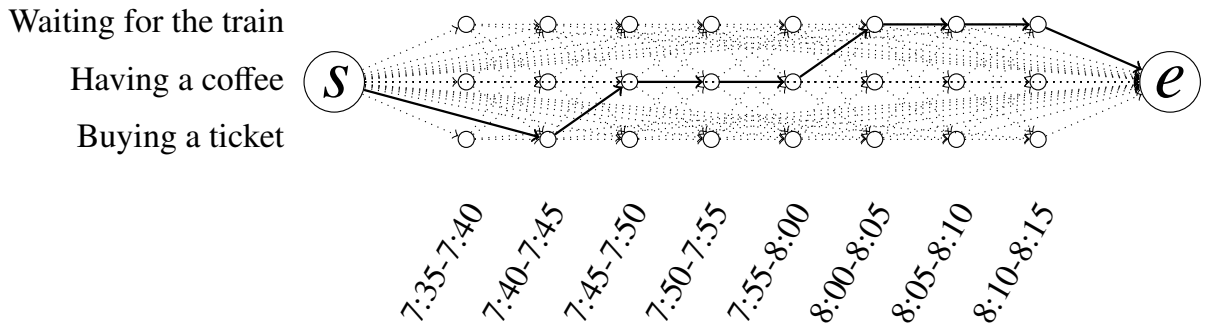

A path choice approach to activity modeling with a pedestrian case study

Antonin Danalet

Michel Bierlaire

EPFL

May 2014

STRC

14th Swiss Transport Research Conference

Monte Verità / Ascona, May 14-16, 2014

EPFL

A path choice approach to activity modeling with a pedestrian case study

Antonin Danalet, Michel Bierlaire
Transport and Mobility Laboratory,
School of Architecture, Civil and
Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

phone: +41 21 693 25 32
fax:
{antonin.danalet,michel.bierlaire}@epfl.ch

May 2014

Abstract

In Switzerland, the largest railway stations are called "RailCities" by the Swiss Railways. It emphasizes their transformation into places to perform different activities, similar to a small-scale city. Concerns are similar than in urban areas: costs of new infrastructure, traffic congestion, land scarcity.

The activity-based approach models the activity participation patterns. Traveling is seen as a derived demand from the need to pursue activities. In the pedestrian context, postulate rules, such as the home-based structure with tours from home, do not hold. The large dimensionality of the problem implies aggregation or hierarchy of dimensions, with priorities of activity types.

We develop a modeling framework based on path choice. The activity-episode sequence is seen as a path in an activity network. The sequence is not home-based nor tour-based. The model can be applied in different contexts, both urban and pedestrian. The large dimensionality is managed through an importance sampling based on Metropolis-Hastings algorithm for the generation of the choice set. The time is discretized. The utility of an activity-episode sequence is the sum of individual trips and activities, including the time-of-day preferences and the satiation effects.

First results of a case study on campus are presented, based on data from WiFi traces.

Keywords

Activity-based modeling, pedestrians, campus, train station

Contents

Notation

a_{ψ_i}	an activity episode, $a_{\psi_i} = (x, t^-, t^+)$, for individual i
$a_{1:\Psi_i} = (a_1, ..., a_{\Psi_i})$	an observed activity-episode sequence
$att(x, t)$	the attractivity for location $x \in POI$ at time t
A_ψ	an activity, $A_\psi = (\mathcal{A}_k, t^-, t^+)$
$A_{1:\Psi_i} = (A_1, ..., A_\psi, ..., A_{\Psi_i})$	an activity pattern with Ψ_i activities, indexed by ψ
$A(a_{\psi_i})$	a function $a_{\psi_i} \mapsto A(a_{\psi_i}) = \mathcal{A}_k \in \{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_K\}$
$\mathcal{A}_k$	an activity type, $k \in K$
$\mathcal{A}_{k,\tau}$	a node in the activity network, corresponding to activity type $\mathcal{A}_k$ and unit of time τ
$\mathcal{A}_{1:T}$	an activity path, i.e., a representation of an activity pattern $A_{1:\Psi_i}$ in an activity network
$b(\Gamma)$	the unnormalized target weights for path Γ
DDR	the domain of data relevance.
β	the parameters of the choice model
δ	a cost function
$\delta_v(v)$	the cost of node v
$\delta_\Gamma(\Gamma)$	the non-link-additive cost of path Γ
$\delta(\Gamma)$	the generalized cost of path Γ
e	the end node of an activity network
$\mathcal{E}$	the set of edges in $SERG$
f	the labeling function, $f : \mathcal{N} \to \tilde{\mathcal{L}}$, in $SERG$
g	a function associating nodes with coordinates in a coordinate system, $g : \mathcal{N} \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}$
Γ	a path in the activity network
i	an individual
I	the set of all individuals in the period of interest
j	the index of a measurement $\hat{m}_j$
J	the total number of measurements
k	the index of an activity type.
$k_{\Gamma n}$	the number of times activity path Γ is drawn
K	the number of activity types.
L	the number of different activity-episode sequences $a_{1:K_i}$ corresponding to observation i
$\mathcal{L}_i$	the set of all activity patterns corresponding to observation i
$\tilde{\mathcal{L}}$	a set of relevant labels for rooms in $SERG$

$\hat{m}$	a raw measurement, containing location $\hat{x}$ and timestamp $\hat{t}$
$\hat{m}_{1:J}$	a set of measurements $\hat{m}_j$
n	a node in SERG, $n \in \mathcal{N}$
$\mathcal{N}$	the set of all nodes in $SERG$
POI	the set of points of interest, $POI \in \mathcal{N}$
$\mathcal{P}_i$	the set of all candidate activity paths for observation i
$\mathcal{P}_I$	the set of all candidate activity paths for all individuals $i \in I$ in the period of interest.
$\mathcal{P}_{A_{1:\Psi_i}}$	the set of candidate paths corresponding to the activity pattern $A_{1:\Psi_i}$
ψ_i	the index of a activity episode a_{ψ_i}.
Ψ_i	the total number of episodes a_{ψ_i} in the activity-episode sequence $a_{1:\Psi_i}$ and the total number of activities A in the activity pattern $A_{1:\Psi_i}$. Ψ_i is individual specific.
$q(j)$	the sampling probability
s	the start node of an activity network
$S_{x,i}(t)$	the instantaneous potential attractivity measure in location $x \in POI$ at time t for individual i
$S_{x,i}(t^-, t^+)$	the potential attractivity measure in location $x \in POI$ between start time t^- and end time t^+ for individual i
$S_{\mathcal{A}_k,\tau}$	the potential attractivity measure for activity type $\mathcal{A}_k$ at time interval τ for individual i, corresponding to node $\mathcal{A}_{k,t}$
$sched_{x,i}(t)$	a dummy variable for time constraints in location $x \in POI$ at time t for individual i
$SERG$	a semantically-enriched routing graph, $SERG := (\mathcal{N}, \mathcal{E}, \mathcal{L}, f, g, POI)$
t	time, continuous
τ	a discrete unit of time in the activity network, $\tau \in 1, 2, ..., T$. τ can also be seen as a time interval between τ_{LB} and τ_{UB}
τ_{LB}	the lower bound of the time interval τ
τ_{UB}	the upper bound of the time interval τ
T	the total number of units of time τ
T_{min}	a minimum time threshold for activity episodes (typically 5 minutes in a pedestrian context)
$\hat{t}$	a timestamp of a raw measurement $\hat{m}$
t^-	the start time of an activity episode a, a continuous random variable

t^+	the end time of an activity episode a, a continuous random variable
$tt_{x_\psi, x_{\psi+1}}$	the travel time from x_ψ to $x_{\psi+1}$.
$\mathcal{U}$	the choice set corresponding to the activity network
v	a node in the activity path Γ, $v \in \Gamma$
$w(\Gamma)$	the target weight of Γ
x	the episode location, $x \in POI$
$\hat{x}$	the position of a raw measurement $\hat{m}$. $\hat{x} \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (x-y coordinates in a coordinate system, and floor or altitude in a multi-floor environment).

1 Introduction

Pedestrian infrastructures such as multimodal transportation hubs (airports, train stations), mass gathering (music festivals) or campuses are facing complex issues. Designing efficient buildings for pedestrian flows, managing congestion, locating new points of interests (ticket machines, shops or restaurants), modifying schedules to improve efficiency and guarantee connections (trains or flights) require pedestrian demand management strategies. These strategies allow to modify individual travel behavior, in particular in terms of activity sequences.

Activity-based modeling at the city scale has been motivated by its sensitivity to different policy issues and demand management measures including congestion pricing, toll lanes, and changing schedules for work or shops (Davidson *et al.,* 2007). While some of these measures, like parking pricing, are irrelevant for pedestrians, changing schedules to minimize the travel time or the flow, or specific design of the walking areas are efficient strategies for pedestrian facilities. Toll tunnel exist for pedestrian for a long time. In Antwerp, Sint-Anna tunnel was not free for pedestrians when it opened in 1933 (Razemon, 2013). More generally, the activity sequence and the location of the points of interest have a large impact on pedestrian flows in multimodal transportation hubs.

The activity-based approach models the interactions shaping the activity participation patterns. Traveling is seen as a derived demand from the need to pursue activities. Several models have been proposed. Activity scheduling model in an entire-day framework is a mix of rule-based algorithm, duration models and discrete choice structures. The biggest drawback of most of these models is the postulated rules: they are structured on home and tours from home, with models applied sequentially according to priorities of activity types. Very often, the large dimensionality of the problem (activity types, continuous time, number of episodes in the day) implies aggregation (broad periods of time, mandatory vs non mandatory, primary vs secondary) or hierarchy of dimensions.

Our modeling approach consider that pedestrians first choose the activity type, timing and duration in their activity patterns in pedestrian facilities. Only then, conditionally on the activity type and time of day, they choose their specific destination. In a railway station, a typical pedestrian chooses first that he needs a sandwich before taking his train, and then only chooses which of the sandwich shops he will visit.

We develop a modeling framework based on path choice. The activity-episode sequence is seen as a path in an activity network. The sequence is not home-based nor tour-based. The model can be applied in different contexts, both urban and pedestrian. The large dimensionality is managed

through an importance sampling based on Metropolis-Hastings algorithm for the generation of the choice set. The time is discretized in regular intervals. The utility of a activity-episode sequence is the sum of individual trips and activities, including the time-of-day preferences and the satiation effects.

In Section 2, a literature review briefly presents some concepts. Section 3 develops the methodology. Finally, in Section 4, first results of a case study on campus are presented, based on data from WiFi traces.

2 Literature review

2.1 Activity choice

Reviews about activity-based travel demand modeling can be found in Ettema (1996), Bhat and Koppelman (1999), Roorda (2005), Habib (2007), Bowman (2009), Feil (2010), Pinjari and Bhat (2011) and Miller (2014).

2.1.1 Location-aware technologies for activity modeling

One recent trend in activity-based travel demand modeling is the usage of data from location-aware technologies (Miller, 2014). Traditionally, data collection are revealed preferences about activity and travel patterns from diary surveys, where people describe a past day (Ettema, 1996). Standard revealed preference data usually avoid from collecting multiple answers from the same individual. GPS-based prompted recall activity-travel survey allows for longitudinal surveys, using GPS devices carried by respondents (Frignani *et al.,* 2010). Recall methods can be implemented on mobile devices (Rindfüser *et al.,* 2003).

Location-aware technologies help improving the quality of explicit surveying. They can also be used alone, from the communication infrastructure side, such as cell tower traces or WiFi access points traces (Bekhor *et al.,* 2013, Calabrese *et al.,* 2013), or from the individuals' devices (Etter *et al.,* 2012, Buisson, 2014). Etter *et al.* (2012) show that it is possible to predict up to 60% of next visited places from passive smartphone data.

Data preprocessing methods are needed to transform these raw observations into data adapted for modeling purpose. First, detection of stops points discriminates places where people spend time and perform activities from moving between these stop points (Rieser-Schüssler, 2012, Jiang

et al., 2013). After cleaning the data, Bekhor *et al.* (2013) define a destination as a cell tower where the device is connected for more than 20 minutes without changing. Using triangulation from cell towers, Calabrese *et al.* (2013) merge all measurements in a time interval ΔT with maximum distance $1km$. ΔT is not given. In Jiang *et al.* (2013), the first step is similar, with a distance of 300 meters and a stay time of 10 minutes. Here, the accuracy of the location is about 200 to 300 meters. The second step associates with each other the different stop points at different times if they are close using a grid-based clustering method. It allows to identify the places that are visited multiple times, despite measurement errors. Based on triangulation from WiFi access points, Danalet *et al.* (2014) associate all measurements to points of interest (POI) in the map, compute the travel time between these POI, define a distribution for arrival and departure time based on travel time and measurement timestamp, and finally remove destinations with expected duration smaller than $T_{min} = 5min$. From WiFi data from smartphones, Buisson (2014) clusters measurements using a Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm.

Pure location-aware technologies lacks the path semantics (Miller, 2014). To overcome this issue and detect activity purpose, localization data are merged with land-use information (Rieser-Schüssler, 2012, Miller, 2014, Danalet *et al.,* 2014). A review of research in this direction before 2012 is available in Miller (2014). Jiang *et al.* (2013) propose a visual example of how land use data could be applied, without a general methodology. In a pedestrian facility, Danalet *et al.* (2014) propose a Bayesian approach:

$$P(a_{1:\Psi_i}|\hat{m}_{1:J}) \propto P(\hat{m}_{1:J}|a_{1:\Psi_i}) \cdot P(a_{1:\Psi_i}) \tag{1}$$

where $a_{1:\Psi_i}$ is an activity-episode sequence for individual i, $\hat{m}_{1_J}$ a set of measurements, $P(\hat{m}_{1:J}|a_{1:\Psi_i})$ is the measurement likelihood and $P(a_{1:\Psi_i})$ is a prior. The prior is proportional to a potential attractivity measure, that allows for merging different land use data sources. By definition, potential attractivity measure is a model of aggregated occupation per POI built on attractivity, such as number of jobs in the area or point-of-sale data in a supermarket, and time constraints, such as opening hours. In urban context, Buisson (2014) also uses a Bayesian approach. For a given cluster of access points:

$$P(\mathcal{A}_k|\hat{t}) \propto P(\hat{t}|\mathcal{A}_k) \cdot P(\mathcal{A}_k) \tag{2}$$

where $\mathcal{A}_k$ is an activity type and $\hat{t}_{1:J}$ is a set of measurement timestamps corresponding to the cluster. $P(\hat{t}|\mathcal{A}_k)$, the probability of generating a signal at a certain time knowing the activity type, is computed using time-use statistics, e.g., from travel diary surveys. The prior is similar to the one defined in Danalet *et al.* (2014), using OpenStreetMap data for list of POI, and census and national statistics for number of residents and employees.

2.1.2 Activity choice for pedestrians

Reviews about activity choice for pedestrians can be found in Timmermans *et al.* (1992), Bierlaire and Robin (2009) and Danalet *et al.* (2014)

Recently, Liu (2013) develops an activity-based travel demand model in the context of an airport, focusing on activity scheduling, destination and route choice and rescheduling models. It is based on revealed and stated preference survey data. The revealed preference survey was sent to faculty and staff from the author's university. They were supposed to describe the activities their performed and the activity with the longest duration the last time they visited an airport in the last 12 months. 359 responses were used for estimating the model. This thesis particularly focuses on congestion and consequent rescheduling. For Liu (2013), the home-based structure of urban activity behavior is replaced in airports by three structuring events: check-in, security check and boarding. Supposedly, in pedestrian context, the level of service (congestion, queues and flight schedules) is more important than people's characteristics, compared to activity choice in urban context. About the model, a nested logit is used. Each choice of activity is considered independently from other activities from a same individual. The nesting structure does not reflect any intuitive behavior.

Ton (2014) studies route and destination choice in train stations based on tracking and counting data. Counting data come from infrared scanners and tracking data come from WiFi and Bluetooth scanners. Counting data allow to apply the model to pedestrians without smartphones. The choice is between destinations for a given activity type. The choice used here is a logit. Kalakou and Moura (2014) apply a similar logit approach for destination choice for a given activity type (which coffee shop knowing that the individual is visiting one). This model includes space syntax in the specification of the utility through "integration", i.e., a measure of accessibility. The case study takes place in an airport.

2.2 Route choice and other discrete choices

2.2.1 Choice set generation

Choice set generation is the process of defining the considered alternatives in an individual decision making. Assumptions must be made about the availability of the different options and the decision maker's awareness of them. Availability or awareness of the alternatives can be deterministically or stochastically defined (Ben-Akiva and Bierlaire, 2003).

In the route choice context, the number of paths connecting an origin and a destination is very large and cannot be enumerated in practice. The universal choice set, containing all possible routes, cannot be used. There are two ways of dealing with it: selecting a choice set that only contains the paths considered by the decision maker (consideration choice set), or sampling a subset of paths large enough to be confident that it contains all important paths for the decisions maker (importance sampling). The consideration choice set is supposed to be consistent with behavior but is very often not available and too small for estimation, while the importance sampling is not very realistic behaviorally but is statistically more efficient.

Van Nes *et al.* (2008) propose a classification of different choice sets, by order of inclusion of alternatives: the chosen alternative, the considered alternatives, the reported alternatives (by the decision makers as considered), the feasible alternatives (i.e., available), the logical alternatives (i.e., no loop) and the existing alternatives (i.e., the universal choice set). They compare choice sets made of reported alternatives by the respondents and choice sets made of feasible alternatives defined by a set of constraints. Having access to reported alternatives is difficult, and even impossible when using localization data from smartphones or antennas. Moreover, when data are available, the size of the choice set is too small for model estimation (average size of 2.8 in Van Nes *et al.* (2008)).

Consideration choice set Depending on the data collection technique, the consideration choice set can be explicitly asked in a survey. This is very often not possible, in particular when using different traces (GPS, WiFi, or other tracking systems). In these cases, the consideration must be modeled and a choice set generation algorithm is defined. It can be seen as a pre-choice before the actual choice. These models are not based on data but on assumptions about how people choose the paths they evaluate.

Repeated shortest path search

These approaches assume that people consider only the shortest paths as possible alternatives. This is less restrictive that it might appear, by using a generalized cost. The repeated shortest path approaches assume that the consideration set is made of a large enough number of shortest paths. These approaches generate very similar paths. In order to represent the heterogeneity of all paths and the variety of choices, van der Zijpp and Fiorenzo Catalano (2005) propose to remove unrealistic paths. Instead of generating a large number of shortest paths and removing the irrelevant ones, they propose algorithms for the constrained shortest path problem, directly generating feasible shortest paths. Constraints are supposed to express relevance, such as attractivity, circuitousness, non-overlapping or detour.

Constrained enumeration

These approaches assume that people do not consider some alternatives due to constraints. They generate all possible alternatives satisfying these constraints using branch-and-bound. Prato and Bekhor (2006) describe the construction of a connection tree between the origin and the destination. It is depth-first built and the branching rule is based on logical constraints: shortest path constraints with tolerance for going backwards or for longer travel times, avoiding detours, loops, overlaps and left turns. Parameters for these constraints are hand-tuned in order to reach behavioral consistency. Consistency is defined as heterogeneity and realism, i.e., ability to reproduce actual chosen routes. 91.1% of chosen paths and 82.6% of reported paths are reproduced.

Importance sampling Flötteröd and Bierlaire (2013) propose a Metropolis-Hastings algorithm for the sampling of paths. Paths are sampled according to an arbitrary distribution, avoiding complete enumeration. In other words, the goal consists in drawing a path with probability $\frac{b(i)}{\sum_{i \in S} b(i)}$, with S the state space. The targets weights are defined as

$$b(i) = \frac{e^{-\mu\delta(\Gamma)}}{|\Gamma|(|\Gamma|-1)(|\Gamma|-2)/6} \tag{3}$$

where δ is the cost function of the path Γ, μ a scale factor and $|\Gamma|$ the number of nodes in path Γ. The denominator in the definition of $b(i)$ is justified by the state variable definition, a tuple (Γ, a, b, c), with a, b and c nodes on Γ.

The transition matrix Q is defined by two main operations. One operations randomly draw a node as a replacement of b and connect a and c through this new node with shortest paths according to the target weights. The other one redistribute a, b and c along Γ. Flötteröd and Bierlaire (2013) propose to draw the new node with a logit distribution using shortest path length through this node, in order to drive the process toward short paths. To guarantee scale-invariance with respect to path cost, they suggest the use of a scale parameter $\mu = \frac{\ln 2}{(\zeta-1)\delta_{SP}}$. In this way, the probability of choosing a path of cost $\zeta\delta_{SP}$ is twice less than the shortest path (with cost δ_{SP}). For the second operation, they propose a uniform ascending choice of a, b and c.

These techniques allow to sample path from a large network according to any sampling probabilities. The sampling probabilities do not need to be defined by link, but can be defined directly for the whole path. Importance sampling is important for an explicit correction in the discrete choice model.

Chen (2013) in Chapter 5 uses the Metropolis-Hastings path sampling technique by Flötteröd and Bierlaire (2013) for a route choice model estimated from GPS data. The weight function is composed of the path's length and of the frequency of observation of the given path. This

"observation score" represent the inclusion of observed GPS data in the sampling process in order to include more relevant observations. This algorithm reduces the needed choice set size.

2.2.2 Correlation structure

In route choice model, the different paths in the choice set are overlapping. Overlapping segments of paths share the same source of errors. For example, if the analyst does not know the high flow of pedestrians on a pedestrian crossing on a certain road segment, the impact of this pedestrian flow on route choice for cars will be absorbed by the error term and shared by all paths going through this road segment. The independence assumption about the error term is violated and the independence from irrelevant alternatives (IIA) property of the logit model does not hold.

Different solutions have been proposed: multinomal probit, cross nested logit, mixed logit.

Another approach to overcome the issue of correlated alternatives makes a deterministic correction of correlation. It is based on the theory of aggregation of alternatives (Ben-Akiva and Lerman, 1985). Already in 1985, the chapter about aggregation is motivated by "destination and other spatial choices". In route choice context, the elemental alternative is a path (Frejinger and Bierlaire, 2007). They are mutually exclusive, collectively exhaustive and the decision maker is choosing one and only one of them. The universal choice set containing all elemental alternatives should be decomposed into nonoverlapping subsets, aggregate alternatives. In route choice context, the aggregate alternatives are the links. Following the notation in Frejinger and Bierlaire (2007), n is the decision maker, C_n is the universal choice set of elemental alternatives, i.e., paths, $a = 1, ..., M$ are the links, C_{an} are the nonoverlapping subsets of aggregate alternatives, i.e., the set of paths using arc a. In the route choice context, the aggregate alternatives represent a group of the elemental alternatives are the same, i.e., path, constrained to go through one given arc.

We define the probability $P'_n(a)$ of choosing an aggregate alternative, i.e., a group of paths going through an arc a, as the probability that the decision maker chooses one of these paths $i \in C_{an}$:

$$P'_n(a) = \sum_{i \in C_{an}} P_n(i) \tag{4}$$

The utility of a path i is $U_{in} = V_{in} + \varepsilon_{in}$. The utility of a group of paths going through an arc a is $U_{an} = \max_{i \in C_{an}}(V_{in} + \varepsilon_{in})$, since only one of the paths with maximum utility is chosen. It can be

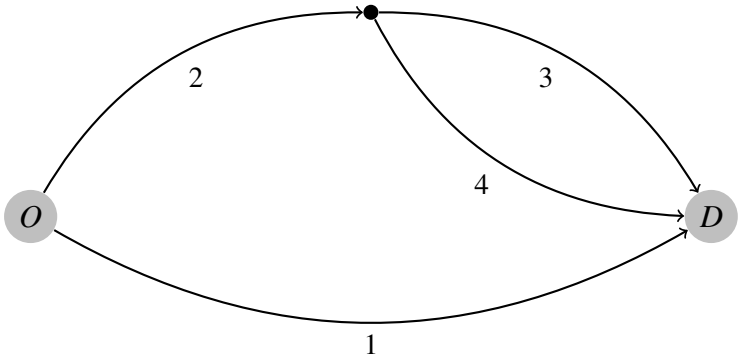

Figure 1: Network for a simple route choice model. $C_n = \{1, 2/3, 2/4\}$, $C_{1n} = \{1\}$, $C_{2n} = \{2/3, 2/4\}$, $C_{3n} = \{2/3\}$ and $C_{4n} = \{2/4\}$.

written as $U_{an} = V_{an} + \varepsilon_{an}$ with $V_{an} = E(\max_{i \in C_{an}}(V_{in} + \varepsilon_{in}))$. The average of the systematic part of the path utility is $\bar{V}_{in} = \frac{1}{M_i} \sum_{i \in C_{an}} V_{in}$.

3 A path choice approach to activity modeling

Our approach to activity-based travel demand modeling decomposes the behavior in two steps. First, a path choice approach models how people choose their activity type in time, taking into account the type of activities (e.g., eating), their sequence (e.g., eating first, then buying a ticket and finally waiting for the train) and their timing/duration (eating for 20 minutes, starting at 12.20 pm). Once the activity type, sequence and timing are chosen, a second step consists in modeling destination choice (e.g., choosing a restaurant, knowing the activity type: eating). We present here only the first step, the choice of activity sequence. An example of destination choice model is presented in Ton (2014) in a similar context.

There are mathematical and behavioral motivations for this decomposition. Mathematically, the problem is very complex. The number of destination is usually very large. The number of sequences of destinations is larger. Including the duration spent at destination makes the problem definitively too large and untractable. Behaviorally, the choice of activity type and time of day precedes the choice of destination. We experience it every day: around lunch break, people start to be hungry, decide they want to eat, decide it is time to go, and only then decide in which restaurant to go.

3.1 Representation of activity type and time

3.1.1 Activity-episode sequences and activity patterns

We define an activity episode $a_{\psi_i} = (x, t^-, t^+)$ as a location, usually a point of interest (POI), where the individual i is spending time. x is the episode location, t^- is the start time, and t^+ is the end time. t^- and t^+ are continuous random variables, expressing the fact that we don't know them perfectly. $t^+ - t^-$ defines the episode duration.

Each individual i performs activity-episode sequences $(a_1, ..., a_{\Psi_i})$, which is abbreviated $a_{1:\Psi_i}$, where Ψ_i is the total number of episodes. Ψ_i is individual specific. Depending on the way the data are collected, there could be one activity-episode sequence per individual/observation i or several different activity-episode sequences representing the ambiguity in the data. In this case, each of the L different activity-episode sequences is associated with the probability $P(a_{1:\Psi_i}|\hat{m})$ of being the actual one (knowing the measurements $\hat{m}$). Danalet *et al.* (2013) derive this probability, explicitly representing the ambiguity in the data.

We assume K activity types $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_k, ..., \mathcal{A}_K$. For each activity episode $a_{\psi_i} = (x, t^-, t^+)$, an activity A_{ψ_i} is defined, $A_{\psi-i} = (A(a_{\psi_i}), t^-, t^+)$. $A(a_{\psi_i})$ is a function mapping activity episode a_{ψ_i} to an activity type $\mathcal{A}_k$. $A(a_{\psi_i})$ depends on the episode location x, its start time t^-, its end time t^+ and the individual i. The list of activities $A_1, ..., A_{\Psi_i}$ corresponding to the activity-episode sequence, with their associated start times t^- and end times t^+, is called an activity pattern and is abbreviated $A_{1:\Psi_i}$. It contains the same number Ψ_i of elements. Its probability of being the actual one is

$$P(A_{1:\Psi_i}) = \sum_{A_{1:\Psi_i}=A(a_{1:\Psi_i})} P(a_{1:\Psi_i}|\hat{m}) \tag{5}$$

Since we have L different activity-episode sequences and some of them may correspond to the same activity pattern, the set $\mathcal{L}_i$ of all activity patterns corresponding to an observation i has less than L elements.

Activity patterns are the behavior we observe.

Illustration In a train station, let's assume an individual has $K = 3$ possible activity types: waiting for the train on the platform, buying a ticket, and having a coffee. The individual enters the station at 7:39 and starts by buying a ticket, from 7:40 to 7:43, then drinks a coffee from 7:47 to 8:01 and goes on the platform, waiting there from 8:03 to 8:12, until the departure of the

train. This activity pattern is represented in Figure 2.

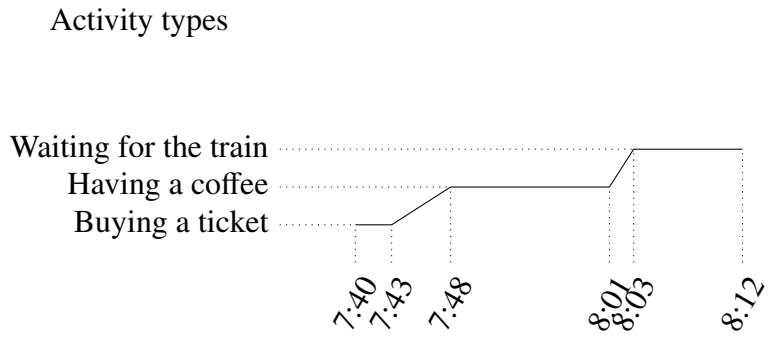

Figure 2: Illustration: an activity pattern on a campus.

A similar example in the context of a day-long urban model is presented in Bowman (1998), p.15.

3.1.2 Activity network

We assume a discretization of time, $\tau \in 1, 2, ..., T$. An activity network represents the choice set and contains all possible activity patterns. It is discrete with respect to activity types $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_K$ and time. It is composed of links and nodes. Nodes $\mathcal{A}_{k,\tau}$ represent the performance by the individual of an activity type k for a unit of time τ. At a given unit of time τ, the number of nodes represent the available activity types K. There are two special nodes, start node s and end node e. They represent the beginning and the end of the observed activity pattern. In total, the activity network contains a maximum of $KT + 2$ nodes. Edges connect some of these nodes and represent the fact that they are successively performed. s is connected to all nodes; it represents the beginning of the observed activity pattern in the time unit of the arrival node. All nodes are connected to e; it represents the end of the observed activity pattern in the time unit of the departure node. All nodes of a given time unit t are connected with the nodes corresponding to the next time unit $t + 1$; it represents the choice of changing activity type or maintaining the activity type for one more time unit. In total, the activity network contains a maximum of $2KT + K^2T$ edges.

The activity network is a representation of all possible activity patterns. It is a representation of the universal choice set. Time constraints for activity types can be included in it by removing some nodes corresponding to a specific activity type and time period.

The timeline must be defined in two ways. First, the time unit must be defined for discretization. Second, the first and last elements, $\tau = 1$ and $\tau = T$, must be defined. They define the earliest

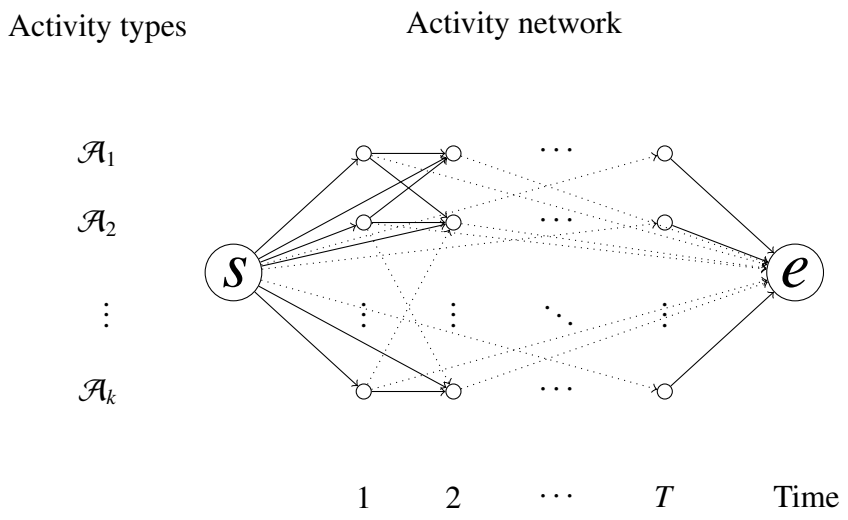

Figure 3: The activity network: the start node s is connected with all nodes and all nodes are connected to the end node e

possible start time and the latest possible end time. In an urban context, it would be defined by a day (τ_1 = 4am), assuming a major period of rest in the night. In a pedestrian infrastructure context, it would be the opening hour of the facility when it applies or a day.

It needs to be stressed that the activity network is cycle-free and the only path between the upstream node and the downstream node of every link is the link. Thus, it complies with the requirements in Flötteröd and Bierlaire (2013), independently of the target weights defined later.

Illustration Coffee shops open at 7:30. In this case, the activity network is shown in Figure 4, with a time discretization of 5 minutes.

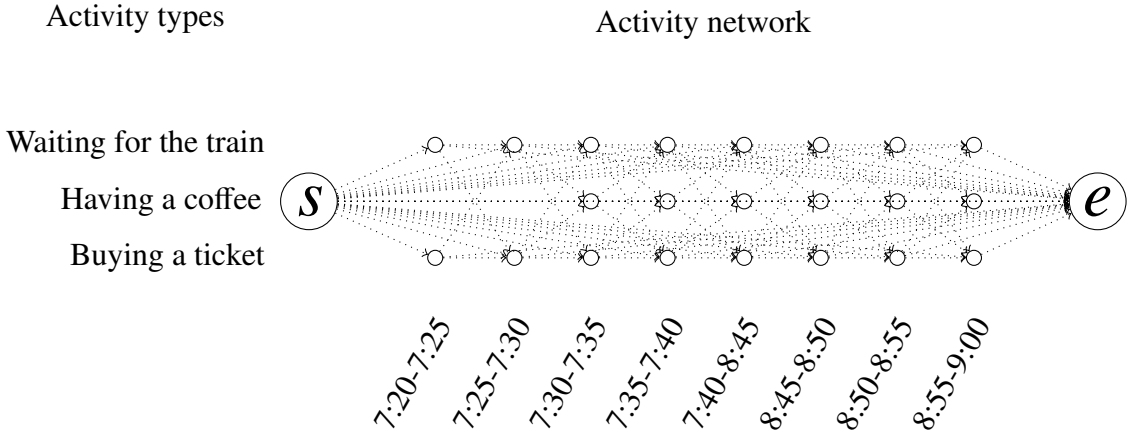

Figure 4: Illustration: an activity network for a train station.

3.1.3 Activity path

Activity paths $\mathcal{A}_{1:T}$ are the representation of an activity pattern $A_{1:\Psi_i}$ in an activity network. One activity pattern can correspond to several activity paths, due to imprecision in the measurement of time. The set of candidate paths corresponding to the activity pattern $A_{1:\Psi_i}$ is called $\mathcal{P}_{A_{1:\Psi_i}}$.

In order to represent the activity patterns described in Section 3.1.1 with continuous time in an activity network, we define a very simple rule. The activity of the time unit in the activity network is the longest activity in the activity pattern for this time interval. In case there are long distances between the activities (i.e, trips longer than a time unit), a "traveling" activity type must be defined.

Activity paths are generated to cover all start and end time support. For each element in $A_{1:\Psi_i}$ with activity type $\mathcal{A}_k$, start time t^- and end time t^+, several nodes $\mathcal{A}_{k,\tau}$ in the activity network are generated with τ such that $P(t^- > \tau) = P(t^+ < \tau) = 0$. The first node corresponding to the first element A_1 of the activity pattern is connected to s, and similarly the last node corresponding to the last element A_{Ψ_i} of the activity pattern is connected to e.

For each activity path $\mathcal{A}_{1:T}$ generated from an activity pattern $A_{1:\Psi_i}$, we need to compute the probability $P(\mathcal{A}_{1:T}|A_{1:\Psi_i})$ that this activity path is the true one based on the distribution of the start time t^- and the end time t^+ of each activity episode. For a given time interval $t \in \{1, ..., T\}$ and a given activity pattern $A_{1:\Psi_i}$:

$$\sum_{k=1}^{K} \mathcal{A}_{k,\tau} = 1 \tag{6}$$

For a given activity path $\mathcal{A}_{1:T}$, the probability that it represents the activity pattern $A_{1:\Psi_i}$ is the probability that each node $\mathcal{A}_{k,\tau}$ represents the activity pattern at this particular time interval τ. For each node, the probability that it is the true one is proportional to the episode duration $t^+ - t^-$ in this time interval τ, compared to the episode duration for other activities different than k.

$$P(\mathcal{A}_{1:T}|A_{1:\Psi_i}) = \prod_{\tau=1}^{T} P(\mathcal{A}_{k,\tau}|A_{1:\Psi_i}) \tag{7}$$

$$= \prod_{\tau=1}^{T} \frac{duration_{k,\tau}(A_{1:\Psi_i})}{duration_{\tau}(A_{1:\Psi_i})} \tag{8}$$

where $duration_{k,\tau}$ is a function computing the time spent performing activity type k during time

interval τ for a given activity pattern, and $duration_\tau$ is a function computing the time spent performing any activity type during time interval τ for a given activity pattern.

$$duration_\tau(A_{1:\Psi_i}) = \sum_{m=1}^{M} \min\left(E(t^+), \tau_{UB}\right) - \max\left(E(t^-), \tau_{LB}\right) \tag{9}$$

$$duration_{k,\tau}(A_{1:\Psi_i}) = \sum_{\substack{m=1 \\ A_m=\mathcal{A}_k}}^{M} \min\left(E(t^+), \tau_{UB}\right) - \max\left(E(t^-), \tau_{LB}\right) \tag{10}$$

where t_{LB} is the lower bound of the interval τ and τ_{UB} is the upper bound of τ.

The set of all candidate activity paths for a given individual i is called $\mathcal{P}_i$ and consists in each set of candidate activity path $\mathcal{A}_{1:T}$ corresponding to each of the L activity patterns $A_{1:\Psi_i}$,

$$\mathcal{P}_n = \bigcup_{\mathcal{L}_i} \mathcal{P}_{A_{1:\Psi_i}} \tag{11}$$

$$P(\mathcal{A}_{1:T}) = \sum_{\mathcal{L}_i} P(\mathcal{A}_{1:T}|A_{1:\Psi_i})P(A_{1:\Psi_i}) \tag{12}$$

Activity paths are the alternatives of the choice process.

Illustration Between 7:40 and 7:45, the main activity is buying a ticket; between 7:45 and 7:50, it is having a coffee; etc. Between 7:35 and 7:40, the individual is not in the area of study and so no activity is executed. This activity path is represented in Figure 5.

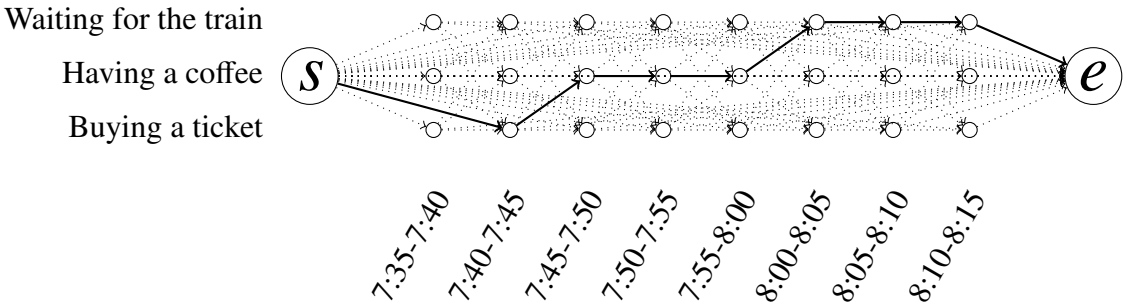

Figure 5: Illustration: an activity path in a train station.

There might be uncertainties in the time measurement. It comes from the error of the respondent in a traditional survey or from the measurement error of a tracking technology. In the previous

example, let's assume there is uncertainty in the departure time from the coffee shop. It happened between 8:01 and 8:04, and the arrival on the platform between 8:03 and 8:06. In this case, there are two possible activity paths for this activity pattern.

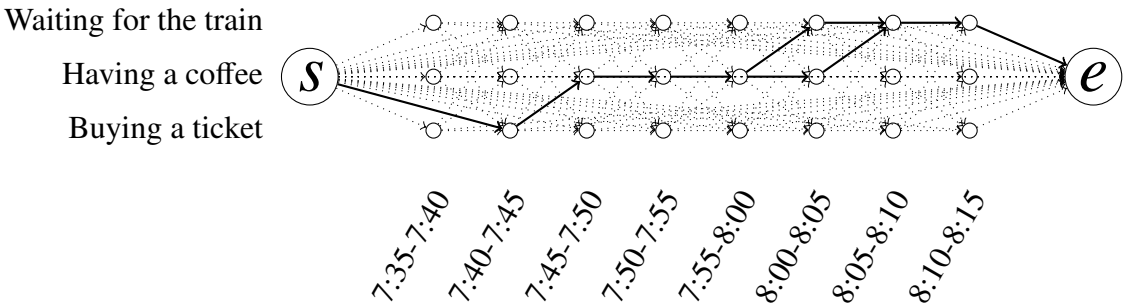

Figure 6: Illustration: an activity network with ambiguity in start and end times.

If we assume a uniform distribution between the bounds, the duration of the activity "Having a coffee" is:

$$\begin{aligned} duration_{\text{"Having a coffee", 8:00-8:05}} &= \min\left(E(t^+), t_{UB}\right) - \max\left(E(t^-), t_{LB}\right) && (13)\\ &= 8{:}02{:}30 - 8{:}00 && (14)\\ &= 2'30 && (15) \end{aligned}$$

And the duration of all activities in the time interval t is:

$$\begin{aligned} duration_{8:00-8:05} &= \sum_{m=1}^{2} \min\left(E(t^+), t_{UB}\right) - \max\left(E(t^-), t_{LB}\right) && (16)\\ &= 2'30 + (8{:}05 - 8{:}04{:}30) && (17)\\ &= 3' && (18) \end{aligned}$$

The probability of following the activity path including having a coffee between 8:00 and 8:05 is:

$$P(\mathcal{A}_{1:T}|A_{1:\Psi_i}) = \frac{2'30}{3'} = \frac{5}{6} \quad (19)$$

while the probability of following the activity path including "waiting for the train" between 8:00 and 8:05 is $\frac{1}{6}$.

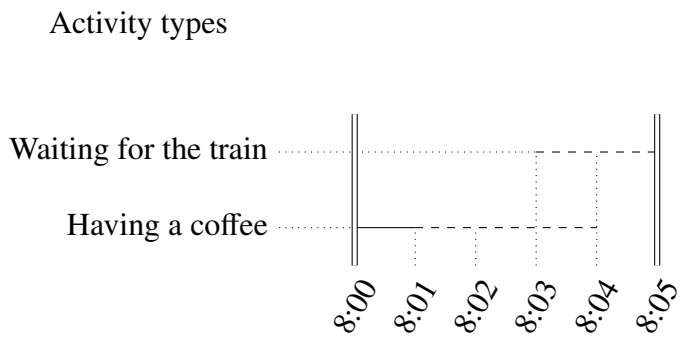

Figure 7: Illustration: an activity pattern in a train station with uncertainty in time, zoom in the time interval 8:00-8:05.

3.2 Choice set generation

The Metropolis-Hastings algorithm defined by Flötteröd and Bierlaire (2013) requests the target weight of a path for importance sampling. It can be defined for each link, and then added up, or defined directly for the whole path. As explained in Frejinger and Bierlaire (2010), "the sample should include attractive alternatives" in order to provide efficient estimators. In the context of activity choice modeling, we propose two measurements of the attractive alternatives: potential attractivity measure and the length and frequency of observed paths in the activity network.

3.2.1 Generation from potential attractivity measure

Potential attractivity measure has already been defined in Danalet *et al.* (2013). We define a potential attractivity measure by merging attractivity and time constraints for the pedestrian context. Formally, we define the potential attractivity measure as a model of aggregated occupation per point of interest (POI). The unit of attractivity is the number of persons. The potential attractivity measure $S_{x,i}(t^-, t^+)$ between a start time t^- and an end time t^+ for $x \in POI$ and individual i is time dependent and may differ across individuals. It depends on the instantaneous potential attractivity measure $S_{x,i}(t)$ at a given time t:

$$S_{x,i}(t^-, t^+) = \int_{t=t^-}^{t^+} S_{x,i}(t)dt \tag{20}$$

The instantaneous potential attractivity measure depends on time-constraints and attractivity:

$$S_{x,i}(t) = sched_{x,i}(t) \cdot att(x,t) \tag{21}$$

where $sched_{x,i}(t)$ is a dummy variable for time-constraints such as schedules or opening hours, with value 1 if the *POI* is open or scheduled and 0 otherwise: opening hours of shops and restaurants, or timetables in the case of conferences, campuses, or public transport infrastructures. Timetables may vary for different individuals, depending on the level of anonymity for localization data (Danalet *et al.,* 2013).

Attractivity $att(x, t)$ is context-specific (Danalet *et al.,* 2013): number of jobs, annual retail sale, population per zone, places at day-nurseries, hospital beds. In the pedestrian facility context, data sources could be checkouts in supermarkets, number of seats in a restaurant, number of employees per office, number of students in class, capacity of different zones in a stadium or a public transport infrastructure.

The potential attractivity measure of an activity type $\mathcal{A}_k$ for a time interval τ (i.e., for a node $\mathcal{A}_{k,\tau}$) is the sum of the potential attractivity measure over all $x \in POI$ corresponding to this activity type:

$$S_{\mathcal{A}_k,\tau} = \sum_{A(x,\tau_{LB},\tau_{UB})=\mathcal{A}_k} S_{x,i}(\tau_{LB}, \tau_{UB}) \tag{22}$$

The potential attractivity measure $S_{\mathcal{A}_k,\tau}$ corresponds to a node $\mathcal{A}_{k,\tau}$ in the activity network. In order to keep the link-additive specification of cost used in Flötteröd and Bierlaire (2013), weight can be associated to the entering edge by convention.

Let's call Γ the generated activity path, to differentiate them from the observed activity path $\mathcal{A}_{1:T}$. In order to keep the shortest path formulation from Flötteröd and Bierlaire (2013), node cost δ_v for node v can be defined as

$$\delta_v(v) = \max_{\substack{k=1,\dots,K \\ \tau=1,\dots,T}} (S_{\mathcal{A}_{k,\tau}}) - S_v + 1 \tag{23}$$

The generalized cost $\delta(\Gamma)$ of the activity path Γ is:

$$\delta(\Gamma) = -\mu_v \cdot \sum_{v\in\Gamma} \delta_v(v) - \delta_\Gamma(\Gamma) \tag{24}$$

where μ_v is a nonnegative real number that defines the weight of the sum in Equation 24, and $\delta(\Gamma)$ represents the non-link-additive cost of path Γ.

If δ is based only on link cost, the most likely output of the Metropolis-Hastings algorithm consists in spending one time unit in the most attractive activity. Link cost is based on attractivity but does not consider the likelihood of spending more than one time unit in an activity, nor the possibility of visiting several activities.

In order to account for time constraints, the non-link-additive cost of Γ can be defined as

$$\delta_\Gamma(\Gamma) = \frac{1}{N} \sum_{\mathcal{A}_{1:T} \in N} \mathcal{I}(|\mathcal{A}_{1:T}| = |\Gamma|) \tag{25}$$

where $|\Gamma|$ is the number of nodes in the generated activity path Γ and $|\mathcal{A}_{1:T}|$ is the number of nodes in the observed activity path $\mathcal{A}_{1:T}$.

This approach can be extended to each activity type:

$$\delta_\Gamma(\Gamma) = \prod_{k=1}^{K} \left(\frac{1}{N} \sum_{\mathcal{A}_{1:T} \in N} \mathcal{I}(|\mathcal{A}_{1:T,k}| = |\Gamma_k|) \right) \tag{26}$$

where $|\Gamma_k|$ is the number of nodes in the generated activity path Γ with activity type k, and $|\mathcal{A}_{1:T,k}|$ is the number of nodes in the observed activity path $\mathcal{A}_{1:T}$ with activity type k.

We use the exponentially decreasing function of the generalized cost as unnormalized target weights $b(\Gamma)$ described in Eq. 3, as proposed by Flötteröd and Bierlaire (2013). The output of the Metropolis-Hastings algorithm is a sample of paths with their sampling probability $q(\Gamma)$, $\Gamma \in \mathcal{U}$, expressed in an unnormalized form $b(\Gamma)$. The choice set $C_i \in \mathcal{U}$ for observation i is the set of the sampled paths with the observed activity path $\mathcal{A}_{1:T}$.

Illustration Let's fix the potential attractivity measure for each activity type: 5 for "Waiting for the train", 1 for "Having a coffee" and 4 for "Buying a ticket". The weight is shown on some edges in Figure 8. The weight of the edge corresponds to the activity type of the arrival node of the edge.

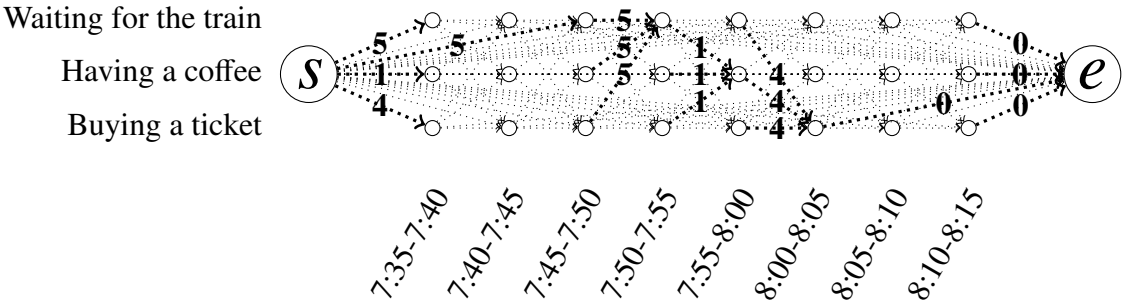

Figure 8: Illustration: target weights in an activity path in a train station.

The algorithm presented in Flötteröd and Bierlaire (2013) is applied with $\mu_v = 1$ and $\delta(\Gamma) = 0$ in Figure 9. 20 paths are generated. We observe that they are "shortest path" in the sense that they spend the smallest amounts of time in the most likely activities. They are not realistic activity paths in the railway station context. In Figure 10, a constraint is defined in order to end the path

not in e but on the platform at 8:12, when the train is leaving. Activity paths are still not very realistic. Figure 11 presents one more constraint, about the non-link-additive cost of the path. Since there is only one observation here, we replace the empirical distribution of length by

$$\delta(\Gamma) = \left| |\Gamma| - \mu_\Gamma | \mathcal{A}_{1:T} | \right| \tag{27}$$

corresponding to Equation 25 with a triangular distribution. It forces the generated paths to have a small difference in overall time spent in the railway station compared to the observed activity path.

This approach introduces endogeneity in the model since it uses observed choices as input for the choice set generation (Chen, 2013). Even it the paths look reasonable, it is recommended to use the length distribution of all observed paths, as showed in the next section.

3.2.2 Generation from length and frequency of observed paths

Another approach to define the attractivity of a node in the activity network for choice set generation purpose consists in using all observed paths in the network. In the route choice context, Chen (2013) calls it the observation score. It is defined here as the sum of all observed path through a given node v.

$$\sum_{\mathcal{A}_{1:T} \in \mathcal{P}} P(\mathcal{A}_{1:T} | \hat{m}_{1:J}) I(v \in \mathcal{A}_{1:T}) \tag{28}$$

where $I(v \in \mathcal{A}_{1:T})$ is an indicative function with value 1 if v is a node of activity path $\mathcal{A}_{1:T}$ and 0 otherwise. In order to keep the shortest path formulation from Flötteröd and Bierlaire (2013), node cost δ_v for node v can be defined as:

$$\delta_v(v) = \max_{\substack{k=1,\dots,K \\ \tau=1,\dots,T}} \left(\sum_{\mathcal{A}_{1:T} \in \mathcal{P}} P(\mathcal{A}_{1:T} | \hat{m}_{1:J}) I(\mathcal{A}_{k,\tau} \in \mathcal{A}_{1:T}) \right) - \sum_{\mathcal{A}_{1:T} \in \mathcal{P}} P(\mathcal{A}_{1:T} | \hat{m}_{1:J}) I(v \in \mathcal{A}_{1:T}) + 1 \tag{29}$$

Similarly to the generation from potential attractivity measure, using only a link cost will leads to very short activity paths. The most likely output is one time unit in the most attractive activity type. The time distribution can be derived from observed activity paths. The number of time units for an activity path $\mathcal{A}_{1:T}$ is equivalent to the number of nodes $|\mathcal{A}_{1:T}|$. We define the non-link-additive cost of Γ as the number of observed paths with length $|\Gamma|$:

$$\delta_\Gamma(\Gamma) = \sum_{\mathcal{A}_{1:T} \in \mathcal{P}} P(\mathcal{A}_{1:T} | \hat{m}_{1:J}) I(|\mathcal{A}_{1:T}| = |\Gamma|) \tag{30}$$

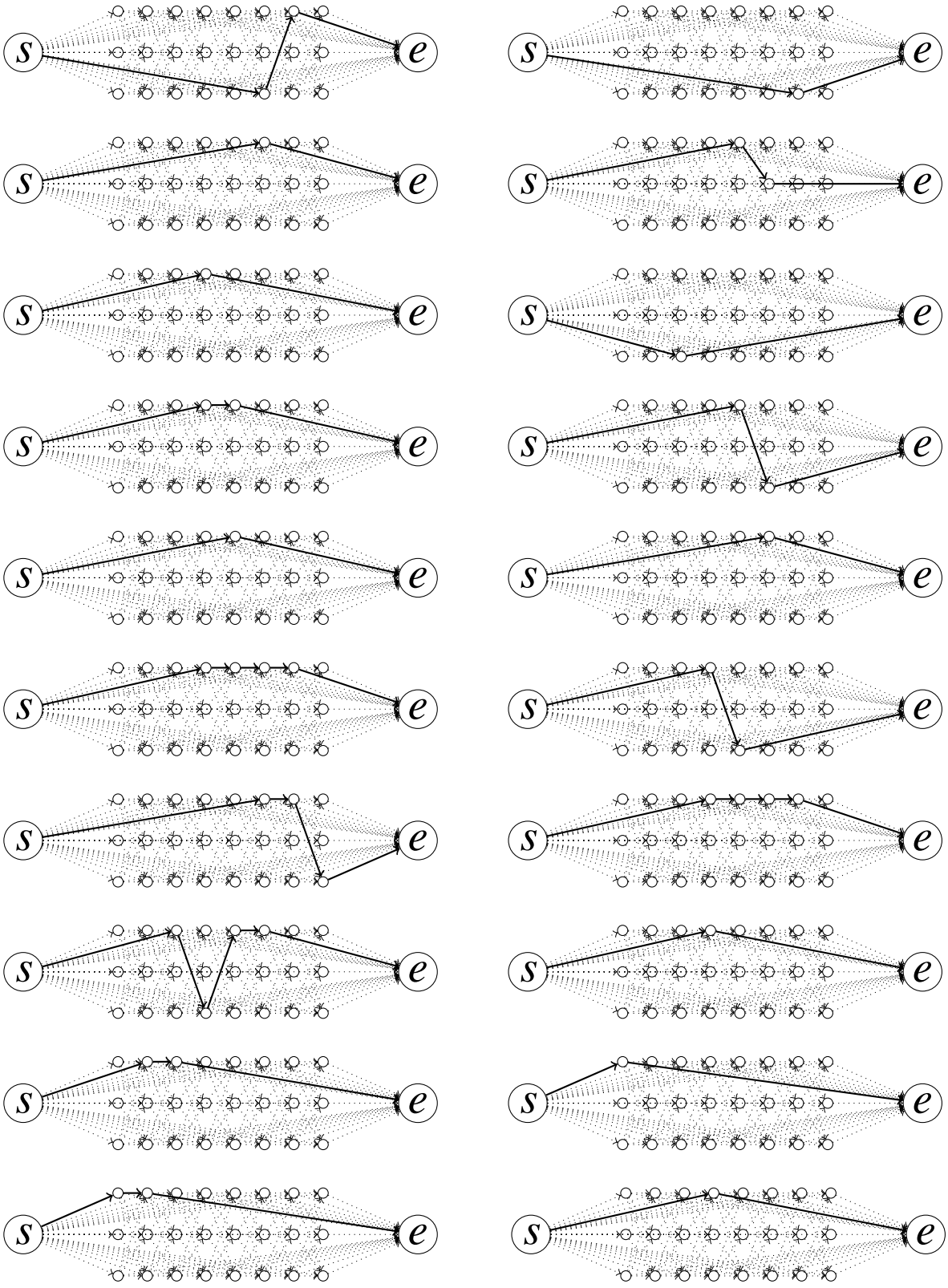

Figure 9: The 20 generated paths based only on link cost.

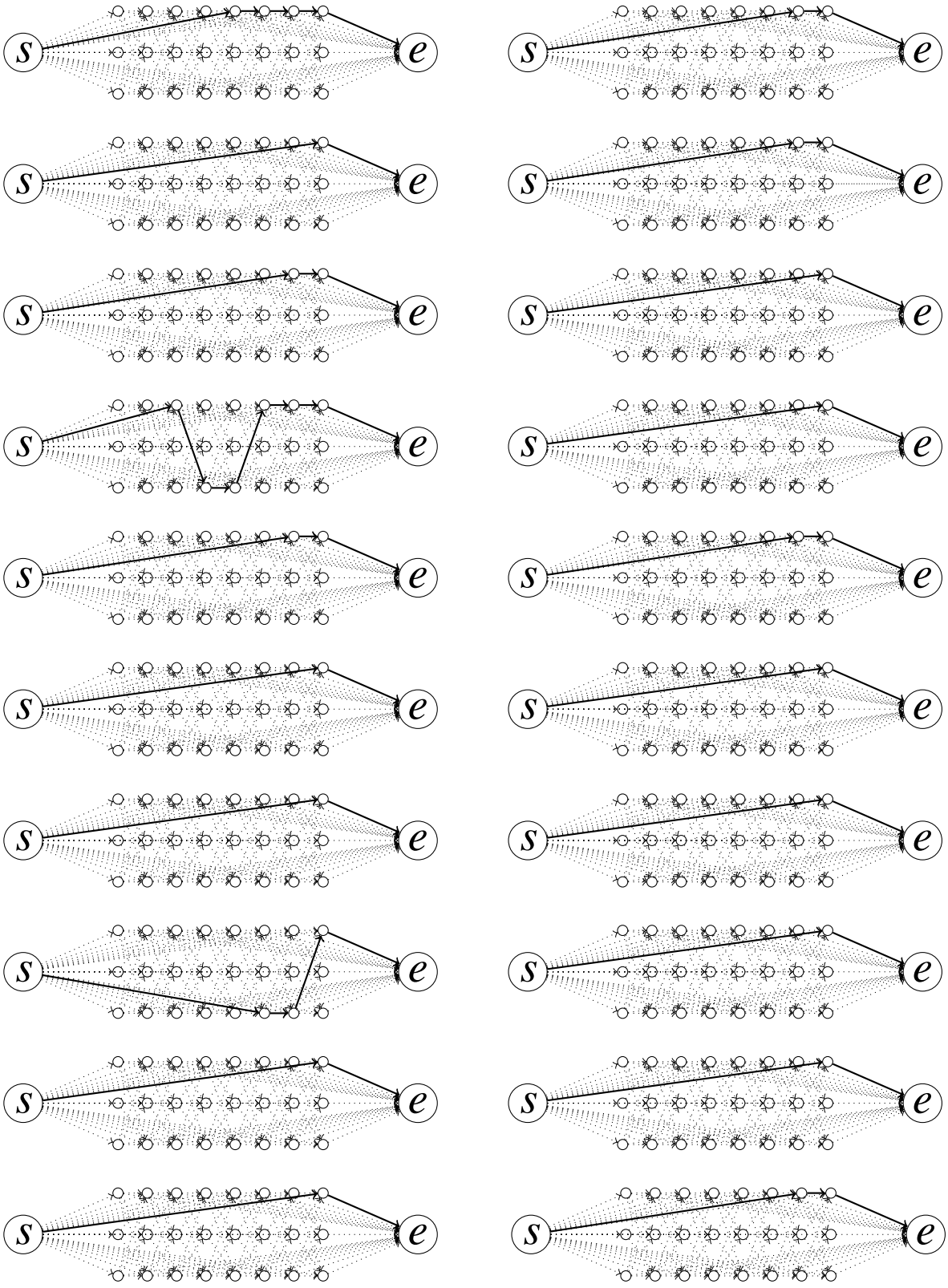

Figure 10: The 20 generated paths with constrain on the link to e.

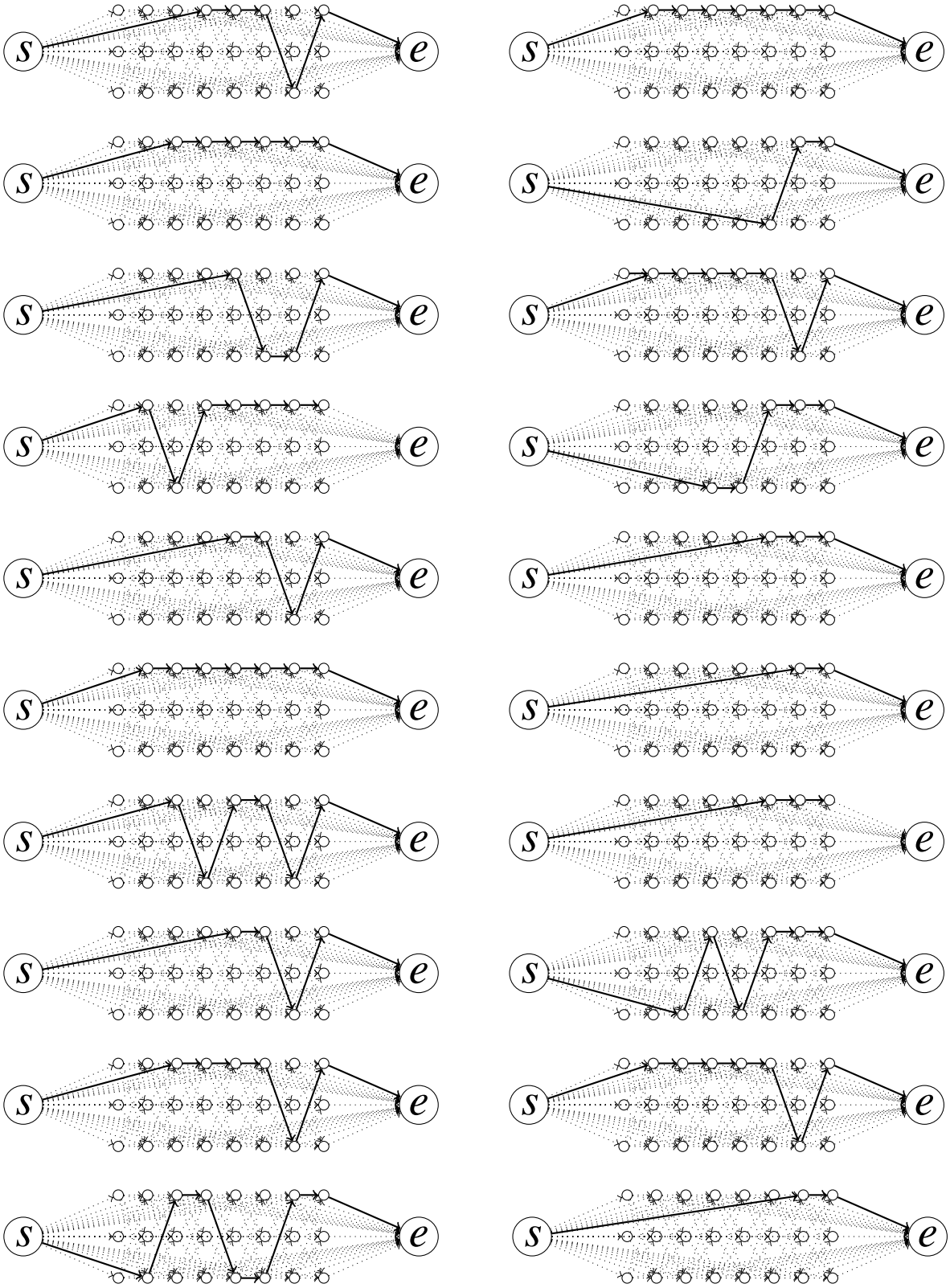

Figure 11: The 20 generated paths with constrain on the link to e and constrain about the length with $\mu_v = \mu_\Gamma = 1$

Similarly to Equation 29, the number of observed paths is weighted by the measurement equation.

3.3 Activity path choice model for WiFi traces

We propose a choice modeling framework for activity paths. This framework uses the probabilistic association between the WiFi data and the activity network with the choice set generated in the previous section, and corrects for importance sampling.

For each individual i, a set of measurements $\hat{m}$ covering the period of interest (one day or the time in the pedestrian infrastructure) is observed. Using the methodology from Danalet *et al.* (2013), signals are generating L activity-episode sequences $a_{1:\Psi_i}$. Each activity episode a_{ψ_i} corresponds to an activity A_{ψ_i}, and activity-episodes sequences can be transformed into an activity pattern $A_{1:\Psi_i}$. Activity patterns are represented in an activity network as activity path $\mathcal{A}_{1:T}$. Discretization of time from continuous random variables possibly leads to several activity paths corresponding to one activity pattern.

Inspired by Bierlaire and Frejinger (2008) and Chen (2013), each individual i generates J "network-free" measurements $\hat{m}_{1:J}$ with some error. Measurements $\hat{m}$ are defined in detail in Danalet *et al.* (2013). The choice probability $P(\hat{m}_{1:J})$ for individual i is

$$P(\hat{m}_{1:J}) = \sum_{\mathcal{A}_{1:T} \in \mathcal{U}} P(\hat{m}_{1:J}|\mathcal{A}_{1:T}) \cdot P(\mathcal{A}_{1:T}|\mathcal{U};\beta) \tag{31}$$

where $P(\hat{m}_{1:J}|\mathcal{A}_{1:T})$ is the measurement likelihood of observing measurements $\hat{m}_{1:J}$ while performing activities $\mathcal{A}_{1:T}$ and $P(\mathcal{A}_{1:T}|\mathcal{U};\beta)$ is a path choice model with unknown parameters β and choice set $\mathcal{U}$.

To be operationalized, the model must correct for the sampling of alternatives from a very large choice set, and for the correlation structure of a path choice. In Section 3.3.1, a measurement likelihood is presented. Then, in Section 3.3.2, a correction term for the sampling of alternatives is presented. The correlation structure is discussed in Section 3.3.3. Finally, Section 3.3.4 describe the structure of the utility function.

3.3.1 Measurement likelihood

For each activity pattern, our goal is to compute the probability that the performed episodes generated the observed measurements $\hat{m}_{1:J}$:

$$P(\hat{m}_{1:J}|\mathcal{A}_{1:T}) \tag{32}$$

This is equivalent to the measurement likelihood of the original activity-episode sequence:

$$P(\hat{m}_{1:J}|a_{1:\Psi_i}). \tag{33}$$

If one activity-episode sequence corresponds to several activity paths in the activity network, the activity paths have the same measurement likelihood. The imprecision in time leading to several different activity paths is related to a lack of information in some points in time. In this period of time without information, there is no measurement, thus no measurement likelihood (see Danalet *et al.* (2013) for an example).

We assume that a measurement $\hat{m}_j$ always corresponds to an activity episode a_{ψ_i}. We denote $\hat{m}_j^{\psi} = (\hat{x}_j^{\psi}, \hat{t}_j^{\psi})$ the measurements in $\hat{m}_{1:J}$ corresponding to $a_{\psi_i} = (x_{\psi}, t_{\psi}^-, t_{\psi}^+)$, i.e. when $t_{\psi}^- \leqslant \hat{t}^j \leqslant t_{\psi}^+$. As a result, $\hat{m}_{1:J} = \cup_{\psi} \hat{m}_{1:J}^{\psi}$.

If the device's owner is performing activity episode a, the probability that it will generate a measurement $\hat{m}$ is a function of the location of the episode location x and the measurement location $\hat{x}$ (e.g., the distance). Thus we can decompose Equation 33 as:

$$\begin{aligned} P(\hat{m}_{1:J}|a_{1:\Psi_i}) &= \prod_{\psi=1}^{\Psi_i} P(\hat{m}_{1:J}^{\psi}|a_{\psi}) && (34)\\ &= \prod_{\psi=1}^{\Psi_i} \prod_{j=1}^{J} P(\hat{m}_j^{\psi}|a_{\psi}) && (35)\\ &= \prod_{\psi=1}^{\Psi_i} \prod_{j=1}^{J} P(\hat{x}_j^{\psi}|x_{\psi}) && (36) \end{aligned}$$

Equality in Equation 34 assumes measurement independence between activities: measurement error of network traces is only related to the corresponding activity episode in time. Equality in Equation 35 assumes measurement independence between signals: measurement error is the

same for different signals while in the same location x_ψ in the same time interval t_ψ^-, t_ψ^+. Equality in Equation 36 assumes no measurement error in time: measurement error is a localization error.

If $\mathcal{A}_{1:T} \notin \mathcal{P}_i$, $P(\hat{m}_{1:J}|\mathcal{A}_{1:T}) = 0$.

3.3.2 Sampling alternatives

We assume the choice set to be the universal choice set containing all possible paths between s and e in the activity network. The sampling strategy for choice set generation presented in Section 3.2 requires the deterministic part of the utility to be corrected in order to estimate unbiased parameters. According to Frejinger *et al.* (2009), a sampling correction term must be added:

$$\ln q(C_n|\Gamma) = \ln \frac{k_{\Gamma n}}{q(\Gamma)} \tag{37}$$

where $k_{\Gamma n}$ is the number of times activity path Γ is drawn in C_n and $q(j)$ is the sampling probability.

The correction term is particularly necessary since there is no alternative specific constant in the utility, similarly to traditional route choice. The alternative specific constant would allow to estimate all other parameters without bias. Alternative specific constants cannot be estimated due to the size of the choice set.

The sampling probability $q(\Gamma)$ is available using the unnormalized target weights $b(\Gamma)$ but require full enumeration for normalization: $q(\Gamma) = \frac{b(\Gamma)}{\sum_{\Gamma' \in \mathcal{U}} b(\Gamma')}$. In practice, the normalizing sum cancels out in the logit formulation:

$$P(\Gamma|C_n) = \frac{e^{\mu V_{\Gamma n} + \ln \frac{k_{\Gamma n}}{q(\Gamma)}}}{\sum_{\Gamma' \in C_n} e^{\mu V_{\Gamma n} + \ln \frac{k_{\Gamma' n}}{q(\Gamma')}}} \tag{38}$$

$$= \frac{\sum_{\Gamma' \in \mathcal{U}} b(\Gamma') \cdot e^{\mu V_{\Gamma n}} \cdot \frac{k_{\Gamma n}}{b(\Gamma)}}{\sum_{\Gamma' \in \mathcal{U}} b(\Gamma') \cdot \sum_{\Gamma' \in C_n} e^{\mu V_{\Gamma n}} \cdot \frac{k_{\Gamma' n}}{b(\Gamma')}} \tag{39}$$

$$= \frac{e^{\mu V_{\Gamma n}} \cdot \frac{k_{\Gamma n}}{b(\Gamma)}}{\sum_{\Gamma' \in C_n} e^{\mu V_{\Gamma n}} \cdot \frac{k_{\Gamma' n}}{b(\Gamma')}} \tag{40}$$

3.3.3 Correlation structure

We propose here two possible correlation structures, inspired from route choice techniques.

Activity path size Let's consider an example, inspired by the case study (Section 4): a student and an employee on campus are going for lunch, in a restaurant, between 12.15 pm and 12.30 pm. In terms of activity paths, they visit the same node (same activity type, same time unit). Let's further assume there is a queue at this time in general in all restaurants on campus. The queue is not measured nor observed by the modeler. It is shared by both the employee and the student, independently of what they have done before and which activity they will perform after lunch.

Ben-Akiva and Bierlaire (1999) proposed the traditional path size logit model, where a path size attribute PS_p corrects the utility for the correlation related to overlapping segments of paths.

$$PS_p = \sum_{a \in p} \frac{L_a}{L_p} \frac{1}{M_a} \tag{41}$$

where a is an arc of path p, L_a and L_p are the length of arc a and path p, and M_a is the number of paths in C_n using link a, i.e., $M_a = \sum_{j \in C_{ps}} \delta_{aj}$ (δ_{aj} is the link-path incidence variable that is one if link a is on path p and 0 otherwise). PS_p is usually computed for a limited number of paths when using a consideration set C_{ps}. When sampling from the universal choice set, M_a must be computed for all paths. Frejinger *et al.* (2009) recommend to use a large set of paths for C_{ps} in practice.

When dealing with an activity network, its symmetry can be used to compute M_a. Assuming that all K activity types are available for each time interval τ, $M_a = K^{\tau-1}$ and does not depend on a. Moreover, the length L_a and L_Γ of arcs and paths correspond to the number of time intervals, $L_a = 1$, $L_\Gamma = |\Gamma| - 1$, and thus we define the activity path size APS_Γ for path Γ as:

$$APS_\Gamma = \frac{1}{K^{\tau-1}} \sum_{a \in \Gamma} \frac{1}{|\Gamma| - 1} \tag{42}$$

$$= \frac{1}{K^{\tau-1}} \tag{43}$$

A second approach considers aggregation similar to route choice, but neglecting time. the elemental alternatives correspond to activity paths and the aggregate alternatives are activity

types. The size of aggregate alternatives, activity types, equals the number of activity paths performing this activity type. C_n is the set of paths considered by individual n. $C_{an} \subseteq C_n, a = 1, ..., M$ is the set of paths using activity type $a.$, and M is the number of activity types. The utility of path i is $U_{in} = V_{in} + \varepsilon_{in}$. The utility of activity type a is $U_{an} = \max_{j \in C_{an}}(V_{jn} + \varepsilon_{jn}), a = 1, ..., M$. M_a is the number of paths using activity type a.

The size of C_{an} is large: all activity types are visited by a lot of different activity paths. Path utilities using an given activity type have equal means and ε_{in} are i.i.d.. The utility associated with activity type a is:

$$U_{an} = \bar{V}_{an} + \frac{1}{\mu} \ln M_a + \varepsilon_{an} \tag{44}$$

M_a is the number of paths using activity type a. $M_a = K^\tau - (K-1)^\tau$.

We believe that other correlation structure might be more realistic. In particular, we could consider specific patterns as sharing unobserved factors. The correlation of activity path is still an open question. Moreover, it seems that a deterministic correction for correlation is not adapted in our context. The correction must be estimated to confirm the different intuitions about correlation structures in activity choice.

3.3.4 Path utility

The path utility is the sum of the utilities of individual nodes $\mathcal{A}_{k,\tau}$:

$$V_\Gamma = \sum_\tau V(\mathcal{A}_{k,\tau}) \tag{45}$$

The utility $V(\mathcal{A}_{k,\tau})$ of a node $\mathcal{A}_{k,\tau}$ represents the individual utility from allocating time to a certain activity type. It does not include the trip (dis)utility (since there is no specified trip between two activity types). Trip utility is included in a submodel of destination choice knowing the sequence of activity types and schedules. The utility includes both the satiation effect and the time of day preference. Satiation effect represents the diminishing marginal utility with duration, $\eta_k \ln(t_k)$, where t_k the duration of activity type k and η_k a satiation parameter for activity type k (Ettema *et al.,* 2007). The time-of-day utility depends on both the activity type k and the time interval τ. It can be generally expressed as $\beta_{k,\tau} I_{k,\tau}$, where $I_{k,\tau}$ is a dummy variable and $\beta_{k,\tau}$ the

corresponding parameter. In practice, some β's might be statistically equal. The path utility is:

$$V(\mathcal{A}_{k,\tau}) = \eta_k \ln(t_k) + \sum_{k,\tau} \beta_{k,\tau} I_{k,\tau} \tag{46}$$

Parameters η_k and $\beta_{k,\tau}$ can be interacted with socioeconomic variables.

Finally, the deterministic part of the utility correcting for both the sampling of alternatives and the correlation due to the physical overlapping of paths in C_{in} is:

$$V_{\Gamma n} = \eta_k \ln(t_k) + \sum_{k,\tau} \beta_{k,\tau} I_{k,\tau} + \ln \frac{k_{\Gamma n}}{b(\Gamma)} + \beta_{PS} \ln APS_{\Gamma} \tag{47}$$

4 Pedestrian case study on EPFL campus

EPFL campus approximately hosts 13'000 people per day. Similarly to transport hubs, some of its users follow schedules (classes instead of train, bus or plane schedules). The activities are diverse. People very often spend a day on campus and work, study or eat during the day.

4.1 Data source

We collected network traces as defined in Danalet *et al.* (2013). Campus users authenticate themselves on the WiFi network through WPA using a Radius server. Accounting is one of the process on the Radius server. It allows to associate a MAC address with a username.

For visitors from University of Lausanne (UNIL), the user name is unique for all users (`unil.ch`). For members of the campus, the username was associated to employee or class attribute through LDAP requests. First, 317 employees were randomly selected, 501 students were selected from 6 random classes[1] and 729 UNIL students visiting EPFL campus. For a party on campus, Vivapoly (`http://vivapoly.epfl.ch/`), the lists of employees and students were modified. 4493 employees were selected, corresponding to all employees who recently connected to the WiFi, and similarly all students in the selected classes, resulting in 298 IDs. After the party, the

[1] Life Science, 1st year students (SV-BA2), Computer Science, 2nd year (IN-BA4), Civil Engineering, 2nd year (GC-BA4), Mathematics, 1st year (MA-BA2), Computer Science, Master students (IN-MA2), Physics, 1st year students (PH-BA2).

lists for students and visitors remained the same, but the list of employees returned to its original 317 IDs. The output of this process is anonymized network traces with known category of users on campus.

Data were grouped in text files on 11 different days[2], and were collected every day at a fixed time. Each time, historical data were collected individually from the CWS system of the Cisco Context Aware Mobility API with the Cisco Mobility Services Engine (MSE) Cisco (2011) that was lent to us. It was impossible to extract all data at once. This was done sequentially, MAC address by MAC address. It resulted in 2'392'973 network traces. For some users and some days, the 11 data collection campaigns would overlap over time. We cleaned the data to avoid broken daily traces related to the time of grouping in text files: if the last signal of a text file for a given individual does not appear in the next text file for the same user, we delete all signals related to this day. Similarly, for the very first signal of the first text files, we remove all signals till 3am. This way, we ensure to have complete days, and no partial sequences of signals for a day. Table 2 shows the number of signals, IDs and days for raw data (without duplicates), and for cleaned data (without partial days).

Table 2: Number of network traces

	Employees	Students						Visitors
		SV-BA2	IN-BA4	GC-BA4	MA-BA2	IN-MA2	PH-BA2	UNIL
Raw data	963'294	204'516	111'199	164'203	178'339	127'911	162'954	446'344
# IDs	4140	221	125	153	168	190	176	729
# days	51	51	50	51	47	49	50	52
Clean data		203'713	110'432	162'583	177'159	127'198	161'930	
# IDs		209	114	138	152	178	158	
# days		51	50	51	47	49	50	

We applied the algorithm described in Danalet *et al.* (2014) to all measurements related to students with $L = 1$. The potential attractivity measure as defined in Danalet *et al.* (2014) is based on different data sources:

- for offices, the attractivity is based on the cumulative work percentage of the different employees of a given office, e.g., if there are 2 full time employees and one working 80% of a full time, the attractivity is defined as 2.8. No schedule is applied on offices, so this attractivity is the same all day long.
- for classrooms with classes, the attractivity equals to the number of registered students to this class. The attractivity is valid only between the start and end time of the class. For the

[2]May 23, 2012, 18:58:36; May 24, 17:10:33; May 25, 16:45:17; May 31, 10:42:24; June 8, 8:42:06; June 14, 13:06:45; June 21, 11:37:40; June 27, 10:20:43; July 2, 10:58:31; July 4, 16:31:49; July 5, 10:33:33.

language classes, the number of registered students is not known but the language center provided an estimation of 13 students per class.
- for the library and the multimedia room of the language center, the attractivity equals to the capacity, i.e., the number of seats, during opening hours.
- for the restaurants, the point-of-sale data are aggregated per quarter of hour and used as attractivity. Note that this is different than in Danalet *et al.* (2014), where the attractivity for restaurants was their capacity, i.e., their number of seats.

The attractivity and time constraints for other points of interest are not available and have to be imputed by the modeler:

- the attractivity of an office for a student must be non-zero and has been fixed to 0.1.
- the attractiviy of classrooms with classes for employees has been fixed to 2 during the class.
- the attractivity of conference rooms has been fixed to 3.
- the attractivity of the post office has been fixed to 13 during opening hours and 3 when only the ATM is available.
- the attractivity of the student union has been fixed to 3.
- the attractivity of all other points of interest has been fixed to 1, by default.

4.2 Activity pattern, activity network and activity path

In Figure 12, an activity pattern is presented as described in Section 3.1.1. This activity pattern is based on real data of one student on campus.

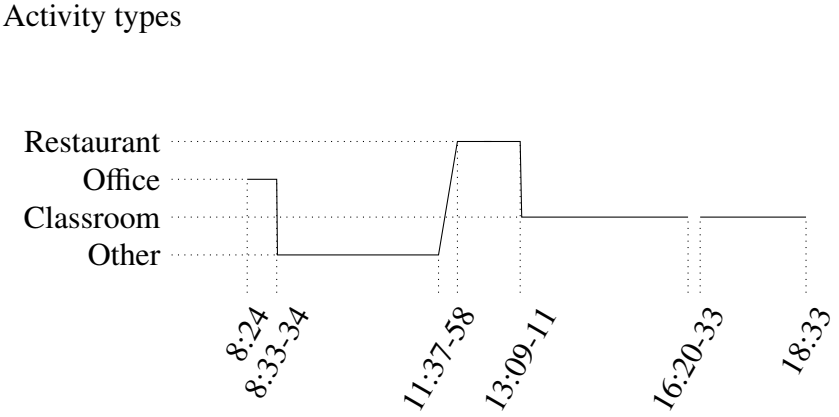

Figure 12: Activity pattern for a student on campus.

As defined in Section 3.1.2, a portion of the activity network on campus is shown in Figure 13. The full activity network covers a complete day. There are seven activity types: classrooms,

shops, offices, restaurant, library, lab and other. The types "Office", "Classroom" and "Lab" are based on norm DIN 277 defined by the *Deutsches Intitut für Normung*. The types "Shops", "Restaurant", "Library" and "Other" are extracted from a list of points of interest from `http://map.epfl.ch`.

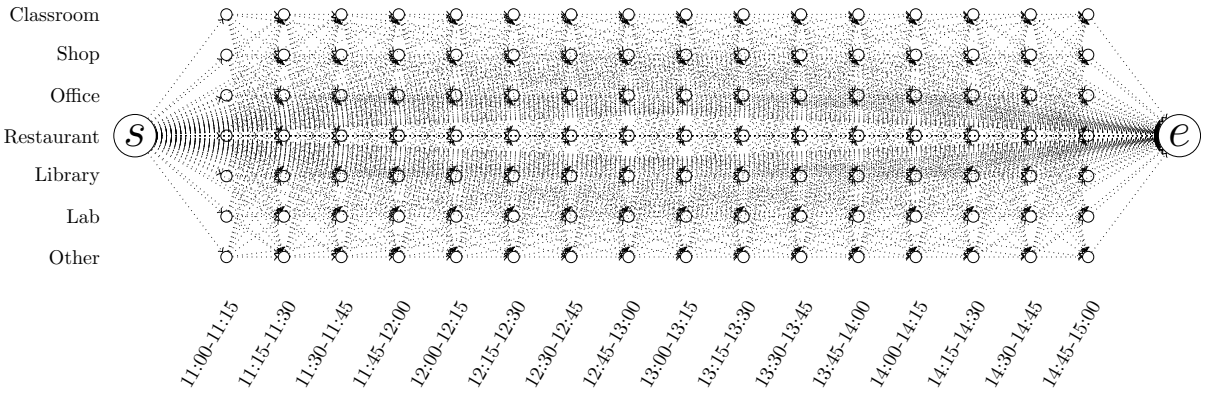

Figure 13: Activity network on campus.

The activity pattern from Figure 12 is expressed as an activity path in the activity network in Figure 14.

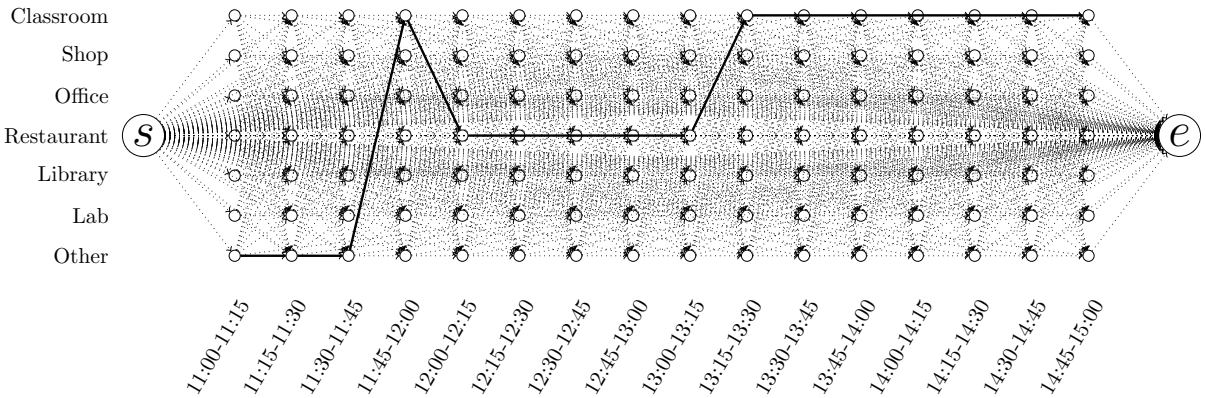

Figure 14: Activity path for a student on campus.

4.3 Choice set

The potential attractivity measure as defined in Section 3.2.1 is based on different data sources described in Section 4.1. Figure 15 shows the cumulated potential attractivity measure per quarter of hour for employees over one week. Lunch break is visible.

Figure 16 shows the importance of each activity type in time based on activity paths from WiFi data for employees on campus. Lunch break is visible. Offices are the most visited activity type.

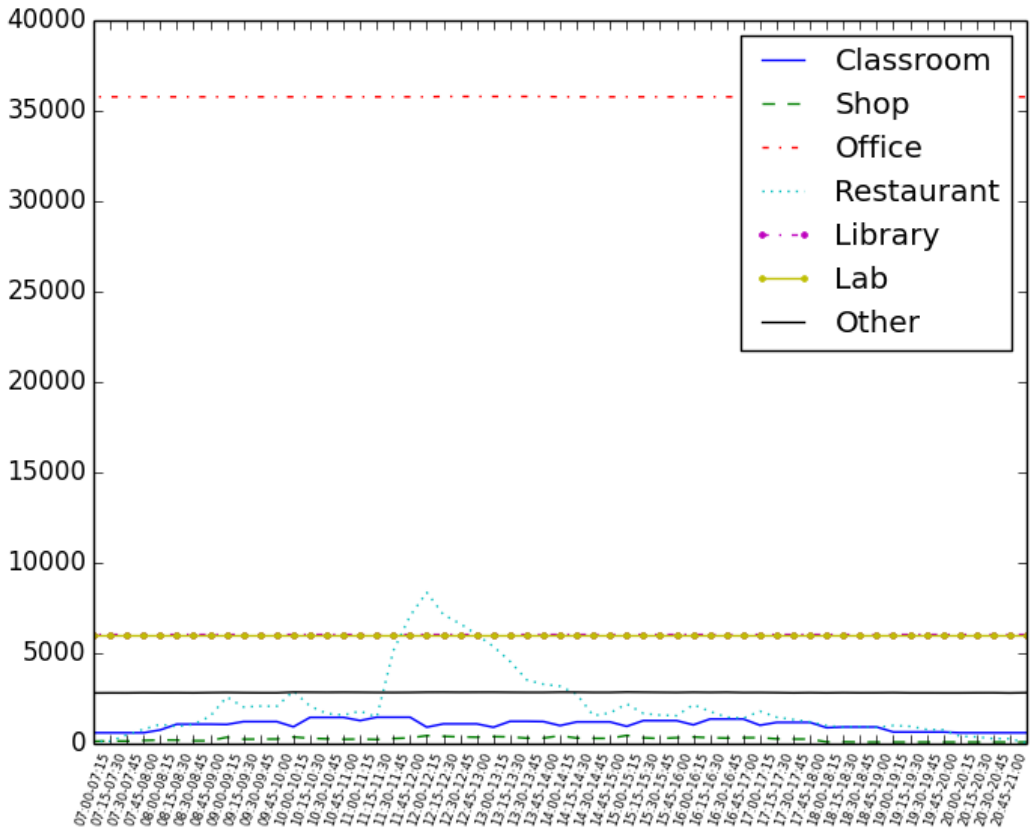

Figure 15: The cumulated potential attractivity measure per quarter of hour for employees. Aggregation is over one week. Y-axis represents the attractivity, expressed as the number of people.

We use this second data, the output of the processing of WiFi measurements, to generate a choice set using the methodology presented in Section 3.2.2. Figure 17 shows the weight of each node in the activity network based on the number of activity paths using each node. This figure is a representation of Figure 16 in the activity network.

Figure 18 shows the length of activity paths for employees. A group of observations have a duration of 8 to 9 hours, which is realistic. Some observations have a short length, about 1 hour, or a very long one, up to 24 hours. These observations are probably due to the source of data, i.e., people turning their device off in office or fix devices.

Results from a mobility survey for 2012 show that the shortest activity sequence for students and employees on campus over 2116 observations is 2h30 per day (Tzieropoulos, 2012). We remove all observations from WiFi that represent less than 2h30. Shorter observations from WiFi are probably due to devices that are turned off part of the day, such as laptops. Results of this filter are presented in Figure 19. This figure shows that activity sequences from WiFi are globally shorter than declared activity sequences from the mobility survey. This might be related

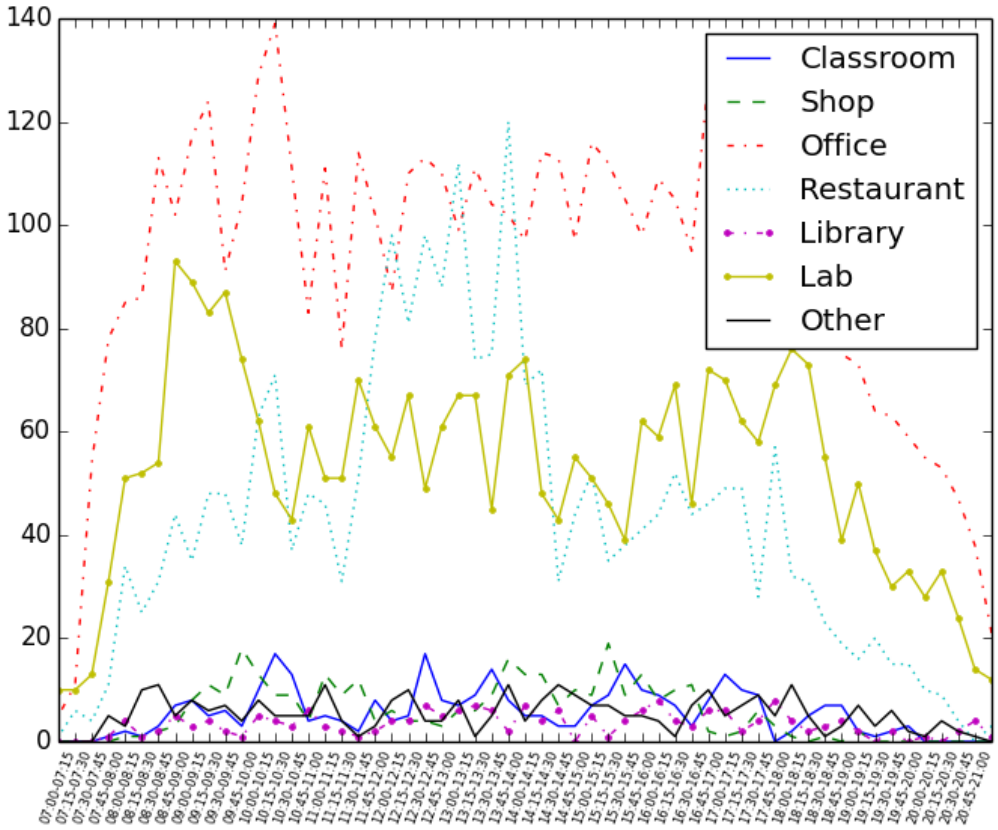

Figure 16: The cumulated number of employees per quarter of hour based on the activity path from activity-episode sequences generated from WiFi measurements. Y-axis represents the number of people.

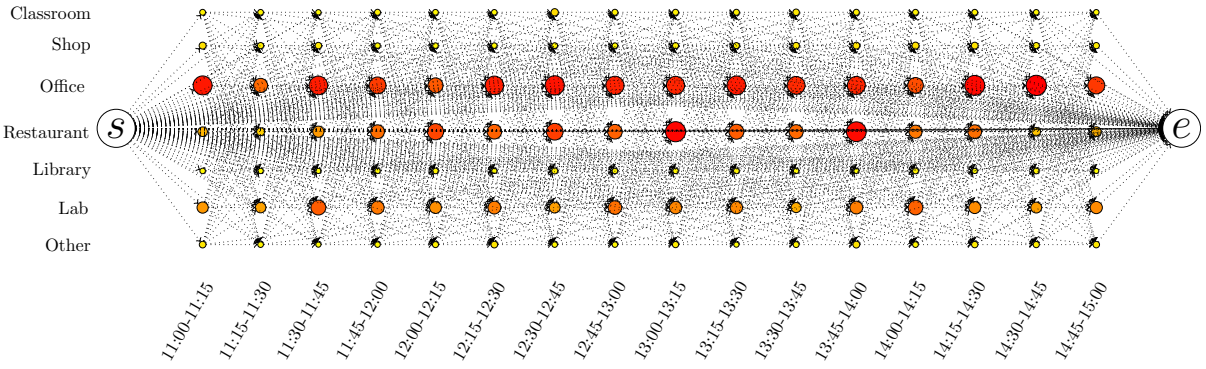

Figure 17: The distribution of activity type and quarter of hour for employees, based on the activity path from activity-episode sequences generated from WiFi measurements. It represents the data of Figure 16 in the (partial) activity network.

to devices with low density of measurements, such as smartphones not regularly connecting to the WiFi.

Danalet *et al.* (2014) conclude that a density of 5.4 measurements per hour is the minimum

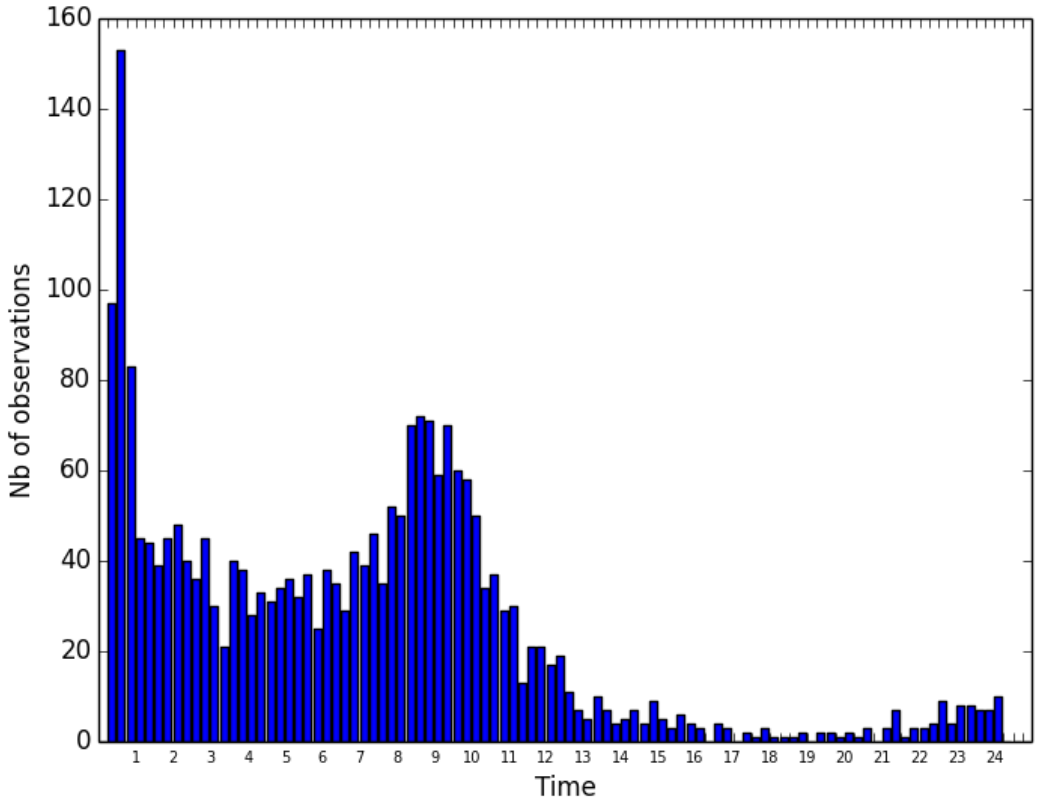

Figure 18: The distribution of length of activity path from activity-episode sequences generated from WiFi measurements. Y-axis represents the number of activity path with a given length.

density of measurements for generating stable and trustworthy activity-episode sequences with respect to the number of episodes, their activity type, their duration and their exact location. We remove here activity paths with a lower mean density. Results are shown in Figure 20. We observe that even with dense measurements, the activity sequences from WiFi are still shorter than declared activity sequences from the mobility survey. Respondents of the mobility survey might have declared longer activity sequences than what they really did. Cognitively incongruent answers are common for declared preferences (Bertrand and Mullainathan, 2001). Students and employees want to show a work-intensive day in their answer to the survey. It might also be that the respondents were asked to answer about a "typical day" in the recent past.

Figure 21 shows when people are present on campus during the day. People arrive between 7 and 10 am and leave starting from 4 pm up to midnight. They arrive earlier based on the mobility survey than based on WiFi measurements.

1200 paths were generated.

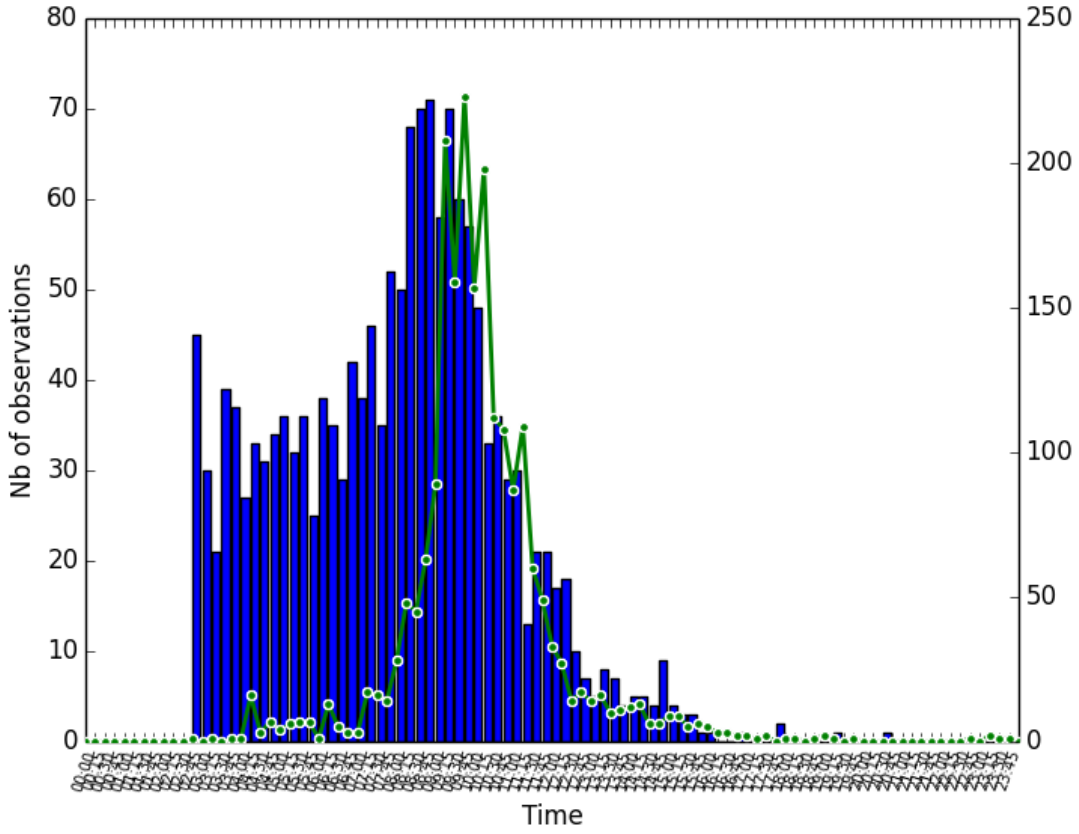

Figure 19: The distribution of length of activity paths from activity-episode sequences generated from WiFi measurements is represented as blue bars. The left Y-axis represents the number of activity path with a given length from WiFi measurements. The distribution of length from (Tzieropoulos, 2012) is represented as dots and green lines. The right Y-axis represents the number of individuals with a given length from (Tzieropoulos, 2012). Activity sequences of less than 2h30 are removed.

4.4 Choice model

We estimate a very simple model. It is a logit model, with no correction for correlation. The utility function is:

$$V_{\Gamma n} = \eta_k \ln(t_k) + \sum_{k,\tau} \beta_k I_{k,\tau} + \ln \frac{k_{\Gamma n}}{b(\Gamma)} \tag{48}$$

It corrects for importance sampling and consider only one β_k per activity type.

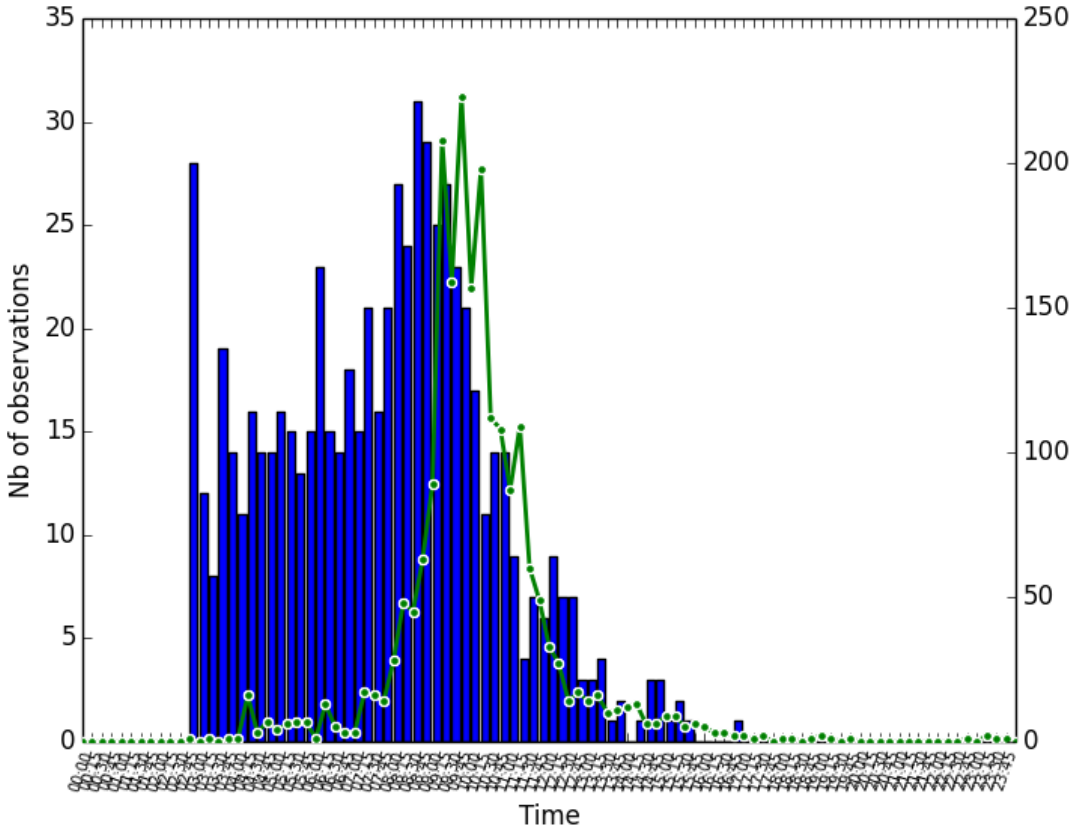

Figure 20: The distribution of length of activity paths from activity-episode sequences generated from WiFi measurements is represented as blue bars. The left Y-axis represents the number of activity path with a given length from WiFi measurements. The distribution of length from (Tzieropoulos, 2012) is represented as dots and green lines. The right Y-axis represents the number of individuals with a given length from (Tzieropoulos, 2012). Activity sequences of less than 2h30 are removed. Activity sequences with a mean density of measurements lower than 5.4 are removed.

4.5 Estimation results

Results are shown in Table 3. $\beta_{\text{Classroom}}$ and β_{Office} are not significantly different from zero and have been removed from the model. Satiation parameters for "Office" η_{Office} is in the model and not significant. It means that the duration of episodes has no impact on their attractivity for this activity.

Offices, shops, the library and classrooms have a negative satiation parameter. The more time is spent at these activities, the less utility is collected for employees. On the other hand, spending more time in labs, restaurants or other points of interest increases their utility. Employees like to avoid the library and other points of interest the most during their day and generally have a negative tendency toward restaurants, the library, labs and other points of interest.

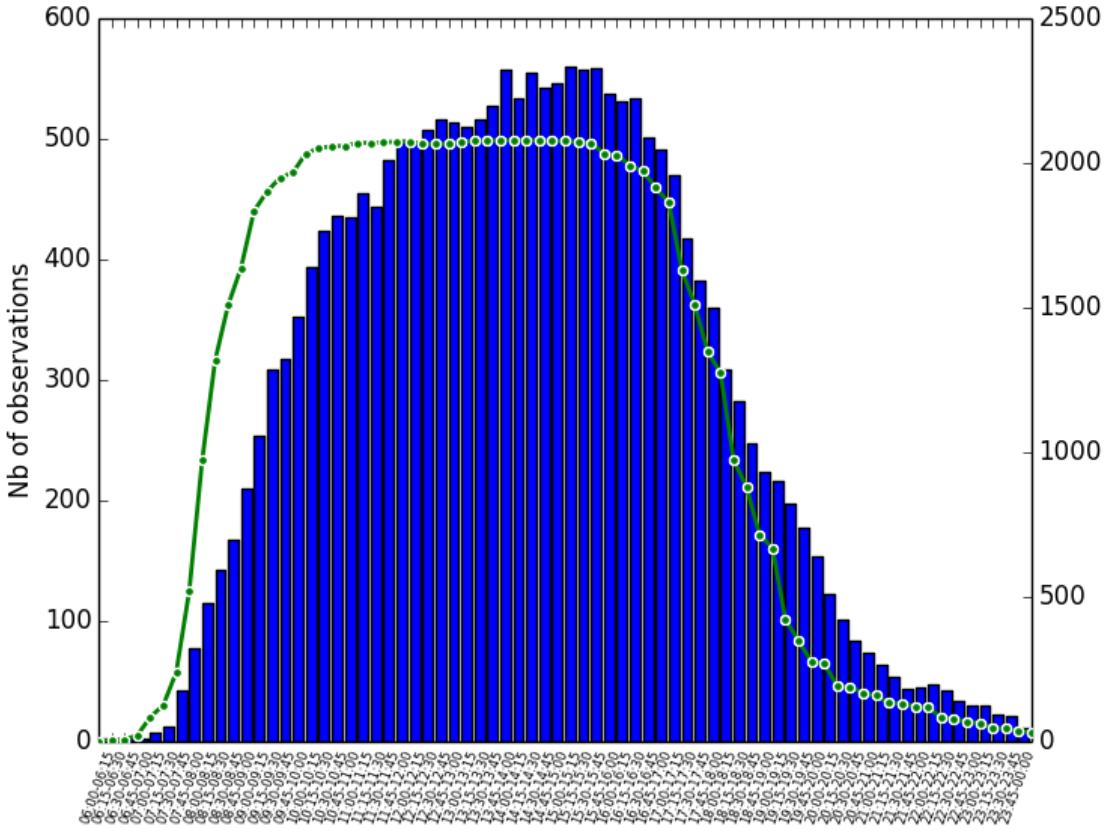

Figure 21: The time-of-day distribution of activity paths from activity-episode sequences generated from WiFi measurements is represented as blue bars. The left Y-axis represents the number of activity path present at this quarter of an hour from WiFi measurements. The time-of-day distribution from (Tzieropoulos, 2012) is represented as dots and green lines. The right Y-axis represents the number of individuals at a given quarter of an hour from (Tzieropoulos, 2012). Activity sequences of less than 2h30 are removed. Activity sequences with a mean density of measurements lower than 5.4 are removed.

More complex models must be implemented. Different betas for different periods of day could be defined, and similarly different satiation parameters could be estimated. For example, the attractivity and satiation are probably not the same in restaurants during lunch break and during the rest of the day.

5 Conclusion and future work

Our approach allows to evaluate and forecast the effects of different policies and designs of pedestrian facilities. It models the choice of activity sequences of individuals and evaluate the attractivity of the different points of interest. Our model is not home-based, nor tour-based. It

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	β_{Lab}	-0.337	0.0949	-3.55	0.00
2	β_{Library}	-2.74	0.0795	-34.45	0.00
3	β_{Other}	-2.78	0.0483	-57.62	0.00
4	$\beta_{\text{Restaurant}}$	-0.725	0.0612	-11.85	0.00
5	β_{Shop}	-0.473	0.103	-4.59	0.00
6	η_{Lab}	2.55	0.895	2.85	0.00
7	η_{Office}	-0.787	0.600	-1.31	0.19
8	η_{Other}	9.66	1.27	7.63	0.00
9	$\eta_{\text{Restaurant}}$	5.56	0.789	7.05	0.00
10	$\eta_{\text{Shop, Library, Classroom}}$	-3.26	0.782	-4.16	0.00

Summary statistics

Number of observations = 2219

Number of estimated parameters = 10

$$\begin{aligned} \mathcal{L}(\beta_0) &= -17952.561 \\ \mathcal{L}(\hat{\beta}) &= -1484.635 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 32935.852 \\ \rho^2 &= 0.917 \\ \bar{\rho}^2 &= 0.917 \end{aligned}$$

Table 3: Estimation results for the model.

can adapt to different contexts and activity types. The large dimensionality of the problem is managed through importance sampling techniques.

The approach presented in this report needs to be associated to a destination choice model conditional on the activity type. A logit model could be used, similar to Ton (2014).

Preliminary results show that the approach is feasible in a realistic context. The model can largely be improved by specification, adding more details and different variables. An important feature of our approach is that it allows to add in the utility function variables that are not specific to activity episodes, but are related to the path itself. Patterns can be included in the utility function and their specific utility or satiation parameters can be estimated.

An important next step in this work is the evaluation of the quality of the generated choice set and its impact on the choice model.

6 References

Bekhor, S., Y. Cohen and C. Solomon (2013) Evaluating long-distance travel patterns in Israel by tracking cellular phone positions, *Journal of Advanced Transportation*, **47** (4) 435–446, ISSN 01976729.

Ben-Akiva, M. and M. Bierlaire (1999) Discrete choice methods and their applications to short-term travel decisions, in R. Hall (ed.) *Handbook of Transportation Science*, Operations Research and Management Science, 5–34, Kluwer, Netherlands.

Ben-Akiva, M. and M. Bierlaire (2003) Discrete choice models with applications to departure time and route choice, paper presented at the *Handbook of Transportation Science, Second Edition*, 7–37, Boston/Dordrecht/London.

Ben-Akiva, M. E. and S. R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Ma.

Bertrand, M. and S. Mullainathan (2001) Do People Mean What They Say? Implications for Subjective Survey Data, *American Economic Review*, **91** (2) 67–72, ISSN 0002-8282.

Bhat, C. and F. Koppelman (1999) Activity-Based Modeling of Travel Demand, *Handbook of transportation Science*.

Bierlaire, M. and E. Frejinger (2008) Route choice modeling with network-free data, *Transportation Research Part C*, **16** (2) 187–198, ISSN 0968090X.

Bierlaire, M. and T. Robin (2009) Pedestrians Choices, paper presented at the *Pedestrian Behavior. Models, Data Collection and Applications*, 1–26, ISBN 978-1-84855-750-5.

Bowman, J. L. (1998) The Day Activity Schedule Approach to Travel Demand Analysis, Ph.D. Thesis, Massachusetts Institute of Technology.

Bowman, J. L. (2009) Historical Development of Activity Based Model Theory and Practice, *Traffic Engineering and Control*, **50** (7) 314–318.

Buisson, A. (2014) Individual activity-travel analysis based on smartphone WiFi data, Master thesis, EPFL.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira and C. Ratti (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C*, **26** (0) 301–313, ISSN 0968090X.

Chen, J. (2013) Modeling route choice behavior using smartphone data, Ph.D. Thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Cisco (2011) Cisco MSE API Specification Guide - Location Service, Release 7.1, *Technical Report*.

Danalet, A., B. Farooq and M. Bierlaire (2013) A Bayesian Approach to Detect Pedestrian Destination-Sequences from WiFi Signatures, *Technical Report*, Transport and Mobility Laboratory, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausanne.

Danalet, A., B. Farooq and M. Bierlaire (2014) A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures, *Transportation Research Part C*.

Davidson, W., R. Donnelly, P. Vovsha, J. Freedman, S. Ruegg, J. Hicks, J. Castiglione and R. Picado (2007) Synthesis of first practices and operational research approaches in activity-based travel demand modeling, *Transportation Research Part A: Policy and Practice*, **41** (5) 464–488, ISSN 09658564.

Ettema, D. (1996) Activity based Travel Demand Modelling, Ph.D. Thesis, Eindhoven Technical University, Holland.

Ettema, D., F. Bastin, J. Polak and O. Ashiru (2007) Modelling the joint choice of activity timing and duration, *Transportation Research Part A*, **41** (9) 827–841, ISSN 09658564.

Etter, V., M. Kafsi and E. Kazemi (2012) Been There , Done That: What Your Mobility Traces Reveal about Your Behavior, paper presented at the *Nokia Mobile Data Challenge 2012 Workshop*, 1–6, June 18-19, Newcastle, UK.

Feil, M. (2010) Choosing the Daily Schedule: Expanding Activity-Based Travel Demand Modelling, Ph.D. Thesis, ETH.

Flötteröd, G. and M. Bierlaire (2013) Metropolis–Hastings sampling of paths, *Transportation Research Part B*, **48**, 53–66, ISSN 01912615.

Frejinger, E. and M. Bierlaire (2007) Capturing correlation with subnetworks in route choice models, *Transportation Research Part B*, **41** (3) 363–378, ISSN 01912615.

Frejinger, E. and M. Bierlaire (2010) On Path Generation Algorithms for Route Choice Models, in S. H. Daly and A. (eds.) *Choice Modelling: The State-of-the-Art and the State-of-Practice*, 307–315, Emerald Group Publishing Limited, ISBN 978-1-84950-772-1.

Frejinger, E., M. Bierlaire and M. Ben-Akiva (2009) Sampling of Alternatives for Route Choice Modeling, *Transportation Research Part B*, **43** (10) 984–994.

Frignani, M. Z., J. Auld, A. K. Mohammadian, C. Williams and P. Nelson (2010) Urban Travel Route and Activity Choice Survey (UTRACS): An Internet-Based Prompted Recall Activity Travel Survey using GPS Data, *Transportation Research Record: Journal of the Transportation Research Board*, **2183**, 19–28.

Habib, K. M. N. (2007) Modelling activity generation processes, Ph.D. Thesis, University of Toronto.

Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira, E. Frazzoli and M. C. González (2013) A review of urban computing for mobile phone traces, paper presented at the *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, New York, New York, USA, ISBN 9781450323314.

Kalakou, S. and F. Moura (2014) Effects of terminal planning on passenger choices, paper presented at the *14th Swiss Transport Research Conference (STRC)*, Monte Verità, Ascona, Switzerland.

Liu, X. (2013) Activity-based pedestrian behavior simulation inside intermodal facilities, Ph.D. Thesis, Mississippi State University.

Miller, H. J. (2014) Activity-Based Analysis, in M. M. Fischer and P. Nijkamp (eds.) *Handbook of Regional Science*, 705–724, Springer Berlin Heidelberg, Berlin, Heidelberg.

Pinjari, A. R. and C. R. Bhat (2011) Activity Based Travel Demand Analysis, in A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds.) *A Handbook of Transport Economics*, chap. 10, 213–248, Edward Elgar Publishing Ltd.

Prato, C. and S. Bekhor (2006) Applying Branch-and-Bound Technique to Route Choice Set Generation, *Transportation Research Record*, **1985** (1) 19–28, ISSN 0361-1981.

Razemon, O. (2013) A Anvers, sous le fleuve les piétons..., *Géomètre*.

Rieser-Schüssler, N. (2012) Capitalising modern data sources for observing and modelling transport behaviour, *Transportation Letters: The International Journal of Transportation Research*, **4** (2) 115–128, ISSN 1942-7867.

Rindfüser, G., H. Mühlhans, S. T. Doherty and K. J. Beackmann (2003) Tracing the planning and execution of activities and their attributes: Design and application of a hand-held scheduling process survey, paper presented at the *10th International Conference on Travel Behaviour Research*, 1–31, August 10-14, Lucerne, Switzerland.

Roorda, M. J. (2005) Activity-based modelling of household travel, Ph.D. Thesis, University of Toronto.

Timmermans, H., X. V. der Hagen and A. Borgers (1992) Transportation systems, retail environments and pedestrian trip chaining behaviour: modelling issues and applications, *Transportation Research Part B*, **26B** (1) 45–59.

Ton, D. (2014) NAVISTATION: a study into the route and activity location choice behaviour of departing pedestrians in train stations, Master thesis, Delft University of Technology.

Tzieropoulos, P. (2012) Mobility Observatory, `http://litep.epfl.ch/page-15350-en.html`.

van der Zijpp, N. and S. Fiorenzo Catalano (2005) Path enumeration by finding the constrained K-shortest paths, *Transportation Research Part B: Methodological*, **39** (6) 545–563, ISSN 01912615.

Van Nes, R., S. Hoogendoorn-Lanser and F. S. Koppelman (2008) Using Choice Sets for Estimation and Prediction in Route Choice, *Transportmetrica*, **4** (2) 83–96, ISSN 1812-8602.