The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility

Neal Lathia, Daniele Quercia, Jon Crowcroft

The Computer Laboratory, University of Cambridge, UK
neal.lathia, daniele.quercia, jon.crowcroft@cl.cam.ac.uk

Abstract. A key facet of urban design, planning, and monitoring is measuring communities' well-being. Historically, researchers have established a link between well-being and *visibility* of city neighbourhoods and have measured visibility via quantitative studies with willing participants, a process that is invariably manual and cumbersome. However, the influx of the world's population into urban centres now calls for methods that can easily be implemented, scaled, and analysed. We propose that one such method is offered by pervasive technology: we test whether urban mobility—as measured by public transport fare collection sensors—is a viable proxy for the visibility of a city's communities. We validate this hypothesis by examining the correlation between London urban flow of public transport and census-based indices of the well-being of London's census areas. We find that not only are the two correlated, but a number of insights into the flow between areas of varying social standing can be uncovered with readily available transport data. For example, we find that deprived areas tend to preferentially attract people living in other deprived areas, suggesting a segregation effect.

Keywords: Mobility, Urban Analysis, Sensors, Well-Being

1 Introduction

An ever-increasing proportion of this globe's 7 billion-strong population is living in or moving into cities; in the United Kingdom, this figure was projected to have already surpassed the 90% mark[1]. In this setting, the ability to design and monitor urban spaces that enable social and economic well-being becomes critical. In the past, urban planners have asserted that the well-being of communities is related to their *visibility* or *imaginability* [1]. The key idea is that the less imaginable a social setting is, the more unnerving experiences within it will be. Sociologists have thus measured urban visibility by asking study participants to draw mental maps of their city [2], the assumption being that urban residents' recall of their city reflects the extent to which different city parts are visible and form a coherent picture in people's minds. More recently, longitudinal studies have been launched (e.g., Understanding Society[2], The Happiness

[1] Data from the World Resources Institute, http://www.wri.org
[2] http://www.understandingsociety.org.uk

Project[3]) to survey participants about the features of their lives that include strong indicators of community well-being. The ongoing studies are being conducted manually and must therefore take great care with continuous sampling of its participants [3]: the inherent labour involved in conducting such enquiries presents a clear challenge that complicates the measurement (and continuous monitoring) of well-being in the cities of the future.

We posit that data from pervasive technology that tracks city residents' movements across a metropolitan area is a valid proxy for urban visibility. To validate this hypothesis, we examine the relation between two independent datasets from the London, England: (1) a month-long sample of public transport mobility data, measured with passive sensors, and (2) publicly available community well-being census data (measured as community social deprivation). In doing so, we find that urban flow correlates with social deprivation. We also uncover facets of flow between communities (Section 4):

- Socially-deprived communities in London tend to be visited more than well-off communities. Similarly, in the US, researchers have found that "people in cities with the least incomes travel slightly more than people in richer cities" [4].
- In general, homophily does not hold: residents of an area with a given deprivation do not travel to equally-deprived areas. At first sight this suggests that Londoners do not segregate themselves with like-minded people. However, by separating deprived communities from less deprived ones, we observe a different picture: well-off areas tend to attract people living in areas of varying social deprivation; by contrast, deprived areas tend to preferentially attract people from other deprived areas: social segregation holds only for socially-deprived areas, and not for well-off areas.

More generally, these results suggest that large-scale and real-time monitoring of community well-being is cheaply available via the passive sensors that urban residents pro-actively carry and use for public transport access.

2 Related Work

Smart phones and embedded sensor systems have given researchers unprecedented access to new and rich datasets, recording detailed information about how people live and move through urban areas. In this section, we describe a select number of examples that highlight how new datasets are lending insight into individuals' lives and urban analysis. Embedded sensors have recently been used to measure the spatio-temporal patterns of an entire city's usage of a shared-bicycle scheme [5]. Smart-phones' sensors have been used to augment psychological research [6]; Bluetooth sensors have been used to measure social interactions [7]; GPS sensors have been shown to aide urban planning and design [8, 9]. Lastly, this paper uses the same dataset from public transport automated

[3] http://www.somervillema.gov/departments/somerstat/report-on-well–being

fare collection systems which was previously used to investigate travellers' perceptions and incentives [10]. Raw sensor readings, however, tend to lack qualitative descriptions of the context of people who are moving about urban spaces: there is a growing awareness that online resources may offer contextually-rich data that is otherwise absent from sensor readings. Recent research includes the use of "check-ins" (where users input their location to their mobile device) [11] and geo-tagged photos [12] to understand the relation between urban space, social events, and mobility.

These new data sources now allow researchers to quantitatively test past assertions made by urban planners, geographers, and social scientists. In 1960, Kevin Lynch published a book titled "The Image of the City" in which he argued that one of the most important conditions for a liveable and enjoyable city is high "imaginability" [1], or the city dwellers' ability to form a coherent representation of the overall structure of the city. Considerable research then went into quantifying imaginability or, more specifically, the recognizability of a city. Milgram did so for New York City [2]. He found that, as expected, the least deprived (i.e., richest) boroughs happen to be the most recognisable ones. More recently, using a nation-wide communication network obtained from telephone data, Eagle *et al.* showed that less-deprived UK neighbourhoods tend to be associated with residents whose social contacts are geographically diverse [13]. Until recently, however, data has not been available to quantify city recognizability at scale: we will use a London's transport dataset, compute two recognizability measures, and correlate them with UK census' community well-being scores.

3 From Mobility to Community Well-Being

To begin with, we describe the data and the methodology that we applied to examine the relation between urban flow and community well-being. Broadly speaking, by analysing a large sample of trips taken with public transport, we infer the communities that different travellers belong to. From this, we derive a *flow matrix* of visit patterns between different communities (i.e., n residents of location i visit location j). This data can then be used to, first, validate our hypothesis by computing its correlation to *IMD* and, second, to investigate the extent that homophily emerges in large-scale travel patterns.

Mobility and Well-Being Datasets. London is the biggest city in the United Kingdom; by most measures, it is also the largest urban area in the European Union. We obtained well-being data from the UK Office for National Statistics[4], as measured (based on national census results) with the Index of Multiple Deprivation (*IMD*). The *IMD* is a composite score derived from the income, employment, education, health, crime, housing, and the environmental quality of each community [14]. We note that the data is normally distributed. Broadly speaking, socially deprived communities have higher *IMD* scores (e.g., Tottenham, Hackney); whilst less deprived the communities have lower scores

[4] http://data.gov.uk/dataset/index_of_multiple_deprivation_imd_2007

(e.g., Mayfair, Belgravia). In this work, we assume that a census area represents a community; we choose such a definition because it has been widely used in recent studies of social deprivation (including the related article by Eagle *et al.* [13]).

While *IMD* data partitions the city according to spatially bounded communities, the Transport for London (TfL) public transport infrastructure forms a network that binds the city together. The transport system is a vast, multimodal network of underground trains (11 interconnected lines with 270 stations), overground trains (5 lines with 78 stations) and buses (about 8,000 buses serving 19,000 stops) as well as trams, river services, and other specialised services. Moreover, TfL operates an automated fare collection system, which uses RFID-based smart card tickets (called *Oyster cards*); by 2009, this system accounted for approximately 80% of all public transport trips in the city [15]. Detailed information about each trip is captured each time an Oyster card is used to both enter and exit the public transport network; most importantly, it allows for individual travellers' trips to be linked [16].

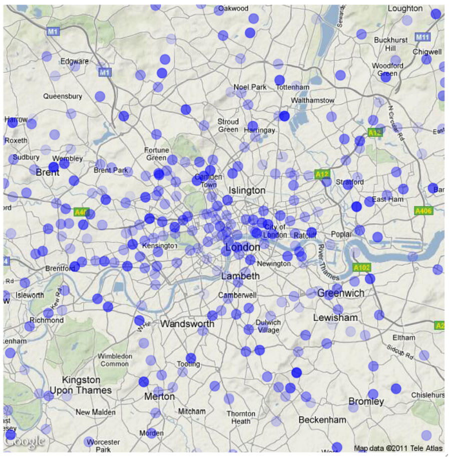

Fig. 1. The geographical distribution of IMD values, mapped using London stations: each circle is a station, darker circles have higher IMD values.

The Oyster card dataset that we obtained contains every single journey taken using smart cards throughout the 31 days of March 2010. This amounts to roughly 89 million journeys, of which 70 million are tube journeys, with the rest made up of trips taken on National Rail, Overground and other rail systems. Each record details the day, anonymised user id, the origin and destination pair, entry time, and exit time (measured only as accurately as the minute of entry/exit). We took two steps to clean the data. First, we removed any entries containing erroneous or inconsistent data, as well as all bus trips (since we do not know the destination for these trips). Entries were removed if the start time was earlier than the end time or if the origin and destination were the same. We are left with 96.4% of the original data, amounting to $76,601,937$ trips by 5.1 million unique users—an average of 2.47 million journeys each day. Lastly, we match stations to census areas by geographical proximity in order to obtain a mapping between stations and *IMD* scores: the resulting geographical layout is shown in Figure 1.

Methodology: We decomposed the process of correlating public transport trips and neighbourhood *IMD* scores into a number of steps:

1. Inferring Travellers' Familiar Locations. This step aims to identify the communities that each traveller is most familiar with. Ideally, we would like to know where each traveller lives; in practice, this data is not available to us. We therefore count the number of entries and exits that travellers have at each station, which allows us to create a ranking of stations for each person. We then pick the top-2 most visited stations by each traveller [17] as their "familiar" locations (which, intuitively, would cover both home and work locations), subject to two conditions. First, the traveller must have had at least 2 trips in the 31 days of our dataset. Second, the inferred locations must also not be major inter-city/international rail stations (e.g., Victoria Station); should both of the top-2 stations fall under this category, the person is dropped from the dataset. Intuitively, this method takes into account typical commuting habits in determining familiar locations [16]; it avoids attributing non-London residents to the communities surrounding intercity train stations, and also prunes people who do not tend to use public transport from the analysis. Note that, for each remaining person, we may have up to two locations that are deemed as familiar locations.

2. Create User-Visit Matrix. Using each trip by traveller u from origin o to destination d, (u, o, d), we produce a binary matrix C which counts the visits (where a visit is broadly defined as a station entry/exit) of travellers to stations. More formally, each matrix entry $C_{i,j}$ is non-zero if traveller j has visited station i, and i is not (one of) j's familiar locations.

3. Create Community Flow Matrix. Now that we have both home locations (Step 1) and visit frequencies (Step 2), we compute a station-by-station flow matrix F which represents which locations community members visit. Each entry $F_{i,j}$ counts the number of people who live in j and who have visited i. If a particular traveller has two inferred familiar locations (h_1, h_2), we count the provenance of each visit to i as 0.5 from h_1 and 0.5 from h_2. Note that the flow matrix does not take into account the frequency of a user's travel to an area; it just accounts for whether or not she visited it. After this step, we have the data we need: a mapping from stations → *IMD* values and a flow matrix of stations → stations. We next investigate what this data can reveal by performing two steps:

1. Correlate *IMD* and Flow. The correlation (or the extent that a linear relationship exists) between two vectors of values is computed using the Pearson correlation coefficient. Given a vector X with mean μ_X (e.g., *IMD* values) and Y with mean μ_Y (e.g., flow values), the correlation is defined as the covariance of the two variables divided by the product of the standard deviations. To perform this, we need to condense our flow matrix F into a vector of values, one per station; we define the flow f_i into an area as the sum of the areas that it receives visitors from:

$$f_i = \sum_i F_{i,j} \tag{1}$$

2. Compute Homophily Indices. We also delve further into the flow matrix by computing two different scores that measure the homophily of each community. The first, which we call the *social equaliser* index, measures the extent to which an area attracts people from areas of varying deprivation:

$$H_1(i) = STD\left(\frac{\sum_j F_{i,j} * IMD_j}{\sum_j F_{i,j}}\right) \tag{2}$$

where STD is the standard deviation of the average enclosed in parenthesis. Intuitively, if $H_1(i)$ is high, then area i is a *social equalizer*: it attracts visitors from areas of varying deprivation (high standard deviation). If it is low, then people in area i tend to flow between areas with people of similar social deprivation. The second, which we call the *heterogeneity* index, measures the extent to which an area attracts people from areas of with similar deprivation:

$$H_2(i) = \frac{\sum_j F_{i,j} \cdot |IMD_j - IMD_i|}{\sum_j F_{i,j}} \tag{3}$$

If $H_2(i)$ is high, then the area i attracts areas different from itself (*heterogeneous* pair of areas having different *IMD* scores); if it is low, then area i attracts areas that are similar to itself. Finally, to examine the relation between community homophily and social deprivation, we computed the correlations between H_1 and IMD as well as H_2 and IMD.

4 Results: Correlating Mobility and Well-Being

We study the Pearson product-moment correlation between *IMD* and metrics of urban flows. Weak, yet statistically-significant, correlations are found between an area's deprivation *IMD* score and the number f_i of areas from which it receives visits (correlation coefficient $r = 0.21$ with $p < 0.001$), suggesting that the more deprived the area, the more it tends to be visited. Considering the *social equaliser* index H_1, we find that it is not correlated with *IMD* ($r = 0.02$ with $p < 0.001$). This means that, in general, there is no homophily effect: Londoners do not tend to visit communities having deprivation scores that are similar to their own communities'. However, we find that *IMD* is negatively correlated with the *heterogeneity* index H_2 ($r = -0.16$ with $p < 0.001$), suggesting that heterogeneity holds only for well-off areas. These areas tend to attract people living in areas of varying deprivation. By contrast, Londoners in well off areas do not tend to visit communities that are deprived. This suggests that segregation effects are observed only in deprived areas, and that has important implications in policy making. Finally, to study how the number of visiting areas and the second (*heterogeneity*) index contribute in explaining the variability of *IMD*, we ran a linear regression of the form:

$$IMD_i \sim \alpha + (\beta_1 \times log(H_2(i))) + (\beta_2 \times f_i) \tag{4}$$

In so doing, we obtain $R^2 = 0.16$, indicating that 16% of the variation in the *IMD* is explained only by the two indicators $H_2(i))$ and f_i. Furthermore, the most important predictor is the *heterogeneity* index ($\beta_1 = -0.51, p < 0.001$) and the contribution of f_i is significantly reduced and becomes negligible ($\beta_2 = 0.001, p < 0.001$).

5 Limitations and Applications

The results above take the first step into examining how data from pervasive technology can be used to investigate social mixing and homophily of urban communities. In this section, we discuss the limitations of our study as well as the theoretical and practical implications of the results we obtained. The public transport data that we have is rife with uncertainty: we do not know the exact home locations of travellers and we had no choice but to drop all bus trips since passengers do not have to use their card when reaching their destination. Our view of the city is also incomplete: we do not have data relating to the penetration of Oyster cards in various communities, which prevents us from knowing the extent that our results are skewed by communities opting for non-public modes of transport (regardless of the reason, e.g., well-off communities using cars). We also do not have data about urban density, in order to normalise against the variability in the number of people who live in different communities. We assume that access to this data would allow us to produce stronger results. Furthermore, we are tied to existing infrastructure: we could only analyse those portions of the city that are covered by the transport network, and the definition of community that we have adopted is in relation to this infrastructure (i.e., each station belongs to one community). We acknowledge that this mapping may not be fully accurate (or indeed capture the entirety of the metropolitan area's communities); a station may sit at the border of two adjacent communities. The results support the emerging research that calls for urban planners [8] and policy makers [10] to leverage mobility data when making and evaluating their decisions. In fact, the lack of coverage limitation of our study may be used alongside *IMD* values to estimate which communities would most benefit from new transport infrastructure. This data may also prove to be invaluable for building tools that monitor the visibility of physical communities, in order to augment longitudinal studies with dynamic and large-scale data.

6 Conclusion

We have used fare collection data to measure how the way people move about cities can be used as an implicit indicator of the visibility of communities. Various fare collection systems are in use in hundreds of other cities throughout the world: repeating this study, as well as discovering novel uses of the data that these systems generate, is a promising area of research. We have three directions of future work. First, we plan on addressing limitations described above by re-examining the relation between home location and travel patterns. We have also measured community visibility from a broad, aggregate view; in practice, the mobility of visitors into a community will be tied to the social events and facilities (work, educational institutions, social venues) in that area. We thus plan to investigate how flows deviate from normal patterns during large-scale events, in order to discover how the dynamics of urban life influence the social well-being of the area. Recent work [18] has also uncovered a relation between *IMD* scores and social media (tweets') sentiment; we plan to enrich the study above by investigating the meeting point of offline physical data and online user-generated content, which increasingly intersect by being geo-located.

References

1. K. Lynch. *The Image of the City.* MIT Press, Cambridge, MA, 1960.
2. S. Milgram. *The Individual in a Social World.* Pinter and Martin, London, UK, third edition edition, 2010.
3. P. Lynn. Maintaining Cross-Sectional Representativeness in a Longitudinal General Population Survey. *Understanding Society Working Paper*, June 2011.
4. Z. Cheng, J. Caverlee, and K. L. D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *AAAI ICWSM*, 2011.
5. J. Froehlich, J. Neumann, and N. Oliver. Sensing and Predicting the Pulse of the City through Shared Bicycling. In *21st IJCAI*, Pasadena, California, 2009.
6. K. Rachuri et al. EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *ACM UbiComp*, 2010.
7. N. Eagle and S. Pentland. Reality Mining: Sensing Complex Social Systems. *Pers. Ubiquitous Computing*, 10:255–268, 2006.
8. Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban Computing with Taxicabs. In *ACM UbiComp*, 2011.
9. V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of Socioeconomic Levels using Cell Phone Records. In *UMAP*, 2011.
10. N. Lathia and L. Capra. How Smart is Your Smartcard? Measuring Travel Behaviours, Perceptions, and Incentives. In *ACM UbiComp*, 2011.
11. A. Bawa-Cavia. Sensing the Urban: Using Location-Based Social Network Data in Urban Analysis. In *Pervasive PURBA Workshop*, 2011.
12. F. Girardin et al. Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Computing*, 7, 2008.
13. N. Eagle, M. Macy, and R. Claxton. Network Diversity and Economic Development. *Science*, 328, 2010.
14. M. Noble et al. The English Indices of Deprivation. *The Department of Communities and Local Government*, Mar. 2008.
15. L. S. Weinstein. Tfl's contactless ticketing: Oyster and beyond. In *Transport for London*, London, UK, Sept. 2009.
16. N. Lathia, J. Froehlich, and L. Capra. Mining Public Transport Usage for Personalised Intelligent Transport Systems. In *IEEE ICDM*, 2010.
17. M.C. Gonzalez, C.A. Hidalgo, and A-L. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196):779–782, 2008.
18. D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking "Gross Community Happiness" from Tweets. In *ACM CSCW*, 2012.

Acknowledgments. This research has received funding from the EC FP7 Programme (FP7-SST-2008-RTD-1; grant agreement n. 234239) and from the RCUK through the Horizon Digital Economy Research grant (EP/G065802/1).