IDENTIFYING COMMONLY USED AND POTENTIALLY UNSAFE TRANSIT TRANSFERS WITH CROWDSOURCING

Elizabeth J. Traut[a], Aaron Steinfeld[b]

[a] (Corresponding Author)
Carnegie Mellon University
Robotics Institute
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
Present Address:
Penn State
Larson Transportation Institute
204 Transportation Research Building
University Park, PA 16802, USA
etraut@psu.edu
Phone: +1 814-865-0500

[b] Carnegie Mellon University
Robotics Institute
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
as7s@andrew.cmu.edu
Phone: +1 412-268-6346

Keywords: public transit; crowdsourcing; transfers; public transportation

Post-Print Version

Traut, E. J., and Steinfeld, A. (2019). Identifying commonly used and potentially unsafe transit transfers with crowdsourcing. *Transportation Research Part A: Policy and Practice*, *122*, 99–111. https://doi.org/10.1016/j.tra.2019.02.005

1 ABSTRACT

Public transit is an important contributor to sustainable transportation as well as a public service that makes necessary travel possible for many. Poor transit transfers can lead to both a real and perceived reduction in convenience and safety, especially for people with disabilities. Poor transfers can expose riders to inclement weather and crime, and they can reduce transit ridership by motivating riders who have the option of driving or using paratransit to elect a more expensive and inefficient travel mode. Unfortunately, knowledge about inconvenient, missed, and unsafe transit transfers is sparse and incomplete.

We show that crowdsourced public transit ridership data, which is more scalable than conducting traditional surveys, can be used to analyze transit transfers. The Tiramisu Transit app merges open transit data with information contributed by users about which trips they take. We use Tiramisu data to do origin-destination analysis and identify connecting trips to create a better understanding of where and when poor transfers are occurring in the Pittsburgh region. We merge the results with data from other open public data sources, including crime data, to create a data resource that can be used for planning and identification of locations where bus shelters and other infrastructure improvements may facilitate safer and more comfortable waits and more accessible transfers. We use generalizable methods to ensure broader value to both science and practitioners.

We present a case study of the Pittsburgh region, in which we identified and characterized 338 transfers from 142 users. We found that 66.6% of transfers were within 0.4 km (0.25 mi.) and 44.1% of transfers were less than 10 minutes. We identified the geographical distribution of transfers and found several highly-utilized transfer locations that were not identified by the Port Authority of Allegheny County as recommended transfer points, and so might need more planning attention. We cross-referenced transfer location and wait time data with crime levels to provide additional planning insight.

2 INTRODUCTION

Public transit is an important contributor to sustainable transportation as well as a public service that makes necessary travel possible for many. Poor transit transfers can lead to both a real and perceived reduction in convenience and safety. Long transfers and traversal between transfer stops can be especially problematic for people with disabilities due to exposure to inclement weather and the risk of theft and other crimes, while tight transfers can be problematic due to inability to run or the need to use a (potentially longer) wheelchair accessible route between stops (National Council on Disability, 2015). This aligns with findings that real-time arrival information led to higher perceptions of safety for riders without disabilities (Ferris et al., 2010) and that personal safety is one of the most important characteristics of a satisfactory transfer (Iseki and Taylor, 2010). Knowledge about safety problems like these and missed transit transfers is sparse and incomplete since it can be time-consuming and expensive to collect travel data from consumers.

Missed transfers and long waits can be inconvenient, suppress optional travel, and expose riders to inclement weather and crime. Riders with the option of driving may abandon transit (or never chose it) due to the perception of unsafe and poor transfers. Poor transfers and the resulting exposure to adverse conditions can also motivate riders who have the option of using paratransit to elect the more expensive travel mode. Good data is needed on exactly how long people are waiting and what pedestrian patterns are occurring for transfers where stops are not co-located.

Some of this information can be deduced from smartcard fare data by looking at points of payment across each piece of the trip. However, for many bus trips this only provides limited insight on the intended trip of the rider due to the lack of origin-destination (O-D) ground truth, since riders may pay at one end of the trip but not reveal the other end. This makes it difficult to determine if a rider used a common stop for the transfer or a wider gap in order to run an errand or another pedestrian activity before transferring. Likewise, it can be difficult to determine the origin or destination stop of the overall trip, thus making it very hard to assess demand for more direct routes. It is possible to make some initial guesses based on historical patterns, but our machine learning analyses using ground truth O-D stop data have shown performance for such estimates to be below the level needed for planning purposes.

Unfortunately, it is also hard to disambiguate intentional and unintentional transfer gaps. As implied above, it is currently not possible with traditional data sources to easily disambiguate intentional gaps between trips from missing a transfer. For example, a gap in a transfer could indicate a delay in the arrival of the second bus, that a bus was missed due to insufficient transfer time, or that the rider stopped to get coffee at a nearby shop. Therefore, it would be useful to have data mining tools for removing intentional gaps from analyses. Similarly, knowledge of which transfers have a high incidence of intentional gaps could be used to identify local resources and infrastructure features that are valuable to riders when transferring.

We use crowdsourced data from the Tiramisu Transit app to create a better understanding of where and when poor transfers are occurring in the Pittsburgh region (Steinfeld et al., 2012; Tomasic et al., 2015, 2014). We compare these locations with local crime data to infer safety risk. The resulting information is provided for use during planning and identification of locations where policing, bus shelters, and other infrastructure improvements may facilitate more safe and comfortable waits. We use generalizable methods to ensure broader value to both science and practitioners.

3 LITERATURE REVIEW

Although bus transit planning is well studied, the literature on transfers and on identifying poor transfers in particular is sparser. The literature on transit planning that includes transfers often assumes or simulates full information about (or probability distributions for) origins and destinations (Benenson et al., 2016; Michel and Chidlovskii, 2016; Ngamchai and Lovell, 2003) or relies on (real or simulated) ridership data without known O-D pairs (Cevallos and Zhao, 2006; Gschwender et al., 2016; Shafahi and Khani, 2010; Zhang et al., 2017). Some studies also relate to specific transfer issues such as multi-model transfers (Chung and Shalaby, 2007; Seaborn et al., 2009) and optimal scheduling and real-time control strategies for timed transfers (Abkowitz et al., 1987; Chung and Shalaby, 2007; Daganzo and Anderson, 2016; Kieu et al., 2017). Some literature focuses on the effect of transfer characteristics and infrastructure on passengers' perception of transit (Fan et al., 2016; Guo and Wilson, 2004; Iseki and Taylor, 2010; Ji et al., 2017).

More recently, some transit systems have started collecting full origin and destination information for many or all trips, making detailed studies possible for those systems, including transfers. Jang used city-wide smart card data from Seoul that included boarding information for all bus trips and alighting information for most, as the local agency requires tapping before exiting the bus in order to get the transfer fare reduction (Jang, 2010). This data allowed Jang to perform a full transfer analysis. The study found that 80% of transfers were under 10 minutes and identified the most popular transfer locations. The study also identified transfer locations where a significant portion of trips involved transfers longer than 10 minutes as places for potential schedule or infrastructure improvement.

Different methods are needed for the many transit systems that do not collect full passenger O-D information. Pinelli et al. (Pinelli et al., 2016) present an alternative method utilizing large-scale location data from mobile phone networks and deriving frequent patterns of movement to generate candidate routes for transit. This method captures movements of large numbers of people between two points but does not capture the detail of each user's exact origin and destination. This approach is also not able to differentiate overlapping bus routes. Another approach is to infer O-D pairs from whatever fare card data is available (boarding or alighting but not both) (Chu and Chapleau, 2008; Gordon et al., 2013; Hanft et al., 2016; Hofmann and O'Mahony, 2005; Ma et al., 2013; Rahman et al., 2016; Sánchez-Martínez, 2017; Seaborn et al., 2009; Wang et al., 2011; Zhao et al., 2017), but this method requires making assumptions about which stops will be involved in transfers and so does not provide ground truth. Data from inference models can be used for some types of transfer analyses such as reducing the number of trips that require multiple transfers (Hanft et al., 2016) but is not useful for fully identifying and characterizing the transfers.

Our study provides an alternate method based on crowdsourcing for analyzing poor transfers in systems where full O-D bus trip information is normally unavailable.

4 METHODS

We combined crowdsourced O-D transit rider trip data from the Tiramisu Transit app with open and publicly available data from several sources to identify and characterize poor bus transfers in the Pittsburgh region. We (1) identified cases when users had provided data on multiple trips in close temporal proximity to identify connecting trips, (2) identified which pairs of trips were transfers versus other events (such as round trips), (3) examined the spatial and temporal gaps during transfers to understand where and when poor transfers were occurring, and (4) combined transfer data with crime data to identify potential safety risks.

4.1 Terminology

We use the terms *transfer* to refer to a situation where a transit rider must disembark from one bus (or light rail) and board another bus (or light rail) in order to travel to the desired destination. The transfer may or may not involve walking between two stops and may or may not include an intentionally longer than necessary gap, such as to run an errand near the stop. Although an intentional gap to run an errand would more properly mean that the situation is not a transfer but rather a *trip chain*, we refer to both situations as transfers because they cannot be disambiguated within the scope of this paper. Unless otherwise specified, we use the word *trip* to refer to one leg of travel by a transit rider from boarding to disembarking. The term *device* corresponds to the unique identifier given by the app to the smartphone it is loaded on. Note that a new number is generated if the user reloads the Tiramisu app on a new phone or after deleting the app. Therefore, a device ID is a good, but not exact, estimate of a unique user. We will refer to *buses* throughout the paper for convenience, although the dataset also includes a small number of trips taken on light rail.

4.2 Data

We combined crowdsourced transit ridership data from the Tiramisu Transit app with open and publicly available data from several sources. We have limited our analyses to the Pittsburgh, PA metro region.

4.2.1 Crowdsourced Bus Ridership Data from Tiramisu

During the data collection period used for this study, the Tiramisu Transit app provided users with real-time bus arrival information via crowdsourcing in addition to any real-time data available from transit agencies. Tiramisu users could trace their trips to contribute to Tiramisu's real-time arrival estimates, which were available before the Pittsburgh bus system started providing AVL (automatic vehicle location) data publicly, and which were still available if and when AVL was not.

Tiramisu collected ground truth O-D stop data whenever riders traced their transit trips to contribute to real-time estimates. Tiramisu users selected their transit stop in the app to see estimated bus arrival times. From there, they could select the bus they would board, select their destination stop from a list, choose the fullness of the bus (from options of Empty, Seats Available, No Seats Available, or Full), and select "Start Recording" to begin tracing, or sharing location information. Tracing automatically stopped when the user neared the destination stop, or they could hit a stop button at any time (Tomasic et al., 2015). For more information on the Tiramisu app, see Steinfeld et al. 2012 (Steinfeld et al., 2012), Tomasic et al. 2014 (Tomasic et al., 2014), and Tomasic et al. 2015 (Tomasic et al., 2015).

The Tiramisu app collected data on where users were, which transit stops they sought data about, and which trips they took (when tracing). The system logged time, date, location, and which data was requested per user whenever the app was opened or new data was requested. When tracing, the app logged the route, the GTFS trip ID, the user-selected origin and destination stops, a user-selected indicator of how full the bus was, and regular timestamped updates on the user's location (trace points) as they progressed along the trip.

In our Pittsburgh case study, Tiramisu trace data was mostly for bus trips, but also included some light rail trips. The Tiramisu database for version 1 of the app was used for this study. It contains 862,564 trace points from July 13, 2011 to October 20, 2015.

Payment on all modes on the Port Authority of Allegheny County system occurred only at the start or end of the trip, depending on direction of travel and time of day. There was also a free fare zone within the downtown core. Therefore, smartcard data does not contain full ground truth data for transfers. Additionally, during the period of data collection for this study, smartcards had not been fully adopted by the Port Authority.

4.2.2 General Transit Feed Specification (GTFS)

Tiramisu used openly available GTFS data to provide bus and light rail schedule information for the Port Authority of Allegheny County (Port Authority of Allegheny County, 2016). We used GTFS data for bus routes, scheduled stop times, and stop locations in our analysis.

According to the GTFS data for the examined time period, there were 7 pairs of stops, all located downtown, between which transfers (in both directions) were recommended by the Port Authority of Allegheny County, and one pair of stops between which transfers were not possible, located in Bridgeville, PA. The GTFS did not specify any timed transfers.

4.2.3 Map Data

We used open map data from OpenStreetMap to provide geographical context (major roads, minor roads, and rivers) when mapping bus ridership data (OpenStreetMap Contributors, 2016a). The Pittsburgh region map data for 2013 was originally obtained from the previous metro extracts site (Migurski and OpenStreetMap Contributors, 2013) and the current version is now available from Mapzen (OpenStreetMap Contributors, 2016b).

4.2.4 Crime Data

We used 2010 crime data for neighborhoods and planning sectors from PGHSNAP (Pittsburgh City Planning, 2012).

4.3 Data Cleaning to Determine Ground Truth O-D Trips

Our first step was to clean the trace data from Tiramisu and obtain traced transit riders' trips with ground truth for origin and destination stop locations. This included removing irrelevant parts of the Tiramisu data and removing data points with GTFS errors, likely caused by changes in the GTFS over time being improperly logged or accounted for in the interface in some past versions of the app. Note that in the Tiramisu data the GTFS agency ID variable is used to indicate the associated GTFS version for each trace point, because only one agency's GTFS was logged and it had frequent version changes (with every route or schedule change). Also note that in the GTFS data a *trip* (which we will call a "GTFS trip") refers to the scheduled travel of a bus on a route, which differs from our use of the term *trip* in this paper to refer to one leg of travel by a rider.

The procedure we used to identify and geolocate bus trips is detailed in Table 1. Once these steps were completed, we had a data set consisting of 60,466 ground truth O-D bus or light rail trips in the Pittsburgh region including origin stop ID and location, destination stop ID and location, trace point locations, and trace point timestamps. Since there is no guarantee that a timestamped trace point will exist in close proximity to the origin or destination, origin and destination timestamps are not included in this data set at this point in the analysis.

4.4 Identifying Transfers from Trips

After bus trips were fully identified, we created a data set of pairs of consecutive trips by the same device and developed a method to determine whether these trip pairs were transfers. We first took the 60,466 trips found after data cleaning and filtered out instances in which a device traced only one trip, removing 7.96% of trips and leaving 55,651 trips from devices that traced multiple trips. We then paired these into 50,834 pairs of consecutive trips by the same device. We then filtered out trip pairs if the first trace timestamps for each trip (usually close to the trip origins) were more than 3 hours apart. Three hours was chosen to be consistent with the Port Authority's fare rules for transfers, which allow a transfer fare if payment for the second ride is made within 3 hours of when payment was made for the first trip, and as a generous time allowance in case a long bus ride may have been followed by a long transfer. Since the transfer is timed (for purposes of applying the 3-hour cutoff) from the beginning of one trip to the beginning of the next trip, it includes the duration of one of the bus trips in addition to the time spent walking between stops and waiting for the next bus. Some Pittsburgh buses have headways of 60 minutes and a long bus ride could easily add another 60 minutes or more. (A note about Port Authority rules: because the Port Authority system has zone-based fares, payment is made at the beginning of most trips but at the end of outbound trips from downtown at certain times of day, though this policy is unevenly enforced. For the purposes of the fare policy, the transfer time might sometimes include both legs of the transfer and the time between legs. This would happen if a rider took a pay-on inbound trip, transferred downtown, and then took an outbound trip during the pay-off time window. For this study we used the 3-hour time cutoff but calculated from the beginning of every trip.)

Table 1. Procedure for cleaning trace point data and converting trace points to trips with ground truth origin-destination data.

Step	Description	Result
1.	We extracted data for trace points from the Tiramisu database, including associated GTFS data.	862,564 trace points
2.	We filtered out trace points without valid destinations selected (origin must be selected to log a trace point).	776,880 trace points (9.93% removed)
3.	We filtered out trace points with no location data (latitude = 0).	774,587 trace points (0.30% removed)
4.	For each trace point we looked up the origin and destination stops in the GTFS stop file, added the stop location coordinates to the trace point, and filtered out trace points for which the given origin or destination stop ID could not be found in the GTFS stop file.	774,162 trace points (0.055% removed)
5.	For each trace point we looked up the route direction (0 =outbound, 1 = inbound) in the GTFS trip file, added the route direction to the trace point data, and filtered out trace points for which the GTFS trip ID logged in the trace data could not be found in the associated GTFS trip file.	763,900 trace points (1.33% removed)
6.	We included only trace points logged while the app was actively tracing by filtering out trace points for which the "source" is not "recording".	380,622 trace points (50.2% removed)
7.	We combined multiple consecutive trace points with the same device, origin, destination, route, and direction into one trip and logged the location and timestamp of each trace that makes up that trip as well as the agency ID, origin ID, destination ID, origin coordinates, destination coordinates, route ID, route long and short names, and GTFS trip ID from the first trace point, since those should be the same for all trace points that make up one trip.	60,466 trips

We created maps with vector and point overlays showing the origins, destinations, and trace point locations for each pair of trips. We performed a visual inspection of the maps to manually label whether each pair of trips was a transfer or some other phenomenon. Some automatic labeling was also used to remove round trips and turnarounds from the data set if they were identical to previously labeled round trips or turnarounds. The labels used were:

- Transfer: including transfers with potentially intentional time gaps
- Round trip: similar starting and ending locations with opposite travel directions

- Single trip or turnaround: for cases where the trip identification failed to combine trace points that were actually part of a single trip, such as when the bus reaches a turnaround at the end of a route and tracing stops automatically but the user restarts tracing
- Other situations: such as trips that are very short and very close together (sometimes within one minute), likely indicating that someone was testing out the app; or trips that are clearly much further apart than walking distance, indicating that there was likely another leg of non-pedestrian travel between them

Due to time restrictions on manual labeling, we were able to label only 3198 trip pairs, or 44% of the available sample. Of the labelled trip pairs, 915 (28.6%) were found to be transfers, 767 (24.0%) were round trips, 1077 (33.7%) were actually single trips or bus turnarounds, and 439 (13.7%) represented some other situation, with many of those being disparate travel legs that probably had another leg of non-pedestrian travel between them.

4.5 Sample Selection for Case Study of the Pittsburgh Region

Since Tiramisu assigns each device a random ID number, and since we labeled the data in order by device ID, using the 44% of the data that we were able to label (given time and resource constraints) was equivalent to taking a random sample of available device data. For the purpose of this study we assume that a unique device ID is equivalent to a unique user, which is the conservative approach for this paper's analyses. We labeled 3198 trip pairs which included 915 transfers from 237 devices.

We further examined the transfers in the sample using visual map inspection to remove problematic data from the sample, including transfers where the origin or destination stop provided by the user appeared to be different from where they actually traveled given knowledge of the region and their trace point locations, so that we were left with only ground truth data for actual O-D pairs involved in transfers. This reduced the sample of transfers by 25.8% to 679 transfers found from 211 unique devices.

4.6 Calculating Transfer Characteristics

4.6.1 Transfer Distances

Origin and destination stop location coordinates were added to the data during the trip identification process. These were used to calculate both a simple straight-line estimate of transfer walking according to the haversine formula (Veness, 2016), which represents the best case scenario and is shown for comparison, and an adjusted walking distance with a correction factor of 2^.5 to represent a walking path that goes around blocks or intersections rather than cutting through them diagonally, which is a more conservative estimate and likely more accurate. Computing walking distances based on walking route maps is beyond the scope of this study.

4.6.2 Transfer Durations

To find the scheduled duration of each transfer, we looked up the scheduled destination stop time for the first trip and the scheduled origin stop time for the second trip in the GTFS trip data and subtracted those times. At this time, we filtered out any transfer for which either of these stops could not be found in the corresponding GTFS schedule (stop_time) data, most likely indicating some mismatch between GTFS versions that were in use at the time the data was collected. This reduced the sample by 48.5% to 350 transfers from 145 devices. (To improve scalability of the labeling procedure in future work, GTFS scheduled stop time data should be added to each trip earlier in the process to avoid manually labelling trip pairs that turn out to be ineligible for transfer duration calculations. To improve overall scalability of this data collection method, careful attention will be needed in matching GTFS data with the crowdsourced data during collection.)

To estimate the real-time duration of each transfer, we needed to estimate the real-time arrival of the bus at the first trip's destination and the real-time departure of the bus from the second trip's origin. Tiramisu periodically logs timestamps and locations, which become trace points, but does not specifically log timestamps at stops. There may also be some delay between when users board a bus and when they start tracing, as well as between when they stop tracing and when they disembark. AVL data was not available from the Port Authority for most of this time period. The estimation procedure was:

- To estimate the destination arrival time for the first trip, we used the last (up to) three trace points of that trip (or fewer than three if three were not available). To estimate the origin departure time for the second trip, we used the first (up to) three trace points of that trip.
- We looked up the stops along each scheduled GTFS trip and selected the stop in closest proximity to each trace point (using the coordinates and the haversine function), then looked up the scheduled stop time at that stop.
- We subtracted to find the time offset between the schedule at the closest stop and the trace point, resulting in an approximation of how early or late the bus was running. This method should be fairly accurate in places where bus stops are only a few minutes apart.
- We averaged the time offset from those (up to) three trace points to find an estimated time offset at the origin/destination stop of interest and adjusted the scheduled arrival/departure time by that amount to estimate the real-time arrival/departure.
- We subtracted the first trip destination arrival time from the second trip origin departure time to find the estimated real-time transfer duration.
- All negative estimated transfer durations were changed to zero. All traced transfers occurred successfully, therefore a negative time estimate, which would have made the transfer impossible, must be incorrect. Negative estimated transfer durations indicate a poor estimate of real-time bus locations, due to either insufficient data (no trace points sufficiently close to the origin/destination) or erroneous data.

4.6.3 Walking and Waiting Durations

We calculated the amount of time during each transfer that was spent walking based on an assumed walking speed of 3.0 mph along the adjusted distance between stops. If the transfer did not have sufficient time available for walking, we used the transfer duration as the walking time, though this may indicate that the rider actually ran between buses.

4.6.4 Removing Outliers

After calculating the transfer distance and scheduled and actual duration, some outliers in the data became apparent. These may indicate anomalies in the data such as user errors in selecting which GTFS trip to trace or other GTFS errors. We removed any transfer that had (1) a scheduled transfer duration more than 3 standard deviations from the mean, (2) an estimated trip time offset more than 2 standard deviations from the mean, or (3) a straight-line distance more than 3 standard deviations from the mean. This reduced the sample by 3.4% to 338 transfers from 142 devices. Some statistics describing transfer characteristics before and after removing outliers are shown in Table 2.

4.6.5 Transit-Relevant Violent Crime Statistics

We combined the annual reported incidences per 100 people of murder, rape, robbery, and aggravated assault to calculate the relative risk of violent crime for riders transferring or waiting for a bus in each neighborhood/sector. We assumed that burglary, auto thefts, and drug violations would be less relevant to pedestrians or people waiting at bus stops. We cross-referenced the crime data with the locations of the stops involved in transfers and with the waiting times at those stops.

Table 2. Descriptive statistics of transfer characteristics before and after removing outliers.

Metric	Before Removing Outliers		After Removing Outliers	
	Mean	Standard Deviation	Mean	Standard Deviation
Scheduled transfer duration	23.2 minutes	48.8 minutes	23.6 minutes	33.8 minutes
Estimated real-time transfer duration	24.1 minutes	33.9 minutes	23.0 minutes	31.7 minutes
Time offset	6.51 minutes	27.4 minutes	6.44 minutes	9.72 minutes
Straight-line transfer distance	0.42 km (0.26 mi.)	1.47 km (0.91 mi.)	0.26 km (0.16 mi.)	0.31 km (0.19 mi.)
Adjusted transfer distance	0.59 km (0.36 mi.)	2.08 km (1.29 mi.)	0.37 km (0.23 mi.)	0.44 km (0.27 mi.)

5 RESULTS AND DISCUSSION

We found that crowdsourced bus ridership data can be used to identify where and when poor transfers are occurring and to characterize those transfers for planning purposes. This has the advantage of providing continuous data collection over the study period and being more scalable than on the ground survey methods.

5.1 Transfer Sample Characteristics

We identified and characterized a sample of 338 transfers in the Pittsburgh region from 142 unique devices spanning the time period of May 19, 2012 to Jun 25, 2013. Some characteristics of these transfers are shown in Figure 1. Figure 1 (a) and (b) show the number of transfers on each day of the week and beginning during each hour of the day, respectively. Figure 1 (c) and (d) show the total transfer duration on each day of the week and during each hour of the day, respectively, with transfers that span multiple hours of the day allocated accordingly. Figure 1 (e) and (f) show boxplots of the distributions of transfer durations on each day of the week and beginning during each hour of the day, respectively, with the mean durations indicated by diamond markers. Due to the lack of buses running in the very early morning hours, we used 3am as the cutoff for the end of the day instead of midnight. The primary effect of this decision was to include transfers that occurred after midnight on Friday as part of the Friday service day instead of allocating them to Saturday.

As shown in Figure 1, the sample contains transfers on every day of the week. Monday had the most transfers and also the shortest mean transfer duration. Monday, Wednesday, Friday, and Saturday had higher total transfer durations then Tuesday, Thursday, and Sunday. The number and total duration of transfers peaked during the evening rush hour and there were no transfers in the middle of the night (when no buses were scheduled). The 75th percentile of transfer durations were longest around lunchtime and on Fridays and Saturdays. Determining the reason for this requires

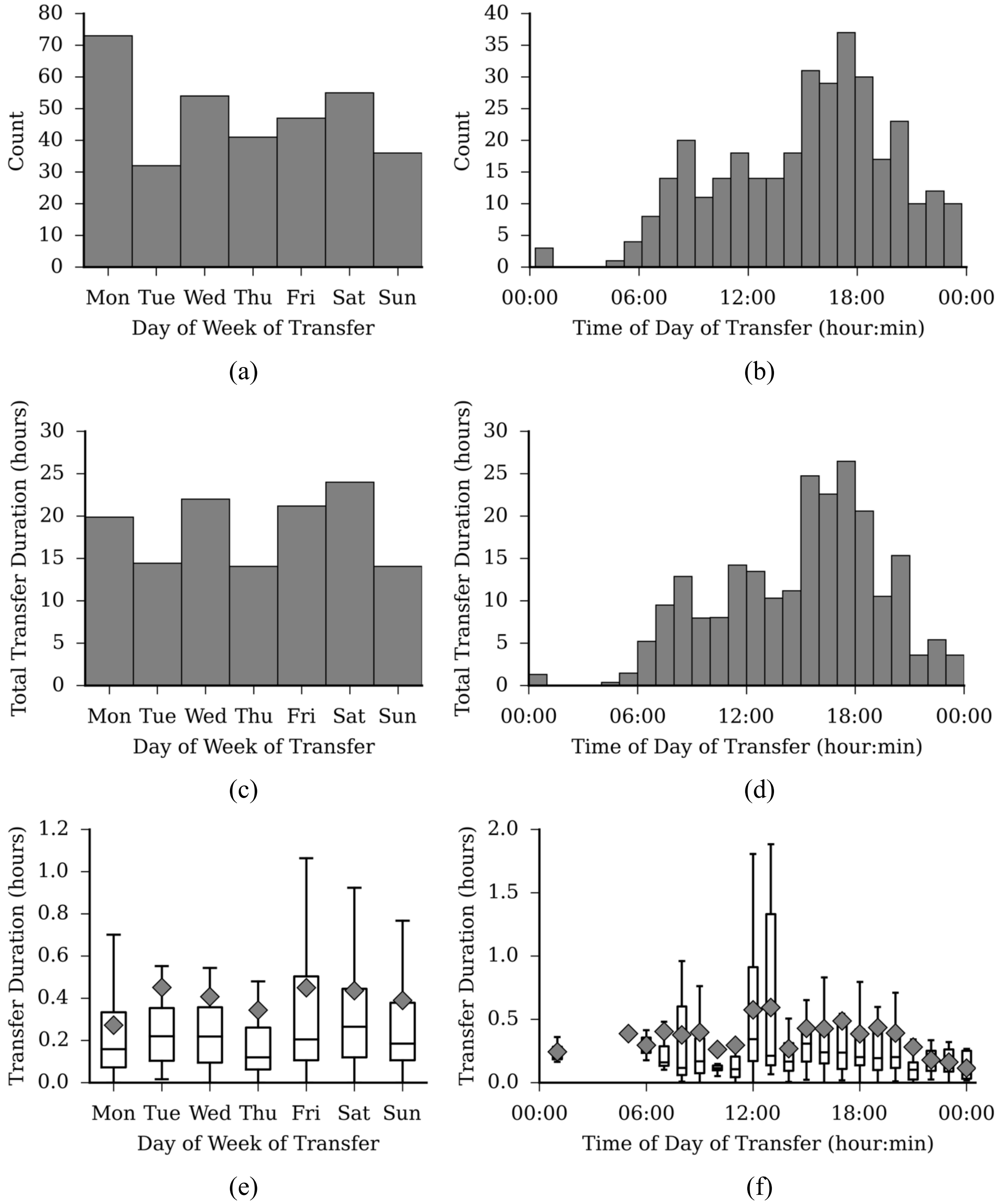
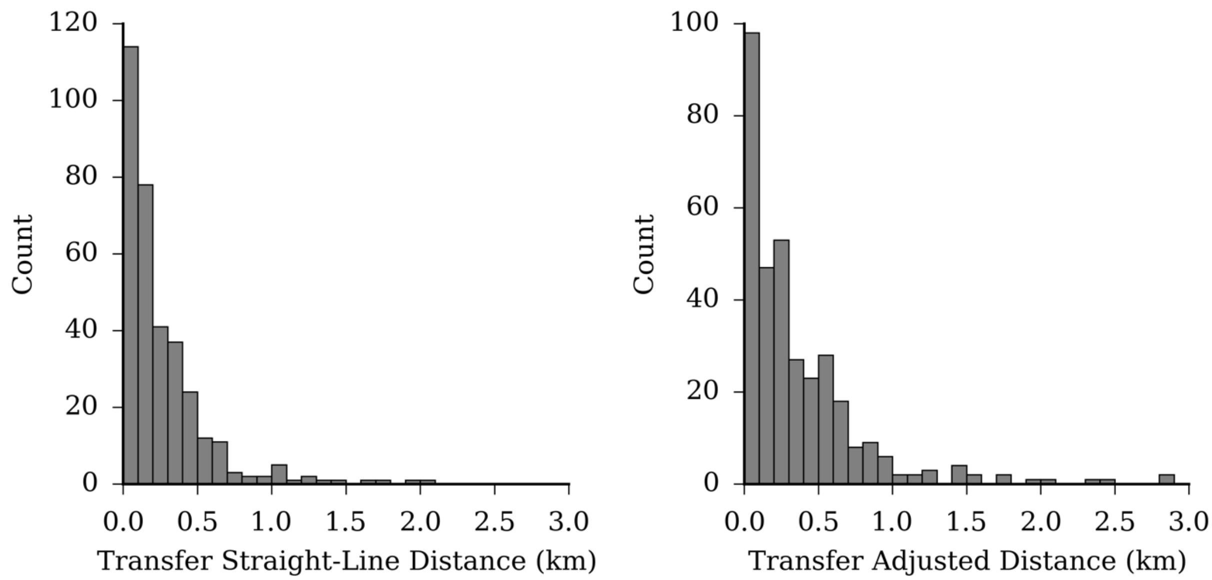


Figure 1. Characteristics of the case study sample of 338 transfers from the Pittsburgh region. Histograms show (a) number of transfers per day of week, (b) number of transfers per hour of the day, (c) total duration of transfers per day of week, and (d) total duration of transfers during each hour of the day. Box plots, with diamond markers at the mean, for quartile distribution of (e) transfer duration on each day of the week and (f) duration of transfers starting during each hour of the day, with whiskers extending to the last datum within 1.5 times the interquartile range. Due to the lack of buses running in the very early morning hours, we used 3am as the cutoff for the end of the (service) day instead of midnight; transfers occurring between midnight and 3am are allocated to the previous day of the week.

further study, but possibilities include poor bus routing or scheduling to reach lunch destinations or night life locations or riders deliberately extending a transfer by stopping for lunch.

5.2 Transfer Distance

Figure 2 shows the distribution of straight-line and adjusted transfer distances. As discussed in the methods section, the adjusted distances are likely more accurate and the straight-line distances are shown for comparison. 79.9% of the transfers were within a straight-line walking distance of 0.4 km (0.25 mi.) and 66.6% were within an adjusted walking distance of 0.4 km (0.25 mi.). The mean straight-line transfer distance was 0.26 km (0.16 mi.) and the standard deviation was 0.26 km (0.16 mi.). The mean adjusted transfer distance was 0.37 km (0.23 mi.) and the standard deviation was 0.44 km (0.27 mi.). Walking distances longer than 0.4 km (0.25 mi) (i.e., after the first four columns of Figure 2) may be poor transfers or may indicate that the rider deliberately detoured, perhaps to run an errand.



(a) (b)

Figure 2. Histogram of (a) straight-line and (b) adjusted transfer distances showing that the majority of trips are within walking distance.

5.3 Transfer Duration, Walking Time, and Waiting Time

Figure 3 shows the estimated time offsets between real-time and schedule for all trips involved in the sample of transfers. The mean time offset was a short delay of 6.44 minutes with a standard deviation of 9.70 minutes. As shown, there is a much larger tail of late buses than early buses. Some of the buses that were estimated to be 20 or more minutes early or late may indicate that the rider mistakenly chose to trace an earlier or later scheduled GTFS trip than the one they actually rode (which is especially likely if the bus was early or late by half or more of the headway time).

Figure 4 shows a comparison of the scheduled and estimated real-time transfer durations. The mean scheduled transfer duration was 23.6 minutes with a standard deviation of 33.8 minutes and the mean estimated real-time transfer duration was 23.0 minutes with a standard deviation of 31.7 minutes. 44.1% of the transfers were estimated to be less than 10 minutes. For comparison, Jang found that in Seoul 80% of transfers were less than 10 minutes, but maximum transfer times were 30 minutes on that system (Jang, 2010). Chu and Chapleau estimated (using an inference-based model) that approximately 65% of transfers in the National Capital Region of Canada were less than 10 minutes, perhaps due to coordinated transfers, on a system with headways similar to Pittsburgh (Chu and Chapleau, 2008). Pittsburgh does not have timed transfers, and so may be expected to have longer transfers than those two systems. The Tiramisu data may also contain a self-selection bias towards tracing more in frustrating situations, such as long transfers.

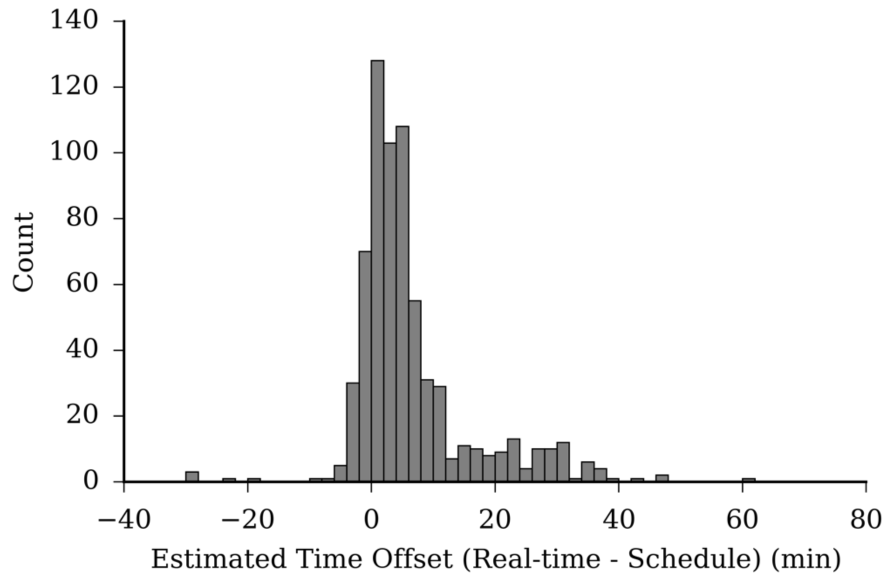
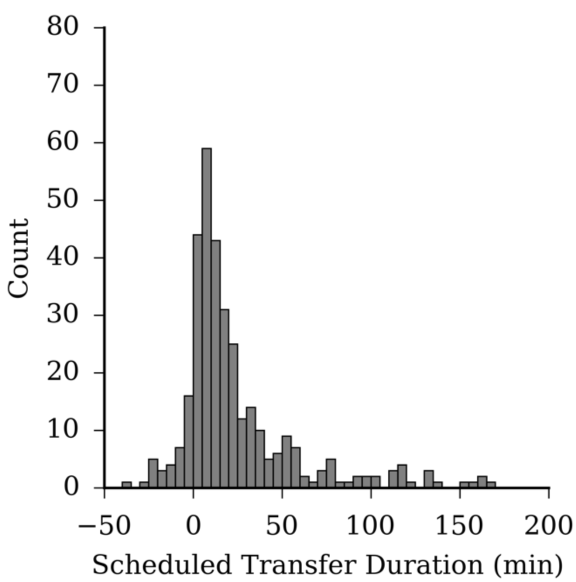
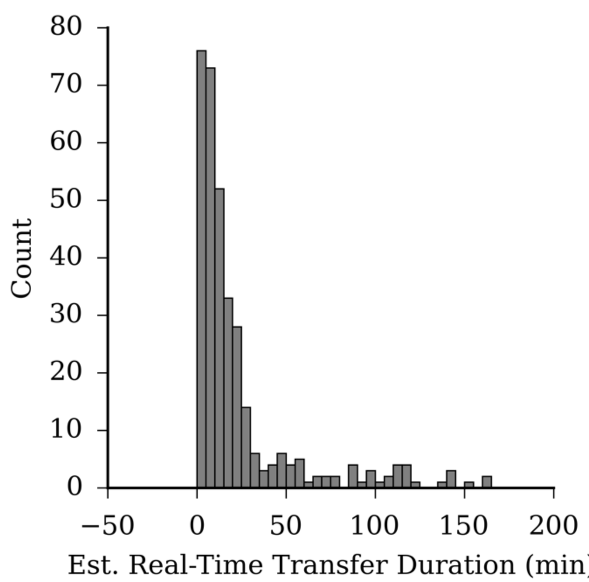


Figure 3. Histogram of estimated time offsets between schedule and real-time. Positive times indicate delayed buses and negative times indicate early buses.



(a)



(b)

Figure 4. Histograms of (a) scheduled transfer durations and (b) estimated real-time transfer durations.

Negative scheduled durations indicate that, according to the schedule, the transfer was not possible, but it was made possible in real-time, likely due to delays. Negative scheduled durations could also indicate that an incorrect scheduled GTFS bus trip was selected by the rider, likely due to the bus being either very early or very late. It is likely that estimated transfers with zero duration indicate real-time good transfers with very short waiting times. Very short estimated durations may also indicate bad transfers that required running instead of walking, a situation that is inconvenient for everyone and impossible for some, especially those with some types of disabilities. Disambiguating tight walking times from transfers that involved running may be aided by a sensitivity analysis on transfer times, but would ideally involve more precise real-time data, such as may be obtained from AVL. Transfer durations longer than the scheduled gap between buses may be poor transfers due to missed or delayed buses or may indicate that the rider had an intentional gap between buses, perhaps to run an errand. For most Port Authority routes, the scheduled headway within a route is 60 minutes or less.

Figure 5 shows histograms of time spent walking between stops (at 3.0 mph along the adjusted distance) and excess time spent waiting. As shown, walking times for all transfers were less than 25 minutes but waiting time has a long tail. Some of the longer waiting times beyond 60 minutes likely represent intentional gaps, though further analysis is needed to verify that.

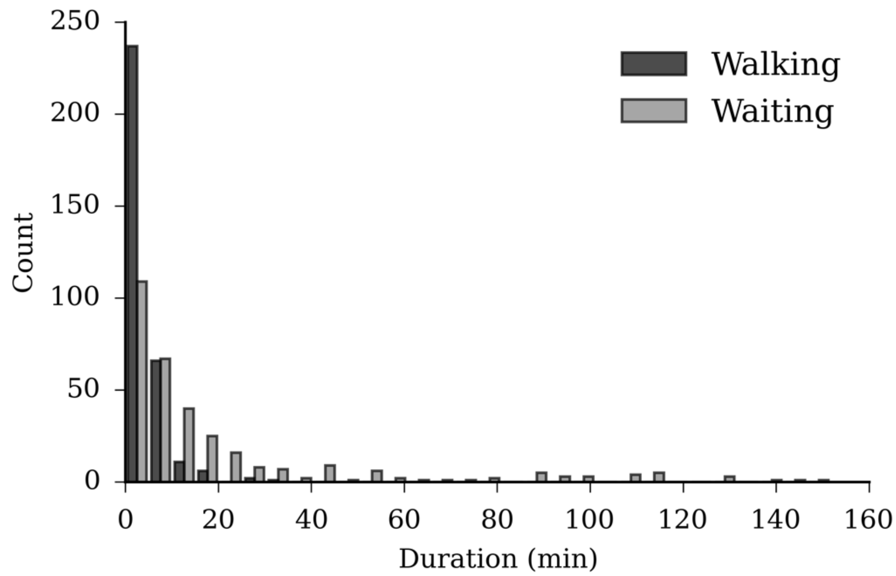


Figure 5. Histogram of transfer time spent walking and transfer time spent waiting. Transfers with negative estimated durations were excluded from this analysis.

5.4 Geographical Distribution and Crime Analysis

Figure 6 shows the geographical distribution of transfers in the form of a heatmap of the stop locations involved in those transfers. As shown, most of the transfers took place in downtown Pittsburgh, which is also where all of the recommended transfer stops are specified in the Port Authority's GTFS data. However, there are also several hot spots for transfers elsewhere, including several significant ones in the neighborhoods to the east (boxes C3 and C4). Since these have not been identified as recommended transfer points, they may not have received as much planning attention.

Figure 6 also shows the rate of violent crimes per 100 people per year in the neighborhoods for which Pittsburgh provides data. As shown in box C2 on the map, the crime level downtown near most of the transfers is moderate. Downtown is also the neighborhood with all of the recommended transfer stops. Some of the popular transfer points to the east of the city (in box C5 and the top of box C4) and to the south (bottom left of C3) occur in or on the edge of neighborhoods with higher crime. Since actual and perceived levels of personal safety are important factors in customer satisfaction with public transit, this should be taken into consideration in planning.

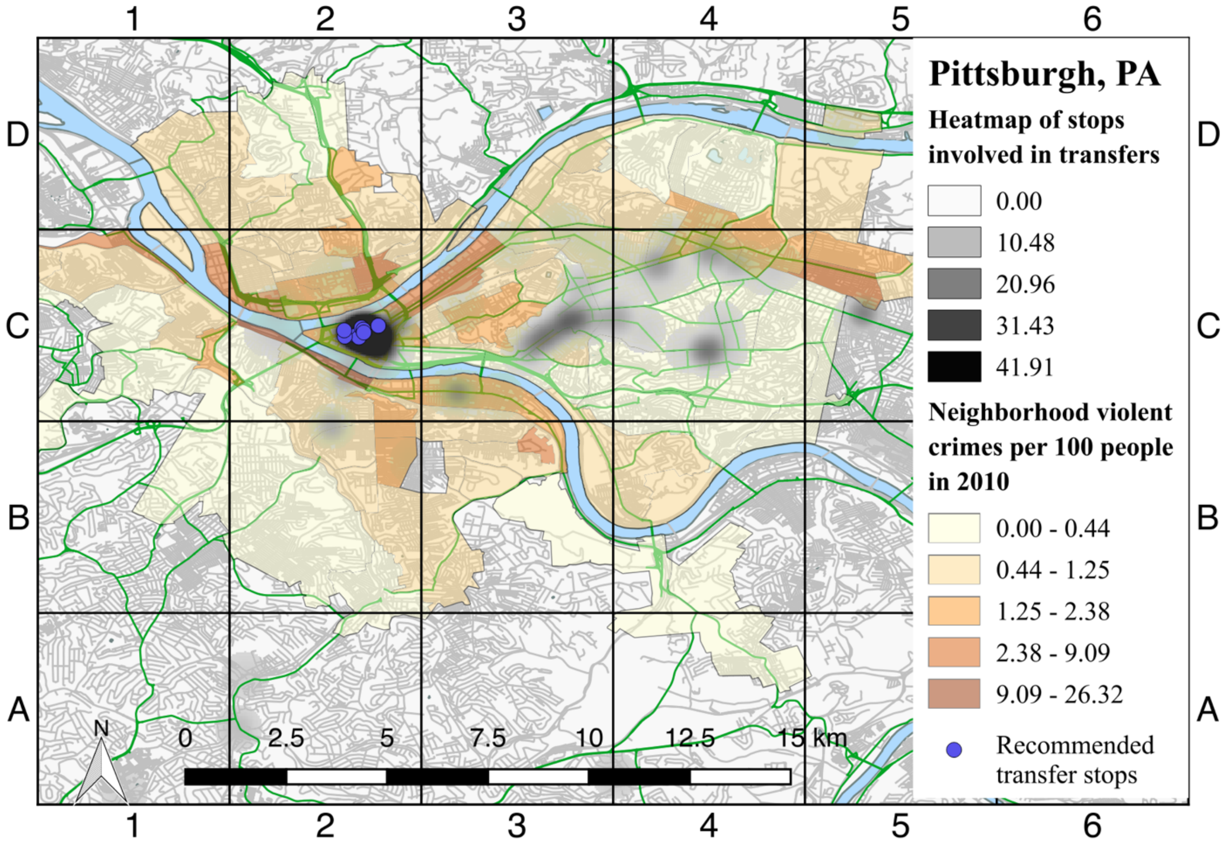
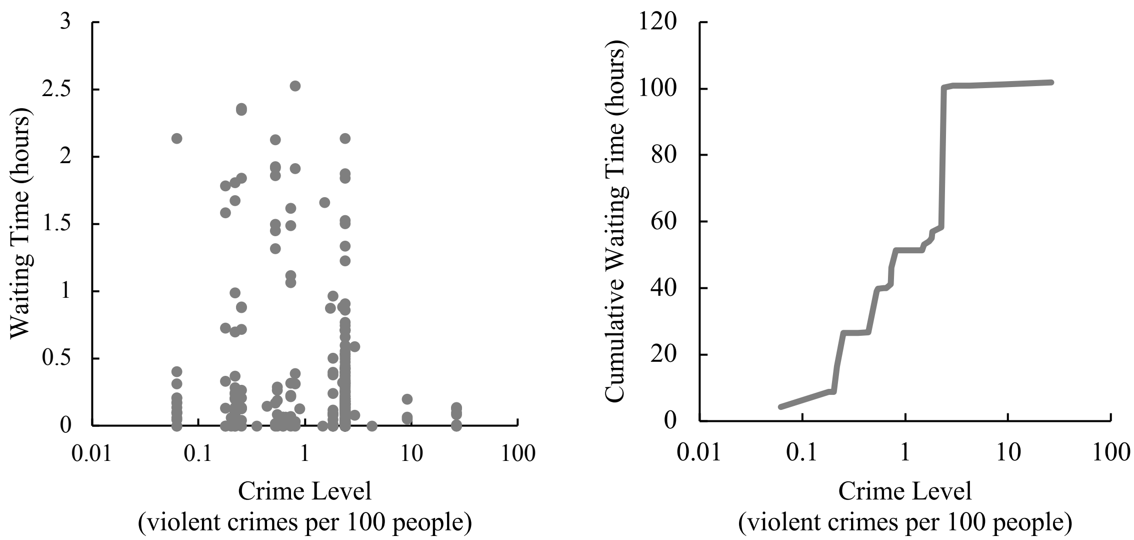


Figure 6. Heatmap of stops involved in transfers overlaid with recommended transfer stop locations and crime data on a map of Pittsburgh, PA. The heatmap legend shows colors corresponding to kernel densities of transfers.

Figure 7 shows waiting time at transfer stops and crime rates in the neighborhood of those stops. As shown, 50% of waiting time is spent at transfer stops with crime rates below 0.8 crime per 100 people per year. 99% of waiting time is spent at crime rates below 3 crimes per 100 people per year. The stops with the highest crime levels also have low waiting times, perhaps not coincidentally. Planning attention may be best spent at the stops that have both a (relatively) high crime level and many long waiting times, such as those shown in the range of 1.5-3 crimes per 100 people per year in Figure 7 (a).

5.5 Transfer Stop Infrastructure

The results from this study can be compared with the current state of infrastructure at the stops involved in transfers to evaluate which stops may be best suited for upgrades to increase comfort, accessibility, and safety during transfers and overall satisfaction with public transit. Further work is needed to identify infrastructure characteristics at all stops, but in August 2016 a visual inspection was performed of one of the clusters of stops involved in a high number of transfers in box C3 of Figure 6, along Fifth Avenue in the North Oakland neighborhood from Tennyson to Craft. This inspection verified that 6 outbound stops and 4 inbound stops along this corridor had no infrastructure except a non-electronic route sign. Street and sidewalk lighting were present but were not verified by the authors to provide good nighttime coverage at all stops. Some stops additionally had a shelter with a bench: Fifth and Tennyson inbound, Fifth and Bigelow inbound and outbound, and Fifth and Chesterfield inbound. Potential infrastructure improvements to better accommodate transfers at these stops could include shelters and benches at the stops that lack them, additional shelters and benches at the more popular stops to accommodate larger numbers of passengers, and electronic signs with real-time arrival information (Ji et al., 2017). Interestingly, the recently proposed bus rapid transit plan for Pittsburgh includes stops near these underdeveloped transfer locations (Port Authority of Allegheny County, 2017).



(a) (b)

Figure 7. Charts showing crime levels and waiting time at the second stop involved in each transfer as a scatter plot (a) and as a cumulative distribution (b). Note that the x-axes are on a log scale to better show variations at the low end of the scale.

6 FUTURE DIRECTIONS, GENERALIZABILITY, AND LIMITATIONS

6.1 Future Directions

This work lends itself well as a basis for further analysis. One possible extension is developing methods for differentiating between intentional gaps and poor or missed transfers to determine where and when the worst transfers take place. Another is to differentiate between fast connections that are walkable and ones that require running, indicating that some riders would not be able to make those transfers.

This work is also a useful basis for transit planning. The results shown in Figure 6 and Figure 7 are an example of how these methods can be used to pinpoint stops for further evaluation. Stops can be identified as frequently utilized transfer stops, as stops with long wait times, and/or as stops with relatively high crime rates. Some of these stops can be targeted for evaluation of existing infrastructure and potential improvements to facilitate comfortable and safe waits. Changes could also be made in route planning to better accommodate those riders, either by eliminating the need to transfer or by reducing wait times. Evaluating when and where long transfers take place during inclement weather and in the dark would also be of interest, as these characteristics may influence desired infrastructure accommodations.

6.2 Generalizability and Scalability

This method based on crowdsourced data has the potential to be more scalable and generalizable than traditional surveys. Once the crowdsourcing platform has been developed and debugged, it can be deployed in new locations with much less effort than the first deployment. The only requirement for integration with a local transit system is that GTFS data should be available, which is widely true. Care should be taken that both the crowdsourcing data and the original GTFS data are stored and matched with each other. Depending on the method used to incentivize crowdsourcing users, there may be varying amounts of self-selection, both of users and of when to trace, but when the primary incentive is to provide information to the transit agency and to other users, this may trend towards more tracing in more frustrating situations, which may be the most useful data depending on the purpose of the study. Transit surveys, even when conducted online, are typically fielded with personnel at bus stops or on transit vehicles, which is unnecessary with crowdsourcing, and crowdsourced data collection can take place continuously whenever riders are using transit without being dependent on timing of survey personnel shifts or limited to specific survey periods.

Manual data labeling to identify transfers was a significant bottleneck in this study. However, with more development time and resources, further automation should be possible (for example using travel speed-based methods as discussed in Jang (Jang, 2010)), with the potential for full automation of the labeling process leading to fully scalable data processing.

6.3 Limitations

There were several limitations to the scope of our case study related to data availability. One limitation is that AVL data was not available during the time period of the study. AVL data for the real-time location of buses would have increased the accuracy of calculating the transfer durations and would facilitate both differentiating missed transfers from intentional gaps in travel and differentiating tight walking transfers from transfers that required running.

Another limitation was the small sample size. This was partially due to the limitation of manual data labeling, which could be further automated in future work, and partially due to a problem with some versions of Tiramisu app causing a large amount of the trace data to have errors in matching with the GTFS data, rendering it unusable without significantly more processing to fix these errors.

Another data limitation is that when users traced trips in the Tiramisu app, they needed to choose not only the route they were taking but also the scheduled GTFS trip. When buses were off-schedule or bunched, it could be hard to tell which GTFS trip to select (is the bus running 5 minutes early or 5 minutes late?). Therefore, some of the trace data may be associated with an incorrect scheduled GTFS trip. This could be partially corrected with error detection after-the-fact, such as by assuming that a bus is more likely to be somewhat late than to be very early, by matching AVL data (if known) with the device's location, or by tracking schedule drift over time (if enough data is available) to determine whether a particular bus is early or late.

Since we did not collect demographic data for Tiramisu users in this data set, we were not able to evaluate how representative our sample was of typical transit riders. Additionally, since Tiramisu users chose when to trace, there may be some selection bias not only in selection of users but in selection of which trips to trace, and the motivation to provide real-time bus arrival information may have prompted more tracing for buses that were off-schedule.

A limitation in our methodology is the use of straight-line walking distances between bus stops with an adjustment factor to account for walking around blocks and intersections. In future work these could be replaced with route-based walking distances, for example using the Google Maps API.

7 CONCLUSIONS

We used crowdsourcing to identify and characterize a sample of 338 transfers in the Pittsburgh region from 142 unique devices spanning the time period of May 19, 2012 to Jun 25, 2013. We obtained O-D ground truth stop information for a system that did not have that data available from fare cards. We found that 66.6% of the transfers were within an adjusted walking distance of 0.4 km (0.25 mi.) and 44.1% of the transfers were less than 10 minutes. The mean scheduled transfer duration was 23.6 minutes, the mean estimated real-time transfer duration was 23.0 minutes, the mean time offset for the individual trips was a delay of 6.44 minutes, and the mean adjusted transfer distance was 0.37 km (0.23 mi.). We identified the geographical distribution of transfers and found several highly-utilized transfer locations that were not identified by the Port Authority of Allegheny County as recommended transfer points, and so might need more planning attention. We cross-referenced transfer location and wait time data with crime levels to provide additional planning insight, finding that 50% of waiting time is spent at transfer stops with crime rates below 0.8 crime per 100 people per year and 99% of waiting time is spent at crime rates below 3 crimes per 100 people per year. The resulting information resource is provided for use during planning and identification of locations where policing, additional bus shelters, and other infrastructure improvements may facilitate more safe and comfortable waits or where routing improvements could be made. We have shown that crowdsourced public transit ridership data, which is more scalable than conducting traditional surveys, can be used to analyze transit transfers, although it likely contains self-selection bias.

8 ACKNOWLEDGEMENTS

For their assistance with data for this study, we thank Alex Haig and M. Zubair Malik. We would also like to thank the Port Authority of Allegheny County for their insights on transit planning. The contents of this paper were partly developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90RE5011). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). This work was also supported by the Technologies for Safe and Efficient Transportation University Transportation Center (T-SET UTC), which is funded by the U.S. Department of Transportation (US DOT grant number DTRT-13-GUTC-26). The contents of this work do not necessarily represent the policy of the Federal Government, and you should not assume endorsement by the Federal Government.

9 REFERENCES

Abkowitz, M., Josef, R., Tozzi, J., Driscoll, M.K., 1987. Operational Feasibility of Timed Transfer in Transit Systems. Journal of Transportation Engineering 113, 168–177. https://doi.org/10.1061/(ASCE)0733-947X(1987)113:2(168)

Benenson, I., Ben-Elia, E., Rofé, Y., Geyzersky, D., 2016. The benefits of a high-resolution analysis of transit accessibility. International Journal of Geographical Information Science 0, 1–24. https://doi.org/10.1080/13658816.2016.1191637

Cevallos, F., Zhao, F., 2006. Minimizing Transfer Times in Public Transit Network with Genetic Algorithm. Transportation Research Record: Journal of the Transportation Research Board 1971, 74–79. https://doi.org/10.3141/1971-11

Chu, K.A., Chapleau, R., 2008. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. Transportation Research Record: Journal of the Transportation Research Board 2063, 63–72. https://doi.org/10.3141/2063-08

Chung, E.-H., Shalaby, A., 2007. Development of Control Strategy for Intermodal Connection Protection of Timed-Transfer Transit Routes. Transportation Research Record: Journal of the Transportation Research Board 2006, 3–10. https://doi.org/10.3141/2006-01

Daganzo, C., Anderson, P., 2016. Coordinating Transit Transfers in Real Time (Caltrans Technical Report No. 2016-TO 021-65A0529), UCCONNECT. University of California Berkeley, Berkeley, CA.

Fan, Y., Guthrie, A., Levinson, D., 2016. Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security. Transportation Research Part A: Policy and Practice 88, 251–264. https://doi.org/10.1016/j.tra.2016.04.012

Ferris, B., Watkins, K., Borning, A., 2010. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit, in: Proceedings of CHI 2010. Atlanta, GA, USA.

Gordon, J., Koutsopoulos, H., Wilson, N., Attanucci, J., 2013. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. Transportation Research Record: Journal of the Transportation Research Board 2343, 17–24. https://doi.org/10.3141/2343-03

Gschwender, A., Jara-Díaz, S., Bravo, C., 2016. Feeder-trunk or direct lines? Economies of density, transfer costs and transit structure in an urban context. Transportation Research Part A: Policy and Practice 88, 209–222. https://doi.org/10.1016/j.tra.2016.03.001

Guo, Z., Wilson, N., 2004. Assessment of the Transfer Penalty for Transit Trips Geographic Information System-Based Disaggregate Modeling Approach. Transportation Research Record: Journal of the Transportation Research Board 1872, 10–18. https://doi.org/10.3141/1872-02

Hanft, J., Iyer, S., Levine, B., Reddy, A., 2016. Transforming Bus Service Planning Using Integrated Electronic Data Sources at NYC Transit. Journal of Public Transportation 19. http://dx.doi.org/10.5038/2375-0901.19.2.6

Hofmann, M., O'Mahony, M., 2005. Transfer journey identification and analyses from electronic fare collection data, in: Proceedings of the 2005 IEEE Intelligent Transportation Systems, 2005. Presented at the 2005 IEEE Intelligent Transportation Systems, 2005., pp. 34–39. https://doi.org/10.1109/ITSC.2005.1520156

Iseki, H., Taylor, B., 2010. Style versus Service? An Analysis of User Perceptions of Transit Stops and Stations. Journal of Public Transportation 13, 23–48. https://doi.org/10.5038/2375-0901.13.3.2

Jang, W., 2010. Travel Time and Transfer Analysis Using Transit Smart Card Data. Transportation Research Record: Journal of the Transportation Research Board 2144, 142–149. https://doi.org/10.3141/2144-16

Ji, Y., Zhang, R., Gao, L., Fan, Y., 2017. Perception of Transfer Waiting Time at Stops and Stations in Nanjing, China, in: Transportation Research Board 96th Annual Meeting Compendium of Papers. Presented at the Transportation Research Board 96th Annual Meeting, Washington, D.C.

Kieu, L.M., Bhaskar, A., Cools, M., Chung, E., 2017. An investigation of timed transfer coordination using event-based multi agent simulation. Transportation Research Part C: Emerging Technologies 81, 363–378. https://doi.org/10.1016/j.trc.2017.02.018

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies 36, 1–12. https://doi.org/10.1016/j.trc.2013.07.010

Michel, S., Chidlovskii, B., 2016. Stochastic optimization of public transport schedules to reduce transfer waiting times, in: Smart Cities Conference (ISC2), 2016 IEEE International. Presented at the Smart Cities Conference (ISC2), 2016 IEEE International, IEEE, Trento, Italy, pp. 1–6. https://doi.org/10.1109/ISC2.2016.7580860

Migurski, M., OpenStreetMap Contributors, 2013. Metro Extracts [WWW Document]. Metro Extracts. URL http://metro.teczno.com/ (accessed 4.27.16).

National Council on Disability, 2015. Transportation Update: Where We've Gone and What We've Learned. National Council on Disability.

Ngamchai, S., Lovell, D.J., 2003. Optimal Time Transfer in Bus Transit Route Network Design Using a Genetic Algorithm. Journal of Transportation Engineering 129, 510–521. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:5(510)

OpenStreetMap Contributors, 2016a. OpenStreetMap [WWW Document]. OpenStreetMap. URL http://www.openstreetmap.org/ (accessed 4.27.16).

OpenStreetMap Contributors, 2016b. Metro Extracts · Mapzen [WWW Document]. Mapzen. URL https://mapzen.com/data/metro-extracts/ (accessed 7.31.16).

Pinelli, F., Nair, R., Calabrese, F., Berlingerio, M., Lorenzo, G.D., Sbodio, M.L., 2016. Data-Driven Transit Network Design From Mobile Phone Trajectories. IEEE Transactions on Intelligent Transportation Systems 17, 1724–1733. https://doi.org/10.1109/TITS.2015.2496783

Pittsburgh City Planning, 2012. PGHSNAP: Pittsburgh's neighborhood data and map resource [WWW Document]. City of Pittsburgh. URL http://www.pittsburghpa.gov/dcp/snap/raw_data (accessed 5.24.16).

Port Authority of Allegheny County, 2017. Bus Rapid Transit [WWW Document]. Port Authority. URL http://www.portauthority.org/paac/CompanyInfoProjects/BRT.aspx (accessed 11.27.17).

Port Authority of Allegheny County, 2016. Developer Resources [WWW Document]. Port Authority. URL http://www.portauthority.org/paac/CompanyInfoProjects/DeveloperResources.aspx (accessed 7.31.16).

Rahman, S., Wong, J., Brakewood, C., 2016. Use of Mobile Ticketing Data to Estimate an Origin–Destination Matrix for New York City Ferry Service. Transportation Research Record: Journal of the Transportation Research Board 2544, 1–9. https://doi.org/10.3141/2544-01

Sánchez-Martínez, G.E., 2017. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data. Transportation Research Record: Journal of the Transportation Research Board 2652, 1–7. https://doi.org/10.3141/2652-01

Seaborn, C., Attanucci, J., Wilson, N., 2009. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. Transportation Research Record: Journal of the Transportation Research Board 2121, 55–62. https://doi.org/10.3141/2121-06

Shafahi, Y., Khani, A., 2010. A practical model for transfer optimization in a transit network: Model formulations and solutions. Transportation Research Part A: Policy and Practice 44, 377–389. https://doi.org/10.1016/j.tra.2010.03.007

Steinfeld, A., Zimmerman, J., Tomasic, A., Yoo, A., Aziz, R., 2012. Mobile transit rider information via universal design and crowdsourcing. Transportation Research Record - Journal of the Transportation Research Board 2217, 95–102. https://doi.org/10.3141/2217-12

Tomasic, A., Zimmerman, J., Garrod, C., Huang, Y., Nip, T., Steinfeld, A., 2015. The Performance of a Crowdsourced Transportation Information System, in: Transportation Research Board 94th Annual Meeting Compendium of Papers. Presented at the Transportation Research Board 94th Annual Meeting, Washington, D.C.

Tomasic, A., Zimmerman, J., Steinfeld, A., Huang, Y., 2014. Motivating Contribution in a Participatory Sensing System via Quid-pro-quo, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '14. ACM, New York, NY, USA, pp. 979–988. https://doi.org/10.1145/2531602.2531705

Veness, C., 2016. Calculate distance and bearing between two Latitude/Longitude points [WWW Document]. Movable Type Ltd: Information Design & Management. URL http://www.movable-type.co.uk/scripts/latlong.html (accessed 6.2.16).

Wang, W., Attanucci, J., Wilson, N., 2011. Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems. Journal of Public Transportation 14, 131–150. http://dx.doi.org/10.5038/2375-0901.14.4.7

Zhang, J., Chen, Z., Liu, Y., Du, M., Yang, W., Guo, L., 2017. Space–time visualization analysis of bus passenger big data in Beijing. Cluster Comput 1–13. https://doi.org/10.1007/s10586-017-0890-8

Zhao, J., Qu, Q., Zhang, F., Xu, C., Liu, S., 2017. Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. IEEE Transactions on Intelligent Transportation Systems PP, 1–12. https://doi.org/10.1109/TITS.2017.2679179