was syphilis incidence rates and our independent variable was the prevalence rate of each ontology definition under examination.

All statistical tests were calculated using R, version 2.12.0 [135]. Time series analyses are calculated using the R-package Time Series Analysis (TSA) version 0.98 [24].

## 5.3.2 Results

### 5.3.2.1 State-level Disease Correlations

At the state level, statistically significant correlations ( $p < 0.05$ ) between HIV incidence, early latent syphilis, and primary/secondary (PS) syphilis were found in ads disclosing Caucasian race. The most correlated behavior term set was PNP, followed by HIV and BB.

Figure 5.6 shows plots of 2011 Caucasian PNP rates vs. 2011 HIV and PS syphilis incidence rates. 2011 HIV incidence was correlated with 2011 PNP rates, with an adjusted r-squared 0.38, coefficient 0.43, 95% CI [0.26,0.59]. 2009 and 2010 rates were similarly correlated with PNP rates. 2011 PS syphilis incidence and PNP rates were also correlated, with an adjusted r-squared 0.55, coefficient 0.60, 95% CI [0.41,0.78]; 2010 PS syphilis looked similar. 2011 Caucasian HIV-disclosure rates were also correlated with 2011 PS syphilis, early latent syphilis and HIV incidence.

2009 BB requests were correlated with 2009 PS syphilis rates in Black males (r-squared 0.21, coefficient 0.20, 95% CI [0.15,0.94]) and self-disclosed HIV status in 2011 ads was correlated with 2011 PS syphilis rates (r-squared 0.15, coefficient 0.23, 95% CI [0.04, 0.42]). There were no statistically significant correlations between disease rates and behavior prevalence rate in ads disclosing Asian or Hispanic/Latin origins. The percentage of MSM ads for a given state was not predictive of any of the diseases under analysis.

### 5.3.2.2 Self-reported HIV Status

For MSM Craigslist postings, we found that county-level-HIV rates and ad rates for our ontology for HIV status and requests for high-risk encounters were highly correlated at an ecological level. Figure 5.7 shows log-log scatterplots for the relationship between HIV/AIDS prevalence and HIV+ and BB ontologies. As the graph indicates, the proportion of ads from the HIV-positive individuals is positively associated with actual HIV rates. In addition, the proportion of unprotected sex requests and PNP requests are also positively associated with HIV rates. Specifically, for HIV+ ads we found r-squared=0.41, coeff. 0.34, and

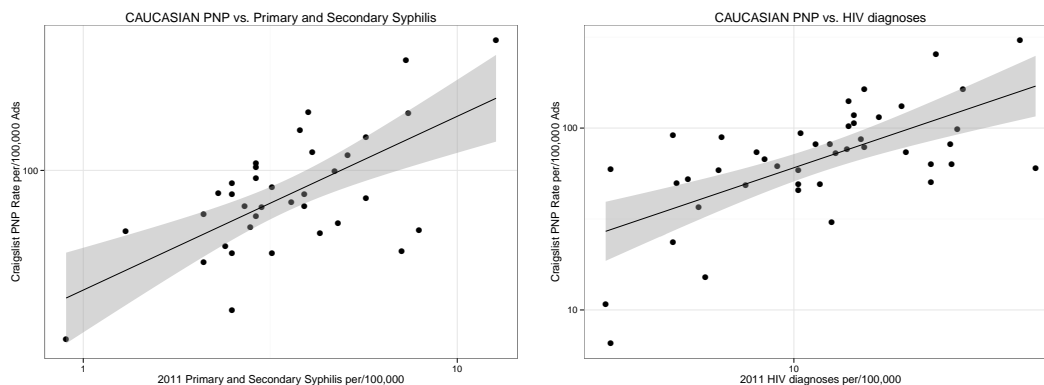


Figure 5.6: U.S. State-level analysis of 2011 Caucasian PNP (party and play) behavior rates vs. 2011 Caucasian PS syphilis (left) and HIV (right) incidence rates using a weighted log-log OLS regression. Disease incidence rates are on the x-axis and PNP rates are on the y-axis. All results are statistically significant ( $p < 0.05$ ). HIV incidence vs. PNP rates: adjusted r-squared 0.38, coefficient 0.43, 95% CI [0.26,0.59]; and PS syphilis incidence vs. PNP rates: adjusted r-squared 0.55, coefficient 0.60, 95% CI [0.41,0.78]. 2009 and 2010 HIV incidence were similarly correlated with PNP rates, as was 2010 PS syphilis.

CI=[0.19,0.49]; PNP r-squared=0.48, coeff.= 0.38, CI=[0.24, 0.52]; and BB r-squared=0.5, coeff.=0.44, CI=[0.28, 0.60]. For all values,  $p < 0.001$ . Thus, data from Craigslist can be used as a proxy measure for HIV prevalence.

### 5.3.2.3 Syphilis Forecasting in California

BB ad rates were positively correlated with syphilis rates in California: a 1% increase in unprotected sex requests was associated with a 0.31% increase in syphilis rates (coeff. 0.31; p-value  $< 0.001$ , 99% CI [0.17, 0.46]). PNP rates were also correlated with syphilis rates (coeff. 0.26, p-value  $< 0.001$ , 99% CI [0.12, 0.40]).

### 5.3.3 Discussion

Our results demonstrate that Craigslist personal ads provide timely behavioral data across a large geographic scale. Specifically, we show that that these ads contain informa-

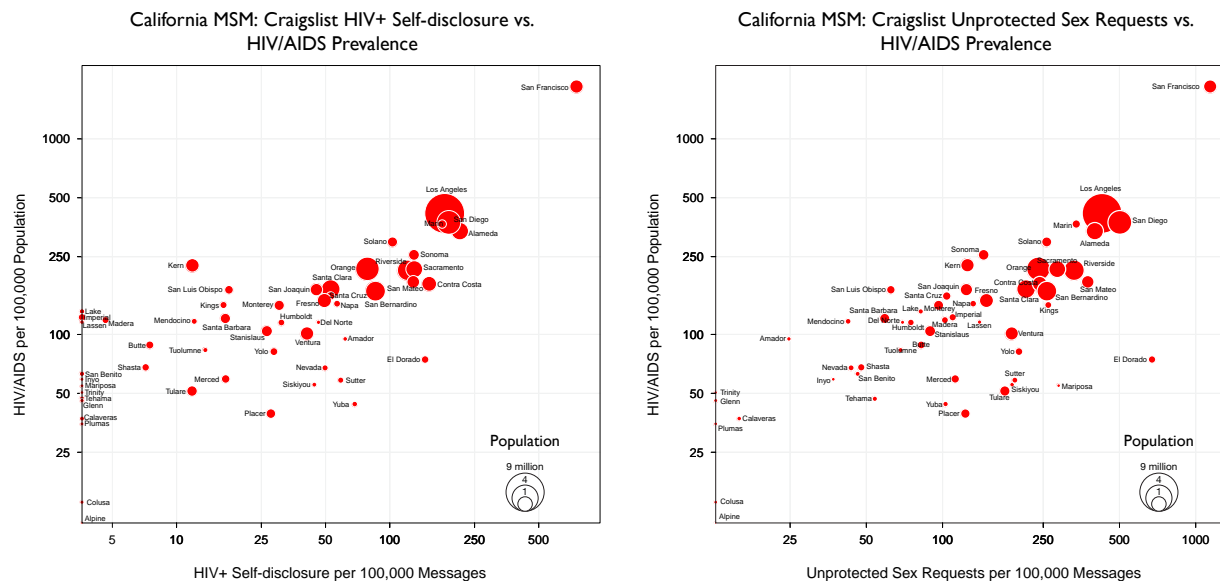


Figure 5.7: Log-log scale scatter plots of the OLS regression model of California MSM HIV+ (left) and BB (right) ontologies vs. combined HIV/AIDS prevalence rates. In these plots the size of the circle corresponds to the population of the county data point. Note the visible clusters between high population and low population counties.

tion relevant to both clinicians and the public health community. In some cases, it would be difficult, if not impossible and certainly expensive, to collect similar information using traditional survey-based approaches. Although these data are de-identified with respect to individuals, they are not de-identified with respect to geography, given the community-based nature of Craigslist. Because all ad content is easily and freely accessible, it represents a practical source of behavioral data for what we call *community risk mining*.

The use of the Internet to seek out sexual partners has had a profound impact on the epidemiology of sexually transmitted infections. Sexual encounters arranged online are likely to be geographically broader, often anonymous, and greater in number. Traditional approaches to control STIs involved contact tracing and partner notification [28,29,99]. However, Internet-mediated encounters undermine these critical prevention and control measures,

although some web sites collaborate with public health in providing risk reduction information [28,152]. Community risk mining promises a general and inexpensive method to collect timely, large-scale, location-specific behavioral surveillance data that will help inform clinicians to ask questions about new emerging risk factors for HIV and STIs.

Many interesting behavioral insights emerge from these data. For example, many ads openly discuss alcohol and illegal drug use during encounters, suggesting a potentially rich data source in the study of the epidemiology of drug abuse. Methamphetamine use has been identified as a strong risk factor for acquiring HIV among MSM [55,68,89,127,148]. We found frequent references to methamphetamine use in our data, often in the form of euphemisms or slang (e.g., “party with my friend tina”). Thus, our risk mining approach could be used for public health officials to better understand risk factors at a local level. Survey data indicates differences in how people in rural and urban areas use the Internet to find partners [17], so it follows that HIV/STI risk factors may vary by setting.

There are a number of limitations to this work. First, Craigslist ads reflect an intention for an encounter, not necessarily an encounter. However, ad content may still give insight into a community’s practices as well as providing some user demographic information. Second, as with surveys, we are unable to directly validate the information contained in these ads (i.e., HIV status). Third, ad posters may misrepresent information about themselves (e.g., their age). Fourth, as our RSS aggregator runs at regular intervals each day, we may fail to capture ads whose active posting lifespan is shorter than our aggregation time window. For this same reason, our corpus also contains some spam, as we capture some ads that will later be flagged by the community for removal.

We manually specified term clusters using as unambiguous term choices as possible; one of the limitations of this approach is that there are a number of terms connoting risk that have mixed word senses. For example, in many parts of California, ads asking about “snow” or “skiing” are in fact euphemisms for cocaine use, mixed-in with other ads containing real requests to go skiing. Terms such as “favors” can be used as slang for bringing drugs to encounters, but that usage must be distinguished from the more common usage (e.g., “sexual favors”). Such word sense disambiguation problems, commonly encountered in natural language processing research, underline the importance of using more sophisticated approaches for information extraction.

Despite these limitations, we demonstrate the ubiquity of Internet-mediated encounters across the US. Although Craigslist is only one site, the number of the ads collected demonstrates the importance of these sites. The ease of finding partners online, the anonymous nature of these encounters, and the associated risky behaviors all have broad implications for HIV and STI transmission. Risk mining will not eliminate the need for traditional surveillance, but it could provide a new supplementary approach.

## CHAPTER 6 SEMANTIC MODELING OF MEDICATIONS IN CLINICAL TEXT

### 6.1 Introduction

In recent years there has been considerable progress in developing multi-layer “deep” neural networks for tasks in speech recognition, computer vision, natural language processing, and other domains. Many challenging NLP tasks like sentiment analysis, question answering, etc. report state of the art performance using deep networks [109, 132]. One of the key advantages of using many hidden layers is the ability to model compact parameterizations of functions whereas a shallow architecture might require exponentially more representational elements [14]. This capacity for representation learning has been particularly successful in the case of *word embeddings*. Evidence suggests that incorporating these embeddings into existing linear or shallow classifiers improves classification performance in tasks like sequence labeling [37, 139]. Thus far, however, there has been little research exploring these techniques in clinical text (CT) settings. This is largely due to the volume of text data required to train models; gaining access to data sets of the required scale is problematic due to regulations governing patient privacy.

This work examines semi-supervised methods for extracting terminology and medical relations directly from a large collection of unlabeled clinical text documents spanning 7 years and 8 million patient visits, consisting of 880M sentences and 6.6B unigram tokens. Biomedical domain experts invest considerable effort into building curated lexicons and ontologies of medical concepts, which provide a common language for encoding and exchanging data for

billing and other purposes. These resources are frequently leveraged for clinical information extraction (IE) tasks like medication extraction.

The primary purpose of this work is to explore unsupervised feature modeling in clinical text and tentatively examine how much performance can be achieved in a sequence labeling task just using text data, in the absence of any of the hand engineered features like lexicons, context-free grammars, or regular expressions commonly employed in clinical IE tasks. We look at 2 motivating questions: 1) can we model meaningful representations of words (i.e., word embeddings) directly from a large collection of clinical text in an unsupervised fashion; and 2) how well do these learned features improve other machine learning and information extraction tasks in clinical text? For this purpose we explore the medication extraction task, which labels mentions of medications and various attributes in unstructured text.

Specifically, we apply recent advances in generating word embeddings from document collections to learn representations of medical terms and phrases. These continuous representations encode interesting semantic and syntactic properties of words [94]. For example, similar words tend to cluster in the embedding space: **knee** is near **elbow**, **hip**, and **shoulder**; **cephalexin** (an antibiotic) clusters with other antibiotics and so forth. More surprising, many syntactic and semantic relational properties can be directly computed using vector operations on embeddings.

Using medication mentions as an example, we show that modeling a low-dimensional representation of a medication name's word context, i.e., its neighboring words, encodes a large amount of non-obvious semantic information. Word embeddings, for example, im-



plicitly encode a large degree of the hierarchical structure of drug families and can model relational attributes such as the generic and brand names of medications. These word embeddings – learned in a completely unsupervised fashion – can then be used as features in other machine learning tasks. We show that incorporating clinical word embeddings in a benchmark classification task of medication labeling leads to a 5.4% increase in  $F_1$  measure over a baseline of random initialization and a 1.9% increase over just using training data. Moreover, we show that embeddings trained on one hospital’s data still lead to improved performance when labeling an unrelated hospital’s discharge summaries. This suggests that embeddings could be shared and distributed across institutions as an empirically-derived lexicon.

## 6.2 Methods & Materials

### 6.2.1 Corpora and Preprocessing

In order to create our word embeddings, we use a very large collection of clinical text documents. The *University of Iowa Hospitals and Clinics Corpus* (UIHC corpus) consists of clinical note data from the UIHC’s Epic-based electronic medical record system. These notes, linked by patient identifier, span 2007 - 2014 and include over 79 different clinical document types (e.g., discharge summaries, clinical notes, progress notes, etc.) Notes are not de-identified, meaning records are linked by patient identifier across time and include protected health information data points like attending healthcare worker, dates of admission, etc. The entire corpus consists of 880M sentences, 6.6B unigram tokens, 34M unique word types (900K when thresholded at a minimum occurrence of 5). For purposes of this work, we

examine 4 corpora in detail: *discharge summaries*, *progress notes*, *clinic notes*, and the set of *all notes*.

To evaluate our second motivating question, we would ideally use a clinical text benchmark data set similar to the Penn Treebank to evaluate information extraction performance in tasks like medication extraction. A benchmark corpus annotated by linguists for part-of-speech tags, shallow parsing (chunks), named entities, semantic role labeling, etc. would allow for better comparisons across proposed clinical text IE methods. However, while at least once such benchmark data set has been proposed [4], legal hurdles regarding patient privacy have thus far prevented the sharing of such corpora with other researchers. Instead, to evaluate the sequence labeling task we use the i2b2 2009 medication extraction challenge data set, described in detail in §3.2.1. Table 6.1 contains summary information on the document collections used in this work. All corpora except i2b2-training (which is considerably smaller in size) are filtered to remove words that occur fewer than 5 times.

As per our stated goals, normalization and preprocessing are deliberately kept to a minimum. We do not subdivide clinical documents into content sections (“Current Medications”, “Patient History”) that other research has shown improves some IE task performance [106]. All words are lowercased and stripped of trailing punctuation marks (e.g., “lasix,” becomes “lasix”) and all digits are replaced with an uppercase N (e.g., “500 mg” becomes “NNN mg”). Finally, since our classifiers operate on sentence input, we use the Python NLTK default Punkt English language module for sentence boundary detection [19].

Table 6.1: UIHC Clinical Document Collections

Corpus Name	Note Type	Words	Sentences	Vocab
i2b2 training	Discharge Summary	1.2M	136K	34K
UIHC-DS	Discharge Summary	258M	28M	133K
UIHC-CN	Clinic Note	2.3B	252M	378K
UIHC-PN	Progress Note	1.8B	293M	378K
UIHC-ALL	All Notes	6.6B	880M	899K

### 6.2.2 Medication Data Set Creation

We use the American Hospital Formulary Service (AHFS) Pharmacologic-Therapeutic Classification [104] system to define medication classes, which is the coding system used in the UIHC. The AFHS is a 4-tier hierarchical classification system; all medications belong to one of 31 top-level families and the majority have finer-grained classifications. We only consider the top 3 tiers due to the sparsity of training examples at the 4th tier. We generate a training set of drug entities by exploiting a syntactic pattern used by UIHC clinical text. Clinic notes frequently prefix drug mentions using a bullet character and a generic name/brand name tuple of the form: `GENERIC_NAME ( BRAND_NAME )`. Using a regular expression, we can create a candidate set of medication tuples. Since this pattern also encodes other medical data (e.g., procedures), the resulting candidate set is validated by matching drug names to a dictionary of AHFS codes. All multiword drug names were removed (e.g., “calcium citrate”) as this work only considers single word embeddings. This results in a final data set of 1076 medication entities, `MEDNAMES`, consisting of a mix of brand names and generic names.

### 6.2.3 Training Word Embeddings

All word embeddings in this document are trained with word2vec using the Skip-gram model [94]. We explore embedding dimensions of 50, 100, 300, 500, 768, and 1000. All models are trained using negative sampling with a value of 10. Since using a large context window  $c$  can lead to better performance at a small cost in training time, we tested window sizes of 2, 5, and 10 using 50-dimensional vectors and found that 10 consistently performed best. We generate separate embeddings for all corpora examined in this work which are trained for 1, 3, or 25 epochs depending on corpus size.

### 6.2.4 Evaluating the Quality of Embeddings

#### **Hierarchical Medication Classification**

Similar medications should occur in similar textual contexts, which capture a variety of information regarding dosage, mode of administration, reason of prescription, duration, etc. Our hypothesis is that by representing medications only using word context, we should gain enough discriminative power to assign medications to their appropriate AHFS class at various hierarchical resolutions. Embeddings that perform well in this classification task are, in some sense, “better” representations of the drugs themselves, providing a proxy quality measure of the learned representation. In order to evaluate the quality of the learned representations of medication words, we build a KNN classifier that assigns medication embeddings to their corresponding AHFS class. We use weighted L2 distance and  $k=3$  for our classifier hyperparameters where  $k$  was chosen through a grid search in the range [1..10]. Instance data consists of all embeddings from the MEDNAMES corpus, using stratified

5-fold cross validation.

## Generic/Brand Name Relations

One interesting property of embeddings is that the computed offsets between word vectors seem to encode a wide range of syntactic and semantic relationships, a property called *linguistic regularities* [95]. For example, using cosine similarity, the computed vector offset (**procedure – procedures**)  $\approx$  (**incision – incisions**), (**dose – doses**)  $\approx$  (**pill – pills**) and so on, suggesting that this offset captures some aspect of the singular/plural relation. More complicated semantic relations can also be computed such as medication brand names. Brand or trade names are usually encoded as a formal relation of the form A `has_tradename` B in medical nomenclatures like the UMLS and RxNorm [24, 80]. This relation can actually be computed directly using embeddings. Consider an analogy of the form **a:b::c:d** (*a* is to *b* as *c* is to *d*) where *d* is unknown. We can compute an approximation of *d* using normalized embeddings:  $y = x_b - x_a + x_c$ . For example, if we wish to determine a brand name for acetaminophen we can transform the analogy **ibuprofen:advil::acetaminophen:?** into the vector operation: (**advil – ibuprofen + acetaminophen**). Surprisingly, the correct answer (**tylenol** in this example) is frequently the nearest neighbor of our computed vector. Using embeddings trained on UIHC-ALL, this is in fact the case: **tylenol**  $\approx$  (**advil – ibuprofen + acetaminophen**).

Using the medication data set described above, we create a test set of 8,340 analogy questions using 248 unique generic/brand name medication pairs. Drug names are selected when both the brand name and generic variant occur with a raw frequency  $\geq 5K$  in the

Table 6.2: Sample Semantic Medication Analogy Questions

Type of Relationship	Generic/Brand Name Pair 1		Generic/Brand Name Pair 2	
Anti-infectives	amoxicillin	amoxil	oseltamivir	tamiflu
Central Nervous Sys. Agents	ibuprofen	advil	acetaminophen	tylenol
Hormones & Synthetics	hydrocortisone	cortef	prednisone	deltasone
Vitamins	ergocalciferol	calciferol	calcitriol	rocaltrol
Blood Formation	alteplase	activase	warfarin	coumadin

UIHC-ALL corpus. Table 6.2 shows some examples of these analogies, which are broken down by AHFS tier 1 classifications in order to evaluate performance across drug families.

### 6.2.5 Sequence Labeling

#### Recurrent Neural Networks

The primary goal of learning high quality word embeddings is to use them in service of other machine learning tasks. We consider the task of sequence labeling in clinical text, specifically identifying medication and their attributes in discharge summaries. Linear chain conditional random fields (CRF) are a popular choice for solving sequence labeling problems, but real-valued embedding features must first be discretized to use as input. This is usually done by clustering the embedding space into word classes or otherwise scaling the embeddings themselves in a preprocessing step [139]. Elman networks [43], a form of Recurrent Neural Network (RNN), can naturally incorporate dense, real valued feature vectors and have been used for other embedding-based sequence labeling tasks [90].

Each word in our target vocabulary is represented as a  $n$ -dimensional vector in a lookup table of  $|Vocab| \times n$  parameters (i.e., our learned embedding matrix). RNN input

features consist of a concatenation of these embeddings to represent a context window of words around our target word. This allows the RNN to model both local dependencies via the immediate context window and longer distance dependencies via the hidden recurrent context layer. For example, given the sentence “zocor ( simvastatin ) NN mg po bedtime” and a content window of 5, the input feature vector,  $ftrvec$ , for the word “simvastatin” is the concatenation (the  $\oplus$  operator) of embeddings:

$$ftrvec = zocor \oplus ( \oplus \mathbf{simvastatin} \oplus ) \oplus NN$$

This lookup table is shared across all input instances and updated during training. Eventually, similar words should have similar parameters, i.e., similar words should be near each other in the learned embedding space.

### Pre-training

By default, all embedding parameters are initialized randomly in the range  $[-1.0, 1.0]$ . However, a variety of research has shown the utility of seeding the initial state of word vectors with parameters trained on different document collections (*pre-training*). Pre-training generally improves classification performance over random initialization and provides a mechanism to leverage large collection of unlabeled data for use in the supervised learning task (*semi-supervised learning*) [45]. We compare random initialization with various pre-training strategies using word embeddings learned on different clinical text corpora.

### Medication Labeling

The original i2b2 medication labeling task consists of an output label set of 6 tags annotated in IOB format: medication/drug name, mode, dose, frequency, duration, and

reason. We evaluate models using all 6 tags as well as models removing duration and reason, which are comprised of longer phrases and tend to be more challenging. See Table 3.3 for more details. The annotated data set is split 60/10/30 into training, validation, and testing sets. We use a hidden layer of 200 units with a sigmoid activation function and softmax output layer, a context window of size 11, a learning rate  $\alpha = 0.1$  and an exponential decay rate  $\lambda = 0.5$ . Our models train for 15 epochs, updating the learning rate after any improvement in the validation set. Longer training periods did not lead to improved performance. Hyperparameters were chosen through a grid search using 50-dimensional embeddings (context window:[3, 5, 11]; hidden layers: [100, 200, 500];  $\alpha$ : [0.01, 0.1]; epochs:[15, 25]) and the corresponding performance on the validation set. We also tested embeddings for 100 and 300 dimensions and 500 hidden layer units, but found no statistically significant performance gains over using 50-dimensional inputs (though larger dimensions did have slightly larger mean scores).

All results are evaluated in terms of precision, recall, and  $F_1$ -score using the `conlle-val.pl` script [137]. Reported results are for testing set performance. For baseline measures, we use a randomly-initialized embedding model and a model pre-trained on unlabeled i2b2 training data. Reported scores are the mean value of 10 runs, with paired t-tests used to determine significant differences between models and the baselines.



Table 6.3: Clinical Text Medication Terms: 10-nearest Neighbors (UIHC-ALL)

<b>tylenol</b> <i>n</i> =1,277,994 Analgesics/Antipyretics AHFS 28:08	<b>amoxicillin</b> <i>n</i> =363,895 Antibacterials AHFS 8:12	<b>prozac</b> <i>n</i> =179,072 Psychotherapeutic AHFS 28:16	<b>lopressor</b> <i>n</i> =102,209 $\beta$ -Adrenergic Blocking AHFS 24:24
motrin	amoxil	fluoxetine	<i>prnivil</i>
acetaminophen	cephalexin	imipramine	<i>vt-cad</i>
ibuprofen	augmentin	paxil	<i>hydrodiuril</i>
<i>pain/fever</i>	ceftin	clomipramine	<i>apresoline</i>
<i>fever/pain</i>	omnicef	luvox	metoprolol
<i>phenergan</i>	cefzil	zoloft	<i>hypertension/cardioprotection</i>
loratab	sepra	buspar	<i>prophylaxis</i>
dilaudid	amoxacillin	desipramine	<i>cad/NNN.NN</i>
tramadol	keflex	bupropion	<i>cardioprotective/ischemia</i>
<i>prn</i>	clavulanic	<i>klonopin</i>	<i>hypertension</i>

## 6.3 Results

### 6.3.1 Medication AHFS-type Classifier

Table 6.3 contains some examples of 10-nearest-neighbors for medication mentions with their corresponding AHFS tier 2 class and raw frequency. Italicized words indicate incorrect members of the query term’s AHFS class. Incorrectly assigned neighbors generally fall into three categories: frequent co-occurring terms (“fever/pain”, “prophylaxis”); hard classification errors featuring drugs from different AHFS tiers (“phenergan” an antihistamine); and soft classification errors (“klonopin”, “apresoline”) which aren’t members of the target class but do share the same major class. Note that many of the errors of the co-occurring type are still related to the query term and often reflect the reason a medication is administered; Tylenol is given for pain/fever and lopressor is given as standard practice

of care for heart attack patients. Beta-blockers (class 24:24) show poor performance in all classification tasks.

Table 6.4: Medication Classifier ( $\geq 5$  Instances-per-class)

Note Type	AHFS Tier	Class N	Embedding Dimensionality / Mean $F_1$					
			50	100	300	500	768	1000
All Notes	Tier 1	30	0.78	0.81	<b>0.86</b>	0.86	0.86	0.86
Discharge Summaries			0.73	0.76	0.79	0.79	0.79	0.79
Clinic Notes			0.78	0.81	0.83	0.84	0.83	0.83
Progress Notes			0.76	0.80	0.84	0.84	0.84	0.84
All Notes	Tier 2	55	0.68	0.74	0.81	0.82	0.82	<b>0.83</b>
Discharge Summaries			0.63	0.69	0.73	0.76	0.75	0.76
Clinic Notes			0.70	0.76	0.79	0.81	0.82	0.80
Progress Notes			0.65	0.73	0.79	0.79	0.80	0.79
All Notes	Tier 3	71	0.67	0.73	0.77	<b>0.79</b>	0.79	0.79
Discharge Summaries			0.58	0.62	0.69	0.71	0.72	0.71
Clinic Notes			0.65	0.72	0.74	0.76	0.77	0.77
Progress Notes			0.60	0.67	0.73	0.74	0.76	0.77

Figures 6.1, 6.2, and 6.3 show 2D visualizations of the medication word embeddings (UIHC-ALL dim=1000) themselves at AHFS tiers 1,2, and 3 respectively. Here numbers correspond to the class id and color indicates one of the top 7 clusters ranked by set size. Note how medications clearly cluster by class and frequently hierarchically within class. Table 6.4 shows mean classification performance of embedding sizes across all classes in each of 3 AHFS tiers. In general, as representation size increases so does the  $F_1$ -score, but performance gains begin flattening out after embedding size of 300. The collection of all notes performed best on the medication classification task across all tiers. Performance

decreases as classification tiers become more specific.

Table 6.5: Medication Classifier: AHFS Tier 1 ( $\geq 10$  Instances-per-class)

AHFS Classification Name	Code	P	R	F <sub>1</sub>	N
Antibacterials	08:12	0.91	0.92	0.92	65
Antifungals	08:14	0.79	0.79	0.79	14
Antivirals	08:18	0.96	0.95	0.95	56
Antiprotozoals	08:30	0.85	0.79	0.81	14
-----					
ANTINEOPLASTIC AGENTS	10:00	0.96	0.97	0.96	88
-----					
Parasympathomimetic (Cholinergic) Agents	12:04	0.92	0.73	0.81	15
Sympathomimetic (Adrenergic) Agents	12:12	0.87	0.93	0.90	14
Skeletal Muscle Relaxants	12:20	0.83	0.83	0.83	12
-----					
Antithrombotic Agents	20:12	0.96	0.96	0.96	24
-----					
Cardiac Drugs	24:04	0.81	0.91	0.86	23
Antilipemic Agents	24:06	0.95	1.00	0.98	20
Hypotensive Agents	24:08	0.80	0.57	0.67	14
Vasodilating Agents	24:12	1.00	0.94	0.97	16
$\beta$ -Adrenergic Blocking Agents	24:24	0.69	0.50	0.58	22
Calcium-Channel Blocking Agents	24:28	1.00	1.00	1.00	12
Renin-Angiotensin-Aldosterone System Inhibitors	24:32	0.69	0.90	0.78	30
-----					
Analgesics and Antipyretics	28:08	0.92	0.88	0.90	40
Anticonvulsants	28:12	0.94	0.94	0.94	34
Psychotherapeutic Agents	28:16	0.87	0.99	0.93	75
Anxiolytics, Sedatives, and Hypnotics	28:24	0.87	0.84	0.85	31
Antiparkinsonian Agents	28:36	0.95	0.91	0.93	22
-----					
Replacement Preparations	40:28	0.87	0.72	0.79	18
-----					
Anti-inflammatory Agents	48:10	0.93	0.93	0.93	15
-----					
Antiemetics	56:22	0.78	0.82	0.80	17
-----					
Antiulcer Agents and Acid Suppressants	56:28	1.00	0.94	0.97	18
-----					
Adrenals	68:04	1.00	0.74	0.85	19
Antidiabetic Agents	68:20	1.00	1.00	1.00	20
-----					
Anti-infectives	84:04	0.89	0.80	0.84	10
-----					
Bone Resorption Inhibitors	92:24	1.00	0.90	0.95	10
Immunosuppressive Agents	92:44	0.75	1.00	0.86	12
-----					
Weighted Average		0.90	0.90	0.89	780

Table 6.6: Antibacterial Classifier ( $\geq 5$  Instances-per-class)

AHFS Classification Name	Code	P	R	F <sub>1</sub>	N
Aminoglycosides	08:12.02	0.50	0.40	0.44	5
Fourth Generation Cephalosporins	08:12.06	0.45	0.71	0.56	14
Monobactams	08:12.07	0.50	0.17	0.25	6
Other Macrolides	08:12.12	0.57	0.57	0.57	7
Extended-spectrum Penicillins	08:12.16	0.25	0.14	0.18	7
Quinolones	08:12.18	0.88	0.88	0.88	8
Glycylcyclines	08:12.24	1.00	1.00	1.00	6
Streptogramins	08:12.28	0.78	0.70	0.74	10
Weighted Average		0.61	0.60	0.59	63

The previous table considers many classes containing few instances ( $\geq 5$  instances). If we restrict our analysis to larger drug families performance improves. Table 6.5 shows Tier 2 classification performance across all classes with  $\geq 10$  instances. The vast majority of classes perform extremely well, with an overall weighted average F<sub>1</sub>-score of 0.89. Larger classes (N  $\geq 40$ ) perform well (F<sub>1</sub>  $\geq 0.90$ ) with many small-sized classes reporting similar performance. Poor performers include medications like  $\beta$ -Adrenergic Blocking Agents, Hypotensive Agents, or Antiemetics. Finally, Table 6.6 shows classification results for the Antibacterials AHFS Tier 3 class, the highest level of resolution considered in our medication classification task. Performance is considerably less than at Tiers 1 and 2, though some classes still report reasonable performance. Quinolones, Glycylcyclines, and Streptogramins all perform better than the remaining antibacterials, with F<sub>1</sub> scores ranging from 0.74 - 1.0 compared to 0.18 - 0.57. This shows that some antibiotic subclasses are difficult to differentiate from each other.

## 6.3.2 Relation Encoding

Table 6.7 shows the accuracy of extracting the `has_tradename` relation from medication embeddings from progress notes, the best performing note type overall. In general, the larger the representation space, the more accurate our performance with diminishing returns seen after size 300 as was the case with the medication classification task. Figure 6.4 shows a line plot of the accuracy of all note types across multiple embedding dimensions.

Table 6.7: Semantic Evaluation: Medication Brand Name Prediction (Progress Notes)

AHFS Tier 1 Class Name	Analogies	Drug $n$	Embedding Dimensionality / Accuracy [%]					
			50	100	300	500	768	1000
Antihistamines	12	4	33.3	58.3	100.0	100.0	100.0	100.0
Anti-infectives	1190	35	17.6	30.3	59.9	65.1	72.4	75.1
Antineoplastics	132	12	37.9	65.2	75.8	76.5	79.5	81.8
Autonomics	156	13	17.3	40.4	69.2	74.4	75.0	76.9
Blood-formation	30	6	40.0	76.7	100.0	93.3	93.3	93.3
Cardiovascular	1190	35	14.5	25.6	55.0	68.4	77.9	81.2
Cntrl Nervous System	4692	69	33.8	54.2	78.4	83.7	86.7	87.6
Repl. Preparations	30	6	10.0	36.7	46.7	56.7	73.3	83.3
Respiratory Tract	6	3	0.0	0.0	66.7	50.0	66.7	50.0
Gastrointestinal	342	19	40.1	57.6	80.1	82.7	83.6	85.1
Hormones & Synthetics	342	19	25.4	44.4	62.3	70.2	74.6	75.7
Oxytocics	2	2	0.0	50.0	50.0	50.0	50.0	50.0
Skin/Mucous Membrane	20	5	30.0	50.0	80.0	80.0	80.0	80.0
Smooth Muscle Relaxants	2	2	100.0	100.0	100.0	100.0	100.0	100.0
Vitamins	12	4	8.3	8.3	83.3	100.0	100.0	100.0
Misc. Therapeutic	182	14	35.2	55.5	78.6	79.1	84.1	85.2
TOTAL	8340	248	28.3	46.3	71.6	77.8	82.4	84.0

## 6.3.3 Sequence Labeling Performance

Table 6.8: i2b2 Medication Labeling: All Tags

Corpus	Testing Set Mean $F_1$ ( $n=10$ )						Overall
	Drug	Dose	Mode	Freq	Duration	Reason	
random-init-50	83.98	91.06	94.06	90.35	<b>39.94</b>	34.74	84.41
i2b2-training-50	86.76	91.39	93.66	90.49	38.94	37.58	85.67
UIHC-DS-50	87.75*	91.51	<b>94.21</b>	<b>90.76</b>	38.47	<b>38.03*</b>	86.15*
UIHC-ALL-50	<b>88.40**</b>	<b>91.74</b>	93.86	90.52	38.02	37.78*	<b>86.37*</b>
Patrick-Li-2009 [106]	88.35	89.26	89.94	89.65	44.64	44.36	85.65

Pre-training with UIHC embeddings improves performance on drug/medication and reason tagging. Table 6.8 shows our system’s performance labeling all tags. As a point of comparison, we also provide the scores of the original i2b2 2009 challenge’s top performing system by Patrick and Li [106].<sup>1</sup> Using UIHC data results in a statistically significant increase in  $F_1$  (indicated by an asterisk in all tables) across both baselines (random-initialization and i2b2-training), with the set of all data performing best overall. Pre-training improves medication as well as reason classification but has no impact on duration classification. Since reason and duration are more challenging labeling tasks usually involving much longer phrases, we also show classification results when ignoring reason and duration tags in Table 6.9. Table 6.10 shows drug labeling in more detail. Note that incorporating 50-dimensional

---

<sup>1</sup>Reported scores are computed as a micro-average across all entities (“system-level”) for all exact matches. Note that the original challenge participants did not have access to the entire collection of 251 annotated discharge summaries, so these measures are provided only as an approximate comparison to a top-performing, manually-engineered system.

embeddings substantially improves recall over a random initialization baseline: 5.3% using i2b2 (non-UIHC) training data and 8.2% using all UIHC note data. Overall, pre-training using UIHC embeddings improves  $F_1$  by 5.4% over random initialization and 1.9% over using non-UIHC training data.

Table 6.9: i2b2 Labeling: Drug, Dose, Mode, Frequency

Embeddings	Mean $F_1$ ( $n=10$ )				<b>Overall</b>
	Drug	Dose	Mode	Freq	
random-init-50	83.95	91.40	94.07	90.32	88.51
i2b2-training-50	86.85	91.89	94.40	90.15	89.83
UIHC-DS-50	88.12*	91.98	94.48	90.52	90.45*
UIHC-ALL-50	88.51*	92.06	94.51	91.26	90.80*
UIHC-ALL-100	88.76*	91.53	94.37	91.61	90.83*

Table 6.10: i2b2 Labeling: Drug

Embeddings	Drug			$\pm\Delta F_1$
	Precision	Recall	$F_1$	
random-init-50	87.22	80.93	83.95	-
i2b2-training-50	88.62	85.19	86.85	+3.5%*
UIHC-DS-50	89.08	87.20	88.12	+5.0%*
UIHC-ALL-50	89.51	87.56	88.51	+5.4%*
UIHC-ALL-100	<b>89.91</b>	<b>87.65</b>	<b>88.76</b>	+5.7%*

## 6.4 Discussion

This work shows that word embeddings capture a large amount of the semantic organization of medications in clinical text. In an unsupervised fashion, using word context alone, embeddings are capable of learning a geometric structure highly predictive of a medication's AHFS drug classification, classifying drugs at three hierarchical tiers with  $F_1$  scores between 0.79 - 0.86, quite reasonable performance for an essentially unsupervised method and 30 - 70 class labels depending on AHFS tier. Discriminating between Tier 1 and 2 AHFS classifications appears to be an easier task than distinguishing between tier 3. While antibacterials can easily be distinguished from antivirals, for example, correctly classifying specific antibacterials is more challenging. We also show that embeddings capture other useful relational properties, demonstrating that the formal relation `has_tradename` can be computed from embeddings directly with 84% accuracy.

Our RNN sequence labeling performance is competitive to the original best performing participant system in the 2009 medication challenge, except that our system does not manually engineer any features. Moreover, incorporating representations learned using UIHC clinical text improves performance in labeling medications in discharge summaries, even when those documents originated from another healthcare institution. Medication labeling results in a 5.4% improvement over random initialization and 1.9% over just using non-UIHC training data. Word embeddings function as empirical lexicons, encoding meaningful semantics and contextual information in an automated fashion. These learned representations could be shared and used for other NLP tasks.

Larger dimension word embeddings after a point do not seem to improve sequence



labeling performance, despite there being clear differences in representational power based on different configurations of dimensionality, content window, and training epochs. This is partly due to the nature of our i2b2 evaluation data which is not annotated with richer label sets (e.g., tagged with medication families) that would benefit from richer embedding representations. Another potential reason for the lack of improvement is that our RNN is relatively shallow (in terms of hidden layers) compared to other existing deep architectures. Multi-layered neural networks common in image analysis domains usually contain many more layers and more sophisticated internal structures like convolutional or max pooling layers. This architecture benefits from richer input features, which allow networks to better learn internal representations at deeper layers.

The i2b2 medication data set itself also has several limitations that potentially impact sequence labeling performance. The annotation corpus is split by line rather than sentence, which split some IOB chunks (incorrectly) across lines and break the input semantics of our RNN. One possible solution to this is to add a sentence boundary detection preprocessing step by creating a multi-layer classifier where the first RNN layer predicts sentence boundary tokens, document structure, etc. which is then used by subsequent layers. Another source of error stems from the i2b2 annotation guidelines, which define drug/medication mentions as any text indicating that the patient took the indicated medication. A small percentage of annotations contain ambiguous or coreferent mentions: “medications”, “home meds”, “that medication.” These expressions are considerably more difficult to correctly model and this performance may reflect more on our choice of sequence labeler than choice of embedding representation.

As is the case with the CRF-based sequence labeling approach used by Patrick and Li’s system, our RNN also performs more poorly at labeling reasons and durations than other tags in clinical text. Durations and reason labels are usually phrases: “upper respiratory and pharyngeal symptoms”, “normalized and potassium became stable at N”; “as long as you have a drain in place”; and so forth, all of which represent more complicated ideas than the semantic unit of a single word. Neural networks, which are capable of modeling phrase level embeddings using more sophisticated internal networks structures (e.g., convolutional layers), have shown improved performance in more complicated tagging tasks such as semantic role labeling. Incorporating these architectural components into our classifier could lead to better performance in these tags.

The ultimate goal of this work is to argue for more data-driven feature generation and more interest in NLP systems rather than manual feature engineering in clinical text. Rather than just building regular expressions, context-free grammars or other (potentially) hospital-specific design patterns, we argue that there is utility in leveraging the scale of clinical text to learn compact, distributable representations of clinical concepts like medications directly from data.

#### 6.4.1 Limitations

There are several limitations to this work. First, the MEDNAMES corpus of drug names only reflect a sample of possible generic and brand names and does not include names with 2 or more tokens. However our data set is not intended to be exhaustive and captures a wide variety of drug family classes and generic / brand name variation. Second, the goal

of this work was to perform as little pre-processing as possible on input documents. As such, we do not utilize several simple and useful features in our RNN, such as letter case or alphanumeric attributes. Letter case is an informative feature for named entity detection and in medications is frequently used to disambiguate orthographically similar medication names (Tall Man lettering [44]). Letter case could easily be incorporated into our classifier by manually specifying features as a preprocessing step.

Third, clinical text records are highly structured documents. However, when exporting clinical notes from the UIHC’s medical record system, notes are stripped of formatting such as tables and section dividers. Partitioning documents into sections as a preprocessing step has been shown to improve classification and labeling tasks and could lead to more targeted training of embeddings, for example, by only training on text found in “Current Medications.”

Finally, embeddings are trained using single word inputs. However, many concepts are better modeled at the phrase-level, with embeddings trained for multi-token words. Naively, this can be accomplished by using an association measure such as pointwise mutual information to merge n-grams and create phrase embeddings. More sophisticated approaches use neural networks structures to model sentence and phrase structure using single words as input. We leave these ideas for future work.

## AHFS Classification Tier 1

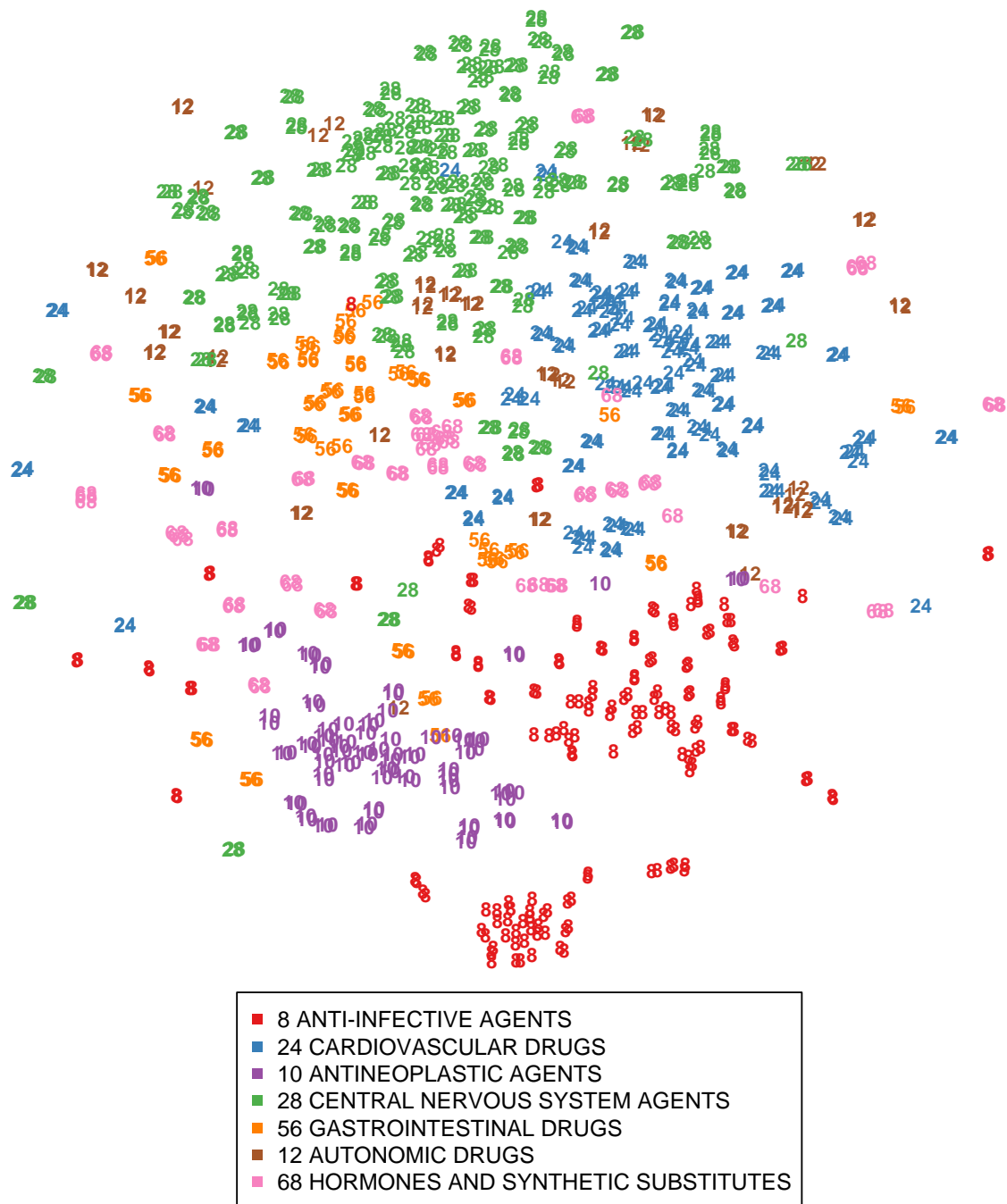


Figure 6.1: T-SNE [147] visualization of medication embeddings, clustered using the AHFS Tier 1 classification system ( $\geq 10$  instances). The number indicates cluster id and color indicates one of the top 7 clusters in terms of set size. Note how *Anti-infectives* (red) and *Central Nervous System Agents* (green) are comprised of multiple sub-clusters. These in fact frequently correspond to the next hierarchical tier of medications (see Figures 6.2 and 6.3).

## AHFS Classification Tier 2

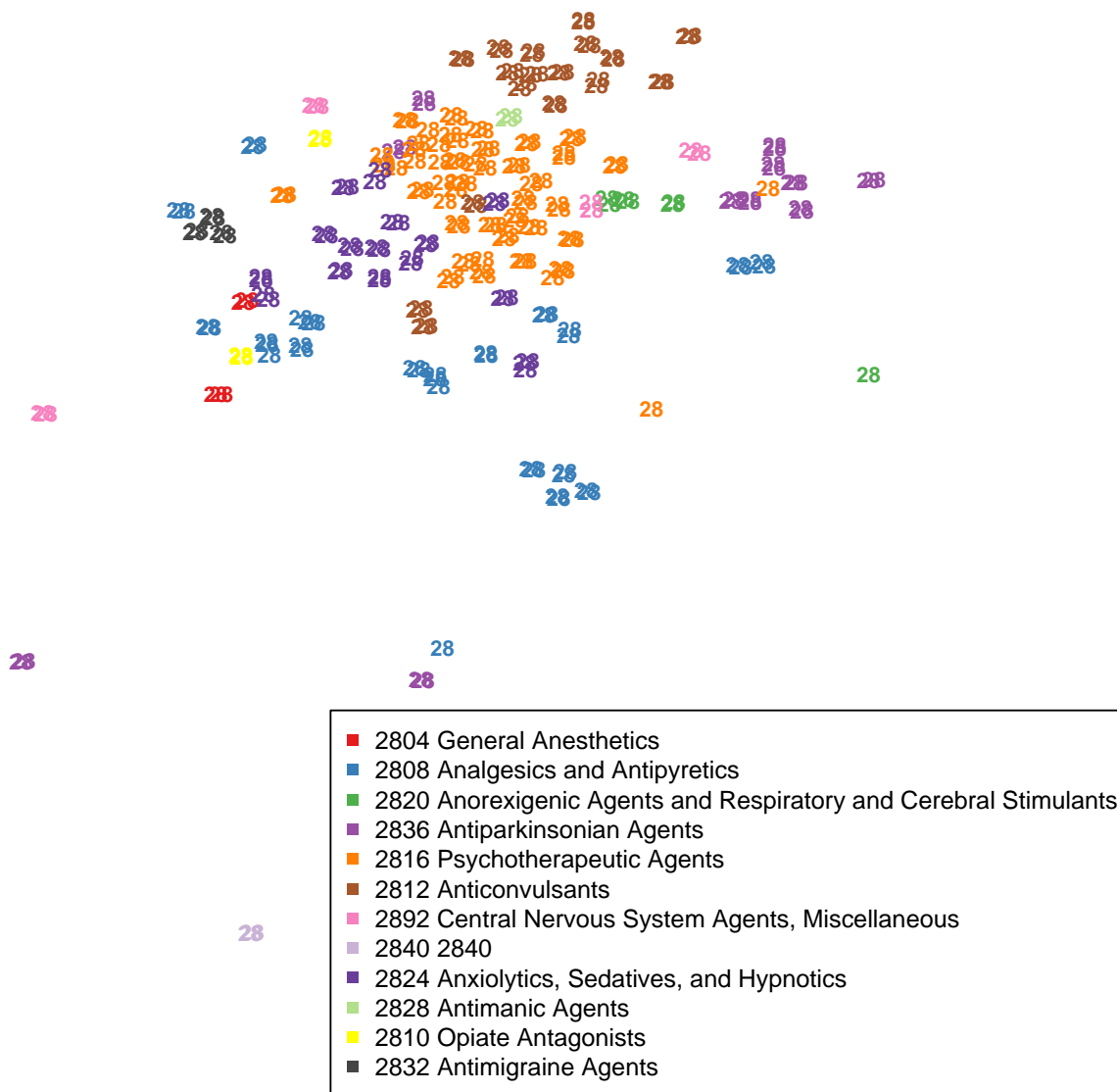


Figure 6.2: Medication clustering for AHFS Tier 2 *Central Nervous System Agents* (green colored points in Figure 6.1). Note the same sub-clustering observed at Tier 1. Numbered points with very close overlap typically correspond to brand name/generic name versions of medication mentions.

### AHFS Classification Tier 3

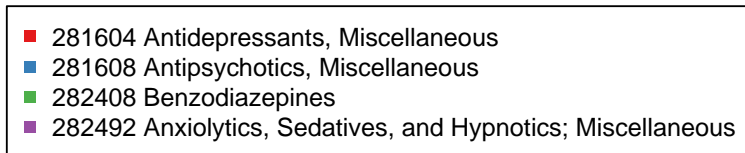
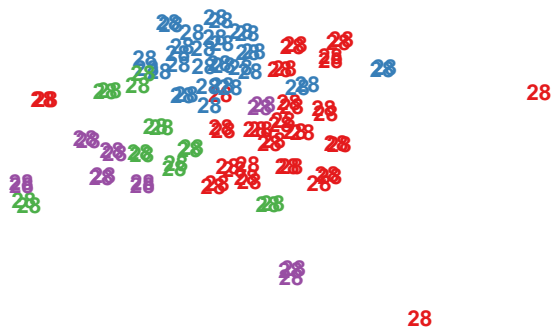


Figure 6.3: AHFS Tier 3 clustering for *Psychotherapeutic Agents* and *Anxiolytics, Sedatives, and Hypnotics* (orange and dark purple respectively in Figure 6.2). Note how much of hierarchical nature of the AHFS medication classification system is implicitly captured by the word embedding training process.

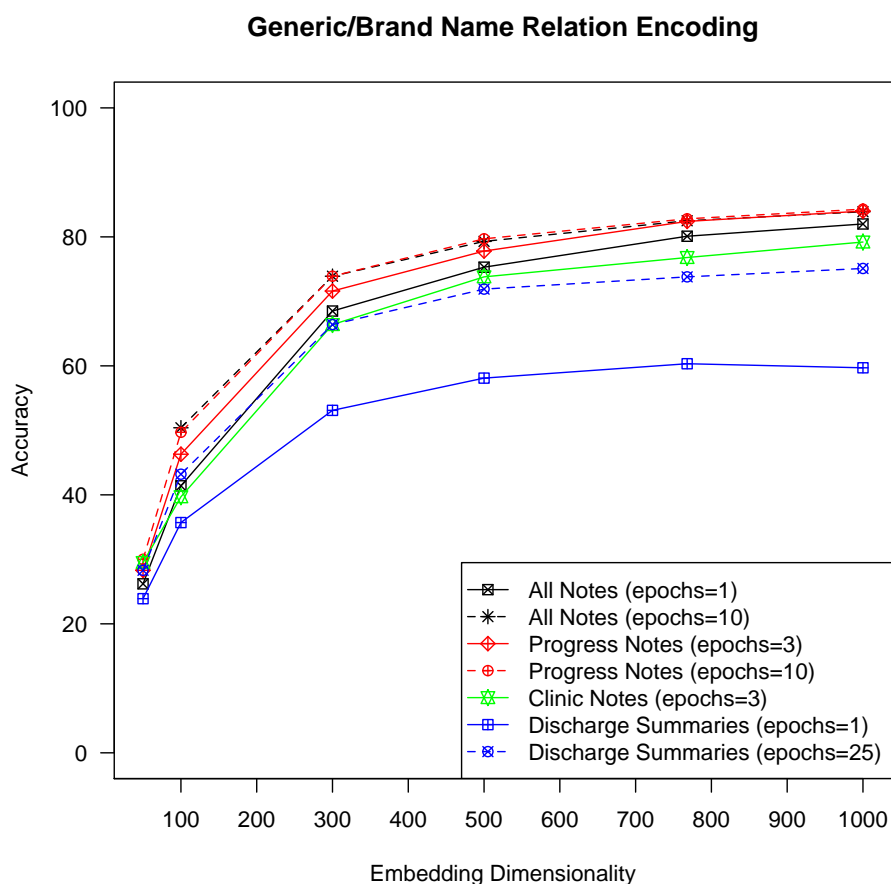


Figure 6.4: Generic/Brand name relation encoding accuracy. Performance gains using larger embedding spaces begin diminishing after 300. Progress notes performed best in capturing generic/brand name relationships, even though that corpus is roughly 27% of the size of the set of all notes. Note that multiple training epochs greatly improved the performance of the discharge summaries corpus, making it competitive to the larger corpora, even though it is substantially smaller (4% the size of all notes).

## CHAPTER 7 CONCLUSION

We have show that building a realistic sexual health surveillance tool involves achieving reasonable performance in four non-trivial NLP applications: 1) geocoding ads; 2) linking anonymous ads by author; 3) extracting detailed author demographic information; and 4) learning the terminology characterizing sexual behaviors. We provide contributions in all of these areas. First, our surveillance pipeline is predicated on accurately geocoding ad text. This is challenging in the context of Craigslist due to the highly informal, ambiguous, and otherwise noisy surface forms of geographic entity mentions in ad text. In our hand-annotated validation set, our geographic entity normalization tool TopoLinker correctly linked 85% of all Craigslist location tag mentions to their canonical knowledge base representations, achieving a mean geocoding error of 5.7 miles. This allows us to link ad text with at least U.S. county-level (and frequently better) spatial resolution.

Second, since Craigslist is an anonymous social network, oversampling is another important challenge that must be addressed. We presented an approximate nearest neighbors system for efficiently identifying sets of similar ads likely written by the same author. Using a validation set of ads disclosing phone numbers (which provide a way of linking authorship), we show that our approach does identify ads written by the same author. In general however, our approach only detects a small subset of an author’s ads; recall suffers as the number of ads written by a single author increases. Despite this, we identify and collapse 15% of all ads into single author clusters.

Third, structural factors like race/ethnicity and age are important variables in the



design and implementation of public health interventions; strategies that work in one population may not work in another. As such, it is important to link surveillance data to demographic variables. We presented work on a sequence labeling-based approach for automatically predicting the race/ethnicity of an ad’s author and their requested partner. Across 6 standardized U.S. Census definitions of race/ethnicity, we achieved a weighted average  $F_1$  of 0.87 in correctly categorizing an author. Using this classifier, we then showed that the distribution of author self-disclosed race/ethnicity fits the underlying census population distribution, validating that we are measuring real signal.

Fourth, one of the central challenges in sexual behavior surveillance is constructing lexical resources that reflect how people actually discuss and negotiate sex online. We discussed computational methods for automatically learning terminology characterizing risk behaviors in the MSM community, e.g., illegal drug use, unprotected sex. Motivated by recent work in deep neural networks and representation learning, we explore an unsupervised method for generating word embeddings – continuous space representations of words - and show that word context alone captures a large amount of semantic information and can be used to model many risk behaviors. We discussed the limitations imposed by this model compared to existing approaches that use keyword-based dictionaries and topic modeling. These methods allow us to gather information similar (in part) to the types of questions asked in risk assessment surveys, but automatically aggregated directly from communities of interest, in near real-time, and at geographically high-resolution.

This dissertation has presented work showing that it is possible to extract useful public health information from the unstructured text of online “hookup” ads. The ability to conduct

public health surveillance in an automated fashion, efficiently collecting large-scale, location-specific demographic data about the anonymous populations using websites like Craigslist is a useful supplemental tool for the public health community. Behavioral surveillance data is especially valuable when the subpopulations in question are difficult to reach through conventional survey approaches. While many public health agencies currently cooperate with hookup sites such as Craigslist, Grindr, etc., the idea of an automated surveillance tool for supplementing conventional survey data is a novel approach. We feel algorithmic surveillance of sexual behaviors can enrich interventions and other tools used to address outbreaks of sexually transmitted infections.

This work is not intended as a solution to the above tasks or even the stated goal of sexual behavior surveillance in general; each of the above components can be improved upon in significant ways. Instead, our hope is that this work serves as a proof of concept for a different approach to gathering data for use by public health departments. Despite limitations imposed by methods and the quality of source data, public health interventions will increasingly need to occur both on and offline, suggesting that automated tools for monitoring sexual behaviors in social media should be incorporated into public health intervention design.

The NLP methods used to model sexual behaviors in Craigslist text have much broader applications than sexual health surveillance. To illustrate the representational power of word embeddings, the last part of this dissertation presents preliminary work modeling medication terminology in clinical text from the University of Iowa Hospitals and Clinics (UIHC). Testing a variety of clinical document types and 1000 different medication generic

and brand names, we show that word embeddings are capable of automatically modeling some of the hierarchical structure of drug pharmacological families. In the best performing word embeddings, a KNN classifier achieves between an  $F_1$ -score between 0.79 - 0.86 in assigning a medication to its correct pharmacological family. Finally, we consider the task of medication extraction in unstructured text. We show that using word embeddings trained on UIHC data in a recurrent neural network leads to increased sequence labeling performance, even when the labeling task involves non-UIHC note data. This shows that embeddings can function as an empirically generated lexicon and could be potentially useful in other classification task across other healthcare institutions.

## 7.1 Future Work

One key question in my future work is this: what does an effective online public health intervention look like? There has been considerable research in the economics, social and political science domains in exploring ways to actually change people's behavior. In a sexual health intervention context, that has always been challenging, since changing behavior is dependent on the the ability to effectively measure individuals who are frequently members of hidden and stigmatized populations. Public health departments conduct surveys of risk behaviors, but are those data actually informative in terms of intervention design? Ecological behavioral data provides one view of high-risk sexual communities, but to what extent do we need to track individuals operating in those communities? Sexual health surveillance of anonymous individuals, rather than just geographic locations, could provide new insight into intervention design, but the ability to conduct such surveillance brings considerable

computational challenges in terms of accurate, large-scale authorship attribution.

Surveillance using Craigslist could be greatly improved upon by focusing on more robust authorship attribution methods. The methods described in this dissertation only capture a small fraction of authors' posting behavior across time. While the general case of authorship attribution at web scale is an very difficult problem, when constrained to the special case of online classified ads, the problem becomes more tractable. Ads display a high degree of similarity across time, meaning that reasonable accuracy might be achievable even using large-scale approximate nearest neighbors coupled with temporal and geographic constraints. We leave this for future exploration.

## REFERENCES

- [1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.
- [2] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.
- [3] Eneko Agirre, Aitor Soroa, and Mark Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, 2010.
- [4] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, pages amiajnl–2012, 2013.
- [5] Alexa: The Web Information Company. Craigslist.org site info, September 2013.
- [6] A.M. Ali, H.M.D. Abdulla, and V. Snasel. Survey of plagiarism detection methods. In *Modelling Symposium (AMS), 2011 Fifth Asia*, pages 39–42. IEEE, 2011.
- [7] Abdulrahman Almuhareb. *Attributes in lexical acquisition*. PhD thesis, University of Essex, 2006.
- [8] Alan R Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
- [9] Alan R Aronson and François-Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
- [10] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.
- [11] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.

- [12] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.
- [13] Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics, 2011.
- [14] Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.
- [15] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Frédéric Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.
- [16] E.G. Benotsch, S. Kalichman, and M. Cage. Men who have met sex partners via the Internet: Prevalence, predictors, and implications for HIV prevention. *Archives of Sexual Behavior*, 31(2):177–183, 2002.
- [17] E.G. Benotsch, A.M. Martin, F.M. Espil, C.D. Nettles, D.W. Seal, and S.D. Pinkerton. Internet use, recreational travel, and HIV risk behaviors in men who have sex with men. *Journal of Community Health*, 36(3):398–405, 2011.
- [18] K.T. Bernstein, F.C. Curriero, J.M. Jennings, G. Olthoff, E.J. Erbelding, and J. Zenilman. Defining core gonorrhea transmission utilizing spatial data. *American journal of epidemiology*, 160(1):51–58, 2004.
- [19] Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.
- [20] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O’Reilly Media, 2009.
- [21] Kim M Blankenship, Sarah J Bray, and Michael H Merson. Structural interventions in public health. *AIDS*, 14:S11–S21, 2000.
- [22] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
- [23] Phil Blunsom, Edward Grefenstette, Nal Kalchbrenner, et al. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

- [24] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
- [25] Sergey Brin. Extracting patterns and relations from the World Wide Web. In *The World Wide Web and Databases*, pages 172–183. Springer, 1999.
- [26] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
- [27] California Department of Public Health. California syphilis elimination surveillance data, 2011. [Online; accessed Dec-2011].
- [28] Alex Carballo-Diéguez, Michael Miner, Curtis Dolezal, BR Simon Rosser, and Scott Jacoby. Sexual negotiation, HIV-status disclosure, and sexual risk behavior among latino men who use the internet to seek sex with other men. *Archives of Sexual Behavior*, 35(4):473–481, 2006.
- [29] W. Cates Jr, R.B. Rothenberg, and J.H. Blount. Syphilis control: The historic context and epidemiologic basis for interrupting sexual transmission of *treponema pallidum*. *Sexually Transmitted Diseases*, 23(1):68, 1996.
- [30] Centers for Disease Control and Prevention (CDC). HIV among African Americans, 2014. [Online; accessed Feb-2014].
- [31] J. Chan and A. Ghose. Internet’s dirty secret: Assessing the impact of technology shocks on the outbreaks of sexually transmitted diseases. 2012.
- [32] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.
- [33] M.S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM symposium on Theory of Computing*, pages 380–388. ACM, 2002.
- [34] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.
- [35] Lee M Christensen, Henk Harkema, Peter J Haug, Jeannie Y Irwin, and Wendy W Chapman. ONYX: A system for the semantic analysis of clinical text. In *Proceedings*

- of the *Workshop on Current Trends in Biomedical Natural Language Processing*, pages 19–27. Association for Computational Linguistics, 2009.
- [36] Lee M Christensen, Peter J Haug, and Marcelo Fiszman. MPLUS: A probabilistic medical language understanding system. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical domain-Volume 3*, pages 29–36. Association for Computational Linguistics, 2002.
- [37] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.
- [38] Jonathan D Cryer and Kung-Sik Chan. *Time series analysis: with applications in R*. Springer, 2008.
- [39] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.
- [40] Elizabeth Cuthill and James McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th National Conference*, pages 157–172. ACM, 1969.
- [41] Kristina Doing-Harris, Olga Patterson, Sean Igo, and John Hurdle. Document sub-language clustering to detect medical specialty in cross-institutional clinical texts. In *Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics*, pages 9–12. ACM, 2013.
- [42] Doug Downey, Oren Etzioni, and Stephen Soderland. A probabilistic model of redundancy in information extraction. Technical report, DTIC Document, 2006.
- [43] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [44] Lynne M Emmerton and Mariam FS Rizk. Look-alike and sound-alike medicines: Risks and ‘solutions’. *International Journal of Clinical Pharmacy*, 34(1):4–8, 2012.
- [45] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660, 2010.
- [46] Jonathan Feinberg. Wordle, 2015. [Online; accessed Feb-2015].



- [47] C.M. Fichtenberg and J.M. Ellen. Moving from core groups to risk spaces. *Sexually Transmitted Diseases*, 30(11):825, 2003.
- [48] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM, 2001.
- [49] J.R. Firth. *A Synopsis of Linguistic Theory, 1930-1955*. 1957.
- [50] Centers for Disease Control. CDC’s national center for HIV/AIDS, viral hepatitis, STD, and TB prevention (NCHHSTP), 2015. [Online; accessed 1-April-2015].
- [51] W Nelson Francis and Henry Kucera. Brown corpus manual. *Brown University Department of Linguistics*, 1979.
- [52] Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Annual Symposium on Computer Applications in Medical Care*, volume 19, pages 347–351. IEEE COMPUTER SOCIETY PRESS, 1995.
- [53] Jason Fries. TAMER: A TIGER/Line shapefile downloader & PostGIS import tool, November 2013. [Online; accessed Nov-2013].
- [54] Vijay N Garla and Cynthia Brandt. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association*, pages amiajnl–2012, 2012.
- [55] R. Garofalo, B.S. Mustanski, D.J. McKirnan, A. Herrick, and G.R. Donenberg. Methamphetamine and young men who have sex with men: understanding patterns and correlates of use and the association with hiv-related sexual risk. *Archives of Pediatrics and Adolescent Medicine*, 161(6):591, 2007.
- [56] D.C. Gesink, A.B. Sullivan, W.C. Miller, and K.T. Bernstein. Sexually transmitted disease core theory: Roles of person, place, and time. *American Journal of Epidemiology*, 174(1):81–89, 2011.
- [57] R.F. Grais, J. Hugh Ellis, and G.E. Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European Journal of Epidemiology*, 18(11):1065–1072, 2003.
- [58] C. Grov. Risky sex-and drug-seeking in a probability sample of men-for-men online bulletin board postings. *AIDS and Behavior*, 14(6):1387–1392, 2010.

- [59] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [60] R.W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.
- [61] Zellig S Harris. Distributional structure. *Word*, 1954.
- [62] Stephan Holl and Hans Plum. PostGIS. *GeoInformatics*, 03/2009:34–36, April 2009.
- [63] H.W. Jaffe, R.O. Valdiserri, and K.M. De Cock. The reemerging HIV/AIDS epidemic in men who have sex with men. *JAMA: The Journal of the American Medical Association*, 298(20):2412–2414, 2007.
- [64] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
- [65] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.
- [66] K. Khan, J. Arino, W. Hu, P. Raposo, J. Sears, F. Calderon, C. Heidebrecht, M. Macdonald, J. Liauw, A. Chan, et al. Spread of a novel influenza a (h1n1) virus via global airline transportation. *New England Journal of Medicine*, 361(2):212–214, 2009.
- [67] Paul Kingsbury and Martha Palmer. Propbank: the next level of treebank. In *Proceedings of Treebanks and Lexical Theories*, volume 3. Citeseer, 2003.
- [68] Beryl A Koblin, Marla J Husnik, Grant Colfax, Yijian Huang, Maria Madison, Kenneth Mayer, Patrick J Barresi, Thomas J Coates, Margaret A Chesney, and Susan Buchbinder. Risk factors for HIV infection among men who have sex with men. *Aids*, 20(5):731–739, 2006.
- [69] Julie Kraut-Becher, Marlene Eisenberg, Chelsea Voytek, Tiffany Brown, David S Metzger, and Sevgi Aral. Examining racial disparities in HIV: Lessons from sexually transmitted infections research. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 47:S20–S27, 2008.
- [70] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [71] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

- [72] Dr. Leo. Pyhyphen: Hyphenation library for python, 2014. [Online; accessed Apr-2014].
- [73] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.
- [74] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Baltimore, Maryland, USA, June. Association for Computational Linguistics*, 2014.
- [75] Qi Li, Haijun Zhai, Louise Deleger, Todd Lingren, Megan Kaiser, Laura Stoutenborough, and Imre Solti. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *Journal of the American Medical Informatics Association*, 20(5):915–921, 2013.
- [76] Percy Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.
- [77] A. Liau, G. Millett, and G. Marks. Meta-analytic examination of online sex-seeking and sexual risk behavior among men who have sex with men. *Sexually transmitted diseases*, 33(9):576, 2006.
- [78] Hongfang Liu, Yves A Lussier, and Carol Friedman. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *Journal of Biomedical Informatics*, 34(4):249–261, 2001.
- [79] Hongfang Liu, Yves A Lussier, and Carol Friedman. A study of abbreviations in the umls. In *Proceedings of the AMIA Symposium*, page 393. American Medical Informatics Association, 2001.
- [80] Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. Rxnorm: prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23, 2005.
- [81] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.
- [82] R Mack, Sougata Mukherjea, Aya Soffer, Naohiko Uramoto, E Brown, Anni Coden, J Cooper, Akihiro Inokuchi, Bhavani Iyer, Yosi Mass, et al. Text analytics for life science using the unstructured information management architecture. *IBM Systems Journal*, 43(3):490–515, 2004.

- [83] Anne E Magurran. Measuring biological diversity. 2004.
- [84] G.S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM, 2007.
- [85] G.S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM, 2007.
- [86] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- [87] Massachusetts General Institute for Patient Care. Describing practice through clinical narratives, 2014. [Online; accessed Jul-2014].
- [88] M. McFarlane, S.S. Bull, and C.A. Rietmeijer. The internet as a newly emerging risk environment for sexually transmitted diseases. *JAMA: The Journal of the American Medical Association*, 284(4):443–446, 2000.
- [89] T.W. Menza, J.P. Hughes, C.L. Celum, and M.R. Golden. Prediction of HIV acquisition among men who have sex with men. *Sexually Transmitted Diseases*, 36(9):547, 2009.
- [90] Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775, 2013.
- [91] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform*, 35:128–44, 2008.
- [92] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [93] Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Cernocký. Empirical evaluation and combination of advanced language modeling techniques. In *INTERSPEECH*, pages 605–608, 2011.
- [94] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

- [95] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013.
- [96] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database\*. *International Journal of Lexicography*, 3(4):235–244, 1990.
- [97] George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
- [98] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.
- [99] M.B. Moore Jr, E.V. Price, J.M. Knox, and L.W. Elgin. Epidemiologic treatment of contacts to infectious syphilis. *Public Health Reports*, 78(11):966, 1963.
- [100] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252. Citeseer, 2005.
- [101] D.A. Moskowitz and D.W. Seal. GWM looking for sex SERIOUS ONLY: The interplay of sexual ad placement frequency and success on the sexual health of men seeking men on craigslist. *Journal of Gay & Lesbian Social Services*, 22(4):399–412, 2010.
- [102] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010.
- [103] National Oceanic and Atmospheric Administration. U.S. daily snowfall and snow depth data, 2011. [Online; accessed 1-Dec-2011].
- [104] American Society of Hospital Pharmacists. *AHFS Drug Information*. Published by authority of the Board of Directors of the American Society of Hospital Pharmacists, 1990.
- [105] D.H. Osmond, L.M. Pollack, J.P. Paul, and J.A. Catania. Changes in prevalence of HIV infection and sexual risk behavior in men who have sex with men in San Francisco: 1997–2002. *Journal Information*, 97(9), 2007.

- [106] Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527, 2010.
- [107] Olga Patterson and John F Hurdle. Document clustering of clinical narratives: a systematic study of clinical sublanguages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1099. American Medical Informatics Association, 2011.
- [108] Marco Pennacchiotti and Patrick Pantel. Entity extraction via ensemble semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 238–247. Association for Computational Linguistics, 2009.
- [109] Hoifung Poon and Pedro Domingos. Deep learning for semantic parsing.
- [110] Hoifung Poon and Pedro Domingos. Unsupervised ontology induction from text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 296–305. Association for Computational Linguistics, 2010.
- [111] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 41–47. Association for Computational Linguistics, 2002.
- [112] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. <http://is.muni.cz/publication/884893/en>.
- [113] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. 2013.
- [114] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1044–1049, 1996.
- [115] Ellen Riloff et al. Automatically constructing a dictionary for information extraction tasks. In *AAAI*, pages 811–816, 1993.
- [116] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

- [117] Benjamin Rosenfeld and Ronen Feldman. Using corpus statistics on entities to improve semi-supervised relation extraction from the web. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 600, 2007.
- [118] B.R.S. Rosser, W. West, and R. Weinmeyer. Are gay communities dying or just in transition? Results from an international consultation examining possible structural change in gay communities. *AIDS Care*, 20(5):588–595, 2008.
- [119] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive Modeling*, 1988.
- [120] Naomi Sager, Carol Friedman, and Margaret S Lyman. Medical language processing: Computer management of narrative data. 1987.
- [121] Naomi Sager, Margaret Lyman, Christine Bucknall, Ngo Nhan, and Leo J Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160, 1994.
- [122] Magnus Sahlgren. An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5, 2005.
- [123] Gerard Salton, Edward A Fox, and Harry Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.
- [124] T. Sanchez, T. Finlayson, A. Drake, S. Behel, M. Cribbin, E. DiNenno, T. Hall, S. Kramer, A. Lansky, Centers for Disease Control, and Prevention (US). *Human Immunodeficiency Virus (HIV) Risk, Prevention, and Testing Behaviors: United States, National HIV Behavioral Surveillance System: Men who Have Sex Men, November 2003 [to] April 2005*. US Department of Health and Human Services, 2006.
- [125] Sunita Sarawagi and William W Cohen. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, pages 1185–1192, 2004.
- [126] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
- [127] S. Schwarcz, S. Scheer, W. McFarland, M. Katz, L. Valleroy, S. Chen, and J. Catania. Prevalence of HIV infection and predictors of high-transmission sexual risk behaviors among men who have sex with men. *Journal Information*, 97(6), 2007.

- [128] Holger Schwenk and Jean-Luc Gauvain. Training neural network language models on very large corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 201–208. Association for Computational Linguistics, 2005.
- [129] Yongzhe Shi, Wei-Qiang Zhang, Jia Liu, and Michael T Johnson. Rnn language model with word clustering and class-based output layer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–7, 2013.
- [130] Chaitanya Shivade, James Cormack, and David Milward. Precise medication extraction using agile text mining. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 75–79, 2014.
- [131] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809, 2011.
- [132] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.
- [133] Sunghwan Sohn, Cheryl Clark, Scott R Halgrim, Sean P Murphy, Christopher G Chute, and Hongfang Liu. MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, pages amiajnl-2013, 2014.
- [134] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. brat: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.
- [135] R Development Core Team. R version 2.12.0. R project for statistical computing, 2009.
- [136] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
- [137] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd Workshop on Learning language in Logic and the 4th Conference on Computational Natural Language Learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.



- [138] Herman D Tolentino, Michael D Matters, Wikke Walop, Barbara Law, Wesley Tong, Fang Liu, Paul Fontelo, Katrin Kohl, and Daniel C Payne. A UMLS-based spell checker for natural language processing in vaccine safety. *BMC Medical Informatics and Decision Making*, 7(1):3, 2007.
- [139] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.
- [140] Peter Turney. Measuring semantic similarity by latent relational analysis. 2005.
- [141] Peter D Turney. Domain and function: A dual-space model of semantic relations and compositions. *arXiv preprint arXiv:1309.4035*, 2013.
- [142] United States Census Bureau. 2000 redistricting data (public law 94-171) summary file, August 2013. [Online; accessed Aug-2013].
- [143] United States Postal Service. Official USPS abbreviations, May 2014. [Online; accessed May-2014].
- [144] U.S. Census Bureau. 2010 TIGER/Line shapefiles, May 2012.
- [145] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.
- [146] Xia Uzuner, Solti. i2b2 medication extraction challenge preliminary annotation guidelines, 2009. [Online; accessed 1-Dec-2014].
- [147] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.
- [148] H.W. Vosburgh, G. Mansergh, P.S. Sullivan, and D.W. Purcell. A review of the literature on event-level substance use and sexual risk behavior among men who have sex with men. *AIDS and Behavior*, pages 1–17, 2012.
- [149] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. Crowd-sourced urban life monitoring: Urban area characterization based crowd behavioral patterns from Twitter. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, page 26. ACM, 2012.
- [150] Wikipedia. Party and play — Wikipedia, the free encyclopedia, 2014. [Online; accessed 1-Dec-2014].

- [151] Wikipedia. Poppers — Wikipedia, the free encyclopedia, 2014. [Online; accessed 1-Dec-2014].
- [152] DHJ Wohlfeiler, HF Raymond, T Kennedy, and W McFarland. How can we improve HIV and STD prevention online for MSM? Assessing the preferences of website owners, website users, and hiv/std directors, 2011.
- [153] C. Xiao, W. Wang, X. Lin, J.X. Yu, and G. Wang. Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems (TODS)*, 36(3):15, 2011.
- [154] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. Spamming botnets: signatures and characteristics. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 171–182. ACM, 2008.
- [155] Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.
- [156] Limin Yao, Sebastian Riedel, and Andrew McCallum. Universal schema for entity type prediction.
- [157] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv.org*, 2015.
- [158] Zillow. Zillow neighborhood boundaries, May 2012. [Online; accessed May-2012].