Modeling the Impact of Lifestyle on Health at Scale

Adam Sadilek
Dept. of Computer Science
University of Rochester
Rochester, NY, USA
sadilek@cs.rochester.edu

Henry Kautz
Dept. of Computer Science
University of Rochester
Rochester, NY, USA
kautz@cs.rochester.edu

ABSTRACT

Research in computational epidemiology to date has concentrated on estimating summary statistics of populations and simulated scenarios of disease outbreaks. Detailed studies have been limited to small domains, as scaling the methods involved poses considerable challenges. By contrast, we model the associations of a large collection of social and environmental factors with the health of particular individuals. Instead of relying on surveys, we apply scalable machine learning techniques to noisy data mined from online social media and infer the health state of any given person in an automated way. We show that the learned patterns can be subsequently leveraged in descriptive as well as predictive fine-grained models of human health. Using a unified statistical model, we quantify the impact of social status, exposure to pollution, interpersonal interactions, and other important lifestyle factors on one's health. Our model explains more than 54% of the variance in people's health (as estimated from their online communication), and predicts the future health status of individuals with 91% accuracy. Our methods complement traditional studies in life sciences, as they enable us to perform large-scale and timely measurement, inference, and prediction of previously elusive factors that affect our everyday lives.

Categories and Subject Descriptors

H.1.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Online social networks, machine learning, computational epidemiology, ubiquitous computing, geo-temporal modeling

WSDM'13, February 4–8, 2013, Rome, Italy.

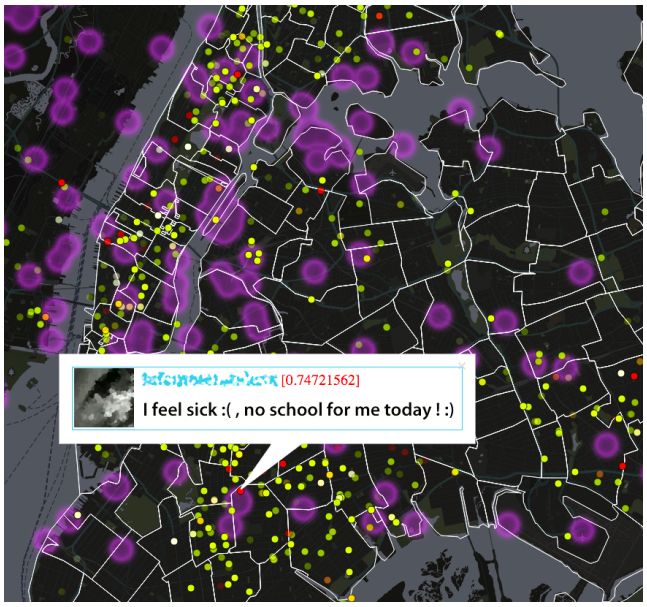

Figure 1: Visualization of the health and location of a sample of Twitter users in New York City. Sick people are colored red, whereas healthy individuals are green. Major pollution sources are highlighted in purple, and ZIP code boundaries are shown with white outlines. This paper explores to what extent online social media can be used to quantify and predict the impact of a large collection of environmental and lifestyle factors on our health. Our web application is available at http://fount.in.

1. INTRODUCTION

How does a new factory affect the health of residents in the city? How does your social status impact your health? Do visits to gyms decrease your susceptibility to communicable diseases? How about visits to bars, or riding the subway? Such questions are traditionally difficult and costly to answer at a population scale. Existing methods resort to surveys of individuals and medical providers, which require extensive amount of human effort to complete, cost large amounts of money, and sample only a small fraction of people in a population. By contrast, we apply machine learning techniques to Twitter data and automatically estimate the health state of any individual on the basis of his or her online communication. Throughout the text, we refer to

the frequency of self-reports of sickness in a user's Twitter updates as user's *health estimate* or simply user's *health* for brevity.

By leveraging the text of geo-tagged tweets, along with the social network structure, we quantify the interplay between a number of important factors and human health. We consider environmental and socioeconomic factors (such as pollution, education, and poverty), as well as social aspects of life (such as encounters and friendships with sick people). Furthermore, we do this at no cost, and without any active user participation. This enables us to operate with a significantly larger number of subjects than previous work in life sciences.

For instance, every thirtieth resident of New York City appears in our dataset.[1] Since Twitter users do not, in general, constitute a representative sample of a population, it is unknown to what extent our results generalize to people who do not participate in online social media. Nonetheless, as we will see, the patterns we find are in agreement with previous epidemiological work. Even if it turns out that the mechanisms we explore here operate in a fundamentally different way within the population at large, our methods still capture a considerable fraction of people. Globally, the prevalence of social media usage is significant, and is increasing: 13% of online adults use Twitter, most of them daily and often via a phone [39].

As a result, this large online population creates a vast "organic" sensor network composed of individuals reporting on their activities, social interactions, and events around them. All of this activity streams in real-time and is often annotated with context including GPS location and images. In this work, we leverage this sensor network to model public health. We capture this important domain in a unified statistical model that measures the impact of various aspects of people's behavior on their health, and allows us to control for a number of confounding factors. Specifically, we show the following:

- Features mined from social media account for up to 54% variance in health of individual Twitter users.
- Physical proximity to pollution sources negatively impacts public health in a measurable way.
- One's online social status has a definite positive association with one's health.
- Encounters and social ties with ill individuals have a large negative impact on one's health.
- The quality of one's neighborhood is associated with one's health.
- One's health state revealed on Twitter can be predicted across individuals with 91% accuracy on the basis of geospatial and relationship data inferred from the Twitter stream.

2. SIGNIFICANCE OF RESULTS

Consider the amount of resources and human effort required to determine how complex social and environmental factors affect the health of millions of people in a large metropolitan area. This paper demonstrates that the expenditure can be, in fact, negligible when we concentrate on users of online social media. We show that fine-grained signals mined from online social media complement traditional coarse-grained offline data, such as census statistics. We view the methods described in this paper as the first step towards real-time public health analysis that enables individuals as well as public officials to make more informed decisions.

[1]More than 19 million people live in the NYC metropolitan area: http://www.census.gov/popest/metro/

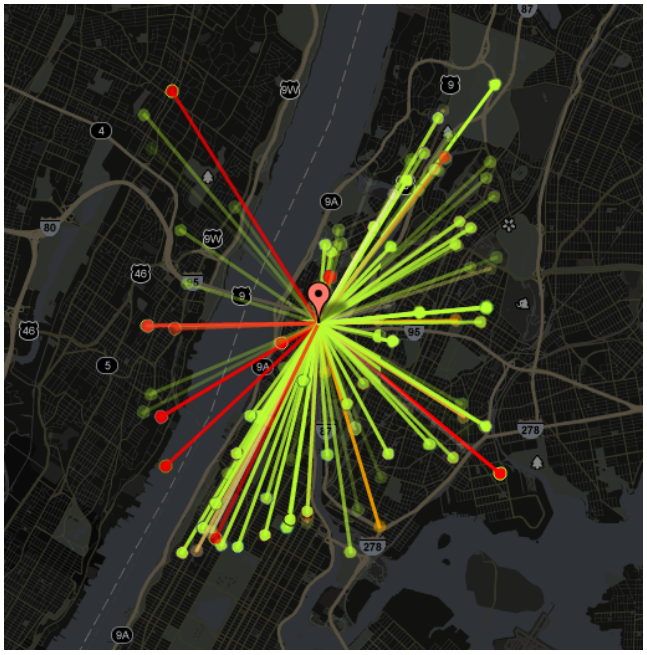

Figure 2: This figure shows the health status of people within a social network of the user u at the center. Each edge represents one of u's friendships, and the color denotes the health of the corresponding friend. Note the patterns in geographical as well as social distribution of sickness. For instance, friends on the east side of the Hudson River tend to be sicker than friends from Manhattan. Is this due to an outbreak of flu in the New Jersey school district? Or is it because the sick individuals lead more stressful lives? In this work, we begin to tease these intimately tied factors apart.

3. MOTIVATION

A large body of research in life sciences studied the effect of a variety of factors on animal and human health. For instance, researchers have established that, among monkeys, immune function can be directly influenced by social status. Experimentally controlled manipulations in social rank were found to lead to widespread changes in gene expression related to immune function. Social rank and immune function were found to be directly correlated, *i.e.*, better social status yields a stronger immune system, as measured by counting anti-bodies in blood samples [42].

Proximity to "green" places, such as urban parks, has been linked to improved resistance to allergies [21] and reduction in stress [40]. Another study showed that even the short-term improvement of pollution levels during Beijing Olympics has a measurable impact on people's health [33].

However, most experiments in this space involved relatively small numbers of individuals—typically less than a

hundred, rarely thousands such as the NHLBI-MESA Study involving 6,500 subjects [5]. Furthermore, because of the amount of human effort required in these controlled studies (physical setup, drawing blood samples, analysis requiring expensive equipment *etc.*), the results and insights come with a significant cost and delay. Additionally, ethical constraints preclude arbitrary controlled experiments with human subjects.

As a result, many important mechanisms remain poorly understood. For example, the link between social status and increased risk of contracting an infectious disease in humans still presents an open question [37]. Parts of the puzzle have been solved. Stress has been found to increase the incidence of common cold, mononucleosis [10], and to activate latent viruses [1]. However, the overall picture remains unclear and incomplete. In this paper, we propose a complementary approach to the traditional small-scale biological and sociological studies. By applying machine learning techniques to large-scale data mined from online social networks, we can quantify the impact of a large collection of social and environmental factors on human health. Since our models are fully unsupervised, they can be updated in real-time as new data and evidence come in.

Let us give a concrete example of how we model the association between social status and health. In this work, we measure a person's social status by network properties, such as PageRank, reciprocity of their relationships, and various centrality measures, as well as by features derived from user interactions, *e.g.*, how many times a person's messages get forwarded or "liked", how many times people mention the person's name. As a proxy for the strength of an immune system of any given person, we use the number of days the person indicates an illness in their online communication. In agreement with [42], we find a strong negative correlation ($R = -0.27$, $p \ll 0.0001$) between people's social status and the frequency with which they get sick (Fig. 4).

Social status is just one of more than sixty factors affecting human health we consider in this paper. These include diverse features ranging from intensity of contact with ill individuals to pollution exposure.

The United States has the world's largest health inequality across society, where the gap of life expectancy of the most and the least advantaged segments of the population is over 20 years [41]. It has been reported that this difference is partly due to differences in social status, but many aspects of the phenomenon remain unexplained [37]. The level of detail and timeliness of the signals we mine from people's online social activities enable us to capture a number of confounding factors that were previously invisible.

For instance, we can explore complex geographical, environmental, and social patterns with fine-granularity and at a large scale as shown in Fig. 2. By fusing this information with public data on pollution sources within a single view, we can begin to quantify the emergent patterns (Fig. 1). This is specific example of our vision of the future of public and personal health management enabled by an automated unification of public and personal data sets. As a result, our work allows governments as well as individuals to model and understand the interplay between health, location, environment, and social factors more effectively.

Before we dive into the details of our approach, we will discuss the limitations that apply to any indirect method of modeling public health.

4. LIMITATIONS

This paper describes our approach to public health modeling that we view as a complement to existing work in epidemiology. Any given dataset carries with it a set of biases and our Twitter dataset is no exception. For example, younger people and minorities are disproportionately present on Twitter as compared to the overall makeup of the population [39]. The results of this study apply directly to the health of Twitter users inferred from their online communication. The health state of each user is captured within a gradient from "sick" to "healthy" with respect to influenza-like illness, but does not further distinguish between specific kinds of ailments. Some users may never self-report (stoics) and others may report being sick when they are not (hypochondriacs). We remove some of this bias by counting the sick *days* for each user, but ground truth is required to control for this effect more explicitly.

Our models are exposed to much noisier data than one finds in a typical survey in life sciences. One component of the noise comes from the text classification process, but evaluation on a held-out set shows that it's small. Additional noise is in the location data and its interpretation. For example, a person tweeting on a sidewalk in front of a bar may be recorded as "visiting" the bar. In most instances, it is next to impossible to determine whether he actually did visit the venue from Twitter data alone. To some extent, we mitigate the rate of false positives by capturing each factor in a number of complementary ways. For example, we also record the distance to the nearest bar and the mean distance to all bars.

While people do not tweet from every location they visit and do not declare every friend online, much of the missing data can be "filled in" by data mining their past behavior [4, 8, 34]. We note that currently used methods suffer from similar confounding effects. For example, infected people who do not visit a doctor, or do not respond to surveys are virtually invisible to the traditional methods. Similarly, efforts such as Google Flu Trends can only observe individuals who search the web for certain types of content when sick. A fully comprehensive coverage of a population will require a combination of diverse methods, and application of AI techniques—like the ones presented in this work—capable of *inferring* the missing information. An important part of our future work, as described at the end of this paper, is to study how estimates based on Twitter users can be adjusted to reflect properties of the general population.

5. DATA

Our experiments are based on data obtained from Twitter, a popular micro-blogging service where people post message updates at most 140 characters long. The forced brevity encourages frequent mobile updates, as we show below. Relationships between users on Twitter are not necessarily symmetric. One can follow (subscribe to receive messages from) a user without being followed back. When users do reciprocate following, we say they are *friends* on Twitter. There is anecdotal evidence that Twitter friendships have a substantial overlap with offline friendships [20]. Twitter launched in 2006 and has been experiencing an explosive growth since then. As of April 2012, over 500 million accounts are registered on Twitter.

Figure 3: A snapshot of Twitter activity overlaid on top New York City public transit map. By merging geo-tagged tweets with the known locations of transit routes, we can model the impact of public transit usage on the health of specific individuals.

Using the Twitter Search API[2], we collected a sample of public tweets that originated from the New York City (NYC) metropolitan area. The collection period was one month long and started on May 19 2010, shortly after U.S. Census 2010 data has been recorded. We periodically queried Twitter for all recent tweets within 100 kilometers of the NYC city center in a distributed fashion. Altogether, we have logged nearly 16 million tweets authored by more than 630 thousand unique users. To put these statistics in context, the entire NYC metropolitan area has an estimated population of 19 million people. Since this work studies the effects environment, location, and co-location have on human health, we concentrate on users that posted more than 100 GPS-tagged tweets during the one-month data collection period. We refer to them as *geo-active users*, and our dataset contains 6,237 such individuals.

5.1 User Privacy

Our research demonstrates that much can be inferred and predicted about specific individuals. This allows users with open profiles to consider the implications of such setting and enables them to make an informed decision about their online behavior. Our methods are useful at an anonymized level as well, as they can extract aggregate information from individuals for the benefit of others. For example, we show that one's social status is significantly tied with one's health. Government officials as well as public advocacy groups can use such insights to make a stronger case for a policy change. Personalized results can be reported directly to the authenticated user.

We recognize that there are substantial privacy questions ahead. We believe the issues ultimately reduce to a cost-benefit analysis. Specifically, by quantifying the trade-offs between the *value* our automated systems create versus loss of user privacy. In the future, we envision each person will be able to set—either explicitly or implicitly via an auction scheme—a dollar valuation on his or her privacy, and online services will take that preference into account when collecting, analyzing, and using customer data. In one extreme, one may decide to risk sharing all data, make money on it, and get more personalized services. On the other hand, one may set tight privacy filters and pay for online services. This may lead to a new open marketplace with public data [3].

[2]http://search.twitter.com/api/

6. BACKGROUND

Support vector machine (SVM) is an established model of data in machine learning [12]. We learn an SVM for linear binary classification to accurately distinguish between tweets indicating the author is afflicted by an ailment and all other tweets.

Linear binary SVMs are trained by finding a hyperplane defined by a normal vector w with the maximal margin separating it from the positive and negative datapoints. Finding such a hyperplane is inherently a quadratic optimization problem given by the following objective function

$$\min_w \frac{\lambda}{2}||w||^2 + \mathcal{L}(w, D), \tag{1}$$

where λ is a regularization parameter controlling model complexity, and $\mathcal{L}(w, D)$ is the hinge-loss over all training data D given by

$$\mathcal{L}(w, D) = \sum_i \max\left(0, 1 - y_i w^T x_i\right). \tag{2}$$

The optimization problem in Equation 1 can be solved efficiently and in a parallel fashion using stochastic gradient descend methods [38].

Class imbalance, where the number of examples in one class is dramatically larger than in the other class, complicates virtually all machine learning. For SVMs, prior work has shown that transforming the optimization problem from the space of individual datapoints $\langle x_i, y_i \rangle$ in matrix D to one over *pairs* of examples $\langle x_i^+ - x_j^-, 1 \rangle$ yields significantly more robust results [23]. (x_i^+ denotes feature vectors from the positive class ($y_i = +1$), whereas x_j^- denotes negatively labeled data points ($y_j = -1$).) This method is often referred to as ROC Area SVM because it directly maximizes the area under the ROC curve of the model.

Measures of centrality mathematically capture the "importance" of a node in a (social) network. Some measures, such as degree centrality, are local and depend only on the immediate neighborhood of a node. Other methods, including PageRank and betweenness centrality, capture the global properties of the network as well. For an excellent general overview of computational analysis of social networks at large see [15].

Regression analysis is a statistical technique of quantifying the relationship between one or more independent variables and a dependent response variable. In this work, we apply regularized least-squares regression model with elastic net algorithm [45]. This formalism encourages grouping of strongly correlated independent variables, and enables variable selection in a principled way.

Decision trees are models of data encoded as rules induced from examples [6]. Intuitively, in our domain, a decision tree represents a series of questions that need to be asked and answered in order to estimate the health quality of a person, based on his or her attributes and contextual features. During decision tree learning, features are evaluated in terms of information gain with respect to the labels and the best candidates are subsequently selected for each inner node of the tree. Our implementation uses *regression* decision trees, where each leaf contains a continuous label.

As described below, we also employ decision trees for feature selection, since they intrinsically rank features by their information content.

7. METHODS

This section presents the methods we use to quantify and predict people's health from their location and activities recorded online. In short, we begin by inferring the health state of any given Twitter user on the basis of the content of his or her online communication. We then data mine a large collection of features for each individual. The features jointly describe the context of people's lives in terms of location, their environment, and social activities. We subsequently capture the associations between the contextual features and people's health via statistical analysis. Finally, we explore the *predictability* of health from the induced factors. The following subsections describe each of these steps in detail.

7.1 Health State Inference

We build upon previous work on classification of short text messages [13, 31, 35] and learn a support vector machine classifier C that accurately identifies tweets that indicate their author is ill. C is trained by directly optimizing the area under the ROC curve, as is therefore robust even in the presence of strong class imbalance, where for every health-related message there are more than 1,000 irrelevant ones. We use C to distinguish between tweets indicating the author is afflicted by an ailment (we call such tweets "sick"), and all other tweets (called "other" or "normal").

As SVM features, we use all unigram, bigram, and trigram word tokens that appear in the training data. For example, a tweet *"I feel sick."* is represented by the following feature vector:

$$\big(i, \textit{feel}, \textit{sick}, i\ \textit{feel}, \textit{feel sick}, i\ \textit{feel sick}\big).$$

Overall, our SVM operates in more than 1.7 million dimensions, where each dimension represents a word or a phrase extracted from training data. Before tokenization, we convert all text to lower case, strip punctuation and special characters, and remove mentions of user names (the "@" tag) and re-tweets (analogous to email forwarding). However, we do keep hashtags (such as "#sick"), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages. Table 1 lists examples of significant features found in the process of learning C.

We use the SVM cascade learning procedure described in [35]. Evaluation of C on a held-out set shows 0.98 precision and 0.97 recall with respect to labels agreed upon by human annotators. Ground truth for each tweet was obtained by asking five Amazon Mechanical Turk workers to label the tweet as either "sick" or "other" and subsequently extracting the majority vote.

7.2 Modeling Associations: Environment, Lifestyle, and Health

From the online activities of each geo-active user, we mine a number of features that describe the context of his or her life in terms of location, environment, and social interaction. We describe the **health** of individuals using two random variables: the expected number of sick days, and the normalized cumulative probability mass of sickness given by

$$P_S = \frac{1}{|M|} \sum_{t \in M} \Pr\big[t \text{ is sick}\big],$$

where M is the set of tweets of a given user. A sick day is defined as a calendar day during which the user wrote at least one "sick" tweet. Throughout the paper, we will use the term *health quality* to denote $-P_S$.

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	−0.4005
headache	0.5249	you	−0.3662
flu	0.5051	lol	−0.3017
fever	0.3879	love	−0.1753
feel	0.3451	i feel your	−0.1416
coughing	0.2917	so sick of	−0.0887
being sick	0.1919	bieber fever	−0.1026
better	0.1988	smoking	−0.0980
being	0.1943	i'm sick of	−0.0894
stomach	0.1703	pressure	−0.0837
and my	0.1687	massage	−0.0726
infection	0.1686	i love	−0.0719
morning	0.1647	pregnant	−0.0639

Table 1: Examples of positively and negatively weighted significant features of our SVM model C.

We quantify one's **social status** with PageRank, reciprocity of following, the number of times other people mention the user, and a collection of centrality measures (degree, communicability, eigenvector, betweenness, load, flow, and closeness). The *reciprocity* of friendships between high school students has been shown to be a strong predictor of social status of individuals [30]. Namely, low-rank individuals frequently (and wishfully) list highly ranked students as their friends, but the relationship is not reciprocated. Most stable and mutual friendships occur between people of the same rank. We find that all measures of social rank are highly cross-correlated ($R > 0.73, p \ll 0.0001$) and have almost identical predictive power, explaining over 24% of the variance in health.

Leveraging the estimated health state of all individuals in our dataset along with their GPS location, we extract the number of **physical encounters** with sick people. The noise in GPS signal in areas with tall buildings throughout New York City can be up to hundred meters. Therefore, we consider two individuals co-located if they appear within hundred meters of each other within a time window (slack) of length T. While this method is likely to overestimate the number of actual encounters with other Twitter users, we find that it serves as a good proxy for the actual level of exposure to infected individuals. In this work, we consider time windows of lengths 1, 4, and 24 hours. Based on declared Twitter friendships (*i.e.*, mutual following of geo-active users), we count the number of **sick friends** each user has.

Using Google Places API,[3] we download GPS coordinates of all bars, night clubs, transit stations, public parks, and gyms in the greater New York City area. Transit stations include major bus stops, subway stations, and train, ferry

[3]https://developers.google.com/places/documentation/

and airport terminals. We consider a total of 25 thousand venues in these categories (see Fig. 3). Using the geo-tags of individual tweets, we calculate the mean shortest distance to each venue type and the number of *visits* to each venue. A visit is measured as tweeting within 100 meters of a venue. These metrics capture aspects of the **behavior** and **lifestyle** of each person. For instance: How often do they frequent bars versus gyms? How much time they spend in crowded public transportation?

Another type of "venue" we consider are major **pollution** sources obtained from the U.S. Environmental Protection Agency.[4] These include factories, power plants, transportation hubs, and other sites emitting significant amounts of volatile chemical compounds, particulate matter, CO, NO_x, SO_2, and other harmful chemicals. We model over 1,700 pollution sites within NYC (see Fig. 1).

We would like to tie in datasets that do not have GPS annotations. For instance the U.S. census contains rich **socioeconomic characteristics** of the entire nation. We tie our geo-tagged tweets with census data by inferring each person's home ZIP code, which then serves as a key into the census dataset. The GPS coordinates of person x home is estimated by fitting a two-dimensional Gaussian to all x's locations between 1am and 6am. The mean of this Gaussian is taken as the most likely home location. We then look up the corresponding ZIP code in the GeoNames.org database.

The average New York City ZIP code zone has an area of 3.6 km^2 and can be walked across in less than 20 minutes. The ZIP code areas are shown in Fig. 1. We can now associate each person with the context derived from the 2010 census—the most recent census available.[5] We focus on three broad characteristics of a person's neighborhood: poverty, education, and race. Poverty is measured in terms of fraction families and individuals below poverty line, the number of abandoned housing units, and the prevalence of social security dependence. Education captures proportion of people over 25 with various levels of education (from elementary school to a doctorate). The race factor includes proportion of different races and ethnic groups.

For each person, we induce 62 features based on the factors described in this section. In Section 8, we will see how these features correlate and associate with people's health, but let us first describe how we use them for health *prediction*.

7.3 Health Prediction

We are interested in how do the features we discuss above generalize across individuals. We explore this by learning a regression decision tree on a training set of subjects and evaluate on the remaining test set. The decision tree additionally reveals the relative importance of the individual features—with the most informative feature at the root and increasingly insignificant features towards the leaves. We prevent overfitting by pruning the tree on the basis of 10-fold cross-validation.

8. EXPERIMENTS AND RESULTS

This section reports the experimental results we obtained and closely follows the structure of the Methods section above. Unless otherwise noted, all results reported are statistically significant at the 0.001 level ($p < 0.001$).

[4]http://www.epa.gov/air/emissions/

[5]http://www.census.gov/main/www/access.html

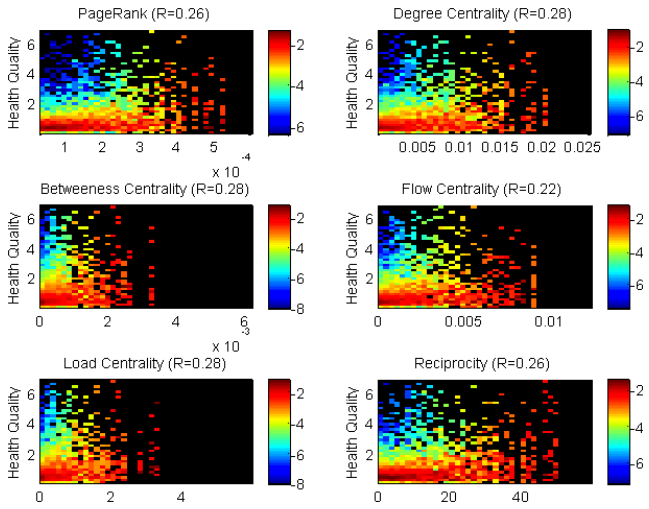

Figure 4: This figure shows the interplay between people's social status and their health. Each pane shows a different measure of social status (plotted on the horizontal axes). The vertical axes show the expected amount of time a person is healthy. A column x in each figure shows the log-probability distribution $\log(\Pr[\text{health}|\text{centrality} = x])$. We see that the people who often get sick (bottom rows in each pane) are more likely of low social status, whereas people with high social rank enjoy better health. The correlation coefficient R for each measure of social status is shown on top. The positive association between rank and health is consistent across all measures of centrality considered.

Fig. 4 shows the interplay between six different measures of social rank and people's health. We see a common pattern independent of the particular centrality measure: the higher the social status, the better the health. While low-rank individuals are concentrated on the "frequently sick" side of the spectrum (on the bottom of each pane), highly ranked subjects (on the right side of each pane) are more uniformly distributed and often attain high health scores.

Applying regression analysis, we quantify the associations between our factors and health (Fig. 5). All measures of social rank are mutually strongly correlated and positively associated with health quality. Proximity to pollution sources is the single most correlated feature with P_s (closer proximity $\Leftrightarrow$ worse health). Visit to polluted sites are also strongly associated with health (more visits $\Leftrightarrow$ worse health).

In agreement with prior work [21], we found a small but significant positive correlation between visits of public parks and health quality.

Fig. 6 shows the fraction of variance in cumulative probability of sickness (P_S) explained by various subsets of our factors (across all subjects).

We see that census data exhibits small correlation scores and explains little variance in health. While the individual contributions of poverty, education, and race are small, they jointly account for 8.7% of variance that is unexplained by other factors. The actual effect of these factors is likely to be even higher, however. In New York City, the ZIP code areas are relatively large compared to the diversity and den-

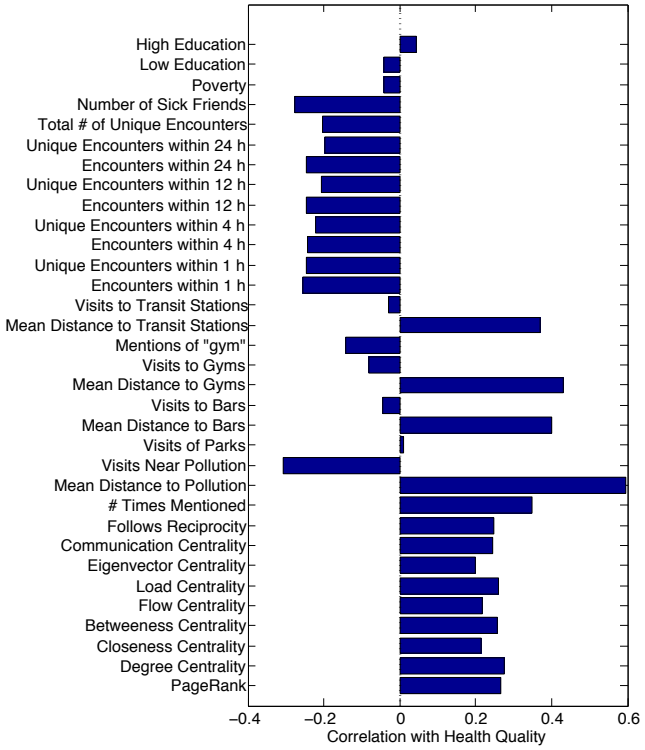

Figure 5: Results of regularized regression analysis with the negative cumulative probability of sickness ($-P_S$) as the dependent variable. Proximity to polluted sites and encounters with sick individuals are negatively correlated with people's health quality, whereas high centrality corresponds with good health. Note the positive association between social rank and health quality, irrespective of the measure of social status. Features derived from census data (education and poverty) have small but significant effect.

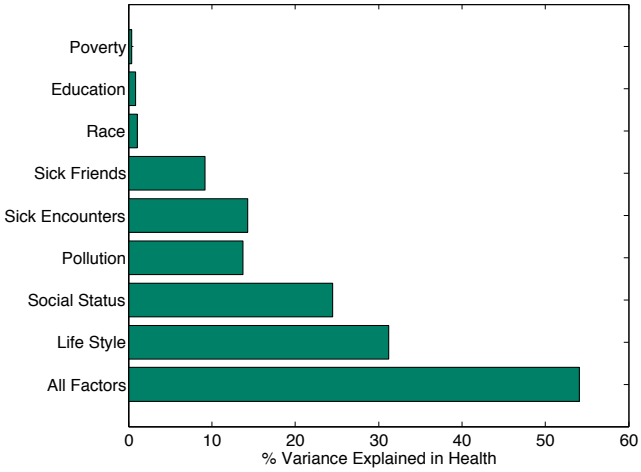

Figure 6: Fraction of variance in health ($-P_S$) explained by various subsets of our factors. All features jointly account for more than 54% of the total variance. The first three factors (Poverty, Education, and Race) are derived from census data for user's home ZIP code. See Section 7.2 for explanation of the composition of the factors.

sity of the city population. As a result, a single ZIP code often contains people on both extremes of any given factor considered in this study. This suggests that the fine-grained data available in online social media does open novel opportunities for increasingly comprehensive models of public health.

From Figures 5 and 6, we see that subjects' activities regarding bars, gyms, and public transit all appear to have a common association with health: avoiding those venues is connected to better health. For gyms, the effect is mildly stronger (in a statistically significant way) as compared to bars and public transit. Interestingly, mentioning the word "gym" in online communication is also associated with worse health. Future work will explore possible confounders and shed more light on the interplay between these important lifestyle factors.

For health prediction, we use 80% randomly selected subjects to induce a regression decision tree D_1 that predicts the expected number of sick days (Fig. 7), and D_2 that predicts the cumulative probability of sickness P_S. Evaluation on the remaining 20% of the subjects shows that D_1 achieves 91% accuracy, and D_2 is within 8% of the actual P_S value more than 95% of the time. We see that the decision tree induced meaningful sequences of features, and it will be interesting to see if such insights can guide the design of a *controlled* epidemiological study that explores the effects further.

We consider two baselines: random and most-frequent. The former draws the predicted number of sick days from a Gaussian distribution learned from training data, whereas the most-frequent baseline always outputs the mode of the labels in training data. They achieve 15% and 64% accuracy, respectively, when predicting the number of individuals' sick days.

9. RELATED WORK

Since the famous cholera study by John Snow (1855), much work has been done in capturing the mechanisms of epidemics. There is ample previous work in computational epidemiology on building relatively coarse-grained models of disease spread via differential equations and graph theory [2, 29], by harnessing simulated populations [16], and by analysis of official statistics [19]. Such models are typically developed for the purposes of assessing the impact a particular combination of an outbreak and a containment strategy would have on humanity or ecology [7]. However, the above works focus on simulated populations and hypothetical scenarios. By contrast, we address the problem of predicting the health of *real-world* populations composed of individuals embedded in a fine social structure. As a result, our work makes a step towards understanding the impact of complex intertwined factors affecting our health.

Traditionally, public health is monitored via surveys and by aggregating statistics obtained from healthcare providers. Such methods are costly, slow, and may be biased. Recently, digital media has been successfully used to significantly reduce the latency and improve the overall effectiveness of public health monitoring. Perhaps most notably, Google Flu Trends models the prevalence of flu via analysis of geolocated search queries[18].

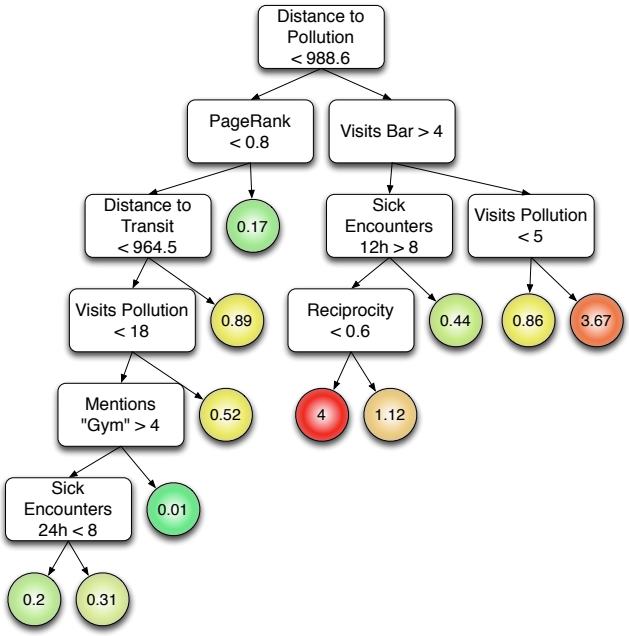

Figure 7: Decision tree predicting the expected number of sick days. The tree has been pruned to an optimal level using 10-fold cross-validation. To estimate the number of sick days for a new individual, we start at the root of the tree and evaluate the inequality in the root. If it evaluates to true, we traverse the left subtree, otherwise we recurse on the right subtree. We see, for example, that highly ranked subjects who avoid pollution sources tend to be healthier than people who frequent bars, encounter too many sick individuals, and have unreciprocated relationships.

Eubank *et al.* began to leverage more fine-grained information, including people's activities [16]. They developed a simulation tool (EpiSims) that leverages synthetic—but statistically realistic—human mobility to study the spread of infectious diseases over a metropolitan area. We see an important continuity from such simulations to real-time tracking of human behavior. For example, as we have seen, our NYC dataset contains 1/30 of the residents. We can simulate the remainder of the population, while "seeding" the system with realistic parameters *learned* from live data. For example, people's home locations or social encounters are no longer drawn from coarse distributions, but rather inferred from online activities.

In the context of social media, [25] explore augmenting the traditional notification channels about a disease outbreak with data extracted from Twitter. By manually examining a large number of tweets, they show that self-reported symptoms are the most reliable signal in detecting if a tweet is relevant to an outbreak or not. This is because people often do not know what their true problem is until diagnosed by an expert, but they can readily write about how they feel. Researchers have also concentrated on capturing the overall *trend* of a particular disease outbreak, typically influenza, by monitoring social media [13, 26, 9]. [17] use information actively submitted by cell phone users to model aggregate public health. However, scaling such systems poses considerable challenges.

Other researchers focus on a more detailed modeling of the *language* of the tweets and its relevance to public health in general [31], and to influenza surveillance in particular [11]. Paul *et al.* develop a variant of topic models that captures the symptoms and possible treatments for ailments, such traumatic injuries and allergies, that people discuss on Twitter. In a follow-up work [32] begin to consider the geographical patterns in the prevalence of such ailments, and show a good agreement of their models with official statistics and Google Flu Trends.

Our previous work has shown that people's interactions recorded in online social networks can be used to learn very specific and fine-grained models of the spread of contagious disease [35, 36]. However, prior work considered only rudimentary features based on immediate user co-location and social ties. In this paper, we include more than forty additional factors based not only on interpersonal interactions, but also on the environment, quality of one's neighborhood, socioeconomic status, and lifestyle. Furthermore, we show that the factors we induce generalize across individuals and have a strong predictive power within our population of Twitter users.

In life sciences, the positive correlation between social rank and strength of an immune system has been found in rhesus monkeys by performing controlled experiments on a small number of subjects [42]. In human studies, social status has been estimated indirectly. Socioeconomic status (SES) has been widely used as a proxy for people's social rank [44]. A strong correlation between SES and health has been documented in a large number of contexts, in countries with egalitarian and socialized medical care, and even for diseases whose outcomes are largely unaffected by the quality of health care. Interestingly, even when controlling for risk factors often associated with low SES, such as smoking and unhealthy diet, SES is still the dominant factor. Subsequent studies have shown that *subjective* SES is a better predictor of a person's health than a global measure SES.

For example, in the US, the higher the income inequality in a given region, the worse the health of the population there [24] (and also the higher the prevalence of firearm ownership [22]). Position in a civil servant hierarchy is another potential proxy for social status. A pioneering study of [28] has found that the risk of heart disease grows as one's occupational rank decreases. In this paper, we use the declared social network of Twitter users and their online interactions to determine the social rank of any given individual.

While we do not yet have a comprehensive understanding of the biological and psychological mechanisms that link social status to disease susceptibility, partial explanations have been proposed. Low social rank often leads to less control over one's life, imbalance between work effort and reward, lack of autonomy, less respect from the general society, fewer means and options available to resolve difficult situations, *etc.* [27]. This renders low-rank individuals more vulnerable to stress, in the developed world mostly *psychological* stress. The negative impact of stress on health is well documented. Stress induces metabolic and endocrine changes that in turn lead to increased risk of disease [37].

However no studies to date have resolved the impact of people's social status on the prevalence of diseases, or the spread of infectious diseases, throughout a large-scale pop-

ulation over extended periods of time. In this paper, we demonstrate that large-scale data mining enables us to fill some of these gaps without any active user participation. As a result, we can improve our understanding of human behavior and begin to quantify—in a scalable fashion—important phenomena affecting our everyday lives. At the same time, our approach complements, but does not replace, controlled longitudinal studies (*e.g.*, [28, 14, 37]) that uncover the detailed biological mechanism behind diseases and capture signal that is, at present, too weak to be detectable online.

10. CONCLUSIONS AND FUTURE WORK

This paper focuses on data mining diverse, noisy, and incomplete sensory data over large numbers of individuals. We show that the induced patterns can be subsequently leveraged in descriptive as well as predictive models of the health of a population of Twitter users at scale. We find that the raw sensory data linked with the content of users' online communication, explicit as well as implicit online social interactions, and relationships are extremely rich information sources. The fine granularity and pervasiveness of the data allows us to model phenomena that have been previously out of reach. We consider environmental factors (such as pollution and poverty) as well as social aspects of life (such as encounters and friendships with sick and healthy people). Furthermore, we do this at no cost, and without any active user participation.

Our methods enable us to shed additional light on important questions in public health that have been either too expensive or outright impossible to answer. In the process, we have drawn parallels to work done in other scientific fields, including epidemiology, immunology, and sociology, and shown how our methods complement previous results and bring new insights.

Our current work concentrates on two areas: effective ways of validating disease models with ground truth about people's health, and expanding the set of environmental factors our models capture. The first area builds on research in traditional epidemiology and active learning, and will enable us to learn more *specific* health models that capture the "offline" part of the population as well.

Consider the following pyramid model of public health. On the base of the pyramid, we have the entire population. In the middle of the pyramid are users of online social media who we can access. At the top of the pyramid is a small—but strategically selected—sample of individuals from the general population (which includes some of the social media users) for whom we have detailed records about their health. These include subjects who respond to online medical surveys, take at-home rapid tests, or even get tested at a nearby medical lab and share the results.

Traditionally, epidemiological studies are based on data collected from the top of the pyramid. This paper addresses the middle of the pyramid. We believe a *hybrid* approach will enable knowledge gained at any level in the pyramid to "trickle down". For example, by applying automated machine learning techniques described in this paper to the large mass of people in online social media, we can bootstrap the top of the pyramid to make well-founded predictions about the general population at the bottom of the pyramid. This will infuse epidemiological models with additional structure and parameters learned from detailed timely data, so that fewer factors need to be modeled via simulation. However, information could also "trickle up", where the latent behavior of the hidden population influences predictions even for individuals on top.

The second area focuses on modeling the interplay between a wider range of personal, social, environmental and health factors. For example: What exact impact eating habits have on one's health, and the health of related people? Animal studies have shown that *personality* plays a significant role in the perception of one's social rank, which in turn modulates the impact of the social status on one's health [43]. Similar processes are likely to occur in human societies as well. For instance, a phlegmatic person may be less affected by his or her low rank. However, these mechanisms prove difficult to capture using current techniques. By contrast, our work builds a foundation that makes such studies possible. For example, the language model used here to infer the health state of an individual can be augmented to estimate the personality type of a person. Finally, our models enable mobile applications that inform the user about health risks around him in real-time, opening opportunities for complex social studies of human behavior at scale.

11. ACKNOWLEDGMENTS

Figures 1-3 in this paper are taken from our web application developed with Sean Brennan, Martin Janda, Andrew Abumoussa, and Hao Chen (http://fount.in). This research was partly funded by ARO grant W911NF-08-1-0242, ONR grant N00014-11-10417, OSD grant W81XWH-08-C07-40, and a gift from the Kodak Company.

12. REFERENCES

[1] R. Ader and N. Cohen. Conditioning and immunity. *Psychoneuroimmunology*, 2:3–34, 2001.

[2] R. Anderson and R. May. Population biology of infectious diseases: Part I. *Nature*, 280(5721):361, 1979.

[3] C. Aperjis and B. Huberman. A market for unbiased private data: Paying individuals according to their privacy attitudes. *Arxiv.org*, 2012.

[4] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM 2011*, pages 635–644. ACM, 2011.

[5] D. Bild, D. Bluemke, G. Burke, R. Detrano, A. Diez Roux, A. Folsom, P. Greenland, R. Kronmal, K. Liu, J. Nelson, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*, 156(9):871–881, 2002.

[6] L. Breiman et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[7] P. Chen, M. David, and D. Kempe. Better vaccination strategies for better people. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 179–188. ACM, 2010.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[9] R. Chunara, J. Andrews, and J. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.

[10] S. Cohen, D. Tyrrell, and A. Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9):606–612, 1991.

[11] N. Collier, N. Son, and N. Nguyen. OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2011.

[12] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[13] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, pages 115–122. ACM, 2010.

[14] F. Destefano, E. Eaker, S. Broste, D. Nordstrom, P. Peissig, R. Vierkant, K. Konitzer, R. Gruber, and P. Layde. Epidemiologic research in an integrated regional medical care system: the marshfield epidemiologic study area. *Journal of clinical epidemiology*, 49(6):643–652, 1996.

[15] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[16] S. Eubank, H. Guclu, V. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[17] C. Freifeld, R. Chunara, S. Mekaru, E. Chan, T. Kass-Hout, A. Iacucci, and J. Brownstein. Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS medicine*, 7(12):e1000376, 2010.

[18] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[19] B. Grenfell, O. Bjornstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[20] A. Gruzd, B. Wellman, and Y. Takhteyev. Imagining Twitter as an imagined community. In *American Behavioral Scientist, Special issue on Imagined Communities*, 2011.

[21] I. Hanski, L. von Hertzen, N. Fyhrquist, K. Koskinen, K. Torppa, T. Laatikainen, P. Karisola, P. Auvinen, L. Paulin, M. Mäkelä, et al. Environmental biodiversity, human microbiota, and allergy are interrelated. *Proceedings of the National Academy of Sciences*, 2012.

[22] D. Hemenway, B. Kennedy, I. Kawachi, and R. Putnam. Firearm prevalence and social capital. *Annals of Epidemiology*, 11(7):484–490, 2001.

[23] T. Joachims. A support vector method for multivariate performance measures. In *ICML 2005*, pages 377–384. ACM, 2005.

[24] I. Kawachi. The health of nations: Why inequality is harmful to your health author: Ichiro kawachi, bruce p. kennedy, publisher: New p. 2006.

[25] M. Krieck, J. Dreesman, L. Otrusina, and K. Denecke. A new age of public health: Identifying disease outbreaks by analyzing tweets. *Proceedings of Health WebScience Workshop, ACM Web Science Conference*, 2011.

[26] V. Lampos, T. De Bie, and N. Cristianini. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases*, pages 599–602, 2010.

[27] M. Marmot. Status syndrome. *JAMA: the journal of the American Medical Association*, 295(11):1304–1307, 2006.

[28] M. Marmot, G. Rose, M. Shipley, and P. Hamilton. Employment grade and coronary heart disease in british civil servants. *Journal of Epidemiology and Community Health*, 32(4):244–249, 1978.

[29] M. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

[30] M. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[31] M. Paul and M. Dredze. A model for mining public health topics from Twitter. *Technical Report. Johns Hopkins University. 2011.*, 2011.

[32] M. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[33] D. Q. Rich, H. M. Kipen, W. Huang, G. Wang, Y. Wang, P. Zhu, P. Ohman-Strickland, M. Hu, C. Philipp, S. R. Diehl, S.-E. Lu, J. Tong, J. Gong, D. Thomas, T. Zhu, and J. J. Zhang. Association between changes in air pollution levels during the beijing olympics and biomarkers of inflammation and thrombosis in healthy young adultsair pollution, inflammation, and thrombosis. *The Journal of the American Medical Association*, 307(19):2068–2078, 2012.

[34] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Fifth ACM International Conference on Web Search and Data Mining*, 2012. (Best Paper Award).

[35] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[36] A. Sadilek, H. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[37] R. Sapolsky. Social status and health in humans and other animals. *Annual Review of Anthropology*, pages 393–418, 2004.

[38] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

[39] A. Smith. Pew internet & american life project. *http://pewresearch.org/pubs/2007/twitter-users-cell-phone-2011-demographics*, 2011.

[40] C. W. Thompson, J. Roe, P. Aspinall, R. Mitchell, A. Clow, and D. Miller. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. *Landscape and Urban Planning*, 105(3):221 – 229, 2012.

[41] B. I. Truman et al. CDC health disparities and inequalities report. *Morbidity and Mortality Weekly Report*, 2011.

[42] J. Tung, L. Barreiro, Z. Johnson, K. Hansen, V. Michopoulos, D. Toufexis, K. Michelini, M. Wilson, and Y. Gilad. Social environment is associated with gene regulatory variation in the rhesus macaque immune system. *Proceedings of the National Academy of Sciences*, 2012.

[43] C. Virgin Jr and R. Sapolsky. Styles of male social behavior and their endocrine correlates among low-ranking baboons. *American Journal of Primatology*, 42(1):25–39, 1997.

[44] R. Wilkinson. *Mind the gap*. Weidenfeld & Nicolson, 2000.

[45] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.