FACES IN PLACES: AN EXPLORATORY METHODOLOGY FOR MEASURING FINE-GRAINED DIVERSITY VIA SOCIAL MEDIA IMAGES

By

SAKET HEGDE

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Electrical and Computer Engineering

written under the direction of

Dr. Vivek K Singh

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2016

ABSTRACT OF THE THESIS

FACES IN PLACES: AN EXPLORATORY METHODOLOGY FOR MEASURING FINE-GRAINED DIVERSITY VIA SOCIAL MEDIA IMAGES

By SAKET HEGDE

Thesis Director:

Dr. Vivek K Singh

In this study, we explore a novel approach to measure fine-grained (photo-level) diversity using Instagram. We compare and contrast these new measures of diversity with traditional metrics (i.e. census). We discuss the merits and shortcomings of supplementing traditional census figures with these new measures. Further, we explore the predictive capacity that this new metric has over socio economic outcomes, namely income inequality.

We find that using our fine-grained metric for measuring diversity in interactions produces very different results compared to traditional census measures. We also determine that diversity (specifically photo based entropy in age and race) are associated with income inequality and the combined model is significantly (though weakly) predictive of inequality. Neighborhoods that have high scores in racial diversity seem to have a correlation with lower inequality, while neighborhoods that have high scores in age diversity seem to have a correlation with higher inequality. We discuss the possible implications of this work on research in sociology and associated areas and suggest further work based on these findings.

Acknowledgments

I would like to thank my advisor, Dr. Vivek K Singh from the School of Communication and Information at Rutgers. The exploratory nature of this work meant we had to learn by doing and I am indebted to Professor Singh for the trust, mutual respect, patience, guidance and the great kindness he showed to a young researcher.

I am also very grateful for the opportunity I had to work with the Behavioral Informatics Research Group at SCI. I was the youngest and most inexperienced member of the group when we first met last summer. Weekly meetings and constant discussion galvanized my thinking for this project and I am grateful for the lasting friendships that developed because of this. I am particularly obligated to Shy and Chris who provided constant assistance and advice. They were instrumental in this project reaching fruition.

I thank Dr. Zoran Gajic, Graduate Director of the ECE Department who saw my request to work with someone outside our department as a rare opportunity for interesting and collaborative research. I thank Dr. Shantenu Jha and Dr. Hana Godrich from the ECE Department for agreeing to be on the thesis committee under short notice as well as their valuable inputs.

I would also like to thank Professor Khadijah White from the JMS department for her valuable consultation.

Last but not least, I thank my family; my parents, who made many sacrifices so I can live my dreams, and my younger brother who will always be a good reason for me to work for a better world and a better future.

Table of Contents

List of Figures

Chapter 1

INTRODUCTION

Diversity is an important socio-economic construct that has associations with multiple aspects of human life including commerce, innovation, wellbeing, criminal justice, civic responsibilities and health among others [30][31][32][33][34][35][36][37][41][105]. Traditionally diversity has been defined as a function of the number of people of different age, gender, ethnicity etc. living in the same neighborhood as observed through long-term data e.g. census [1][122]. However, there exist multiple nuances in the notion of diversity, some of which remain hidden if we work with coarse, long-term, residential measures of diversity. For example, zooming further into the location aspect, it has been reported that the areas which appear to be most diverse at city scale are often also the most segregated, when observed at a neighborhood scale [2].

Census data (such as from the American Community Survey [6]) provide an important public service by collecting and preserving demographics data going back for decades[4][7]. Despite its inherent value, census based data collection and analysis has several shortcomings including being (comparatively) expensive, time consuming and invasive. In much previous work, it has also been argued that studying informal interactions is important, but it was hard to study them all along [124].

There is hence much motivation to complement traditional census metrics with a novel methodology that leverages the increasing amount of information that is available from new sources of data such as social networks.

We propose to develop a more spatio-temporally fine grained diversity score for different locations using social media (Instagram images [10]). We focus on neighborhoods in New York City and quantify the mixing between people based on individual photos from each neighborhood. Specifically, we focus on photos that have multiple people interacting (two

or more faces), and we employ automatic methods for race, age, and gender estimation from such photos. Based on the facial analytics from individual photos we calculate measures of diversity for each neighborhood

Using this more fine-grained representative data, we aim to enrich the interpretation of diversity further in two ways:

1) Diversity varies over time: We posit that diversity changes not just over years and months but also over days, hours, and minutes and real-time social media data provides a unique opportunity to quantify diversity as an evolving property across space and time.

2) Diversity via relationships, not residence: It is often reported that different sections of society living in the same zip code, street or even the same building, never talk to each other [38][39] and block level segregation is often not indicative of these pervasive social barriers. Thus, we want to identify a method that quantifies the interconnections and communications between people rather than just their residential addresses.

Since the last decade, researchers increasingly turned to social media analysis to complement traditional survey methods in areas such as public health[80], politics [81], and most recently, in marketing [82]. The growing popularity of social media, particularly photo sharing communities, presents a huge compendium of data that motivates a different approach to studying diversity. As of June 2016, Instagram has over 500 million users globally of which 300 million use the app every day [55]. Also, as of September 2015, it was reported that over 80 million photos are shared daily on Instagram [56]. This large user base, coupled with a massive amount of data that is easily accessed via Instagram's public API [57] has presented a unique opportunity for researchers seeking to urban studies [123], demographics (of selfie takers) [59] and visualizations of cityscapes [58] [60].

Instagram presents a unique facet for studying diversity due to the assumed level of interconnection between individuals in group photos. We elaborate more on the reasoning behind using this platform (especially in terms of relationships) and the demographic represented by the users of this application in the Literature review section.

Keeping in mind the aim of enhancing the interpretation of diversity and with the growing popularity and utility of social media for research in neighboring disciplines, we develop our first research question:

RQ1: Can social media be used to study diversity at a fine grained resolution?

We assume that people sharing the same "frame" in a single photo have some kind of interconnection among themselves. Previous work has shown that photos are important in social relationships [62]. The content of photos shows who is part of a group and telling stories about photos helps nurture relationships [63]. In the section on methodology, we elaborate on what exactly constitutes a neighborhood and also why we selected New York City for this study including the unique opportunities and challenges this presents.

Building such a diversity index and tracking it over time may yield two benefits: 1) From an epistemological perspective, these fine-grained real-time diversity scores might reveal certain phenomena in their own right. For example, how does the notion of diversity change when measured over short and long-term durations? Or what do daily or weekly cycles in such a diversity measure indicate? 2) From a socio-economic impact perspective, such a diversity index may also yield predictive power on socio-economic outcomes including levels of innovation, economic activity, crime, civic engagement, and happiness levels.

This leads to our second and third research questions:

RQ2: Does the metric of diversity derived from fine grained (photo-level/personal space) interactions differ from traditional measures of diversity at higher granularity in densely populated urban environments?

RQ3: Does studying diversity using fine grained interactions in social media photos yield a different predictive capacity over socio economic outcomes (as compared to traditional measures of diversity)?

One of the major issues to emerge in the current presidential primary season is income inequality, This is unsurprising, given the lack of income growth for households in the middle and lower parts of the distribution since 1970 [3][18]. Due to this, we concentrate on New York city neighborhoods and measure inequality for each of these areas in terms of mean versus median income by using data from the American Community Survey. We also measure the diversity in fine grained interactions in these neighborhoods and employ statistical methods to see if there is any association between diversity and income inequality at the neighborhood level. There is very little work on studying diversity via social media at

the very fine level (personal space) of geospatial granularity (the foursquare locations study being a welcome exception [17]) and very little in terms of correlation to socioeconomic outcomes [58].

The rest of this document is organized as follows. In Section II, we first discuss related work in demographics studies that harness social media platforms and data. We elaborate on diversity and inequality as well as their relationship and the context that New York City provides in this respect with a literature review. In Section III, we examine the methodology we have employed along with the steps to mitigate some of the biases. Finally in Section IV, we discuss our results and significant findings along with a discussion on threats to validity and suggestions for a future exploration of similar work.

Chapter 2

LITERATURE REVIEW AND RELATED WORK

2.1 Social Media's role in Studying Diversity

Social media represents a rich repository of information and the theme of analyzing social media data to identify interesting patterns in urban environments has been around for some time. In 2013, Adnan et al. at the University College London used Twitter's API to examine a million geotagged tweets in London and produced a map of the city's ethnic groupings [15]. Using Twitter usernames and the Onomap methodology developed by Mateos et al. [84] they exploit the distinctive naming practices in different societal groups for classification.

Indeed, "diversity" as a construct varies not only geospatially but also based on the size of the grain metric. Caetano et al. [17] finds that while neighborhoods may seem diverse, people sort by age and by gender across venues that are very close to each other [17]. This shows that social interactions are more complex and that diversity in venues does not necessarily ensure diversity in participation at the venue level. This study is particularly relevant because the findings were fairly consistent across samples from dense, older cities such as New York to sprawling, younger cities such as Dallas. Similar work [85] has found that the concept of homophily also extends to personality and association in social networking contexts. Both these studies exploit the Location Based Social Network (LBSN) called Foursquare [86]. Researchers have also used Twitter to study multicultural diversity via language detection in Milan [16].

These studies highlight that social media can provide a rich dataset, conveniently provided via an API to make reasonably accurate demographic predictions at a neighborhood level in an ethnically diverse city. We hence consider our work to be a further exploration of these ideas.

2.2 Diversity as a Socio-Economic Construct

An interconnection between diversity scores and socio-economic variables has been studied in the past [87][88][89][90][91]. For example, [87] has argued for the economic value of cultural diversity and [89] has connected immigration and the associated diversity with economic prosperity. However, these works focused on city or neighborhood level residential statistics rather than the interconnections between people. Trying to measure interactions among people, [92] has reported that cities in the UK that had higher diversity in terms of connections with other cities, as measured via phone-calls placed to other cities, had higher levels of economic development. Looking at equity markets, [94] found that diversity can thwart prominent failures like price bubbles that may arise out of a tendency for conformity. Lastly, it has been found that crime patterns in London can be predicted reasonably accurately by quantifying the changes in diversity scores across day and night [93].

Given the importance of the topic, there is also a growing interest in plotting and visualizing such diversity data. For example, both LA Times and ESRI have released visualizations of such neighborhood level diversity indices in the recent past [95][96]. However, the authors have again focused on residential statistics, rather than the mixing of people in terms of talking to each other or sharing the same photo frame.

2.3 Income Inequality in the Context of New York City

New York City presents a unique urban setting to study economic inequality. A study from the Brookings Institution [53] finds that inequality is particularly high in economically vibrant cities (like New York and San Francisco) than in less dynamic ones like Columbus, Ohio and Wichita, Kansas. Low-inequality cities, the study found, tend to be large, "spread-out" cities located in the South and Midwest. This suggests that among other factors, New York's size, and affluence may affect income inequality in the area. Census data covering the year 2013 [6] for example shows that the top 5 percent of households earned $864,394, or 88 times as much as the poorest 20 percent in Manhattan, the largest wealth gap in the US. A Harvard study [25] however, finds that despite this, New Yorkers have a high chance of upward mobility in income levels. This is often because public housing residents

in neighborhoods with high and increasing income have better social outcomes [54]. Despite this, few New Yorkers speak to their neighbors [14] and we were interested to see what we would learn by studying the diversity in interactions within neighborhoods in the city, where the rich and poor often live within blocks of each other [117].

2.4 Instagram and the Demographic that it represents

A recent study of "The Demographics of Social Media Users" by Pew Internet Research [75] found that 72% of online American adults use Facebook, 31% use Pinterest, 28% use Instagram and about a quarter use LinkedIn and Twitter. Social media usage is skewed towards women for all of the aforementioned platforms, except for LinkedIn, where it is almost equal and Twitter, where men form a greater majority. The same study found that Instagram is extremely popular with non-whites and young adults: 55% of online adults aged 18 to 29 use Instagram, as do 47% of African Americans and 38% of Hispanics. One can easily confirm that these demographics as well as the demographic represented by users of other social media do not correlate well with New York City demographics [74]. Indeed, many studies have found that social media users do not form a representative sample of the population (for example by over-representing minorities and urban population) [76]. Some researchers have attempted to adjust for this bias [77][78]. Instagram has the same biases in user demographics inherent in other social media. Several instagram photos contain facial images of users without Instagram accounts. Since we include these in our analyses, we do end up adding at least a portion of individuals who do not use Instagram.

Although liking, commenting and sharing these photos represent interactions in the social media world, the true data in this study is the faces of real people in the real world. The unique facet presented by this is the assumed level of interconnection between individuals in group photos. Research has shown that photos are important in social relationships [62]. The content of photos shows who is part of a group and telling stories about photos helps nurture relationships [63].

Chapter 3

METHODOLOGY

In this section, we describe the collection of our dataset, the filtering procedures we employ to reduce biases and how we build diversity and inequality measures for each neighborhood. We also detail what constitutes a "neighborhood" in the context of this work.

3.1 Sampling and Filtering of Data

We sampled content from Instagram (photos and associated metadata) over a 14 day period in April 2016 in the New York City area (see Fig. 1). We used Instagram's public API to gather these photos and employed random location-based sampling. The method commonly employed is to select a location at random within a latitude-longitude range encompassed by the city. The most recent photos in the immediate vicinity are then sampled and this process is repeated several times over. At the end of the sampling period, we had a corpus of about 34,382 unique photos. After filtering out tourists and visitors, we determined that 8,067 images or about 23 percent of these images were facial images. About 3,688 images or about 9.32 percent of all images had multiple faces with an average of about 3.5 faces per image.

Previous studies of large Instagram samples in large urban areas have shown that approximately one-fifth of photos contain faces while less than 5% of photos contain selfies [59]. Our statistics seem to concur with these findings.

Along with the photos gathered from the API, we also collected metadata for each photo such as location and timestamp information.

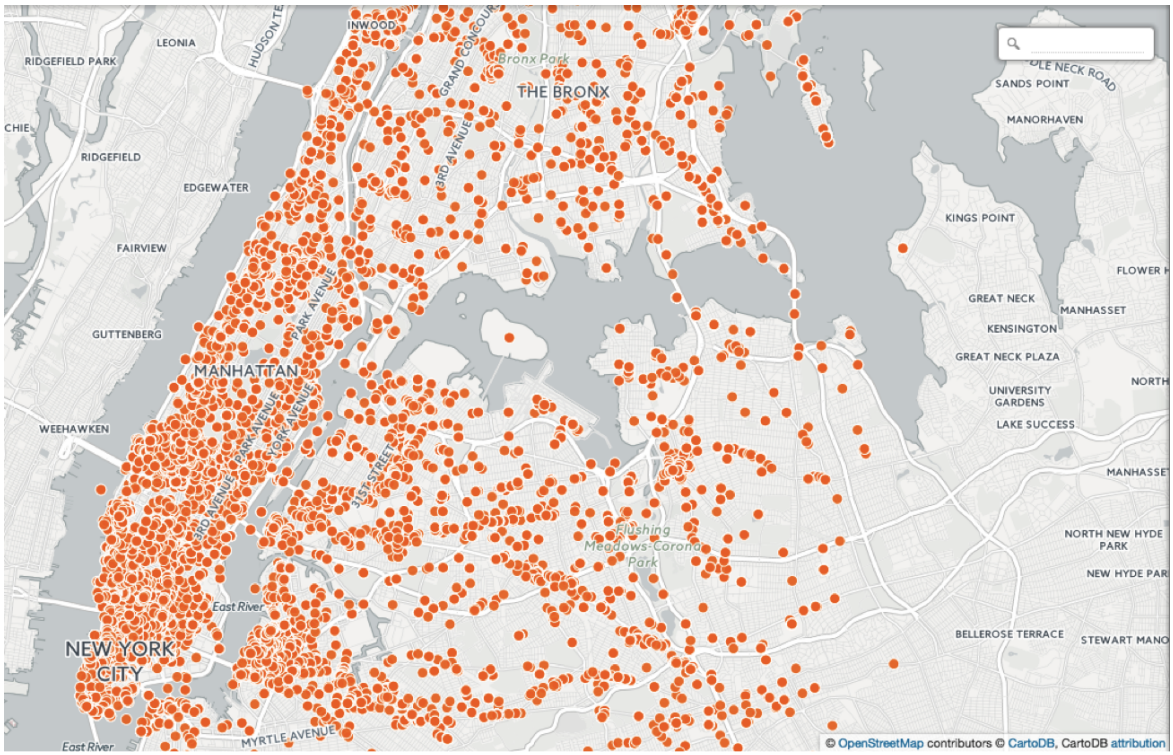

Figure 1: Image of New York City Area showing location of sampled photos over a 2 week period (made using CartoDB visualization)

3.1.1 Tourists and Visitors

In 2015, New York City received over 55 million tourists and this number has been growing every year[98][99]. In order to capture demographics only for New Yorkers, it is necessary to minimize biases introduced by tourists and visitors who also upload photos from within the city. The discovery and visualization of "tourist" photos has been studied in the past [97] and we employ a similar technique. We use a simple methodology by examining the media of user accounts that posted photos included in our sample set. If over 50% of recent user media (taken within one month before the sampled photo) was posted from outside New York City, then we label the user and all his/her photos as tourist/visitor photos and discard them from our analysis.

3.2 New York City Neighborhoods

It is known that social media activity, at least in terms of production of information is unequal in New York City [58]. In order to control for the unequal distribution of photos across New York, we sample over a two week period and only consider sample areas with at least 30 faces. We therefore had to select a level of spatial granularity that ensured that each region had a sufficient sample size of faces, while also attempting to include as many regions in New York City as possible. By setting a threshold that each region must contain at least 30 faces, we found that characterizing New York City neighborhoods by Zip Code Tabulation Areas [26] fits our requirements for an appropriate spatial metric. These are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas. We also select this level of granularity since ACS figures on demographics and income are easily available. The tradeoff ensures a statistically significant sample size while ensuring that most of the area of New York City is covered. As can be seen from the sample map in Fig 1, areas in lower Manhattan have a disproportionately higher number of photos while Upper Manhattan and other boroughs have a more sparse distribution. It is worth mentioning that this variation in social media activity or "social media-divide" has been interpreted by some to be indicative of the transcendence of social inequality and the digital divide [58]. The median number of faces sampled from each neighborhood is 114,

with a range of 30 to 657 (after excluding those ZCTA's with a lower number of detected faces in photos). Despite this, we are able to include some 79 ZCTA's in our analysis. As seen in the heat maps of figures 2 to 4, we have succeeded in covering most of New York City by using the aforementioned characterizations and thresholds.

3.3 Income Inequality

There are several ways to measure income inequality such as the Gini coefficient (arguably the most commonly used measure) [120] [44] or the ratio of mean household incomes in the top and bottom quintiles [43]. Gini coefficients at low granularity and particularly in New York, may not be a representative measure of inequality. Since it takes a substantial change in incomes to generate small changes in the Gini coefficient, this measure can hide large differences in relative incomes, particularly at smaller levels of spatial granularity [43]. New York City also provides housing for low income residents via housing lotteries [51] and affordable housing [52]. This often places low income residents in otherwise affluent neighborhoods which results in higher Gini coefficients that may not accurately represent inequality in those areas.

In our analysis, we choose to utilize the ratio of mean to median income. In the United States mean incomes have been consistently rising since the worst of the Great Recession, while median income has been falling, indicating the accrual of a greater share of income to those at the top of the income distribution [121]. Moreover, long term trends in income inequality (in America) show a clear divergence in these two metrics of prosperity for at least the last three decades [42]. Determining the ratio of mean to median incomes therefore provides a simple measure of the skewness in the distribution [44], has long been used in literature [45][46] and is less sensitive to underreporting of income at the top of distributions.

3.4 Face Detection and Analysis

Face detection refers to the computer vision technique of determining the presence and segmenting human faces in a provided photo. Typically facial features are detected while background “clutter” is ignored. Facial analysis builds on detection, using the rich set of

features provided to deduce age, gender, race, pose, emotions, etc.

The past few years have seen an explosion in facial recognition research by several research groups in areas ranging from biometric authentication to interactive learning through video and gesture recognition [67][68][69][70][73].

We determine the demographics of individual photos using facial analysis from the publicly available face detection API Face++ [13]. The Face++ API has already been validated for accuracy in predicting age and gender in social media photos both on Twitter [12] and Instagram [11]. However since the accuracy for race classification using this API has not been validated in prior work, we validate the software on a prelabeled database of images as described later in this section. Face++ provides an API that accepts the URL of an Instagram image and returns information about detected faces. This information includes the position of the face in the image, as well as the detected gender and age range of all faces. Recall that, we filter for only those images with multiple (i.e. 2 or more faces).

3.5 Data

For each photo, we calculate the values of the variables provided by the Face++'s API (we use a methodology quite similar to [11]). For gender data, we maintain two counts, one of the number of female faces detected and another of the number of male faces detected in an image. To reduce dimensionality for age, we categorize faces in an image into various age ranges. The three age ranges we consider in this paper are (1) children and teens- younger than 18, (2) young adults- faces with age between 18 and 35, and (3) older adults- older than 35. Finally, for race, we consider the three major racial groups in New York City-White, Black and Asian (this is based on ACS Demographics and Housing Estimates 2010-2014 American Community Survey 5-Year Estimates [74]). We maintain counts for the number of White, Black and Asian faces detected in each photo.

A formal description of all the variables recorded for individual photos is provided below-

Number of faces: For each Instagram photo, we count the number of faces detected by the API. This is a whole number that is greater than or equal to 2.

Variables for age, gender and race are calculated as follows:

Number of Male Faces: Count of the number of Male faces detected by the API. It can be any whole number between 0 and the number of faces.

Number of Female Faces: Count of the number of Female faces detected by the API. It can be any whole number between 0 and the number of faces.

Number of Faces <18 years old: Count of the number of faces in the photo below 18 years of age as identified by the API

Number of Faces >18 and <35 years old: Count of the number of faces in the photo identified to be between 18 and 35 years of age

Number of Faces >35 years old: Count of the number of faces in the photo that are identified to be older than 35 years by the API

Number of White Faces: Count of the number of White faces in the photo as identified by the API

Number of Black Faces: Count of the number of Black faces in the photo detected by the API

Number of Asian Faces: Count of the number of Asian faces in the photo as identified by the API

3.6 Developing a Neighborhood measure of Diversity

Based on the values of the aforementioned variables (calculated for each photo), we develop a measure of diversity for age, gender and race. This is then calculated for each neighborhood (i.e. ZCTA as defined earlier). A popular method, frequently cited in studies of diversity in the United States and at varying levels of granularity[100][101][102][103][105] is Shannon Entropy also known as Shannon's Diversity Index. The measure was originally proposed by Claude Shannon to quantify the entropy (uncertainty or information content) in strings of text [104].

It is calculated as follows:

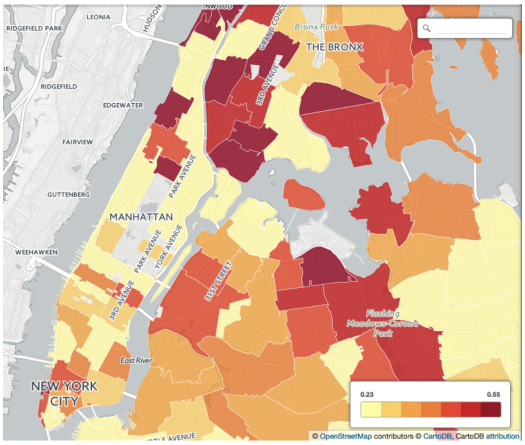

Figure 2: Heat Map of New York City showing the Racial Diversity in Photo-level interactions (SEraceMedian) in various neighborhoods

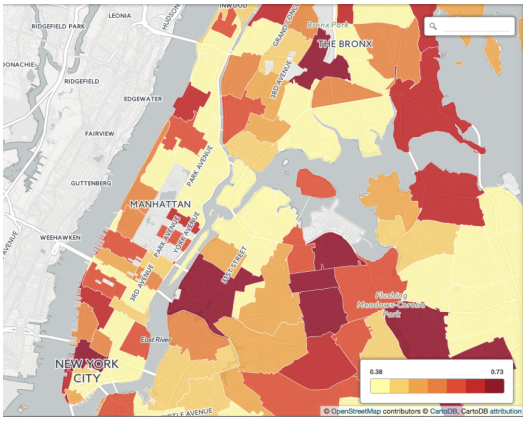

Figure 3: Heat Map of New York City showing the Age Diversity in Photo-level interactions (SEageMedian) in various neighborhoods

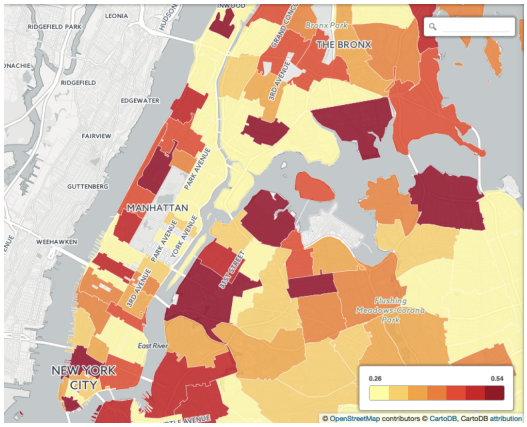

Figure 4: Heat Map of New York City showing the Gender Diversity in Photo-level interactions (SEgenderMedian) in various neighborhoods

$$-\sum_{i=1}^{R} p_i ln p_i \tag{3.1}$$

where pi is the proportion of characters belonging to the ith type of letter in the string of interest. In ecology, pi is often the proportion of individuals belonging to the ith species in the dataset of interest. Then the Shannon entropy quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset. To begin with, we first calculate three measures of Shannon Entropy for each individual image-

SEage- As a measure of age diversity for the photo, with 3 values of i representing the 3 age groups defined above.

SEgender- As a measure of gender diversity for the photo, with 2 values of i representing male or female gender.

SErace- As a measure of race diversity for the photo, with the 3 values of i representing New York's 3 major racial groups as described above.

For each neighborhood we then determine the median value among the photos for each of the above 3 variables as SEageMedian, SEgenderMedian and SEraceMedian. These then serve as our measure of diversity in terms of age, gender and race respectively, for that neighborhood. Figures 2 to 4 display the spatial distribution of these values by neighborhood using a heat map. All maps were made using CartoDB [71].

For the benefit of comparison, we also calculate measures of neighborhood level diversity using Census figures [74]. We do so by using ACS measurements on race, gender and age to calculate similar Shannon Entropy figures from census. The 3 variables we calculate for each neighborhood are SEageCensus, SEgenderCensus and SEraceCensus, also computed as defined in equation (1).

In the section on results, we discuss whether these new measures allow us to model diversity as a socio-economic construct and the kind of correlations that are gleaned by looking at income inequality in relation to both of the diversity metrics that we calculate. We develop a measure of dependence between diversity measures and income inequality by using multiple linear regression.

3.7 Validation of Race Classification

As mentioned earlier, despite being validated for high accuracy in age and gender classification, the Face++ API has not been validated for race classification. In order to do so, we use the prelabeled MORPH database [106]. It contains 55,000 unique images of more than 13,000 subjects along with labels for race. This database has been used in previous studies to validate race classification for a project that used facial analysis of social media images [12]. Using a corpus of 1,154 (500 Black, 500 White and all 154 available Asian) randomly selected images, we validate race classification for New York's three major racial groups with a mean average precision (MAP) of 92.98%. This is comparable to the classification accuracy for race reported in other recent work that uses social media [12].

Chapter 4

RESULTS

4.1 Comparing Census and Social Media based Measures of Diversity

We find that measuring diversity based on fine grained social media interactions results in a very different portrait of the city. Due to reasons mentioned earlier, such as the perceived level of interaction in photos, this measure of diversity is quite different from traditional census measurements.

ZCTA	SEageMedian	SEraceMedian	SEgenderMedian	SEageCensus	SEraceCensus	SEgenderCensus
10012	0.636	0.5	0.562	0.94	0.55	0.69
10017	0.693	0.693	0.28	0.88	0.657	0.688
10280	0.318	0.318	0	0.99	0.571	0.69
10475	0.636	0.379	0.636	0.583	0.898	0.689
10805	0.655	0	0.661	0.993	0.663	0.692

Table 1: Census and Photo Based Entropy for a sample of New York neighborhoods (ZCTA's)

As seen in Table 1, it is no surprise that Diversity as measured by census and photo entropy provides very different statistics. In some cases, the difference is as high as 100% due to zero entropy in the level of interaction at the photo level (for sampled photos) for that neighborhood. In general, photo based entropy is almost always lower than census based entropy for all three measures. The lack of photo based entropy can be explained by the phenomenon of homophily seen at lower levels of spatial granularity as also noted by [17]. We have also elucidated on other reasons for this expected outcome in earlier sections.

These results corroborate earlier efforts that have also reported that different sections of society living in the same zip code, street or even the same building, never talk to each other [38][39] and block level segregation is often not indicative of these pervasive social barriers.

4.2 Effect of different measures of Diversity on Income Inequality

We employ statistical methods, namely multilinear regression to determine the association between the two measures of diversity (i.e. photo based entropy and census based entropy) and income inequality.

Using multiple linear regression, we ran calculations for ANOVA (Analysis of Variance) to form a basis for tests of significance. Using the three photo based metrics of diversity as predictors, we find that a combined model is significantly (though weakly) predictive of inequality. We find that adjusted R square =0.116 or that 11.6% of the variation in income inequality is explained by this model with a significance value of 0.005 indicating at least one of the constituent predictors is significant in the association.

Digging deeper, we find that for SEageMedian there is a beta value of 0.301 and significance value of 0.005. For SEraceMedian, there is a beta value of -0.272 and significance value of 0.011. While for SEgenderMedian, the association with income inequality is neither strong nor significant.

Upon building a census based entropy model, we find that all three predictors yield a model that is not significantly predictive of inequality. Not surprisingly, individual features are also not predictive. Hence, looking at census data alone, one might assume there is no correlation between diversity in urban neighborhoods and income inequality, but the fact that we see a significant correlation between the diversity in fine grained interactions and income inequality motivates the utility of social media to study diversity along with associated socio-economic outcomes.

The predictive capacity of both photo based diversity metrics as well as census based metrics is summarized below.

Diversity Measure	Beta (Photo Based Entropy)	Sig Value (Photo Based Entropy)	Beta (Census Based Entropy)	Sig Value (Census Based Entropy)
SEageMedian	0.301	0.005	0.001	0.996
SEraceMedian	-0.272	0.011	-0.149	0.161
SEgenderMedian	0.007	0.948	-0.160	0.269

Table 2: Diversity Measures as Predictors of Inequality (Photo Based Entropy and Census Based Entropy)

The major finding of this work is that measuring diversity through fine grained (photo-level interactions) is justified and can be used to augment traditional measures of diversity

such as census based surveys. We also find that while traditional diversity metrics may not be indicative of certain socio-economic trends at low spatial granularity, studying diversity through interactions on social media has the potential to yield associations with income inequality at the neighborhood level. This encourages a more in-depth exploration of such a methodology both for studying diversity as well as to gain a better understanding of phenomena of social interest.

Chapter 5

Discussion

In this study we attempt to explore a new methodology to determine diversity measures using fine grained interactions in a diverse, densely populated urban environment. We employ the canvas of social media to study diversity as a socio-economic construct. Our first research question asks if we can use social media to study diversity at a fine grained resolution. To answer this, we use the Instagram platform to gather photos from the New York City area. We then use facial analytics to label faces for age, gender and race in individual geotagged photos. Further, we use these labels to develop diversity scores for neighborhoods (based on where photos were uploaded) in New York while minimizing the effect of tourists and visitors on these scores. We find that our novel methodology allows us to develop diversity scores that are quite different from traditional measures. We discuss why this is likely to happen.

Our second research question concerns the the new metric of diversity studied using photo level interactions and a comparison with traditional census figures. We find that our new methodology is a simple, cheap and faster method for determining diversity in interactions in neighborhoods in New York City, although it is not without inherent biases. For example, it is circumscribed by the Instagram user base and may be slightly skewed due to tourists and visitors. We do attempt to mitigate some of these effects, for example, by filtering out tourist and visitor images.

Finally, we ask if this new method of measuring diversity in fine grained interactions yields a different predictive capacity over socio economic outcomes as compared to traditional measures of diversity. We find some significant correlations though we must concede that they they are able to explain relatively modest levels of variation in the variable of interest.

We find that racial diversity (measured using our new metric) is associated but negatively correlated with income inequality. Areas with higher racial mixing appear to have lower income inequality. With age diversity (as measured in photo based entropy), the effect is opposite with a positive association between age diversity and income inequality.

We attempt to explain these phenomena (without justifying a causal effect) in our discussion.

These correlations disappear when considering Census based measures of diversity. A census-based entropy model (with all three features combined) yields a model that is not significantly predictive of inequality. Unsurprisingly, considering individual features leads to the same outcome.

We also find that gender diversity whether measured via census or photo based entropy is not indicative of income inequality.

Racial Diversity in photos is negatively correlated with Inequality(sub under discussion) (done till end of this section except blue) One of the findings in this study is the association between racial diversity in Instagram photos and reduced inequality in the neighborhoods that these photos were uploaded from. At the neighborhood level, racial diversity as measured by traditional census data was not found to be associated with inequality. However using the newer method, we found a significant and negative correlation between racial diversity and income inequality. The Census Bureau itself has studied how income inequality varies spatially over the United States, but at spatial resolution that is much higher than our metric [116]. To the best of our knowledge, the relationship between fine-grained (personal space) level interactions between persons from different racial groups and income inequality has never been studied, particularly because it is difficult to quantify the former.

One possible explanation for our findings stems from the idea of societal integration. The causes for racial and economic inequalities are in the United States are by now well understood. Persistent segregation by race and income aggravates racial and income inequalities by disinvestment in neighborhoods with low income residents and a lack of access to good quality housing and schooling [108][107][109]. In US metro areas in particular, this problem has been worsening with (income-based) residential segregation worsening even at the census tract level [114]. Integration has the potential to reduce such inequalities by

providing residents access to similar resources and amenities and by offering opportunities to interact [113][107]. There is therefore a desire to promote racial and economic integration to prevent the adverse consequences of segregation [107]. As discussed earlier, we know that census figures, even at the block level and especially in New York City, are often not indicative of the true level of integration in a community. However, since our metric of ?diversity? involves measuring the racial mixing in fine grained (photo level interactions), it is probably more indicative of the actual level of integration that occurs in New York City neighborhoods.

Age Diversity in photos and Inequality (sub under discussion) A particularly surprising finding is the association between age diversity in photos and higher income inequality. While this appears to be counterintuitive in the first instance, a possible interpretation is as follows.

In 2007 it was reported that the borough of Manhattan is experiencing a "baby boom" that is unique among U.S. cities [118]. Since 2000, the number of children under age 5 living in Manhattan has grown by more than 32% with the sharpest increase seen in affluent white families with median household incomes over $300,000 [118]. More recent data has also shown similar trends suggesting that birth rates are higher among the more affluent, particularly in Manhattan [119]. Although inequality is a city wide occurrence in New York, the issue is magnified in Manhattan, to the extent that it has the highest income gap among any large county in the United States [72]. It is possible that our results are the reflection of the presence of family photos in affluent but unequal areas in Manhattan. A future expansion of this study over larger areas may shed more light on this phenomenon.

5.1 Threats to Validity

One of the goals of this project is to define a simple and more robust methodology for measuring diversity. However, despite being cheaper and more real-time, social media studies such as ours are not without flaws. Arnaboldi, Michela et al.[16] note that the correspondence between figures from Twitter and census data (such as the percentage of residents from a certain ethnicity) is often infrequent and shallow primarily because the two

sources describe very different phenomena (residents versus social media authors including tourists and visitors). To account for this, we have already discussed the issue of tourists and visitors in the methodology section, elaborating on how we minimized the skew introduced by their presence in photos. The highly biased nature of the Instagram population is something we have acknowledged in the section on Literature Review and as in other, similar studies [16], we cannot completely discount its effects. Adoption rates of Instagram, the consideration of only public, geotagged photos are some other factors that similar studies have conceded as shortcomings [16] and that we would do well to concede as well. It can also be plainly seen that the location in a geotagged photo may not necessarily represent the neighborhood that the individuals in the photo are resident in. With these deficiencies in mind, we present our methodology as a way of augmenting (rather than replacing) research that uses census data. At the same time, this is the first attempt to leverage social media to not explore diversity but also understand associated social phenomenon. In future work, we can refine our methodology, collect more photos and apply this work to a larger swathe thus yielding varied results.

5.2 Future Work

The exploratory nature of this work entails that we do not argue for a causal effect in the relationship between income inequality and diversity in interactions in social media photos. However, we do consider these associations worthy of exploration in further studies. For example, gentrification is one aspect of the modern city that may be better understood by undertaking such an approach.

Rapid gentrification is radically changing the face of many neighborhoods in the nation. Nowhere is this as apparent as New York City. The NYU Furman Center recently released a special report on gentrification as part of their annual report on New York City?s Housing and Neighborhoods [8] identifying “gentrifying” neighborhoods as areas characterized by low income residents in 1990 but with higher than usual rent increases in the last 20 years. While census data provides yearly estimates of demographics, monthly statistics (supplemented by data on rent and income increases [9]) can help identify areas where sharp inequality is prevalent and displacement is likely to be occurring or may occur in the future.

The use of social media to augment traditional census studies may thus be complementary and we hope our results motivate the use of the proposed methodology for studying similar questions pertaining to social mobility, racial inequality and segregation. Answering these questions will be of great value to anthropologists, urban planners, policy makers and sociologists. Further, although we compare our methodology with traditional diversity scores, a more in-depth exploration of how the notion of diversity changes in neighborhoods over time is another interesting path on which we plan to undertake future work.

References

[1] M. Maly, ?The Neighborhood Diversity Index: A Complementary Measure of Racial Residential Settlement? Journal of Urban Affairs, Volume 22, Issue 1, pages 37-47, Spring 2000. Available: http://maps.latimes.com/about/

[2] N.Silver, ?The Most Diverse Cities Are Often The Most Segregated?, FiveThirtyEightEconomics [Online]. Available: http://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/

[3] Chad Stone, Danilo Trisi, Arloc Sherman, and Brandon Debot. 2006. A Guide to Statistics on Historical Trends in Income Inequality. (October 2015). Retrieved March 2, 2005 from http://www.cbpp.org/research/poverty-and-inequality/a-guide-to-statistics-on-historical-trends-in-income-inequality

[4] Jason Gauthier, PIO. "American Community Survey - History - U.S. Census Bureau". *Census.gov.* N.p., 2016. Web. 1 July 2016. https://www.census.gov/history/www/programs/demographic/american_community_survey.html

[5] Sharkey, Patrick Thomas. 2007. The enduring inequality of race and place: Racial inequality in the neighborhood environment over the life course and across generations. Retrieved March 2, 2005 from http://gradworks.umi.com/32/85/3285544.html

[6] Bureau, US. "American Community Survey (ACS)". Census.gov. N.p., 2016. Web. 1 July 2016. https://www.census.gov/programs-surveys/acs/

[7] Goering, John. "Segregation, Race, And Bias: The Role Of The US Census:". *Graduate Center and Baruch College, CUNY (2004)*: n. pag. Print. https://www.census.gov/housing/patterns/publications/goering.pdf

[8] Maxwell Austensen, Ingrid Gould Ellen, Luke Herrine, Brian Karfunkel, Gita Khun Jush, Shannon Moriarty, Stephanie Rosoff, Traci Sanders, Eric Stern, Michael Suher, Mark A. Willis, and Jessica Yager. 2016. State of New York City?s Housing & Neighborhoods. (June 2016). Retrieved March 2, 2005 from http://furmancenter.org/research/sonychan

[9] Sharkey, Patrick Thomas. 2016. Locating Displacement Hot Spots in NYC. (November 2016). Retrieved March 2, 2005 from http://research.prattsils.org/blog/coursework/locating-displacement-hot-spots-in-nyc/

[10] "Instagram". *Instagram.com.* N.p., 2016. Web. 1 July 2016. https://www.instagram.com/

[11] Saeideh Bakhshi, David Ayman Shamma, Eric Gilbert. 2014. Faces engage us: photos with faces attract more likes and comments on Instagram. (April 2014) . Retrieved March 2, 2005 from https://www.researchgate.net/publication/266655817

_Faces_engage_us_photos_with_faces_attract_more_likes_and_comments
_on_Instagram

[12] Yu Wang, Yuncheng Li, Jiebo Luo. 2016. Deciphering the 2016 U.S. Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists. (March 2016) . Retrieved March 2, 2005 from http://arxiv.org/pdf/1603.03097v1.pdf

[13] "Face++: Leading Face Recognition On Cloud". *Faceplusplus.com.* N.p., 2016. Web. 1 July 2016. http://www.faceplusplus.com

[14] Jackie Stockinger. 2016. FACT: 46% of New Yorkers Only Talk to Their Neighbors About Noise Complaints. (February 2016) . Retrieved March 2, 2005 from http://spoilednyc.com/percent-forty-six-new-york-neighbors-talk-noise-complaints/

[15] Adnan, Muhammad, Guy Lansley, and Paul A Longley. "A Geodemographic Analysis Of The Ethnicity And Identity Of Twitter Users In Greater London". University College London, Department of Geography, Gower Street, London, WC1E 6BT, 2013. Print. http://www.geos.ed.ac.uk/ gisteac/proceedingsonline/GISRUK2013/ gisruk2013_submission_50.pdf

[16] Arnaboldi, Michela et al. "Studying Multicultural Diversity Of Cities And Neighborhoods Through Social Media Language Detection". *The Workshops of the Tenth International AAAI Conference on Web and Social Media CityLab: Technical Report WS-16-16 (2016)*: n. pag. Print. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download /13227/12851

[17] Caetano, Gregorio and Vikram Maheshri. *Homophily And Sorting Within Neighborhoods.* 2015. Print. http://www.gregoriocaetano.net/resources/Research/Homophily _Neighborhoods.pdf

[18] The White House,.*Economic Report Of The President.* 2015. Print. https://www.whitehouse.gov/sites/default/files/docs/cea_2015_erp.pdf

[19] Corak, M. ?Do poor children become poor adults? Lessons from a cross-country comparison of generational earnings mobility?, in *Dynamics of Inequality and Poverty.* Emerald, pp. 143?188 http://www.emeraldinsight.com/doi/abs/10.1016/S1049-2585

[20] Corak, Miles. Generational Income Mobility In North America And Europe. Cambridge: Cambridge University Press, 2004. Print. https://books.google.com/books?hl=en&lr=&id=fQC7Pdr5FLMC&oi=fnd&pg =PA38&dq=solon+2004+inequality&ots=7oos66Unj9&sig =e_YpdevQgQ4JCuSgK43y2RPhkG0#v=onepage&q=solon%202004 %20inequality&f=false

[21] Andrews, Dan and Andrew Leigh. "More Inequality, Less Social Mobility". Applied Economics Letters 16.15 (2009): 1489-1492. Web. http://www.tandfonline.com/doi/abs/10.1080/13504850701720197#. V1sH4FUrKUk

[22] Rothstein, Richard. "The Racial Achievement Gap, Segregated Schools, And Segregated Neighborhoods: A Constitutional Insult". Race Soc Probl 7.1 (2014): 21-30. Web. http://www.epi.org/publication/the-racial-achievement-gap-segregated-schools-and-segregated-neighborhoods-a-constitutional-insult/

[23] Chetty, Raj, Nathaniel Hendren, and Lawrence Katz. "The Effects Of Exposure To Better Neighborhoods On Children: New Evidence From The Moving To Opportunity Experiment". Harvard University, 2015. Print. http://www.equality-of-opportunity.org/images/mto_exec_summary.pdf

[24] Chetty, Raj and Nathaniel Hendren. "The Impacts Of Neighborhoods On Intergenerational Mobility". Harvard University, 2015. Print. http://www.equality-of-opportunity.org/images/nbhds_exec_summary.pdf

[25] "Equality Of Opportunity". *Equality-of-opportunity.org.* N.p., 2016. Web. 1 July 2016. http://www.equality-of-opportunity.org/

[26] "ZIP Code Tabulation Areas (Zctas) - Geography - U.S. Census Bureau". *Census.gov.* N.p., 2016. Web. 1 July 2016. https://www.census.gov/geo/reference/zctas.html

[27] Sawhill, Isabel V. and Scott Winship. "Pathways To The Middle Class: Balancing Personal And Public Responsibilities". Social Genome Research Project (2012): 48-51. Print. http://www.brookings.edu/research/papers/2012/09/20-pathways-middle-class-sawhill-winship

[28] Chetty, Raj et al. Where Is The Land Of Opportunity? *The Geography Of Intergenerational Mobility In The United States.* Harvard University, UC-Berkeley and NBER, 2014. Print.http://www.equality-of-opportunity.org/images/mobility_geo.pdf

[29] Abt Associates,. *The Effects Of Neighborhood Change On New York City Housing Authority Residents.* The NYU Furman Center for Real Estate and Urban Policy, 2015. Print. http://www.nyc.gov/html/ceo/downloads/pdf/nns_15.pdf

[30] Page, Scott. *Diversity Powers Innovation.* 2007. Print. http://inclusive.nku.edu/content/dam/Inclusive/docs/diversity_powers_innovation.scot%20page.pdf

[31] Duranton, Gilles and Diego Puga. "Nursery Cities: Urban Diversity, Process Innovation, And The Life Cycle Of Products".*American Economic Review* 91.5 (2001): 1454-1477. Web. http://www.jstor.org/stable/2677933?seq=1#page_scan_tab_contents

[32] Van der Vegt, G. S. and O. Janssen. "Joint Impact Of Interdependence And Group Diversity On Innovation". *Journal of Management* 29.5 (2003): 729-751. Web. http://jom.sagepub.com/content/29/5/729.short

[33] Hero, Rodney E. Racial Diversity And Social Capital. New York: Cambridge University Press, 2007. Print.*Racial diversity and social capital*: Equality and community in America

[34] Lichter, D. T. (2012), Immigration and the New Racial Diversity in Rural America. Rural Sociology, 77: 3?35. doi: 10.1111/j.1549-0831.2012.00070.x Immigration and the New Racial Diversity in Rural America*

[35] Hopkins, Daniel J. "The Diversity Discount: When Increasing Ethnic And Racial Diversity Prevents Tax Increases".*The Journal of Politics* 71.1 (2009): 160-177. Web.The diversity discount: When increasing ethnic and racial diversity prevents tax increases

[36] HOWARD ECKLUND, ELAINE. "Models Of Civic Responsibility: Korean Americans In Congregations With Different Ethnic Compositions".*J Scientific Study of Religion* 44.1 (2005): 15-28. Web.Models of civic responsibility: Korean Americans in congregations with different ethnic compositions

[37] Maguire, Mike, Rodney Morgan, and Robert Reiner. *The Oxford Handbook Of Criminology.* Oxford: Oxford University Press, 2007. Print.Ethnicities, racism, crime and criminal justice

[38] "Do You Know Your Neighbors?".*Pew Research Center*. N.p., 2010. Web. 30 June 2016.http://www.pewresearch.org/daily-number/do-you-know-your-neighbors/

[39] Dunkelman, Marc J. *The Vanishing Neighbor*. Print. http://books.wwnorton.com/books/The-Vanishing-Neighbor/

[40] NYU Furman Center. The Furman Center for Real Estate and Urban Policy. The Changing Racial and Ethnic Makeup of New York City Neighborhoods http://furmancenter.org/files/sotc/The_Changing_Racial_and_Ethnic_Makeup_of_New_York_City_Neighborhoods_11.pdf

[41] Jonas, M. The downside of diversity- A Harvard political scientist finds that diversity hurts civic life. What happens when a liberal scholar unearths an inconvenient truth? The Boston Globe.Available :http://archive.boston.com/news/globe/ideas/articles/2007/08/05/the_downside_of_diversity/

[42] MacDonald, L. Median Income as a Better Measure of Development Progress?Nancy Birdsall and Christian Meyer. Available: http://www.cgdev.org/blog/median-income-better-measure-development-progress-nancy-birdsall-and-christian-meyer

[43] Haveman, J. Research Brief on Income Inequality in the San Francisco Bay Area. Silicon Valley Institute for Regional Studies. Available: https://www.jointventure.org/images/stories/pdf/income-inequality-2015-06.pdf

[44] Birdsall, N. and Meyer, C. The Median Is the Message: A Good-Enough Measure of Material Well-Being and Shared Development Progress. Center for Global Development, Working Paper 351, January 2014. Available: http://www.cgdev.org/sites/default/files/median-message-good-enough-measure-shared-development-progress_final_0.pdf

[45] Wolfson, M. C. (1997), DIVERGENT INEQUALITIES: THEORY AND EMPIRICAL RESULTS. Review of Income and Wealth, 43: 401?421. Doi: 10.1111/j.1475-4991.1997.tb00233.x. Available:http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4991.1997.tb00233.x/abstract

[46] Michael C. Wolfson. The American Economic Review Vol. 84, No. 2, Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association (May, 1994), pp. 353-358. Available: http://www.jstor.org/stable/2117858?seq=1#page_scan_tab_contents)

[47] Hurst E.,Li G. and Pugsley B. Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-Employed* Forthcoming in the Review of Economics and Statistics. October 2012. Available: https://www.newyorkfed.org/medialibrary/media/research/economists /pugsley/income_underreporting_10262012.pdf

[48] Maldonado C.,Mehrhoff R., Zawrotniak E., McNiff E., Rodriguez C., St. Preux B. Reporting of Billboard Income. New York State Office of the State Comptroller Thomas P. DiNapoli Division of State Government Accountability. Available: http://www.osc.state.ny.us/audits/allaudits/093013/11n2.pdf

[49] Marcelo Medeiros and Pedro H. G. Ferreira de Souza. The Rich, the Affluent and the Top Incomes: a Literature Review. IRLE WORKING PAPER #105-14 April 2014 Available: http://www.irle.berkeley.edu/workingpapers/105-14.pdf

[50] Jeffrey C. Moore , Linda L. Stinson , and Edward J. Welniak, Jr. Income Measurement Error in Surveys: A Review. Census Bureau. Available: https://www.census.gov/srd/papers/pdf/sm97-05.pdf

[51] New York City Housing Lottery. Available: http://www1.nyc.gov/nyc-resources/service/2076/new-york-city-housing-lottery

[52] New York City Affordable Housing. Available: http://www1.nyc.gov/nyc-resources/service/1021/affordable-housing

[53] Berube, A. Metropolitan Opportunity Series, Number 51 of 67 Paper, February 20, 2014. Available: http://www.brookings.edu/research/papers/2014/02/cities-unequal-berube

[54] Samuel Dastrup, Ingrid Ellen, Anna Jefferson, Max Weselcouch, Deena Schwartz, Karen Cuenca. The Effects of Neighborhood Change on New York City Housing Authority Residents. NYC Center for Economic Opportunity. Available: http://www.nyc.gov/html/ceo/downloads/pdf/nns_15.pdf

[55] Instagram Press Release, Instagram Today: 500 Million Windows to the World. June 21, 2016. Available: http://blog.instagram.com/post/146255204757/160621-news

[56] Instagram Press Release, Celebrating a Community of 400 Million. September 22, 2015. http://blog.instagram.com/post/129662501137/150922-400million

[57] Instagram API Documentation. Available: https://www.instagram.com/developer/

[58] Daniel Goddemeyer,Moritz Stefaner,Dominikus Baur, Lev Manovich. The On Broadway Project. Available: http://on-broadway.nyc

[59] Lev Manovich, Moritz Stefaner, Mehrdad Yazdani, Dominikus Baur, Daniel Goddemeyer, Alise Tifentale, Nadav Hochman, Jay Chow. SelfieCity. Available: http://selfiecity.net

[60] Nadav Hochman, Lev Manovich. Zooming Into an Instagram City: Reading the local through social media. Available: http://firstmonday.org/ojs/index.php/fm/article/view/4711/3698

[61] Bakshi, Shamma, Gilbert. Faces engage us: photos with faces attract more likes and comments on Instagram. CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Pages 965-974. Available: http://dl.acm.org/citation.cfm?id=2557403&dl=ACM&coll=DL&CFID=800238042&CFTOKEN=96656337

[62] Van House, N. A., Davis, M., Takhteyev, Y., Ames, M., Finn, M. The Social Uses of Personal Photography: Methods for Projecting Future Imaging Applications, 2004. http://www.sims.berkeley.edu/- vanhouse/vanhouseetal2004b.pdf

[63] Nancy Van House, Marc Davis, Morgan Ames, Megan Finn, Vijay Viswanathan. The Uses of Personal Networked Digital Imaging: An Empirical Study of Cameraphone Photos and Sharing. Available: http://people.ischool.berkeley.edu/ vanhouse/van_house_chi_short.pdf

[64] Haxby, J., Hoffman, E., and Gobbini, M. The distributed human neural system for face perception. *Trends in cognitive sciences 4*, 6 (2000), 223?233.

[65] Viola, p., and jones, m. j. Robust real-time face detection. *International Journal of Computer Vision 57*, 2 (2004), 137?154.

[66] Wright, J., Yang, A., Ganesh, A., Sastry, S., and Ma, Y. Robust face recognition via sparse representation. *Pattern analysis and machine intelligence, IEEE Transactions on 31*, 2 (2009), 210?227.

[67] PILAB Website, Bogazici University, Department of Computer Science. Perpetual Intelligence Laboratory (PILAB)

[68] Cylab Website, Carnegie Mellon University, Security and Privacy Insititute, CyLab

[69] VASC Website, Carnegie Mellon University, Robotics Institute, Vision and Autonomous Systems Center (VASC)

[70] "CSU Face Recognition Homepage". Cs.colostate.edu. N.p., 2016. Web. 30 June 2016.http://www.cs.colostate.edu/facerec/index10.php

[71] "Map Your World's Data ? Cartodb". Cartodb.com. N.p., 2016. Web. 1 July 2016. https://cartodb.com/

[72] Roberts, Sam. "Poverty Rate Is Up In New York City, And Income Gap Is Wide, Census Data Show". *The New York Times* 2013: n. pag. Print. Link: http://www.nytimes.com/2013/09/19/nyregion/poverty-rate-in-city-rises-to-21-2.html

[73] Koester, Daniel. "FIPA - FIPA". *Face.cs.kit.edu.* N.p., 2016. Web. 30 June 2016.Facial Image Processing and Analysis Group (FIPA)

[74] (DADS), Data. "American Factfinder - Results". *Factfinder.census.gov.* N.p., 2016. Web. 30 June 2016.http://factfinder.census.gov/bkmk/table/1.0/en/ACS/14_5YR/DP05/1600000US3651000

[75] Duggan, Maeve. "The Demographics Of Social Media Users". *Pew Research Center: Internet, Science & Tech.* N.p., 2015. Web. 30 June 2016.http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

[76] Mislove et al. Understanding the Demographics of Twitter Users. Available: http://dougleschan.com/the-recruitment-guru/wp-content/uploads/2014/01/Understanding-the-Demographics-of-Twitter-Users-Jukka-Pekka-....pdf

[77] Gayo-Avello, Daniel, Panagiotis T. Metaxas, and Eni Mustafaraj. Limits Of Electoral Predictions Using Twitter. 2011. Print.https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper /viewFile/2862/3254 (Gayo-Avello (2011) is an exception.)

[78] Ehsan Mohammady and Aron Culotta Department of Computer Science Illinois Institute of Technology. Using county demographics to infer attributes of Twitter users Available: http://www.aclweb.org/anthology/W14-2702

[79] http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0158161 Boy JD, Uitermark J (2016) How to Study the City on Instagram. PLoS ONE 11(6): e0158161. doi:10.1371/journal.pone.0158161

[80] Michael J. Paul and Mark Dredze Human Language Technology Center of Excellence Department of Computer Science Johns Hopkins University. A Model for Mining Public Health Topics from Twitter http://www.spatialcapability.com/Library/Capstone/Social%20Media/A%20Model%20for%20Mining%20Public%20Health%20Topics%20from %20Twitter.pdf (Dredze, 2012)

[81] O?Connor, Brendan et al. From Tweets To Polls: Linking Text Sentiment To Public Opinion Time Series. Carnegie Mellon University, 2010. Print.http://homes.cs.washington.edu/ nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf (O?Connor et al., 2010)

[82] Gopinath, Shyam, Jacquelyn S. Thomas, and Lakshman Krishnamurthi. "Investigating The Relationship Between The Content Of Online Word Of Mouth, Advertising, And Brand Performance". Marketing Science 33.2 (2014): 241-258. Web.https://scholar.google.com/citations?view_op=view_citation&hl=en&user=DDpOZqIAAAAJ&citation_for_view=DDpOZqIAAAAJ: UeHWp8X0CEIC (Gopinath et al., 2014)

[83] Islam, Mohammad T., Scott Workman, and Nathan Jacobs. *FACE2GPS: ESTIMATING GEOGRAPHIC LOCATION FROM FACIAL FEATURES.* Department of Computer Science, University of Kentucky. Print.http://cs.uky.edu/ tarik/papers/face2gps.pdf

[84] "ONOMAP". ONOMAP. N.p., 2016. Web. 30 June 2016.http://onomap.co/

[85] "Talk ? Personality And Homophily In Online Social Networks - FTS". *Cs.cf.ac.uk.* N.p., 2016. Web. 30 June 2016.http://www.cs.cf.ac.uk/fts/talk-personality-and-homophily-in-online-social-networks/

[86] "San Antonio — Food, Nightlife, Entertainment". *Foursquare.com.* N.p., 2016. Web. 30 June 2016.https://foursquare.com

[87] G. Ottaviano, G. Peri, ?The Economic Value of Cultural Diversity: Evidence from US Cities,? National Bureau of Economic Research, NBER Working Paper No. 10904, Nov 2004. Available: http://www.nber.org/papers/w10904

[88] G. Ranis, ?Diversity of Communities and Economic Development: An Overview,? Economic Growth Center, Yale University, New Haven, CT. Available: http://www.econ.yale.edu/growth_pdf/cdp1001.pdf

[89] A. Alesina et al. (2013, August 22).*Immigration, Diversity and Economic Prosperity* [Online]. Available: http://www.voxeu.org/article/immigration-diversity-and-economic-prosperity

[90] R. Florida. (2011, December 12).*How Diversity Leads to Economic Growth* [Online]. Available: http://www.citylab.com/work/2011/12/diversity-leads-to-economic-growth/687/

[91] Badal, Sangeeta Bharadwaj. "The Business Benefits Of Gender Diversity". *Business Journal* (2014): n. pag. Print.http://www.gallup.com/businessjournal/166220/business-benefits-gender-diversity.aspx

[92] Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981), 1029-1031

[93] "Predicting Next Month Crime From Mobile Network Activity - "Datathon For Social Good" Competition In London — Fondazione Bruno Kessler". *Fbk.eu*. N.p., 2016. Web. 30 June 2016. http://www.fbk.eu/news/predicting-next-month-crime-mobile-network-activity-datathon-social-good-competition-london

[94] Levine, Sheen S. et al. "Ethnic Diversity Deflates Price Bubbles".*Proceedings of the National Academy of Sciences* 111.52 (2014): 18524-18529. Web.http://www.pnas.org/content/111/52/18524.abstract

[95] 9. D. Smith *et al., Mapping L.A.* [Online]. Available: http://maps.latimes.com/about/

[96] *ArcGIS Neighborhood Diversity Index for Chicago*, Esri USA Diversity Index and Demographics, 2012. Available: http://www.arcgis.com/home/item.html?id=62317e37c8714f51b869d1172b27dd73

[97] "Mapbox: Locals & Tourists". *Mapbox.com*. N.p., 2016. Web. 30 June 2016.https://www.flickr.com/photos/walkingsf/4671594023/in/album-72157624209158632/][http://on-broadway.nyc/

[98] "NYC Statistics". *nycgo.com.* N.p., 2016. Web. 30 June 2016. http://www.nycandcompany.org/research/nyc-statistics-page

[99] "Travel And Tourism".*NYCEDC.* N.p., 2016. Web. 30 June 2016.http://www.nycedc.com/economic-data/travel-and-tourism

[100] Schaller, Gary. Entropy Measurements Of Economic And Racial Inequality In New Jersey. Print.https://books.google.com/books?id=fgGA8plP7pkC&pg=PT4&lpg=PT4&dq=shannon+entropy+to+measure+race+diversity&source

=bl&ots=TPIyZr_hWq&sig=TKvttYw4knUSNSFMdwrZD-Z60Eo&hl=en&sa=X&ved=0ahUKEwiyhIW2pMTNAhWFlB4KHdV3AGIQ6AEIKzAC#v=onepage&q=shannon%20entropy%20to%20measure%20race%20diversity&f=false

[101] SCHILLING, MARK. "Measuring Diversity in the United States." *Math Horizons* 9.4 (2002): 29-30. Web.https://www.csun.edu/ hcmth031/MDITUS.pdf

[102] Iceland, John. *The Multigroup Entropy Index.* U.S. Census Bureau, 2004. Print.https://www.census.gov/housing/patterns/about/multigroup_entropy .pdf

[103] Roberto, Elizabeth. *Measuring Inequality And Segregation.* Department of Sociology, Princeton University, Princeton, NJ, USA, 2015. Print.https://arxiv.org/pdf/1508.01167.pdf

[104] Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27, 379?423 and 623?656.

[105] Acevedo-Garcia, D. et al. "Toward A Policy-Relevant Analysis Of Geographic And Racial/Ethnic Disparities In Child Health". *Health Affairs* 27.2 (2008): 321-333. Web.http://www.hec.unil.ch/documents/seminars/deep/233.pdf

[106] "MORPH — I3S". *Faceaginggroup.com.* N.p., 2016. Web. 1 July 2016. http://www.faceaginggroup.com/morph/

[107] Tach, Laura M. Diversity, Inequality, And Microsegregation: *Dynamics Of Inclusion And Exclusion In A Racially And Economically Diverse Community.* U.S. Department of Housing and Urban Development, 2014. Print.https://www.huduser.gov/portal/periodicals/cityscpe/vol16num3/ch1 .pdf

[108] Sharkey, Patrick. *Stuck In Place.* Chicago: The University of Chicago Press, 2013. Print. Stuck in Place, Patrick Sharkey

[109] Acevedo-Garcia, D. et al. "Toward A Policy-Relevant Analysis Of Geographic And Racial/Ethnic Disparities In Child Health". *Health Affairs* 27.2 (2008): 321-333. Web.http://www.ncbi.nlm.nih.gov/pubmed/18332486

[110] Jacobs, Jane. *The Death And Life Of Great American Cities.* New York: Vintage Books, 1992. Print.The Importance of Death and Life of Great American Cities (1961)

[111] Nyden, Philip. *Weaving Social Seams: Stable, Racially And Ethnically Diverse Communities As Places Of Social Innovation.* 2012. Print.https://www.semanticscholar.org/paper/Weaving-Social-Seams-Stable-Racially-and-Nyden/68500f408305a5dde1bef72853291b1d3a63f9fd/pdf

[112] CHASKIN, ROBERT J and MARK L JOSEPH. "SOCIAL INTERACTION IN MIXED-INCOME DEVELOPMENTS: RELATIONAL EXPECTATIONS AND EMERGING REALITY". *Journal of Urban Affairs* 33.2 (2011): 209-237. Web.http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9906.2010.00537.x/abstract?userIsAuthenticated=false& deniedAccessCustomisedMessage=

[113] Joseph, Mark L., Robert J. Chaskin, and Henry S. Webber. 2007. The theoretical basis for addressing poverty through mixed-income development. *Urban Affairs Review* 42(3): 369-409. Read the article

[114] Fry, Richard and Paul Taylor. "The Rise Of Residential Segregation By Income". *Pew Research Center* (2012): n. pag. Print.http://www.pewsocialtrends.org/2012/08/01/the-rise-of-residential-segregation-by-income/

[115] *Exploring Racial Segregation And Income Inequality Patterns And Relationships.* U.S. Department of Housing and Urban Development, 2016. Print.https://www.huduser.gov/portal/pdredge/pdr_edge_research_032212.html

[116] Weinberg, Daniel H. U.S.*Neighborhood Income Inequality In The 2005?2009 Period.* U.S. Census Bureau, 2011. Print.https://www.census.gov/prod/2011pubs/acs-16.pdf

[117]]Arnott, David A. "Unequal Cities: Where Rich And Poor Live Within Blocks Of Each Other". *The Business Journals* (2015): n. pag. Print.http://www.bizjournals.com/bizjournals/news/2015/11/19/unequal-cities-where-rich-and-poor-live-nearby.html

[118] Roberts, Sam. "In Surge In Manhattan Toddlers, Rich White Families Lead Way". *The New York Times* 2007: n. pag. Print.http://www.nytimes.com/2007/03/23/nyregion/23kid.html?_r=0

[119] Bellafante, Ginia. "Baby Boom Among New York?S Affluent". *The New York Times* 2015: n. pag. Print.http://www.nytimes.com/2015/05/03/nyregion/baby-boom-among-new-yorks-affluent.html

[120] Desilver, Drew. "The Many Ways To Measure Economic Inequality". Pew Research Center (2013): n. pag. Web. 30 June 2016.http://www.pewresearch.org/fact-tank/2015/09/22/the-many-ways-to-measure-economic-inequality/

[121] Board of Governors of the Federal Reserve System,. Changes In U.S. *Family Finances From 2010 To 2013: Evidence From The Survey Of Consumer Finances.* Federal Reserve, 2014. Print.http://www.federalreserve.gov/pubs/bulletin/2014/pdf/scf14.pdf

[122] http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9906.2007.00333.x/full HAYS, R. A. and KOGL, A. M. (2007), NEIGHBORHOOD ATTACHMENT, SOCIAL CAPITAL BUILDING, AND POLITICAL PARTICIPATION: A CASE STUDY OF LOW- AND MODERATE-INCOME RESIDENTS OF WATERLOO, IOWA. Journal of Urban Affairs, 29: 181?205. doi:10.1111/j.1467-9906.2007.00333.x