A Network-based Study of Built Environments and their Effect on Physical Activity

Boris Ivanovic
Department of Computer Science, Stanford University
450 Serra Mall, Stanford, CA 94305
borisi@cs.stanford.edu

NOTE: Materials related to this project are not allowed to be uploaded to the course website or any public repository of projects.

Abstract

Physical inactivity is a major global pandemic responsible for millions of deaths per year. Understanding the factors affecting physical activity is a crucial part of determining how to encourage activity and influence healthy behaviors. In this work, we determine how physical activity is affected by the built environment using quasi-experimental network analysis methods. Specifically, we determine how a user’s physical activity level (as measured by their daily steps) is affected by changes to their local environment (measured quantitatively by neighborhood walkability metrics). We base our analyses on prior work in population migration, taking it to much finer timescales and using fine-grained user localization to identify the effect of location characteristics on physical activity. Our main contributions in this work are the characterization of the network defined by human mobility across 1600+ locations in the USA.

1. Introduction

Over 5 million deaths per year are related to physical inactivity[6]. Understanding quantitatively what factors affect physical activity is a necessary part of understanding how to prevent inactivity and influence healthy behaviors. The main focus of our work is determining how physical activity is affected by the built environment using quasi-experimental network analysis methods. Specifically, we wish to determine how a user’s physical activity level (as measured by their daily steps) is affected by changes to their local environment (measured quantitatively by neighborhood walkability metrics). We base our analyses on prior work in population migration, taking it to much finer timescales and using fine-grained user localization to identify the effect of location characteristics on physical activity. Our main contributions in this work are the characterization of the network defined by human mobility across 1600+ locations in the USA.

2. Related Work

2.1. Migration Flows

Abel and Sander estimate migration from sequential stock tables published by the United Nations (UN) over the past 30 years[1]. They use maximum likelihood to estimate the number of movements required to meet the observed changes over time in migrant stock data, using an iterative proportional fitting algorithm. In essence, they are analyzing differences in snapshots of the world’s mobility graph across time. Fundamentally, Abel and Sander analyze the difference between snapshots of a graph at differ-
ent times and make inferences about the time between, rather than directly observing movement and its effect on people. Further, and perhaps more seriously, such long timescales between data points (the UN publishes stock tables every 10 years) makes their inference between points hard to believe, especially since this paper only uses data from 3 data points (the 1990, 2000, and 2010 stock tables).

Works studying migration flow usually focus on inter-country or intercontinental migration, which is a limitation in understanding where people move and potentially why (e.g., cities within the same country have vastly different characteristics and accounting for these unique affects is important to understanding overall effects). Tying this back to our graph idea of the world's mobility network, this work purely analyzed edges and how they change over time. Comparatively, our work uses finer granularity data and analyses to identify both when people move and to where.

2.2. Physical Activity as a Function of Location

The seminal work of Sallis et al. showed that where one lives has an effect on their physical activity, as measured by accelerometer units provided for the study [9]. The variety of locations chosen throughout the world helped illustrate that significant built environment differences lead to significant physical activity differences. This paper is akin to a detailed analysis of node characteristics in the world’s mobility graph, informing the hypothesis of our work and indicating that accelerometer data is a sensible choice for data-gathering. This work would have severely benefited from a much larger dataset, as manual data collection suffers from diminishing returns with respect to cost and does not scale to the size of country populations. Further, the study is limited in that there is no information about the effects of different locations on the same population. For example, are users in Seattle, WA more physically active than those in Chicago, IL because Seattle is more conducive to being physically active or because the people in Seattle are intrinsically more prone to being active? This question remains unanswered after reading the paper, but is a central theme of our work.

2.3. Prior Large-Scale Analyses

Recently, large-scale data science approaches have been applied to understanding and showing the existence of worldwide physical activity inequality [2]. Althoff et al. similarly identified that the basic principles that govern physical activity require more study and that there is a lack of large-scale measurements of physical activity patterns across free-living populations. With respect to the world’s mobility graph, this work provides a large-scale study of nodes, specifically how men and women differ in activity in fine-grained locations. This work’s dataset catapults our abilities to analyze large-scale physical activity and moving patterns forward, addressing many concerns with the above mentioned works and limited data availability. However, further research questions remain: What would happen if men and women from highly activity-unequal locations moved to activity-equal locations? Would their behavior in the new location stay unequal or would they be affected by the new environment and become more equal? Our work aims to address these questions.

Another large-scale data scientific approach was presented in [11], where Zagheni et al. address the same problem as [1], but use much finer timescales. This addresses the long timescale inference issue we highlighted previously. Zagheni et al. utilize free data from Facebook’s Advertising Platform to identify groups of migrants in specific locations and compare the counts across time. While this work dives deeper into location granularity, the origins of the identified population groups are still on an inter-country or inter-continental scale, leaving more detailed source location information to be desired.

[5] researches the similar migration estimation problem of [1, 11], using more direct geolocation information from Twitter posts. They focused
only on locations within the USA, which was justified as a choice to not involve the difficulties associated with modeling and accounting for different country behaviors. Unlike [1, 11], Fioro et al. focus on detailed locations (county-level) in the US and estimate migration rates ignoring the original user ethnicity or birthplace. We agree with this choice and feel the approach better characterizes a location and its physical activity-affecting properties. This paper is similar in method to our idea, following a large amount of individual users as they travel from location to location. The only piece missing from this work is a consideration of physical activity, which our work covers.

3. Method

3.1. Raw Data

Our raw dataset is a collection of data from fitness apps owned by Azumio from 2014 through to early 2016 [2]. In it, user physical activity as counted by cellphone accelerometers, a method validated in prior studies [9]. It contains 68 million days of physical activity from 717,527 people [2]. This raw data is in the form of row entries for each day of user activity, geolocated and timestamped.

3.2. Data Filtering and Augmentation

We first filtered our dataset down to the 147K users that have 25 or more location-containing check-ins, 5 or more location-containing check-ins at a second location, and both locations being in the USA. We focus only on locations within the USA as it is the country with the largest amount of data and we do not wish to consider cross-country behaviors, similar to [5].

We then identified and added new data resolving available location information (in the form of wunderground.com URLs) in our dataset with concrete US cities via distributed web scraping of wunderground.com. For example, a daily steps check-in with the location URL https://www.wunderground.com/weather/us/il/shiloh is located in Shiloh, FL (after visiting the link and verifying the indicated location).

To obtain walkability metrics for cities, we web scraped walkability scores[4] for 2,500 locations from all 50 US states, more than any previous study. These scores are an experimentally-validated measure of neighborhood walkability that take into account local features and distance to amenities [4].

In order to use quasi-experimental methodology, we additionally designed and implemented an efficient algorithm which detects users who moved within the US. Our definition of a move is illustrated in Fig. 1.

In order to remove miscellaneous stops and immediate move-effects, we ignore data from 5 days before and after a move (e.g. you may be more active after moving since you are shopping for furniture, carrying boxes, etc).

Finally, in order to ensure that we have enough statistical power when averaging a user’s daily

\[\text{From https://www.walkscore.com/}\]
steps before and after moves, we additionally require that a user has at least 10 steps check-ins within 30 days before and after the move.

We used SNAP.py, pandas, and seaborn for all data processing, analysis, and visualization [7, 8, 10].

3.3. Resulting Dataset and Graph

From our original 147K two-location users, we identified 7,430 moves from 5,410 users using the criteria described above. Summary statistics are shown in Table 1. A plot of the locations captured in our dataset is shown in Fig. 2.

We defined a graph $G = (V, E)$ over these moves where $V$ is the set of locations for which we have a move going to or leaving from and $E$ is the set of directed edges $(i, j) \in E$ such that there exists a user that moved from location $i$ to $j$. After creating this graph, we have 1,607 nodes and 6,319 edges, where the number of nodes and edges corresponds to “Cities” and “City Combinations,” respectively, in our dataset summary table (Table 1).

4. Results

4.1. Graph Structure

After defining the above graph, one can analyze its structure and properties. Fig. 3 shows the graph’s degree distribution, and fits a power law distribution to it. It is intuitive that the node degree distribution follows a power law because the underlying population growth follows a preferential attachment model as larger cities will attract more people because there’s more jobs, people, restaurants, music venues, etc.

Since we can link nodes to the cities they represent, we can directly give real-world meaning to these degree counts. The largest degree node in our mobility graph is New York City, NY with a degree of 212. This means there were 212 moves to or from New York City captured in our dataset. The top 10 in- and out-degree nodes are listed in Tables 2 and 3, respectively. As one might expect, they are all large cities (following our prior intuition).

We also wanted to see how this graph compares to other real-world social networks. The average clustering coefficient of our graph is 0.10 which is very similar to the MSN network from class. Our graph’s average shortest path length is 3.7 which is much shorter than MSN’s value of 6.6. This makes sense because our graph has many more people in one node than in a social network where a node is usually one person. Thus, it only takes a few people moving from one city to connect it to US travel hubs or span large geographic

Table 1: Dataset counts. “Users” is the number of moving users. “Days” is the number of daily steps check-ins from moving users. “Moves” is the number of moves satisfying our move definition and filtering. “Cities” is the total number of cities. “City Combinations” is the number of distinct city pairs.
Figure 3: **Left:** Node degree distribution visualized. A power law distribution was fit with least squares, resulting in $\alpha_{ls} = 1.35$. **Right:** Node degrees visualized on a map of the contiguous US. Node size and color are proportional to node degree, higher degree nodes are larger in size and darker in color. Colorbar values indicate node degrees. This detailed figure is best viewed digitally at high magnification.

Table 2: Top 10 in-degree nodes in our graph. For us, in-degree represents how many users moved to a location.

<table>
<thead>
<tr>
<th>Location</th>
<th>In-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York, NY</td>
<td>113</td>
</tr>
<tr>
<td>Chicago, IL</td>
<td>106</td>
</tr>
<tr>
<td>Las Vegas, NV</td>
<td>95</td>
</tr>
<tr>
<td>Orlando, FL</td>
<td>95</td>
</tr>
<tr>
<td>San Diego, CA</td>
<td>84</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>83</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>75</td>
</tr>
<tr>
<td>Atlanta, GA</td>
<td>70</td>
</tr>
<tr>
<td>Dallas, TX</td>
<td>58</td>
</tr>
<tr>
<td>Seattle, WA</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 3: Top 10 out-degree nodes in our graph. For us, out-degree represents how many users moved out of a location.

<table>
<thead>
<tr>
<th>Location</th>
<th>Out-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York, NY</td>
<td>99</td>
</tr>
<tr>
<td>Orlando, FL</td>
<td>91</td>
</tr>
<tr>
<td>Chicago, IL</td>
<td>90</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>86</td>
</tr>
<tr>
<td>San Diego, CA</td>
<td>85</td>
</tr>
<tr>
<td>Las Vegas, NV</td>
<td>80</td>
</tr>
<tr>
<td>Houston, TX</td>
<td>77</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>75</td>
</tr>
<tr>
<td>Atlanta, GA</td>
<td>71</td>
</tr>
<tr>
<td>Boston, MA</td>
<td>64</td>
</tr>
</tbody>
</table>

distances across via airplanes, whereas in a social network it would be rare to have many long-distance connections to people that your friends’ friends’ friends’ ... friends do not know.

To further understand the graph’s structure, we analyze and visualize its connected components. The directed human mobility graph we defined has a maximum strongly-connected component (SCC) size of 1,062 nodes, covering 66.1% of the graph’s nodes. It has a maximum weakly-connected component (WCC) size of 1,575 nodes, covering 98.0% of the graph’s nodes. Fig. 4 shows which nodes are in the largest SCC. It appears that most of the red nodes are in the eastern US, with a majority of high degree nodes in the east being in the largest SCC. The reason the maximum SCC size is only 66.1% of all nodes is because our graph is directed over longer-term user moves, and not short-term trips which would create “to” and “from” edges for
nearly all travel, essentially creating an undirected graph. If one were to treat the edges in this graph as undirected, then the maximum SCC size balloons to 98.0% of all nodes (matching values from social networks, e.g. MSN’s value of 99%).

To understand which locations are a part of the SCC, we visualized nodes in our mobility network in Fig. 4 and colored them by membership in our largest SCC. It is evident that many edges are between east coast cities and between west coast cities, with long moves across the US connecting the coasts. Thus, a natural clustering of nodes is to split them east-west by geographical location. Although we were hoping for more “high-level” semantic groupings such as by industry, this result makes more intuitive sense as a location’s geography and its resulting industrial landscape are tightly coupled. East coast cities are generally older and have been business hubs for many old and new industries. West coast cities are relatively newer and attract younger industries with the ability to be distanced geographically from conventional business hubs.

In summary, we identified that the US human mobility network behaves similarly to a real-world social network. However, special characteristics such as the frequency of long-distance travel and hub locations allow our network to have a smaller average shortest path length.

4.2. Changes in Physical Activity vs. Changes in Location Walkability

With moves defined, we can analyze how a user’s physical activity changes vs. how their location walkability changes. To perform this analysis, we

1. Averaged all moves’ daily steps before and after the move.
2. Subtracted all moves’ source daily steps average from their destination daily steps average (creating the dependent variable).
3. Subtracted all moves’ source walkability from their destination walkability (giving us a notion of how much of a change there was to the user’s built environment, creating the independent variable).

Finally, we plotted the dependent variable against the independent variable and Fig. 5 was obtained. It reveals that a user’s daily physical activity is linearly related to the user’s location walkability! This indicates that a user’s physical activity habits are not necessarily intrinsic to them, but are affected by their surroundings as otherwise a change in location would not affect physical activity. We verified that there are few self-selection effects, if any, by comparing metrics from our movers to results from the 2016 Current Population Survey Annual Social and Economic Supplement (CPS ASEC). Fig. 6 illustrates the age and gender distribution of our movers compared to the CPS ASEC results.

Now that we have identified that a change in the built environment affects physical activity, deeper analysis into these changes and how they manifest through time can be performed. Two questions stand out:

1. After a move, does a user’s physical activity change immediately or does it take time affect the user?
2. What kinds of physical activity is affected, and how much does it change by?

To answer the first question, we looked at how our movers’ physical activity changes with respect to time around a move. Fig. 7 shows a few examples of this longitudinal analysis. It is evident that an individual’s physical activity changes immediately following a move, hinting that physical activity habits are not entrenched and there is no internal “resistance” to changes in physical activity when the surrounding built environment changes.

To answer the second question, we looked at the underlying dataset that generated our daily steps data. We looked at minute-level step counts

---

2Summary data available through the US Census.
Figure 4: **Left:** Nodes in red are part of the largest SCC. **Right:** Our mobility graph visualized above a map of the contiguous US. Each line denotes a single edge. Nodes in red are part of the largest SCC. Moves to and from Alaska and Hawaii are not visualized here. Edge directionality is not shown. Node size is proportional to node degree, higher degree nodes are larger in size. This detailed figure is best viewed digitally at high magnification.

Figure 5: How much a user’s daily steps change vs how much their location walkability changes. Δ Walkability is the difference between the destination’s and source’s walkability. For example, if a user moved from New York City, NY with walkability 89 to Chicago, IL with walkability 78 they would have ΔWalkability = 78 − 89 = −11. All error bars are 95% bootstrapped confidence intervals.

for each moving user and compared their histogram of step activity before and after a move. Fig. 8 summarizes our findings. Moving individuals gain and lose active exercise as defined by current guideline[3] hinting that the built environment has a strong influence on whether users exercise. Since we can now relate movers and the amount of time they perform active exercise, we additionally looked into how the amount of time exercising changes with respect to walkability as well as how many users met physical activity guidelines before and after moves with respect to the move’s walkability change. Fig. 8 summarizes our findings and reinforces the notion that the built environment has a strong influence on an individuals’ exercise.

5. Conclusion

We present significant results relating physical activity and the built environment using quasi-experimental network analysis methods. Specifically, we determined how a user’s physical activity level (as measured by their daily steps) is affected by changes to their local environment. We characterized the network defined by human mobility and found that it matches social networks well, with slight differences owing to how the graph is defined. We analyzed its structure, identifying coastal areas as two major communities with long arcs connecting them. We also looked

at which regions have the most movers, and found intuitive results that show population is the dominant factor determining the amount of movers, making large cities the biggest sources of mobility. These results can be used in public health and urban planning to identify how to build better cities and quantify the benefits that increasing location walkability would have on resident physical activity.

In the future, it would be interesting to characterize the effect of weather on physical activity since it is a factor within locations, leading to even more detailed models of physical activity. This would allow us to identify how much a location’s walkability even matters if the weather is not conducive to exercise or how much it is boosted if the weather is conducive to exercise. Additionally, using the linear models in this work to predict how physical activity changes as cities grow over time would develop interesting notions of how urban sprawl may benefit or harm physical activity, depending on how a location’s walkability changes as a result of urban growth.

References


Figure 7: Participants physical activity levels undergo significant changes following relocation to and from specific locations of different walkability. Examples show moving from/to New York, NY, San Jose, CA, and Albuquerque, NM. Physical activity levels change significantly by about 1000-2000 daily steps depending on the location. Note the symmetry between moving from (left) and to (right) specific locations.


Figure 8: Walkability differences lead to differences in moderate-to-vigorous physical activity (MVPA). **Top:** Changes in physical activity following relocation to higher (left) and lower (right) walkability environments, stratified by intensity of physical activity. Note the symmetry in effects when walkability increases (left) or decreases (right). **Middle:** Physical activity is essentially unchanged across all intensities when walkability remains relatively constant. **Bottom Left:** Differences in walkability versus change in time spent in MVPA (minutes/week). Large increases in walkability are associated with an increase of one hour in MVPA. **Bottom Right:** The increases in time spent in MVPA are associated with significantly more participants meeting national physical activity guidelines of 150 min/week or more of MVPA. All error bars correspond to bootstrapped 95% confidence intervals.