

EigenTransitions with Hypothesis Testing: The Anatomy of Urban Mobility

Ke Zhang

School of Information Sciences
University of Pittsburgh
kez11@pitt.edu

Yu-Ru Lin

School of Information Sciences
University of Pittsburgh
yurulin@pitt.edu

Konstantinos Pelechris

School of Information Sciences
University of Pittsburgh
kpele@pitt.edu

Abstract

Identifying the patterns in urban mobility is important for a variety of tasks such as transportation planning, urban resource allocation, emergency planning etc. This is evident from the large body of research on the topic, which has exploded with the vast amount of geo-tagged user-generated content from online social media. However, most of the existing work focuses on a specific setting, taking a statistical approach to describe and model the observed patterns. On the contrary in this work we introduce *EigenTransitions*, a spectrum-based, generic framework for analyzing spatio-temporal mobility datasets. *EigenTransitions* capture the anatomy of the aggregate and/or individuals' mobility as a compact set of latent mobility patterns. Using a large corpus of geo-tagged content collected from Twitter, we utilize *EigenTransitions* to analyze the structure of urban mobility. In particular, we identify the *EigenTransitions* of a flow network between urban areas and derive hypothesis testing framework to evaluate urban mobility from both temporal and demographic perspectives. We further show how *EigenTransitions* not only identify latent mobility patterns, but also have the potential to support applications such as mobility prediction and inter-city comparisons. In particular, by identifying neighbors with similar latent mobility patterns and incorporating their historical transition behaviors, we proposed an *EigenTransitions*-based *k*-nearest neighbor algorithm, which can significantly improve the performance of individual mobility prediction. The proposed method is especially effective in "cold-start" scenarios where traditional methods are known to perform poorly.

Introduction

Urban and transportation planners, as well as, city officials have been trying to understand the way people act and behave in our cities for many years now. This will allow them to design cities that can deliver a livable, resilient and sustainable urban environment that is relevant to the city-dwellers' needs. Identifying the pulse of a city through the mobility of its dwellers and visitors has been central to geographical and social sciences as well as to urban and transportation planning since the seminal work on migration from Ravenstein (Ravenstein 1885). Nevertheless, it is only recently that an unprecedented amount of data on urban activ-

ities has become available to researchers and can facilitate these efforts.

Using data from cellular networks and geo-tagged social media content a large volume of research has attempted to build models that describe the statistical properties of urban human mobility. Contrary to existing work on modeling of the statistical properties of the urban human mobility patterns, in this work we aim to provide a generic framework for analyzing mobility data that is able to also tie the mobility with the context within which it emerges. These patterns are affected by the underlying urban geography (Isaacman et al. 2010; Noulas et al. 2012), and are also shaped by the activities possible in the various parts of the city as well as the dwellers' interests. Hence, our study context can refer either to characteristics of the dweller's themselves (e.g., demographic information, interests etc.) or to the urban form of a neighborhood in the city.

As a proxy for the urban mobility we use geo-tagged content generated from Twitter users. Using the transitions observed we build a network between urban areas (e.g., neighborhoods) in the city that can reveal their underlying connectivity. We further propose a generic, spectrum-based method, *EigenTransitions*, that can analyze and capture the network dynamics generated by the underlying human mobility by reducing its effective dimensionality. *EigenTransitions* utilize Principal Component Analysis (PCA) as its core building block to identify the latent urban transition patterns. These mobility patterns are often shaped by individuals with their regular travel needs and interests across space and time. Depending on the transition matrix on which we apply PCA we can analyze different aspects of the mobility traces and in different (temporal and spatial) granularities.

As an example, Figure 1 presents the original transition flow between neighborhoods in New York City (NYC), where we can see the incoming and outgoing transitions to/from Manhattan dominate the urban flow. Figure 2 shows the four major *EigenTransitions* when we apply PCA on a transition matrix where each row corresponds to a day and each column is a transition between specific neighborhoods. The mobility pattern underlying component (a) resembles the original flow pattern which essentially explains the most popular transition pattern that most people tend to follow. Components (b)-(d) capture less popular but still

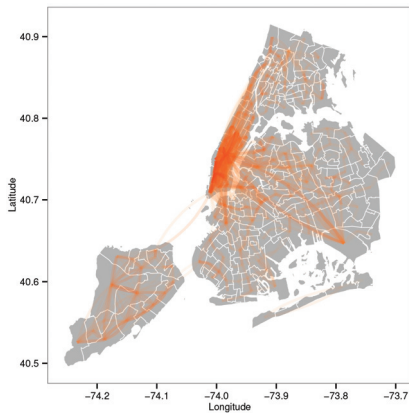


Figure 1: Human urban transition flow between NYC neighborhoods. The incoming and outgoing transitions to/from Manhattan dominate the structure of urban mobility.

strong transition flows to other urban areas. For example, component (c) represents people visiting Central Park. Since this leisure activity might be taking place during specific days only (e.g., weekends) this pattern is less popular overall as compared to component (a).

Using *EigenTransitions* we show that we can identify differences in the mobility of sub-populations that are not observable when using the original, high, dimensionality. By leveraging *EigenTransitions* into a rigorous hypothesis testing framework, we are able to identify mobility differences between different demographic parts of the population. We further present that *EigenTransitions* are able to facilitate location prediction, especially in cold-start scenarios, i.e., predicting transitions that have never been observed in the past. In particular, we propose an *EigenTransitions*-based nearest neighbor algorithm that leverages the mobility behavior of similar neighbors in the latent space. Our experiments using data from two cities show that our proposed algorithm can significantly improve the prediction performance with the help of only a small fraction of neighbors among the whole population. Furthermore, we show how *EigenTransitions* can be used to compare (and group) different cities based on their mobility patterns. Inter-city comparisons are important in order to understand how solutions can be transferable between cities. For example, cities with similar mobility patterns can potentially benefit from sharing ideas and solutions to transportation problems.

The **key contributions** of this work include: (1) We propose a generic framework, *EigenTransitions*, to identify the latent structure of large-scale urban transition flow patterns. (2) Through leveraging *EigenTransitions* into a rigorous hypothesis testing framework, we are able to identify demographic and temporal differences in the mobility dynamics that are not *visible* in the original data. In particular, we show statistically significant differences exist in the transition flow of different gender and ethnicity groups. Moreover, temporal differences are also identified. (3) We further apply *EigenTransitions* in location prediction

application and demonstrate its effectiveness over baseline methods, with a particular improvement over cold-start scenarios.

Moreover, to showcase the generalizability of our proposed method, we further provide results on using *EigenTransitions* to compare different cities with respect to their mobility predictability. Our preliminary results demonstrate that *EigenTransitions* are not only effective in sub-populations comparisons but can be used in cross-population comparison and more border context.

Related Work

In this section we will review studies related to our work.

Urban Mobility Literature: Despite the long interest in urban mobility, it is only recently, with the advancements in mobile technology and computing, that we have been able to obtain large-scale, real-world mobility data. Using data from cellular networks and geo-tagged social media content a large volume of research has attempted to build models that describe the statistical properties of urban human mobility (e.g., (Noulas et al. 2012; Isaacman et al. 2012; Song et al. 2010b; 2010a) - with the list of course being non-exhaustive).

In terms of mobility models there are two big classes. The first one is inspired by Newton’s law of gravity and supports that mobility is impeded by distance. Movements over long distances cost more than moves over short distances. In particular, the flow of people from a given starting location ℓ_s to a destination location ℓ_j decreases with the distance between these two locations (Carrothers 1956; Wilson 1967; Erlander and Stewart 1990; Krings et al. 2009). The second class of models is based on Stouffer’s law of intervening opportunities (Stouffer 1940). As Stouffer posits it “The number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities”. Simply put, displacements are driven by the spatial distribution of places of interest. While existing literature seems to favor Stouffer’s theory (Miller 1972; Haynes, Poston, and Schnirring 1973), both models are extensively used.

Urban Activity Literature: In a different line of research, data from a variety of sources (e.g., location-based social networks, call detailed records from cellular networks, GPS traces etc.) have been used to quantify and model the activities that people engage in the urban space (e.g., (Noulas, Mascolo, and Frias-Martinez 2013; Yuan, Zheng, and Xie 2012; Reades, Calabrese, and Ratti 2009; Becker et al. April 2011; Girardin et al. October December 2008; Froehlich, Neumann, and Oliver 2009; Zhang and Pelechris 2014; Noulas et al. 2011; Cranshaw et al. 2012; Jiang, Jr., and Gonzalez 2012)). The common motivation behind these studies lays on the fact that understanding the spatial and temporal properties of urban activities can facilitate data-driven urban planning operations such as urban re-development and resource allocation.

Origin-Destination (OD) Flow Estimation: In transportation and operations research, there have been studies on OD flow estimation and prediction (Hazelton 2001;

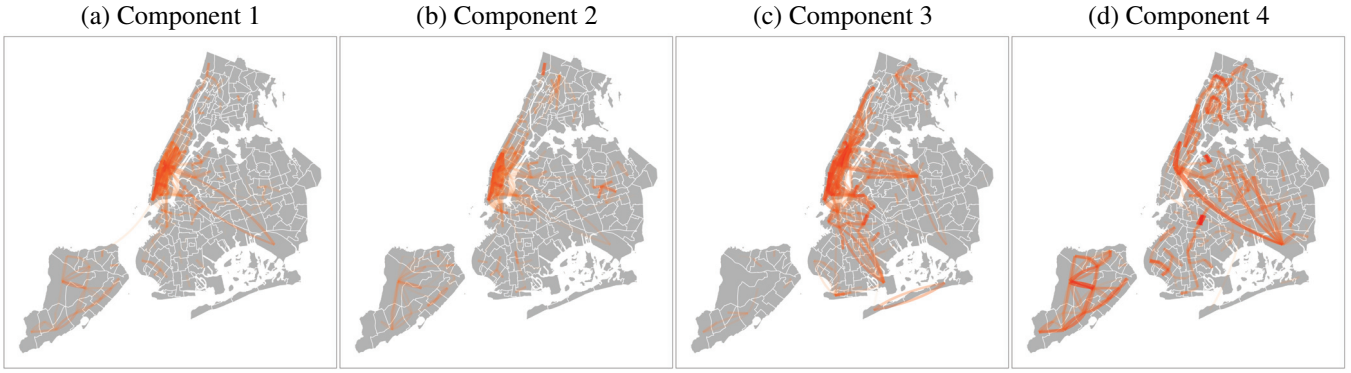


Figure 2: Urban transition flow recovered by EigenTransitions. The first component captures the main structure of urban mobility, indicating that most incoming and outgoing flow concentrates in Manhattan. On the contrary, the rest of the components represent less popular, but still important, patterns. In component (b), frequent transitions happen between neighborhoods in Bronx and Queens; component (c) captures transitions to/from Central Park, while component (d) represents a sub-structure that captures the mobility between major transportation hubs (i.e., Penn Station and JFK Airport).

Ashok and Ben-Akiva 2000; Li et al. 2015), mainly using traffic data from vehicle and bicycle commuting. The work by (Djukic, van Lint, and Hoogendoorn 2012) is closer to EigenTransitions. In particular, they use PCA that dramatically reduces the computational cost of the OD matrix prediction. However, the OD flow estimation can be considered as a special application/case of EigenTransitions, since it focuses on mobility prediction at the aggregate level.

Contrary to the existing literature our study aims at developing a generic framework that can analyze the mobility patterns at a reduced dimensionality space. EigenTransitions can form the core of a number of applications beyond the aggregate mobility flow prediction that is the focus of existing literature.

Dataset and Experimental Setup

Data Collection: We collected geo-tagged Tweets generated within the area covering New York City and Pittsburgh from Jul 15, 2013 to Nov 09, 2014. Each tweet has a tuple format $\langle \text{user Id, place Id, time, latitude, longitude} \rangle$. In total, we have 27,664,594 geo-tagged tweets from 274,933 users in NYC, and 1,988,569 geo-tagged tweets from 19,763 users in Pittsburgh. In our analysis, we consider the municipal neighborhoods as the basic spatial granularity. For our study we also need the population in each urban area (e.g., neighborhood). For this we obtain the neighborhoods boundaries and Census Demographics data at the neighborhood level from NYCOpenData and from Pittsburgh’s Department of City Planning. In summary, there are 195 and 91 municipal neighborhoods in NYC and Pittsburgh, separately.

Urban Region Flow Network: In the urban region flow network¹ $\mathcal{G}_U = (\mathcal{U}, \mathcal{E})$, the set of nodes \mathcal{U} is a collection of non-overlapping areas/neighborhoods in the city under examination. Furthermore, a directed edge $e_{ij} \in \mathcal{E}$ between

¹For simplicity, we will refer to this network as flow network for the rest of the paper.

two areas $u_i, u_j \in \mathcal{U}$ exists if there has been observed a transition by a city-dweller from u_i to u_j . The definition of the urban region can be arbitrary (e.g., municipal neighborhood borders, grids etc.). In our analysis, we divide the whole city using municipal neighborhood borders. We can also annotate every edge e_{ij} with a weight $w(e_{ij})$, which captures the number of such transitions between the two urban regions i and j . However, we will need to calibrate the absolute number of transitions to account for the population in every neighborhood, since the population size of two urban areas indicates a baseline degree of interaction between them. In particular,

$$w(e_{i,j}) = \frac{\tau_{i,j}}{\sqrt{\kappa_i} \cdot \sqrt{\kappa_j}} \quad (1)$$

where $\tau_{i,j}$ is the absolute number of transitions from a location in u_i to one in u_j and κ_i is the population in u_i .

In order to obtain the structure of \mathcal{G}_U for NYC and Pittsburgh we use the geo-tagged Tweets. In particular, we generate an edge $e_{ij} \in \mathcal{E}$ if the same Twitter user has generated two consecutive tweets in locations $\ell_i \in u_i$ and $\ell_j \in u_j$ within a predefined time interval Δ_t and the distance between these two locations is greater than a threshold Δ_d . The edges of \mathcal{G}_U describe the dynamic interaction between urban neighborhoods as captured by the underlying human mobility. In our experiments, we set $\Delta_t = 4$ hours, and given the typical accuracy of GPS technology in urban areas we set $\Delta_d = 100m$. This value for Δ_d also ensures that potential movements within the same building are not considered as transitions. Finally, we have **3,791,072** such transitions in NYC and **260,284** in Pittsburgh. Note here that, the above definition allows for self-edges in \mathcal{G}_U .

The calculation of the weights for the edges in \mathcal{G}_U , requires the estimation of the home neighborhood of a Twitter user. We define the home neighborhood of a user as the one that is most frequently visited by the user. Consequently we estimate the (Twitter) population of neighborhood u_i by counting the number of Twitter users with home location u_i . Using the estimated population allows us to compute

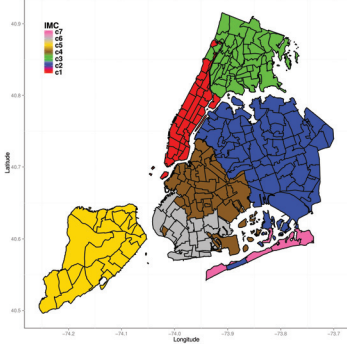


Figure 3: Communities detected using the dynamic urban human flow are spatially concentrated and similar to the five areas defined by the municipality of NYC.

$\kappa_i, \forall u_i \in \mathcal{U}$. Note here that, one could have used the information from the Census Demographics, but this would only be appropriate if Twitter users were a uniformly sampled subset of the actual population, which is not necessarily true (Mislove et al. 2011).

Neighborhood Communities: As alluded to above the flow network can be defined using different spatial divisions of the city. For example, one can aggregate neighborhoods to communities, based on \mathcal{G}_U and then define a higher level network, where the nodes represent a set of neighborhoods belonging to the same community, while the edges represent transitions between communities (as compared to neighborhoods). In fact, our framework, *EigenTransitions*, is generic and can analyze flow networks at different spatial levels as we show later.

Given the urban neighborhood flow network, we apply a community detection algorithm, namely Infomap (Rosvall and Bergstrom 2008) to cluster the neighborhoods into different communities. To reiterate, the community represents a higher-level unit as compared to the predefined neighborhoods. Figure 3 presents the community structure of NYC captured by \mathcal{G}_U , where different colors indicate different communities. The interesting thing to note is that the identified communities are spatially concentrated, while they are similar to the five well known areas of NYC, namely, Manhattan, Brooklyn, Queens, Bronx and Staten Island.

In our work we analyze both the neighborhood-based urban flow network \mathcal{G}_n , as well as, the community-based flow network \mathcal{G}_c .

EigenTransitions

In this section we will formally present *EigenTransitions*. Given a set of N non-overlapping urban areas and a set of transitions between these areas, we define the $N \times N$ matrix T as the adjacency matrix of the corresponding flow network \mathcal{G}_U . The adjacency matrix T can be constructed in a variety of ways. For instance, we can use the transitions of a single user over the whole period that our dataset covers. Alternatively, we can use the transitions from all the users but only during a specific time period. In general, matrix T allows for different aggregation

levels/entities. For a given entity i , we vectorize T and get a transition profile for this entity, which is essentially a $N^2 \times 1$ vector. Considering m instances of this entity (e.g., m users if each matrix T corresponds to a user) we define the transition matrix X as the matrix where each row represents a separate transition profile. Simply put, X is an $m \times N^2$ matrix.

Our goal with *EigenTransitions* is to develop a generic framework that *analyzes* and *summarizes* the urban mobility in a smaller dimensionality, which will allow for spotting persistent patterns in the data by filtering out the *noise*. As we will show in detail later, *EigenTransitions* is indeed able to spot differences in the mobility of different parts of the population that are not “visible” in the original, higher dimensionality. Towards this objective we apply Principle Component Analysis (PCA) on matrix X to get the spectrum of its covariance matrix. We consequently use the eigenvectors and eigenvalues obtained to define the *EigenTransitions*.

With $X \in \Pi^{m \times n}$, where m is again the number of instances (i.e., users) and $n = N^2$ is the number of features (i.e., the original dimensionality of the transition profile), we first calculate the covariance matrix S . In particular,

$$S = \frac{1}{m-1} X^T X \quad (2)$$

Then, we calculate the eigenvectors and eigenvalues of matrix S . This process is computational expensive especially when n is large. However, there is an interesting connection between Singular Value Decomposition (SVD) and PCA. In particular, let the SVD of matrix X be:

$$X = U \Sigma V^T \quad (3)$$

Then the eigenvalue decomposition for S is

$$S = \frac{1}{m-1} V \Sigma^T U^T U \Sigma V^T = V \Lambda V^T \quad (4)$$

where Λ is a diagonal matrix containing the eigenvalues λ_i of S in descending order. In particular, $\Lambda = \frac{1}{m-1} \Sigma^T \Sigma$, since $U^T U = I$. Based on Equation (4) the eigenvectors of S are the right singular vectors of X , and the eigenvalues of S are the square of the singular values of X divided by $m-1$. Matrix V includes the *EigenTransitions*, which are essentially the proto-mobility patterns present in the original dataset. These *EigenTransitions* correspond to a latent mobility space. More specifically, the columns of matrix V correspond to the basis of this latent space, while the rows correspond to the original columns of matrix X , namely, the features. Simply put, matrix V encodes the latent, proto-mobility, patterns of the population as a linear transformation of the original space (that of the full transitions) to the latent space of *EigenTransitions*. U is the coefficient matrix, where each row corresponds to an instance of the original matrix X (e.g., a user) and each column captures the *coordinates* of this instance in the latent *EigenTransitions* space. For example, element $U_{i,j}$ is the coordinate of instance i in the *EigenTransitions* dimension j with base vector the j -th right eigenvector of X . In other words, matrix U captures how much an instance

contributes to each latent pattern depending on the value of coefficients.

However, not all of the columns are necessary to reconstruct the original dataset. In fact, a small number of proto-mobility patterns might be enough to explain a pre-defined level of the variance in the dataset (i.e., reconstruct the covariance matrix S). A principled way of choosing the number of proto-patterns involves the calculation of the ratio:

$$\phi_k = \frac{\sum_{i=1}^k \lambda_i^2}{\sum_{i=1}^n \lambda_i^2} \quad (5)$$

This ratio represents the percentage of variance in the dataset that can be explained by the k first principal components. A typical value for the variance explained is 95% and hence, the minimum value of k that provides a ratio $\phi_k > 0.95$ is the number of `EigenTransitions` we consider. Given the number of `EigenTransitions`, the transition profile of the various instances in the latent space can be represented by the reduced matrix U^r , that is, the first k columns of matrix U .

Scalability consideration: For a large matrix (i.e., an extremely high-dimensional space), we are interested in keeping only those principal components whose eigenvalues are greater than 1, as components with eigenvalues greater than 1 explain at least the same amount of variance as a single transition dimension. In practice, we can compute a partial eigenvalue decomposition using the augmented implicitly restarted Lanczos bidiagonalization (irlba) algorithm (Baglama and Reichel 2005), which allows for fast and scalable eigenvalue decomposition using a few approximate singular values and the corresponding singular vectors. This method can also work on a large sparse matrix, which is typically the case for a transition matrix X that corresponds to a finer spatial granularity.

Temporal & Demographic Mobility Dynamics

In this section we will explore the benefits of `EigenTransitions` in identifying differences in sub-groups of the total population that cannot be observed in the original, high dimensionality. For example, based on the gender of users, we are interested in examining the difference between the mobility patterns of male and female population. In particular, are these two groups different with regards to their mobility patterns? To answer this question we build two transition matrices X_{male} and X_{female} . Row i of X_{male} (X_{female}) corresponds to the original transition profile of male (female) user i .

Previous work (Bagrow and Lin 2012) using phone call records has indicated some connections between individual mobilities and demographics. With the application of `EigenTransitions` in our study, demographic information (e.g., ethnicity, gender, age etc.) is not the only way to define and compare sub-populations. Temporal dynamics of urban mobility can also be examined. For example, we can study the different travel behaviors during the weekdays and weekends by defining and using daily-wise transition matrix X_{day} that captures the transitions of users (rows) on specific days. Further population segmentation can be achieved

by combining demographic and temporal dimensions, e.g., we can compare the mobility behavior during weekdays and weekends for the male population by building a daily-wise transition matrix $X_{day,male}$ using transitions only from the male population.

In principle, `EigenTransitions` are not necessary for comparing the mobility of two populations. Focusing, for presentation reason, on comparing the male and female mobility patterns, one could simply perform a statistical hypothesis test between the two populations using as features the full transition profiles. In other words, one could perform a Hotelling’s T^2 test (Hotelling 1931) on the datasets described by matrices X_{male} and X_{female} . The Hotelling’s T^2 test is the generalization of the t-test for the case of multidimensional variables. In a nutshell, with \bar{X}_{male} and \bar{X}_{female} being the multivariate means for the male and female sub-populations respectively, Hotelling’s T^2 examines the following hypothesis test:

$$H_0 : \bar{X}_{male} = \bar{X}_{female} \quad (6)$$

$$H_1 : \bar{X}_{male} \neq \bar{X}_{female} \quad (7)$$

However, an alternative way to compare the two populations is applying again the Hotelling’s T^2 test but instead of using the original feature space, we can use the `EigenTransitions`, which have much lower dimensionality but at the same time can capture the majority of the variance in matrices X_{male} and X_{female} . To reiterate each row in the coefficient matrix U captures the “coordinate” of an instance in the latent space. We can then compare the coefficients matrix U of each sub-population, i.e., U_{male} and U_{female} . The hypothesis test now becomes comparing the multivariate means \bar{U}_{male} and \bar{U}_{female} :

$$H_0 : \bar{U}_{male} = \bar{U}_{female} \quad (8)$$

$$H_1 : \bar{U}_{male} \neq \bar{U}_{female} \quad (9)$$

Note here that, in practice we are using the reduced coefficient matrix U^r given the top- k principle components extracted.

The premise is that the noise present in the high dimensionality of the original transition space can affect the performance of the test. For example, when the size of the datasets is *small* compared to the dimensionality of the features, the statistical power of the test can be reduced and therefore, it might be unable to identify (small) differences between the populations compared at a pre-defined significance level. In fact, when the dimensionality is strictly larger than the total size of the two populations the Hotelling’s T^2 test cannot be applied at all! Using a space of reduced dimensionality can overcome this problem and hence, `EigenTransitions` are crucial in similar settings. Furthermore, depending on the dataset, a high dimensionality can potentially lead to the null hypothesis being rejected due to differences in a small number of “secondary” elements of the feature vector. The reduced dimensionality space that `EigenTransitions` offer can again alleviate this problem since they capture the most important mobility patterns in the dataset.

Table 1: p -value of the Hotelling T^2 test comparing the mobility patterns between two populations from different temporal and demographic perspectives.

Spatial granularity		Neighborhood		Community	
		NYC	PITT	NYC	PITT
Gender	<i>female/male</i>	0.0	0	0	0.033
	<i>White/Black</i>	0.0	0.542	0.012	0.988
Ethnicity	<i>White/Asian</i>	0.001	0.005	0.388	0.014
	<i>White/Hispanic</i>	0.0	0.039	0.0	0.496
	<i>Black/Asian</i>	0.0	NA	0.0	0.692
	<i>Black/Hispanic</i>	0.098	NA	0.002	0.998
	<i>Asian/Hispanic</i>	0.0	0.815	0.0	0.849
Temporal	<i>Weekday/Weekend</i>	0	0	0.0132	2e-04
	<i>Daytime/Night</i>	0	0	2e-192	0

In what follows we use `EigenTransitions` to compare different demographic parts of the population. In particular, we infer the ethnicity and gender of each user in our dataset (see Appendix A for details) and compare their mobility patterns. We use the `EigenTransitions` identified by both the neighborhood-based urban flow network \mathcal{G}_n as well as the community-based urban flow network \mathcal{G}_c , which essentially give us the transition profiles at two different spatial granularities. Apart from the demographics comparisons we also compare the temporal patterns of the urban mobility captured from our data.

Demographic and temporal dynamics comparisons:

Table 1 presents the results of the Hotelling’s T^2 test for different divisions of the populations for NYC and Pittsburgh. As we can see for NYC, in (almost) all of the cases the null hypothesis is rejected (at the significance level of 0.01), i.e., there is strong evidence against the hypothesis that the two populations exhibit the same `EigenTransitions` on average. For Pittsburgh, in some cases the test fails to reject the null hypothesis. This most probably can be attributed to low statistical power of the test since the size of the various subgroups is fairly small (e.g., we were able to only identify 7 “Black” users). In a few cases we were not even able to perform the Hotelling test at all since the dimensionality of the feature space was greater than the total sample set size. As we can see these cases appear when we consider the neighborhood-based flow network where the number of nodes (and hence, the dimensionality of X) is much larger.

In order to ensure that the rejection of the null hypothesis is not an artifact of the large size of our dataset, leading to the rejection of H_0 due to irrelevant differences between the two populations, we perform a “within” population test. In particular, we randomly split each sub-population into two parts and perform the Hotelling’s T^2 test on these random splits. One would expect that since both parts come from the same population, the Hotelling’s T^2 test will fail to reject the null hypothesis. Indeed this is the case for all the demographic and temporal sub-populations as we can see in the results presented in Table 2. In particular, for every case we perform 100 random splits and present the median p -value.

We further examine the temporal mobility dynamics for each population. Tables 3 and 4 present the results comparing the patterns during weekdays and weekends, as well as

Table 2: Hotelling T^2 tests within a population by randomly splitting the population into two groups. Each p -value reported is the median of 100 different random splits.

Spatial granularity		Neighborhood		Community	
		NYC	PITT	NYC	PITT
Gender	<i>female</i>	0.55	0.65	0.55	0.59
	<i>male</i>	0.545	0.45	0.545	0.48
Ethnicity	<i>White</i>	0.44	0.58	0.558	0.62
	<i>Black</i>	0.465	NA	0.575	0.485
	<i>Asian</i>	0.49	0.425	0.413	0.475
	<i>Hispanic</i>	0.46	0.625	0.518	0.47
	<i>Weekdays</i>	0.445	0.515	0.449	0.513
Temporal	<i>Weekends</i>	0.515	0.455	0.480	0.428
	<i>Daytime</i>	0.44	0.57	0.453	0.48
	<i>Nighttime</i>	0.658	0.545	0.579	0.503

Table 3: Each population sub-group presents significantly different mobility patterns during weekdays and weekends.

Spatial granularity		Neighborhood		Community	
		NYC	PITT	NYC	PITT
Gender	<i>female</i>	0	0	4.166e-05	0
	<i>male</i>	0	0	0.006	0.005
Ethnicity	<i>White</i>	0	0	0.003	0
	<i>Black</i>	0.001	0.003	0.4	0.004
	<i>Asian</i>	0	0	3.441e-06	0.034
	<i>Hispanic</i>	0	0.319	0.0131	0.055

during daytime (4am-6pm) and nighttime. As we can see, in (almost) all of the cases there is a strong temporal component, i.e., the groups change their behavior over time.

Discriminative patterns: Hotelling’s T^2 test provides us with a sense of whether two populations are heterogeneous across the whole latent mobility space. The more interesting question is which `EigenTransitions` really differentiate them. To answer this question we compare the two sub-groups under consideration with regards to the individual `EigenTransitions` using bootstrap hypothesis test (Efron and Tibishirani 1993). We choose to rely on bootstrap for the hypothesis testing rather than on the t -test to avoid any assumption for the distribution of the data. In particular, the i -th column of the reduced coefficient matrix U^r contains the coefficient for the i -th `EigenTransitions` for the entities described by its rows. Hence, for example, by considering the i -th column of the U_{male}^r and U_{female}^r reduced coefficient matrices we can identify whether the i -th `EigenTransitions` discriminates the two populations.

Table 5 presents the results for the top 3 components for NYC (the result for Pittsburgh are omitted due to space limitations and since they exhibit the same behavior). The most interesting observation is that most of the tests for the first `EigenTransitions` fail to reject the null hypothesis. This indicates that the two populations are similar with regards to the strongest mobility pattern. Intuitively the first component from PCA always captures a large fraction of the variance and resembles the main artery of the urban mobility, that everybody (every time) tends to follow. What really differentiate the two populations are usually the secondary

Table 4: The mobility behavior of a population sub-group differs between daytime and nighttime.

Spatial granularity		Neighborhood		Community	
		NYC	PITT	NYC	PITT
Gender	<i>female</i>	0	0	0.001	0
	<i>male</i>	0	0	0.021	0
Ethnicity	<i>White</i>	0	0	0.013	0
	<i>Black</i>	0	0.004	0.361	0.002
	<i>Asian</i>	0	0	0	0.002
	<i>Hispanic</i>	0	0.097	0	0.032

patterns which capture the different interests of the individuals. These discriminative *EigenTransitions* play an important role in understanding and targeting a specific population of interest.

Neighbor facilitated mobility prediction

In this section, we examine how *EigenTransitions* can facilitate the mobility prediction problem, focusing especially in the so-called “cold-start” scenarios where traditional methods have been shown to be ineffective. These cases correspond to the situations where a user visits an area for the first time and hence, any methods that are purely based on individuals historical trails will fail. Existing methods that utilize the gravity (Erlander and Stewart 1990) and/or the intervening opportunity model (Stouffer 1940) mainly take advantage of the aggregate level travel demand, but they do not consider the interest of individuals. Recent work consider the historical travel behavior of individual users to predict their movement in the future, since most individuals are highly predictable given enough historical trails (Song et al. 2010b). Social features have also been proven to help improve the location prediction of individuals (Cho, Myers, and Leskovec 2011), since mobility patterns are homophilous (Zhang and Pelechris 2014) and users’ movement can be influenced by their social connections (Wang et al. 2011).

In this work, we consider a different setting. In particular, we are interested in predicting the next destination area (neighborhood or community depending on the scenario) when the origin is an urban area that the user has not visited before. This is a cold-start problem in the sense that we do not have any historical travel information for the user so as to build the transition probability distribution to destinations. One way to solve this problem is to simply take advantage of the transition behavior of the whole population. This forms an intuitive baseline, since if the majority of the population is following specific transition patterns, then there should also be a high probability that the user under consideration will follow the same patterns. However, this method is limited since it utilizes the same transition distribution regardless of individuals’ interests and mobility structures.

Instead of using the overall population, we propose to leverage the mobility behaviors of the top-K nearest neighbors (**KNN**) to facilitate the location prediction for “cold” users. This is similar to the idea of collaborative filtering in recommender systems, e.g., users with similar interests in

some types of products may also have similar interests in other types of products. In the setting of location prediction, users with similar mobility behavior across some urban areas will be more likely to have similar behaviors across other urban areas. In particular, given that a target user currently moves to urban area u_i for the first time, we first identify the top-K nearest neighbors with similar historic transition profile. Leveraging neighbors’ historical transitions starting from area u_i to other areas, we then build the transition distribution for the target user, with the underlying assumption that the target user tends to have similar transition behaviors originating from u_i as the identified neighbors. The probability that the user will travel to destination area u_j consequently depends on this transition distribution.

In order to find neighbors with similar mobility behavior, we calculate the distance between users’ original historic transition profiles and then select the top-K nearest neighbors. However, the noise present in the original transition profiles may distort the distance calculation, leading to a non-robust set of nearest neighbors. In this work, we propose a *EigenTransitions*-based k-nearest neighbors (**eKNN**) algorithm. By using the *EigenTransitions*, we identify nearest neighbors with similar mobility patterns in the latent *EigenTransitions* space. In particular, we select the top-K nearest neighbors by calculating the Euclidean distance between the users’ coefficients from the reduced coefficients matrix U^r .

To evaluate our algorithm, we utilize our NYC and Pittsburgh datasets and focus on the mobility prediction task at the community level. We first split the 16-month data into two parts: the first 11-months are used for training and the rest for testing. We keep users who have at least 10 transitions in the whole dataset and at least 1 transition in the training set. The latter ensures that we have some historic information about the user and hence, we can obtain a basic view for the mobility interests of the user. This is necessary for locating neighbors with similar mobility profiles. In the training stage, we build the transition matrix X using the corresponding transitions. Consequently we can find the nearest neighbors by either using the original transition profile or the *EigenTransitions*. In the testing stage, we only consider the “cold”-start scenarios.

We compare our **eKNN** algorithm to four baselines: (1) random guess, that is the user selects the destination uniformly at random; (2) population-based, the destination is selected based on the transition distribution from the whole population. This is essentially a special case of **eKNN**, where we set the number of nearest neighbors to the size of whole population. We refer this method as population-based nearest neighbors (**pNN**); (3) K neighbors are uniformly at random selected (**rKNN**); (4) neighbors are selected based on the distance between the original transition profiles (**oKNN**).

Figure 4 presents the prediction accuracy for our experiments. As we can see even with a small number of nearest neighbors **eKNN** is outperformed by the population-based algorithm. However, as the number of nearest neighbors considered increases and exceeds a certain threshold, i.e., the percentage of the nearest neighbors used exceeds 1.5%

Table 5: Our individual bootstrap hypothesis tests for the top-3 EigenTransitions indicate that the secondary latent mobility patterns are important for differentiating between sub-groups of the population. k indicates the number of EigenTransitions for the specific sub-population.

Spatial granularity		Neighborhood				Community			
		k	E1	E2	E3	k	E1	E2	E3
Gender	<i>female/male</i>	161	0.314	0.315	0.531	6	0.324	0.715	0.482
Ethnicity	<i>White/Black</i>	130	0.404	0.489	0.133	6	0.983	0.388	7.119e-07
	<i>White/Asian</i>		0.035	0.090	0.383		0.579	0.620	0.022
	<i>White/Hispanic</i>		0.129	0.061	0.115		0.017	3.112e-05	4.167e-10
	<i>Black/Asian</i>		0.411	0.179	0.161		0.894	0.316	0.400
	<i>Black/Hispanic</i>		0.773	0.106	0.930		0.600	0.578	0.481
	<i>Asian/Hispanic</i>		0.002	0.039	0.047		0.033	0.0003	0.612
Temporal	<i>Weekday/Weekend</i>	1	0.01322	NA	NA	4	0.219	0.002	1.326e-23
	<i>Daytime/Night</i>	1	2.174e-192	NA	NA	4	0.253	1.133e-07	2.416e-36

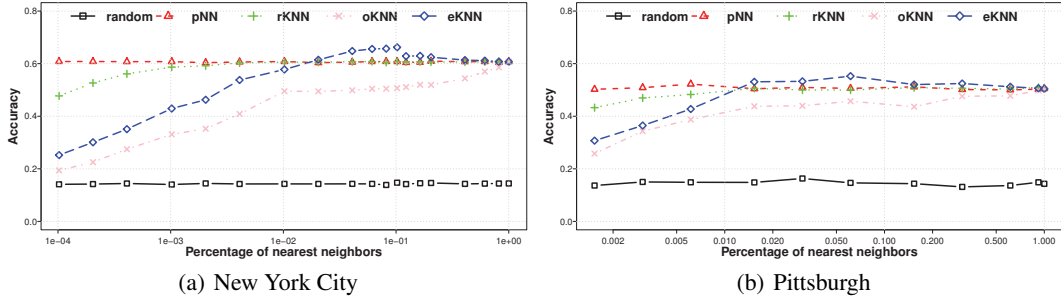


Figure 4: Performance of location prediction comparing different algorithms. The x-axis represents the percentage of the nearest neighbors used from the whole population. Our results indicate that the proposed EigenTransitions-based k-nearest neighbors algorithm outperforms all the baselines considered, given the percentage exceeds a small threshold (i.e., 1.5%).

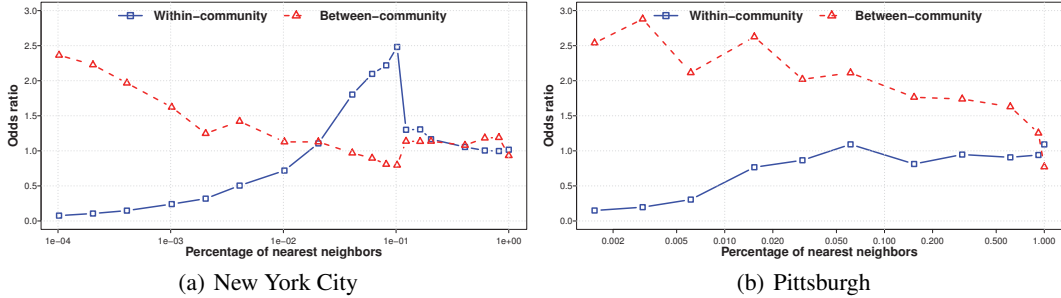


Figure 5: Odds ratio between the prediction performance of **eKNN** and **pNN**. **eKNN** exhibits a diverging trending as the percentage of nearest neighbors increases, with respect to predicting transitions within communities and between communities.

of the whole population, **eKNN** outperforms **pNN**. The accuracy reaches its peak when the percentage of the nearest neighbors used is about 5% – 10% of the whole population. Further increasing the number of neighbors considers, does not significantly improve the performance of **eKNN** over **pNN** and finally converges to that of **pNN** as one might have expected. These results confirm the intuition that a subset of neighbors with similar mobility interests can facilitate the location prediction in cold-start scenarios. Furthermore, **eKNN** is always better than **oKNN** and **oKNN** is always outperformed by the population-based method. This indicates that similar neighbors identified in

the EigenTransitions latent space are more robust and effective than that in the original high-dimensionality noisy space. **rKNN** outperforms **eKNN** when a small percentage of neighbors is used. We speculate that this is due to the fact that when a small number of neighbors is used, it is very likely that they do not include “cold” transitions of the user, since his neighbors will be similar to him. On the other hand **rKNN** will include at least the most popular “cold” transitions with high probability. However, as the number of neighbors consider by **eKNN** increases, our scheme is able to diversify more and hence, outperform **rKNN**.

Figure 5 further presents the odds ratio between the pre-

diction performance of **eKNN** and **pNN**. The odds ratio is calculated as:

$$OR(eKNN, pNN) = \frac{p_{eKNN}/(1 - p_{eKNN})}{p_{pNN}/(1 - p_{pNN})} \quad (10)$$

where p is the prediction accuracy of the algorithm. **eKNN** outperforms **pNN** if the odds ratio is greater than one.

In particular, we are interested in examining what is the prediction performance when considering different transition types, e.g. self-transitions (i.e., transitions within community) versus transitions between communities. As we can see **eKNN** exhibits a diverging trending as the percentage of nearest neighbors considered increases. This implies that nearest neighbors do not help with predicting individual mobility within communities. However, it is necessary to have a certain number of similar neighbors to enhance transition prediction between communities. This might be due to the fact that transitions within communities are much more popular than transitions between communities, thus, naturally more predictable. So neighbors' transition profiles are not that helpful in this situation.

Discussion

EigenTransitions is a generic analytical framework and hence, potential population biases associated with the Twitter dataset used do not affect the core of our study. EigenTransitions can form the building block in a variety of applications, and of course in this case the data that drive the application are crucial. While in this study we have focused on its applicability and benefits when comparing mobility patterns between sub-populations and facilitating the cold-start mobility prediction, there are many different scenarios where EigenTransitions can be helpful.

For instance, cross-city comparisons are crucial for understanding what policies might be transferable between cities. EigenTransitions can facilitate a comparison between cities with respect to the underlying mobility and its predictability. For example, given a city c , the matrix X_c captures the transition profiles of all its dwellers. The number of EigenTransitions that explains 95% of the variance of the underlying data can provide us with an estimate of the predictability of the aggregate urban mobility. For example, a city that includes a small number of EigenTransitions can be deemed fairly more “predictable” in terms of transportation needs as compared to one that requires a large number of EigenTransitions. Of course, transportation and mobility patterns are mutually dependent but the point is that the developed framework can be used to perform cross-city comparisons as well.

For example, Table 6 presents the number of components needed to explain 95% of the variance for NYC and Pittsburgh and for different transition matrices X . Focusing on the matrix where the rows correspond to specific users, we see that for the dataset from the city of Pittsburgh less EigenTransitions are required to explain 95% of the variance, translating to more “stable” patterns. While our experiments here are very small-scale they are clearly illustrating the potential of EigenTransitions to be used as a

cross-city comparison metric. In the future, we opt to further explore this direction.

Table 6: The number of EigenTransitions needed to capture 95% of the variance for the mobility dataset of different cities and for different matrices X .

Entity	Neighborhood		Community	
	NYC	PITT	NYC	PITT
Users	204	80	6	4
Days	4	7	1	2

Conclusions

In this work, we introduce EigenTransitions, a generic framework to analyze and summarize mobility datasets. We demonstrate that EigenTransitions can be applied in a variety of settings. In particular, we utilize EigenTransitions to compare the temporal and demographic dynamics of the observed mobility. EigenTransitions are able to identify differences that are not observable in the original transition space. Furthermore, we develop an **eKNN**-based mobility prediction method, which as we show outperforms various baselines in “cold-start” prediction scenarios.

In the future, we opt to incorporate into our analysis mobility traces from more cities and explore how EigenTransitions can be used to examine and compare mobility patterns across difference cities. Finally we plan to extend our methodology of matrix factorization to the high-dimension tensor decomposition, that enables to capture multiple facets of urban mobility simultaneously.

Acknowledgments

This work was partially supported by the ARO Young Investigator Award W911NF-15-1-0599 (67192-NS-YIP) and CRDF from the University of Pittsburgh.

References

- Ashok, K., and Ben-Akiva, M. E. 2000. Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. *Transportation Science* 34(1):21–36.
- Baglama, J., and Reichel, L. 2005. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing* 27(1):19–42.
- Bagrow, J. P., and Lin, Y.-R. 2012. Mesoscopic structure and social aspects of human mobility. *PloS one* 7(5):e37676.
- Becker, R.; Caceres, R.; Hanson, K.; Loh, J.; Urbanek, S.; Varshavsky, A.; and Volinsky, C. April 2011. A tale of one city: Using cellular network data for urban planning. In *IEEE Pervasive Computing*, vol. 10, no. 4.
- Carrothers, V. 1956. A historical review of the gravity and potential concepts of human interaction. *Journal of the American Institute of Planners* 22:94–102.
- Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090. ACM.

Cranshaw, J.; Schwartz, R.; Hong, J.; and Sadeh, N. 2012. The livelihoods project: Utilizing social media to understand the dynamics of a city. In *AAAI ICWSM*.

Djukic, T.; van Lint, J.; and Hoogendoorn, S. 2012. Application of principal component analysis to predict dynamic origin-destination matrices. *Transportation Research Record: Journal of the Transportation Research Board* (2283):81–89.

Efron, B., and Tibishirani, R. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Erlander, S., and Stewart, N. 1990. *The Gravity Model in Transportation Analysis: Theory and Extensions*. CRC Press - Topics in Transportation.

Froehlich, J.; Neumann, J.; and Oliver, N. 2009. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*.

Girardin, G.; Calabrese, F.; Fiore, F.; Ratti, C.; and Blat, J. October - December 2008. Digital footprinting: Uncovering tourists with user-generated content. In *IEEE Pervasive Computing*, vol. 7, no. 4.

Haynes, K.; Poston, D.; and Schnirring, P. 1973. Inter-metropolitan migration in high and low opportunity areas: indirect tests of the distance and intervening opportunities hypotheses. *Economic Geography* 49(1):68–73.

Hazelton, M. L. 2001. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological* 35(7):667–676.

Hotelling, H. 1931. The generalization of student’s ratio. In *Annals of Mathematical Statistics* 2(3), 360–378.

Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Rowland, J.; and Varshavsky, A. 2010. A tale of two cities. In *ACM HotMobile*.

Isaacman, S.; Becker, R.; Cáceres, R.; Martonosi, M.; Rowland, J.; Varshavsky, A.; and Willinger, W. 2012. Human mobility modeling at metropolitan scales. In *ACM MobiSys*.

Jiang, S.; Jr., J. F.; and Gonzalez, M. 2012. Discovering urban spatial-temporal structure from human activity patterns. In *ACM UrbComp*.

Krings, G.; Calabrese, F.; Ratti, C.; and Blondel, V. 2009. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* L07003.

Li, Y.; Zheng, Y.; Zhang, H.; and Chen, L. 2015. Traffic prediction in a bike-sharing system.

Miller, E. 1972. A note on the role of distance in migration: costs of mobility versus intervening opportunities. *Journal of Regional Science* 12(3):475–478.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of twitter users. *ICWSM* 11:5th.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *SMW*.

Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; and Mascolo, C. 2012. A tale of many cities: universal patterns in human urban mobility. In *PLoS ONE* 7(5): e37027. doi:10.1371/journal.pone.0037027.

Noulas, A.; Mascolo, C.; and Frias-Martinez, E. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *International Conference on Mobile Data Management*.

Ravenstein, E. 1885. The laws of migration. *Journal of the Statistical Society of London* 48(2):167–235.

Reades, J.; Calabrese, F.; and Ratti, C. 2009. Eigenplaces: analysing cities using the spacetime structure of the mobile phone

network. *Environment and Planning B: Planning and Design* 36(5):824–836.

Rosvall, M., and Bergstrom, C. T. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4):1118–1123.

Song, C.; Koren, T.; Wang, P.; and Barabási, A.-L. 2010a. Modelling the scaling properties of human mobility. In *Nature* 6, 818–823.

Song, C.; Qu, Z.; Blumm, N.; and Barabási, A.-L. 2010b. Limits of predictability in human mobility. In *Science* 327, 1018–1021.

Stouffer, S. 1940. Intervening opportunities: A theory relating mobility and distance. *American Sociological Review* 5(6):845–867.

Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; and Barabási, A.-L. 2011. Human mobility, social ties, and link prediction. In *ACM SIGKDD*.

Wilson, A. 1967. A statistical theory of spatial distribution models. *Transportation Research* 1:253–269.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering regions of different functions using human mobility and pois. In *ACM SIGKDD*.

Zhang, K., and Pelechris, K. 2014. Understanding spatial homophily: the case of peer influence and social selection. In *Proceedings of the 23rd international conference on World wide web*, 271–282. ACM.

Appendix A: Demographics Inference

A large volume of methods have been proposed to infer the demographics of Twitter users. In our work, we utilize a simple and reliable method reported in (Mislove et al. 2011), which infers the gender and ethnicity of a user from the self-reported names.

Table 7: Number of users with identified demographics.

City		New York City	Pittsburgh
Gender	<i>female</i>	13,059	982
	<i>male</i>	14,090	1,050
Ethnicity	<i>White</i>	7,340	738
	<i>Black</i>	175	7
	<i>Asian</i>	783	32
	<i>Hispanic</i>	1,953	27

Inferring gender from first names: As per (Mislove et al. 2011), we first download the list of top 1000 male and female baby names for each year between 1900 and 2013, as reported by U.S. Social Security Administration. Then we aggregate the names together and calculate the corresponding frequency for each name, which results in 3,757 female names and 3,114 male names. We only keep the names that are at least 95% predictive, that is, given a first name, the proportion of a gender (male or female) is at least 95% (e.g., we remove the name Taylor for which 26% are males and 74% are females). Finally, to infer the gender, we compare the first word of a user’s name to the compiled list of first names.

Inferring ethnicity from last names: Similar to the gender identification, we identify the ethnicity of Twitter users using their last names. To achieve this, we download the list of last names from U.S. Census 2000, where each last name is associated with a distribution for its ethnicity. Then we keep the last names that are at least 90% predictive and identify the ethnicity by matching the last name of users with the ones in the compiled list. Table 7 presents the number of users with demographics identified for NYC and Pittsburgh.