Social networks and decision-making: food and influence in the age of Twitter

Nicole A. Pangborn
Darwin College

UNIVERSITY OF CAMBRIDGE

This dissertation is submitted to the University of Cambridge in partial fulfillment of the requirements for the degree of Master of Philosophy in Applied Biological Anthropology

University of Cambridge
Department of Archaeology and Anthropology
Division of Biological Anthropology
Pembroke Street
Cambridge, CB2 3DZ

email: np378@cam.ac.uk

15 July, 2012

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

It does not exceed the 20,000 word limit of the degree committee.

Acknowledgements

First, I would like to thank my supervisor, **Dr. Peter D. Walsh**, for allowing me to pursue my interests and for continued advice and support throughout the process.

Next, I would like to acknowledge **Fabio Lahr**, whose willingness to help with the technical aspects of this project went above and beyond any of my expectations. I sincerely thank him for his efforts.

I would also like to thank **Topsy Labs, Inc.** for the use of their exploratory tools and for their understanding during the writing process.

Finally, my peer editors and friends **Robert Attorri**, **Alison Macintosh**, **Tessa Stewart**, **Borja Moreno Fernández**, **Frederica Stahl**, and **Alice Durieux** have spent hours of their time giving me feedback and support for this project. I very much appreciate their generosity.

Abstract

BACKGROUND: The "Web 2.0" of today presents a highly interactive version of the original Internet. An example of this newly interactive setting comes from the popular microblogging service Twitter, which provides short, real-time updates of user-generated content, called "tweets". Twitter makes the opinions and stated behaviors of users accessible to researchers on a wide scale. One behavior of current interest, especially in the United States, involves eating patterns, as rates of metabolic disease continue to rise. Obesity rates in particular display interesting geographical patterns, and certain regions are more obese than others. Could we use the information available in tweets from US cities to describe and predict real-world trends in obesity and food discussion?

METHODS: Queries to Twitter's Streaming API were made to obtain 38,039,682 unique food-related tweets from the continental United States over a week long period in May 2012. Simple Python scripts processed the text output to reveal relative frequencies of target words (associated with obesogenic environments, such as fast food or hunger mentions) from Twitter's most popular U.S. cities. These frequencies were compared to obesity rates of the cities as reported in surveys administered by the Centers for Disease Control and Prevention. Differences in word frequencies and obesity rates were also compared to differences in physical distance to better understand the role of geography in obesity culture.

RESULTS: Significant positive correlations were found between regional obesity rates and relative mentions of McDonald's ($r = .648$, $p < .001$) and hunger ($r = .572$, $p < .01$) on Twitter, with negative correlations found between regional obesity rate and relative mentions of 'healthy' ($r = -.641$, $p < .001$). Mean relative food frequencies were significantly different by region, with the Western United States clearly distinguishable in all cases. Word frequencies were sufficient to classify an unidentified set of tweets by region in 63.8% of all cases, and a reasonably accurate model for prediction of obesity rate based on tweets was established ($F(3, 43) = 16.7$, $p < .001$). Finally, plots are presented which describe increasing differences in food culture as a function of physical distance between cities. Differences appear to increase in cities between 1000-3000 kilometers apart and are consistent with a model of food idea "transmission" which becomes slower at longer distances from either coast.

CONCLUSIONS: Overall, this study has shown that the ways in which people in the United States discuss food on Twitter are consistent with corresponding real-world patterns in aspects of both "obesogenic" environment and regional metabolic disease. Results presented here point to Twitter's potential usefulness as a more widespread tool in the public health sector.

Contents

Chapter 1

Introduction

With the birth of the Internet came extraordinary new means of human communication. Its near-instant emailing services and information-gathering powers were previously unparalleled [1]. Despite these critical advances, in its early years, the Web was still relatively simple. It was a network of networks used by many, but modified by few. Web pages were static to visitors, and the majority of users hopped between them only to passively gather information. By the early 2000s, however, we began to see an enormous shift in the 'feel' of the Internet: websites began focusing on active user participation rather than simple content absorption. User comments, social networks, blogs, wikis, and other cooperative content became commonplace, and this significant transformation led to the very natural coinage of the term 'Web 2.0' in referring to the Internet of today [2].

Since the way we co-create information on the Web today is completely new, the set of facts and opinions at our disposal to help us make decisions has been altered. Classic models of decision-making and social influence must be revised in light of this interactive Internet: we are now exposed to the opinions of an exponentially larger group of individuals, and this new type of communication is bound to both exhibit and affect corresponding real-world behavior.

Perhaps the most fruitful way to use the mounds of communication on record in the depths of the social web is to track the real-world behavioral changes that matter. One such change is displayed in eating patterns, as these are directly related to current trends in metabolic disease — generally considered to hold an 'epidemic' status (especially in Western nations). As paper surveys of eating patterns have been notoriously unreliable [3], it would be particularly useful if we had a widespread, real-time personal 'updater' to track what is happening in the everyday lives of the people of the world.

Conveniently, in 2006, the microblogging service Twitter was born. Twitter, founded on the principle that 'creativity comes from restraint' [4], allows its users to post short 140-character messages — "tweets" — designed to answer the question: "what's happening?" [5]. These updates appear on the user's profile in reverse chronological order, and other Twitter users can "follow" them — in a manner akin to a subscription — to receive realtime updates from all users he or she is interested in. The tweets can contain linked "`#hashtags`", which users can click on to track the way people are discussing certain widespread ideas or opinions. They can also be directed towards specific users (via @user) or can be forwarded from a previous tweet (called a "retweet"). Public tweets are accessible via Twitter's search interface, so users participate in a public, traceable, and purposeful exchange of information and opinion—all from the convenient location of a web browser or mobile phone application. To give readers a better sense of the Twitter interface, my own Twitter profile (Figure 1.1a) and home page (Figure 1.1b) are displayed on the next page.

In the words of Hermida [6], Twitter is really somewhat of a collective "awareness system". It embodies a new type of information-sharing and news-reporting — in real-time — that is by the people and for the people. But how accurately can we use this type of personal reporting to examine eating behavior? Do people regularly tweet about the food they eat?

The answer to the above question, it seems, is an emphatic 'yes'. Billions of Twitter users per day post updates mentioning some type of food, and recently it has been shown that approximately half of young adult "Millenials" tend to tweet while they eat [7]. This makes sense when one considers how intertwined eating is with other events of daily life. Within a microblogging service designed to report 'what's happening' in real-time, food is bound to be mentioned quite often. Why not take advantage of the constant conversation about food being held across the social Web?

In this work, I intend to use the wealth of food information provided by personal tweets to describe and predict real-world trends in metabolic disease and eating ideas. Specific hypotheses are presented in detail in Chapter 3.

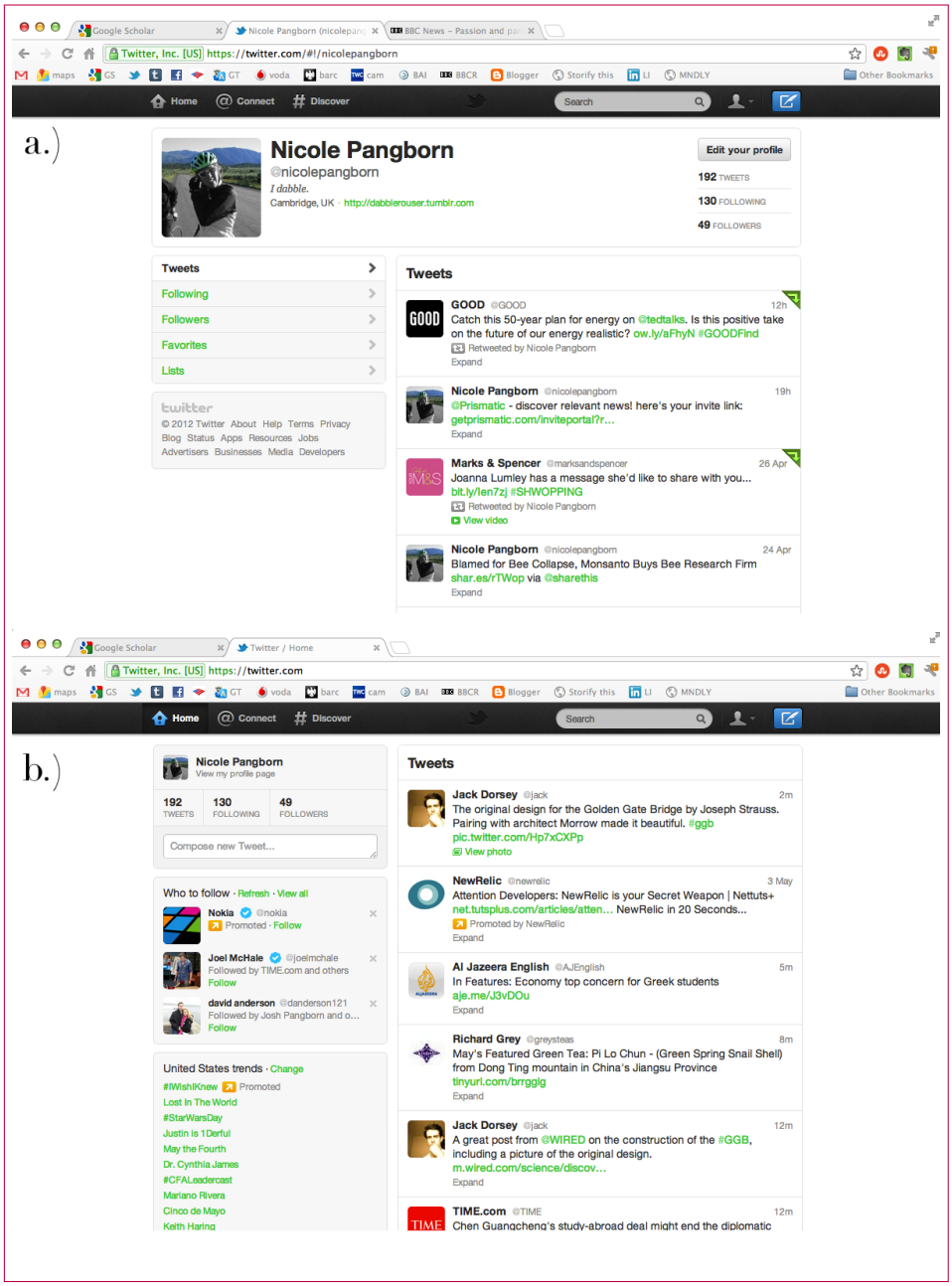


Figure 1.1: Twitter environment

(a.) *My own Twitter profile. Note the timeline stream of short status updates combining my own 140-character posts and 'retweets' - posts originally from other users that I deemed important enough to re-post.* ***(b.)*** *My own Twitter homepage. This displays the real-time stream of 140-character tweets from the users I follow (i.e., subscribe to).*

Chapter 2

Literature review

This work as a whole attempts to combine separately well-established realms in order to gather new information from the ways people discuss food in the Digital Age. These realms, as reviewed by this chapter, are:

- Twitter: properties and previous uses
- Obesity and food environment in the United States

Each of the above sections will serve to point out relevant research as well as what is missing from the current body of work.

2.1 Twitter: properties and previous uses

2.1.1 User behavior and network structure

From the beginning, Twitter was recognized for its unique user behavior and network topology. With its 140-character update limit, Twitter forces users to be brief and focused — usually resulting in a single idea per update. A high influx of these concise and targeted messages (each of which require relatively little time and thought to generate) allows for a much more rapid exchange of ideas than that observed in traditional blogging systems. In addition, the constraints set on the users cause a unique set of tweeting practices to emerge. When these real-time practices are combined with Twitter's distinctive one-way following mechanisms, the resulting structural properties create an ideal environment for the spread of information.

Early assessment and Tweeting Practices

In one of the earliest surveys of the then-new type of microblogging network, Java et al. [8] determined through means of word frequency and intent analysis that most Twitter updates were centered around three main categories: daily routine or activity updates, short conversations with specific users, and the exchange of information or news via summaries and URL-sharing. As such, they claimed that users themselves could also generally fall into one of three categories:

- Friends - those posting mainly activity updates for their general followers or participating in directed conversations by use of the '@user' mention syntax;
- Information sources - those frequently posting informative or URL content; and
- Information seekers - those rarely posting themselves, but simply following the posts of many other users.

While the update categories of Java et al. still hold well five years later, the lines between their relatively simple and distinct three user categories have recently become a bit more hazy. Many users today display a mixture of behavioral characteristics that span across all three categories [9]. In particular, the practice of "retweeting" (re-posting another user's tweet) — according to Boyd [4] — has purposes that combine multiple aspects of those mentioned above. In his 2010 paper, Boyd refers to retweeting as "[both] a form of information diffusion and a means of participating in a diffuse conversation", and it is perhaps Twitter's most powerful sharing mechanism.

Retweeting another user's tweet is a way to "validate and engage with others" [4]. Originally, the retweet arose with Twitter's first user base: typical syntax would include *"RT @originaluser 'text of original tweet' retweeter's commentary"*. When this traditional syntax is used, a retweet could serve to comment on another's post by adding new content. Eventually, as the 'RT' syntax became more popular, Twitter built it into their interface — now, a retweet button can be found below any tweet, and clicking it would re-post that tweet onto the user's own stream. A user might retweet a message to allow it to be spread to new audiences, to publicly agree with the original user, or as an act of friendship in drawing attention to the original tweet. On the other hand, a retweet might also be used for more selfish purposes: to gain followers, to save tweets for future personal access, to make one's prescense visible, or as a type of 'shout out' for political gain.

The types of posts that are retweeted tend to include breaking news, trending `#hashtags` (public conversations), or call-to-action messages (e.g. for donation or crowdsourcing). In all cases, retweets are most commonly meant to be seen — they allow users to participate in spreading some type of thought to as many other users as possible [4]. The result is a different type of overall conversation: as Boyd puts it, "rather than participating in an ordered exchange of interactions, people instead loosely inhabit a multiplicity of conversational contexts at once". The high exposure level created by these varied contexts increases the chances of a single idea spreading to a much larger audience. In the context of this study, retweets would potentially allow an idea about food to be spread quickly and to a high number of users.

Network Topology

When the mixture of these most common intentions and sharing practices of users are coupled with Twitter's one-way connections (i.e., user A can "follow" the updates of user B, but B does not have to follow A in return), an interesting overall structure emerges. This structure is important to understand, as it can in turn affect the manner and speed with which information through the network is spread by users.

Twitter's topology has changed in the years since its inception. In 2007, Java et al. [8] found a high rate of follower-following reciprocity: in general, if user A followed user B, user B also followed user A. However, this was most likely due to the nature of Twitter's first user base and the high probability that members of the small start-up community shared similar interests. A more recent topological analysis by Kwak et al. [10] shows that as the network grew larger and the user base became broader, much as it is the day I write this, the level of reciprocity in user connections plummeted. In 2010, 77.9% of user links were one-way following, and only 22.1% were reciprocated. In addition, 67.6% of users were found to not be followed by any of the users they were following.

Kwak et al. suggest that this one-way following mechanism plays an important role in other emergent properties of the network as a whole. A one-way connection between user A and user B indicates more of a "subscription" than a "friendship". In their 2010 paper, a plot of the complementary cumulative distribution function of the number of followers is shown to deviate from the predicted power-law distribution when x (no. of followers) is greater than 10^5 (Figure 2.1).

This indicates that unlike traditional social networks, which consistently display a power-law degree distribution (no. of connections per person), some Twitter users have a much higher number of followers than would be pre-

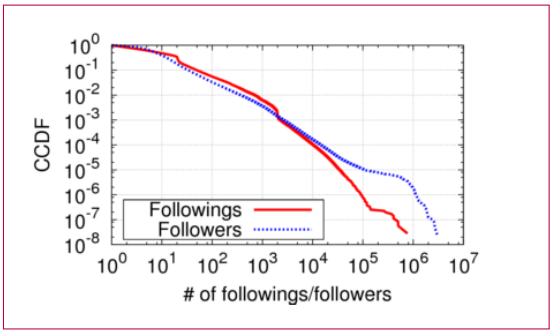


Figure 2.1: Follower-Following distribution

Figure from Kwak et al., 2010

dicted. Kwak et al. attribute this to the fact that there are many celebrities and public figures with Twitter accounts, and these types of users typically have millions of followers. Consequently, many people are now exposed to the previously inaccessible thoughts of popular personalities.

With this unique set of one-way connections, one would imagine that the average *path length* — i.e., the average number of 'hops' it takes to get from any user to another — between any two Twitter users might be longer than those found in traditional social networks, since sometimes direct reverse connections are nonexistent. When this is the case, paths might follow very different (and most likely longer) routes between users in the reverse direction. However, Kwak et al. [10] found the opposite: the Twitter user network has an average path length of 4.12. This is much less than expected given that path lengths of approximately 6 are found in offline social networks [11] and even other online social networks [12] whose sole purposes are to allow users to connect and interact with others. This finding suggests that Twitter's use is much more geared towards optimal spreading and exposure of important information and opinions than it is towards directed and personal interaction.

Summary

The new focused and brief user posts, tweeting mechanisms, and interesting one-way follower-following connections combine to make Twitter a one-of-a-kind type of social media technology — one that is seemingly set up to be an extraordinarily powerful information spreader. This spreading power applies to all popular shared topics (food being among them).

2.1.2 Translation into superspreading

Given the structure of the Twitter network and the types of posts coming from its users, how can such static properties translate into the dynamic and powerful superspreading and information cascades seen over time?

Speed and Scale of Dissemination

Perhaps when it comes to actually measuring the speed of spreading information, Twitter's most important property is its ability to show user posts in real-time. When followers of users can receive updates almost immediately, the speed with which information can spread throughout the entire network will necessarily increase. When this property is combined with the retweet and following mechanisms discussed in the previous section, the result is an extraordinarily fast diffusion of information after the first retweet. Indeed, in the network survey by Kwak et al., it was found that any retweeted tweet on average reaches about 1,000 users: after the initial retweet, the tweet is retweeted almost instantly on the 2nd, 3rd, and 4th hops from the original user [10].

Of course, the content of these tweets most certainly plays a role in their spreading power, but a study by Yang et al. [13] confirms that properties of the users themselves are just as - if not more - important in determining whether or not a tweet will travel far. In their predictive model, it was shown that the number of times a user was previously 'mentioned' (via @user) can serve as a highly accurate indicator of whether or not a tweet will be spread to many other users: tweets of users with more previous mentions tend to travel much farther in the network and reach more people. When considering how many celebrities occupy the Twittersphere — and how much more likely these public figures are to have a high number of followers and mentions than other users — it is easy to label them as our 'superspreaders'. Their thoughts and posts, due to the nature of the network, will reach many more people than would be the case for a 'normal' user. Since following requires no reciprocation, people get information immediately to and from these important users, who then can increase the chances of the tweet being spread widely. In addition, the users exposed to their tweets are themselves much more densely connected to each other (given the small path length of the network discussed above), and thus are more likely to see and retweet tweets.

Geographical Considerations

When examining the spread of information on the Twitter network, it is also useful to consider how the spread between users corresponds to the spread between their respective geographical locations. Does the widespread nature of the Internet take away the effects of physical distance? Is information passed on even more quickly between users who are located in the same region? How might this display or affect regional patterns in ideas passed through the network?

In investigating these effects, a previous study by Yardi et al. [14] found that local Twitter #hashtag networks are significantly more dense than those that are non-local. In their measurements, during a local event, users geographically close had a distinct advantage in receiving information about that event. These results could simply indicate that the certain local events chosen only happened to be important to local individuals, and that quite possibly a more "important" widespread event would show an increased diffusion past the initial locale. Even so, the high density of local #hashtag networks indicates a potentially significant regional dynamic at play.

In another more recent study by Takhteyev et al. [15], a similar result in network density was found: 39% of social ties on Twitter were concentrated between users less than 100km apart. This is somewhat surprising given the ease of connection that the Internet provides across long geographical distances. Takhteyev et al.'s data refute the long-held claim that "distance is dead" [16]. It seems, then, that the case of geography will be particularly important in my own project, since I will be discussing the spread of information about food — especially given how varied eating patterns are by region in the United States (Section 2.2).

Summary

When Twitter's embedded 'retweet' mechanism is combined with a short average path length and real-time sharing capabilities, the result is user-generated content that can spread to a wide range of people extraordinarily quickly. Despite this overall online speed and spread, it seems that geography still plays a role: when the information is specific to a certain locale (as food often is), the information is spread even more quickly among users within that locale. This could result in regional #hashtag and topic clustering, which will be especially important in my examination of regional food and metabolic disease differences.

2.1.3 Twitter and prediction

Because of its broad user base, geographic spread, and real-time nature, Twitter posts have been used by researchers in the recent past to predict aspects of the real world, such as events, sentiment, and biological phenomena.

Predicting an Event

In 2010, Sakaki et al. [5] devised a tweet classifier based on keywords associated with earthquakes in an attempt to detect the occurrence of earthquakes in Japan in real-time. By treating each Twitter user as a type of earthquake sensor — measuring when, where, and how frequently the sensors emitted signals (i.e., keywords associated with earthquakes mentioned in tweets) — they created a spatiotemporal probabilistic model for prediction. Their model came in two main parts: in the time series, they calculated the probability that an event was occurring based on the changes in the frequency of earthquake-word signals mentioned in tweets throughout time. Spacially, they then used user location data and GPS tagging (available for some tweets) to extract the center of a tweet trajectory. This combined model successfully predicted occurence and location of 96% of earthquakes stronger than a level 3 on a seismic intensity scale, and 100% of those stronger than a level 4. These results demonstrate the power of social signal on Twitter in gathering real-world information [17].

Predicting Sentiment

Twitter has also proven to be a powerful tool in predicting how the public feels about aspects of the real world. Jansen et al. in 2009 investigated the effectiveness of what they termed 'eWOM' (electronic word of mouth) on Twitter for acquiring information about a company's position in the consumer market [18]. They found that approximately 10% of all tweets contain some mention of a brand and an opinion about that brand. Their sentiment algorithms measured the relative frequencies of positive and negative terms in tweets to classify customer brand perceptions, which companies could then include in their marketing strategies. In a similar analysis of sentiment, Tumasjan et al. [19] used tweets to analyze how the German public felt about candidates in the weeks leading up to the federal election of the national parliament in Germany (September 2009). When sentiment results were combined with each candidate's 'share of Twitter traffic', or how many total mentions each candidate had relative to the others, the Twitter ranking of the candidates was identical to the ranking in the actual election results. (The actual percentages themselves were only slightly different, with a low

mean absolute error of 1.65%.) Innovative ways of using Twitter to explore public sentiment are becoming more and more important as the usage of social media continues to grow across the globe. Twitter and Facebook Revolutions — such as those so central to the 'Arab Spring' protests of 2011 — might well have been predicted upon closer analysis of online sentiment. As one Tunisian explained to Guardian reporter Beaumont in 2011, social media is "how we tell the world what's happening" [20].

Predicting Biological Phenomena

A third relevant application of the data from social communication on Twitter comes in the realm of predicting trends in health. People do discuss health issues online with each other, and it would be beneficial from a public health standpoint to make use of the large amounts of social data that are so readily available. Scanfeld et al. [21] reviewed Twitter status updates mentioning 'antibiotics' (and derivatives) to understand the ways in which people discuss their use of antibiotics with each other online. A random sample of 1000 was chosen for content analysis, in which human workers categorized tweets into types of use. Despite the potential error associated with these manual efforts and the small sample size, 97.1% of the sample tweets containing the word 'antibiotics' were able to be classified into a category of relevant biological discussion. (In some of the largest categories, 29.8% of users discussed general antibiotic use, and 16.2% sought advice from one another.)

In 2012, Sadilek et al. [22] took the notion of a Twitter health analysis one step further and attempted to use mentions of user sickness to predict the real-time spread of disease. Since certain tweets (though a relatively low percentage) are geo-tagged, there is a potential to acquire information about a user's precise whereabouts. Their goal was to achieve a higher level of accuracy than traditional disease-surveying methods, such as paper questionnaires and Google flu trends, which both have a substantial amount of missing data: any sick person who either misses a paper survey or happens to not type their symptoms into the Google search bar is essentially invisible. Though Twitter analyses would also have some missing data, no previous disease prediction methodology has taken into account the ways in which people discuss their sickness with each other in real-time. After a machine-learning process which allowed automatic differentiation between just a general mention of sickness on Twitter vs. a specific indication that users were themselves sick, Sadilek and his group found a clear exponential relationship between probable physical encounters with sick friends and ensuing sickness. An important point to make is that these trends reflect not

the friendships themselves between sick users, but are merely indicators of a more complex set of phenomena — e.g. being exposed to the same atmospheres, sharing drinks, etc. — which might not be directly attainable. The same will apply to my own analyses, which will serve as an indicator of possible eating behavior outside of the realm of Twitter.

Summary

It is clear that researchers are just beginning to pay attention to the substantial amount of social data available in online communication. Twitter can be (and has been) used to predict the real-time spread of event information, ideas/sentiment, and even disease. The question then becomes: what can Twitter tell us about the information and ideas that *affect* disease?

2.1.4 What's missing?

Its network properties have been assessed, its superspreading power has been ascertained, and its correspondence to events in the real world has been documented. But to what extent does Twitter reflect a more behavioral human reality? Can the geographically-relevant spread of ideas on Twitter help us understand how and why different people might make different decisions? In particular, when ideas are spread about food—a distinctly cultural and environmental topic—through such a new medium, could they be translated into metabolic disease patterns? The realms covered in my own project will cross those of communication, biology, and psychology alike, and the next section of this literature review will briefly address the aspects of obesity research and social influence which will be necessary for my approach.

2.2 Obesity and food environment in the U.S.

It is a truth universally acknowledged that the people of the United States are some of the most obese in the world [23]. Noticeable yearly increases in obesity prevalence began to emerge here in the 1970s, but by the 1990s, those increases began to skyrocket. With the overall prevalence rising nearly 6 percentage points between 1991 and 1998 [24], obesity started to become a focal point of health-related research. By the mid 2000s, as the rates continued to increase in an even more alarming fashion (up to their current figures of 67% overweight, 34% obese [23][25][26][27]), the term 'epidemic' became inextricably tied to all mentions of obesity in the United States.

Among its many aims, this work attempts to link the realms of obesity research and digital communication. Though the issue of obesity is fundamentally biological in nature, a number of external factors have also proven themselves as important instigators [28]. Some of these could very well be displayed in the ways in which people discuss their food choices with each other on Twitter and other outlets in the Digital Age. As the full range of such factors would require a substantially longer and more in-depth literature review, this section will mainly address those necessary given the scope of my project: regional differences in obesity rates, aspects of "obesigenic" environments, fructose consumption, and food idea transmission.

2.2.1 Regional differences

Obesity Rates

While the enormous increase in overall obesity prevalence described above applies to the United States (and, indeed, the world [23]) as a whole, it is clear that different regions of the country display markedly different rates. Using data from NHANES [29], BRFSS [30], and the National Survey of Children's Health [31], Wang et al. [25] and Singh et al. [32] have led the way in recent years with their in-depth geographical analyses of US adult and child obesity. Both studies found substantial and consistent differences across states, with the Southeastern region significantly heavier than the Midwestern and Northeastern states, and the Western states significantly lighter (general relationship by weight: Southeast > Midwest > Northeast > West). In the below map from the Centers for Disease Control and Prevention [33], we can observe these striking regional differences (Figure 2.2).

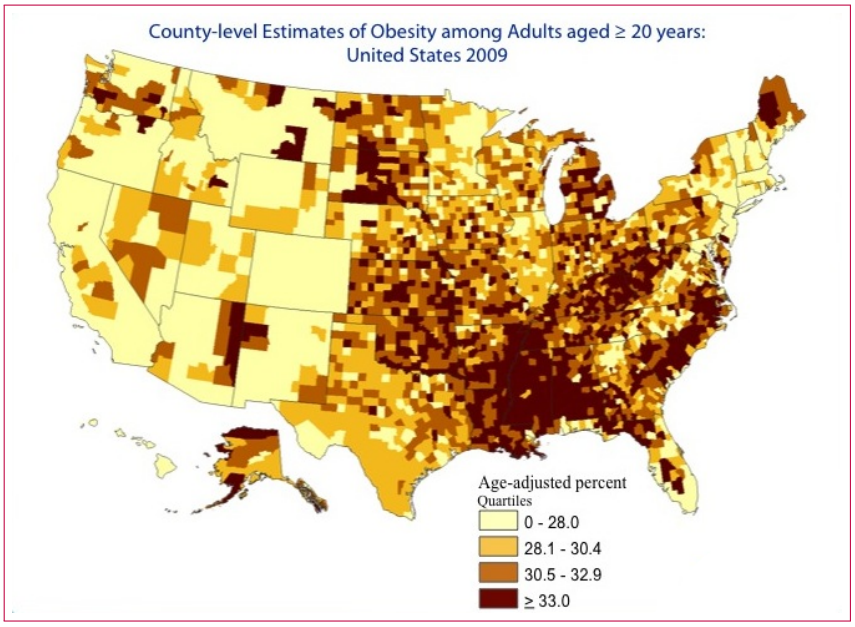


Figure 2.2: Regional Obesity Prevalence

Most recent available regional map of obesity prevalence in the US from the Centers for Disease Control and Prevention, taken from http://www.cdc.gov/

Obesity has frequently been tied to factors such as socioeconomic status [34][35][25][36], race or ethnicity [37][38], neighborhood social capital [39][40], inactivity [40][41][42], and crime rate of the region [43][32][44], all of which might vary slightly from state to state; however, even with all of these factors 'corrected' for in multivariate logistic regression analyses [25][32], we still observe significant regional differences in adjusted obesity prevalence [45]. In fact, the disparities among states actually became clearer over the period of time from 1990-2005 [25], during which differences in economic inequality became more pronounced in all regions, particularly the Southern, but showed very similar rates of increase in the Northeast, Midwest, and West alike [46]. Essentially, though the personal/demographic factors listed are all strongly associated with obesity, and most certainly play a role, they themselves do not fully explain the heightened geographic variance.

Environment

Since the classic demographic characteristics described above cannot completely account for the regional disparities in obesity rate, we might turn to more environmental explanations. A body of recent work suggests that the obesity problem might be less of a personal issue and perhaps more of a

general problem of exposure to "obesogenic" environments.

Grocery store availability seems to play an important role in the examination of how differences in regional environment may relate to health disparities. Large chain supermarkets and grocery stores, often containing healthier food options [47], tend to be much more highly concentrated in suburban and wealthier regions, while non-chain small markets and convenience stores tend to litter the streets of urban and inner city areas [48][49]. This is interesting, particularly given the fact that obesity rates in general tend to be higher in inner city regions as well [25]. A number of works in fact suggest that grocery store availability is directly correlated with regional obesity rate: in regions where supermarkets are in low density (but convenience stores are in high density), the prevalence of obesity and overweight is higher, while in high-density supermarket regions, the prevalence of obesity and overweight is lower [50][51][52].

The same types of correlations seem to be present when the densities of fast food restaurants are examined. In the US, fast food consumption has been increasing rapidly each year [53][54], and poorer urban areas tend to have a higher concentration of fast food outlets than wealthier suburban neighborhoods [55][56]. Those consuming large amounts of fast food have been shown to consistently be much heavier than those who do not [57][58], and it seems that such easy access to fast food is also reflected in the rates of obesity: in the Southeastern US especially, prevalence of obesity and overweight is significantly higher in areas with more fast food restaurants [59][28].

The above urban vs. suburban studies, while meaningful, still do not address the issue of regional differences. After all, every region of the United States has its fair share of urban and suburban counties. What would have a regional effect, however, is if the overall concentrations of fast food restaurants were significantly different in different areas. Indeed, Powell et a. [56] found exactly that: percentages of zip codes containing fast food and other restaurants mirror the obesity rates by region, with the Southeast having the highest percentage (34.0%), Midwest following behind (29.8%), then the Northeast (19.5%), and finally the West (16.6%).

These regional environment-health correlations say nothing themselves, of course, unless people actually tend to eat what's around them. Luckily, that also seems to be the case: Cheadle et. al [47] found that surveys of individual eating habits in different areas accurately reflected the different types of food available in the stores of the area. If the widespread increased prevalence of obesity—differing by region, but still increasing across the country—can't be fully explained by simultaneously emerging individual patterns of choice (i.e., everyone across the country decided at the same time to make poor food choices), it seems that regional changes in the types of food available to large

groups of individuals might provide some more answers. In this sense, the US is a very classically "obesogenic" place to live — in some regions more than others [51][56][60].

Cultural Considerations

Another aspect of food differences by region is perhaps inextricably tied to food culture—which, as Michael Pollan put it, is really just another word for "your mother" [60]. These types of differences in food preference are not quite as easily quantified as other aspects of the obesogenic environment, but attempts to describe regional cuisine have been made [61][62]. In general, Southern and Midwestern states do tend to incorporate many fried and fatty dishes into their cuisine, whereas the Northeastern and Western regions are known for their many seafood dishes [62]. While these cultural differences are important, every region does have characteristically 'unhealthy' foods. Full descriptions are perhaps a bit outside of the scope of my own Twitter analysis, which will focus on the more quantifiable and comparable aspects of regions (references to fast food availability via mentions on Twitter in particular).

Summary

For various reasons—including the density of grocery stores and fast food restaurants within a neighborhood—regions of the United States display markedly different obesity rates and eating patterns. Regional obesity rate will be one of the main variables assessed in this work as well as mentions on Twitter of aspects of "obesogenic" environments (fast food vs. supermarkets, etc.) (Chapter 3).

2.2.2 Fast food, fructose, and fat

Though behavioral and environmental factors can instigate obesity, the issue is fundamentally biological in nature. What is it specifically about the food sold in fast food restaurants that might trigger the deposition of large amounts of fat?

Despite those who claim the problem is just a matter of energy balance from eating too much fast food (a "kcal in - kcal out = kcal stored" mentality [63][64]), recent works suggest that such a simplified equation cannot completely account for persistent weight gain. While, crudely, application of the laws of thermodynamics must be correct, the translation between kcal consumed / kcal expended and weight gain does not appear to be a linear

one in any sense; as Wells points out [65], for some, the energy contained in "half a biscuit per day" is enough to bring about substantial weight gain over time. This mode of thought might aid in explaining the high percentage of failed diets (in the long term) involving only calorie restriction and exercise [66].

If the problem is not simply overeating the readily available (and inexpensive [67]) fast food, could it involve a chemical or biological aspect of the food itself? Some proponents of "discordance theory" maintain that our current food is mismatched with the biology of our bodies. Founders of discordance theory such as Eaton and Cordain claim that many types of modern foods are so drastically different from the human diet even hundreds of years ago — what would be considered a fraction of a split second in an evolutionary timeframe — that our bodies have not caught up biologically [68][69][70]. However, if the problem was a very general discord involving a wide range of modern foods, as the early proponents of discordance theory held, one would think the rises in extreme obesity rates might have begun much earlier than they actually did. Curiously, the sharp rises in the United States began in the 1970s, coinciding precisely with the introduction of one novel sugar substitute into the US food supply [71].

The specific role of this compound — high fructose corn syrup (HFCS) — in the obesity epidemic is widely debated within the scientific community. As a substitute for traditional table sugar, sucrose, which is a 50:50 mixture of glucose and fructose, HFCS is a cheaper, more easily distributed sweetener with a 45:55 glucose-fructose ratio. Since its first widely-viewed association with the rising obesity trends [71], it has come under fire multiple times — with mixed results [72]. Recently, however, simply fructose itself has become an object of scrutiny due to its interesting metabolic properties.

Fructose is, of course, found in many natural foods (fruits in particular), but in the wild it is consistently accompanied by natural fiber and is never in concentrations as high as observed in processed sweets, soda, and fast food [73]. The fructose model of obesity centers around the fact that fructose is metabolized quite differently than glucose and other types of sugars, and when removed from the fibrous bodies of natural fruits, there is some evidence that it may cause extensive fat deposition, insatiety, and hyperinsulinemia in both humans [73][65][74][75], and non-human mammals [76][77][78].

The consensus among this group is that high concentrations of processed foods containing fructose use an inslin-independent receptor (GLUT5) for intestinal absorption, which happens to be activated solely in the liver [73][65]. In the liver, it skips the "checking" phase of glycolysis that would be present in most other metabolically active sites, and as a result, glycolysis intermediates are rapidly accumulated, which are soon converted to fatty acids,

very low-density lipoproteins, and triglycerides (i.e., fat). In addition, its rapid metabolism uses up ATP stores very quickly, tricking cells into a false state of "starvation" [74], activating ATP-kinase, and causing an excess of glucose and insulin to be released into the bloodstream [79][74][65]. The entire process leaves one feeling deceptively hungry while actually having accummulated excess fat [73][65][80][75]. In Lustig's terms, fructose creates a "feed-foward" loop of obesity and hunger [81].

There are some that disagree with the validity of the fructose model, as the proposed biochemical pathway of fructose metabolism and its effects have yet to have been reproduced on a mass scale during a human clinical trial [82][83]. Even so, the widespread hike in fructose consumption in the United States during the past 40 years is cause for suspicion. Its co-occurrence with rising obesity rates points to an at least interesting relationship between the two, and its presence in widely-eaten fast food makes it a convenient marker for regional differences in consumption. It is for this reason that I chose to include food known to contain high amounts of fructose in my examination of regional tweets.

Summary

New research points to a complex association between biochemical components of fast food and obesity. These are the foods which will be included in my monitored term list while searching tweets.

2.2.3 Obesity and transmission

When discussing aspects of environment (such as the above differences in the types of food available to particular regions), it is also important to consider the potential effects that regional social environments might have on the growth of the obesity epidemic. Are ideas about food and eating spread between individuals? Does distance or geography play a role?

One attempted answer to these questions came from Christakis and Fowler in 2007 [84], who published a controversial study about the spread of obesity within a real-world social network. They followed the weights and friendships of 12,067 adults within the Framingham Heart Study from 1971 to 2003, and used aspects of social network analysis to describe the trends in weight gain among connected individuals over time. Significant clustering was found among groups of obese individuals and groups of non-obese individuals, as well as what appeared to be significant "peer effects": i.e., that the likelihood of an individual becoming obese was increased after a connected individual became obese. These effects depended on the nature of the

relationship between individuals, with directional friendship ties showing the strongest increase in likelihood (57%).

Christakis and Fowler also claimed that this type of "*social* distance" — degrees of separation — was more important in determining the probability of becoming overweight than *physical* distance. However, it is important to note that the entire cohort was located within close boundaries throughout the timespan (in or around Framingham, Massachusetts), and the majority of the distance connections examined (5 of 6 distance groups) were within a 16 km radius. When comparing to national distances and differences in obesity rates, 16 km is quite trivial. Indeed, a subsequent study from Cohen-Cole and Fletcher [85] used econometric techniques to test Christakis and Fowler's claims of peer effects on a larger national sample, and found that shared environmental factors played a much larger role in 'transmission' of weight status. Cohen-Cole and Fletcher's results suggest that when comparing across nationally-relevant distances, group-level characteristics, rather than direct peer effects, were behind the more likely mechanism of weight influence.

Summary

In both of the cases presented above, there is a strong correlation between interconnectivity of individuals and subsequent "transmission" of health ideas and metabolic patterns. While the separation of peer effects from the effects of group environment will not be addressed here, this project will test whether or not Twitter can display their combined effects over national distances (see Chapter 3, Hypothesis 2).

2.2.4 What's missing?

While extensive amounts of research have been dedicated to the "obesity epidemic", particularly in the United States, none so far have attempted to use online social networks — which, as we have seen, are powerful modes of communication and information-spreading — as tools for prediction and interpretation of metabolic disease patterns. It is important to note here that this work is not an investigation into causes of obesity, but rather it is an attempt to use the wealth of information provided by online social communication to reflect real-world eating, weight, and environmental trends.

Chapter 3

This project

This work attempts to tie together the realms discussed in the last chapter. It will not solve the problem of the true biological or economic origin of obesity and regional differences in health, but will rather attempt to answer the following questions: Can the above regional obesity phenomena be reflected in the way people talk to each other via Twitter? How accurately would we be able to use social media as a tool for metabolic disease prediction? And finally, if online social networks provide another medium through which users can influence the ideas and decisions of others (with a speed unmatched by traditional offline networks — especially within geographic clusters), could this new way of discussing food highlight the transmission of food opinion and eating behavior?

3.1 Hypothesis 1

I hypothesize that the relative frequencies for mentions of the below terms in tweets from designated cities of the continental United States will be positively correlated with city obesity rates:

- **Fast food** - chosen because of its label as an aspect of "obesogenic" environment and association with fructose consumption;
- **Soda/pop** - chosen because of its connection to fructose;
- **Candy** - chosen because of its connection to fructose; and
- **Hunger** - chosen because insatiety is seen as a symptom of high fructose consumption.

I also predict that relative mentions of this next set of terms will be negatively correlated with regional obesity rate:

- **'Healthy'** - chosen as a marker of healthy eating culture; and
- **Supermarkets & grocery stores** - chosen because of their associations with non-"obesigenic" environments.

I also think that all of the above categories of mentions will differ significantly enough by region to serve as a tool for prediction and classification.

3.2 Hypothesis 2

Given both that:

- friends on Twitter are more likely to be found within 100 km of each other, and
- tweets travel much faster within these dense local networks (see section 2.1.2),

an interesting set of metrics to examine in my analysis of food tweets would be the relationship between obesity or obesogenic foods mentioned and physical distance. In testing this relationship, I would simultaneously observe the effects of both proposed "transmission" mechanisms presented earlier (peer influence and group environment). This project will not separate the two, and will not track the mechanisms themselves, but rather will test whether or not Twitter can display their combined effects over distance.

More specifically, here I hypothesize that differences between cities in both obesity rate and the relative frequencies of words mentioned in Hypothesis 1 will be positively correlated with differences in physical distance.

Chapter 4

Methods

4.1 Twitter streaming API

All tweets in my sample were acquired through use of Twitter's Streaming API. Twitter provides extensive and easily accessible documentation for developers on their website, `https://dev.twitter.com/docs/streaming-api/methods`. For this section, two types of queries to Twitter's API were made. In both queries, I was only granted simple developer access to the stream, which encompasses a random sample of about 1% of all tweets. (The Twitter "firehose" and other streams grant access to much higher percentages, but only certain corporate partners are allowed to pay for these privileges [86].)

4.1.1 Short run for #hashtags

First, I ran a general query of the `statuses/filter` streaming API. The `statuses/filter` stream returns a random sample of public tweets that match certain filtering parameters, which can be altered by adding or removing constraints on the initial command. The goal in this shorter run was to use a random sample of all tweets within the continental United States in order to obtain the top food-related terms and #hashtags tweeted by US users. These 100 food tags would then be used in the longer run for data in which only US tweets containing those tags would be included in the output, giving me a random sample of all food-related tweets from my regions of choice.

I inserted location parameters by defining a bound box — a set of bounding longitude and latitude coordinates, beginning in the southwest corner and ending in the northeast — of the continental United States (bound box = -124.51, 24.54, -66.95, 49.00). I ran the full command over an 18-hour pe-

riod on 01 May 2012 — throughout the span of conventional 'eating hours' in all time zones of the continental US — and obtained 486,503 unique tweets. The command saved the output to a text file, which I then processed using simple Python scripts. All Python scripts used for text file processing are included in Appendix A.

The output of the curl command to the Twitter API is always in JSON format, which is highly readable, but also contains some extraneous information. A simple script processed this large output in pieces to obtain only the text of the tweets. Next, another script analyzed word frequencies and obtained the top food-related #hashtags. I included the top food-related words by frequency in the output ($n = 87$) as well as supplementary food-related #hashtags ($n = 13$) which were either mentioned in recent articles [87][88] or observed as current trending tags on Twitter. These top 100 #hashtags (along with the counts of those found in the sample of 486,503 tweets) are listed below:

food 3571
eat 3128
lunch 2852
dinner 2241
Starbucks 2094
coffee 2063
hungry 1843
pizza 1700
eating 1579
Cafe 1529
Restaurant 1356
chicken 1220
breakfast 1124
cake 954
cream 933
taco 906
McDonalds 829
mayo 820
subway 809
chocolate 788
sushi 745
burger 743
ate 680
cheese 669
yum 622
treat 554

kitchen 550
bread 544
yummy 537
chipotle 521
cook 510
juice 484
delicious 371
Mmm 364
Steak 363
sandwich 352
healthy 350
salad 349
cheeses 349
milk 340
cookies 326
bbq 322
foods 313
apple 298
feed 294
cooking 289
meal 283
candy 281
diet 269
starving 263
donuts 260
tacos 259

Panera 256
sauce 254
fries 251
chips 247
bacon 246
Salt 243
Deli 237
beef 235
Dunkin 235
craving 233
ham 232
butter 232
Wendys 229
burgers 225
cookie 225
rice 223
sugar 210
fried 207
fruit 206
soup 201
plate 201
Bistro 193
pie 188
Applebees 187
Steakhouse 186
meat 185

pancakes 184
cereal 182
popcorn 176
Diner 169
snack 165
CHickfilA 164
banana 161
jelly 152
eggs 152
omnomnom
omnomnomnom
tasty
vegan
vegetarian
tweetwhatyoueat
groceries
nutrition
grocery
supermarket
cuisine
takeout
eats

4.1.2 Long data run

My second query of the Twitter Streaming API used the 100 food-related #hashtags found above to obtain only tweets in which food was mentioned. This run was substantially longer than the initial run for #hashtags, as a longer timespan was necessary to gather a sufficient number of food-related tweets for comparison by city. I ran another simple `curl` command with the `track` parameter replacing the `location` parameter used above. The `track` parameter, according to the Twitter developer documents, will match `AnyWord` with the following semantic correspondents in the text of tweets: ANYWORD, anyword, "Anyword", anyword., #anyword, @anyword, and http://anyword.com.

For the Twitter API, bounding boxes are logical ORs, meaning that if I had run a `curl` command with both `track` and `location` parameters included in the statement, the resulting output would be of tweets *either* containing the food-related #hashtags *or* within the continental US bound-box. Since this would give me quite a large number of irrelevant tweets, I opted to include only the `track` parameter to track the 100 #hashtags, and processed the output after the initial command was stopped to include only geo-tagged tweets from cities within the continental US. My post-processing method is described in the city categorization section below.

Between 22 May 2012 and 30 May 2012, I obtained **38,039,682 unique tweets** containing any of the 100 food-related #hashtags above. I ran a simple Python script to extract the essential bits of information from the JSON output — tweet text and location — for use in further analyses (see Appendix A).

4.2 City categorization

Of the 38,039,682 total tweets, I used only those properly labeled with a geographic location in one of Twitter's top trending US cities ($n = 48$) scattered throughout the continental US. Only a small portion of users tend to list their geographic location [89], so by using top trending cities, where user base is naturally high, I would be guaranteed the largest possible sample size from tweets containing location information. After the long data run described above, I filtered the JSON data to include only tweets from these cities (matched by county, as obesity rate is described by county in the US):

Atlanta, GA (Fulton County)
Austin, TX (Travis County)
Baltimore, MD (Baltimore City, independent)
Baton Rouge, LA (East Baton Rouge Parish)
Birmingham, AL (Jefferson County)
Boston, MA (Suffolk County)
Charlotte, NC (Mecklenburg County)
Chicago, IL (Cook County)
Cincinnati, OH (Hamilton County)
Cleveland, OH (Cuyahoga County)
Columbus, OH (Franklin County)
Dallas-Ft. Worth, TX (Wise, Denton, Collin, Hunt, Delta, Parker, Tarrant, Dallas, Kaufman, Johnson, Ellis, Rockwall Counties)
Denver, CO (Denver County)
Detroit, MI (Wayne County)
Greensboro, NC (Guilford County)
Harrisburg, PA (Dauphin County)
Houston, TX (Harris County)
Indianapolis, IN (Marion County)
Jackson, MS (Hinds County)
Las Vegas, NV (Clark County)
Los Angeles, CA (Los Angeles County)
Memphis, TN (Shelby County)
Miami, FL (Miami-Dade County)
Milwaukee, WI (Milwaukee County)
Minneapolis, MN (Hennepin County)
Nashville, TN (Davidson County)
New Haven, CT (New Haven County)
New Orleans, LA (Orleans Parish)
New York, NY (Bronx, Kings, New York, Queens, Richmond Counties)
Norfolk, VA (Norfolk City, independent)
Orlando, FL (Orange County)
Philadelphia, PA (Philadelphia County)
Phoenix, AZ (Maricopa County)
Pittsburgh, PA (Allegheny County)
Portland, OR (Multnomah, Washington, Clackamas Counties)
Providence, RI (Providence County)
Raleigh, NC (Wake County)
Richmond, VA (Richmond City, independent)
Sacramento, CA (Sacramento County)
Salt Lake City, UT (Salt Lake County)
San Antonio, TX (Bexar, Medina, Comal Counties)
San Diego, CA (San Diego County)
San Francisco, CA (San Francisco County)
Seattle, WA (King County)
St. Louis, MO (St. Louis City, independent)
Tallahassee, FL (Leon County)
Tampa, FL (Hillsborough County)
Washington, DC (DC, independent)

In total, 2,746,381 unique food-related tweets (7.22% of the 1% sampled) listed locations in one of these cities.

These cities were eventually separated into four distinct regions as defined by the U.S. Census Bureau: Northeast, South, Midwest, and West, as seen on the map below (Figure 4.1).

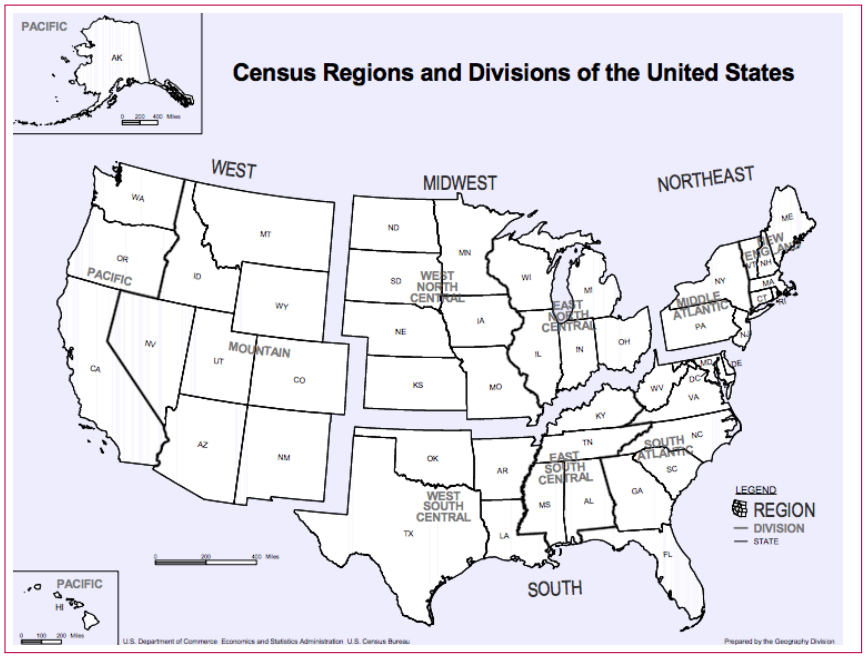


Figure 4.1: U.S. Regions

Depiction of the four main regions of the United States as defined by the U.S. Census Bureau (available online at www.census.gov).

4.3 Final data format

4.3.1 Tweet mentions: relative frequencies

The output file (containing the 2,746,381 unique food-related tweets from the above listed trending cities) was then processed separately 48 times, with each iteration producing a separate text file containing the text of tweets from one city on the list (see Appendix A for script). Each of those 48 text files — one for each city — was then processed again to obtain wordcounts for the tweets of that city. The output from this step produced data in a list format, with each element in the list itself a nested list: `[('word1', # of times word1 mentioned in city i), ('word2', # of times word2 mentioned in city i),...]`. With these wordcounts, and assuming that each tally for a particular word most likely came from a single tweet (i.e., that most tweets mentioning 'mcdonalds' tend to mention it only once given the 140 character tweet limit), I calculated the relative frequencies rf of my chosen food-related #hashtags in all food-related tweets from city i:

$$\boxed{rf_x(i) = \frac{\text{no. of tweets mentioning food word x in city i}}{\text{total no. of food-related tweets in city i}}} \tag{4.1}$$

As stated at the end of Chapter 2, I chose to focus my analysis on a subset of the food tags/categories that were most relevant to the obesity epidemic given the assertions of the current literature. Relative frequencies were calculated for each city in the following six word categories, which make up the bulk of my variable list, and which will henceforth be referred to as the words in Courier typeface:

- `McDonald's` = 'mcdonalds' and all derivatives ('mcdonalds' + 'McDs' + 'mickey ds'),
- `Soda` ='soda' + 'sprite' + 'coke'
- `Candy` ='candy'
- `Hungry` ='hungry'
- `Healthy` ='healthy'
- `Grocery` ='supermarket' + 'grocery' + 'groceries'

4.3.2 Obesity rates

Since I gathered very recent Twitter data by using posts from May 2012, I attempted to match them with the most recent obesity rates for the listed regions. The County Health Rankings team at the University of Wisconsin Health Population Institute assembled calculations made by the Dartmouth Institute from surveys administered by the Centers for Disease Control and Prevention [33] (mainly the National Vital Statistics System [90] and the Behavioral Risk Factor Surveillance System [30]) into one convenient location, `www.countyhealthrankings.org` [91]. Here is where I obtained most recent (2012) obesity rate data by county. Though the County Health Rankings team cautions researchers against using certain aspects of their county 'ranking' systems to compare states because ranks were calculated within each state, adult obesity is listed not as a ranking but as a 'Health Outcomes' measure, which can be accurately used for comparison across states [91].

As seen in the above city/county list, certain large cities in my study lie geographically in multiple counties. In order to obtain a single obesity rate for

these large cities (Dallas-Ft. Worth, TX; New York City, NY; Portland, OR; and San Antonio, TX) I obtained county populations from the US Census [46] available at `http://www.census.gov`. I weighted the obesity rates of counties in these large cities by population to obtain a weighted average of obesity prevalence in the city as seen below.

Let LC be a large city with population P_{LC}, obesity rate OB_{LC}, and n multiple counties c_i each with obesity rates ob_{c_i} and populations p_{c_i}. Then:

$$\boxed{OB_{LC} = \frac{\sum_{i=1}^{n} ob_{ci} * p_{ci}}{P_{LC}}} \tag{4.2}$$

4.3.3 Distance calculations

To measure differences in obesity or relative frequencies of food terms and their relations to differences in physical distance between cities, I constructed separate full distance matrices for each variable. Each distance matrix described the differences in a given variable between each city and all other cities. Differences in physical distance were calculated by obtaining latitude and longitude coordinates for the centroids of each city via Google Earth, converting to decimal degrees, and inputting the .csv file of decimal coordinates into QGIS (an open-source geographic information system, available at `www.qgis.org`). Using the spheroid reference surface WGS-84, QGIS returned distances in decimal degrees between each city and all other cities. I used an approximation of the Haversine formula to convert these decimal degree distances to kilometers for graphical representation ($1DD \approx 111$ kilometers).

To construct each matrix D, each describing differences in given variable x (e.g. obesity rate, physical distance, or rf of a given food term) between two cities i and j, entry d_{ij} was defined as:

$$\boxed{d_{ij} = |x_i - x_j|} \tag{4.3}$$

Since the results were symmetrical matrices (leaving the main diagonal entries all equal to zero), the upper triangles were discarded in analyses so as to not have a duplicate set of points.

Chapter 5

Results

The results assembled in this section address whether or not social media communication via microblogging platform Twitter can *describe* and *predict* real-world regional trends in metabolic disease and food ideas.

After tallying the tweet mentions of different food categories for each city and running some preliminary tests, I observed that Las Vegas was a consistent and extreme outlier in every analysis. Since the city is notoriously a frequent holiday spot (one for which living in excess and extravagance is considered the norm), I decided that the city would most likely not be reflective of a true permanent population. For this reason, I have excluded Las Vegas from all analyses. Results incorporating all other 47 cities are presented below.

Raw data tables can be found in Appendix B.

5.1 Relationships between foods mentioned and obesity rate

In order to understand the relationships between the relative frequencies of different foods mentioned on Twitter and regional obesity rate, I ran separate correlation analyses to detect individual associations. Before doing so, I checked for normally distributed samples (in all variables, within each region) with the Kolmogorov-Smirnov test. All p values were nonsignificant, so I proceeded to use Pearson's correlation coefficient for interpretation.

Relative frequencies of mentions of `McDonalds`, `Soda`, `Hungry`, and `Healthy` in tweets from a given city were all significantly related to obesity rate, but `Candy` and `Grocery` mentions were not. `Mcdonald's` was positively related ($r = .648, p < .001$), as was `Hungry` ($r = .572, p < .01$). Mentions of `Healthy`

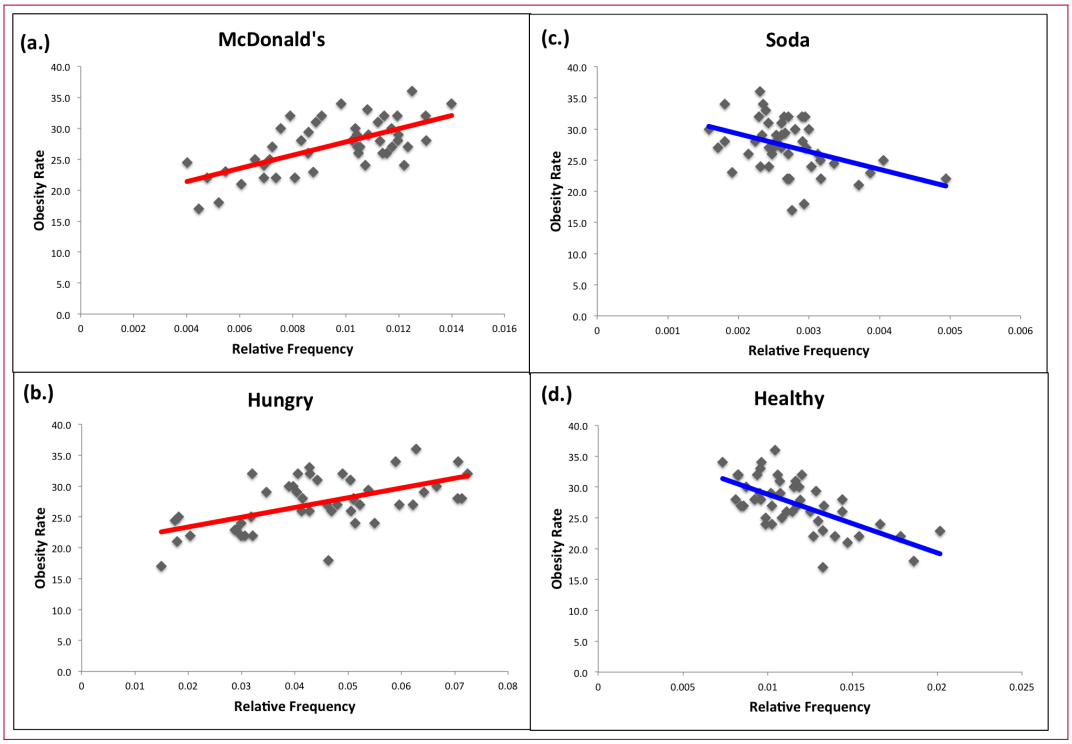


Figure 5.1: Obesity rate vs. Tweets

Significant correlations between city obesity rate and (a.) McDonalds mentions ('mcdonalds', 'McDs' and other derivatives) per US city; (b.) 'hungry' mentions; (c.) 'soda', 'coke', or 'sprite' mentions; and (d.) 'healthy' mentions.

were negatively related to obesity rate ($r = -.641$, $p < .001$). Surprisingly, `Soda` mentions were also negatively related to obesity rate ($r = -.412$, $p < .01$).

Three of the four significant relationships presented above confirm their respective portions of my hypothesis. It seems that, in general, tweets from US cities with higher relative frequencies of fast food and hunger mentions are associated with higher obesity rates, while tweets mentioning healthy options are in higher concentrations in cities with lower obesity rates.

5.2 Mean differences by region

The interesting significant relationships seen above between relative mentions of `McDonald's` / `Hungry` / `Soda` / `Healthy` and regional obesity rates led me to delve deeper into regional differences.

I conducted univariate one-way ANOVAs for the dependent variables obesity rate, rf `McDonald's`, rf `Hungry`, rf `Soda`, and rf `Healthy` mentions in all 47 cities grouped by region.

ANOVA 1: Obesity rate by region To confirm the statistics from the CDC and other data sources cited in Chapter 2, I first examined the mean differences in regional obesity rates, using my selected top-trending Twitter cities as the sample. Levene's test statistic was nonsignificant, so equal variances were assumed. Significant differences in obesity rate between the four regions (South, Midwest, Northeast, and West) was found ($F(3, 43) = 6.26$, $p = .001$). I used Hochberg's GT2 post hoc test to hone in on these differences, since the sizes of sample groups of cities within regions were unequal. I found that the mean obesity rate for cities in the West was significantly different from the means of both the South ($p = .001$) and the Midwest ($p < .01$), but that all other comparisons between regions were nonsignificant. (To aid in visualizing these findings, Figure 5.2 below presents a comparison of trends among regional means for all dependent variables examined.)

ANOVA 2: McDonalds mentions by region Again, Levene's test statistic was nonsignificant, so equal variances were assumed. Significant differences in mean rf `McDonald's` mentions by region were found ($F(3, 43) = 14.30$, $p < .001$). In addition, Hotchberg's post-hoc tests showed that the mean of `McDonald's` mentions from the West was significantly different from the means of the South ($p < .001$), Midwest ($p < .001$), and Northeast ($p = .003$).

ANOVA 3: Hungry mentions by region Levene's test statistic was nonsignificant, so equal variances were assumed. Significant differences in mean rf `Hungry` mentions by region were found ($F(3, 43) = 10.05$, $p < .001$). Hotchberg's post-hoc tests showed that the mean of rf `Hungry` mentions from the West was significantly different from the means of the South ($p < .001$) and Midwest ($p < .05$), but not the Northeast ($p\ ns$).

ANOVA 4: Soda mentions by region Levene's test statistic was nonsignificant, so equal variances were assumed. Significant differences in the

mean relative frequency of `Soda` mentions between regional groups were not found ($F(3, 43) = .701$, p ns).

ANOVA 5: Healthy mentions by region Levene's test statistic was nonsignificant, so equal variances were assumed. Significant differences in the mean relative frequency of `Healthy` mentions were found between regions ($F(3, 43) = 4.34$, $p < .01$). Hochberg's post-hoc test again confirmed that the mean relative frequency of `Healthy` mentions from the West was significantly different from both the South ($p < .01$) and the Midwest ($p < .05$).

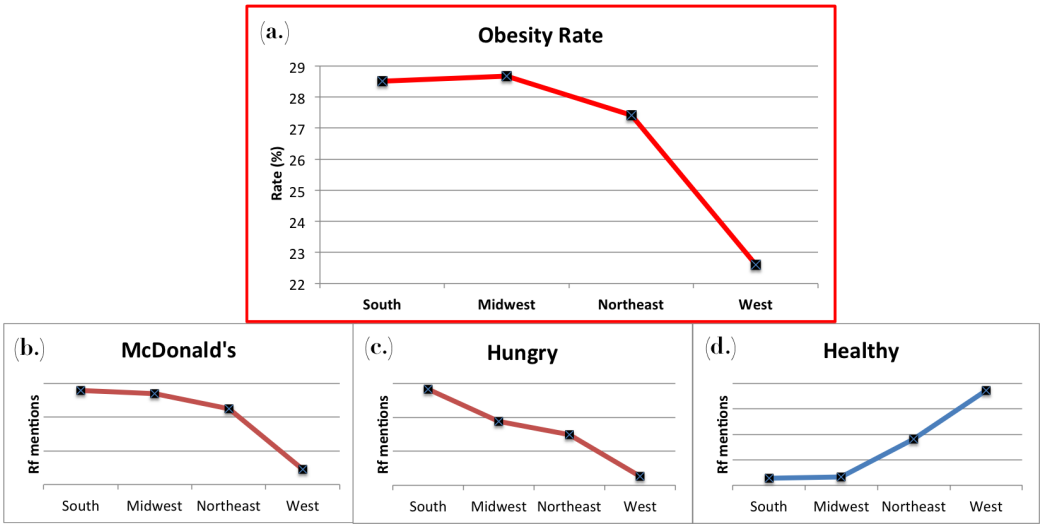


Figure 5.2: Regional Mean Trendlines

Mean trends by region (significantly different in the West, as described above) in a.) obesity rate, b.) rf McDonalds mentions, c.) rf Hungry mentions, and d.) rf Healthy mentions. Trendlines in red were expected to look similar to trends in mean obesity by region, whereas trendlines in blue were expected to display the opposite pattern.

In the mean plots of significant dependent variables above, notice the similarities between the mean trends in obesity by region and `McDonald's` / `Hungry` tweet mentions. Also recognize the differences between mean trends in obesity by region and `Healthy` tweet mentions. (Note: by including line graphs, I do not mean to imply that the x axis, region, is continuous. Trendlines were used among the four regions (in the same order) solely for the purpose of easy visual comparison between variables.)

5.3 Can Twitter classify?

Discriminant Function Analysis Since the mean regional differences from the univariate ANOVAs were significant for mentions of `McDonald's`, `Hungry`, and `Healthy`, I decided to use only these variables in a Discriminant Function Analysis attempting classification by region.

The DFA revealed three discriminant functions. The first explained 85.8% of the variance, canonical $R^2 = .52$; the second explained 13.3% of variance, canonical $R^2 = .14$; and the third explained only 0.9% of the variance, canonical $R^2 = .01$. In combination, the three functions significantly differentiated the regional groups, $\Lambda = 0.41$, $\chi^2(9) = 37.9$, $p < .001$. However, removing the first function indicated that the second and third functions did not significantly differentiate the regions (p ns).

We can see the role of each of the first two discriminant functions ('Function 1' and 'Function 2') in Figure 5.4 on the next page. The first function clearly and successfully differentiated the West from other regions (South, Midwest, and Northeast). Figure 5.3 plots the values of the centroids from each region, highlighting the distance between the Western centroid and the main cluster of remaining regions.

In terms of overall accuracy of classification, **63.8%** of all 47 cases (cities) were correctly classified into their respective regions by only the relative frequencies of `McDonald's`, `Hungry`, and `Healthy` in the text of tweets from the cities. Western cities were classified the most accurately (88.9% of the time), followed by Midwestern cities (66.7%), Southern cities (63.6%), and finally Northeastern cities (28.6%). Table 5.1 below presents all percentages.

Region	South	Midwest	Northeast	West	**Total**
South	63.6%	13.6%	13.6%	9.1%	100.0%
Midwest	22.2%	66.7%	0.0%	11.1%	100.0%
Northeast	14.3%	42.9%	28.6%	14.3%	100.0%
West	0.0%	0.0%	11.1%	88.9%	100.0%

Table 5.1: Predicted Group (Regional) Membership

Generally, then, it seems that Western cities are fairly easily classified as such by the relative frequencies of `McDonald's`, `Hungry`, and `Healthy` mentioned in the text of their tweets. The Southern and Midwestern cities have intermediate classification accuracies based on their tweets, which is perhaps related to the fact that their obesity rates are quite similar in this sample. Northeastern cities were most often classified incorrectly as Midwestern (42.9% of the time), and their centroids are the closest in distance in Figure 5.3. This inaccuracy is discussed in more detail in Chapter 6.

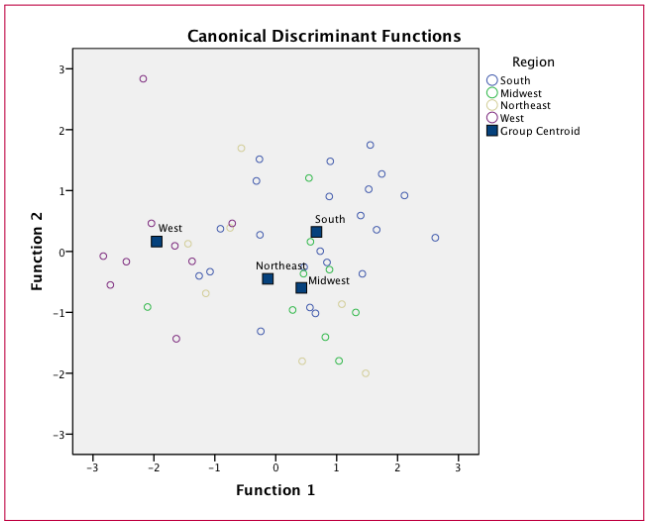


Figure 5.3: Regional centroids

Centroid plot for each of the four color-coded regions. We can see the Western centroid is away from the main cluster containing the centroids of the South, Northeast, and Midwest.

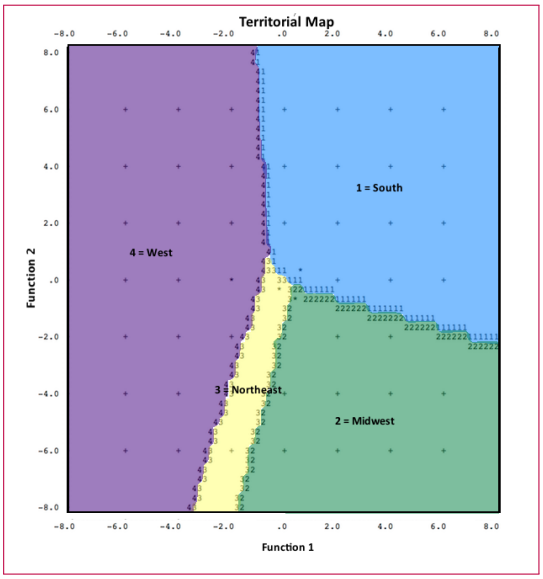


Figure 5.4: Territories by Discriminant Function

Territorial map for the four regions, highlighting the ability of Discriminant Function 1 to clearly differentiate the West from other regions.

5.4 Can Twitter predict?

Given the above examinations of Twitter's descriptive power, I set out to find whether or not obesity rate of a given new city x could be reasonably predicted based on the values of significant independent variables — rf `McDonald's`, rf `Hungry`, and rf `Healthy` — in the text of tweets from the city.

Multiple Linear Regression I carried out a multiple linear regression using the 'Enter' method with obesity rate as the dependent variable and rf `McDonald's`, rf `Hungry`, and rf `Healthy` as my predictor variables. The maximum Cook's distance was .158, meaning that there were no significant outliers in my sample of $n = 47$. VIFs were all less than 10, so it was likely that multicollinearity was not an issue. In addition, the Durbin-Watson statistic was 2.06, indicating that residuals were approximately uncorrelated.

The run produced a significant model for prediction, $F(3, 43) = 16.7$, $p < .001$, with adjusted $R^2 = .506$. Coefficient values for the constant, rf `McDonald's`, and rf `Healthy` only were significant in the overall model, which is presented in the table below:

	B	**t**	**p-value**
Constant	27.35	7.30	.000***
rf `McDonalds`	640.3	2.39	.021*
rf `Hungry`	15.49	.342	.734
rf `Healthy`	-587.9	-3.08	.004**

Table 5.2: Obesity Rate prediction based on Twitter mentions

Overall, the above model was associated with a reasonable amount of standard error of estimation: Given a new city x, and armed with only the relative frequencies of `McDonald's` and `Healthy` mentioned in tweets from the city, this model — calculated with an extremely limited sample of cities and tweets — could estimate an obesity rate within an average of **+/- 2.96** percentage points (95 % CI: -5.80 to +5.80).

5.5 Does distance make a difference?

The following analyses test my second hypothesis that the difference in obesity rate or food mentions between cities would be positively correlated with physical distance. Raw values are from the distance matrices formulated via the methods described in section 4.3.3.

Raw correlations For the raw bivariate correlations between all difference values, I truncated the matrix data at roughly half the maximum distance between cities. This was done both for normalization and also to minimize the exaggerated effects the very few long-distance points would have on the overall correlation coefficients.

Significant positive correlations were found among differences in distance and differences in all other variables between cities (in obesity rate, $r = .178$; rf `McDonald's`, $r = .295$; rf `Hungry`, $r = .190$; and rf `Healthy`, $r = .159$; all $p < .001$). The farther apart cities were from each other, the more likely their obesity rates were farther apart in percentage points, as well as the more likely their values of tweet mentions were farther apart in relative frequency. This in general highlights the change in aspects of the "obesogenic" environment and culture while moving away from any particular city.

Trends by binning To aid in visualizing these trends in differences across distance, I graphed binned data by averaging across 20 samples in categories ordered by distance. (In other words, I arranged my raw difference values in order of increasing distance and averaged all variables in chunks of 20 cases at a time.) No statistical tests were carried out on this binned data, but their graphical results are very much worth examining.

First pictured below (Figure 5.5) are plots of the binned values through the entire range of distances with their best-fitting trendlines, which happened to be quadratic. As shown in the graphs, differences in all variables tend to rise through about 3000 - 4000 kilometers, but then almost appear to drop down again at the longest distances between cities. Locations at these long distances apart would necessarily be points along opposite coasts. This has interesting implications for the spread of obesity-related ideas (to be discussed in Chapter 6).

Next, I again truncated the data halfway for accuracy, zooming in on differences between cities less than 3000 kilometers apart. As seen in the second figure below (Figure 5.6), the data appear to have a flat, uncorrelated segment out to about 1000 kilometers, and then seem to steadily rise to 3000 kilometers:

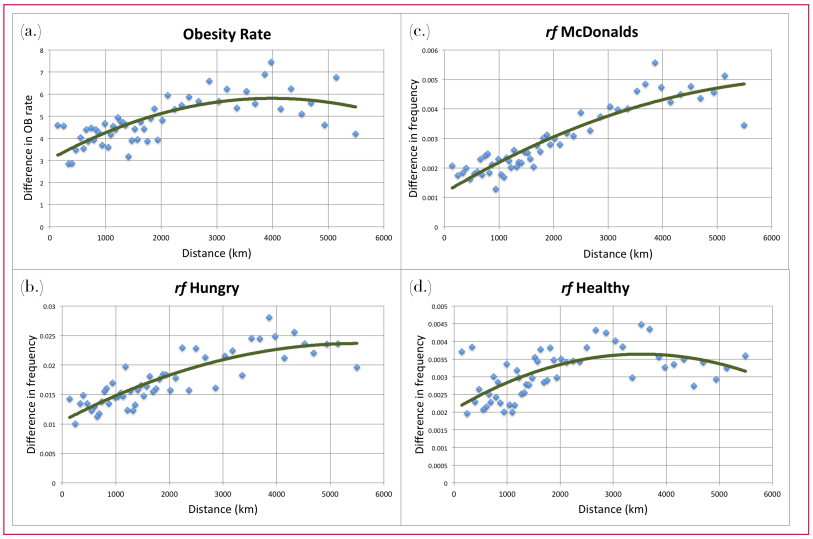


Figure 5.5: Differences by distance - full range

Plots of binned data showing the differences in various variables when distance between cities is increased: (a.) obesity rate difference vs. distance; (b.) rf Hungry difference vs. distance; (c.) rf McDonald's difference vs. distance; and (d.) rf Healthy difference vs. distance.

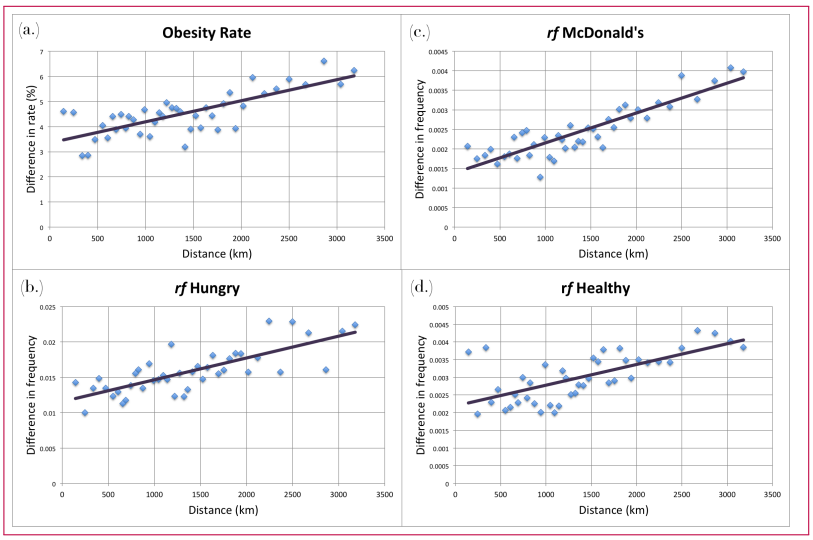
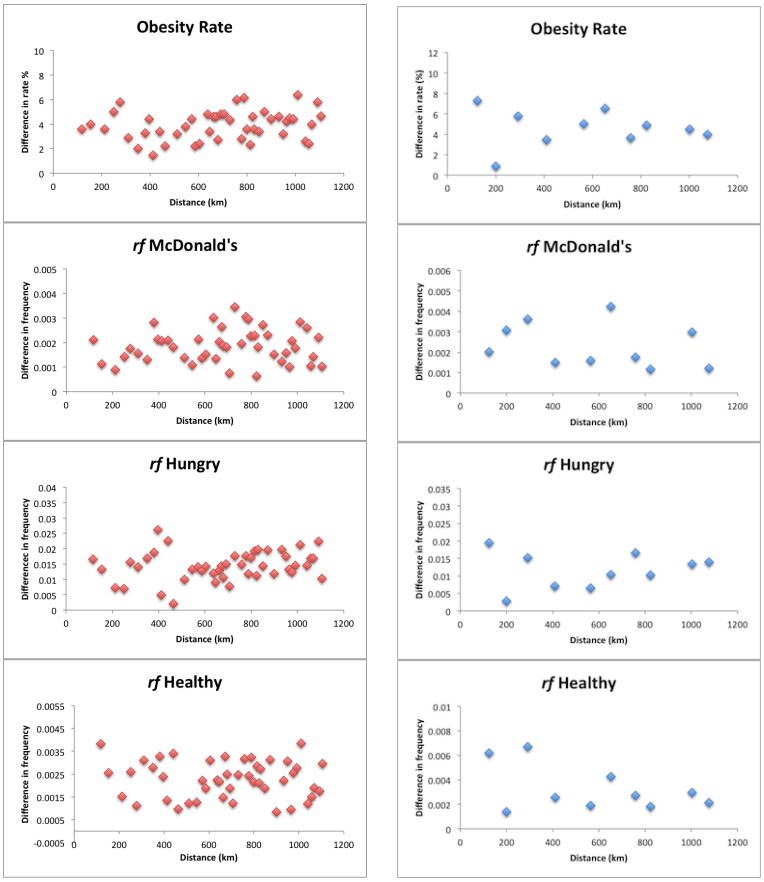


Figure 5.6: Differences by distance - half range

Plots of binned data showing the differences in various variables when distance between cities is increased: (a.) obesity rate difference vs. distance out to 3000km; (b.) rf Hungry difference vs. distance out to 3000km; (c.) rf McDonald's difference vs. distance out to 3000km; and (d.) rf Healthy difference vs. distance out to 3000km.

Finally, to address whether or not the flat-looking trends presented in the first 1000 kilometers of the half-distance set of plots were indeed flat, I graphed this last set to hone in on shorter distances between cities. Given that the first set of full distance plots showed some similarities between cities on opposite coasts, in this figure I separated cities in the far coastal regions (Northeast and West) from those in the Middle (Midwest and South):



(a) Middle cities **(b) Coastal cities**

Figure 5.7: Short distance differences by region

Differences in all variables by distance out to 1000 kilometers between both (a.) cities in the Middle regions of the country and (b.) cities in the coastal regions. Trends are seemingly flat, indicating low levels of difference between cities which are close together.

The trends out to 1000 km above show no correlation in either set of regions, indicating that within 1000 kilometers, cities can be quite similar in obesity prevalence and food tweets. In combination with Figures 5.5 and 5.6, these results point to differences in obesity culture in cities past 1000 km apart, which increase with distance from either coastline.

Chapter 6

Discussion

Overall, this study has shown that the ways in which people in the United States discuss food on Twitter are consistent with corresponding real-world patterns in aspects of both "obesogenic" environment and regional metabolic disease. Results indicated that cities with higher obesity rates are likely to have higher mentions of fast-food-related terms and lower mentions of health-related terms on Twitter. These trends in tweets accurately reflected differences in regional mean obesity rate, and were sufficient to classify by region in a majority of cases. In addition, tweets from the relatively small sample set of cities used here could predict obesity rate with a reasonable amount of error. Finally, graphical representations of differences between cities in all variables examined showed interesting patterns with increasing distance, consistent with a model of food or health idea "transmission" which becomes slower at longer distances from either coast.

6.1 Interpretations & Implications

6.1.1 Twitter as an obesogenic mirror

The analyses conducted above partially support my first hypothesis and overall confirm the use of Twitter as a valid reflection of real-world phenomena.

Significant relationships First, relative mentions of `Hungry` were positively correlated with obesity rate, which could indicate one of two things: either a.) Twitter users are simply more comfortable tweeting about their hunger in obese regions as compared to non-obese, but the hunger levels of both regions are actually equal (a social/cultural phenomenon); or b.) users

are indeed hungrier in regions with higher obesity rates (a biological/food-related phenomenon). A comprehensive examination of intent and culturally-specific content of tweets (rather than simple raw relative word frequencies) would be necessary to distinguish which of these mechanisms is at work in the `Hunger`-obesity rate correlation. If, upon further study, the mechanism was in fact related to a higher level of biologically-based hunger in obese regions, the other significant positive correlation found between relative mentions of `McDonald's` and obesity rate might point to Lustig's fructose-fast-food model as a potential explanation [81]. Are people hungrier in obese regions because the food most easily available to them makes them hungrier?

Apart from a possible higher level of consumption of McDonald's food in obese regions, the differences in number of nearby McDonald's locations by region available to be discussed on Twitter most likely has also played a role in the positive correlations found. Figure 6.1 is a visualization of the continental United States based on distances to the nearest McDonald's. As seen below, the densities of the restaurant are much higher overall in the Eastern half of the country (the South, Northeast, and most of the Midwest as defined by the US Census) than they are in the Western half:

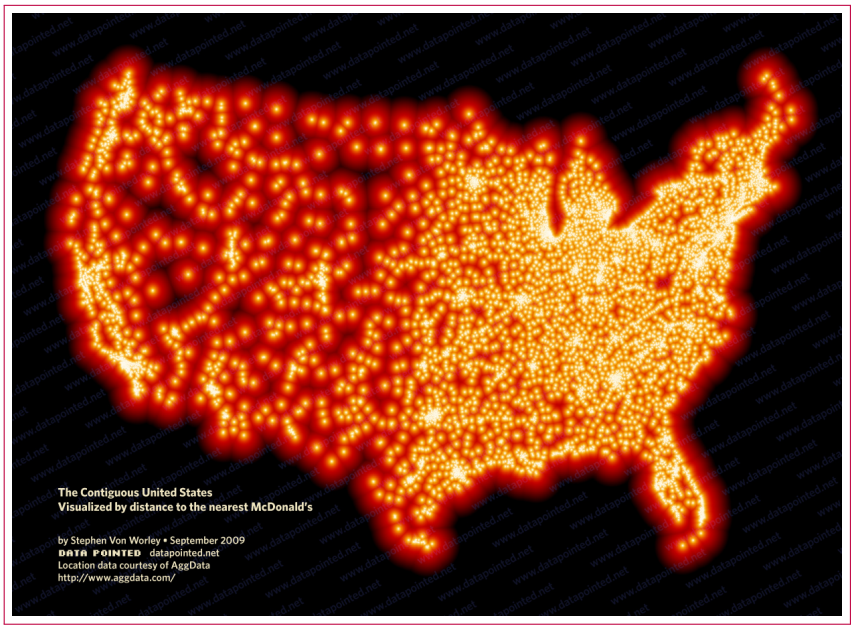


Figure 6.1: Density of nearest McDonald's in the US

A map of the United States constructed by distances to the nearest McDonald's restaurant. Brighter points indicate higher densities.

It is useful to point out here that though the Eastern and Western regions are overall quite different in terms of McDonald's densities, the cities I have chosen are the most populated in all areas. Thus, the McDonald's densities of the cities in my sample are all represented by very bright lights in the above map. This could potentially indicate that the amount of fast food available in every city examined here was in fact approximately equal, yet we still observe higher mentions of fast food in the cities which happened to be more obese. In either case, whether higher fast food mentions are due to food preference/consumption or food availability, the argument for Twitter as a reflection of real-world obesogenic environment is upheld.

The significant negative correlation found between `Healthy` and obesity rate could similarly be interpreted in multiple ways. Either people in less obese regions are in fact eating the healthy options they tend to tweet about (and are overall more concerned about their health), or people in these regions simply want their followers to *believe* that they are more concerned about health. Again, whether the correlation marks an accurate eating preference difference or a social/cultural difference cannot be determined here, but the combined effects of both cases can be observed simply by tallying relative frequencies of terms.

The last significant correlation found — the negative relationship between `Soda` mentions and obesity rate — is a surprising one; however, it could potentially be explained by either a.) aspects of my processing scripts, or b.) recent soda-related events. Firstly, my scripts analyzed counts by single word, and it is reasonable to suggest that some 'soda', 'coke', or 'sprite' mentions might have been accompanied by a 'diet' modifier: diet soda, diet coke, and diet sprite would have simply been counted as tallies for their full-calorie, full-fructose counterparts. The second possibility involves New York City's Mayor Bloomberg, who proposed a city-wide super-size soda ban in May 2012 [92]. It is possible that hikes in soda mentions could be due to the controversy surrounding the ban. Because of these confounding factors, I hesitate to interpret the negative soda relationship as an accurate representation of metabolic phenomena. More about this and other limitations is discussed in Section 6.3.

Nonsignificant relationships It is interesting to note that while one aspect of local food environment and choice — fast food — was significantly related to regional obesity rate, mentions of its counterpart — grocery stores — were not. This could be explained by the fact that there are still just as many unhealthy items as there are healthy items to choose from in grocery stores, whereas the reverse is not necessarily true; though large-chain

fast food outlets are attempting to incorporate healthier items into their menus, in general they are still dominated by energy-dense, high fructose options, and at times even the 'healthier' options are not quite as healthy as consumers might initially believe [93]. Since the options at grocery stores are much more variable in terms of nutritive quality, regional mentions of groceries might provide a much less accurate measure of differences in food choice.

The second nonsignificant relationship between `Candy` and obesity rate could potentially be explained by the ways in which people refer to the food they are eating. After conducting this experiment, it occurred to me that perhaps more often than not people tend to use the actual name or brand of the candy they are eating rather than the generic term 'candy' itself. This makes sense particularly given the Twitter environment — tweets are meant for others to read, and unique ideas about specific foods being eaten are probably more interesting and discuss-able than those which keep terms as generic as possible.

Regional divisions Notably, the stark regional differences in obesity rate seen in the literature (see Chapter 2: South > Midwest > Northeast > West) are not fully displayed by the measurements from cities I have chosen. In my results, only the West consistently displayed a significantly different mean of obesity rate. This is perhaps due to the nature of my sample: though it was necessary via my methodology for me to pick the largest cities in each region (to acquire a suitably comparable number of tweets - see Section 6.3.1), these large cities themselves are often are not perfect representatives of their respective regions as a whole. However, it is important (given my argument) only that results from Twitter can reflect the differences that do exist among the cities chosen here (with less emphasis on the extent of the differences themselves). Judging by the simple facts that a.) the trendline of mean `McDonald's` mentions by region fits almost the exact same shape as the curve of obesity rate by region (Figure 5.2.1), b.) `Hungry` mentions follow a highly similar pattern, and c.) `Healthy` mentions follow a distinctly opposite pattern, Twitter's reflective capabilities are upheld by region.

Classifications Given a set of tweets — and armed with only their relative frequencies of `McDonald's, Hungry, and Healthy` — an overall regional classification accuracy of 63.8% points to Twitter's usefulness as a sorting tool: people from different areas do generally tweet about different things (or similar things at different relative frequencies). The only region with a higher percentage of innacurately-classified relative to accurately-classified

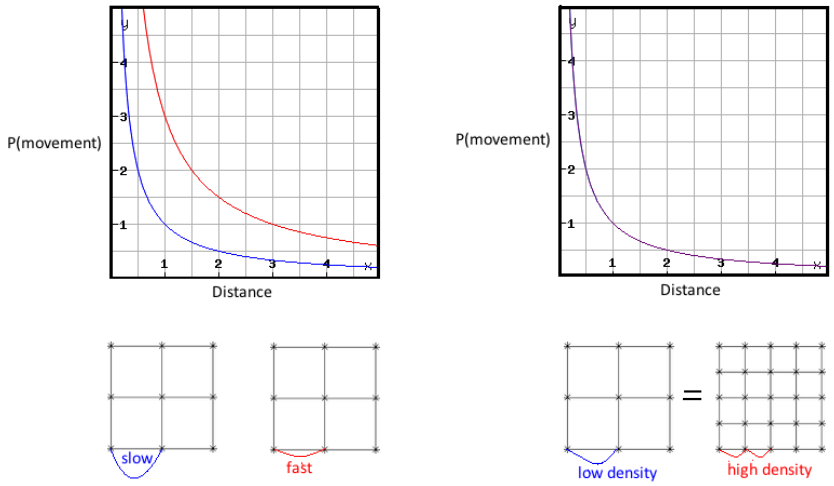
cases was the Northeast, whose cities were most often grouped as Midwestern by the discriminate functions. I attribute this inaccuracy to the fact that nearly half of the cities in the CDC's 'Northeast' region from this sample — the western Pennsylvanian cities in particular — are in areas directly adjacent to Midwestern states. This serves as a good reminder that while regional boundaries are present, the set range of people on either side of those boundaries is really more of a cultural spectrum than a distinct and sharply-cut line.

Prediction At this point, with my extremely limited city sample of 47, the rate of accuracy for predicting regional obesity rate given only the tweets from the area (+/- 2.96%) is reasonable, but might not be actually more useful than the methods currently put in place by organizations such as the CDC. However, with a higher level of access to geo-tagged tweets and a wider range of cities examined throughout the country (see Section 6.3), I would anticipate that the level of prediction accuracy would become just as high (if not higher) than traditional surveying methods.

6.1.2 The culture-distance spectrum

The last set of graphical representations presented in Chapter 5 reviewed differences in obesity rates and food mentions as a function of continuous distance (rather than comparison between discrete regional chunks). The general trends seem to support my second hypothesis, which posited that cultural differences between cities would rise with increasing distance. Plots revealed similarities between cities less than 1000 kilometers apart, with a steady increase in cultural/food differences between 1000 - 3000 kilometers apart. One aspect of the graphs that was not explicitly predicted — but perhaps should have been — was the final drop in cultural differences seen between cities that were the longest distances away from each other (necessarily the opposite coasts). If these trends in differences with distance from either coastline could be verified at a higher level of detail, they could hold strong implications for the spatial transmission of health- and food-related ideas.

If, for interpretation, we assume a traditional gravity model of movement of individuals between cities, we could begin to speculate why differences in food ideas might increase with distance from either coast. This traditional type of model presents the probability of moving away from a current position as an inverse function of distance, and is a well-known approach to both mobility and infectious disease transmission [94][95]. When applied to the transmission of food ideas and obesity culture here, the differences between



(a) Individual speed **(b) Density of cities**

Figure 6.2: Potential gravity models of information diffusion

Probability of movement between cities in coastal regions (red), middle regions (blue), or both regions (purple). In part a.), probabilities of individual/personal movement (and thus idea movement) between cities in the coastal regions are higher than the individual rates in the middle regions, resulting in faster transmission. In part b.), probabilities are equal, but the density of cities in coastal regions is higher, and thus faster transmission can still occur.

coastal and middle regions of the country might be explained by one of two applications of the gravity model. Either a.) the people themselves are simply more mobile in coastal regions, and thus might transmit ideas about food and health to each other at a much faster rate than those in the middle regions, showing higher probabilities of moving longer distances (as seen in part a. of Figure 6.2), or b.) the movement functions of both regions are similar, but the densities of people and cities available to move to are simply higher in the coastal regions relative to the middle regions (part b.).

Could the similarities of the coastal regions found in the parabolas of Figure 5.5 be explained by more mobile individuals (and thus more mobile health ideas)? Or could they be caused by a higher population and city density on the coasts relative to the middle portion of the United States? In looking at a recent population density map from the US Census Bureau (2010), we see the highest population densities in the far Western and Northeastern regions, potentially consistent with a model of ideological mobility dependent on density:

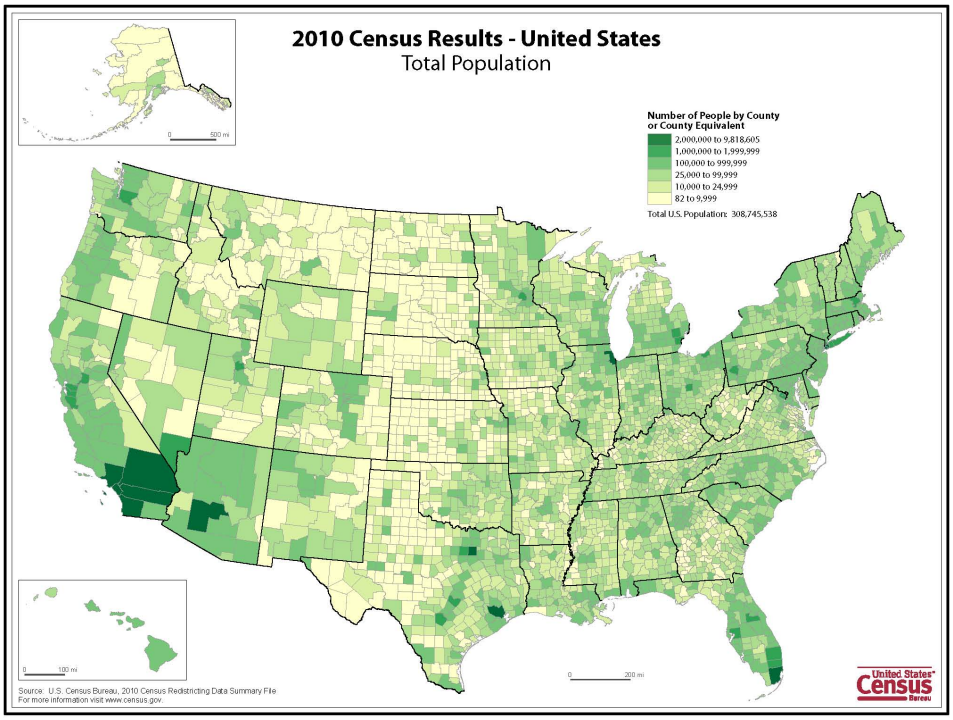


Figure 6.3: U.S. Regional Population Density

2010 population density by county of the United States, available from 2010.census.gov. Counties on the far Northeast and Western coasts have higher population densities relative to those away from the coasts.

This could point to explanation (b.) in Figure 6.2 as the correct option. However, in either case, discussion of food ideas on Twitter tends to be similar in cities less than 1000 kilometers apart, then rises steadily until the maximum distance from either coast. This interesting result is worthy of its own set of future studies.

6.2 Applications

Twitter's metabolically-reflective capabilities described above — garnered from the tweets of only 47 cities — point to its potential usefulness as a more widespread tool in two realms of the public health sector.

6.2.1 For disease monitoring

In considering the results of the first four statistical analyses presented in Chapter 5, it is clear that differences in food discussion by region can be monitored externally via user updates on Twitter. In an extraction of merely 1% of tweets over a week-long period, my uncut sample contained over 38 million unique updates mentioning some type of food. Ostensibly, with higher access levels and wider location parameters, public health researchers could have billions of unique food tweets per week at their disposal. With an estimated 300,000 new Twitter users per day [96], the wealth of eating information publicly available in the form of online discussion will continue to rise. Publicly tracking these millions of updates per day about food eaten, desired, or available in various regions could become a new tool for public health officials in understanding the growth of metabolic disease and identifying future "problem regions" where certain food mentions might be on the rise.

These monitoring capabilities become even more useful when considering the demographics of Twitter users. Surprisingly, Twitter has higher percentages of low-income users than other popular online social networks, such as Facebook or LinkedIn [97][98]. It was estimated that nearly 17% of Twitter users were from families making less than $25,000.00/year [98]. As more low-income households continue to gain Internet access (and, increasingly, mobile Internet access [97]), the percentage of those using Twitter will undoubtedly also increase. Monitoring the everyday food choices of this low-income group in real-time could become especially valuable given the associations between poverty and obesity reviewed in Chapter 2.

6.2.2 For health idea transmission

As we saw in Chapter 2, local Twitter user networks are more dense than non-local. The Internet has surprisingly not cleared all of the effects of physical distance, and thus it is still crucial to consider distances and the spread of ideas between users on a macroscopic scale. However, in addition to this examination of geographical implications, a consideration of pure virtual information diffusion would also aid in our interpretation of any health results from Twitter [99]. With a deeper understanding of how food information is passed through the social web, public health officials could harness its spreading power in order to broadcast any important health messages.

In particular, I might suggest attempting to use the "super-spreading" positions of celebrities on the Twitter network to break geographical food-idea boundaries by having them tweet more frequently about healthy foods.

Through their high connectivity levels and the re-tweets of their followers, celebrities could feasibly begin to fix regional disparities in food tweets and even possibly influence their admirers to make healthier choices. The key to this method would be the un-sponsored nature of a celebrity's tweet: a tweet is an example of an everyday, 'normal', direct update from an admired figure, and might serve as a more intimate reminder to followers that food choice is important.

6.3 Limitations

6.3.1 Sample

Perhaps the most blatant limitation in the analyses presented here involves the set of cities considered. Because as a public user I was granted access to only 1% of public tweets via the API, and also since many users do not list their location willingly, I was somewhat forced into using tweets from Twitter's top trending cities so as to have enough per city for comparison amongst them. These large cities have relatively high populations and business concentrations, and hence also have a higher percentage of new residents originally from other areas. They are not evenly distributed throughout the continental United States, and the sets are of different sizes per region. Counties holding the highest obesity rates happened to not be included in this city set from Twitter. Therefore, they most likely present a less accurate representation of their respective regions as a whole. An ideal analysis would have more Twitter access privileges, with a higher percentage of overall sample and more specific geo-location identifiers. That way, the entire range of county obesity rates and national distribution would be accounted for.

6.3.2 Missing information

In simply tallying relative frequencies, my analyses missed out on information about the intent of user messages. Relative frequencies did not distinguish between mere mentions of fast food and proclamations of actually eating said fast food. In addition, with my simple Python text counting methods, some of the words analyzed could have been accompanied by modifiers which alter their meaning. (As an example, some of the words I've discussed above — such as 'candy' or 'soda' — could have been accompanied by modifiers like '*eye* candy' or '*baking* soda', and would still have been tallied as their nutritional counterparts.) With a longer time span and higher computing power, machine-learning algorithms could potentially help detect the semantic dif-

ferences in each tweet and hone in on those explicitly discussing food being eaten (or desired to be eaten).

6.3.3 Bias

The third most noticeable limitation is psychological in nature. Are users being honest in their posts? Are people more likely to share things that they would want other people to read? Aspects of "self categorization" theory [100] would predict that tweets might be inherently geared towards an audience of admired followers, and thus might not necessarily be as open in mentioning foods the users deem as unhealthy. However, considering the sheer volume of tweets I received about fast food (particularly in comparison to those about healthy foods), I would actually categorize this limitation as a very minor one.

6.4 Future Work

It is clear even from this very basic analysis that there exists an abundance of information about eating patterns on the social web; however, to the best of my knowledge, this study is the very first to address the public food conversation being held on Twitter in light of metabolic disease prediction and potential public health usages. Future studies might first improve upon the methodologies used here. With higher computing capabilities, access to the full Twitter Firehose, a longer timespan, and a fuller set of geo-tagged city locations, much more accurate trends in food ideology and eating patterns could be captured. In addition, machine-learning techniques would prove to be extremely useful in discerning the intent of food-related tweets to distinguish those merely mentioning food from those about eating said food. More detailed queries keeping these improvements in mind would produce better representations of the differences in food culture behind the obesity epidemic.

Secondly, in the area of information diffusion, previous studies have very much focused on the spreading trajectories of popular news stories or other nationally- and globally- widespread links. Understanding more about the diffusion of local opinion and personal updates on Twitter — not simply in the United States, but through regions worldwide — would undoubtedly bring forth more culturally-specific information to aid in the interpretation of ideological differences in any realm. Additional examinations of the static qualities of local environments and how they correspond to the types of tweets displayed by users would shed light on the external factors behind these differences.

Thirdly, an understanding of online influence — beyond the simple algorithms used by web applets today (such as the popular "Klout") — would be useful in determining the potential for health information to stand out amidst the flood of conversations on the web. Finding universal patterns in influential users would be instrumental in identifying the best places in social networks to insert health messages. In addition, future studies might consider the ways in which online influence translates into real-world behavioral influence (via survey or other methods).

Finally, Twitter is not the only online social network in which the people of the world discuss their everyday lives. Partnerships between public health or research organizations and other big-name online social networks — Facebook, Google+, Tumblr, etc. — will prove to be practical and powerful in coming years, as Internet access gradually spreads around the globe.

***

The Internet of today is the largest warehouse of human-related (and human-generated) information in existence. The continual conversation about food on the social web in particular makes it an invaluable and incredibly convenient resource for studying food-related disease, and it is my sincere hope that in the near future, health researchers begin paying more attention to it.

Chapter 7

Bibliography

[1] R. Schein, K. Wilson, and J. Keelan, "Literature Review on Effectiveness of the Use of Social Media: A Report for Peel Public Health," Tech. Rep. 1, Carleton University, 2011.

[2] T. O'Reilly, "What Is Web 2.0?," 2005.

[3] E. A. Krall, J. T. Dwyer, and K. Ann Coleman, "Factors influencing accuracy of dietary recall," *Nutrition Research*, vol. 8, pp. 829–841, July 1988.

[4] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10, Jan. 2010.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *WWW Conference*, pp. 851–860, 2010.

[6] A. Hermida, "From TV to Twitter: How Ambient News Became Ambient Journalism," *Media Culture Journal*, vol. 13, no. 2, 2010.

[7] The Hartman Group, "Culture of Millenials," tech. rep., 2011.

[8] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, ACM, 2007.

[9] S. Macskassy, "On the Study of Social Interactions in Twitter," *Association for the Advancement of Artificial Intelligence*, 2012.

[10] H. Kwak, C. Lee, and H. Park, "What is Twitter, a Social Network or a News Media?," *Proceedings of the 19th international conference on the World Wide Web*, pp. 591–600, 2010.

[11] S. Milgram, "The Small World Problem," *Psychology Today*, vol. 1, pp. 61 – 67, 1967.

[12] A. Mislove, M. Marcon, and K. Gummadi, "Measurement and analysis of online social networks," *Internet measurement*, pp. 29–42, 2007.

[13] J. Yang and S. Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 355–358, 2010.

[14] S. Yardi and D. Boyd, "Tweeting from the Town Square: Measuring Geographic Local Networks," *Proceedings of the International Conference on Weblogs and Social Media*, pp. 194–201, 2010.

[15] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of Twitter networks," *Social Networks*, vol. 34, pp. 73–81, Jan. 2012.

[16] F. Cairncross, *The Death of Distance: How the Communications Revolution Will Change Our Lives*. Harvard Business Press, 1997.

[17] J. Ritterman and M. Osborne, "Using prediction markets and Twitter to predict a swine flu pandemic," *Workshop on Mining*, pp. 1–9, 2009.

[18] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169–2188, 2009.

[19] A. Tumasjan, T. Sprenger, and P. Sandner, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185, 2010.

[20] P. Beaumont, "The truth about Twitter, Facebook and the uprisings in the Arab world," 2011.

[21] D. Scanfeld, V. Scanfeld, and E. L. Larson, "Dissemination of health information through social networks: twitter and antibiotics.," *American journal of infection control*, vol. 38, pp. 182–8, Apr. 2010.

[22] A. Sadilek and H. Kautz, "Modeling Spread of Disease from Social Interactions," *Artificial Intelligence*, 2012.

[23] World Health Organization, "Global Database on Body Mass Index," tech. rep., 2012.

[24] A. H. Mokdad, "The Spread of the Obesity Epidemic in the United States, 1991-1998," *JAMA: The Journal of the American Medical Association*, vol. 282, pp. 1519–1522, Oct. 1999.

[25] Y. Wang and M. a. Beydoun, "The obesity epidemic in the United States–gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis.," *Epidemiologic reviews*, vol. 29, pp. 6–28, Jan. 2007.

[26] Y. Wang, M. A. Beydoun, L. Liang, B. Caballero, and S. K. Kumanyika, "Will all Americans become overweight or obese? estimating the progression and cost of the US obesity epidemic.," *Obesity*, vol. 16, pp. 2323–30, Oct. 2008.

[27] K. M. Flegal, M. D. Carroll, C. L. Ogden, and L. R. Curtin, "Prevalence and Trends in Obesity Among US Adults, 1999-2008," *JAMA: The Journal of the American Medical Association*, vol. 303, pp. 235–241, Jan. 2010.

[28] F. Li, P. Harmer, B. Cardinal, M. Bosworth, and D. Johnson-Shelton, "Obesity and the Built Environment: Does the Density of Neighborhood Fast-Food Outlets Matter?," *American Journal of Health Promotion*, vol. 23, no. 3, pp. 203–209, 2009.

[29] Centers for Disease Control and Prevention, "National Health and Nutrition Examination Survey," 2012.

[30] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System," 2012.

[31] The Child & Adolescent Health Measurement Initiative, "National Survey of Children's Health," 2012.

[32] G. K. Singh, M. D. Kogan, and P. C. van Dyck, "A multilevel analysis of state and regional disparities in childhood and adolescent obesity in the United States.," *Journal of community health*, vol. 33, pp. 90–102, Apr. 2008.

[33] Centers for Disease Control and Prevention, "Obesity and Overweight for Professionals: Data and Statistics: Data, Maps, and Trends," tech. rep., 2012.

[34] J. Sobal and A. J. Stunkard, "Socioeconomic status and obesity: A review of the literature.," *Psychological Bulletin*, vol. 105, no. 2, pp. 260–275, 1989.

[35] Q. Zhang and Y. Wang, "Socioeconomic inequality of obesity in the United States: do gender, age, and ethnicity matter?," *Social Science & Medicine*, vol. 58, pp. 1171–1180, Mar. 2004.

[36] C. L. Ogden, M. M. Lamb, M. D. Carroll, and K. M. Flegal, "Obesity and socioeconomic status in adults: United States, 2005-2008.," *NCHS data brief*, vol. 127, pp. 1–8, Dec. 2010.

[37] S. Paeratakul, J. Lovejoy, D. Ryan, and G. Bray, "The relation of gender, race and socioeconomic status to obesity and obesity comorbidities in a sample of US adults.," *International Journal of Obesity and Related Metabolic Disorders*, vol. 26, no. 9, pp. 1205–1210, 2002.

[38] N. Cossrow and B. Falkner, "Race/ethnic issues in obesity and obesity-related comorbidities.," *The Journal of clinical endocrinology and metabolism*, vol. 89, pp. 2590–4, June 2004.

[39] I. Kawachi, B. P. Kennedy, and R. Glass, "Social capital and self-rated health: a contextual analysis.," *American Journal of Public Health*, vol. 89, pp. 1187–93, Aug. 1999.

[40] D. Kim, S. V. Subramanian, S. L. Gortmaker, and I. Kawachi, "US state- and county-level social capital in relation to obesity and physical inactivity: a multilevel, multivariable analysis.," *Social science & medicine (1982)*, vol. 63, pp. 1045–59, Aug. 2006.

[41] P. Dubbert, T. Carithers, A. Sumner, K. Barbour, B. Clark, J. Hall, and E. Crook, "Obesity, physical inactivity, and risk for cardiovascular disease," *American journal of the Medical Sciences*, vol. 324, no. 3, pp. 116–126, 2002.

[42] A. H. Mokdad, "The Continuing Epidemics of Obesity and Diabetes in the United States," *JAMA: The Journal of the American Medical Association*, vol. 286, pp. 1195–1200, Sept. 2001.

[43] H. Burdette, "Neighborhood playgrounds, fast food restaurants, and crime: relationships to overweight in low-income preschool children," *Preventive Medicine*, vol. 38, pp. 57–63, Jan. 2004.

[44] M. Stafford, S. Cummins, A. Ellaway, A. Sacker, R. D. Wiggins, and S. Macintyre, "Pathways to obesity: identifying local, modifiable determinants of physical activity and diet.," *Social science & medicine (1982)*, vol. 65, pp. 1882–97, Nov. 2007.

[45] E. Hughes, M. McCracken, H. Roberts, A. H. Mokdad, B. Valluru, R. Goosdon, E. Dunn, L. Elam-Evans, W. Giles, and R. Jiles, "Surveillance for Certain Health Behaviors Among States and Selected Local Areas - BRFSS, United States, 2004," tech. rep., 2004.

[46] US Census Bureau, "Household Income Inequality Within U.S. Counties: 2006 2010," Tech. Rep. February, 2012.

[47] A. Cheadle, B. M. Psaty, S. Curry, E. Wagner, P. Diehr, T. Koepsell, and A. Kristal, "Community-level comparisons between the grocery store environment and individual dietary practices," *Preventive Medicine*, vol. 20, pp. 250–261, Mar. 1991.

[48] C. Chung, "Do the Poor Pay More for Food? An Analysis of Grocery Store Availability and Food Price Disparities," *Journal of consumer affairs*, vol. 33, no. 2, pp. 276–296, 1999.

[49] K. Morland, S. Wing, A. Diez Roux, and C. Poole, "Neighborhood characteristics associated with the location of food stores and food service places," *American Journal of Preventive Medicine*, vol. 22, pp. 23–29, Jan. 2002.

[50] K. Morland, A. V. Diez Roux, and S. Wing, "Supermarkets, other food stores, and obesity: the atherosclerosis risk in communities study.," *American journal of preventive medicine*, vol. 30, pp. 333–9, Apr. 2006.

[51] S. Cummins and S. Macintyre, "Food environments and obesity–neighbourhood or nation?," *International journal of epidemiology*, vol. 35, pp. 100–4, Feb. 2006.

[52] L. M. Powell, S. Slater, D. Mirtcheva, Y. Bao, and F. J. Chaloupka, "Food store availability and neighborhood characteristics in the United States.," *Preventive medicine*, vol. 44, pp. 189–95, Mar. 2007.

[53] S. Nielsen, "Trends in Food Locations and Sources among Adolescents and Young Adults," *Preventive Medicine*, vol. 35, pp. 107–113, Aug. 2002.

[54] B.-H. Lin, E. Frazao, and J. F. Guthrie, "Away-From-Home Foods Increasingly Important to Quality of American Diet," tech. rep., 1999.

[55] B. Swinburn, I. Caterson, J. C. Seidell, and W. P. T. James, "Diet, nutrition and the prevention of excess weight gain and obesity," *Public health nutrition*, vol. 7, pp. 123–146, Feb. 2004.

[56] L. M. Powell, F. J. Chaloupka, and Y. Bao, "The availability of fast-food and full-service restaurants in the United States: associations with neighborhood characteristics.," *American journal of preventive medicine*, vol. 33, pp. S240–5, Oct. 2007.

[57] R. W. Jeffery and S. A. French, "Epidemic obesity in the United States: are fast foods and television viewing contributing?," *American Journal of Public Health*, vol. 88, no. 2, pp. 277–280, 1998.

[58] M. A. Pereira, A. I. Kartashov, C. B. Ebbeling, L. Van Horn, M. L. Slattery, D. R. Jacobs, and D. S. Ludwig, "Fast-food habits, weight gain, and insulin resistance (the CARDIA study): 15-year prospective analysis.," *Lancet*, vol. 365, no. 9453, pp. 36–42, 2005.

[59] K. B. Morland and K. R. Evenson, "Obesity prevalence and the local food environment.," *Health & place*, vol. 15, pp. 491–5, June 2009.

[60] M. Pollan, *In defense of food: an eater's manifesto*. Penguin Press, 2008.

[61] Danforth, Feierabend, and Chassman, *Culinaria: The United States - A Culinary Discovery*. 1998.

[62] M. F. Nenes, *American Regional Cuisine*. 2006.

[63] B. Swinburn, G. Sacks, and E. Ravussin, "Increased food energy supply is more than sufficient to explain the US epidemic of obesity.," *The American journal of clinical nutrition*, vol. 90, pp. 1453–6, Dec. 2009.

[64] A. M. Prentice and S. A. Jebb, "Obesity in Britain: gluttony or sloth?," *BMJ*, vol. 311, pp. 437–439, Aug. 1995.

[65] J. C. K. Wells and M. Siervo, "Obesity and energy balance: is the tail wagging the dog?," *European journal of clinical nutrition*, vol. 65, pp. 1173–89, Nov. 2011.

[66] T. Mann, A. J. Tomiyama, E. Westling, A.-M. Lew, B. Samuels, and J. Chatman, "Medicares Search for Effective Obesity Treatments: Diets are Not the Answer," *American Psychologist*, vol. 62, no. 3, pp. 220–233, 2007.

[67] A. Drewnowski, "Obesity and the food environment: dietary energy density and diet costs.," *American journal of preventive medicine*, vol. 27, pp. 154–62, Oct. 2004.

[68] S. Eaton and L. Cordain, "Evolutionary aspects of diet: old genes, new fuel," *World Review of Nutrition and Dietetics*, pp. 26–37, 1997.

[69] S. B. Eaton, "The ancestral human diet: what was it and should it be a paradigm for contemporary nutrition?," *Proceedings of the Nutrition Society*, vol. 65, pp. 1–6, Mar. 2007.

[70] S. B. Eaton, B. I. Strassman, R. M. Nesse, J. V. Neel, P. W. Ewald, G. C. Williams, A. B. Weder, S. B. Eaton, S. Lindeberg, M. J. Konner, I. Mysterud, and L. Cordain, "Evolutionary health promotion.," *Preventive medicine*, vol. 34, pp. 109–18, Feb. 2002.

[71] G. A. Bray, S. J. Nielsen, and B. M. Popkin, "Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity," *Am J Clin Nutr*, vol. 79, no. 4, pp. 537–543, 2004.

[72] R. A. Forshee, M. L. Storey, D. B. Allison, W. H. Glinsmann, G. L. Hein, D. R. Lineback, S. A. Miller, T. A. Nicklas, G. A. Weaver, and J. S. White, "A critical examination of the evidence relating high fructose corn syrup and weight gain.," *Critical reviews in food science and nutrition*, vol. 47, pp. 561–82, Jan. 2007.

[73] E. Isganaitis and R. H. Lustig, "Fast food, central nervous system insulin resistance, and obesity.," *Arteriosclerosis, thrombosis, and vascular biology*, vol. 25, pp. 2451–62, Dec. 2005.

[74] S. H. Cha, M. Wolfgang, Y. Tokutake, S. Chohnan, and M. D. Lane, "Differential effects of central fructose and glucose on hypothalamic malonyl-CoA and food intake.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 16871–5, Nov. 2008.

[75] K. L. Teff, S. S. Elliott, M. Tschöp, T. J. Kieffer, D. Rader, M. Heiman, R. R. Townsend, N. L. Keim, D. D'Alessio, and P. J. Havel, "Dietary fructose reduces circulating insulin and leptin, attenuates postprandial suppression of ghrelin, and increases triglycerides in women.," *The Journal of clinical endocrinology and metabolism*, vol. 89, pp. 2963–72, June 2004.

[76] C.-Y. O. Chen, J. Crott, Z. Liu, and D. E. Smith, "Fructose and saturated fats predispose hyperinsulinemia in lean male rat offspring.," *European journal of nutrition*, vol. 49, pp. 337–43, Sept. 2010.

[77] H. Jürgens, W. Haass, T. R. Castañeda, A. Schürmann, C. Koebnick, F. Dombrowski, B. Otto, A. R. Nawrocki, P. E. Scherer, J. Spranger, M. Ristow, H.-G. Joost, P. J. Havel, and M. H. Tschöp, "Consuming fructose-sweetened beverages increases body adiposity in mice.," *Obesity research*, vol. 13, pp. 1146–56, July 2005.

[78] M. E. Bocarsly, E. S. Powell, N. M. Avena, and B. G. Hoebel, "High-fructose corn syrup causes characteristics of obesity in rats: increased body weight, body fat and triglyceride levels.," *Pharmacology, biochemistry, and behavior*, vol. 97, pp. 101–6, Nov. 2010.

[79] I. Hwang, H. Ho, B. Hoffman, and G. Reaven, "Fructose-induced insulin resistance and hypertension in rats," *Hypertension*, vol. 10, pp. 512–516, 1987.

[80] R. J. Johnson, M. S. Segal, Y. Sautin, T. Nakagawa, D. I. Feig, D.-H. Kang, M. S. Gersch, S. Benner, and L. G. Sanchez-Lozada, "Potential role of sugar (fructose) in the epidemic of hypertension, obesity and the metabolic syndrome, diabetes, kidney disease, and cardiovascular disease," *Am J Clin Nutr*, vol. 86, pp. 899–906, Oct. 2007.

[81] R. H. Lustig, "Childhood obesity: behavioral aberration or biochemical drive? Reinterpreting the First Law of Thermodynamics.," *Nature clinical practice. Endocrinology & metabolism*, vol. 2, pp. 447–58, Aug. 2006.

[82] K. J. Melanson, L. Zukley, J. Lowndes, V. Nguyen, T. J. Angelopoulos, and J. M. Rippe, "Effects of high-fructose corn syrup and sucrose consumption on circulating glucose, insulin, leptin, and ghrelin and on appetite in normal-weight women.," *Nutrition (Burbank, Los Angeles County, Calif.)*, vol. 23, pp. 103–12, Feb. 2007.

[83] L. Tappy and K. Lê, "Metabolic effects of fructose and the worldwide increase in obesity," *Physiological reviews*, vol. 90, no. 1, pp. 23–46, 2010.

[84] N. a. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years.," *The New England journal of medicine*, vol. 357, pp. 370–9, July 2007.

[85] E. Cohen-Cole and J. M. Fletcher, "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic.," *Journal of health economics*, vol. 27, pp. 1382–7, Sept. 2008.

[86] T. Developers, "Streaming API Methods — Twitter Developers," 2012.

[87] J. McLaughlin, "6 Twitter Hashtags Every Restaurant Should Use," *sproutsocial*, 2011.

[88] C. Matyszczyk, "Almost half of millennials tweet while they eat, survey says," *Cnet News*, 2012.

[89] B. Hecht, L. Hong, and B. Suh, "Tweets from Justin Biebers Heart: The Dynamics of the Location Field in User Profiles," *CHI*, pp. 237–246, 2011.

[90] Centers for Disease Control and Prevention, "National Vital Statistics System Homepage," 2012.

[91] University of Wisconsin PHI, "County Health Rankings," 2012.

[92] H. Goldman and D. D. Stanford, "NYC Mayor Bloomberg Seeks Ban on Super-Size Soft Drinks," 2012.

[93] M. Berman, "Obesity prevention in the information age," *JAMA: the journal of the American Medical Association*, vol. 355, no. 3, pp. 2006–2008, 2008.

[94] G. J. Boender, T. J. Hagenaars, A. Bouma, G. Nodelijk, A. R. W. Elbers, M. C. M. de Jong, and M. van Boven, "Risk maps for the spread of highly pathogenic avian influenza in poultry.," *PLoS computational biology*, vol. 3, p. e71, Apr. 2007.

[95] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, 2009.

[96] K. Bodnar, "The Ultimate List: 100+ Twitter Statistics," 2010.

[97] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, "Social Media & Mobile Internet Use Among Teens and Young Adults," 2010.

[98] Digitalsurgeons, "Facebook vs Twitter Infographic," 2010.

[99] E. Bakshy, I. Rosenn, and C. Marlow, "The Role of Social Networks in Information Diffusion," *Arxiv preprint arXiv:1201.4145*, 2012.

[100] Z. Birchmeier, B. Dietz-Uhler, and G. Stasser, *Strategic Uses of Social Technology: An Interactive Perspective of Social Psychology(Google eBook)*. Cambridge University Press, 2011.

Appendix A

Python scripts

A.1 Extracting tweets from JSON

This script will extract just the text and location from the JSON output in hugetextfile.txt:

```
file = open('/Location/hugetextfile.txt')

for line in file:
    start_text = line.find('"text"')
    start_tweet = line.find(':', start_text)
    end_tweet = line.find('","', start_tweet + 1)
    tweet = line[start_tweet+2:end_tweet]
    print tweet

    start_location = line.find('"location"', end_tweet)
    end_location = line.find('",', start_location)
    location = line[start_location:end_location]
    print location}
```

A.2 Obtaining tweets from city i

This will return the text of tweets from a given city (here, San Francisco):

```
file = open('/Location/tweetsfromallcities.txt')
tweet = file.readline()
location = file.readline()
while tweet:
    location_lowercase = location.lower()
    if 'san francisco' in location_lowercase:
    print tweet
tweet = file.readline()
location = file.readline()
```

A.3 Word frequencies

This will return the frequencies of every word in the text of tweets from a given city:

```
file = open('/Location/tweetsfromcityi.txt')

wordcounts = {}

for line in file:
    words = line.split()
    for word in words:
        word = word.lower()
        while not word.isalpha() and len(word) > 1:
            word = word[0:len(word)-1]
        if word in wordcounts:
            wordcounts[word] = wordcounts[word] + 1
        else:
            wordcounts[word] = 1

import operator

x = wordcounts
sorted_x = sorted(x.iteritems(), key=operator.itemgetter(1))

import csv

writer = csv.writer(open('/Location/counts_cityi.csv', 'wb'),
dialect = "excel")
writer.writerows(sorted_x)
```

A.4 Line counts

This will count the number of lines in a given file (useful here because the default JSON output puts each streamed tweet from Twitter's API into a single line):

```
def file_len(file):
    with open('/Location/tweetsfromcityi.txt') as file:
        for i, l in enumerate(file):
            pass
    return i + 1

print file_len('/Location/tweetsfromcityi.txt')
```

Appendix B

Raw data

(See next page for landscape-oriented tables.)

Table B.1: Food Tweet Tallies per City

City Sample	Region	Ob rate	Total tweets	McD's	Soda	Candy	Grocery	Healthy	Hungry
Atlanta	South	24.000	186945	2006	433	1682	930	1915	9601
Austin	South	25.000	50278	359	159	419	325	496	1600
Baltimore	South	27.000	50571	528	123	389	191	433	3015
Baton Rouge	South	32.000	20093	230	46	232	114	166	1455
Birmingham	South	32.000	63571	578	172	345	120	597	2037
Boston	Northeast	22.000	88058	608	237	650	445	1229	2828
Charlotte	South	26.000	43861	460	94	357	255	503	2058
Chicago	Midwest	26.000	203403	2348	635	1688	1072	2260	8398
Cincinnati	Midwest	27.000	25310	267	75	230	152	297	1164
Cleveland	Midwest	28.000	38250	392	69	356	178	367	1952
Columbus	Midwest	31.000	39836	447	104	324	187	427	2010
Dallas	South	29.353	89093	766	237	807	426	1144	4790
Denver	West	18.000	27673	144	81	235	182	515	1281
Detroit	Midwest	34.000	72619	714	131	621	412	700	4275
Greensboro	South	28.000	14451	173	42	117	64	117	1019
Harrisburg	Northeast	32.000	4151	54	11	39	24	44	178
Houston	South	29.000	136689	1425	357	1317	723	1305	8783
Indianapolis	Midwest	30.000	29028	301	87	245	195	335	1153
Jackson	South	36.000	41666	521	96	382	196	435	2614
Los Angeles	West	22.000	193108	1424	526	2155	1038	2972	5904
Memphis	South	34.000	45082	631	106	395	221	331	3184
Miami	South	24.000	132448	1616	322	1014	535	1310	7276
Milwaukee	Midwest	32.000	26457	316	78	212	150	218	1296
Minneapolis	Midwest	21.000	24273	147	90	161	178	357	435
Nashville	South	30.000	38118	288	107	278	311	452	1484
New Haven	Northeast	27.000	5274	62	9	39	22	54	253
New Orleans	South	30.000	44928	527	71	433	218	392	2992

Continued on next page

Table B.1 – continued from previous page

City Sample	Region	Ob rate	Total tweets	McD's	Soda	Candy	Grocery	Healthy	Hungry
New York	Northeast	22.916	312192	2742	1208	3105	1205	6294	8986
Norfolk	South	33.000	20506	222	49	110	83	196	876
Orlando	South	27.000	50995	630	133	483	213	679	3171
Philadelphia	Northeast	32.000	58269	460	169	558	270	700	2362
Phoenix	West	24.000	26332	182	80	226	124	438	788
Pittsburgh	Northeast	29.000	39871	433	93	381	209	429	1382
Portland	West	24.458	33401	134	112	272	186	433	585
Providence	Northeast	27.000	10800	78	27	76	47	91	564
Raleigh	South	26.000	20999	180	52	165	128	263	1061
Richmond	South	31.000	27636	245	67	253	148	322	1222
Sacramento	West	28.000	20919	174	53	238	106	301	868
Salt Lake City	West	25.000	9127	60	37	76	48	99	166
San Antonio	South	27.931	39057	441	99	385	165	466	1993
San Diego	West	23.000	61716	337	118	1013	271	818	1794
San Francisco	West	17.000	63858	285	176	412	301	847	955
Seattle	West	22.000	53310	254	169	480	438	677	1084
St. Louis	Midwest	29.000	22114	265	56	174	131	225	892
Tallahassee	South	28.000	17879	233	40	145	83	165	1275
Tampa	South	26.000	28801	328	78	194	154	415	1232
Washington DC	South	22.000	93365	755	461	709	507	1666	2790

End of table

Table B.2: Food Tweet Relative Frequencies per City

City	Region	RF McDonalds	RF Soda	RF Candy	RF Grocery	RF Healthy	RF Hungry
Atlanta	South	0.010730429	0.002316189	0.008997299	0.004974725	0.010243655	0.051357351
Austin	South	0.0071403	0.003162417	0.008333665	0.00646406	0.00986515	0.031823064
Baltimore	South	0.010440766	0.002432224	0.007692156	0.003776868	0.008562219	0.059619149
Baton Rouge	South	0.011446773	0.002289355	0.01154631	0.005673618	0.008261584	0.072413278
Birmingham	South	0.009092196	0.002705636	0.005427003	0.001887653	0.009391075	0.032042913
Boston	Northeast	0.00690454	0.002691408	0.007381499	0.005053487	0.01395671	0.032115197
Charlotte	South	0.010487677	0.002143134	0.008139349	0.005813821	0.011468047	0.046920955
Chicago	Midwest	0.011543586	0.003121881	0.008298796	0.005270325	0.011110947	0.041287493
Cincinnati	Midwest	0.01054919	0.002963256	0.009087317	0.006005531	0.011734492	0.045989727
Cleveland	Midwest	0.010248366	0.001803922	0.00930719	0.004653595	0.009594771	0.05103268
Columbus	Midwest	0.011221006	0.002610704	0.008133347	0.004694246	0.010718948	0.050456873
Dallas	South	0.008597757	0.002660142	0.009057951	0.00478152	0.012840515	0.053764044
Denver	West	0.005203628	0.002927041	0.008492032	0.006576808	0.018610198	0.046290608
Detroit	Midwest	0.009832138	0.001803936	0.008551481	0.005673446	0.009639351	0.058868891
Greensboro	South	0.01197149	0.002906373	0.008096326	0.004428759	0.008096326	0.070514151
Harrisburg	Northeast	0.013008914	0.002649964	0.009395326	0.005781739	0.010599855	0.042881233
Houston	South	0.010425126	0.002611768	0.009635011	0.00528938	0.00954722	0.064255353
Indianapolis	Midwest	0.010369299	0.002997106	0.008440127	0.006717652	0.011540582	0.03972027
Jackson	South	0.0125042	0.002304037	0.009168147	0.004704075	0.010440167	0.062737004
Los Angeles	West	0.007374112	0.002723864	0.011159558	0.00537523	0.015390352	0.030573565
Memphis	South	0.013996717	0.002351271	0.008761812	0.004902178	0.007342176	0.070626858
Miami	South	0.012201015	0.002431143	0.007655835	0.004039321	0.009890674	0.054934767
Milwaukee	Midwest	0.011943909	0.00294818	0.008013002	0.005669577	0.008239785	0.048985146
Minneapolis	Midwest	0.006056112	0.003707824	0.006632884	0.007333251	0.0147077	0.017921147
Nashville	South	0.007555486	0.002807073	0.007293142	0.008158875	0.011857915	0.038931738
New Haven	Northeast	0.011755783	0.001706485	0.007394767	0.004171407	0.010238908	0.047971179
New Orleans	South	0.011729879	0.001580306	0.009637642	0.004852208	0.008725071	0.066595442

Continued on next page

Table B.2 – continued from previous page

City	Region	RF McDonalds	RF Soda	RF Candy	RF Grocery	RF Healthy	RF Hungry
New York	Northeast	0.008783057	0.003869414	0.009945803	0.003859804	0.02016067	0.028783569
Norfolk	South	0.0108261	0.002389545	0.005364284	0.004047596	0.009558178	0.042719204
Orlando	South	0.012354152	0.002608099	0.009471517	0.00417688	0.013315031	0.062182567
Philadelphia	Northeast	0.007894421	0.002900342	0.009576276	0.004633682	0.012013249	0.040536134
Phoenix	West	0.006911742	0.003038129	0.008582713	0.004709099	0.016633754	0.029925566
Pittsburgh	Northeast	0.010860024	0.002332522	0.009555818	0.005241905	0.0107597	0.034661784
Portland	West	0.004011856	0.003353193	0.008143469	0.005568696	0.012963684	0.017514446
Providence	Northeast	0.007222222	0.0025	0.007037037	0.004351852	0.008425926	0.052222222
Raleigh	South	0.008571837	0.002476308	0.007857517	0.006095528	0.012524406	0.050526216
Richmond	South	0.008865248	0.002424374	0.009154726	0.005355334	0.011651469	0.044217687
Sacramento	West	0.008317797	0.002533582	0.011377217	0.005067164	0.014388833	0.041493379
Salt Lake City	West	0.006573902	0.004053906	0.008326942	0.005259121	0.010846938	0.018187794
San Antonio	South	0.01129119	0.002534757	0.009857388	0.004224595	0.01193128	0.051027985
San Diego	West	0.005460496	0.001911984	0.016413896	0.004391082	0.013254261	0.029068637
San Francisco	West	0.004463027	0.002756115	0.006451815	0.004713583	0.013263804	0.014955057
Seattle	West	0.004764585	0.003170137	0.009003939	0.008216095	0.012699306	0.020333896
St. Louis	Midwest	0.011983359	0.002532332	0.007868319	0.005923849	0.01017455	0.040336438
Tallahassee	South	0.013032049	0.002237262	0.008110073	0.004642318	0.009228704	0.071312713
Tampa	South	0.011388493	0.002708239	0.006735877	0.005347037	0.014409222	0.042776292
Washington DC	South	0.008086542	0.00493761	0.007593852	0.0054303	0.017843946	0.029882718

End of table