Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Towards Measuring Fine-Grained Diversity Using Social Media Photographs

Vivek K. Singh,[1] **Saket Hegde,**[1] **Akanksha Atrey**[2]
[1]Rutgers University, [2]State University of New York, Albany
{vs451, ssh105}@scarletmail.rutgers.edu; aatrey@albany.edu

Abstract

Diversity is an important socio-economic construct that influences multiple aspects of human lives from the prosperity of a city to corporate earnings and from criminal justice to health and social engagement. Large, heavily populated urban areas can be highly diverse at the city or even neighborhood level, but we know very little about how much people from diverse demographics (such as age and race) interact with each other. Previous work has shown that photos are important in social relationships. The growing presence of photos online and on social media, therefore presents a unique opportunity to study diversity in interactions. In this paper, we explore a novel approach to measure *p-diversity* i.e. a *personal*, *photo-level* diversity metric computed using social media data. Specifically, we focus on Instagram photos of multiple people interacting, and employ automatic methods for race, age, and gender estimation to quantify mixing in such photos. We compare and contrast this new measure of diversity with traditional (i.e. census-based) metrics using a dataset for New York City. Results obtained motivate the use of social media photos to complement census data to develop cheaper, faster, mechanisms for studying diversity and applying them in social, economic, political, and urban planning contexts.

Diversity is an important socio-economic construct that has associations with multiple aspects of human life including commerce, innovation, well-being, criminal justice, civic responsibilities and health among others (Page 2007). 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 (Maly 2000). 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 (Silver 2015).

Census data (such as from the American Community Survey) provide an important public service by collecting and preserving demographics data going back for decades. Despite its inherent value, census based data collection and analysis has several shortcomings including being expensive, time consuming and labor-intensive. In much previous work, it has also been argued that studying informal interactions is important to understand social diversity, but it was hard to study them all along (e.g. (Hays and Kogl 2007)). 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.

Previous work has shown that photos are important in social relationships. The content of photos shows who is part of a group and telling stories about photos helps nurture relationships (Van House et al. 2005). Hence, here we assume that people sharing the same "frame" in a single photo have some kind of interconnection among themselves and proceed to define a new fine grained metric for measuring diversity. While similar in spirit to the traditional measures of diversity, the proposed *p-diversity* metric quantifies the level intermixing occurring between people of different demographic descriptions in their *personal* spaces as observed via social media *photographs*.

The proposed *p-diversity* metric allows the notion of diversity to vary in (near) real-time rather than assumed to be a constant between census updates and focuses on diversity of relationships between people rather than their residential addresses. In this short article, we describe a methodology to compute this new *p-diversity* metric using social media (Instagram) data and undertake a case-study analysis by comparing the results obtained by the proposed metric with those obtained via traditional census based metrics in New York City. The obtained results motivate and ground the use of social media *photographs* for studying diversity.

Methodology

The *p-diversity* metric: A popular method, frequently cited in studies of diversity in the United States and at varying levels of granularity is Shannon Entropy also known as Shannon's Diversity Index (Schilling 2002).

It is calculated as follows:

$$E = -\sum_{i=1}^{R} p_i \cdot ln(p_i) \quad (1)$$

where p_i is the proportion of individuals belonging to the i^{th} demographic description (e.g. male) in the dataset of interest and R is the number of distinct 'bins' (e.g. male and female) considered. *The p-diversity score for a neighborhood*

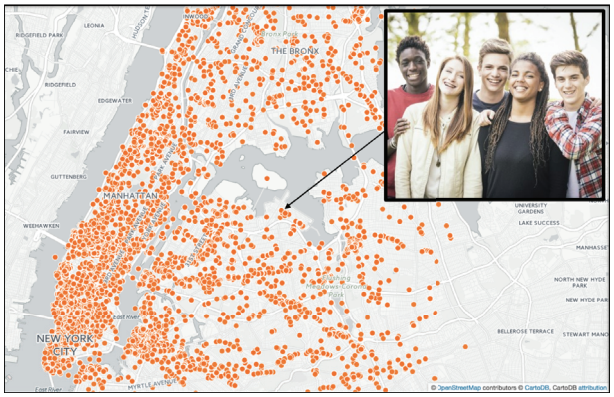

Figure 1: Image of New York City Area showing location of sampled photos over a 2 week period (made using CartoDB visualization. *Inset*: A sample image highlighting the social mixing (gender, race) as observable via social media posts. Note: Image is not from the data set for privacy reasons.

is defined as the median level of (Shannon entropy-based) diversity as observed in photographs shared from that neighborhood based on each demographic facet (age, gender, and race).

Here we calculate three measures of Shannon Entropy for each individual image where the i is defined on bins based on gender (male and female), age (< 18; 18-35; > 35), and race (White, Black, and Asian). Computing such diversity scores for each image in a neighborhood allows us to quantify the median level of diversity in terms of age, gender, and race in each neighborhood, thus allowing for computation of the *p-diversity* score in terms of age, gender, and race for the neighborhood. For race, we consider the three major racial groups in New York City-White, Black and Asian (this is based on ACS (American Community Survey) Demographics and Housing Estimates: 2010-2014 5-Year Estimates).

Defining Neighborhoods: We use Zip Code Tabulation Areas (ZCTAs) as our definitions for neighborhoods. ZCTAs 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 readily available.

Variable of interest - Income Inequality: Given the focus on income inequality in recent political discussions, we consider income inequality as a socio-economic variable of interest that may be related to diversity. In our analysis, we utilize the ratio of mean to median income to quantify income inequality. Determining the ratio of mean to median incomes provides a simple measure of the skewness in the distribution, has long been used in literature (Birdsall and Meyer 2015) and is less sensitive than Gini coefficients to under-reporting of income at the top of distributions.

Data Corpus

Data Collection and Cleaning: To compute the diversity metrics mentioned above, 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. We first selected a location at random within a latitude-longitude range encompassed by the city. Next, the most recent photos in the immediate vicinity were 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 (explained later), 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. The median number of faces sampled from each neighborhood was 114, with a range of 30 to 657 (after excluding those ZCTAs with a lower number of detected faces in photos). Despite the exclusion criteria, we are able to include 79 ZCTAs in our analysis. As seen in Figure 2 the ZCTAs matching the abovementioned criteria cover most of New York City.

Facial Analysis: We determine the demographics of individual photos using facial analysis from the publicly available face detection Face++ API. This API provides age, race, and gender estimates for each face found in the image. Since our goal is to quantify social mixing between people we filter for only those images with multiple (i.e. 2 or more faces). For each photo, we calculate the values of the variables provided by the Face++'s API (we use a methodology quite similar to (Bakhshi, Shamma, and Gilbert 2014)) and compute the diversity scores as shown in Eq. 1.

Validation of Race Classification: Despite being validated for high accuracy in age and gender classification as reported in previous works (Wang, Li, and Luo 2016) (Bakhshi, Shamma, and Gilbert 2014), the Face++ API has not been validated for race classification. In order to do so, we use the pre-labeled standardized MORPH database. 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 (Wang, Li, and Luo 2016). 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 (Wang, Li, and Luo 2016).

Removing Tourist Photos: 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. We employ a method similar to (Fischer 2014) to remove tourist photos. We examine the posts of users that posted photos included in our sample set. If over 50% of recent user posts (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.

Census Data: For comparison, we also calculate measures of neighborhood level diversity using Census figures. We do so by using ACS measurements on race, gender and

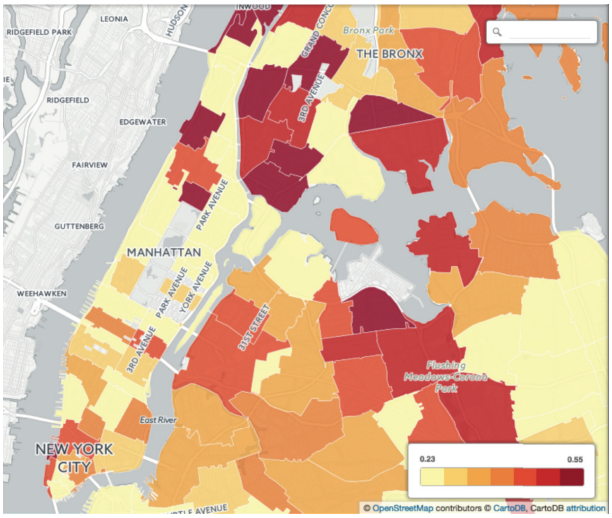

Figure 2: Heat Map of New York City showing the Racial Diversity in Photo-level interactions (*p-diversity:race*) in various neighborhoods

age for each neighborhood and using Shannon Entropy metric as defined in Eq. 1.

Case Study and Results

As a case study, we compute the proposed *p-diversity* scores for different neighborhoods in New York City. A heat map of the *p-diversity* metric (based on race) as computed using the Instagram dataset is shown in Figure 2.

Upon comparison of diversity and *p-diversity* scores for each neighborhood, we find that measuring diversity metrics based on personal photo-based interactions results in a very different portrait of the city. In some cases, the difference is as high as 100% due to zero entropy in the level of interaction at the photo level for that neighborhood. In general, *p-diversity* is almost always lower than census based diversity for all three measures (age, gender, race).

Type	Diversity (census-based)	*p-Diversity* (personal, photo-based)
Age	0.99	0.58
Gender	0.69	0.33
Race	0.69	0.27

Table 1: Average levels of *diversity* and *p-diversity* scores (in terms of age, gender, and race) for New York neighborhoods

Table 1 summarizes the average values for diversity and *p-diversity* scores obtained for different neighborhoods in New York City. Clearly, the *p-diversity* scores are much lesser on average than the diversity scores. A Wilcoxon-test for pairwise comparison of diversity and *p-diversity* scores indicated significant differences in the means for all three facets (age, gender, and race) at a significance threshold of $p < 0.01$. This difference is obvious in all three metrics and most pronounced in terms of race. This suggests that level of diversity as measured by census may be a quite optimistic in terms of quantifying the actual personal interactions taking place between the residents.

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 (Howard Ecklund 2005) and block level segregation is often not indicative of these pervasive social barriers. The lack of photo based entropy can be explained by the phenomenon of homophily as also noted by (Caetano and Maheshri 2016). While leaving further analysis of reasons outside the scope of this article, the results motivate the need and the value for a more fine-grained metric (and easy to deploy methodology) for understanding the mixing of populations.

The different interpretations of diversity may also influence the associations studied and reported between social outcomes and the notion of diversity. For example, income inequality is an important social outcome often studied in the context of diversity of a neighborhood. However, employing multilinear regression to determine the associations between income inequality and the two variants of diversity metrics (i.e. photo based entropy and census based entropy) yielded very different results.

Using the three photo based metrics of diversity (i.e. *p-diversity*) as predictors, we find that a combined model is significantly (though weakly) predictive of income 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 for an ANOVA test indicating at least one of the constituent predictors is significant in the association. However, a similar model with census-based diversity metrics yielded a model that was found to be *not* significantly predictive of inequality based on an ANOVA test. Not surprisingly, the coefficients for individual features were also not found to be significantly related.

The coefficients for the multilinear regression for both photo-based diversity metrics as well as census-based metrics are summarized in Table 2.

Diversity Measure	β (Social Media)	β (Census)
Age	0.301**	0.001
Gender	0.007	-0.160
Race	-0.272*	-0.149

Table 2: Diversity measures (photo-based entropy and census-based entropy) as predictors of Inequality based on multilinear regression . * indicates that the coefficient is significant at 0.05 level; **indicates significance at 0.01 level

This result again underscores the methodological motivation behind this work. For example, the notion of diversity has been widely studied for its interconnections with socio-economic factors like income-inequality, crime, disease, political participation and so on. However, the use of a similar but arguably more personal-space driven metric as defined here might yield very different results from those obtained via census-based diversity metrics. For instance, as seen in in Table 2, looking at census data alone, one might assume

there is no significant association between racial diversity in urban neighborhoods and income inequality, but the fact that we see a significant correlation between racial diversity in fine grained interactions and income inequality suggests that digging deeper might reveal some interesting patterns.

The Census Bureau itself has studied how income inequality varies spatially over the United States, but at spatial resolution that is much coarser than our metric (Weinberg 2011). 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.

Since the last decade, researchers have increasingly turned to social media analysis to complement traditional survey methods in areas such as public health, politics, and marketing. Multiple recent efforts have looked at text-based social media for demographic analyses (e.g. (Arnaboldi et al. 2016)). *However, there is as yet very little work on utilizing social media photographs for urban-scale diversity studies*. The growing popularity of photo-based social media as well as image analysis APIs presents a unique opportunity for researchers seeking to engage in urban studies and develop visualizations of cityscapes (e.g. (Hochman and Manovich 2013)). The current work hence encourages a more in-depth exploration of visual data both for studying diversity as well as to gain a better understanding of different phenomena of social interest.

Despite being cheaper and more real-time, social media based approaches such as ours are not without flaws. Arnaboldi et al. (2016) 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 in the methodology section. Further, the users (or rather the people captured in the photos by Instagram users) may not be representative of the underlying population. Adoption rates of Instagram, the consideration of only public, geotagged photos, and the privacy implications of photos used, are some other factors that need to be considered. With these limitations 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 one of the first attempts to leverage social media photos to not only study diversity but also understand the associated socio-economic phenomena. Our hope is that our group and the wider research community will be able to quantify and/or ameliorate these shortcomings in the future, thus paving way for a new way to study and understand diversity and mixing in the physical world by the use of geolocated social photographs.

Conclusion

This work motivates and grounds the use of a new social media photo-based methodology to study diversity as the mixing between people with different demographic descriptions in a neighborhood. A case study based on Instagram data in New York City suggests that a fine-grained representation of social mixing as captured via social media photos may paint a very different picture of the city in comparison to traditional census-based diversity metrics. For example, the level of mixing was found to be consistently lower at personal photo-scale spatial resolutions than what would be expected by looking at coarse neighborhood level diversity metrics. The results motivate the use of social media photos to complement census based metrics to study multiple phenomena in social, economic, political, and urban planning contexts.

References

Arnaboldi, M.; Brambilla, M.; Cassottana, B.; Ciuccarelli, P.; Ripamonti, D.; Vantini, S.; and Volonterio, R. 2016. Studying multicultural diversity of cities and neighborhoods through social media language detection. In *Tenth International AAAI Conference on Web and Social Media*.

Bakhshi, S.; Shamma, D. A.; and Gilbert, E. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proc. ACM CHI*, 965–974. ACM.

Birdsall, N., and Meyer, C. J. 2015. The median is the message: A good enough measure of material wellbeing and shared development progress. *Global Policy* 6(4):343–357.

Caetano, G., and Maheshri, V. 2016. Gender homophily and segregation within neighborhoods.

Fischer, E. 2014. Mapbox. Online: https://www.mapbox.com/bites/00245/locals/.

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(2):181–205.

Hochman, N., and Manovich, L. 2013. Zooming into an instagram city: Reading the local through social media. *First Monday* 18(7).

Howard Ecklund, E. 2005. Models of civic responsibility: Korean americans in congregations with different ethnic compositions. *Journal for the Scientific Study of Religion* 44(1):15–28.

Maly, M. T. 2000. The neighborhood diversity index: a complementary measure of racial residential settlement. *Journal of Urban Affairs* 22(1):37–47.

Page, S. 2007. Diversity powers innovation. Online: https://www.americanprogress.org/.

Schilling, M. 2002. Measuring diversity in the united states. *Math Horizons* 9(4):29–30.

Silver, N. 2015. The most diverse cities are often the most segregated. Online: https://fivethirtyeight.com/features/the-most-diverse-cities-are-often-the-most-segregated/.

Van House, N.; Davis, M.; Ames, M.; Finn, M.; and Viswanathan, V. 2005. The uses of personal networked digital imaging: an empirical study of cameraphone photos and sharing. In *CHI'05 extended abstracts on Human factors in computing systems*, 1853–1856. ACM.

Wang, Y.; Li, Y.; and Luo, J. 2016. Deciphering the 2016 us presidential campaign in the twitter sphere: A comparison of the trumpists and clintonists. *arXiv preprint arXiv:1603.03097*.

Weinberg, D. H. 2011. Us neighborhood income inequality in the 2005–2009 period. *American Community Survey report. Washington, DC: US Census Bureau*.