Modeling Human Dynamics and Lifestyles using Digital Traces

by

Sharon Xu

B.S. Statistics, University of California, Los Angeles

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author ....

Sharon Xu
Sloan School of Management
May 18, 2018

Signature redacted

Certified by ....

Marta C. González
Visiting Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .... Signature redacted ....

Patrick Jaillet
Dugald C. Jackson Professor, Department of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

THIS PAGE INTENTIONALLY LEFT BLANK

Modeling Human Dynamics and Lifestyles using Digital Traces

by

Sharon Xu

Submitted to the Sloan School of Management
on May 17, 2018 in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

In this thesis, we present algorithms to model and identify shared patterns in human activity with respect to three applications.

First, we propose a novel model to characterize the bursty dynamics found in human activity. This model couples excitation from past events with weekly periodicity and circadian rhythms, giving the first descriptive understanding of mechanisms underlying human behavior. The proposed model infers directly from event sequences both the transition rates between tasks as well as nonhomogeneous rates depending on daily and weekly cycles. We focus on credit card transactions to test the model, and find it performs well in prediction and is a good statistical fit for individuals.

Second, using credit card transactions, we identify lifestyles in urban regions and add temporal context to behavioral patterns. We find that these lifestyles not only correspond to demographics, but also have a clear signal with one's social network.

Third, we analyze household load profiles for segmentation based on energy consumption, focusing on capturing peak times and overall magnitude of consumption. We propose novel metrics to measure the representative accuracy of centroids, and propose a method that outperforms standard and state of the art baselines with respect to these metrics. In addition, we show that this method is able to separate consumers well based on their solar PV and storage needs, thus helping consumers understand their needs and assisting utilities in making good recommendations.

Thesis Supervisor: Marta C. González
Visiting Associate Professor of Civil and Environmental Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

First and foremost, I thank my advisor, Professor Marta C. González. When the methods were unclear and the data had no signal, she offered support and guidance, while still allowing me to find my own way and grow as a researcher. She has taught me not only how to listen to the story the data is telling, but also how to then go on and tell it myself. I will forever be grateful to have had the opportunity to join her lab and learn about this field.

Thank you to my friends and fellow students in the Operations Research Center and the Human Mobility and Networks Lab. I've learned so much from you in the past two years—you've made my time here truly memorable. I would specifically like to thank Phil Chodrow, Edward Barbour, Riccardo DiClemente, and Steven Morse giving guidance on research and being a sounding board for ideas. A special thanks to Ted Grunberg for being there to debug all the little problems in life. These past few years you've shown me that I really can have my cake and eat it too.

And finally, I would like to thank my family—Yanqi Xu, Xiufei Ye, and Amanda Xu—for their unconditional support throughout my journey here. Your help and encouragement made this path possible for me.

Contents

Introduction 1

Within the last decade, the digital age has sharply redefined the way we study human behavior. With the advancement of data storage and sensing technologies, electronic records now encompass a diverse spectrum of human activity, ranging from location data [65, 30], phone [36, 4] and email communication [48] to Twitter activity [67] and open-source contributions on Wikipedia and OpenStreetMap [76, 75].

In particular, the rising ubiquity of passively collected data allows for new opportunities to understand the behavior of individuals on a granular time scale. In this thesis, we use passively collected data to model human activity in three applications:

Shopping Activity. The shopping patterns of individuals have the potential to give deeper insight into lifestyles and communities. The main work studying credit card records (CCRs) has centered around measuring similarity in purchases through affinity algorithms [56, 64]. Recently they have also been used to connect transaction types with metrics of social and mobile activity [19]. Through these records, individuals have been found to have inherent regularity in shopping patterns [39], indicating a promising avenue for models of shopping behavior.

Mobility. With regards to mobile computing, the pervasive use of cellular phones has generated a wealth of data. Call detail records (CDRs) document the social activity and mobility of their users with high temporal resolution, presenting new opportunities to understand human mobility [27], analyze wealth [12], and model social network dynamics [51]. Regarding the analysis of CDR data, there exists a wide body of work characterizing human mobility patterns. As a notable example, [27] describes the temporal and spatial regularity of human trajectories, showing that each individual can be described by a time independent travel distance and a high probability of returning to a small number of locations. Further, the authors are able to model individual travel patterns using a single spatial probability distribution.

Energy. Through the large-scale deployment of advanced metering infrastructure (AMI), companies have access to high granularity smart meter data with detailed records of energy consumption in hourly or sub-hourly intervals. Due to the scale of this smart meter data, it has the potential to offer significant benefits in terms of power system operation, emissions reduction and monetary rewards for utilities

[23]. As such, categorizing households in terms of their consumption behavior is a problem of interest to both companies and policymakers alike.

In this thesis, we focus on designing descriptive models and algorithms for such passively collected data. Throughout applications in shopping, mobility, and energy, we propose interpretable algorithms that can be used not only for prediction and recommendation, but also to understand the mechanisms underlying how and why individuals generate this passive data.

1.1 Modeling the dynamics of human behavior

Human actions drive a range of complex social, urban, and economic systems, yet understanding the patterns and dynamics of human behavior is still an open question in modern-day science. As can be seen from mobility traces, credit card transactions, and communications data, such behavior is inherently non-Poissonian and tends to exhibit bursts of activity throughout time, or alternating periods of high and low activity. As shown in Fig. 1.1), this leads to a power-law scaling on the inter-event time distribution, or the distribution on the time between consecutive events.

Current research in this area tends to hypothesize that this burstiness is well described by a mechanism ranging from task prioritization to circadian rhythms. It then proceeds to construct a model based on that hypothesis and test for statistical goodness-of-fit. This class of models includes:

- Queuing process models of waiting times. Shown in Fig. 1.1, these models attribute burstiness to the execution of tasks based on their priority, with some assuming finite memory [18, 7]. Known as priority list models, they describe the waiting time of a task, or the time period before a task is executed. This depends on the cumulative time needed to perform all tasks before it, and accordingly this model leads to heavy tails in the waiting time distribution [3, 70]. This has been found to agree well with empirical observations [54].

- Memory-driven models that depend on past events. The main contributions of this type are centered around a one-dimensional self-exciting process known as a Hawkes process. The memoryless property of the Poisson process means that it is unable to capture a dependence on history; however, we would like the event of an arrival to increase the probability of arrivals in the next small interval of time. In this stochastic process, the memory kernel parametrizes the increase in the arrival rate incurred by each event. [49] uses a kernel with exponential decay, while in [37] a power law kernel is used to mimic a sequential memory loss mechanism. In Fig. 1.2, we can see temporal clustering that results from this self-excitement.

- Poissonian models. [31] shows that a power law scaling on inter-event time distributions can be achieved using Poissonian agents with varying rates. At the individual level, [48] argues that such distinctly non-homogeneous event sequences are due solely to circadian rhythms, proposing a non-homogeneous Poissonian model for cascades of Poissonian activity depending on hour and day of week. This model is shown in Fig. 1.3. Extensions of this model uses a Markov process with multiple states to modulate transitions between Poisson models with different rates, thus reflecting periods of high and low activity seen in human communications [35, 58]. Although these methods give a close approximation to observed data, it does so using a large set of parameters. The resulting method is not descriptive and gives no generative explanation for diversity in human dynamics [38]. In addition, research [36, 84] shown that even after removing periodic effects, signals remain bursty.

These are single-dimensional models; accordingly, they give no descriptive generative mechanism capturing the actual transitions between different types of activity. In contrast, in Chapter 1 we propose a model, the multidimensional periodic Hawkes process, that couples the excitation structure between human activity types with weekly cycles and circadian rhythms. This model explicitly characterizes the priority and order of specific activity types, describing the varied dimensions of human behavior as well as the interdependence between them.

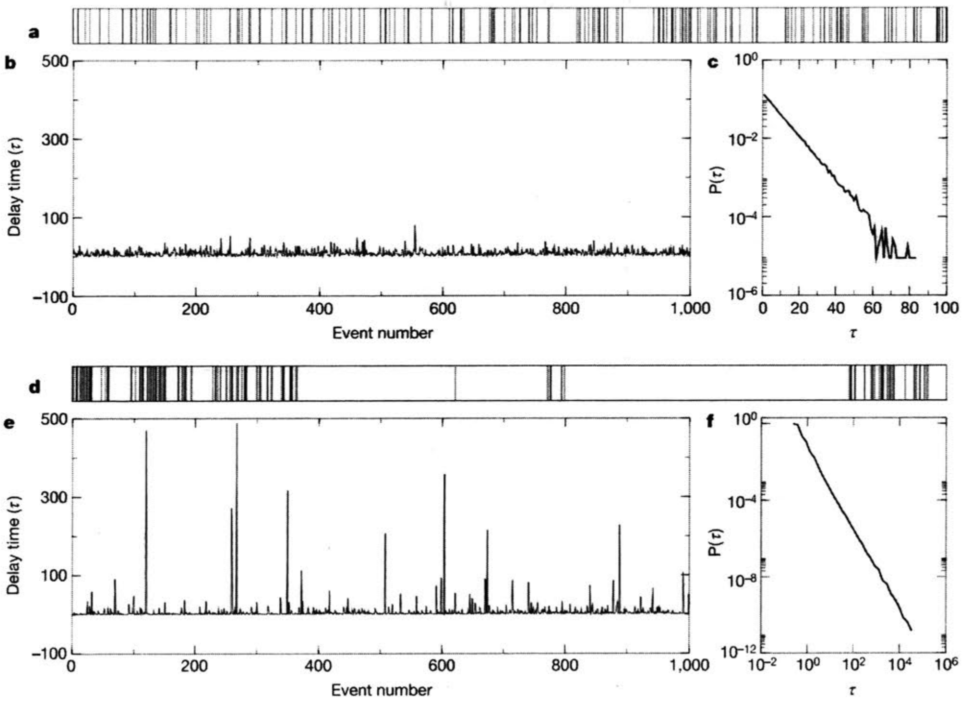


Figure 1.1: The difference between the activity patterns predicted by a Poisson process and the heavy-tailed distributions observed in human dynamics. **a**, Succession of events predicted by a Poisson process, which assumes that in any moment an event takes place with probability q. The horizontal axis denotes time, each vertical line corresponding to an individual event. Note that the inter-event times are comparable to each other, long delays being virtually absent. **b**, The absence of long delays is visible on the plot showing the delay times t for 1,000 consecutive events, the size of each vertical line corresponding to the gaps seen in **a**. **c** and **e** respectively show the inter-event time distribution corresponding to a Poisson process and a priority list model. We see heavy tails in the latter. **e**, The waiting time t of 1,000 consecutive events, where the mean event time was chosen to coincide with the mean event time of the Poisson process shown in a–c. Note the large spikes in the plot, corresponding to very long delay times. (*Source*: Adapted from [7] under Copyright (2005) Nature Publishing Group, United Kingdom.)

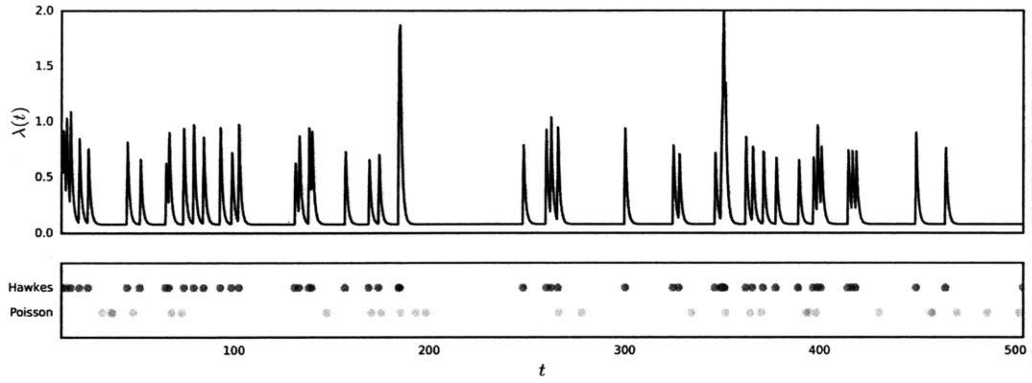


Figure 1.2: Comparison between a one-dimensional Poisson process and Hawkes process with exponential triggering kernel. Note the clear "burstiness" or temporal clustering apparent in event sequence generated by the Hawkes process, which are not present when using the Poisson process. Top figure shows the corresponding intensity function of the Hawkes process. (*Source*: Adapted from [52] under Copyright (2017) Massachusetts Institute of Technology, United States.)

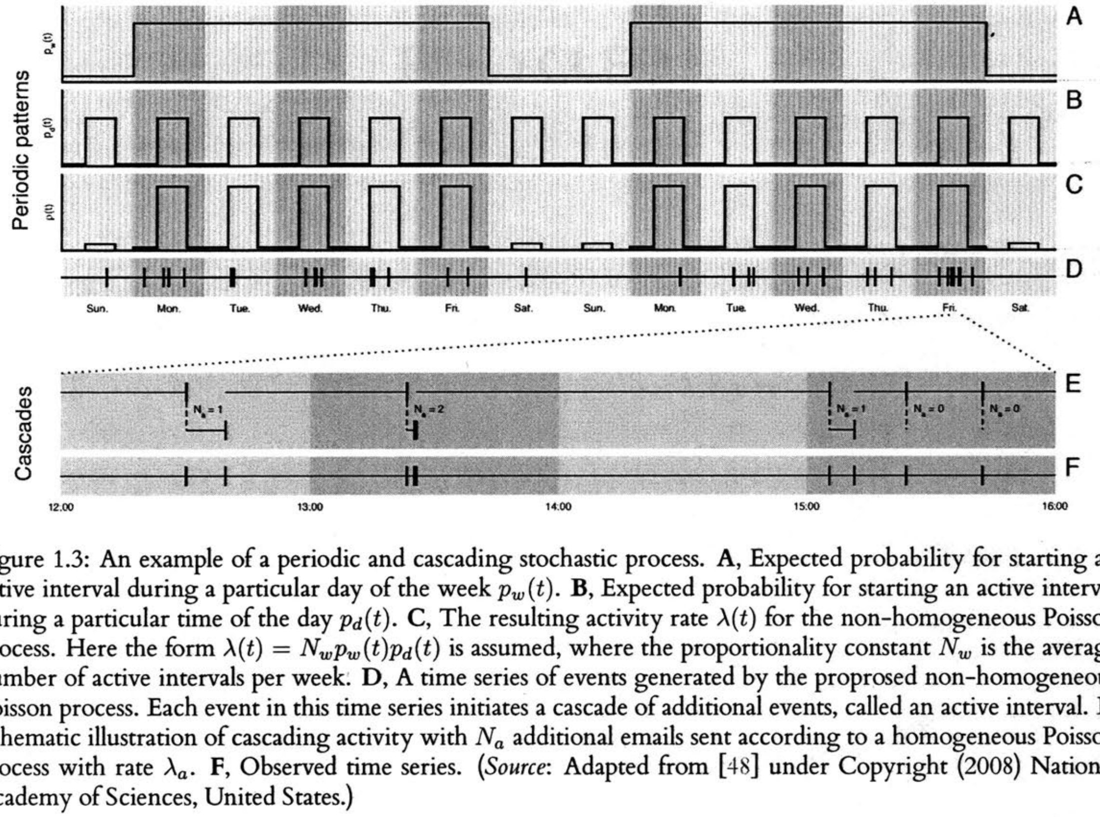


Figure 1.3: An example of a periodic and cascading stochastic process. **A**, Expected probability for starting an active interval during a particular day of the week $p_w(t)$. **B**, Expected probability for starting an active interval during a particular time of the day $p_d(t)$. **C**, The resulting activity rate $\lambda(t)$ for the non-homogeneous Poisson process. Here the form $\lambda(t) = N_w p_w(t) p_d(t)$ is assumed, where the proportionality constant N_w is the average number of active intervals per week. **D**, A time series of events generated by the proprosed non-homogeneous Poisson process. Each event in this time series initiates a cascade of additional events, called an active interval. **E**, Schematic illustration of cascading activity with N_a additional emails sent according to a homogeneous Poisson process with rate λ_a. **F**, Observed time series. (*Source*: Adapted from [48] under Copyright (2008) National Academy of Sciences, United States.)

1.2 Temporal lifestyles

Work at the intersection of different passive datasets includes the inference of friendships from call data records [20], or the analysis of credit card records in relation to metrics on spending behavior such as diversity, engagement, and loyalty [63]. Recent work [19] uses the Jaccard distance as a similarity measure on motifs among spending categories, then applies community detection algorithms to find communities of users that connect to social behavior and mobility.

In Chapter 2, we add temporal context to these communities, recovering definitive lifestyles in urban regions, and revealing consistent relationships between one's social network, shopping behavior, demographics, and expenditure. We describe these lifestyles temporally using the multidimensional periodic Hawkes process described in Chapter 1.

1.3 Dual view of lifestyles: perspectives from shopping and mobility

Coupled collaborative filtering methods, also known as collective matrix factorization methods, have been successfully applied in a variety of urban computing applications for data fusion and prediction [80, 82, 81], from location-based activity recommendations [79, 78] to travel speed estimation on road segments [61]. Recent work includes methods that use Laplacian regularization [14] to leverage social network information, as well as others that use geometric deep learning methods for matrix completion to model nonlinearities [13].

However, the only known paper that connects shopping and mobility behavior [32] frames its analysis only on an aggregate scale of city regions. Using collective matrix factorization (CMF) [62] methods, we relate the shopping and mobility patterns of consumers on an individual level for the first time. This analysis is also presented in [74]. We use credit card records and call data records, adding context to call records using points of interest (POIs). Our method results in an increase in prediction and recovers interesting relationships between shopping and mobility.

1.4 Consumer segmentation using smart meter data

The widespread global deployment of smart meters offers a unique opportunity for utilities to understand the energy use lifestyles and needs of their consumers. It also presents a new challenge: mining these massive datasets and translating fine-grained consumption data into meaningful insights for purposes of efficiency and interpretability.

Much of pre-existing literature on segmentation relies on self-reported information regarding attitudes and behavior, approaching the problem from the perspective of marketing or psychology fields [60], [66], [72]. Indeed, utility companies have been increasingly employing such psychographic segmentation strategies in the last decade [53]. However, actual energy usage data has been utilized much more rarely [1], [22], despite its ability to eliminate the assumptions and bias from self-reporting.

In an academic context, a significant amount of work has been done in the area of load forecasting and load profiling [21], [72], [57], [33]. Additionally, several recent works have made efforts toward consumer segmentation; these studies have mainly focused on clustering consumers by the shape of their load profiles. In [16], self-organizing maps were applied to the normalized load profiles for dimensionality reduction, and in [25], on the average load profile for each consumer. Both methods then applied k-means to cluster the low-dimensional load shape representations. In lieu of incorporating daily consumption into their clustering methodology, [25] attempts to account for consumption levels by manually separating winter, summer, workday and weekday loads. Recently, [40] and [41] focus on the shape of the normalized load profile as well, encoding daily time series by constructing a preprocessed dictionary on standardized data.

Prior works have indicated many applications for consumer segmentation. [8] uses electricity to infer socio-economic class, then suggests that utilities could use this information to target customers of a certain socio-economic status without survey information. [40] and [26] note that segments would be useful for designing customized tariffs. [29] focuses on producing ten robust clusters, such that there is a high degree of certainty for the cluster to which each consumer belongs. Recommended applications include distribution network planning, as well as the design of future trials so as to include representative customers. Despite the wide range of possible applications implied among these papers, no work has explicitly demonstrated a segmentation strategy on such an application.

Thus, in Chapter 4 we present work from [73] on the segmentation of consumer lifestyles based on energy consumption data, proposing a structured methodology that considers both the shape of the daily consumer load profile as well as its total energy consumption. We show that this distinction is critical to target recommendations of solar photo-voltaics (PV) and batteries for each consumer segment.

Modeling Bursty Human Dynamics 2

Bursts in activity characterize the dynamics of many natural and human-driven phenomena, from earthquakes and neural impulses to social systems, technological advances, and economic markets. In particular, the analysis of the "burstiness" of human behavior has been of great interest in the interdisciplinary sciences, encompassing a wide range of studies in human mobility, email and phone communications, purchase transactions, and other digital records. This burstiness, defined by alternating periods of high and low activity, has been shown to be a fundamental property of human dynamics [48, 70, 49, 37]. Previous literature focuses on characterizing the dynamics of a specific action of one type. These one-dimensional models do not consider multiple types of activity, nor their mutual influence. In contrast to past work, the proposed model captures the patterns and interdependencies between different activity types, lending insight into not only the temporal correlation between these types, but also the order in which individuals tend to participate in each activity.

Corresponding to past work, we consider two fundamental mechanisms which generate these interdependencies: rational decision-making, and periodicity:

Rational decision-making. Much of past work focuses on showing that task prioritization results in the burstiness observed in human dynamics. [7, 54, 3, 70]. Similarly to [49], we hypothesize that certain activities occur tend to excite each other, leading to short inter-event times, or the time between consecutive events. For example, a person running weekly errands will make many purchases in a short time period, a taxi ride may frequently result in a restaurant transaction and department store purchases, or regular payments for network, phone, and cable services may often be made together. In communication networks, similar patterns have been uncovered [52]. A call from mother to son may excite a call from son to father; an email from the manager may excite more communication between team members. Capturing the structure of these excitation patterns gives important insight into the priorities of individuals, extending current work, which only considers temporal behavior without the context of different activities.

Periodicity. Past studies have shown that human behavior tends to follow daily and weekly cycles, leading to heavy tails. Such patterns have been found throughout electronic records of activity ranging from location data [65, 30], phone [36, 4] and email communication [48] to Twitter activity [67] and open-source contributions on Wikipedia and OpenStreetMap [76, 75]. A number of factors contribute to this periodicity, including the day-night cycle, employment status, work schedules and commuting patterns [46, 47], and the activity of one's social contacts [68].

2.1 Model

Here we propose a model which capture these fundamental mechanisms, describing both the order and rate of transition between tasks, as well as fluctuations due to weekly cycles and circadian rhythms. Unlike previous models, we not only show statistical fit for a hypothesized mechanism, but also interpret human behavior with respect to that mechanism; that is, we actually quantify levels of periodicity and the transitions between specific types of activity.

The proposed model is based on the *Hawkes process*, a stochastic process which captures self-exciting behavior, allowing for temporal clustering or "burstiness" as opposed to the memoryless Poisson process. These stochastic processes can be defined by their conditional intensity. Let H_t be the past history and $N(t)$ be the number of events up to and including time t. Then the conditional intensity function is

$$\lambda(t_0) = \lim_{\varepsilon \downarrow 0} \frac{\mathbb{E}[N(t_0 + \varepsilon t) - N(t_0) | H(t_0)]}{\varepsilon t} \tag{2.1}$$

$$= \left. \frac{\partial \, \mathbb{E}[N(t) | H(t_0)]}{\partial t} \right|_{t=t_0} . \tag{2.2}$$

The conditional intensity is naturally understood as the infinitesimal arrival rate in the process $N(\cdot)$. For a Hawkes process, this takes the form

$$\lambda(t; \mu, \theta | H_t) = \mu + \sum_{i:t_i<t} f(t - t_i; \theta) \tag{2.3}$$

where μ is the background intensity, θ is the set of parameters, and f denotes the triggering kernel which modulates excitation.

To capture the bursty and cyclical nature of human activity, we propose a periodic multidimensional Hawkes process model (MPHP). This model learns directly from data both the periodic effects on human behavior as well as the interdependencies between activity types. Consider a sequence of events $\{(t_i, v_i)\}_{i=1}^{N}$, where the ith event occurs at time t_i and is of event type v_i. Consistent with our observations in Fig. 2.2, we relate the rate $\lambda_v(t)$ for event type v at time t to both (1) excitation caused by previous events (parametrized by matrix $A = \{a_{ij}\}$) and (2) weekly periodicity and circadian rhythms (parametrized by $\delta_{d(t)}$ for the dth day of the week, and $\delta_{h(t)}$ for the hth hour):

$$\lambda_v(t) = \mu_v \delta_{d(t)} \delta_{h(t)} + \sum_{i:t_i<t} a_{v_i,v} g(t_i - t_j) \tag{2.4}$$

where μ_v is the background intensity for event type v. The triggering kernel $f(\Delta t; \theta)$ is a function of A which measures the effect of previous activity types on the type of the current activity. The triggering function g controls the magnitude and decay rate for the influence of past events on future events. In terms of periodicity, the model is flexible and can be easily modified (see Methods) to capture periodic effects of arbitrary intervals of time (for example six hour intervals describing morning, afternoon, evening and night), or solely whether the weekend has an effect. We adopt the commonly used scaled exponential form for interpretability and tractability:

$$g(\Delta t; \omega_1, \omega_2) = \omega_1 e^{-\omega_2(\Delta t)} \tag{2.5}$$

As such, our model not only captures bursty and self-exciting behavior, but also allows for direct inference on the structure of influence between types of activity. We can see that the parameters on the influence between activity types closely follow actual bursts in purchases. In contrast to past work on human dynamics, we add context to human activity, offering simple, interpretable insights into human actions. We show that this context lends insights into lifestyles obtained by grouping individuals according to their transaction similarity.

2.2 Methods

2.2.1 Parameter Estimation

Consider a sequence of events $\{\tau_i\}_{i=1}^{N}$ where each $\tau_i = (t_i, u_i)$ corresponding to the time t_i of the event and the stream u_i upon which it occurred. Under the framework of the proposed model, arbitrary types of periodicity can be defined; however, for notational transparency we describe here a slight simplification of the multidimensional periodic Hawkes Process in Eq. 2.4 with only daily periodicity. Other types can be defined analagously.

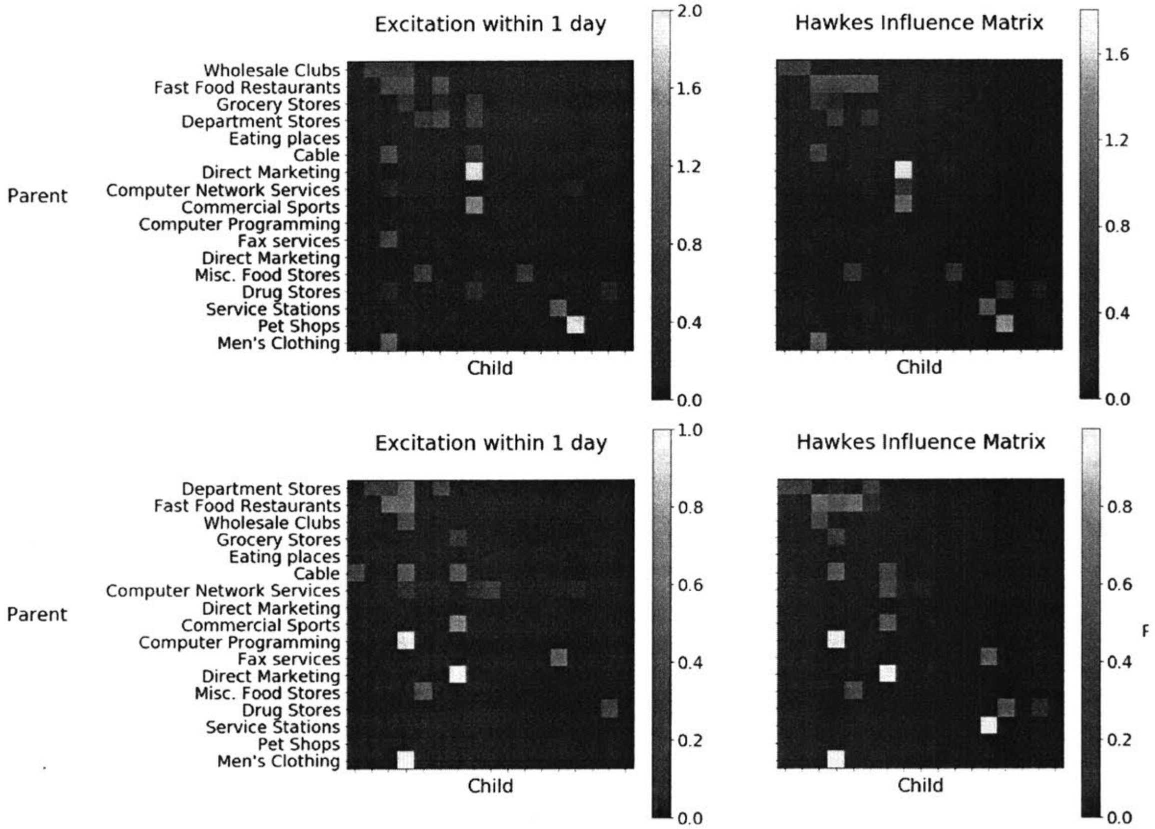
For this stochastic process, the likelihood of a given sequence $\tau = \{\tau_i\}$, where $G(t) = \int_0^t g(s)\mathrm{d}s$, is given by

$$\begin{aligned}\mathcal{L}(A, \mu) &= \sum_{i=1}^{N} \log\left(\mu_{v_i}\delta_{d_i} + \sum_{t_j<t} a_{v_i v_j} g(t_i - t_j)\right) \\ &- \frac{1}{7}T \sum_{v=1}^{V}\sum_{d=1}^{7} \mu_v \delta_d \sum_{v=1}^{V}\sum_{j=1}^{N} a_{v v_j} G(T - t_j)\end{aligned} \tag{2.6}$$

Following [52], we adapt a novel maximum aposteriori expectation-maximization (MAP EM) algorithm which avoids convergence problems with maximum likelihood. Furthermore, this algorithm allows for regularization on the influence matrix and periodicity parameters.

Bayesian Expectation-Maximization

Following [52], we define the latent variables $Q = [q_{ij}]$:


Excitation within 1 day

Hawkes Influence Matrix

Excitation within 1 day

Hawkes Influence Matrix

Figure 2.1: On the left, we see the average number of occurrences where a hypothesized "child" event occurred within the same day of a "parent" event for two sample users. On the right is the Hawkes influence matrix. We observe that this influence matrix corresponds closely to the transitions between activities in empirical data.

- $q_{ij} = \mathbb{1}$(event i is the parent of event j)
- $q_{ii} = \mathbb{1}$(event i is a background event)

Also known as the branching matrix, Q gives rise to a natural interpretation of the branching structure of a Hawkes process, as shown for credit card transactions in 2.3.

Now, given $\tau = \{(t_i, v_i)\}$, we maximize the log of the complete data posterior subject to constraints on δ_d with respect to the set of parameters $\Theta = \{\mu, A, \delta\}$:

$$\log p(\Theta|\tau, Q) \propto \log p(\tau, Q|\Theta) + \log p(\Theta) + \beta \left(\sum_{i=1}^{7} \delta_d - 7 \right) \tag{2.7}$$

We place a Gamma prior on the scaling parameter for day of week, $\delta = [\delta_d]$, and influence matrix entries $A = [\alpha_{ij}]$. This reduces the effect of days of the week that have not been seen in the sequence, and regularizes a potentially large number of influence parameters:

$$p(\delta_d) = \prod_{i,j} \text{Gamma}(\delta_d;\ x_d, y_d) \tag{2.8}$$

$$p(A) = \prod_{i,j} \text{Gamma}(\alpha_{ij};\ s_{ij}, t_{ij}) \tag{2.9}$$

Incorporating these priors, Bayesian expectation-maximization (EM) algorithm then alternates between finding the expected value of $P = [p_{ij}]$ of Q in the expectation step (*E-step*), and maximizing the posterior to with respect to Θ in the maximization step (*M-step*). Assume an integer number of weeks is observed. Then the updates are as follows:

E-step.

Compute $P^{(k+1)} = \mathbb{E}[Q\ |\tau, \Theta^k]$ as

$$p_{ii}^{(k+1)} = \frac{\mu_{v_i}^{(k)} \delta_{d_i}^{(k)}}{\mu_i^{(k)} \delta_{d_i}^{(k)} + \sum_{j=1}^{i-1} \alpha_{v_i v_j}^{(k)} g(t_i - t_j)} \tag{2.10}$$

$$p_{ij}^{(k+1)} = \frac{\alpha_{v_i u_j}^{(k)} g(t_i - t_j)}{\mu_i^{(k)} \delta_{d_i}^{(k)} + \sum_{j=1}^{i-1} \alpha_{v_i v_j}^{(k)} g(t_i - t_j)} \tag{2.11}$$

M-step.

Compute $\Theta^{(k+1)} = \left(\mu^{(k+1)},\ A^{(k+1)},\ \delta_d^{(k+1)}\right)$, with $G(t) = \int_0^t g(s)\ \mathrm{d}s$, as

$$\mu_v^{(k+1)} = \frac{\sum_{i:v_i=v} p_{ii}^{(k)}}{T} \tag{2.12}$$

$$\delta_d^{(k+1)} = \frac{\sum_{i:d_i=d} p_{ii}^{(k)} + x_d - 1}{\frac{1}{7}\sum_i p_{ii} + x_d - 1} \tag{2.13}$$

$$\alpha_{uu'}^{(k+1)} = \frac{\sum_{i:v_i=v} \sum_{j:v_j=v',j<i} p_{ij}^{(k)} + s_{vv'} - 1}{\sum_{i=1}^{N} \sum_{j:\ v_j=v',j<i} G(T - t_j) + t_{vv'}} \tag{2.14}$$

2.2.2 Simulation

We adapt the improvements of [52] for faster simulation. Namely, given the rates at the last event t_k, we can calculate $\lambda(t)$ for $t > t_k$ by

$$\lambda_v(t) = \mu_v \delta_d + e^{-\omega(t-t_k)} \left(a_{vv'_k} \omega + (\lambda_v(t_k) - \mu_v \delta_d)\right) \tag{2.15}$$

Algorithm 1: Simulation of event sequences from an MPHP

```
Input: $\mu = \{\mu_v\}$, $A = (a_{ij})$, $\omega_1$, $\omega_2$, $\delta = \{\delta_d\}$, horizon
Output: Sequence of event types $\{(t_i, v_i)\}_{i=1}^{N}$
Simulate first event:
$I^* \leftarrow \sum_v \mu_v$
$D^* \leftarrow \sum_{i=1}^{7}(\delta_d)$
repeat
    $t_0 \sim \text{Exp}(1/M)$
    $d_0 \leftarrow \text{dayofweek}(t_0)$
    $U \sim \text{Unif}(0, 1)$
until $U < \frac{d_0}{\sum_{d=1} 7\delta_d}$
$v_0 \leftarrow v$ w.p. $\mu_v / I^*$
$\lambda(t_0) \leftarrow \mu\delta_d$

General procedure: $k \leftarrow 0$
Step 1
    $I^* \leftarrow \max(\delta_d) \sum_{\mu_v} \max(\delta_d)\lambda_v(t_k) + \omega \sum_v a_{vv_k}$
Step 2
    $t' \leftarrow t_k + s,\ s \sim \text{Exp}(1/I^*)$
if $t' >$ horizon
    return $\{(t_i, v_i)\}$
end
$\lambda(t') \leftarrow \delta_{d'}\mu + e^{-\omega(t'-t_k)}\left(A_{v_k}\omega + \lambda(t_k) - \delta_{d_k}\mu\right)$
Step 3
$u' \leftarrow u$ w.p. $\mu_v / I^*$
$d' \leftarrow d_k$
if $v'$ is $v + 1$
    Reject
else
    Attribute:
        $t_{k+1} \leftarrow t'$
        $v_{k+1} \leftarrow v'$
        $d_{k+1} \leftarrow d'$
        $\lambda(t_{k+1}) \leftarrow \lambda(t')$
        $k \leftarrow k + 1$
    Step 1
```

2.3 Results

We test our model on six months of credit card data (individual credit card records, or CCRs) in Mexico City, one of the most populated cities in Latin America. In this dataset, the activity types are Merchant Category Codes (MCCs) or purchase categories. The granularity of time stamps are one day, and thus we model periodicity depending on the day of week only. In Fig. 2.2, empirical patterns in credit card transaction history show clear burstiness for two sample users.

To fit the model, we learn the parameters through a novel expectation-maximization algorithm (see *Methods*) and choose ω_1 and ω_2 to maximize the out-of-sample likelihood in cross-validation. We show that this model outperforms standard baselines in prediction, and fits human behavior well as evidenced by statistical testing. As shown in Fig. 2.3, the estimation of influence between activity types results in the explicit explanation of the dependencies on past events.

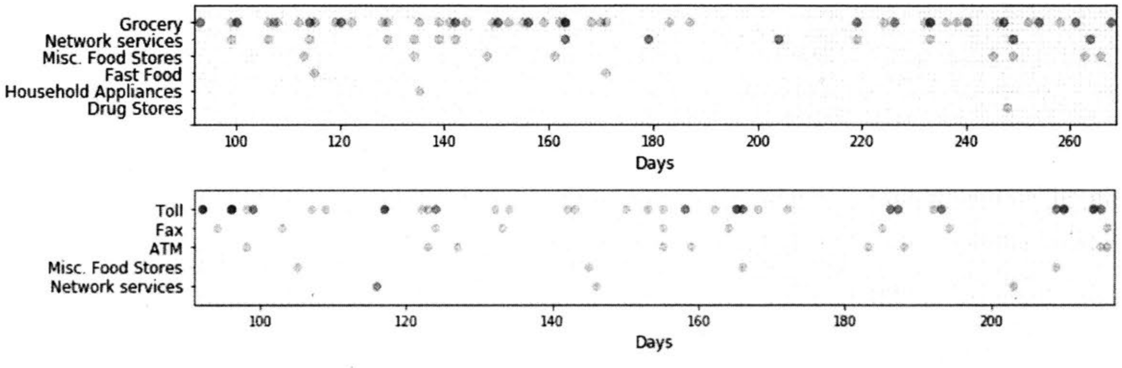


Figure 2.2: The transaction history for the shopping behavior of users indicates alternating periods of high and low activity.

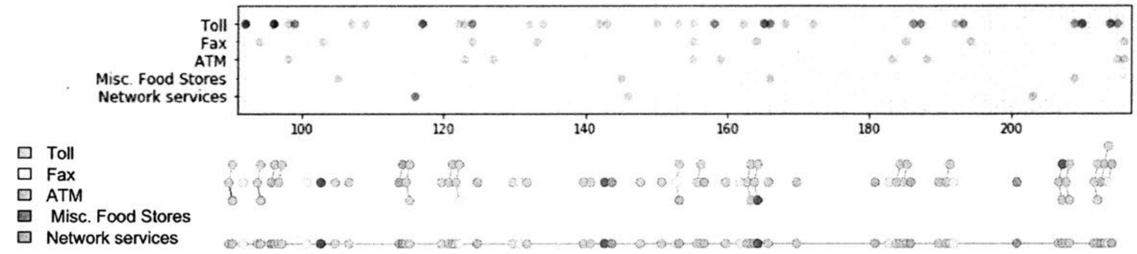


Figure 2.3: Branching process of the multidimensional periodic Hawkes model for a single user's credit card transaction history. Each edge denotes a parent event exciting a child event. We see that toll fees tend to both self-excite and excite purchases in other categories.

2.3.1 Prediction

We compare the predictive performance of the multidimensional periodic Hawkes model with the multidimensional periodic Poisson process as a null model. Following [55], we also consider latent Dirichlet allocation (LDA) [9], a generative statistical model commonly used in the context of natural language processing. This is a widely used model that identifies shared patterns across users, but does not consider the temporal dimension.

To evaluate the predictive power of our model, we consider a binary classification task: given all purchases of a user until time t, will the user make a new purchase type i in the next time period, $[t, t+\epsilon]$? For each user, t is a randomly chosen time within the last 10% user's history. We choose a small time window of $\epsilon = 2$ to measure each model's ability to capture self-exciting behavior in addition to general patterns in activity. For almost all MCCs, an overwhelming majority of users will not make a purchase of that type (90% - 97%). Due to the imbalanced nature of the data, we use precision and recall as metrics to evaluate prediction performance.

Using a stochastic process, we can repeatedly generate a sequence of events in the time window $[t, t+\epsilon]$ and record the percent of sequences containing a purchase of the specified type. Let N be the total number of events in the training data (all events within $[0, t]$) and T be the total time encompassed. Similarly, for latent Dirichlet allocation (LDA), we repeatedly generate a sequence by drawing $N\epsilon/T$ events using the generative process of LDA.

As we can see from Fig. 2.4, the multidimensional periodic Hawkes process (MPHP) outperforms both the multidimensional periodic Poisson process (MPP) and LDA in predicting a range of event types. This shows that the temporal component is necessary, and that more than periodicity is needed; indeed, the self-exciting behavior captured by the MPHP model is more representative of human shopping patterns.

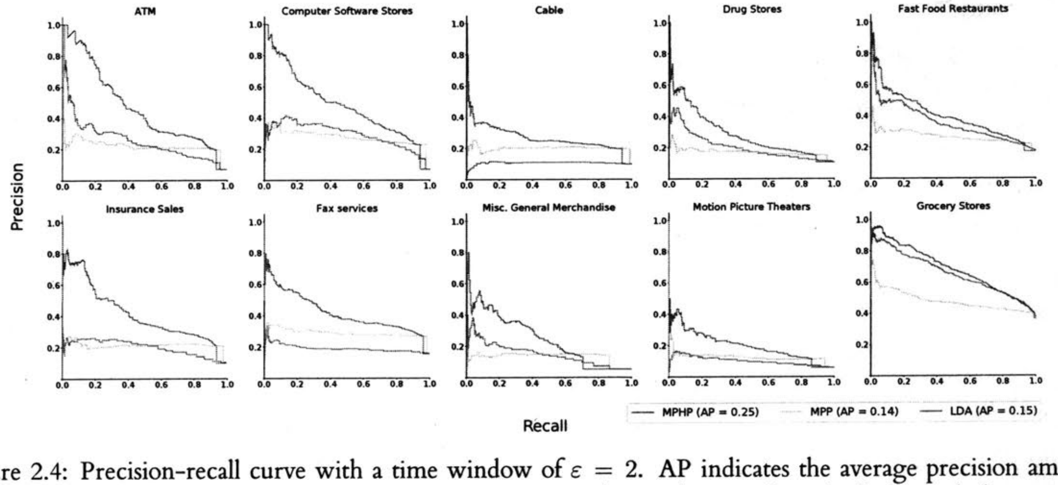
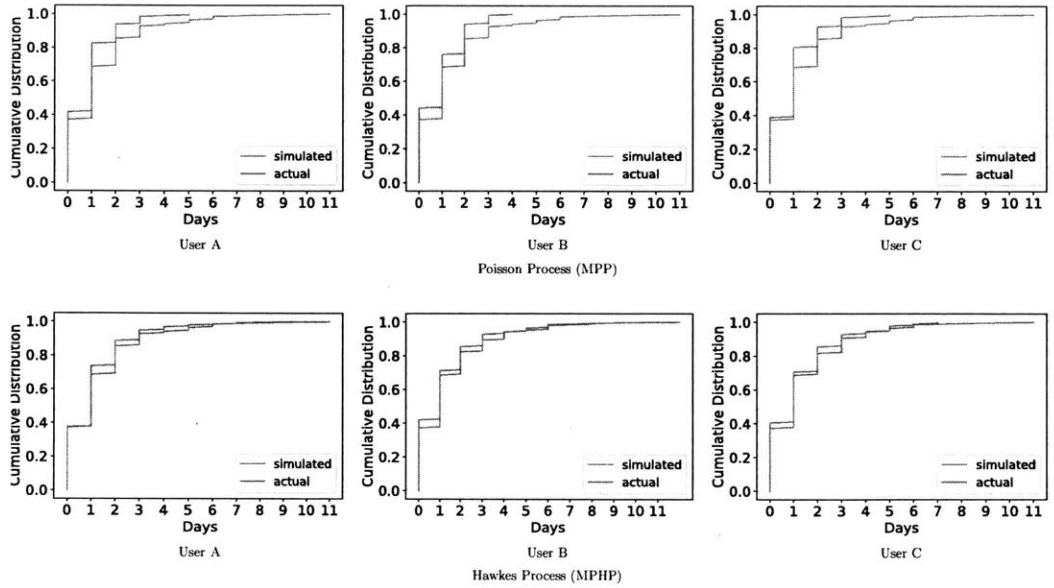


Figure 2.4: Precision-recall curve with a time window of $\varepsilon = 2$. AP indicates the average precision among all thresholds. The multidimensional periodic Hawkes process (MPHP) outperforms both the multidimensional periodic Poisson process (MPP) and latent Dirichlet allocation (LDA).



User A

User B

User C

Poisson Process (MPP)

User A

User B

User C

Hawkes Process (MPHP)

Figure 2.5: We simulate event sequences using the proposed model and the multidimensional Poisson process for three users. We compare the cumulative distributions of inter-event times of each simulated sequence and the empirical data. We see from the orange curves that empirical data shows heavy tails in the inter-event distribution, which, in contrast to the Hawkes process, the Poisson process is not able to generate.

2.3.2 Monte Carlo Hypothesis Testing

For each user, we learn a multidimensional periodic Hawkes model and compare the model's predictions with the empirical cumulative distribution of inter-event times (see Alg. 2). Due to the inherent burstiness of human activity, we expect heavy tails in these distributions. The empirical distributions of several shoppers in Fig. 2.5 illustrate how the proposed multidimensional periodic Hawkes process better captures bursty inter-event time distributions than a multidimensional periodic Poisson process.

As the estimated parameters depend on the empirical data, we use Monte Carlo hypothesis testing to assess the significance of the agreement for each of the 23,317 users. At the 5% significance level, our model can only be rejected for 29.8% of users. For this minority, the probability of an a one day inter-event time was comparable to that of a same day inter-event time, indicating that their excitation function is not exponential decaying. This suggests that for these cases, a better fit may be achieved with the substitution of a triggering kernel that does not start decaying until after one day. A nonparametric triggering kernel [83, 6] could result in a closer fit for all users, but would result in losses in both interpretability and scalability.

In comparison to the multidimensional periodic Hawkes model, a multidimensional periodic Poisson model is rejected for 100% of users at the 5% significance level,. We see that the proposed model is complex enough to capture a wide range of human behavior, while remaining simple enough to lend insight into patterns within individual activity. Following [48], we assess statistical goodness of fit of a model using the area statistic, or the area between the cumulative distribution function, between the empirical data and event sequences simulated from the model.

Algorithm 2: Monte Carlo hypothesis testing

```
Input: Data D
Output: p-value for rejection of MPHP as a model for D
$m \leftarrow 0$
while $m < M$
    Learn stochastic process S1 from D
    Simulate D1 from S1
    $A^m \leftarrow$ Area between inter-event CDFs of (D, D1)

    Learn stochastic process S2 from D1
    Simulate D2 from S2
    $A_0^m \leftarrow$ Area between inter-event CDFs of (D1, D2)
    $m \leftarrow m + 1$
end
Compute t-statistic between groups $\{A^m\}_{m=1}^M$ and $\{A_0^m\}_{m=1}^M$.
```

Temporal Lifestyles 3

3.1 Mining Shopping Patterns

Our spending habits reflect patterns in behavior, capturing an essential aspect of our lifestyles. Within the computational social science community, the question remains whether pervasive trends exist among different groups at urban scale. A recent paper [19] reveals lifestyles of urban regions through shopping behaviors discovered by both latent Dirichlet allocation (LDA) and text compression methods. Here, we extend these findings and characterize those lifestyles temporally, recovering new insights in not only patterns of MCC sequences, but also the temporal transitions between them. Here, we extend these findings, characterizing for each discovered lifestyle both periodic cycles and the temporal transition between purchases. Additionally, we show that these shopping lifestyles have a strong connection with both demographics and social contacts.

We first use LDA to identify behavioral patterns of co-occuring MCCs among individuals, representing each individual's spending lifestyle as a finite mixture of an underlying set of behaviors. Each behavioral pattern, in turn, is modeled as a mixture of a set of Merchant Category Codes (MCCs). In addition, LDA regularizes the number of behaviors per individual, as well as the number of MCCs per behavioral pattern. Thus each individual is represented by a small number of behaviors, and each behavior involves making a small set of purchase categories with high frequency.

From Table 3.1, we see the highest weighted MCCs pertaining to each shopping behavior. We discover natural patterns in credit card usage, including behaviors related to dining (Behavior 8), vacation and travel (Behavior 7), and regular technology service purchases (Behavior 3). In addition, we recover outside errands of a related type (Behaviors 6, 9 and 10), and modes of transportation (Behaviors 1 and 2). As transportation is necessary across the population, we expect purchases in mainly one form of transportation, and that this will be independent of other shopping habits. Accordingly, we see separated behaviors related solely to the mode of transport in Behavior 1 (Tolls) and Behavior 2 (Taxis).

To recover the temporal dimension of these shopping behaviors, we represent each behavior as a mixture of the individual-level Hawkes models. This is done directly by taking the weighted average of each user's parameters, where each individual's weight is determined by the proportion of that individ-

Behavior	Merchant Category Codes	Proportion
1	Toll and Bridge Fees	94.4%
2	Taxicabs and Limousines	78.8%
3	Computer Software, General Merchandise, Direct Marketing, Cable	59.9 %
4	ATM, Insurance, Fax and Telecommunication, Business Services, Cable	74.3%
5	Computer Network Services, Fax and Telecommunication	78.5%
6	Grocery Stores, Department Stores, Drug Stores, Wholesale Clubs	59.1%
7	Airlines, Hotels, Specialty Retail, Travel Agencies, Family Clothing	37.0%
8	Eating places and Restaurants, Fast Food, Department Stores	65.9%
9	Misc. Food and Convenience Stores, ATM, Grocery Stores, Fast Food	82.5%
10	Gas Stations, Grocery Stores	60.6%

Table 3.1: The highest weighted MCCs for each shopping behavior in decreasing order of proportion (the probability of the MCC given the topic). All MCCs with proportion over 4% are shown for their respective topic. Proportions on the right show the total proportion of the shopping behavior described by these MCCs.

ual's lifestyle described by the behavior. In Fig. 3.1 we depict the background intensities μ_v for a subset of the highest weighted MCCs in each shopping behavior. As expected, these background intensities correspond closely to the behaviors in Table 3.1, indicating that these temporal rates describe well the purchasing patterns within each behavior.

For each shopping shopping behavior, we also examine the patterns of excitation between MCCs. Fig. 3.2 contains a subset of the Hawkes influence matrices for MCCs with the highest influence parameters $\{a_{ij}\}$. Despite the wide diversity in shopping behaviors, there is strong similarity in the MCCs that tend to excite other purchases. These MCCs generally indicate that outside errands, such as purchases in Grocery Stores (behavior 6), Department Stores (behavior 4), and Restaurants (behaviors 4, 7, and 8) excite other purchases that involve purchases outside the home. Those that indicate leaving the house (Tolls, Taxis, Gas/Service Stations in behavior 9) tend to excite other purchase types at an even higher rate as seen in shopping behaviors 1, 3 and 10. From the diagonal elements of these matrices it is clear that many purchase types strongly tend to self-excite, including transportation (Toll, Taxis), shopping (Department Stores, Women's Clothing), and business transactions or scheduled transactions (Computer Network Services, Fax).

For each shopping shopping behavior, Fig. 3.2 contains a subset of the Hawkes influence matrices for MCCs with the highest influence parameters $\{a_{ij}\}$. From the diagonal elements of the Hakwes influence matrices, it is clear that many purchase types tend to self-excite. This includes transportation (Toll, Taxis), shopping (Department Stores, Women's Clothing), and business transactions or scheduled transactions (Computer Network Services, Fax and Telecommunication). Despite the diversity in shopping behaviors seen in Table 3.1 and Fig. 3.1, there is strong similarity in the MCCs that tend to excite other purchases. As seen in shopping behaviors 1, 3 and 10, these MCCs indicate that outside errands (Grocery, Department Stores, Restaurants), and transportation related purchases (Tolls, Taxis, Gas/Service Stations) tend to excite other purchase types at an even higher rate. This corresponds with

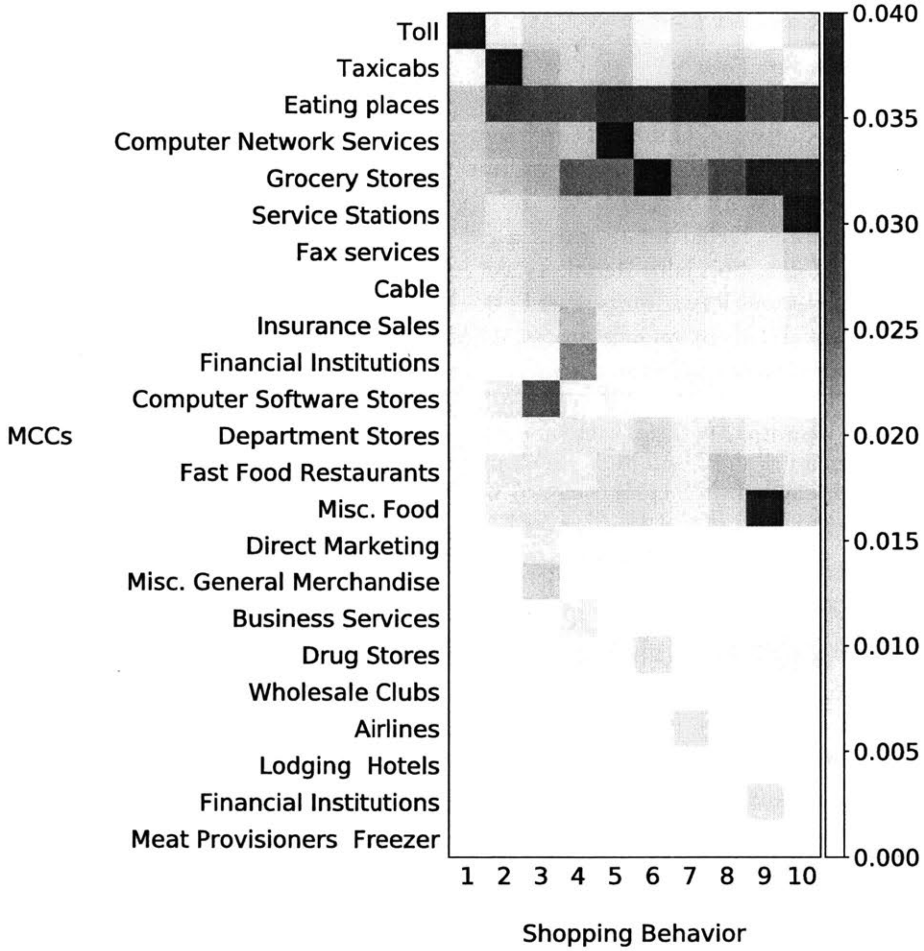


Figure 3.1: The background intensities for each shopping behavior learned from LDA. We show the set of MCCs containing the fifteen highest rates for each behavior.

research in mobility; leaving the home results in a higher probability of additional outside stays [34].

3.2 Discovering Lifestyles

As each individual is a mixture of shopping behaviors, we use the Jensen-Shannon divergence [43] on these probability distributions to calculate the similarity matrix between users. Using this matrix, we construct a weighted graph between users and identify lifestyles using the Louvain community detection algorithm [11] for computational efficiency.

From Fig. 3.3 and Table 3.2, we see that our method reveals consistent relationships between demographics, expenditure, and shopping behavior. Altogether, we recover definitive lifestyles in urban regions, behaviors connected across multiple aspects of activity. In Fig. 3.3, Lifestyle 1 characterizes primarily older male users whose main spending (shown in Table 3.2) is composed of toll fees, followed

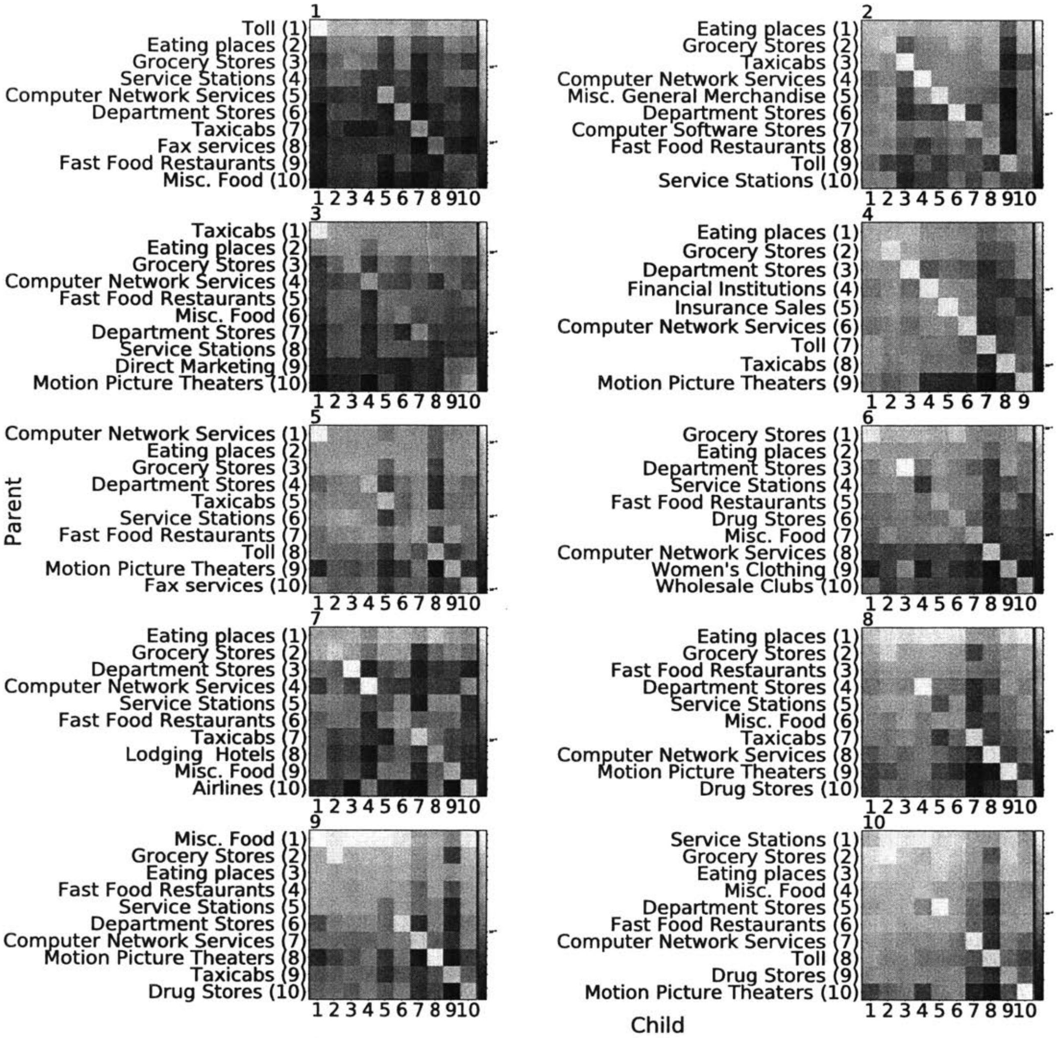


Figure 3.2: Hawkes influence matrix for each of ten shopping behaviors. We see clear patterns of excitement within these behaviors. **1, 2, 3**, Purchases relating to modes of transportation (Tolls, Taxis, and Service Stations) tend to result in fast transitions to a range of purchase in other categories. **4,6,9** We see similar results for errands outside

by dining purchases. Given that 57% of Mexico City residents do not own a car, these users tend to have a relatively high expenditure, as expected. We label this lifestyle *Car Owners*. The second lifestyle we denote as *Families* and describes older, married individuals with higher spend, primarily on groceries. In contrast, the younger group *Tech Users* mainly spend on technologies such as computer services, fax, and cable, in addition to restaurant and fast food purchases. Within the younger population, we find two additional groups without technology purchases: a lower expenditure group we label *Low Income Youth*, with transactions mainly in grocery stores, convenience stores, and ATMS, and a higher expenditure group *Higher Income Youth* with frequent purchases in taxis and restaurants.

Car Owners

Behavior	Merchant Category Codes	Proportion
1	Toll and Bridge Fees	52.9%
8	Eating places and Restaurants, Fast Food, Department Stores	5.2%
5	Computer Network Services, Fax and Telecommunication	3.6%

Families

Behavior	Merchant Category Codes	Proportion
6	Grocery Stores, Department Stores, Drug Stores, Wholesale Clubs	42.8%
8	Eating places and Restaurants, Fast Food, Department Stores	11.9%
10	Gas Stations, Grocery Stores	8.5%

Tech Users

Behavior	Merchant Category Codes	Proportion
3	Computer Software, General Merchandise, Direct Marketing, Cable	25.0%
8	Eating places and Restaurants, Fast Food, Department Stores	12.8%
5	Computer Network Services, Fax and Telecommunication	7.8%

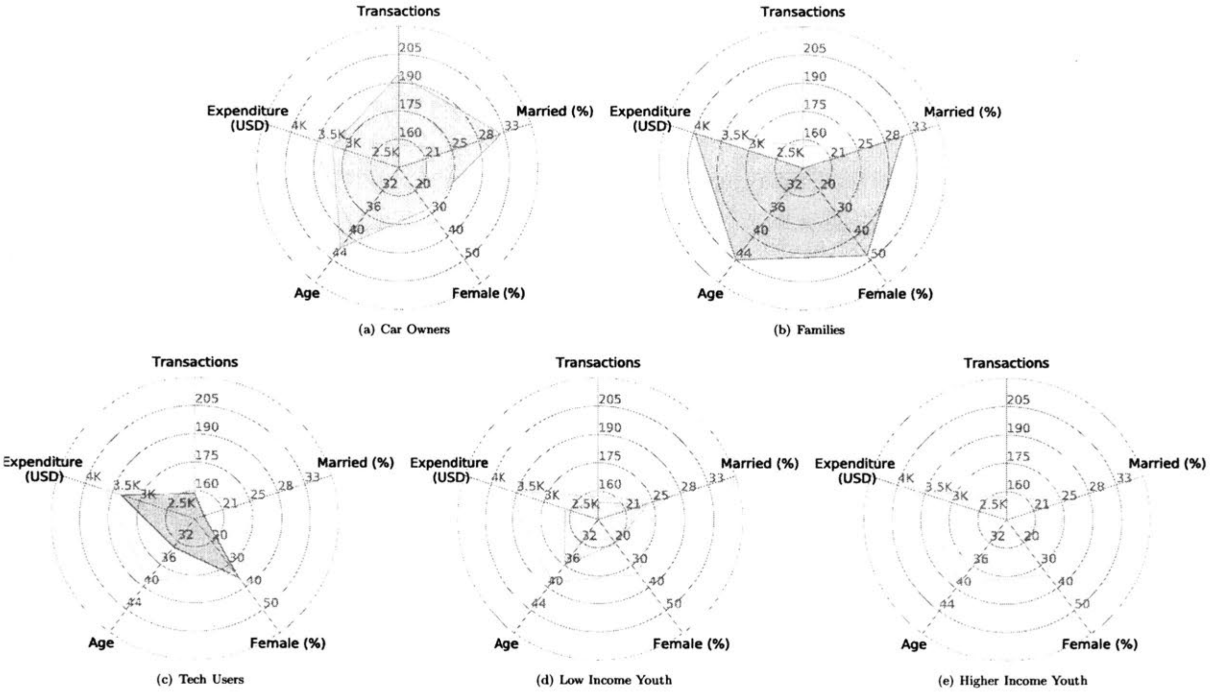
Low Income Youth

Behavior	Merchant Category Codes	Proportion
9	Misc. Food and Convenience Stores, ATM, Grocery Stores, Fast Food	41.1%
10	Gas Stations, Grocery Stores	8.6%
6	Grocery Stores, Department Stores, Drug Stores, Wholesale Clubs	5.6%

Higher Income Youth

Behavior	Merchant Category Codes	Proportion
8	Eating places and Restaurants, Fast Food, Department Stores	42.5%
6	Grocery Stores, Department Stores, Drug Stores, Wholesale Clubs	8.1%
10	Gas Stations, Grocery Stores	7.5%

Table 3.2: The most highly represented shopping behaviors within lifestyles discovered by Louvain community detection. From Fig. 3.3, we see that these communities have a decisive relationship with demographics and expenditure.



(a) Car Owners

(b) Families

(c) Tech Users

(d) Low Income Youth

(e) Higher Income Youth

Figure 3.3: Lifestyles found by Louvain. Median values for all users are indicated by the gray lines.

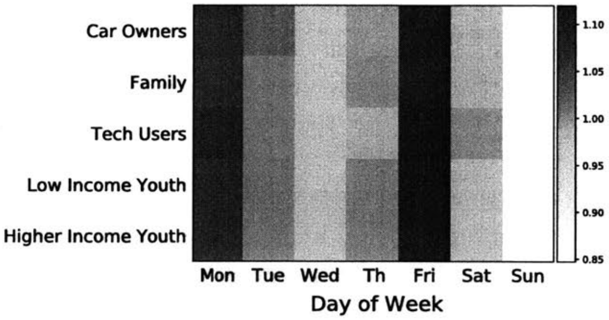


Figure 3.4: The scaling parameters for day of week by lifestyle.

Fig. 3.4 shows that weekly rhythms are constant across lifestyles in most respects. We see the highest increase in purchases on Fridays, consistently with our expectation that people tend to dine out, shop in addition to their normal working commute. As the first working day, Mondays also tend to lead to an increase in purchases. In addition, we see that *Car Owners* and *Tech Users*, lifestyles that may be more associated with working professionals, tend to make significantly fewer purchases on Sunday, just before the working week begins.

Method	Vertices	Edges	Max Degree
unfiltered	1.5×10^6	2.0×10^6	14.6×10^3
5 calls	1.5×10^6	2.0×10^6	14.6×10^3
10 calls	7.9×10^5	9.9×10^5	5.4×10^2
mutual call	2.7×10^5	4.1×10^5	280

Table 3.3: Call network properties for different levels of filtering.

3.3 Shopping Lifestyles and Social Networks

Literature suggests that friends tend to be similar in terms of shopping and other aspects of their lifestyles [45, 17, 68]. We test this hypothesis through the lens of community detection, quantifying the relationship between communities obtained from shopping behavior and communities discovered in actual communications data.

We first construct a social network using call detail records. We consider the set S of $23,317$ users for which we have credit card records. An edge exists between two users, or vertices, if there is a call between them. We then extract records of calls for all users with a path length of 2 or less between users in set S, and construct a directed weighted network based on the number of calls. This results in a network of about 3 million vertices. However, these vertices include companies, telemarketers, and weak connections, as we can see from the maximum degree in Table 3.3. We use multiple methods to filter out weak connections, including keeping only edges with at least 5 or 10 calls, or alternatively employing the common practice [42] of considering only reciprocal, or mutual calls, where each user has called the other at least once.

Given the large scale of this call network, computational efficiency is a key issue in applying community detection. The Louvain algorithm [11] is known to be scalable and produce communities of high quality with respect to modularity [69, 11]. In contrast, Infomap [59] is less computationally efficient, but accounts for the directedness of the network through an information theoretic objective function describing how well information obtained through communities can compress random walks. We compare the results of these two algorithms on social networks obtained through various filtering methods, and evaluate the strength of their relationship with the communities of shopping lifestyles found previously. In addition, we construct a multiplex network as shown in Fig. 3.5, with the social network as one layer, and the distributions of shopping behavior forming another. Specifically, since each user is a mixture of shopping behaviors, we create a weighted edge from each user to the ten previously discovered shopping behaviors, including edges only if the corresponding weight is greater than a certain threshold (here 0.1).

We compare both the multiplex network and the single layer social networks with respect to their normalized mutual information (NMI) evaluated on shopping lifestyles. NMI is a standard method to evaluate similarity between communities [50], and here is scaled to range from 0 to 1, analagously to

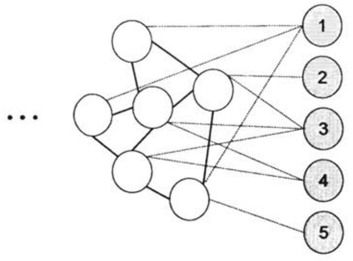


Figure 3.5: Combining shopping lifestyles and call networks. We start with a social network, and add a weighted edge from each user to the ten previously discovered shopping behaviors, including edges only if the corresponding weight is greater than a certain threshold.

Method	Filter	Number of Clusters	NMI
Infomap, combined	mutual	6.7×10^4	0.41
Infomap	10	9.0×10^4	0.40
Infomap	5	1.2×10^5	0.40
Louvain	10	4.4×10^4	0.20
Louvain, combined	mutual	2.3×10^3	0.12
Louvain	5	2.3×10^3	0.11

Table 3.4: Normalized mutual information (NMI) of shopping and call network communities, where the left most column describes methods applied to the call network. From the NMI we observe a clear relationship between one's social network and shopping behavior.

the correlation measure. In Table 3.4, we see that Infomap achieves the highest NMI of 0.41 with the multiplex network, though we also observe that different filtering methods make little difference with respect to Infomap's ability to find meaningful communities. However, filtering methods do impact the results obtained from the Louvain algorithm, with the filter for at least 10 calls achieving the best results. We hypothesize that the success of the Infomap algorithm is due mainly to its ability to leverage the directed nature of the network. We see also that many weak connections do not affect the algorithm as much as Louvain, likely because random walks have low probability of crossing the corresponding edges.

Thus, we further validate the urban lifestyles obtained using the social network of these individuals. We observe a clear signal between these communities, and correspondingly, identify a significant relationship between one's social network and one's demographics and shopping preferences.

3.4 Discussion

Our results clearly show that the excitation structure between events, when coupled with weekly cycles, precisely characterizes bursty activity in human behavior. In addition to accurately describing heavy tails in inter-event time distributions, the proposed model solidly outperforms baseline models in difficult prediction tasks. Furthermore, the descriptiveness of the model easily lends itself to interpretation, giving insight into the actual priority individuals place on tasks, and the order in which they tend to execute them. Through this model, we add context to lifestyles found in urban regions, and describe the temporal behavior of disparate groups. Due to its generality and flexibility, the multidimensional periodic Hawkes process can describe a wide range of activity. The detailed generative mechanisms derived from the proposed model have the potential to lend insight into the diverse aspects of not only human dynamics, but also the dynamics of natural, technological and economic phenomena.

Furthermore, we show that the discovered lifestyles connect to both demographics and expenditure of different groups. We use the social network of these individuals to further validate these lifestyles, observing a clear signal between communities found in communications data and those obtained through the analysis of purchasing patterns. Correspondingly, we identify a significant relationship between social contacts, demographics, and shopping preferences.

There are many possible avenues for future work. One direct extension of this work would be to include the mobility information obtained from Call Detail Records in this analysis; that is, connecting how people move to how they spend, as well as who they communicate with. Furthermore, we can study the dynamics of opinion formation with respect to purchasing behavior and mobility. For example, we can model the social network as a dynamical system using SI-type models, or alternatively study voter models or evolutionary graph theory methods as possible models for opinion dynamics.

Mining Urban Lifestyles 4

In this chapter, we jointly model the lifestyles of individuals, a more challenging problem with higher variability when compared to the aggregated behavior of city regions. Using collective matrix factorization, we propose a unified dual view of lifestyles. Understanding these lifestyles will not only inform commercial opportunities, but also help policymakers and nonprofit organizations understand the characteristics and needs of the entire region, as well as of the individuals within that region. The applications of this range from targeted advertisements and promotions to the diffusion of digital financial services among low-income groups.

4.1 Mining Shopping and Mobility Patterns

Location and transactional data offer valuable perspectives on the lifestyles of each user. For example, we may expect the shopping purchases of middle-aged parents to include groceries and fuel, while their mobility patterns may center around localities near home and work locations, in addition to points of interest such as supermarket, laundry, and so on. We use mobility information to aid in the prediction of shopping behavior, connecting the two views using collective matrix factorization [62]. In this way, we discover representative patterns relating shopping and mobility, characterizing behavior for a richer understanding into urban lifestyles and improved prediction of behavior.

The high granularity of such digital records allows modeling at the level of the individual, providing a new framework in which to relate movement and spending. However, in using CDR data for data on individuals, we must deal with issues of sparsity and lack of contextual information on the user's activities. In proposing this dual view of lifestyles, our contributions can be summarised as follows:

Prediction of Shopping Behavior with Data Sparsity

There are many individuals for which we have no CDR data. To deal with this data sparsity issue, we construct a framework that uses mobility patterns as supplementary information in the prediction of shopping behavior. We connect the two perspectives on lifestyles using collective matrix factorization (collective matrix factorization). In comparison to modeling only shopping behavior, we find that incorporating mobility information in the prediction of shopping lifestyles leads to a significant reduction in root mean square error (RMSE). This enables more precise recommendations for products based on consumer preferences.

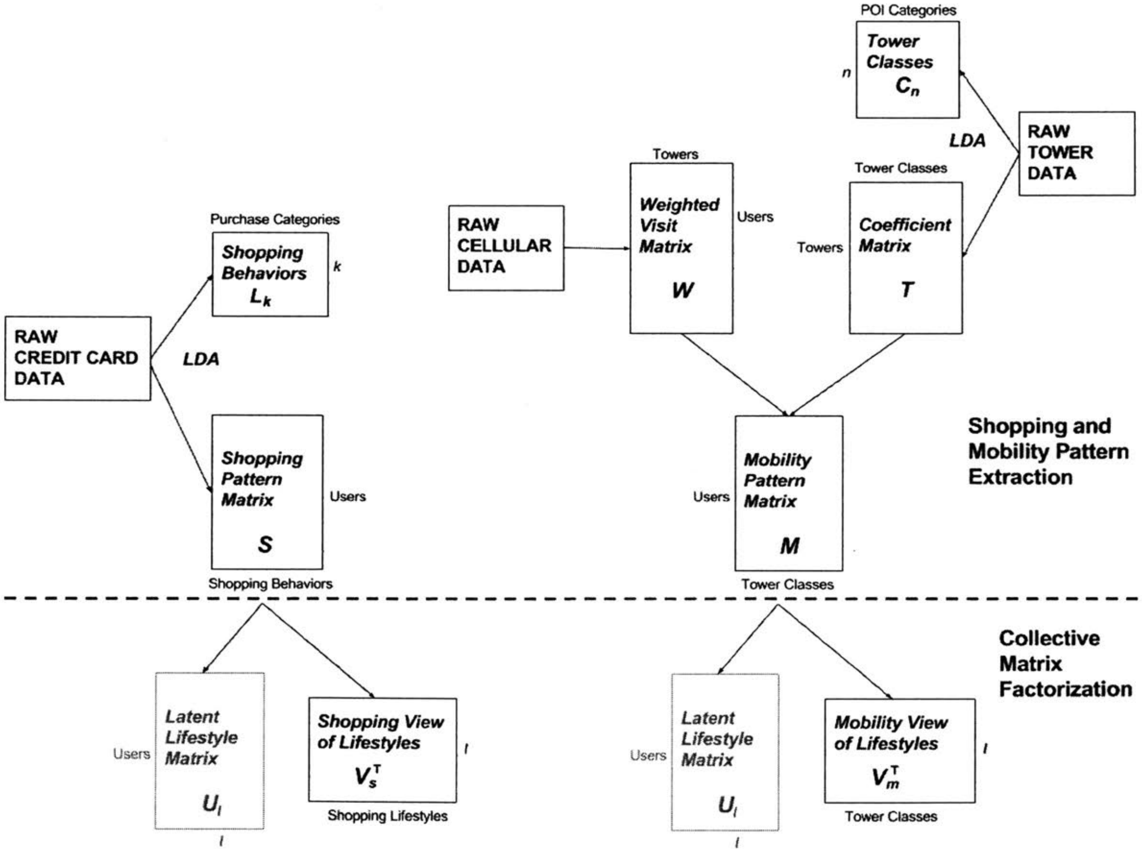
Adding Contextual Information to Location Data

We transform mobility data using external data sources to better relate CCR to CDR Data. Although CCRs provide high granularity at the level of the individual user, spatial granularity can range from a radius of 200 - 1000 meters, and there is no contextual information for the user's activities within that region. Thus, there has been little previous work leveraging CCR data for prediction with CDR data.

Multi-Perspective Lifestyles

We describe the mappings between shopping and mobility patterns, connecting the two views to provide a novel understanding of consumer behavior in urban regions.

Figure 4.1: Our framework



Data

The primary datasets used in this chapter consist of two sets of anonymised data for residents in Mexico throughout five months in 2015:

- Call detail records (CDRs). CDRs are produced with each telephone exchange, These kocation records give the nearest cellular tower at the time of a placed call. There are 1192 cell towers throughout Mexico City – as users tend to visit a small subset of these towers, this mobility data is extremely sparse. In a count matrix denoting user visits to towers, 98% of entries indicate zero visits.

- Credit card records (CCRS). CCRs are recorded with each purchase and denote the purchase category, or Merchant Category Code (MCC), of the transaction as well as the amount spent. Each month, we have on the order of 10 million financial transactions and 200 million location records.

4.2 Discovering Shopping Patterns

Our spending habits reflect our lifestyles, capturing an essential aspect of our behavior. Within the computational social science community, the question remains whether pervasive trends exist among disparate groups at urban scale [19]. In this chapter, we use latent Dirichlet allocation (LDA) [10] to identify topics (behavioral patterns) among individuals, representing each individual's spending lifestyle as a finite mixture of an underlying set of behaviors. Each behavioral pattern, in turn, is modeled as a mixture of a set of words (Merchant Category Codes, or MCCs). These topics are determined by co-occurrences of words within a document. For example, in an article database, we may uncover a topic containing the words "data", "processing", "computer", and so on because these words frequently appear in an article together.

By putting a Dirichlet prior on the per-user behavior distribution and per-behavior MCC distribution, LDA controls the sparsity of the number of topics per document (the number of behaviors per individual), as well as the number of words per topic (the number of MCCs per behavioral pattern). In this way, each individual is represented by a small number of behaviors, and each behavior involves making a small set of purchase categories with high frequency.

As a generative model, LDA allows us to calculate the probabilities (assignments to shopping behaviors) of previously unseen users. We train the model on 40% of the users, and generate the matrix S for the remaining 60%. In so doing, we set up the prediction of lifestyles for unseen users, assessing the LDA model itself in addition to the relation of shopping with mobility patterns. We experiment with the choice of number of behaviors to learn, as well as adding a categorical variable describing amount spent to each MCC. To maximise interpretability, we choose five topics while using MCCs as input only.

In Fig. 2 we plot the twenty most highly weighted MCCs of the five shopping behaviors. The first shopping behavior describes credit card usage that is centered on food-related purchases such as Grocery Stores, Misc. Food Stores and Restaurants. The second shopping behavior seems to be associated primarily with business purchases, with spending within MCCs such as Fax Services and Financial Institutions. The third shopping behavior is dominated by relative "luxuries" such as purchases in the Cable and Department Store categories, and is characterised by a relatively high proportion of Air Travel and Hotel Lodging MCCs. The fourth shopping behavior contains primarily purchases in Computer Network Services and Service Stations (gas stations). The third and fourth shopping behavior describe a slightly wealthier portion of the population, as only 35% of Mexicans owned a computer in 2010 [71], and only 44.2% own a car [28]. Lastly, the fifth shopping behavior captures purchase primarily for toll fees and subscription services.

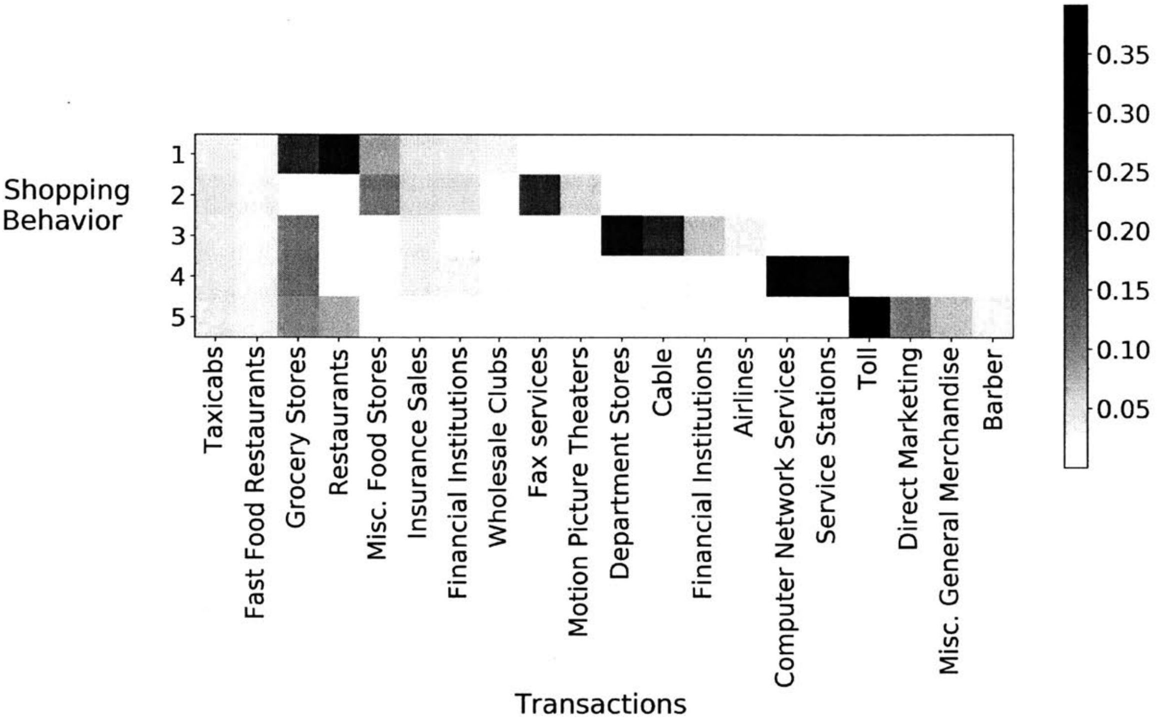


Figure 4.2: The top weighted purchase categories of the five shopping behaviors learned from LDA.

4.3 Mobility Pattern Extraction

Extracting Cellular Tower Location Types

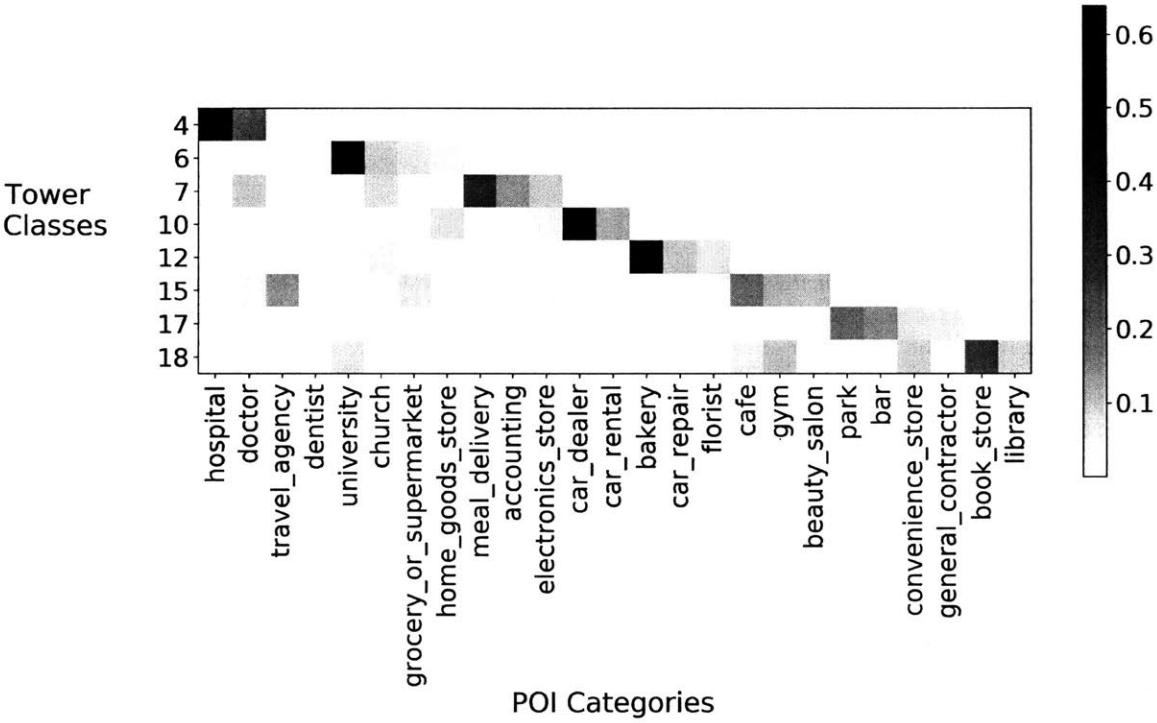
Within the CDR data, each tower is the site for a corresponding cell within the Voronoi diagram; i.e., it is the closest tower to any point within this cell. We define a "visit" to a cellular tower as a call placed within its corresponding cell. In order to relate cellular towers to spending behavior, for each tower we crawl Google's API for points of interest within a certain radius. To determine this radius, we use

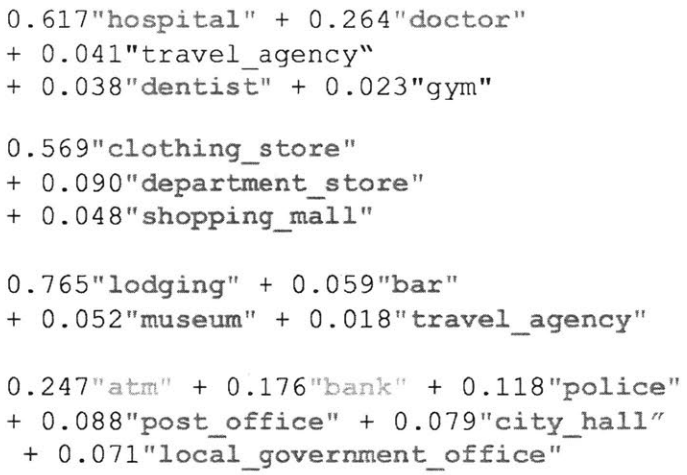
Delaunay triangulation, a widely used method in computational geometry. Delaunay triangulation gives the dual graph to the Voronoi diagram, maximizing the minimum angle among all the triangles within the triangulation, and connecting the sites in a nearest-neighbor fashion [5]. For each tower, we set the crawling radius to be half the average distance from the site to its neighbors.

Treating each of the Voronoi cells as a document and the POI categories as words, we use latent Dirichlet allocation to discover underlying tower "classes" that will be more informative of shopping behavior. We remove from the vocabulary any POI categories that occur with over 25% frequency. These removed categories are uninformative classifications such as "point of interest" and "establishment". For purposes of interpretability, we learn the LDA model with twenty classes on the 1192 towers.

In Fig. 4.3, we show a subset of tower classes highly weighted within our final lifestyles (see Section 6), and the corresponding points of interest with the highest probabilities. From the sample topics in Fig. 4.4, we see each tower class puts specific emphasis on related points of interest, such as "hospital" and "doctor", "car rental" and "car repair", or "book store" and "library". In this way, we cluster the towers in terms of nearby POI categories, obtaining contextual information more directly related to shopping.

Figure 4.3: The top weighted POI categories of a subset of tower classes learned from LDA.





```
0.617"hospital" + 0.264"doctor"
+ 0.041"travel_agency"
+ 0.038"dentist" + 0.023"gym"

0.569"clothing_store"
+ 0.090"department_store"
+ 0.048"shopping_mall"

0.765"lodging" + 0.059"bar"
+ 0.052"museum" + 0.018"travel_agency"

0.247"atm" + 0.176"bank" + 0.118"police"
+ 0.088"post_office" + 0.079"city_hall"
 + 0.071"local_government_office"
```

Figure 4.4: Sample topics from learned from LDA, treating each tower as a document and each POI as a word.

Baseline Methods

Before introducing our model, we present the results of several baseline methods, illustrating the challenges of incorporating CDR data into the prediction of shopping patterns.

Regression on Average Amount Spent

Using the columns of the per tower count matrix W directly as features, we use regression with L1 regularization to predict the average amount spent by the user per week. As we increase regularization we increase the test R-squared, but due to a combination of sparsity and lack of signal achieve a maximum test R-squared of 0 as the coefficients shrink to 0.

Classification of Primary Shopping Behavior

For each user, we take as our outcome the highest weighted shopping behavior from the topic proportions learned from LDA. This is the user's primary behavior. Again using the columns of W as our features, we employ a range of classifiers including SVM and AdaBoost to predict primary behavior. We find that the best classifier achieves only 21.6% accuracy, when already 21.9% of users fall into a single class.

Characterizing Mobility Patterns

From the Voronoi diagram of the p cell tower locations, we construct a matrix $W \in \mathbb{R}^{nxp}$ where each entry w_{ij} is the number of days individual i visited tower j throughout five months. We weight these counts using TF-IDF, a common method for text representation [77]. Using TF-IDF, we offset the tower counts by the frequency of the tower in the data, so that a user's visit to an uncommonly visited tower is assigned a higher weight. We now have a matrix $W \in \mathbb{R}^{nxp}$ that characterizes users in terms of tower visits, and a matrix $C_m \in \mathbb{R}^{pxd}$, where d is the chosen number of tower classes. We define our mobility pattern matrix as $M = WT$, achieving a significant dimensionality reduction with $M \in \mathbb{R}^{nxd}$. In this manner, we obtain a representation of mobility more closely related to shopping behavior, as users are now characterized by their visits to tower classes defined by POI categories.

4.4 Collective Matrix Factorization

For many users, we have access to data on mobility patterns (M) but not shopping patterns (S). In this section, we describe our methodology for incorporating mobility information in addition to shopping information for the matrix completion problem of predicting the shopping behavior of unseen users.

We use *collective matrix factorization* [62] to recover latent representations underlying patterns in shopping behavior and mobility. Denote S as the matrix of behavior proportions obtained from latent Dirichlet allocation, and M as the matrix of weighted visit frequencies to the different tower classes. Modeling each user's shopping and mobility behavior as two views of the same lifestyle, we assume that S and M are generated from a matrix U_l containing the latent lifestyle information of each user.

$$S \approx U_l V_s^T$$

$$M \approx U_l V_m^T$$

Traditionally, the objective function under this model is represented as

$$\mathcal{L}(U_l, V_s, V_m) = ||S - U_l V_s^T||^2 + ||M - U_l V_m^T||^2 + \lambda_1 ||U_l||^2 + \lambda_2 ||V_s||^2 + \lambda_3 ||V_m||^2$$

4.5 Prediction

In our problem, credit card data is unknown for many users, but we would like to use mobility information to predict their shopping behavior; i.e., S contains many empty rows. Thus, to test the performance within this setting, we remove rows from the shopping behavior S to predict the shopping behavior of users for which we have no credit card information. We use 10-fold cross validation and compare our collective matrix factorization predictions with the actual values. We use the popular metric root mean square error (RMSE) to evaluate our model.

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{i,j} (S_{i,j} - \hat{S_{i,j}})^2}$$

Using cross-validation to determine the rank (number of lifestyles), we find that the inclusion of mobility data leads to a 1.3% decrease in RMSE and obtain a test error of 21.6%.

4.6 Dual Lifestyles

Using collective matrix factorization, we also obtain both the dual shopping and mobility views of these latent lifestyles, in V_s and V_m respectively.

Lifestyle 1 is connected with wealthier shopping behavior typical common to urban white collars. The top weighted shopping patterns indicate spending on cable, air travel, hotels and at department stores as well as gas stations and computer network services (Fig. 2: behaviors 3 and 4, respectively). This suggests that people who can afford to spend on relative luxuries tend to have vehicles and thus higher mobility, visiting a wider range of tower classes. The mobility patterns of this lifestyle focus on areas with points of interest such as universities, accounting, electronics, bakeries and car repair (Fig. 4.3: tower classes 6, 7, 12, 17 and 20).

Lifestyle 2 is extremely food-oriented, with high weight on shopping behavior 1. Mobility patterns suggest visits to cafes, gyms and convenience stores.

Lifestyle 3 primarily captures the transportation aspect of lifestyles. Top weighted mobility patterns indicate visits to areas with car rental and car repair (tower classes 10 and 12), while shopping patterns include gas stations in behavior 4 and food in behavior 1.

4.7 Discussion

In this study, we relate the shopping and mobility patterns of consumers on an individual level for the first time. Viewing these perspectives as aspects of the same underlying lifestyle, we set up a framework to incorporate call detail records in the prediction of shopping patterns for unseen users. We achieve a significant increase in prediction, allowing for greater accuracy in consumer recommendations. Additionally, we lend insight into lifestyles in urban regions by establishing interesting relationships between shopping and mobility.

There are many directions for future work. In terms of modeling formulation, it would be interesting to introduce a temporal dimension into the task of shopping prediction, as human behavior and needs vary over time. There is also the opportunity to include social regularization in the collective matrix factorization formulation, constraining each user to be similar to his or her neighborhood. In addition, stronger prediction methods may be achieved by modeling nonlinear relationships using geometric deep learning methods described by [13].

Household Segmentation by Load Shape and Daily Consumption 5

5.1 Introduction

In this chapter, we propose an algorithm to segment consumers using their daily load profiles, which, in contrast to previous work, clusters both by peak times in energy consumption as well as the overall magnitude of consumption. Such household segmentation not only yields a deeper understanding of consumer behavior, but can also be used to develop meaningful insights for a wide range of applications. Some of these include: targeting consumers who are most suitable for the adoption of certain distributed energy technologies (such as PV and storage), finding out which consumers can effectively provide demand response at particular time periods, and helping Distribution Network Operators (DNOs) to plan for robust low voltage networks.

The focus of this chapter is to develop an interpretable number of consumer segments from high variability smart meter data. In the proposed segmentation methodology, we take a structured approach that leverages the detailed information within smart meter records to cluster load profiles by both shape and total energy consumption. We show that this distinction is critical to explicitly determine PV and battery suitability for our consumers, and show that smart meter-based segmentation enables personalized recommendations for both PV and battery adoption.

Our main contributions are summarized as follows:

- *Segmentation by both peak time and overall consumption*: As we will demonstrate, novel techniques such as [40] are unable to segment consumers by both peak time and overall consumption. We present two-stage *k*-means, a simple but effective methodology that achieves both.

- *Development of metrics to measure representativeness of clustering*: Previous works in household segmentation [40, 41, 8, 26] do not attempt to measure the representativeness of each cluster centroid for its respective members. We propose metrics to capture two quantities that are of significant interest for a wide range of applications: similarity in peak time and magnitude, and similarity in overall consumption. We conduct a survey of popular and recently published methods for electric load shape segmentation, and evaluate each method with respect to these metrics.

- *Demonstration of value to PV battery systems*: While extensive research works have implied applications for their segmentation methods [8, 26, 29, 40, 41], they have not explicitly shown how their methods could be used in these applications. In this chapter we demonstrate the use of our proposed methodology for battery and PV systems, making a direct comparison with surveyed methods. The results are of interest to both companies and informed consumers seeking a precise mapping from consumption patterns to best selection of solar and storage capacity.

5.1.1 Problem Statement

In contrast with previous work, the current paper proposes an alternate approach for segmentation that considers both the shape of a load profile – the time and magnitude of its peaks – as well as its *overall consumption* – *i.e.*, the daily total energy consumption in kWh. We survey four unsupervised learning methods with respect to their ability to segment consumer load profiles, analyzing the segmentation results using residential energy consumption data. Building on these results, we propose two-stage *k*-means, a scalable methodology for segmentation that improves upon surveyed methods in terms of both peak time and overall consumption. We then show this method's ability to group consumers based on their suitability for PV and storage.

By clustering thousands of daily load profiles into an interpretable number of consumption patterns, we obtain a dictionary of centroids representative of a daily consumption pattern of a household. Each household is ideally characterized by a small fraction of centroids within this dictionary. As opposed to segmenting on the average load shape per household, this allows us to preserve the granularity of the data describing a household's day to day transitions across centroids.

As has been done in previous work [40, 41], to bring measurements to a similar scale for pattern comparison we introduce normalized load profiles which are obtained through a simple decomposition of the load profile $l(t)$. These are expressed as:

$$s(t) = \frac{l(t)}{T}, \quad T = \sum_{i=1}^{N} l(t),$$

where $s(t)$ is the normalized load profile, or load shape, and N is the number of measurements within a 24 hour period.

5.2 Metrics

In order to evaluate the representative accuracy of the different segmentation methods considered, we propose several metrics.

5.2.1 Peak Overlap

Accurate forecasts of both peak times and peak magnitudes are essential for ensuring the smooth operation and planning of electric utility companies [24]. To measure the performance of an algorithm in these respects, we propose a metric that quantifies the percentage of overlap for peaks in each daily load shape with those of its assigned cluster centroid. As described in Alg. 3, for each load profile this metric measures the area shared by its peaks and the peaks of the assigned centroid. It then divides the sum of these overlapping areas by the sum of the union of the areas covered by the peaks of both time series. With additional considerations for noise, the peak start index is determined by the point at which the gradient becomes positive, and the end index by the point at which the gradient becomes nonnegative[1].

5.2.2 Overall Consumption

An important measure of representative accuracy is the similarity of the centroid with respect to the overall consumption of the load profile. To measure this, we calculate the total energy consumption (kWh) for both the daily load profile and the total of the assigned cluster centroid, taking the absolute value of the difference as the error. We then divide this error by the energy consumption of the load profile to obtain the percent error in consumption.

5.2.3 Entropy

While some information can be gleaned from average household load shapes, a household's day-to-day variability is an important behavioral characteristic. Understanding this is crucial in predicting extreme load conditions and as well as for other applications — for example, households with low variability maybe be more reliable targets for demand response or may be able to discover higher rewards from PV adoption. We use the metric of entropy to capture the consumption pattern variability of a household j:

$$H_j = -\sum_{i=1}^{k} p(C_i) \log_2 p(C_i),$$

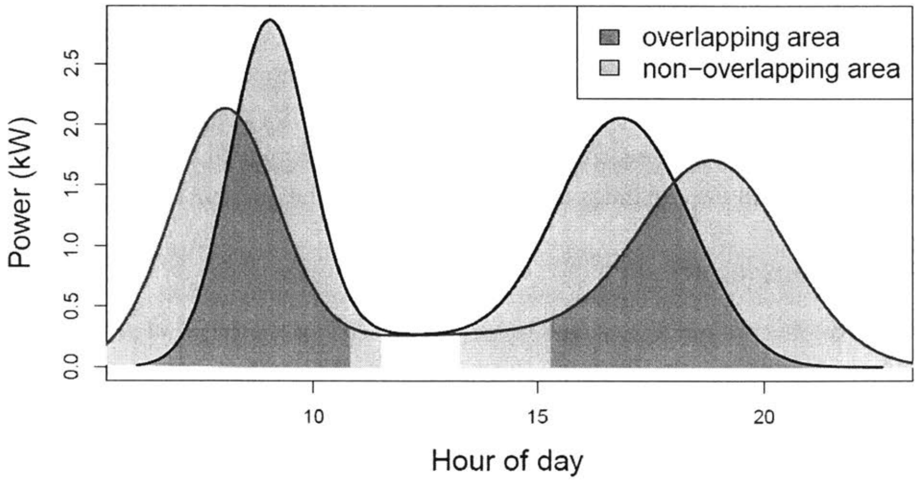
Here $p(C_i)$ is the probability that a load profile of household j will be encoded by cluster C_i, given by the proportion of its load profiles assigned to cluster C_i among the k possible clusters. The highest entropy occurs when $p(C_i) = 1/k$ for all i, and the lowest when the household's load profiles fall in only one cluster.

5.3 Baseline Methods

- *Standard k-means*: We implement the standard k-means algorithm on load shapes, clustering daily usage patterns by the Euclidean distance between normalized load profiles. As Section 3.1 will

[1] code available online at: https://github.com/humnetlab/household_segmentation

Figure 5.1: Illustration of the peak overlap metric described in Alg. 3. Let the red curve represent a smoothed centroid, and the blue a smoothed daily load shape. The peak overlap is given by the overlapping area divided by the sum of the overlapping area and non-overlapping area indicated above.



show, the method distinctively segments load profiles by peak times. Due to the use of normalized load profiles, standard k-means inherently ignores overall consumption when clustering.

- *Adaptive k-means and hierarchical clustering*: The adaptive k-means method first establishes k clusters using standard k-means. It then continues to further split each of the k clusters using standard k-means with $k = 2$ if any load shape violates the mean squared error threshold:

$$\sum_{t=1}^{N}(s(t) - C_s(t))^2 \leq \theta \sum_{t=1}^{N} C_s(t)^2,$$

where $C_s(t)$ is the cluster centroid of load shape $s(t)$, and θ is the threshold choice ($0 \leq \theta \leq 2$). When the threshold is satisfied, the method then uses hierarchical clustering to recombine the adaptive k-means clusters with the closest centroids until the desired number is achieved [40].

This method aims to achieve simpler segmentation with a small tradeoff in representative accuracy. As the adaptive method also focuses on Euclidean distance between load shapes, overall consumption does not factor into the segmentation results.

- *SAX k-means*: Symbolic Aggregate Approximation (SAX) [44] is a flexible method for dimensionality reduction that gives a symbolic representation of a time series. SAX first transforms the time series into its Piecewise Aggregate Approximation (PAA) representation, then converts this representation into a discrete string. Thus, distance measures defined on the symbolic approach are a lower bound on the corresponding distance measures given by the original time series. In SAX k-means, we substitute this method for the Euclidean distance as used by standard k-means.

Algorithm 3: Peak Metric: Percentage of Overlapping Area

```
Input  : Smoothed load profiles and the corresponding centroids obtained from clustering
Output: Peak overlap classification metric for each load profile (peak overlap percentage)
foreach load profile do
    For each load profile, find the indices of the peak intervals for each peak: (peak start, peak end)
    overlapping area = 0
    total area = 0
    foreach peak in time series peaks do
        Search for overlapping centroid peaks - a centroid peak index within (peak start, peak end)
        if no overlapping centroid peak exists
            total area += Area(peak)
            continue
        else
            $a_0$ = max(peak start, centroid peak start)
            $b_0$ = min(peak end, centroid peak end)
            overlapping area += $\int_{a_0}^{b_0}$ min(peak, centroid peak) $dt$

            $a_1$ = max(peak start, centroid peak start)
            $b_1$ = min(peak end, centroid peak end)
            total area += $\int_{a_1}^{b_1}$ max(peak, centroid peak) $dt$
        end
    end
    foreach non-overlapping centroid peak do
        total area += Area(centroid peak)
    end
    peak overlap percentage = overlapping area / total area
end
```

- *Integral k-means*: Due to the use of normalized load shapes, prior work has primarily clustered load profiles by the times of peak occurrences, largely neglecting the overall energy consumption of the load profile. As an alternative method to consider this factor, we introduce integral k-means. In this method we perform k-means on both the integral of the load shape and an additional feature which is described by the maximum observed power in the daily load profile.

  Specifically, we integrate the area under the load shape from $s(0)$ to $s(t)$ for $t = 1, \cdots, N$, constructing a new sequence $I(n)$ of the same length which gives the area under the signal from time 1 to time n:

$$I(n) = \int_0^n s(t)\,dt = \frac{n}{2N}(s(0) + \sum_{t=0}^{n-1} 2s(t) + s(n)),$$

  where $n = 1, \cdots, N$. Thus defined, normalized consumption levels and peak magnitudes propagate throughout the new signal to consider the load profile's overall consumption.

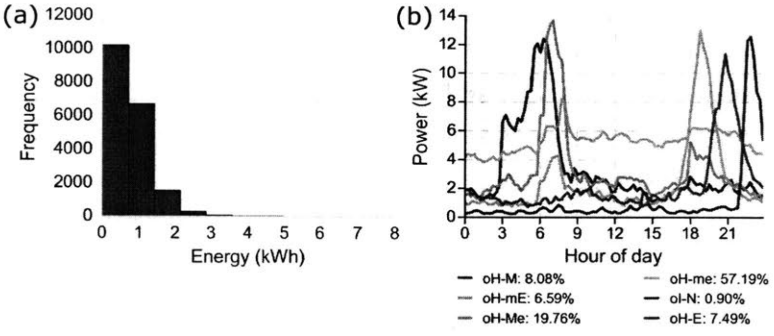
5.4 Results

Description of data

We use 15-minute smart meter data from January 2015 obtained from the Pecan Street project [2], where we have a full month of data for each household. This data contains 19,070 daily load profiles corresponding to approximately 600 households in 18 cities but primarily covering Austin, Texas (75.67% of data). Although seasonality affects overall consumption, we do not address this and choose instead to focus on successful segmentation within a single season, taking advantage of the wide range of cities covered to survey the capability of our methods to handle outliers and high variability within the segmentation process.

We apply a low-pass filter (moving average) to the data to smooth erroneous readings of 0 kW and then normalize to obtain the load shape. As seen in Fig. 5.2, the distribution of maximum peak demand is right-skewed. Only 1.8% of daily time series have a maximum peak energy usage greater than 2.5 kWh in a 15 minute interval. These outliers can have a strong effect on the learning results, especially if the objective is to obtain an interpretable number of clusters that are segmented by magnitude of overall consumption. Thus, we remove them from our primary analysis, instead segmenting them separately with standard k-means (Fig. 5.2).

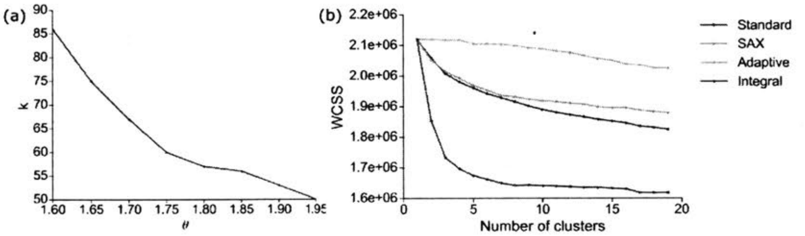
Figure 5.2: (a) Histogram of the maximum energy values for 15 minute intervals with respect to each smoothed load profile. Outliers are load profiles with maximum peak energy usage greater than 2.5 kWh. (b) Cluster centroids for outliers obtained through standard k-means on load shape ($k = 6$ determined through the elbow heuristic with WCSS). Each centroid is the average of all the readings in its cluster. The legend shows the proportion of load profiles in each cluster. See Section 6.2 for labeling conventions.



Clusters were determined using the elbow criterion as a heuristic, with the objective of minimizing the within-cluster sum of squares (WCSS) while maintaining a low number of clusters. From Fig. 5.3(b), $k = 9$ for standard k-means, $k = 6$ for SAX k-means, and $k = 9$ for the adaptive method are reasonable choices by this heuristic. For adaptive k-means, we initialize the cluster centroids using standard k-

means with the previously chosen $k = 9$ and determine an optimal threshold of $\theta = 1.75$ (Fig. 5.3a). The results are then used for hierarchical clustering with $k = 9$.

Figure 5.3: Within-cluster sum of squares for standard, SAX, adaptive and integral k-means



From Fig. 5.4, we observe that the centroids resulting from standard and SAX k-means are very clearly distinguished by peak time. This is not the case with adaptive k-means, which produces several times more clusters than desirable for interpretability, even as θ approaches its maximum value of two (Fig. 5.3). The size of these clusters varies widely, and when grouped together with hierarchical clustering, the largest clusters are combined to create a very uneven distribution with one cluster containing nearly 98% of the load shapes (Fig. 5.4). Thus, in this case where the objective is an interpretable number of clusters for a set of load profiles with high variability, the adaptive methodology performs worse than baseline k-means. It is worth noting here that previous work using this method opted to produce many more clusters than $k = 9$ and subsequently focused their interpretations on the set of highly populated clusters rather than the entire dataset [40]. In Table 5.1, we compare the results of each method with respect to the various metrics introduced in Section 2. Methods that focus on peak time such as standard and SAX k-means perform well in terms of peak overlap percentage, but have relatively high percent error in daily consumption.

As we can see from Fig. 5.5, integral k-means clusters load profiles by the magnitude of energy consumption instead of by peak time. When compared to the results of standard k-means, the segmentation results have a lower percent error in daily consumption, but also have a lower average peak overlap percentage (Table 5.1).

5.5 Two-stage k-means

Due to the high variability in energy usage across households, segmenting load profiles by overall energy consumption is an important aspect of producing a representative set of clusters. Of course, the ideal segmentation would also include information on the time of peak occurrences, which is important in many applications. We therefore suggest that by first using integral k-means to obtain n groups, then further clustering each of these n groups with standard k-means. This can segment the load profiles according to both overall consumption and time of the peak, which we show is very important for better targeting the different needs per consumer type in amount of solar PVs and battery capacity needed.

Figure 5.4: Cluster centroids for standard *k*-means, SAX *k*-means and adaptive *k*-means methods (methods that segment by peak time). The legend shows the proportion of load profiles in each cluster.

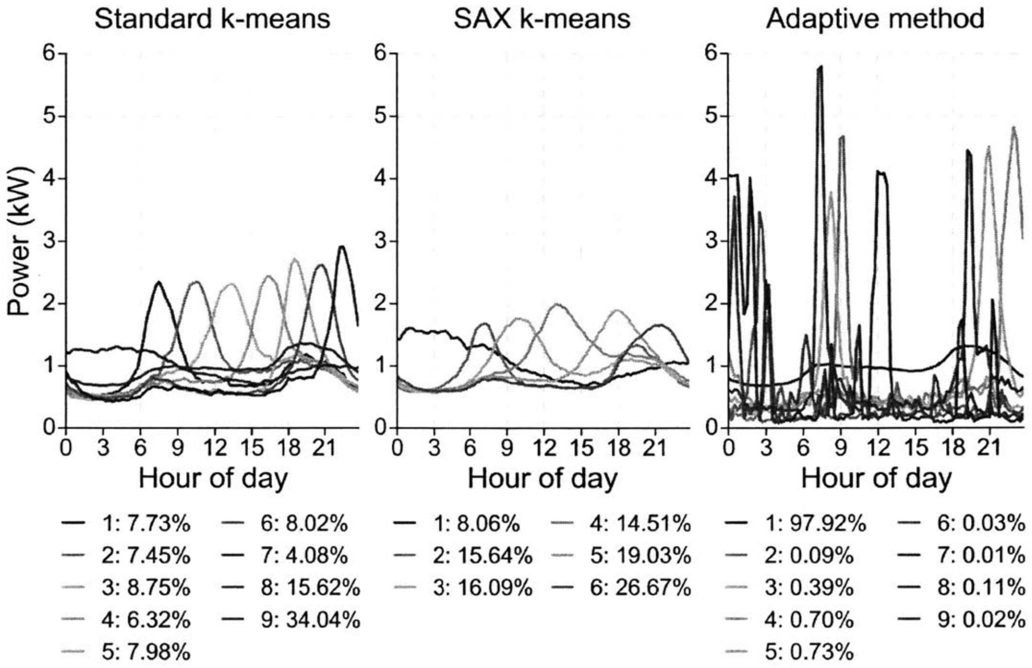
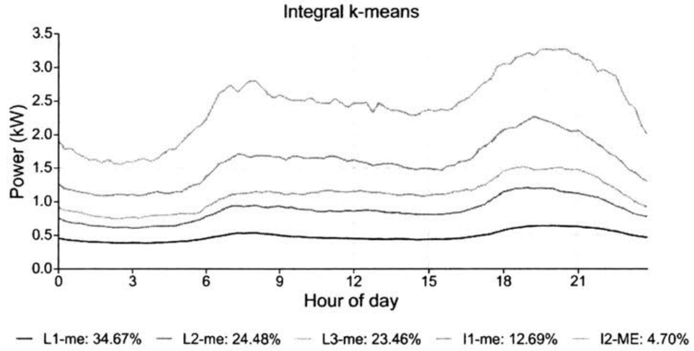


Figure 5.5: Cluster centroids for integral *k*-means. The legend shows the proportion of load profiles in each cluster.



Results

Exploiting these two clustering methods in series it is easy to produce a large number of clusters, which violates one of our primary goals of maintaining interpretability. Therefore, a lower *k* is preferred for the first stage. From the integral *k*-means elbow plot (Fig. 5.3b), we observe that three clusters is a reasonable initial number that gives suitable representative accuracy. Thus we first cluster daily time series into three groups using the integral k-means method (Fig. 5.6a), and then subsequently separate each of these groups further into clusters with $k = 4$, using the standard k-means method to obtain twelve clusters in total (Fig. 5.7). From the elbow plots showing the further splitting of the three integral k-means clusters (Fig. 5.6b), we see that $k = 4$ seems like a reasonable choice. If we wish to maintain interpretability, there is a natural tradeoff between segmenting by peak time and segmenting by consumption. We find that two-stage k-means is able to segment the load profiles relatively well in both respects: the segmentation results for two-stage k-means have an average peak overlap percentage

Figure 5.6: (a) Cluster centroids obtained from step one of two-stage k-means. (b) Within-cluster sum of squares for each group from Fig. 5.6. Using the elbow heuristic, we choose $k = 4$ for each group in the second step of two-stage k-means.

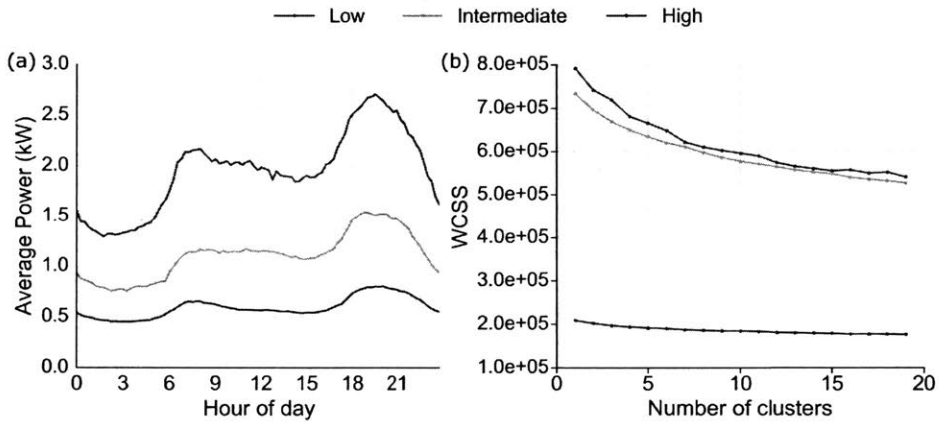
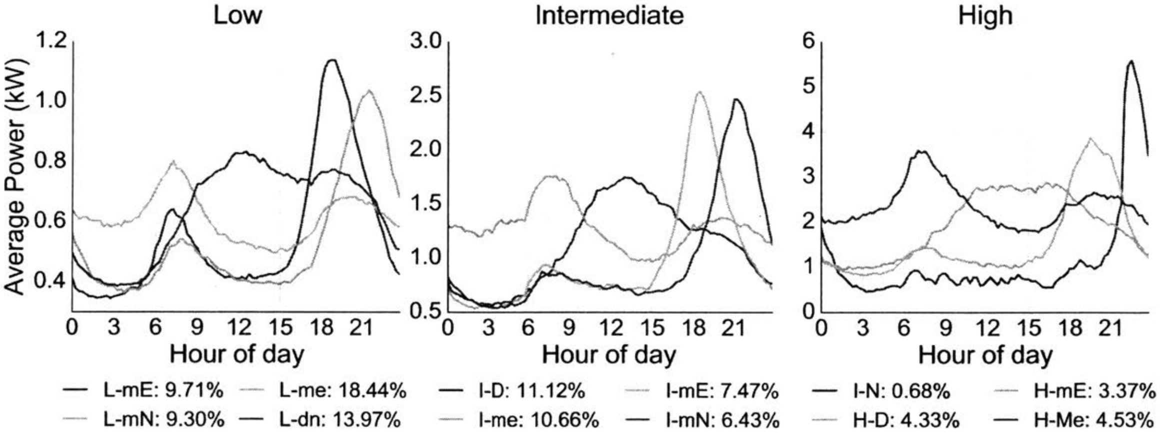


Figure 5.7: Results of the second step of two-stage k-means. Cluster centroids for each group seen in Fig. 5.6. The legend shows the proportion of load profiles in each cluster.



Method	Avg load shape entropy	Avg peak overlap %	Avg % error in consumption
Standard	1.99 ± 0.005	26.55% ± 0.1%	77.00% ± 0.8%
Adaptive	0.0978 ± 0.002	19.63% ± 0.1%	76.80% ± 0.8%
SAX	1.87 ± 0.003	24.35% ± 0.1%	76.20% ± 0.8%
Integral	1.27 ± 0.004	23.32% ± 0.1%	42.59% ± 0.4%
Two-stage	2.13 ± 0.004	26.95% ± 0.1%	47.10% ± 0.5%

Table 5.1: Method performance metrics and standard errors

comparable to standard k-means, and an average percent error in daily consumption only slightly higher than integral k-means (Table 5.1).

5.5.1 Interpretation of Segments

We label each of the resulting clusters in the proposed method based on the magnitude of the peak and their time of occurrence. Namely we define the consumption levels:

L: Low $\leq$ 20 kWh
I: Intermediate $>$ 20 kWh and $\leq$ 35 kWh
H: High $>$ 35 kWh
oI: outlier, Intermediate
oH: outlier, High

and the time of the peaks:
M, m: Morning peak (4:00-10:00)
D, d: Daytime peak (10:00-16:00)
E, e: Evening peak (16:00-21:00)
N, n: Night peak (21:00-4:00)

Within our labeling convention, lowercase letters denote low magnitude peaks, which we restrict to have a maximum value of 0.5 kW higher than the baseline power (given by the minimum power of the load profile). Otherwise, if a peak is higher than the baseline consumption by more than 0.5 kW, it is denoted by a capital letter. For example, the H-mE cluster represents a group whose centroid has a small morning peak and a large evening peak, and its total daily consumption is in the high usage bracket. Table 5.2 shows the different clusters created under this nomenclature convention. The average total daily consumption, entropy, and the proportion of load shapes assigned to the cluster are also shown. As outlined in Section 5.1, the outliers were determined by the maximum peak energy demand in the load profile, not by the total energy consumption. So the oI-N cluster contains load profiles with a maximum energy greater than 2.5 kWh, but it has low consumption elsewhere such that its total daily consumption is less than 35 kWh.

To derive further insight into lifestyles from the two-stage segmentation, we relate the results to relevant features such as average daily consumption and consumption pattern variability (entropy). As seen in

Cluster	Avg daily consumption (kWh)	Avg load shape entropy	Proportion
L-mN	12.89	2.08	9.14%
L-mE	13.32	2.19	9.54%
L-me	14.74	2.04	18.11%
L-dn	15.27	2.07	13.72%
I-mN	23.36	2.43	6.31%
I-mE	23.68	2.45	7.34%
I-D	26.55	2.26	10.93%
oI-N	27.77	2.27	0.02%
I-N	28.02	2.16	0.67%
I-me	30.96	2.17	10.47%
H-mE	37.71	2.47	3.31%
H-D	46.31	2.30	4.25%
oH-E	53.61	2.79	0.13%
H-Me	56.90	2.09	4.45%
oH-mE	58.60	2.46	0.12%
oH-Me	77.69	2.03	0.35%
oH-M	80.82	1.87	0.14%
oH-me	127.39	1.69	1.0%

Table 5.2: Statistics of households in Fig. 5.7

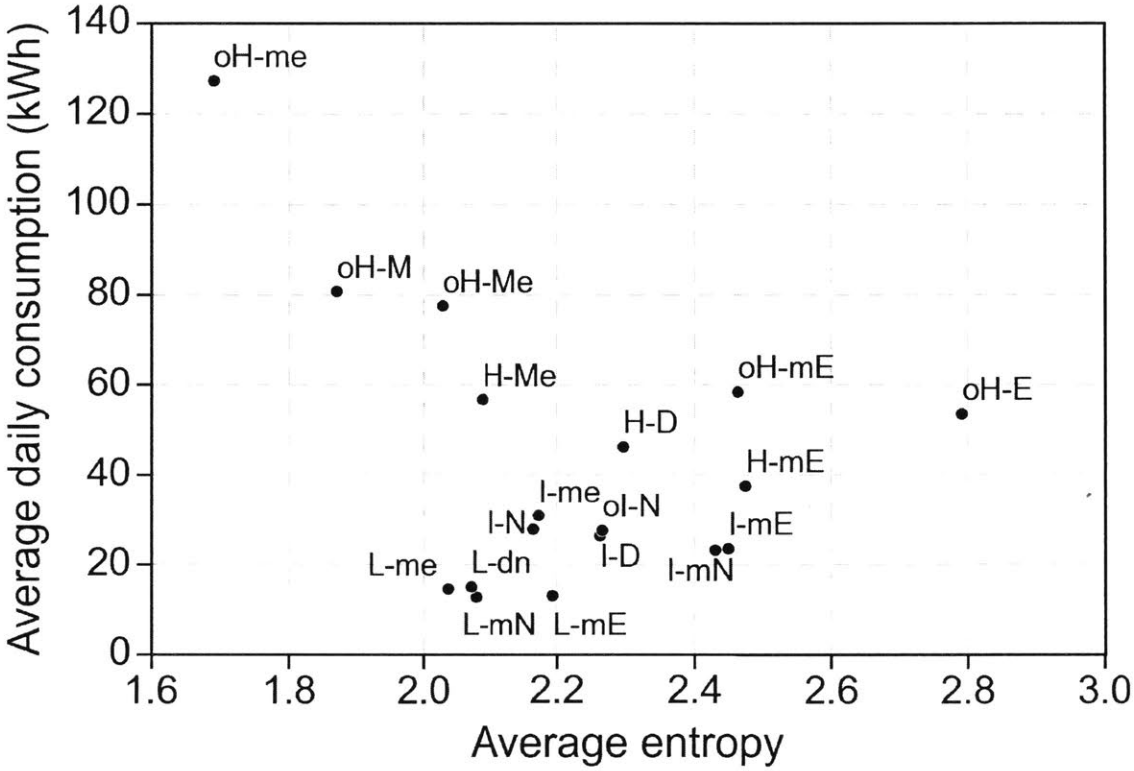
Table 5.2 and Fig. 5.8, households from clusters with low overall consumption tend to have lower entropy. These clusters also represent a high proportion of the load profiles.

Most clusters exhibit morning and evening peaks, the expected shape for residential loads. However, the presence of other loads shapes does imply interesting heterogeneity in behavior, with some households having more common demand during the daytime period. We also find higher consumption load profiles to be more variable. This may be explained by households with a higher number of occupants, while the low consumption households typically have fewer occupants with more regular behavior. It is also worth noting that we include both weekends and weekdays in our analysis, which increases the variability of households when compared to analyses which just consider weekdays.

5.6 Targeted recommendations for consumers of PV-Battery systems

In this section we illustrate the applicability and performance of the customer segmentation results. We return to the original data and calculate standard amounts for solar *PV* and battery capacity each customer may be offered in the market based on their current consumption patterns. Then we show how the knowledge of our proposed segmentation is able to separate consumers with different demand. We show that segmenting daily load shapes with the proposed two-stage k-means allow us to better group customers according to similar needs for both solar and battery capacity. To this end, we first estimate

Figure 5.8: Average daily consumption and consumption pattern variability scatter plots of lifestyle clusters



two properties of significant interest to both consumers and utilities. Namely; (1) the self-sufficiency of a consumer with PV and (2) the battery size required to increase their self-sufficiency above 75%. Consumer j's self-sufficiency (ss_j) is defined below, where E_{PVused} is the electricity generated by their PV system that they self-consume and $E_{consumed}$ is the actual energy demand.

$$ss_j = \frac{E_{PVused}}{E_{consumed}}$$

To simulate the output of a PV system we exploit the NREL PV Watts calculator [2]. We specify the system location as Austin, Texas, the DC system size as 5kW, choose a standard fixed roof-mount type and keep the rest of the parameters as default (20° tilt, 180° azimuth and 14% losses). This yields hourly results and we interpolate these to 15-minute resolution, using data for January which corresponds to the same period as our load data, defining $PV(t)$. Finally, in order to make each PV system consumer specific, we sum their total monthly usage and scale the PV output so that each consumer has a PV system which produces electricity equivalent to their total use over the period. Consumer j's PV output at the period t is expressed as:

[2] available online at: http://pvwatts.nrel.gov/pvwatts.php

$$PV_j(t) = \frac{\sum_{t=1}^{N} l(t)}{\sum_{t=1}^{N} PV(t)} PV(t)$$

To calculate each consumers required battery size for 75% self-sufficiency we devise a similar method to that used by [15] to estimate the economically optimal battery size for a microgrid. For each consumer this involves increasing the size of their battery in 1 kWh increments until their self-sufficiency increases above 75%. At each incremental battery size, the battery is then scheduled to maximize self-sufficiency. We subsequently calculate the consumers net load and resulting self-sufficiency with that battery size. Consumer j's net load at period t is then expressed as:

$$l_j^{Net}(t) = l_j(t) - PV_j(t) - P_j^{batt}(t)$$

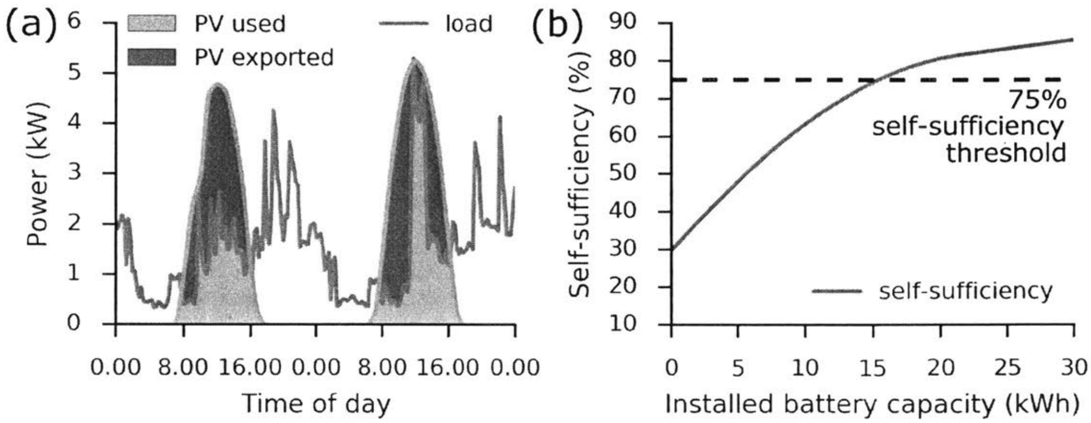
The self-sufficiency is easily calculable directly from the net load time series – it is the total net load greater than zero divided by the total consumption without PV or storage. The battery's operation P_j^{batt} for each consumer is scheduled according to Alg. 4.

Algorithm 4: Schedules the charging and discharging operation of consumer j's battery module

```
Input  : Consumer $j$'s load profile, $l_j(t)$, consumer $j$'s PV generation profile, $PV_j(t)$
Output: Schedule of consumer $j$'s battery operation, $P_j^{batt}(t)$
foreach time period t do
    if $t=1$
        if $PV_j(t) > l_j(t)$
            the battery is charged at the minimum of the set $\{PV_j(t) - l_j(t), P_{chg}^{max}, \frac{SOC_j^{max}}{\eta_{chg}\Delta t}\}$
        end
    else
        if $PV_j(t) \geq l_j(t)$
            the battery is charged at the minimum of the set
            $\{PV_j(t) - l_j(t), P_{chg}^{max}, \frac{SOC_j^{max} - SOC_j(t-1)}{\eta_{chg}\Delta t}\}$
        else
            the battery is discharged at the maximum of the set
            $\{PV_j(t) - l_j(t), P_{dischg}^{max}, \eta_{dischg}\frac{SOC_j^{min} - SOC_j(t-1)}{\Delta t}\}$
        end
    end
    Update the battery's operation $P_j^{batt}(t)$ and state of charge, $SOC(t)$ at each period.
end
```

In Alg. 4, P_{chg}^{max} and P_{dischg}^{max} are the maximum charging and discharging rates (P_{dischg}^{max} is negative), SOC_i^{max} and SOC_i^{min} are the maximum and minimum State Of Charge of consumer i's battery, and η_{chg} and η_{dischg} are the battery's charging and discharging efficiency. We see that when $PV_j(t) = l_j(t)$ then the battery neither charges or discharges. Similarly, when the storage capacity $SOC_j(t-1) = SOC_j^{max}$ then the battery cannot charge at period t and when $SOC_j(t-1) = SOC_j^{min}$ the battery cannot discharge. Each kWh of battery storage can cycle at 85% depth-of-discharge, fully charge or

Figure 5.9: (a) Illustrating solar generation (PV) self-sufficiency - the area of the yellow region divided by the integral of the blue line. Red illustrates export (b) A specific consumers self-sufficiency against battery size.



discharge within 2 hours (a C-rate of 0.5C) and has an average round trip efficiency of 90%. These characteristics are typical of Li-ion batteries for residential energy storage [3]. Figure 5.9 shows how PV contributes to consumer self-sufficiency and the effect of a battery.

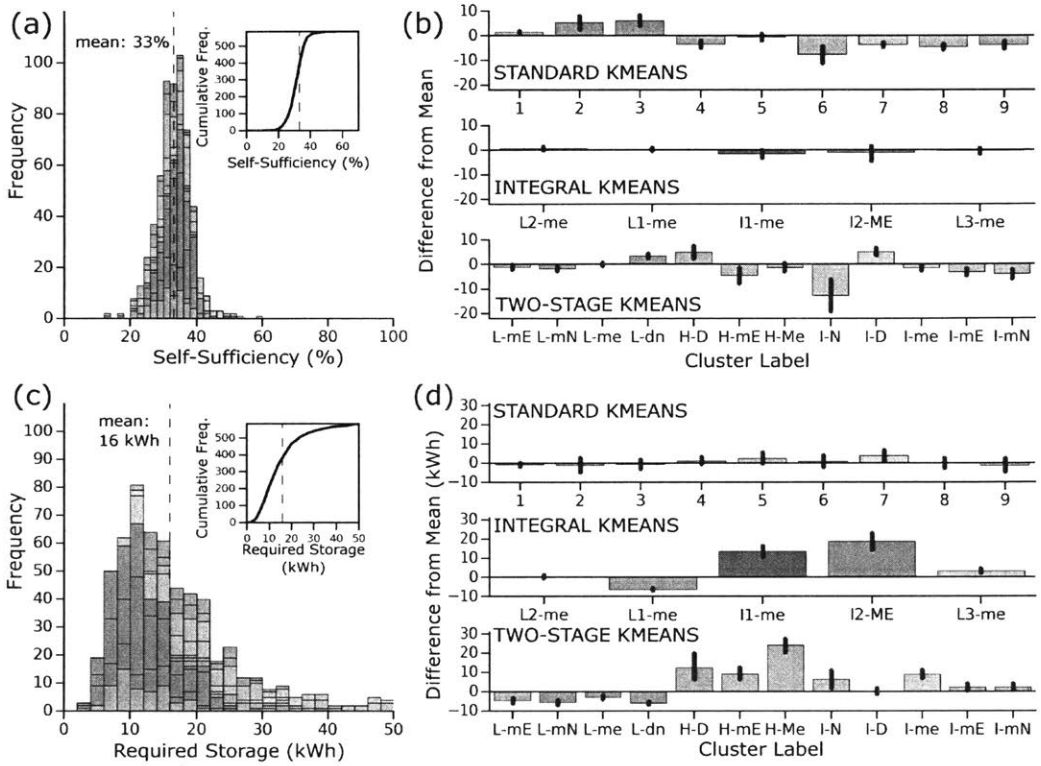
After calculating the self-sufficiency and storage values for each consumer, we separate these results by segment, assigning each consumer to the mode of the segments assigned to their daily load shapes. The best segmentation method is the one that groups consumers that separate from the mean in their solar and battery needs. For each segmentation method, we assign one non-outlier cluster for the 585 consumers (each with over 25 days of data).

Figure 5.10 b-d shows the results for all the consumers and the three different segmentation methods: standard *k*-means, integral *k*-means and the proposed two-stage *k*-means. Figure 5.10 a shows the distributions of self-sufficiency with PV only, and 5.10 b the required battery capacity for all consumers (colored by cluster from the two-stage *k*-means method). Note that the bar plots in 5.10 b-d illustrate the separation of each cluster from the mean of the respective property for each method. Both standard *k*-means and two-stage *k*-means find clusters with significant separation in terms of their self-sufficiency (Figure 5.10b), while integral *k*-means provides little insight for PV. Conversely, for battery size, integral *k*-means gives good separation (Figure 5.10d), as does the two-stage method, while the standard *k*-means provides little insight. Overall the two stage methods is performs better to separate customers to be targeted for both battery and solar services.

As would be expected, the clusters with daytime peaks (L-dn, H-D and I-D) have the highest self-

[3] see https://www.tesla.com/powerwall

Figure 5.10: (a) Distribution of monthly self-sufficiency values for all consumers (PV-only). (b) Segmentation of self-sufficiency for each segmentation method. (c) Distribution of required battery size for all consumers for 75% self-sufficiency. (d) Segmentation of the battery sizes for each segmentation method. In (a) and (c) the colors under the curve indicate the clusters of the two-stage k-means. This method separates well customers respect to their solar capacity required via their self-sufficiency and their storage needs.



sufficiency and those with peaks at night (L-mN, I-N and I-mN) have the lowest. The I-N cluster is the least suitable for only PV. Indeed, as seen in Fig. 5.7, the centroid has very little electricity use in daylight hours and belongs to the least populous non-outlier cluster. For the standard *k*-means method, we note that centroids 1-3 have higher self-sufficiency, corresponding to the daytime peaks while centroids 4-9 have lower self-sufficiency. The fact that integral *k*-means produces clusters with no real separation in misalignment demonstrates that self-sufficiency is independent of consumption magnitude when net solar production is equivalent to usage.

For estimating the battery size required for a consumer, consumption magnitude rather than peak time is the most relevant property. All of the high consumption clusters (H) require the largest batteries irrespective of peak time, followed by the intermediate clusters (I) and then the low consumption clusters (L). This is clear when inspecting the integral *k*-means clusters. Within these use brackets, the clusters that have peaks during the daytime (L-dn and L-me) generally require less storage. However, this is a second order effect when compared with the use bracket.

These results show that it is important to consider both peak time and overall consumption within the segmentation methodology. The proposed two-stage method accomplishes this, providing a useful mechanism for utilities to identify the best system size for their consumers' needs. We successfully segment consumers for PV based only on each consumer's most frequent cluster. These findings help utilities to make a good recommendation on PV suitability or battery size based on daily load shapes.

5.7 Conclusion

In this chapter, we introduced metrics to measure the representative accuracy of a segmentation methodology in terms of peak time and overall consumption. We surveyed several methods in their ability to cluster well with respect to these metrics, proposing a two-stage household energy segmentation method that, in contrast to prior work, is able to successfully segment consumers by both peak time and overall consumption.

In regards to the application of PV battery systems, we found that only the two-stage method gives clusters with significant separation for both solar self-sufficiency (most strongly affected by peak time) and required battery size (most strongly affected by overall consumption), thus illustrating the importance of clustering in terms of both aspects of the consumer load profile.

In addition to PV, these results have implications for utility policies ranging from demand response to energy efficiency. Using these consumer segments, we can effectively assess consumer needs for various programs and pricing packages. For instance, a segmentation based on peak times and consumption levels enables us to efficiently target consumers for demand response, concentrating on those with the highest potential to generate savings.

While in this chapter we analyzed the PV and storage suitability of different residential consumer segments, there is a need to do the same for non-residential energy consumers, for whom solar PV and storage may hold an especially high savings potential. In addition, though we made no assumptions regarding the structure of electricity tariffs in this chapter, having access to the exact tariff structure and pricing levels would allow the evaluation of direct financial benefits for the consumer segments obtained from the clustering methodology.

Conclusion and Future Work 6

We have proposed several descriptive methodologies to understand individual behavior in the context of shopping behavior, mobility, and energy segmentation. This includes methods that focus on interpretability in order to understand the underlying mechanisms in human dynamics (Chapter 1), mine lifestyles by connecting separate passive datasets (Chapters 2 and 3), and provide recommendations to consumers according to their needs(Chapters 3 and 4).

6.1 Summary

6.1.1 Modeling Bursty Human Dynamics and Lifestyles

Human actions determine the dynamics of a wide range of complex systems, ranging from social to economic phenomena. Understanding what drives these actions is still an open question in the field, with many possibilities for future work. Current models of human dynamics focus solely on assessing goodness of fit for a hypothesized mechanism of behavior. In contrast to these one-dimensional models, we propose the multidimensional periodic Hawkes process, coupling periodicity and the interdependence between varying types of activity. This model actually describes the transition between tasks, and even lends insight into their priority by capturing the order in which they occur. It captures the strength of dependencies between these tasks directly from data, and also learns the nonhomogeneous rates due to periodicity. We have seen that the proposed model is a good statistical fit for the shopping behavior of most individuals in empirical data, and performs well in prediction.

In addition, we find that we can connect multiple perspectives of human behavior, revealing actual lifestyles in urban regions. We see that these lifestyles extend beyond shopping behavior, with a strong connection to mobility, social behavior, and demographics of individuals.

6.1.2 Energy Segmentation

In this chapter, we introduced much-needed metrics find the representative accuracy of a segmentation methodology, where previously none existed. These metrics measure representativeness with respect to both the peak times of energy consumption, as well as the overall magnitude of consumption. These aspects are crucial to applications ranging from demand response to PV and storage recommendation. Surveying several standard and recent methods using these metrics, we see that these methods group households only by peak time, disregarding overall consumption. We introduce an algorithm, two-stage k-means, which is able to successfully segment consumers with regards to both metrics. Additionally, we find that this two-stage method results in segments that have significant differences for both PV self-sufficiency and required battery size. Thus, using these groups, we can effectively understand consumer needs.

6.2 Future Work

There is a rich history in the modeling and analysis of passively collected data. As we have seen, we can construct good predictive models for this data that simultaneously lead to a deeper understanding of individual behavior. One such interesting extension of Chapter 1 would be to model human mobility using a comparable stochastic process, comparing temporal patterns in movement to patterns found in shopping behavior here.

Beyond the level of the individual, the multidimensional periodic Hawkes process has the potential to extend directly to the broader field of complex systems. Recent work [52] uses the proposed model without periodicity to find cascades of activity in communication networks. Other applications range from predicting traffic in road segments to understanding the interdependency of events in economic markets.

In terms of model formulation, it would be interesting to experiment with algorithms that combine generative models with the Hawkes process, as does with [55] a Markov modulated Poisson process and latent Dirichlet allocation. Such a method is able learn from shared behaviors while operating in continuous time.

Our work in modeling human lifestyles and dynamics is part of a growing field. This field is of continuing significance as the underlying patterns of human actions touches on the wide range of work in transportation, urban growth, complex systems and beyond. We expect this interest to hold, and hope to explore the many open questions that remain.

Bibliography

[1] Final segmentation report–california public utilities commission. Technical report, Opinion Dynamics Corp., 2010.

[2] Dataport, 2016.

[3] J. Abate and W. Whitt. Asymptotics for m/g/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25(1-4):173–233, 1997.

[4] T. Aledavood, E. López, S. G. Roberts, F. Reed-Tsochas, E. Moro, R. I. Dunbar, and J. Saramäki. Daily rhythms in mobile telephone communication. *PloS one*, 10(9):e0138098, 2015.

[5] F. Aurenhammer. Voronoi diagrams&mdash;a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23(3):345–405, Sept. 1991.

[6] E. Bacry, K. Dayri, and J.-F. Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):157, 2012.

[7] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207, 2005.

[8] C. Beckel, L. Sadamori, T. Staake, and S. Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[13] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[14] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[15] S. Chen, H. B. Gooi, and M. Wang. Sizing of energy storage for microgrids. *IEEE Transactions on Smart Grid*, 3(1):142–151, 2012.

[16] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems*, 19(2):1232–1239, May 2004.

[17] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[18] A. Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[19] R. Di Clemente, M. Luengo-Oroz, M. Travizano, B. Vaitla, and M. C. Gonzalez. Sequence of purchases in credit card data reveal life styles in urban populations. *arXiv preprint arXiv:1703.00409*, 2017.

[20] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, Sept. 2009.

[21] M. Espinoza, C. Joye, R. Belmans, and B. D. Moor. Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. *IEEE Transactions on Power Systems*, 20(3):1622–1630, Aug 2005.

[22] L. L. et al. Behavioral assumptions underlying california residential energy efficiency programs. Technical report, CIEE Energy & Behavior Program, Berkeley, CA, 2009.

[23] H. Farhangi. The path of the smart grid. *IEEE power and energy magazine*, 8(1), 2010.

[24] E. A. Feinberg and D. Genethliou. Load forecasting. In *Applied mathematics for restructured electric power systems*, pages 269–285. Springer, 2005.

[25] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems*, 20(2):596–602, May 2005.

[26] C. Flath, D. Nicolay, T. Conte, C. van Dinther, and L. Filipova-Neumann. Cluster analysis of smart metering data. *Business & Information Systems Engineering*, 4(1):31–39, 2012.

[27] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[28] E. Guerra. The geography of car ownership in mexico city: a joint model of households' residential location and car ownership decisions. *Journal of Transport Geography*, 43:171–180, 2015.

[29] S. Haben, C. Singleton, and P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7(1):136–144, Jan 2016.

[30] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2013.

[31] C. A. Hidalgo R. Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems. *Physica A: Statistical Mechanics and its Applications*, 369(2):877–883, 2006.

[32] T. Hu, R. Song, Y. Wang, X. Xie, and J. Luo. Mining shopping patterns for divergent urban regions by incorporating mobility data. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 569–578. ACM, 2016.

[33] G. W. Irwin, W. Monteith, and W. C. Beattie. Statistical electricity demand modelling from consumer billing data. *IEE Proceedings C - Generation, Transmission and Distribution*, 133(6):328–335, September 1986.

[34] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

[35] Z.-Q. Jiang, W.-J. Xie, M.-X. Li, W.-X. Zhou, and D. Sornette. Two-state markov-chain poisson nature of individual cellphone call statistics. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(7):073210, 2016.

[36] H.-H. Jo, M. Karsai, J. Kertész, and K. Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.

[37] H.-H. Jo, J. I. Perotti, K. Kaski, and J. Kertész. Correlated bursts and the role of memory range. *Physical Review E*, 92(2):022814, 2015.

[38] M. Karsai, H.-H. Jo, and K. Kaski. *Bursty human dynamics*. Springer.

[39] C. Krumme, A. Llorente, M. Cebrian, E. Moro, et al. The predictability of consumer visitation patterns. *arXiv preprint arXiv:1305.1120*, 2013.

[40] J. Kwac, J. Flora, and R. Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1):420–430, Jan 2014.

[41] J. Kwac, J. Flora, and R. Rajagopal. Lifestyle segmentation based on energy consumption data. *IEEE Transactions on Smart Grid*, PP(99):1–1, 2016.

[42] M.-X. Li, Z.-Q. Jiang, W.-J. Xie, S. Micciché, M. Tumminello, W.-X. Zhou, and R. N. Mantegna. A comparative analysis of the statistical properties of large mobile phone calling networks. *Scientific reports*, 4:5132, 2014.

[43] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[44] J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[45] F. Liu and H. J. Lee. Use of social network information to enhance collaborative filtering performance. *Expert systems with applications*, 37(7):4772–4778, 2010.

[46] A. Llorente, M. Garcia-Herranz, M. Cebrian, and E. Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

[47] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4:5276, 2014.

[48] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, 2008.

[49] N. Masuda, T. Takaguchi, N. Sato, and K. Yano. Self-exciting point process modeling of conversation event sequences. In *Temporal Networks*, pages 245–264. Springer, 2013.

[50] A. F. McDaid, D. Greene, and N. Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

[51] S. Morse, M. C. González, and N. Markuzon. Persistent cascades: Measuring fundamental communication structure in social networks. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 969–975. IEEE, 2016.

[52] S. T. Morse. *Persistent cascades and the structure of influence in a communication network.* PhD thesis, Massachusetts Institute of Technology, 2017.

[53] S. J. Moss. Market segmentation and energy efficiency program design. Technical report, CIEE Energy & Behavior Program, Berkeley, CA, 2008.

[54] J. G. Oliveira and A.-L. Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251, 2005.

[55] J. Pan, V. Rao, P. Agarwal, and A. Gelfand. Markov-modulated marked poisson processes for check-in data. In *International Conference on Machine Learning*, pages 2244–2253, 2016.

[56] D. Pennacchioli, M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi. The retail market as a complex system. *EPJ Data Science*, 3(1):33, 2014.

[57] B. D. Pitt and D. S. Kitschen. Application of data mining techniques to load profiling. In *Proceedings of the 21st International Conference on Power Industry Computer Applications. Connecting Utilities. PICA 99. To the Millennium and Beyond (Cat. No.99CH36351)*, pages 131–136, Jul 1999.

[58] G. J. Ross and T. Jones. Understanding the heavy-tailed dynamics in human behavior. *Physical Review E*, 91(6):062809, 2015.

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

[60] T. F. Sanquist, H. Orr, B. Shui, and A. C. Bittner. Lifestyle factors in u.s. residential electricity consumption. *Energy Policy*, 42:354 – 364, 2012.

[61] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu. Inferring gas consumption and pollution emission of vehicles throughout a city. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1027–1036. ACM, 2014.

[62] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 650–658, New York, NY, USA, 2008. ACM.

[63] V. K. Singh, L. Freeman, B. Lepri, and A. S. Pentland. Predicting spending behavior using socio-mobile features. In *Proceedings of the 2013 International Conference on Social Computing*, SOCIALCOM '13, pages 174–179, Washington, DC, USA, 2013. IEEE Computer Society.

[64] M. R. Solomon, D. W. Dahl, K. White, J. L. Zaichkowsky, and R. Polegato. *Consumer behavior: Buying, having, and being*, volume 10. Pearson, 2014.

[65] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[66] B. Sütterlin, T. A. Brunner, and M. Siegrist. Who puts the most energy into energy conservation? a segmentation of energy consumers based on energy-related behavioral characteristics. *Energy Policy*, 39(12):8137 – 8152, 2011. Clean Cooking Fuels and Technologies in Developing Economies.

[67] M. Ten Thij, S. Bhulai, and P. Kampstra. Circadian patterns in twitter. *Data Analytics*, pages 12–17, 2014.

[68] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.

[69] V. Traag. *Algorithms and Dynamical Models for Communities and Reputation in Social Networks.* Springer, 2014.

[70] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

[71] L. Villagram. For most mexicans, the digital age is still out of reach. 2012.

[72] C. F. Walker and J. L. Pokoski. Residential load shape modelling based on customer behavior. *IEEE Transactions on Power Apparatus and Systems*, PAS-104(7):1703–1711, July 1985.

[73] S. Xu, E. Barbour, and M. C. González. Household segmentation by load shape and daily consumption. *ACM SIGKDD International Workshop on Urban Computing*, 2017.

[74] S. Xu, R. Di Clemente, and M. C. González. Mining urban lifestyles: urban computing, human behavior and recommender systems. In O. Khalid, S. U. Khan, and A. Y. Zomaya, editors, *Big Data Recommender Systems: Recent Trends and Advances.* Institution of Engineering and Technology (IET), Submitted.

[75] T. Yasseri, G. Quattrone, and A. Mashhadi. Temporal analysis of activity patterns of editors in collaborative mapping project of openstreetmap. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 13. ACM, 2013.

[76] T. Yasseri, R. Sumi, and J. Kertész. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1):e30091, 2012.

[77] W. Zhang, T. Yoshida, and X. Tang. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.

[78] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *AAAI*, volume 10, pages 236–241, 2010.

[79] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, pages 1029–1038. ACM, 2010.

[80] Y. Zheng. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data*, 1(1):16–34, 2015.

[81] Y. Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

[82] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[83] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309, 2013.

[84] T. Zhou, Z.-D. Zhao, Z. Yang, and C. Zhou. Relative clock verifies endogenous bursts of human dynamics. *EPL (Europhysics Letters)*, 97(1):18006, 2012.