Inferring Individual Attributes from Search Engine Queries and Auxiliary Information

Luca Soldaini[*]
Georgetown University
Washington, D.C., USA
luca@ir.cs.georgetown.edu

Elad Yom-Tov
Microsoft Research
Herzliya, Israel
eladyt@microsoft.com

ABSTRACT

Internet data has surfaced as a primary source for investigation of different aspects of human behavior. A crucial step in such studies is finding a suitable cohort (i.e., a set of users) that shares a common trait of interest to researchers. However, direct identification of users sharing this trait is often impossible, as the data available to researchers is usually anonymized to preserve user privacy. To facilitate research on specific topics of interest, especially in medicine, we introduce an algorithm for identifying a trait of interest in anonymous users. We illustrate how a small set of labeled examples, together with statistical information about the entire population, can be aggregated to obtain labels on unseen examples. We validate our approach using labeled data from the political domain.

We provide two applications of the proposed algorithm to the medical domain. In the first, we demonstrate how to identify users whose search patterns indicate they might be suffering from certain types of cancer. This shows, for the first time, that search queries can be used as a screening device for diseases that are currently often discovered too late, because no early screening tests exists. In the second, we detail an algorithm to predict the distribution of diseases given their incidence in a subset of the population at study, making it possible to predict disease spread from partial epidemiological data.

Keywords

Query log analysis; Disease screening; Health informatics

1. INTRODUCTION

Identifying people with specific demographics, interests, or traits is a topic long of interest for researchers interested in online behavior and communities [12, 13]. The ability to identify a cohort—that is, a group of people with a common defining characteristic—is a critical phase of the research process. For example, when studying how the internet is used to seek medical advice, researchers have employed diverse heuristics to identify medical queries [19, 31] or health information seekers [23, 37]. Such heuristics are usually sufficient at identifying common queries and conditions, but fail to capture small cohorts, such as users suffering from an uncommon disease. While such groups could be identified using personal information, demographic data, or health records, much of the data available to researchers is anonymous, in an effort to preserve the privacy of individuals.

In this paper, we introduce an algorithm for inferring individual attributes of a population of users by relying on a small set of examples with known labels and statistical information about the entire population. In other words, we show how to identify a cohort of interest by learning from a small set of users—which we identified using a very effective yet low-recall heuristic—and information about the distribution of the cohort of users in the entire population.

We validate the proposed algorithm by identifying the political affiliation of Twitter users: given a set of users and their tweets, we predict their political affiliation using a small set of users with known political orientation and statistics about the outcome of the elections. Our algorithm determines the affiliation of such users more effectively than other methods when the fraction of known users is small.

Finally, we present two applications of the proposed algorithm. Both use the proposed system to create a cohort of users whose search patterns indicate they might be suffering from specific forms of cancer. No personal data or patient history is used; rather, we combine a low recall, high precision heuristic with epidemiological data about incidence of the cancer. Once the cohort has been determined, we show how to use it to train and evaluate a classifier to pre-screen for users who suffer from the cancer at study. This shows how search queries can be used as a screening device for diseases that are often discovered too late, because no early screening tests currently exists. Furthermore, we present a classifier that uses the cohort identified by the algorithm to predict the incidence of disease in regions where it is not known. Such application could be useful to estimate the spread of a disease in regions where the number of reported cases is not sufficient to carry out a statistical analysis.

In summary, our contribution is threefold:

- We study the problem of **identifying cohorts of users who share a common trait** (e.g., they suffer from the same medical condition) from a population;

[*] Work carried out during internship at Microsoft Research.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052629

- We **propose and evaluate an algorithm** that uses fine-grained data on users and coarse-grained population statistics to identify cohorts for research purposes;
- We **describe and solve two possible applications** of the proposed algorithm: identification of users who might suffer from certain types of rare cancers and predicting the distribution of a disease in regions where it is unknown.

2. RELATED WORKS

Traditionally, most of the medical research exploiting internet data has focused on population-level disease incidence. The questions therein are of the form "*how many people in a given area are currently suffering from influenza?*" [11]. Because of the large number of people involved, it is superfluous to identify each individual with the condition. Instead, it is sufficient to find correlations between disease incidence and specific keywords [22, 24] or even website visits [17].

More recently, researchers have begun attempting to identify anonymous search engine users suffering from conditions of interest, either to provide individual level predictions or to learn from individual behaviors. For example, Yom-Tov et al. [37] identified people suffering from mood disorders according to their queries of drugs used to treat the disorder, as well as changes in their behavior near the time of mood disorder events. In other work, Ofran et al. [18] used a threshold on the number of cancer-specific queries to identify people who were likely diagnosed with cancer and then track their information needs over time. Good correlation was found between the number of people searching for cancer and disease incidence (but not prevalence) in the USA. A more fine-grained approach was taken in Yom-Tov et al. [35] where a small subset of users was found to have identified themselves as suffering from a condition of interest. The queries of this population were used to construct a classifier that predicted whether the condition a user was asking about most often was one they were suffering from. The ability to identify users with specific conditions was then used to analyze their search histories for precursors of disease. More recently, Paparrizos et al. [20] used people who self-identified as suffering from pancreatic cancer to predict their diagnosis ahead of time.

The task of determining labels for individuals from population statistics relates to the ecological inference problem. Ecological inference aims at inferring characteristics about individuals from ecological data (i.e., of the entire population). As an example, it might be used to answer the following question: "*Given the number of votes for political parties A and B in a precinct and the number of men and women in the precinct, how many women voted for party A?*" Ecological inference has a long history in the fields of statistics and social studies [15]. Recently, Flaxman, et al. [10] used kernel embeddings of distributions to predict which demographics groups supported Barack Obama in the 2012 US Presidential Election. Park and Gosh [21] introduced LUDIA, a low-level rank approximation algorithm designed that leverages ecological inference to predict hospital spending for individuals based on their length of stay. Culotta, et al. [6] used website traffic data to predict demographics of Twitter user. Ultimately, our problem differs from ecological inference in that we are interested in identifying individuals whose distribution is known rather than inferring behaviors at an individual level from population data.

Another area of study that bears a similarity with our proposed algorithm is Learning with Label Proportions (LLP). In LLP, the training data is provided to the classifier in groups on which only the distribution of classes in each group is known. Many solutions have been proposed for the problem [16, 26]; yet—to the best of our knowledge—none of them is designed to bias the learning process by incorporating individuals with known labels. Keerthi, et al. [29] introduced a semi-supervised SVM classifier that uses a small labeled dataset in conjunction to class proportion on the training data to predict labels on test data. While sharing some similarity with our algorithm, their method is less generalizable, as it does not handle learning from training data drawn from sets with different class distributions. Instead, our proposed approach solves this issue by conjunctively optimizing correlation with all sets the training data is drawn from.

Finally, many have studied semi-supervised learning (SSL), the problem of learning when a combination of labeled and unlabeled examples are available [4]. For example, Druck, et al. [7] proposed a framework that leverages labeled features—that is, features that are highly representative for a class—to learn constrains for a multinomial logistic regression. More recently, Ravi and Diao [27] have proposed a graph model to efficiently use SSL on large datasets. Compared to a classic SSL model, we not only leverage individual level features, but also take advantage of population data.

3. METHODOLOGY

3.1 Notation

Throughout this paper, we will adhere to the following notation: scalars are identified by lowercase italic letters (e.g., s), vectors by lowercase bold letters (e.g., $\boldsymbol{v}$), and matrices by uppercase italic letters (e.g., M). Calligraphy uppercase letters (e.g., $\mathcal{X}$) are used to denote sets.

Let $\mathcal{X}$ be a population of size $n = |\mathcal{X}|$. To each element of $\mathcal{X}$, we associate the following: a features vector $\boldsymbol{x}_i = \{x_{ij}\}_{j=1}^{m}$, a label $y_i \in \{0, 1\}$, and a property vector $\boldsymbol{p}_i = \{p_{ik}\}_{k=1}^{t}$. y_i has value "1" if the i-th example belongs to the cohort of interest, "0" if its membership is unknown. We refer to the $n \times m$ matrix of all features vector as X. A feature could be, for example, the use of a certain phrase by a user. $\boldsymbol{p}_i$ represent a set of properties for an individual we directly take advantage in the proposed method. For example, a property of an individual could be the US state where they are located; in this case, $\boldsymbol{p}_i$ would be a 1×51 vector whose k-th position equals to "1" if the i-th individual is located in the k-th state, "0" otherwise. While a property vector $\boldsymbol{p}_i$ is a feature vector for the i-th element of $\mathcal{X}$, it is convenient to consider it separately from $\boldsymbol{x}_i$, as it simplifies the definition of the algorithm introduced in Section 3.2.

We denote by $\boldsymbol{y}$ a $n \times 1$ vector holding all labels, while P is a $n \times t$ matrix whose element (i, k) represent the value of the k-th property for the i-th element of the population.

We encode the known statistical information as a $1 \times t$ vector $\boldsymbol{\pi}$ containing statistical information about the property of individuals in $\mathcal{X}$. For example, given a disease and a population of users located in the USA, $\boldsymbol{\pi}$ could be a vector containing the incidence of the disease in each state.

Finally, we establish the notation for functions that will be used extensively in the remainder of the paper. $\mathrm{H}(a, b)$ indicates the harmonic mean between the values a and b. We

Algorithm 1: The proposed SGD algorithm.

```
Data: Features matrix $X$, labels vector $\boldsymbol{y}$, property matrix
      $P$, statistical information function $\boldsymbol{\pi}$, number of
      iteration $\eta$, and learning percentile $\delta$.
Result: Vector $\boldsymbol{l} = \{l_i\}$ of confidence values of element in $\mathcal{X}$
        to be in the cohort of interest.
begin
    $\boldsymbol{w} \leftarrow$ initializeHyperplane( ); $o^* \leftarrow 0$
    for $j = 1, \ldots, \eta$ do
        $i \leftarrow$ randomSample($\{1, \ldots, n\}$); $\boldsymbol{x}_i \leftarrow X[i]$
        $\boldsymbol{c}^+ \leftarrow (\boldsymbol{w} + \boldsymbol{x}_i)/||\boldsymbol{w} + \boldsymbol{x}_i||$
        $\boldsymbol{c}^- \leftarrow (\boldsymbol{w} - \boldsymbol{x}_i)/||\boldsymbol{w} - \boldsymbol{x}_i||$
        $\boldsymbol{d}^+ \leftarrow X \cdot \boldsymbol{c}^+$; $\boldsymbol{d}^- \leftarrow X \cdot \boldsymbol{c}^-$
        $o^+ \leftarrow \mathrm{H}\left(\mathrm{Corr}(\boldsymbol{\pi}, P, \boldsymbol{d}^+, \delta), \mathrm{Recall}(\boldsymbol{y}, \boldsymbol{d}^+, \delta)\right)$
        $o^- \leftarrow \mathrm{H}\left(\mathrm{Corr}(\boldsymbol{\pi}, P, \boldsymbol{d}^-, \delta), \mathrm{Recall}(\boldsymbol{y}, \boldsymbol{d}^-, \delta)\right)$
        if $o^+ > o^-$ and $o^+ > o^*$ then
            $\boldsymbol{w} \leftarrow \boldsymbol{c}^+$; $o^* \leftarrow o^+$
        else if $o^- > o^*$ then
            $\boldsymbol{w} \leftarrow \boldsymbol{c}^-$; $o^* \leftarrow o^-$
    $\boldsymbol{d}^* \leftarrow X \cdot \boldsymbol{w}$
    $\boldsymbol{l} \leftarrow$ SoftmaxNormalize($\boldsymbol{d}^*$)
```

represent Spearman's rank correlation coefficient between values of vectors $\boldsymbol{r}$ and $\boldsymbol{s}$ as $\rho_{\mathrm{s}}(\boldsymbol{r}, \boldsymbol{s})$. $\mathrm{Perc}(\boldsymbol{r}, \alpha)$ returns the value in $\boldsymbol{r}$ corresponding to the α^{th} percentile; building on the previous notation, we define the following operator:

$$\mathrm{PercSel}(\boldsymbol{r}, \alpha) = \{i \mid r_i \geq \mathrm{Perc}(\boldsymbol{r}, \alpha) \quad \forall r_i \in \boldsymbol{r}\} \qquad (1)$$

PercSel selects the set of indices of $\boldsymbol{r}$ whose corresponding values are in the α^{th} percentile. The result of such function can be used to extract the matching components of any vector $\boldsymbol{s}$:

$$\lceil \boldsymbol{s} \rceil_{\boldsymbol{r}, \alpha} = \{s_i \mid i \in \mathrm{PercSel}(\boldsymbol{r}, \alpha)\} \qquad (2)$$

We will take advantage of the notation $\lceil \boldsymbol{s} \rceil_{\boldsymbol{r}, \alpha}$ to identify the α^{th} percentile of vector $\boldsymbol{s}$ with respect to weight vector $\boldsymbol{r}$ throughout the manuscript.

3.2 Proposed Algorithm

Recall that, given a population $\mathcal{X}$, we wish to identify a subset of $\mathcal{X}$—i.e., a cohort—such that all members of the cohort share a property of interest. A solution for such problem should return a vector $\boldsymbol{l}$ of real values between 0 and 1 expressing the likelihood of each individual in $\mathcal{X}$ of being part of the cohort of interest. A naïve solution consists of using a classifier trained on the set of known members in the cohort. However, as we will describe in Section 4.2 and Table 1, this approach does not work well when the size of the set of users with known labels is small.

The algorithm we propose in this paper addresses this issue by conjunctively maximizing two quantities: (i) the correlation between the counts of properties in the δ^{th} percentile of users and the statistical information vector $\boldsymbol{\pi}$ (e.g., the correlation between the number of users in the δ^{th} percentile for each state and the incidence of the disease in each state), and (ii) the fraction of known positive users (i.e., users whose label is "1") in the δ^{th} percentile. By optimizing for both quantities at the same time, we exploit the individual features of users that whose label is known, as well as statistical information about the distribution of the cohort of interest.

The algorithm works by finding a linear separating hyperplane which assigns a predicted label to each user given the features thereof. We formally define the two aforementioned quantities as follows: Let $\boldsymbol{d} = X \cdot \boldsymbol{w}$ be the vector of signed distances of elements from the decision hyperplane $\boldsymbol{w}$; then, $\lceil P \rceil_{\boldsymbol{d}, \delta}$ is the $n' \times t$ property matrix associated with elements whose distances from $\boldsymbol{w}$ are in the δ^{th} percentile. In other words, $\lceil P \rceil_{\boldsymbol{d}, \delta}$ contains the property vectors of those elements with distance from the decision hyperplane greater or equal than $\mathrm{Perc}(\boldsymbol{d}, \delta)$. Thus, we can define quantity (i) as:

$$\mathrm{Corr}(\boldsymbol{\pi}, P, \boldsymbol{d}, \delta) := \rho_{\mathrm{s}}\left(\boldsymbol{\pi},\ \mathbf{1}^{1 \times n'} \cdot \lceil P \rceil_{\boldsymbol{d}, \delta}\right) \qquad (3)$$

where $\mathbf{1}^{1 \times n'}$ is the unit vector of size $1 \times n'$.

The fraction of known positive users (ii) is the recall of the algorithm on known users; that is, the fraction of positive users whose distance from $\boldsymbol{w}$ is in the δ^{th} percentile:

$$\mathrm{Recall}(\boldsymbol{y}, \boldsymbol{d}, \delta) := \frac{\left|\{y_i \mid y_i = 1 \quad \forall y_i \in \lceil \boldsymbol{y} \rceil_{\boldsymbol{d}, \delta}\}\right|}{\left|\{y_i \mid y_i = 1 \quad \forall y_i \in \boldsymbol{y}\}\right|} \qquad (4)$$

The two quantities determined by functions "Corr" and "Recall" are combined by considering the harmonic mean of the two as the objective function. We chose this mean as it penalizes the algorithm if the two quantities diverge significantly.

We used a modification of the Perceptron algorithm [28] in which a stochastic gradient descent learns the hyperplane $\boldsymbol{w}$ separating elements of the positive and negative classes that maximizes the objective function described previously.

The details of the procedure are shown in Algorithm 1. The algorithm iterates η times over all elements in the population; at each iteration, it randomly samples an element i; then, it generates two candidate hyperplanes $\boldsymbol{c}^+$ and $\boldsymbol{c}^-$ by respectively adding and subtracting $\boldsymbol{x}_i$ from $\boldsymbol{w}$. If any of the two candidate hyperplanes increases the value of the objective function, it then replaces $\boldsymbol{w}$. Finally, the confidence vector $\boldsymbol{d}^*$ is calculated by multiplying the feature matrix X with $\boldsymbol{w}$, and normalized by applying a Softmax normalization [25] to obtain the likelihood vector $\boldsymbol{l}$.

We note that the objective function defined in Algorithm 1 is not convex. While this implies that we cannot provide strong theoretical guarantees about the stability of the algorithm, we experimentally verified that the classifier leads to consistent results (Section 6.1).

4. VALIDATION

We validate the proposed algorithm on a dataset containing US Twitter users with known political affiliation. This task has been studied in the past (e.g., [1, 5, 6]); in this paper, we use it as a benchmark for the proposed algorithm.

We show how the algorithm introduced in Section 3.2, when combined with statistical data on the outcome of the 2012 US presidential election, can be used to infer the political affiliation of users. To do so, we hide a fraction γ of users with known political affiliation by assigning them the label "0"; then, we measure the ability of the stochastic gradient descent in identifying these hidden users.

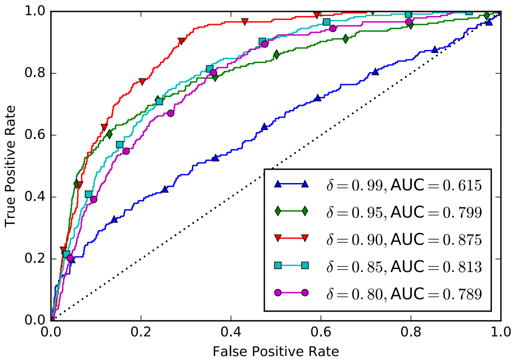


Figure 1: ROC curve of the stochastic gradient descent algorithm. The fraction of "hidden" Republicans is kept constant at $\gamma = 0.75$, while the learning percentile δ is varied.

4.1 Data Description

Similarly to [8], we took advantage of a set of Twitter users with known political affiliation to evaluate our system. Our dataset contains 372,769 users who explicitly expressed support for Barack Obama and 22,902 users who expressed preference for Mitt Romney during the 2012 US presidential election. For the remainder of the paper, we will refer to the two groups as "Democrats" and "Republicans", while the set of all users will be identified as $\mathcal{U}$.

The political affiliation of members of $\mathcal{U}$ was determined by two sets of hashtags used by the supporters of the two parties during the election (e.g. "*#romneyryan2012*", "*#voteobama*"; the complete list is available in [8]). This heuristic was found to have over 95% accuracy [8].

The set $\mathcal{T}$ of all tweets generated by users in $\mathcal{U}$ between August 1$^{\text{st}}$, 2012 and November 15$^{\text{th}}$, 2012 was extracted. We discarded all users for whom no location data was available (i.e., none of their tweets was geotagged), tweeted from two or more US states, or less than 30 times. We identify the set of the $15,472$ remaining users (900 Republicans and $14,572$ Democrats) as $\mathcal{X}$.

We use the set $\mathcal{T}_{\mathcal{X}}$ of tweets associated with users in $\mathcal{X}$ to construct the feature matrix X. For each user, we extracted the following features from their tweets: hashtags, mentions, domain name of URLs, and words occurring 10 or more times in the corpus (except stopwords). Prior works found such features to be effective at predicting the political affiliation of users [1, 5, 6, 8]; in this work, we investigate their effectiveness when paired with the proposed algorithm.

We represent the state each user in $\mathcal{X}$ belongs to through property matrix P; in other words, P is a $15,472 \times 51$ matrix where position (i, k) is equal to 1 if the i-th user tweets from the k-th state (according to geo-tagging), 0 otherwise.

The population statistic vector $\boldsymbol{\pi}$ was derived from the total count of votes casted for the Republican and Democrat candidates in each US state as disclosed by the official Federal Election Commission report [9]. Specifically, the k-th value of $\boldsymbol{\pi}$ represents the number of Republican voters for each inhabitant in the k-th state. We normalized $\boldsymbol{\pi}$ by the number of active users in each state within the time frame of data collection; this gave us the expected number of Republican Twitter users in each state.

4.2 Results

In this section, we illustrate the performance of the proposed algorithm in identifying Republican users whose label have been hidden. Recall that, in order to quantify the ability of Algorithm 1 to correctly label users in $\mathcal{X}$, we remove the label of a fraction γ of republican users; that is, we assign them the label "0". Therefore, we tested the algorithm for various values of γ, as well as multiple values of learning percentile δ. The exploration of different parameters allows us to test how the algorithm behaves when number of users in the cohort is not known (if available, such number could be used to tune δ). All the experiments were executed under five-fold stratified cross validation; the number of iterations η was set to $30,000$ to ensure reaching a stable point.

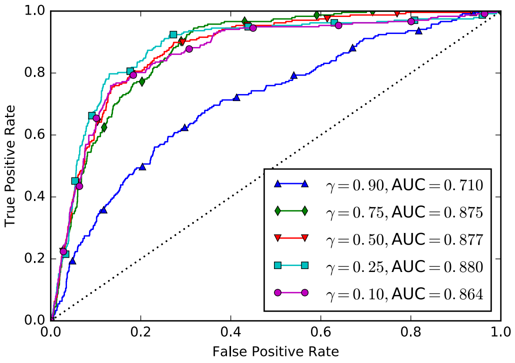


Figure 2: ROC curve of the stochastic gradient descent algorithm. The learning percentile is kept constant at $\delta = 0.9$, while the fraction of "hidden" Republicans γ is varied.

We report the results of our experiments in Figures 1 and 2, as well as in Tables 1 and 2. Specifically, Figure 1 shows the Receiver Operating Characteristic (ROC) curves produced by varying values of the learning percentile δ. For this experiment, γ is fixed at 0.75.

Two observations can be made about the results. First, we note that the performance of the classifier, as measured by the Area Under the Curve (AUC), increases as δ approaches 0.9; then it starts declining. We explain this behavior by observing that the number of elements in the 10$^{\text{th}}$ percentile is close to the number of Republican in $\mathcal{X}$; therefore, as δ approaches this value, the performance of the proposed algorithm increase. Past the 10$^{\text{th}}$ percentile, more noise is introduced, thus affecting the quality of classification outcome. We remark that the decline in performance is not abrupt; this characteristic is desirable, as in applications of the proposed algorithm (such as those shown in Section 5) the exact size of the cohort of interest is often unknown.

Second, we observe that classifiers with larger values of δ (e.g., $\delta = 0.95$) have a higher true positive rate associated with lower false positive rate (bottom left of Figure 1). This is likely due to the fact that such classifiers make fewer mistakes on elements they have high confidence in (i.e., the values in $\boldsymbol{l}$ for with high-confidence elements are close to 1).

For the second experiment (Figure 2) we varied the fraction of hidden users γ between 0.9 (i.e., only 10% of Republicans are disclosed) and 0.1 (90% of Republicans are disclosed). We observe the AUC increases as γ decreases; that is to be expected, as less hidden republicans equals a more diverse pool of training examples. However, to our initial surprise, we also noticed that the performance of the classifier show little improvement for values of $\gamma < 0.75$. Such behavior is beneficial for the applications where this algorithm will be used, where typically only a small set of users of the cohort of interest is known.

Classifier	AUC
Linear Stochastic Gradient Descent (LSGD)	0.667
LSGD + property vectors as features	0.614
Support Vector Machine (SVM) from [5]	0.703
SVM [5] + property vectors as features	0.629
Proposed SGD ($\delta = 0.9$)	**0.875**

Table 1: Comparison of the proposed algorithm to previously-proposed baselines. When a small fraction of Republican users is used for training ($\gamma = 0.75$), the algorithm outperforms a linear SGD baseline and the system from [5] (difference is statistically significant, Wilcoxon signed-rank test, $p < 0.05$).

Rank	Feature	Weight
1	#4moredays	0.0517
2	#landslide	0.0490
3	#loveofcountry	0.0377
4	#whyiamnotvotingforobama	0.0244
5	#whyimnotvotingforobama	0.0229
6	#bengahzi	0.0148
7	anncoulter.com	0.0129
8	searchnc.com	0.0112
9	#bengha	0.0111
10	personalliberty.com	0.0110

Table 2: Top ten features for classifier $\delta = 0.9, \gamma = 0.75$. Websites ranked 7, 8, and 10 are right-leaning publications.

We compared the proposed system with a simple Linear Stocastic Gradient Descent (LSGD), as well as with the Support Vector Machine (SVM) classifier proposed by Conover, et al. in [5]. For all three systems, we kept the ratio of hidden Republican users set to $\gamma = 0.75$, as we were interested in studying how the proposed algorithm compares to other algorithms when only a small set of users in the cohort is known. As shown in Table 1, the proposed algorithm outperforms both baselines, confirming that combining statistical information about the distribution of the cohort in the population with weights learned from individual features is an effective strategy to solve the task introduced in this paper. For the two baselines, we experimented with using just the features in X to train the classifiers, as well as concatenating $P \cdot \boldsymbol{\pi}$ with the features matrix X. Interestingly, the performance of the two baselines decrease when augmenting X with $P \cdot \boldsymbol{\pi}$, suggesting that the naïve approach of expanding the feature set with population statistics is not effective at identifying the cohort of users.

The features that were assigned the highest weights are the most indicative phrases used by positive (Republican) users. We report features with the highest weight in w for the classifier $\delta = 0.9, \gamma = 0.75$ in Table 2. We note that the hashtags ranked in 1^{st}, 3^{rd}, 4^{th}, 5^{th}, 6^{th}, 9^{th} places are typically used in right-wing circles; the remaining hashtag ("*#landslide*") while related to the election, is not unique to the rhetoric on any of the two political parties. Finally, we note that the all URLs shown in Table 2 are of websites leaning on the right side of the political spectrum.

5. APPLICATIONS

We present two applications of the algorithm introduced in the previous sections. The first (Section 5.2) deals with identifying users of a search engine whose search patterns suggest a higher risk of developing a certain type of cancer; the second (Section 5.3) is concerned with predicting the incidence of two forms of cancer in regions of the USA.

Features in X	Features in Z
list of ovarian/cervical cancer symptoms*	list of ovarian/cervical cancer symptoms*
q most common terms in queries $\mathcal{Q}$ **after** disease mention	q most common terms in queries $\mathcal{Q}$ **before** disease mention
list of symptoms*	list of symptoms*
list of diseases*	
list of names of drugs	
list of names of US hospitals	

Table 3: Features used to construct matrices X and Z. Features in X are extracted from queries issued *after* the first query mentioning the disease, while features in Z are extracted from queries issued *before* the first query mentioning the disease. q was set to $2,000$ after empirical evaluation.

We focus on ovarian cancer and cervical cancer. These relatively rare cancers (affecting approximately 12 and 10, respectively of $100,000$ women in the USA), are also quite deadly: Indeed, though ovarian cancer accounts for only 3% of all cancers in women, it is the deadliest cancer of the female reproductive system [32]. One reason for this is that symptoms of these cancers are relatively benign, which means that many women are diagnosed at late stages of the cancer, though treatment is most effective in early stages. Additionally, no simple screening test is available for these cancers [3]. Thus, the ability to pre-screen for these cancers using Internet data could be of significant importance.

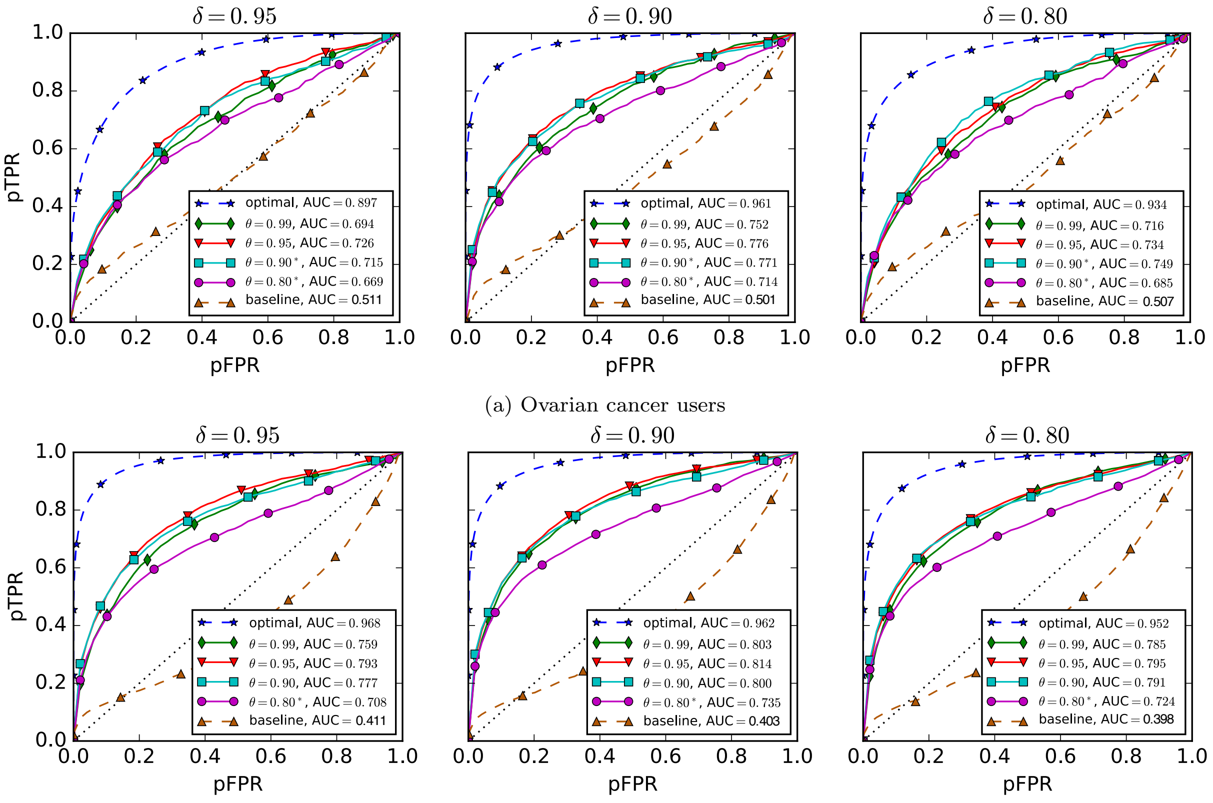
5.1 Data Description

Two sets of Bing users were considered to evaluate the applications introduced in this section. The first population, which we identify as $\mathcal{X}^{ov}$, consists of users who are likely to be suffering from ovarian cancer. Our interest in studying this disease is due to the fact that, despite its low incidence (it only accounts for 3% of all cancers in women), ovarian cancer is the deadliest cancer of the female reproductive system [32]. Furthermore, while treatment for ovarian cancer is the most effective in early stages, no screening test is available yet [3]. The second population of users is of users who are potentially suffering from cervical cancer. Such disease, while less deadly than ovarian cancer, has a similar incidence, thus being another useful dataset to validate the applications presented in this section. We refer to this group as $\mathcal{X}^{cr}$.

We stress that it is traditionally very challenging to identify those users in $\mathcal{X}^{ov}$ or $\mathcal{X}^{cr}$ who are affected by the aforementioned conditions using Internet data, because both diseases have a low incidence rate, thus causing any heuristic—such as extracting all users who issue a specific query—to retrieve too few individuals. Thus, we apply the algorithm introduced in Section 3 to obtain an estimate of the probability of each user of suffering from cancer.

For the remainder of the paper, we will refer to the set of users $\mathcal{X}$ to describe all procedures that are common to both $\mathcal{X}^{ov}$ and $\mathcal{X}^{cr}$; differences will be pointed out when necessary.

To obtain $\mathcal{X}$, we proceed as follows: First, using the websites of the Center for Diseases Control (CDC) and the American Cancer Society, we produced a list of symptoms and drugs commonly associated with the each of the two diseases. The two lists were expanded using two experts-to-laypeople synonym mappings, *MedSyn* [34] and *Behavioral* [36]. This expansion was made so as to bridge the gap be-



(a) Ovarian cancer users

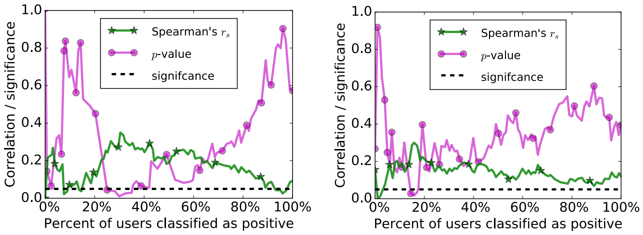
(b) Cervical cancer users

Figure 3: pROC curves for ovarian cancer (top) and cervical cancer (bottom). The values of the learning percentile δ are reported above each figure. The AUC of $\mathbb{C}_1$ under multiple values of θ are shown alongside the optimal and baseline classifiers. Values of θ maked as $*$ are statistically different from the best runs (Wilcoxon signed-rank test, $p < 0.05$).

tween the vocabulary used by health experts and expressions preferred by laypeople [30]. We extracted all Bing users in the United States who have queried in English in a span of five months: Ovarian cancer from April to August 2015 and cervical cancer from June 2015 to October 2015, for any of the symptoms or drugs associated with the cancer and the name of the cancer itself. Finally, we extracted the US state of origin of each user through reverse IP address lookup to take advantage of the state-level incidence statistics for the two types of cancer. Users who were associated with two or more US states were discarded. This heuristic identified $3,167$ users who potentially have ovarian cancer and $9,327$ users who might have been diagnosed with cervical cancer. Not all users in the two sets are affected by the respective diseases; rather, the heuristic was used to reduce class imbalance before using the proposed algorithm to derive their likelihood of having the condition. We refer to the set of all queries issued by all users in $\mathcal{X}$ as $\mathcal{Q}$.

For both conditions, we identify a set of users who are known (by their own admission) to be affected by cancer, as in [35]. This was done by finding all users who issued a query starting with "*i have* `<condition>`" or "*i was diagnosed with* `<condition>`", where `<condition>` is either "*ovarian cancer*" or "*cervical cancer*". We will refer to these users as "self-identified users" or SIUs. Through this heuristic, we extracted 140 users for ovarian cancer, and 41 users for cervical cancer. We assigned the label "1" to this subset of $\mathcal{X}$, while the rest of the users were labeled as "0".

We define two features matrix X and Z using the queries in $\mathcal{Q}$. For each user, X contains features extracted from queries issued after the first query mentioning the disease. For example, the i-th row of X^{ov} contains features extracted from all queries submitted by the i-th user in $\mathcal{X}^{ov}$ after searching for *"ovarian cancer"* for the first time. Conversely, Z contains features extracted from all queries issued before the first query mentioning the disease. A full list of features used in X and Z is reported in Table 3. The features matrix Z is used by the classifiers introduced in Sections 5.2 and 5.3. Matrix Z is comprised of queries mentioning symptoms and most common tokens, excluding stopwords, numbers, or names of the top one hundred websites in the US as ranked by Alexa (`http://alexa.com`). The latter was used so as to remove navigational queries. The number of tokens in Z exceeded, for both datasets, fifty thousand. In an effort to remove noise, we decided to keep only the top q tokens; q was set to $2,000$ after empirical evaluation. The feature matrix X is used by the stochastic gradient descent to infer, for each user, their likelihood of being affected by cancer; therefore, we also consider names of diseases, drugs, and US hospital as features. Upon completion of the feature extraction phase, matrices X^{ov}, X^{cr}, Z^{ov}, and Z^{cr} contain 7605, 8766, 2176, and 2170 features respectively.



(a) Ovarian cancer users

(b) Cervical cancer users

Figure 4: In **green**, Spearman's rank correlation coefficient ρ_s between the users identified by classifier $\mathbb{C}_2$ and disease incidence as a function of users identified by $\mathbb{C}_2$. p-values are shown in **purple**. For ovarian cancer, $\mathbb{C}_2$ achieve a statistically significant correlation ($\rho_s = 0.35$, $p < 0.05$) when 963 users in $\mathcal{X}_{ov}$ are classified as positive (for cervical cancer: $\rho_s = 0.35$, $1,483$ individuals).

As in Section 4.1, we represent the state each user in $\mathcal{X}$ belongs to through the property matrix P. The population statistic vector $\boldsymbol{\pi}$ was obtained from the CDC [33]. Similarly to Section 4.1, the vector was normalized by the number of active Bing users in each state during the data period.

5.2 Suggesting Medical Pre-Screening to Users

Here, we introduce a classifier designed to identify search engine users who show signs of being potentially affected by cancer. The classifier is designed to assess, for each searcher, their risk of developing cancer based on their query logs. The classifier is based on the labels inferred using the proposed algorithm, and uses past queries to assess if users will later be classified as suffering from the cancer of interest.

A logistic classifier $\mathbb{C}_1$ is trained to achieve the desired goal. The classifier uses the feature matrix Z; to obtain labels to train the system on, we proceed as follows: first, we run the stochastic gradient descent on input $(X, P, \boldsymbol{y}, \boldsymbol{\pi}, \eta, \delta)$, where X, P, and $\boldsymbol{y}$ are as defined in Section 5.1, η is set to $10,000$ for ovarian cancer and to $30,000$ for cervical cancer, and δ is varied between 0.95 and 0.80. Then, once obtained the confidence vector $\boldsymbol{l}$ for elements in $\mathcal{X}$, we extract users whose risk factor is in the θ^{th} percentile of $\boldsymbol{l}$, as well as those users whose risk factor is in the λ^{th} percentile of $\boldsymbol{l}$. The former are used as positive training examples, while the latter are used as negative training examples.

Since we expect the number of users with no cancer to be greater than the number of users with cancer, we fix $\lambda = 3(1-\theta)$. Therefore, the training set contains three negative examples for each positive example, somewhat mitigating the class imbalance probelm. The weighting of each class was adjusted accordingly when training the classifier.

As a baseline, we consider a linear SVM trained solely on self-identified users. This baseline was adapted from the classifier introduced by Yom Tov, et al. [35] to identify search engine users who have specific medical issues. Specifically, we use SIUs as positive training examples and a sample of users from the remainder of the population as negative examples. Similarly to $\mathbb{C}_1$, we sample three times the number of SIUs as negative examples.

5.2.1 Results

Standard ROC methodology plots the fraction of correctly classified positive instances as a function of the fraction of incorrectly classified negative instances; however, in this setting, such technique cannot be applied, as the true labels of the examples are not known. Instead, only a probability of the labels' correctness is known.

Techniques to adapt ROC analysis to probabilistic labels have been proposed in the literature; in this work, we take advantage of the methodology introduced by Burl, et al. in [2]. Let $\boldsymbol{c}$ be the probabilistic output of classifier $\mathbb{C}_1$, $\boldsymbol{c} = \{c_1, \ldots, c_n\}$. Recall that $\boldsymbol{l} = \{l_i\}_{i=1}^{n}$ (where $l_i \in [0,1]$ for all i) is the likelihood of each element in the population of being in the cohort of interest, i.e., for this application, of suffering from cancer. Then, for each decision threshold τ_i of classifier $\mathbb{C}_1$, the following two quantities can be defined:

$$\text{pTPR}(\tau_i) = (\Sigma_{j=1}^{i} l_j)/(\Sigma_{j=1}^{n} l_j) \tag{5}$$

$$\text{pFPR}(\tau_i) = (\Sigma_{j=1}^{i}(1-l_j))/(\Sigma_{j=1}^{n}(1-l_j)) \tag{6}$$

The set of points $(\text{pTPR}(\tau_i), \text{pFPR}(\tau_i))$ for all values of τ_i define a curve in the ROC plane, which we refer to as probabilistic Receiving Operating Curve, or pROC.

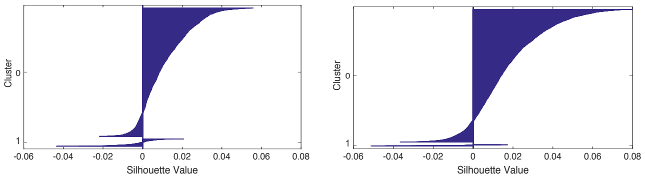
A few observations can be made regarding any pROC curve. First, we point out that, unlike in standard ROC analysis, the maximum AUC achievable by any $\mathbb{C}_1$ is less than one. This is due to the fact that, even in case of perfect classification, the true and false positive rate are bounded by the probabilistic labels in $\boldsymbol{l}$. A corollary is that for any value of the false positive rate, the probabilistic true positive rate of the classifier is a lower bound on the actual true positive rate. This explains why, in Figures 3, the optimal pROC curve—which is obtained when the labels are known with complete accuracy—is described by a curve rather than the segments $[(0,0),(0,1)]$ and $[(0,1),(1,1)]$.

Results of classifier $\mathbb{C}_1$ on the dataset of ovarian cancer and cervical cancer users are reported in Figures 3a and 3b. Each run is evaluated using five-fold stratified cross validation. For both dataset, we present three groups of pROC curves, each one associated with a different value of the learning percentile δ. Each group consists of the optimal classification pROC curve, four curves associated with four values of training percentile θ, and the pROC curve of the baseline classifier.

First, we note that all configuration of the classifier perform substantially better than the baseline. This is to be expected, as the baseline classifier is trained on very few examples. Furthermore, we observe that the baseline classifier for cervical cancer is decisively worse than the baseline classifier for ovarian cancer. We believe that the phenomenon is due to the fact that the number of self-identified cervical cancer users is substantially smaller than the number of self-identified ovarian cancer users. Thus, both a small number of SIUs and the population-level data are needed to correctly identify users.

Third, we note that, for all values of δ, not all classifiers are significantly different from each other (Wilcoxon signed-rank test, $p < 0.05$) with the exception of $\theta = 0.80$. This is a desirable outcome: since the size of users in the cohort of interest is unknown, a classifier that is resistant to small variations of the tuning parameters is beneficial.

Lastly, we study the differences in classification outcomes for fixed values of θ. We notice that, once again, there are no significant differences between $\theta = 0.99$ and $\theta = 0.95$. For the cervical cancer dataset, this is the case for $\theta = 0.90$ as



(a) Ovarian cancer users

(b) Cervical cancer users

Figure 5: Silouette charts for the similarity between SIUs (denoted by "1") and non-SIUs identified as patients (denoted by "0"). Negative values imply similarity between classes and positive values dissimilarity.

	SIUs	non-SIUs	SIUs to non-SIUs
Cervical	0.942	0.918	0.934
Ovarian	0.938	0.926	0.936

Table 4: Average cosine similarities among SIUs, among non-SIUs identified as suffering from the condition of interest, and between SIUs and non-SIUs.

well. However, we observe that, as θ decreases, the performance of the classifier becomes less stable. In particular, the classification outcomes associated with different values of δ significantly differ (Wilcoxon signed-rank test, $p > 0.05$). This event is likely to be caused by that fact that, as δ and γ decrease, users who are not affected by cancer might be part of the learning or training percentile, which naturally decreases the accuracy of the classifier.

5.3 Predicting Disease Distribution

In this section we show how the probabilistic labels computed by the proposed algorithm can be exploited to predict the incidence of diseases in areas for which it is not known. Specifically, we introduce a logistic classifier $\mathbb{C}_2$ that identifies search engine users affected by the disease of interest in states with unknown incidence. The incidence in each region can then be determined by dividing the number of users identified by the number of active search engine users. Thus, we can infer disease incidence in areas where it is unknown, which is an important utility for epidemiologists interested in the spread of a disease.

The procedure to train $\mathbb{C}_2$ is not dissimilar from the one used to train $\mathbb{C}_1$ (Section 5.2). However, unlike $\mathbb{C}_1$, matrices X and Z were combined to train the system.The learning percentile δ and the training percentile θ were set to 0.90 and 0.95, respectively; these value were chosen based on the results described in Section 5.2.1. The classifier is evaluated using the dataset introduced in Section 5.1 under five-fold cross validation.

5.3.1 *Results*

The results of the classifier $\mathbb{C}_2$ on the ovarian and cervical cancer datasets are shown in Figure 4. We report Spearman's rank correlation coefficient ρ_s between the number of users identified by classifier $\mathbb{C}_2$ and disease incidence as reported by the Center of Disease Control as a function of the percentage of users identified as positive by $\mathbb{C}_2$. Before calculating the correlation, counts of users identified by the classifier in each state were normalized by the number of total search engine users in the state.

For both datasets, the classifier is able to obtain a statistically significant correlation (Spearman's rank correlation test, $p < 0.05$) between the normalized number of users identified and the incidence of the disease. $\mathbb{C}_2$ reaches the highest correlation of $\rho_s = 0.35$ when 30% of users are labeled as positive on the ovarian cancer dataset (Figure 4a); similarly, it obtains a correlation of $\rho_s = 0.30$ when 16% of users are label as positive on the ovarian cancer dataset (Figure 4b). We note that the large difference in percentage of positively label users between the two datasets is mostly due to the fact that $\mathcal{X}^{ov}$ and $\mathcal{X}^{cr}$ are of different sizes; in fact, $\mathbb{C}_2$ classifies a similar number of users as positive at the at the point of maximum correlation: 963 for ovarian cancer dataset and 1,483 for the cervical cancer dataset. The smaller difference in terms of individuals classified as positive is more consistent with the US incidence provided by the CDC, which is similar for the two diseases.

We also point out that, for both datasets, correlation follows a similar pattern: when $\mathbb{C}_2$ labels very few users (left side of Figures 4a and 4b) the correlation with CDC data is low and not significant; then, as the number of positively classified users increases the correlation value improves up to reaching statistical significance. However, it declines and looses significance as the number of users classified as positive approaches the size of the population (right side of Figures 4a and 4b).

Finally we note that, while the correlation values are modest, previous research [35] that used only SIUs found a correlation of 0.45 between HIV incidence and number of users. Thus, our correlations are close to those achieved using only users who are known to be suffering from a condition.

6. ADDITIONAL OBSERVATIONS

6.1 Stability of the Algorithm

As the stochastic gradient descent algorithm attempts at separating positive and negative users in $\mathcal{X}$ with hyperplane $\boldsymbol{w}$, it is natural to ask whether the solution it identifies for a given dataset is stable. This is especially the case with the formulation defined in this work, as the objective defined in Algorithm 1 is not convex. To answer this question, we ran Algorithm 1 ten times and measured (i) the rank correlation between the scores of users from any two runs and (ii) the inter-run agreement between all runs. Results show that, for the dataset introduced in Section 4.1, the Spearman's rank correlation between any two runs is at least 0.8 (statistically significant, $p < 0.05$); furthermore, the inter-run agreement is 0.73, which suggests high agreement among runs. Similar results are obtained for the datasets introduced in Section 5.1. This shows that the stochastic gradient descent achieves very similar prediction despite the sampling process in Algorithm 1, lending additional credence to the hypothesis that users identified by the algorithm do indeed share the trait of interest.

6.2 The Similarity of SIUs to Other Patients

Previous work [20, 35] used anonymous self-identified users (SIUs) to identify behaviors associated with other users suffering from the condition of interest. Having noted the poor performance of predicting diseases using only SIUs, in this section we ask whether this performance could be because the behavior of SIUs is not representative of other users suffering from the condition of interest.

We compared all SIUs of a condition (ovarian or cervical cancer) with non-SIUs in the top 10% of users found by the Perceptron run at a 95% threshold (i.e., $\delta = 0.95$).

Table 4 shows the average cosine similarity (computed from X) between users within the two classes and between users of different classes. As the table demonstrates, SIUs are most similar among themselves. Non-SIUs are least similar, and the similarity between non-SIUs and SIUs is in between the two user classes. A different way to observe these similarities is through the silhouette graphs [14] shown in Figure 5. As the graphs show, there are some similarities between groups (SIUs vs. non-SIUs), as demonstrated by the negative values on the charts, but also significant amounts of dissimilarities (positive values on the graphs).

These results imply that SIUs are different from other users identified as suffering from the conditions of interest. This lends additional support to the importance of using the proposed algorithm to identify additional users beyond the small number of self-identified users.

7. CONCLUSION

In this paper, we introduced a novel algorithm for identifying cohorts of interest among internet users. Our approach exploits a small set of users whose membership to the cohort of interest is known (e.g., they self identified themselves) alongside statistics on the entire population. The algorithm was validated on a political dataset of tweets in Section 4. Then, in Section 5, we introduced two applications of the proposed algorithm. First, we discussed a classifier designed to pre-screen for specific forms of cancer using search engine queries. This system could be of great help in detecting diseases that have a set of nonspecific symptoms, no screening test, or may have increased risk if not diagnosed early. However, further research is required to validate whether the sensitivity and specificity of this approach is high enough to be of practical purpose. The second application we investigated dealt with predicting the incidence of a disease in regions in which it is not known. The proposed classifier would be of high value in cases where the incidence of a disease is too low to be measured in a specific region by traditional surveillance methods, or when a disease is spreading within a population. Alternatively, such system could also be helpful in those cases where, for technical reasons, incidence of a disease was not reported.

An important observation stemming from our work is that, when studying anonymous users, SIUs are insufficiently representative of the population. This is both because of the dearth of SIUs, but also, possibly, because there is something unique in the behavior of those users who self-identify. However, SIUs are crucial for identifying the cohort. This observation means that algorithms such as the one proposed herein are needed for the study of anonymous users.

8. ACKNOWLEDGMENTS

We are grateful to Yishay Mansour for his insights and enlightening conversations throughout this work.

9. REFERENCES

[1] P. Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015.

[2] M. C. Burl, U. M. Fayyad, P. Perona, and P. Smyth. Automated analysis of radar imagery of venus: handling lack of ground truth. In *Image Processing*, volume 3, pages 236–240. IEEE, 1994.

[3] S. S. Buys, E. Partridge, A. Black, C. C. Johnson, L. Lamerato, C. Isaacs, D. J. Reding, R. T. Greenlee, L. A. Yokochi, B. Kessel, et al. Effect of screening on ovarian cancer mortality: the prostate, lung, colorectal and ovarian (PLCO) cancer screening randomized controlled trial. *Journal of the American Medical Association*, 305(22):2295–2303, 2011.

[4] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[5] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In *PASSAT*, 2011.

[6] A. Culotta, N. K. Ravi, and J. Cutler. Predicting Twitter user demographics using distant supervision from website traffic data. *JAIR*, 55:389–408, 2016.

[7] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *SIGIR*, pages 595–602, 2008.

[8] K. Dyagilev and E. Yom-Tov. Linguistic factors associated with propagation of political opinions in Twitter. *Social Science Computer Review*, 32(2):195–204, 2014.

[9] Federal Election Commission. Federal Elections 2012. Technical report, jul 2013.

[10] S. R. Flaxman, Y.-X. Wang, and A. J. Smola. Who supported Obama in 2012?: Ecological inference through distribution regression. In *KDD*, 2015.

[11] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[12] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*, 2012.

[13] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW*, 2007.

[14] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[15] G. King. *A solution to the ecological inference problem*. Princeton University Press, 1997.

[16] H. Kuck and N. de Freitas. Learning about individuals from group statistics. *arXiv:1207.1393*, 2012.

[17] D. J. McIver and J. S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Computational Biology*, 10(4):e1003581, 2014.

[18] Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, and E. Yom-Tov. Patterns of information-seeking for cancer on the Internet: an analysis of real world data. *PLoS One*, 2012.

[19] J. Palotti, A. Hanbury, H. Müller, and C. E. Kahn Jr. How users search and what they search for in the medical domain. *Information Retrieval Journal*, 19(1-2):189–224, 2016.

[20] J. Paparrizos, R. W. White, and E. Horvitz. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice*, page JOPR010504, 2016.

[21] Y. Park and J. Ghosh. LUDIA: An aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data. In *KDD*, 2014.

[22] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLoS Currents*, 6, 2014.

[23] M. J. Paul, R. W. White, and E. Horvitz. Search and breast cancer: On episodic shifts of attention over life histories of an illness. *ACM Transactions on the Web (TWEB)*, 10(2):13, 2016.

[24] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using Internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008.

[25] K. L. Priddy and P. E. Keller. *Artificial neural networks: an introduction*. SPIE Press, 2005.

[26] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *JMLR*, 10, 2009.

[27] S. Ravi and Q. Diao. Large scale distributed semi-supervised learning using streaming approximation. *arXiv:1512.01752*, 2015.

[28] F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[29] S. K. Selvaraj, B. Bhar, S. Sellamanickam, and S. Shevade. Semi-supervised SVMs for classification with unknown class proportions and a small labeled dataset. In *CIKM*, 2011.

[30] L. Soldaini, A. Yates, E. Yom-Tov, O. Frieder, and N. Goharian. Enhancing web search in the medical domain via query clarification. *Inf. Retr. Jour.*, 2016.

[31] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*, 21(1):44–51, 2004.

[32] US Cancer Statistics Working Group. United States cancer statistics: 1999–2010 incidence and mortality web-based report. *Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, 2013.

[33] US Cancer Statistics Working Group et al. United States cancer statistics: 1999–2015 incidence and mortality web-based report. *Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute*, 2015.

[34] A. Yates and N. Goharian. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *ECIR*. 2013.

[35] E. Yom-Tov, D. Borsa, A. C. Hayward, R. A. McKendry, and I. J. Cox. Automatic identification of web-based risk markers for health events. *JMIR*, 17(1), 2015.

[36] E. Yom-Tov and E. Gabrilovich. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *JMIR*, 15(6):e124, 2013.

[37] E. Yom-Tov, R. W. White, and E. Horvitz. Seeking insights about cycling mood disorders via anonymized search logs. *JMIR*, 16(2), 2014.