RUSH! Targeted Time-limited Coupons via Purchase Forecasts

Emaad Manzoor
Carnegie Mellon University
emaad@cmu.edu

Leman Akoglu
Carnegie Mellon University
lakoglu@cs.cmu.edu

ABSTRACT

Time-limited promotions that exploit consumers' sense of urgency to boost sales account for billions of dollars in consumer spending each year. However, it is challenging to discover the right timing and duration of a promotion to increase its chances of being redeemed. In this work, we consider the problem of delivering time-limited discount coupons, where we partner with a large national bank functioning as a commission-based third-party coupon provider. Specifically, we use large-scale anonymized transaction records to model consumer spending and forecast future purchases, based on which we generate data-driven, personalized coupons. Our proposed model RUSH! (1) predicts both the time and category of the next event; (2) captures correlations between purchases in different categories (such as shopping triggering dining purchases); (3) incorporates temporal dynamics of purchase behavior (such as increased spending on weekends); (4) is composed of additive factors that are easily interpretable; and finally (5) scales linearly to millions of transactions. We design a cost-benefit framework that facilitates systematic evaluation in terms of our application, and show that RUSH! provides higher expected value than various baselines that do not jointly model time and category information.

KEYWORDS

transaction data, targeted promotions, point processes, cost-benefit

1 INTRODUCTION

321 billion coupons were delivered to consumers in 2015 [17], and draw in billions of dollars in spending every year [5]. Digital coupons transmitted via email or smartphones remain nascent, comprising only 0.6% of the total volume delivered, despite the fact that coupons delivered digitally were found to be *over 30 times* more likely to be redeemed than traditional paper alternatives [5]. As such, opportunities to further improve redemption rates via data-driven personalization and targeting appear plentiful, but remain untapped. This is primarily due to the barriers in information-sharing between competing businesses that limit the construction of complete consumer profiles.

In this work, we tap into the power of comprehensive and large-scale transaction data in order to *promote data-driven, personalized discount coupons* to users by forecasting their future purchases *based on their spending history*.[1] To do so, we partner with a national bank that provides us with a large, anonymized database that comprises transactions from ~200,000 customers over a period from September 2013 to January 2016. Transactions include timestamps, merchant category codes, and dollar amounts. Building on the transaction history, we design a predictive model called RUSH![2] to estimate *both the next transaction time as well as its purchase category* for a given customer. More specifically, we estimate the time-of-next-purchase and generate a digital coupon (to be delivered on a smartphone) with 1-3 merchants from the most probable purchase categories as estimated by our model (see, for example, Fig. 3). Informally, the problem we address is as follows.

Informal Problem 1. *Given a sequence of millions of transaction triplets (customer, purchase category, time stamp), e.g., (Alice, Grocery, 2/18/17 14:10); for a given customer like Alice,*

- *Predict a time interval that contains her next purchase.*
- *Identify relevant discount coupons to deliver.*

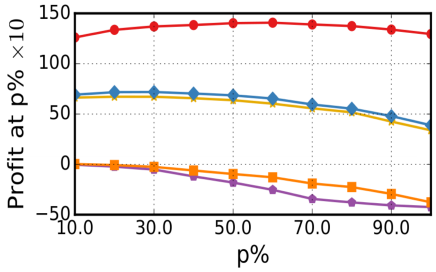
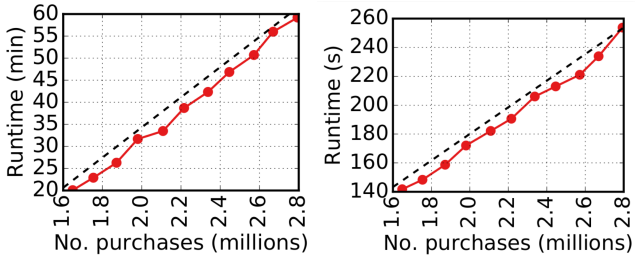


Figure 1: RUSH! (red) provides higher expected profit than several baselines, when delivering coupons 1 hour long containing 2 purchase categories, as the *p%* most confident predicted purchases are selected for coupon delivery. See §4.5 for details.



(a) Training time by data size **(b)** Precomputation time

Figure 2: RUSH! scales linearly to large datasets containing millions of transactions, for both the time to (a) train and the (b) precomputation before training.

With our proposed approach, we aim to empower both our partnering bank as well as its customers—we expect timely personalized

[1] Transaction, purchase and spending are used interchangeably in this paper.
[2] Code available at http://github.com/emaadmanzoor/rush/

KDD '17, August 13-17, 2017, Halifax, NS, Canada

DOI: 10.1145/3097983.3098104

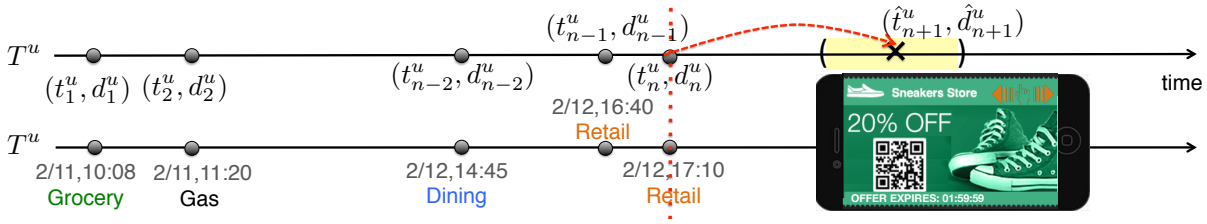


Figure 3: (top) Timeline for user u with ordered transactions $\{(t_i^u, d_i^u)\}_{i=1}^n$, where t_i^u are timestamps and d_i^u are merchant categories, and (bottom) an example realization. Our model RUSH! constructs an interval (highlighted in yellow) around the predicted next purchase time ($\hat{t}_{n+1}^u$, black cross), as well as a predictive distribution over merchant categories, based on which it delivers a personalized, time-limited digital coupon to user u.

coupons to help customers save on their purchases, while the bank to earn commission-based revenue from the redemption of each delivered coupon at participating merchants. Our key objective hence is to maximize the redemption rate of delivered coupons via data-driven personalization and targeting. We focus specifically on *time-limited* coupons characterized by a delivery time, duration, and a merchant category. Time constraints are widely believed to accelerate coupon redemption by invoking feelings of urgency and scarcity [1, 16]. A challenge with limiting usage time is the sensitivity of redemption rates to the exact delivery time and coupon duration. A coupon delivered too early with respect to the intended purchase time will be forgotten, one delivered too late will be unused, and one valid for too long triggers no urgency. Intuitively, a useful time-limited coupon would (*i*) be delivered close to the intended purchase time, (*ii*) provide a coupon for the intended purchase category, and (*iii*) last just long enough to capture the uncertainty in the intended purchase time.

Accurately targeting time-limited coupons is a challenging prediction problem, one of estimating both the (continuous) time and the (discrete) category of purchases. Consumer behavior is highly non-stationary in time and susceptible to various exogenous factors such as time-of-day, receipt of pay checks, occurrence of holidays and other personal occasions. Non-stationarity in purchase categories further exacerbates this problem: if no causal pattern exists (for example, between insurance payments and having coffee), consumers may arbitrarily swap the order of purchase categories over time, thus confusing ill-equipped models. Moreover, individual consumers typically make only a few purchases a week, leading to very sparse consumption timelines that pose difficulties in estimating complex models. In short, our goal is to design a method that can predict both event time and category, adapt to temporal dynamics, handle event sparsity, that is also scalable and interpretable.

The proposed work makes the following notable contributions:

(1) **Problem Formulation: Promoting Digital Coupons.** As an example for "turning data into business value via predictive analytics", we formalize the problem of promoting personalized time-limited coupons to bank customers, which puts to use a large collection of transactions data from a national bank. Done effectively, such an application is expected to benefit all parties; by helping the customers save, the bank to raise profit from commissions, and the participating merchants to receive foot traffic.

(2) **Modeling Transaction Data for Purchase Forecasting.** We model the purchase time and category inference problem using continuous-time point processes, which are a natural fit to the continuous temporal nature of our data. Analyzing the purchases in our data reveals (*i*) *time-varying aspects* of consumer spending behavior (such as increased purchase rates on Fridays, see Fig. 6) and (*ii*) *triggering effects* or excitation among purchases (such as shopping purchases triggering dining purchases in the near future, see Fig. 7). We capture both phenomena via additive linear augmentations of the conditional intensity function. Thanks to this additive nature, RUSH! is interpretable and provides insights that are valuable for consumer profiling. Moreover, RUSH! scales (linearly) to datasets with millions of transactions (see Fig. 2).

(3) **Real-Data Experiments and Cost-Benefit Analysis.** For our specific application, we cannot use traditional performance evaluation measures such as the mean absolute or root mean squared errors often used with real-valued prediction tasks. First, there are different cost-benefit trade-offs for under and over-predicting the purchase time (analogous to false positives and negatives in discrete predictions). Moreover, we do not simply make point predictions; our coupons come with a duration in which they can be redeemed. Second, we predict both purchase time (continuous, real-valued) and category (discrete), which complicates the matter further (a coupon at the right time but for merchants in the wrong category is useless). Therefore, we design a *cost-benefit framework* for quantifying model performance for our intended application, where we try to come as close as we can to simulating our model in production[3]. We show that RUSH! provides better trade-offs over baseline methods that ignore time-varying factors or purchase history (see Fig. 1).

2 PROBLEM OVERVIEW

We now formalize the problem of targeting time-limited coupons. We also introduce the notation that is used through the rest of this paper. Let $[T_0, T)$ be the window of observation wherein we observe all the transactions of every consumer. We may assume $T_0 = 0$ without loss of generality. Each consumer u is represented

[3]Claudia Perlich on evaluation in the real world: http://www.kdnuggets.com/2016/12/interviews-data-scientists-claudia-perlich.html

by a continuous-time sequence of transaction timestamps (interchangeably called events) $T_u = \{t_1, \ldots, t_n\}$ and their associated categories (interchangeably called dimensions) $D_u = \{d_1, \ldots, d_n\}$ (for example, grocery, dining, etc.). Transaction categories are derived from the Merchant Category Codes associated with transactions and form a finite set $\mathcal{D}$. We also associate with time t a set of binary *temporal features* $f_j \in \mathcal{F}$ where $f_j(t) \in \{0, 1\}$, $\forall t$. These features are functions of the wall-clock time at t; for example, if t lies on a weekend, or if t lies on a federal holiday. We now define the Coupon problem as that of one-step-ahead, next purchase prediction.

Problem 1 (Coupon). ***Given*** *a collection of purchase timestamps $\mathcal{T}$ and categories $\mathcal{D}$ for U consumers observed in a window $[0, T)$, let t_n be the last observed purchase for any consumer and t_{n+1} be the unobserved future purchase. At t_n, for each consumer,* ***forecast*** *a coupon $C = \{\hat{t}^{start}_{n+1}, \hat{t}^{end}_{n+1}, D_{n+1}\}$ that is valid from $\hat{t}^{start}_{n+1}$ to $\hat{t}^{end}_{n+1}$ and contains k offers from merchants chosen from categories $D_{n+1} = \{d_1, \ldots, d_k\} \subset \mathcal{D}$; such that its redemption probability is maximized.*

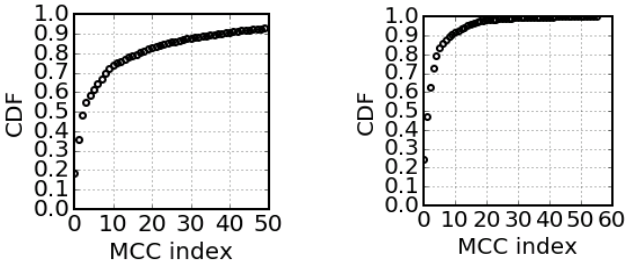
Given a coupon forecasted by any method solving Coupon, for the probability of the coupon being redeemed to be high, each of the following intermediate results must be accurate:

(1) The probability densities of the predicted purchase times $\hat{t}_{n+1}$ must be high at the actual purchase times t_{n+1}.
(2) Point predictions of the predicted purchase time must minimize the absolute prediction error $|\hat{t}_{n+1} - t_{n+1}|$.
(3) The predicted coupon interval $(\hat{t}^{\text{start}}_{n+1}, \hat{t}^{\text{end}}_{n+1})$ must trap the actual purchase time, i.e., $t_{n+1} \in (\hat{t}^{\text{start}}_{n+1}, \hat{t}^{\text{end}}_{n+1})$.
(4) The probability density of the predicted purchase category $\hat{d}_{n+1}$ must be high at the actual purchase category d_{n+1}.

3 DATA

Description. Our partner bank provides a variety of financial services to customers across the United States, such as checking, savings, loan, credit and debit accounts. In this work, we focus on customers who hold *prepaid card* accounts. Prepaid cards expand the reach of banking services to the "underbanked", including 26.9% of all U.S. households (over 60 million adults) as of 2015 [14]. The underbanked are restricted from holding traditional checking and savings accounts due to poor liquidity and credit history. Such customers tend to rely on a single prepaid card for all of their transactions, including deposits of income. Since these individuals are less likely to have other accounts (unlike others who could potentially have a number of credit cards across different banks), data collected from prepaid card customers represent a *near-complete picture of their financial activities* and is thus very suitable for modeling spending behaviors. This near-complete history of transaction logs from prepaid card users was earlier leveraged for studying the existence of a "payday effect" on spending behavior [29].

Our data contains 199,109 prepaid card accounts with a total of 24,858,748 transactions (excluding non-purchases such as cash withdrawals and service charges), spanning a time period from September 2013 to January 2016. The length of the transaction history of a customer varies between an hour to 2.3 years, and the total number of transactions per customer varies between 1 and 4,047. Each transaction is associated with a dollar amount, a timestamp (at the granularity of seconds) and a Merchant Category Code (MCC). MCCs are 4-digit numbers assigned to businesses by credit card companies to classify their primary industry. A comprehensive listing of MCCs and their descriptions are available at https://github.com/greggles/mcc-codes.



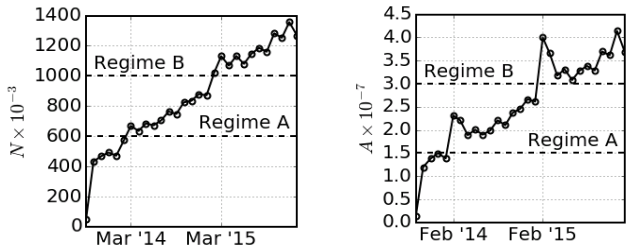
(a) MCCs with 4 digits **(b)** MCCs grouped by first 2 digits

Category	Example Purchases	MCCs	Avg. Amt.
DINING	fast-food, restaurants	5811-5814	$12.60
SERVICES	laundry, postal	7, 60, 72 73, 80, 82 83, 86, 93, 94	$60.22
RETAIL	non-grocery shopping	5815-5818 5832, 50, 52 56, 57, 59	$31.70
GROCERIES	produce, bakeries	53, 54	$34.23
AUTO	gas-station	55, 75	$20.96
UTILITIES	electricity, rent, internet	48, 49, 63	$97.60
ENTERTNMT	theatre, bowling, lottery	78, 79	$20.32
TRAVEL	air and rail travel	30-34,40 44-47	$131.00
HOTEL	hotels, B&Bs	35-38, 65	$170.10
COMMUTE	public buses	41	$21.04

(c) Purchase categories and MCCs; also given are the average dollar amount spent on each category per purchase.

Figure 4: Cumulative fraction of transactions associated with each category derived from (a) all 4 and (b) the first 2 MCC digits. Categories are indexed from 0 in order of their frequency of occurrence in the data. (c) Our category-MCC mapping: 2-digit codes in the table represent all MCCs having the same first 2 digits.

Preprocessing. MCCs can be used to derive purchase categories, however with several challenges. Purchases in our data encompass 557 unique MCCs with a highly skewed distribution: the top 50 most frequent MCCs in the data account for over 90% of all purchases (Fig. 4 (a)). We also observe that MCCs follow a hierarchical structure: for example, MCCs 3000 and 3001 correspond to United Airlines and American Airlines respectively, and all MCCs beginning with 30, 31 and 32 (close to 300 of them) can be grouped under a broad "airlines" category. Deriving categories by grouping MCCs using their first 2 digits would result in 56 unique categories (distributed as in Fig. 4 (b)). However, this grouping fails to separate certain important and frequently occurring categories (such as dining and retail shopping), and fails to combine certain categories that exhibit similar spending behavior (such as theatre and orchestras, broadly "entertainment"). Therefore, we start with the 2-digit grouping and manually split or combine categories as



(a) Total no. N of purchases/month **(b)** Total amount A spent/month

Figure 5: Discontinuity in overall spending behavior between different years, with notable increases in February and March each year, possibly caused by promotions at work or tax-refunds.

appropriate to construct an MCC-category mapping that comprises 10 broad purchase categories (listed in Fig. 4 (c)).

There is also a significant change in overall spending behavior across years. Fig. 5 visualizes the total number of purchases and total amount spent by all customers over the entire time-frame of our data. The general increasing trend in both plots is due to new customers being acquired. However, there is a discontinuity in March (for total number of transactions) and in February (for total amount spent) each year that splits overall purchase behavior into different regimes. To avoid issues with data spanning different regimes, we select purchases taking place in regime A from February 2014 to February 2015. During this period, new customers are acquired and some become inactive. To avoid issues stemming from anomalous purchase behavior when customers first open their accounts or stop using them, we select customers with at least one transaction before February 1, 2014 and at least one transaction after February 1, 2015. The final subset of the data contains 2,808,360 transactions from 7,719 unique customers.

4 MODELING PURCHASE BEHAVIOR

To model purchases taking place in continuous-time, we adopt the framework of temporal point processes [8]. Unlike various discrete sequence models, point processes naturally accommodate events in continuous time, and can be augmented to capture (i) time-variation in event-occurrence rates and (ii) sequential correlation among events of different types. Formally, a temporal point process is a stochastic process, the realization of which is an ordered sequence of event timestamps $\{t_i\} \subset [0, \infty)$. When extended to D dimensions, each timestamp t_i lies on dimension $d_i \in \{1, \ldots, D\}$. In our scenario, events are purchases and dimensions are purchase categories. Temporal point processes are characterized by their conditional intensity function $\lambda^*(t)$ which, informally, denotes the probability that an event occurs in a small interval $\mathrm{d}t$ conditioned on the history of events before t, $\mathcal{H}_t = \{(t', d') | t' < t\}$,

$$\lambda^*(t)\mathrm{d}t = \mathbb{P}\{\text{event in } (t, t + \mathrm{d}t) | \mathcal{H}_t\}. \tag{1}$$

The simplest point process is the homogeneous Poisson process parameterized by a constant positive base-rate λ_0, independent of the purchase history. It has a conditional intensity function,

$$\lambda^*(t) = \lambda_0, \qquad \lambda_0 > 0. \tag{2}$$

Table 1: Temporal features

Feature index j	**Binary Feature** $f_j(t)$
1-24	Hour of the day at t = 00-23
25	Day of the week at $t \in$ {Mo, Tu, We, Th}
26	Day of the week at t = Fri
27	Day of the week at $t \in$ {Sa, Su}
28	Day of the month at $t = 1$ (often pay-day)

While the homogenous Poisson process (P) is particularly well-suited to model recurring purchases such as rent and insurance payments, it does not account for two key factors that affect the probability of a purchase at a given time: (i) the wall-clock date/time (time-variation), and (ii) the purchase history up to that time (memory). In the rest of this section, we pave the way towards our proposed model by gradually increasing the modeling complexity by incorporating time-variation, memory, and the ability to predict categories.

4.1 Incorporating Time-Variation

The probability of a purchase varies widely with the absolute time, as shown in Fig. 6. Consumers make 5 times as many purchases in their most active hours (11AM-12PM, 5PM-6PM, 2AM-3AM) than they do in their least active hours (4AM-9AM). The day of the week has a similar effect, with 1.5 times as many purchases made on Friday, Saturday and Sunday compared to other days of the week.

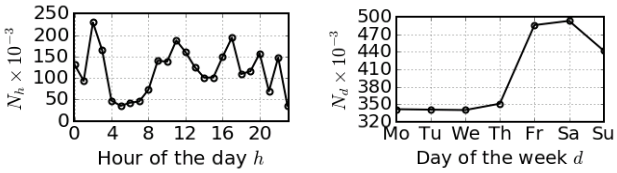


Figure 6: Spending behavior varies by time. Total number of purchases in the dataset (left) at each hour of the day and (right) on each day of the week.

To incorporate time-variation, we augment the conditional intensity with a feature-based time-varying component. Thus, a time-varying Poisson process (TVP) has conditional intensity,

$$\lambda^*(t) = \lambda_0 + \sum_{f_j \in \mathcal{F}} \mu_j f_j(t) \qquad \lambda_0 > 0,\ \mu_j \geq 0,\ f_j(t) \in \{0, 1\}. \tag{3}$$

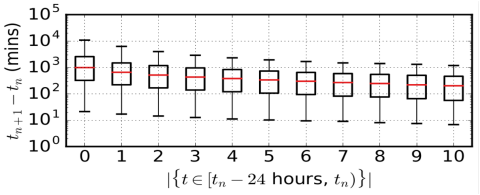
Each feature $f_j(t) \in \mathcal{F}$ is a binary function of the wall-clock time/date at t; for example, if t lies on a weekend, or on the 1st of the month (which is often the pay-day, when spending on utilities, for example, tends to increase [29]). Each μ_j is a non-negative weight that is added to the total conditional intensity at t if $f_j(t) = 1$. We define 28 such temporal features, listed in Table 1.

4.2 Incorporating Memory

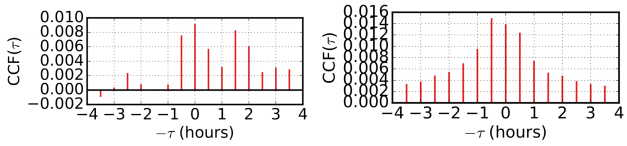
Purchases in time may be triggered due to "self-excitation", a phenomenon wherein the occurrence of a purchase (in any category) increases the probability of another in the near future. A related phenomenon is "mutual-excitation", wherein a purchase in one category increases the probability of a purchase in another category in the near future. We see evidence of such phenomena in our dataset.

Fig. 7(a) shows, given a purchase at time t_n, the distribution of the inter-purchase time $(t_{n+1} - t_n)$ conditioned on the number of purchases in the 24 hours prior to t_n. We observe that having more purchases in the previous 24 hours shifts the distribution towards smaller inter-purchase times.

We also investigate the sequential correlation between purchase categories using the cross-correlation coefficient (CCF), which measures the similarity of two time series at different lags. Fig. 7 (b) and (c) show example CCF plots at various lags (-4 to 4 hours in 30 minute increments). We see from (b) that people often Commute 30 mins before and up to 2 hrs after Dining, and (c) suggests the correlation is symmetric between Retail vs. Grocery. From these case studies, we conclude that a strong sequential correlation exists between purchases in different categories.



(a) Having more purchases in the previous 24 hours indicates a smaller inter-arrival time until the next purchase.



(b) Commute vs. Dining as reference **(c)** Retail vs. Grocery as reference

Figure 7: Evidence of self/mutual-excitation in purchase behavior.

To incorporate this sequential correlation information, or in other words the excitation effect of past purchases (memory) into our models, we adopt Hawkes processes (HP) [19]. HPs explicitly parameterize excitation, and have been used to model phenomena ranging from earthquakes [21] to financial contagion [2] and information diffusion in social networks [13]. The conditional intensity contains a self-exciting component dependent on the process history,

$$\lambda^*(t) \quad = \quad \lambda_0 + \beta \sum_{t' \in \mathcal{H}_t} e^{-\alpha(t-t')} \qquad \lambda_0, \beta, \alpha > 0. \tag{4}$$

β is the magnitude of excitation caused by events in the purchase history, and α is the rate of decay of the excitation effect. Note that augmenting (4) with the time-varying component in (3) results in a time-varying Hawkes process (TVHP).

4.3 Proposed Model

The HP as such does not account for different degrees of excitation for events of different types. To incorporate this variation in excitation as well as the variation in purchase rates over time, we augment multidimensional Hawkes processes with the time-varying component introduced in (3). Our proposed model RUSH! has a conditional intensity specified for each category $m \in \mathcal{D}$ as,

$$\lambda^*_m(t) \quad = \quad \lambda_{m0} + \mu_m(t) + \sum_{m'=1}^{D} \sum_{t' \in \mathcal{H}_t, d(t')=m'} \beta_{mm'} e^{-\alpha(t-t')} \,,$$

$$\mu_m(t) \quad = \quad \sum_{f_j \in \mathcal{F}} \mu_{mj} f_j(t) \,. \tag{5}$$

where λ_{m0} is the base rate for category m, $\beta_{mm'}$ is the magnitude of mutual-excitation (or self-excitation, if $m = m'$) of category m' on category m, and α is the rate of decay for excitation effects. The time-varying terms for each category are analogous to (3).

Extension to D dimensions allows each category to have its own base rate and time-varying component. This enables capturing, for example, the fact that groceries are often purchased during the day while entertainment purchases often occur at night. It also allows varying degrees of mutual-excitation between purchases of different categories in our data as shown in Fig. 7; we will see later (§6) that this effect is often asymmetric.

The intermediate models specified by (2)-(5) gradually increase in complexity and expressiveness. Hence, it is instructive to consider them as baselines in evaluation, to understand the contribution of each additional phenomenon on predictive performance. We differentiate the salient features of each model in Table 2 and visualize examples of their conditional intensities in Fig. 8.

Table 2: Summary of proposed and baseline models

Model	Time Varying	Has Memory	Predicts Categories
Poisson (P)	✗	✗	✗
Time-Varying Poisson (TVP)	✓	✗	✗
Hawkes (HP)	✗	✓	✗
Time-Varying Hawkes (TVHP)	✓	✓	✗
RUSH!	✓	✓	✓

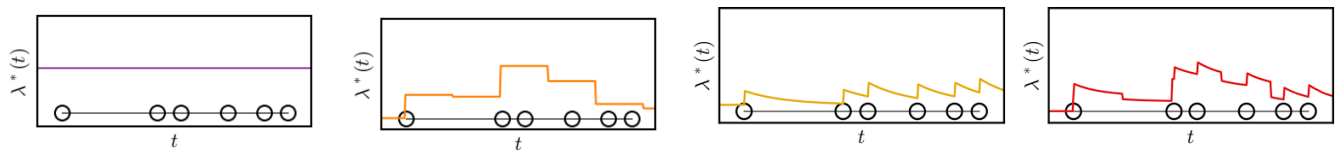
4.4 Learning and Prediction

4.4.1 Parameter Estimation. Our model is parameterized by $\boldsymbol{\Theta} = \{\boldsymbol{\lambda_0}, \boldsymbol{\beta}, \boldsymbol{\mu}\}$; dimension-specific base rates $\boldsymbol{\lambda_0}$, $(D \times D)$ excitation matrix $\boldsymbol{\beta}$ and dimension-specific time-varying feature weights $\boldsymbol{\mu}$. Consider a sequence of timestamps $\{t_1, \ldots, t_n\}$ in an observation window $[0, T)$ where the dimension of each timestamp t_i is denoted by $d(t_i) \in \{1, \ldots, D\}$. The loglikelihood of this sequence with respect to any temporal point process can be written directly in terms of its conditional intensity,

$$\mathcal{L}(\{t_1, \ldots, t_n\}) \quad = \quad \sum_{i=1}^{n} \log \lambda^*_{d(t_i)}(t_i) - \sum_{m=1}^{D} \int_0^T \lambda^*_m(\tau) \mathrm{d}\tau \,. \tag{6}$$

Loglikelihood of a dataset of sequences is the sum of the loglikelihoods for each sequence. Since it is concave in all parameters, we maximize the loglikelihood regularized with $-\gamma \|\boldsymbol{\Theta}\|_2^2$ directly using the L-BFGS-B optimization method [6]. Regularization penalty γ and decay rate α in (5) are considered as hyperparameters, selected based on a validation set.

Speeding up Learning. Though computing the loglikelihood (and gradient) is linear in the number of transactions, finding its maximum involves a large number of L-BFGS-B iterations (over 1,000 for the multivariate models), each of which traverses the entire training data.To speed up model fitting, we precompute terms in



(a) Poisson (P) **(b)** Time-Varying Poisson (TVP) **(c)** Hawkes Process (HP) **(d)** TVHP ≡ RUSH! w/ $D = 1$

Figure 8: Conditional intensities over time $\lambda^*(t)$ of the point process models and six example events (black circles).

the loglikelihood and gradient that are independent of the parameters (for example, $\int_0^T f_j(s)\mathrm{d}s$ in the gradient). For each α, we also precompute terms that depend only on α, since they do not change while optimizing the parameters. Precomputation time also scales linearly in the size of the data (see Fig. 2 (b)). We further exploit the separability of the loglikelihood across different sequences and compute it in parallel for each sequence across 100 CPU cores. The optimizations have a drastic impact: specifically, time to fit RUSH! reduces from over 52 days (estimated) to 60 minutes.

4.4.2 Prediction. Given a fitted model and the purchase history $\mathcal{H}_t$ with the last observed purchase at t_n for a specific customer, our goal is to predict the time and category of their next purchase $\{t_{n+1}, d_{n+1}\}$. We first obtain the empirical posterior distributions of the next timestamp $\hat{h}^*(t_{n+1}) = \hat{h}(t_{n+1}|\mathcal{H}_t)$ and category $\hat{g}^*(d_{n+1})$ predictions via simulating Ogata's modified thinning algorithm ([8], algorithm 7.5.IV) 100 times. The posterior distributions can then be used to obtain point predictions. We use the median of $\hat{h}^*$ to predict $\hat{t}_{n+1}$, which we find empirically minimizes the mean absolute error $|t_{n+1} - \hat{t}_{n+1}|$. We generate a coupon that can be redeemed within period $(\hat{t}_{n+1} - \Delta, \hat{t}_{n+1} + \Delta)$ where $2\Delta = \tau$ denotes its duration, and that contains k offers with the largest probabilities in $\hat{g}^*$. We investigate the impact of Δ and k on performance in §5, within a cost-benefit framework that we discuss next.

4.5 Cost-Benefit Framework

In general, mean absolute error (MAE) or root mean squared error (RMSE) are used to evaluate real-valued prediction models. In our application, however, we cannot directly use these measures, due to different cost-benefit trade-offs for under and over-predicting the purchase time. Moreover, we predict both the (real-valued) next purchase time and its category (discrete) simultaneously, which complicates the matter further (a coupon at the right time but with offers in the wrong purchase categories is useless).

In practice, the right performance measure depends on the application that determines a model's intended use[3]. For most business applications, including ours, what ultimately matters is the business value, often quantified in terms of profit. For any new model, a good performance metric is thus the excess revenue generated. To this end, we develop a cost-benefit framework [24] to quantify model performance for our intended application, which is designed to closely mimic *simulating our model in production*.

Let us consider a customer with last purchase time t_n, for which our model generates a coupon to be redeemed within $(t^{\text{start}}, t^{\text{end}})$, with duration $\tau = t^{\text{end}} - t^{\text{start}}$ and containing offers from categories $D \subset \mathcal{D}$. We model the probability that a coupon will be redeemed by a decaying function $c(\delta)$, where δ denotes the time that has elapsed since the coupon begins, in order to capture a "forgetting-rate". One such function is $c(\delta; \gamma) = \exp(-\gamma\delta)$ for $\gamma > 0$. Note that we truncate the function such that $c(\delta) = 0$ for $\delta < 0$ and for $\delta > \tau$: the probability of redemption is non-zero only while the coupon is active. The decay rate γ penalizes overly-long coupons; if τ is too large, even perfectly predicted coupons where $\hat{t}_{n+1} = t_{n+1}$ would not always be redeemed, due to being forgotten.

Success scenario. If the true purchase time is trapped within coupon's duration ($t^{\text{start}} \le t_{n+1} \le t^{\text{end}}$) *and* the true purchase category is among the predicted categories ($d_{n+1} \in D$), we incur a benefit $B(d_{n+1})$. The benefit may be different for each category based on category-specific commission-rates. The expected value of the coupon is thus $c(t_{n+1} - t^{\text{start}}) * B(d_{n+1}) - C$ where C is some constant cost of coupon generation.

Error scenarios. There are various cases when the prediction is inaccurate.

Case 1: $t_{n+1} < t^{\text{start}}$. If the true purchase occurs before the coupon begins, we simply cancel the coupon release and re-estimate the next purchase time. This incurs no benefit, but a cost of $-C$.

Case 2: $t^{\text{start}} \le t_{n+1} \le t^{\text{end}}$, $d_{n+1} \notin D$. If the coupon traps the true purchase, but the true category is not among the ones predicted, we incur a loss of customer trust ("spam"). Given a constant "spam-cost" S, the overall cost is $-(1 - c(t_{n+1} - t^{\text{start}})) * S - C$.

Case 3: $t^{\text{end}} < t_{n+1}$. If the true purchase occurs after a coupon's end time, the customer is again exposed to unusable "spam" offers. We incur no benefit, but a cost of $-(1 - c(t_{n+1} - t^{\text{start}})) * S - C$, which is proportional to the magnitude of prediction error.

5 MODEL EVALUATION

We split the data into training (February - July, 2014), validation (August - October, 2014) and test (November, 2014 - January, 2015) subsets. The train window contains 1,415,895 purchases, the validation window contains 703,505 purchases and the test window contains 688,960 purchases.

Data Likelihood. We first evaluate the accuracy of the posterior distribution $h^*(t_{n+1})$. Fig. 9 shows the predictive loglikelihood for each model, computed using Eq. (6) on the test data. Note that we do not show RUSH! here since its loglikelihood includes category information and is not comparable to univariate models. With the Poisson model as a baseline, incorporating time-variation (P to TVP) improves the predictive loglikelihood by 9.53 units, while incorporating memory (P to HP) improves it by 16.28 units. Combining both factors (P to TVHP) improves it by 30.52 units: 4.71 units greater than sum of the improvement by each factor separately.

This reveals an important marketing insight. Consumer purchases tend to be "clustered" in time, with rapid bursts of purchases separated by longer periods of inactivity. Clustering could either result from higher purchase rates at certain times (time-variation)

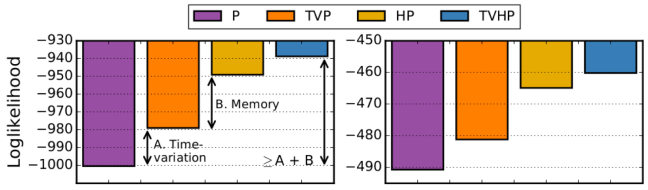


Figure 9: (left) training and (right) test data loglikelihoods of models (note: numbers not comparable due to different train/test sizes).

or from self-excitation. The aforementioned results indicate that self-excitation alone explains clustered purchases better than time-variation alone; while combining the two results in better explanation of the data than considering each of them separately.

Expected Value Analysis. The cost-benefit framework introduced in §4.5 facilitates evaluating our model systematically for the application, and helps us carefully account for the benefits gained for correct predictions on different categories as well as costs incurred by different types of errors. Our setup is as follows[4].

- Benefit B_d for category d is the average amount spent in purchases of category d (see Fig. 4(c)), multiplied by a commission rate of 1%.
- Spam cost $S = \$0.01$
- Fixed cost $C = \$0.001$
- Forget-rate $\gamma = \ln(2)/(2 * 3600)$; intuitively, 2 hours from the start of the coupon, its benefit drops from B to $B/2$.

We make our predictions on the test data and compute cost and benefit based on the success and error scenarios in §4.5. For baseline models that do not predict categories, we choose the top k most frequently occurring categories in the training data. We show the total benefit vs. total cost for all the models in Fig. 10 for coupon durations $\tau = \{1 \text{ hour}, 2 \text{ hours}\}$ and number of predicted categories in a coupon $k = \{1, 2, 3\}$. We see that RUSH! provides the highest benefit, significantly higher than P and TVP, at a cost comparable to the HP and TVHP models. The memory-less models often over-predict, as they do not account for excitation from the recent past, and end up with low cost but no benefit (case 1 in §4.5). RUSH! yields the highest overall gain.

Similar observations can be made by looking at the absolute errors of the models. Fig. 11 shows the fraction predictions by each model that obtained different values of the absolute error v in hours, and also captured the true purchase category within the top $k = \{1, 2, 3\}$ predicted categories. We see that memory-less models P and TVP make larger time prediction errors that cannot be compensated with more offers per coupon. On the other hand, memory-based models HP and TVHP catch up to RUSH! as the number of offers is increased suggesting that their mispredictions are mainly due to their inability to predict the correct category.

We also consider the impact of only delivering coupons for which we are confident about redemption. Intuitively, delivering only high-confidence coupons conservatively avoids the costs of spamming consumers, but trades off against potentially larger gain from serving more coupons. Fig. 12 shows this trade-off for 1-hour long coupon and $k = 1, 2$ when serving the top $p\%$ most confident

[4]We engaged with domain experts at the bank to determine appropriate ranges of values for the commission rates, forgetting-rate γ, spam cost S and fixed cost C. The presented results are are for a single configuration of values, but are robust to values within the determined ranges.

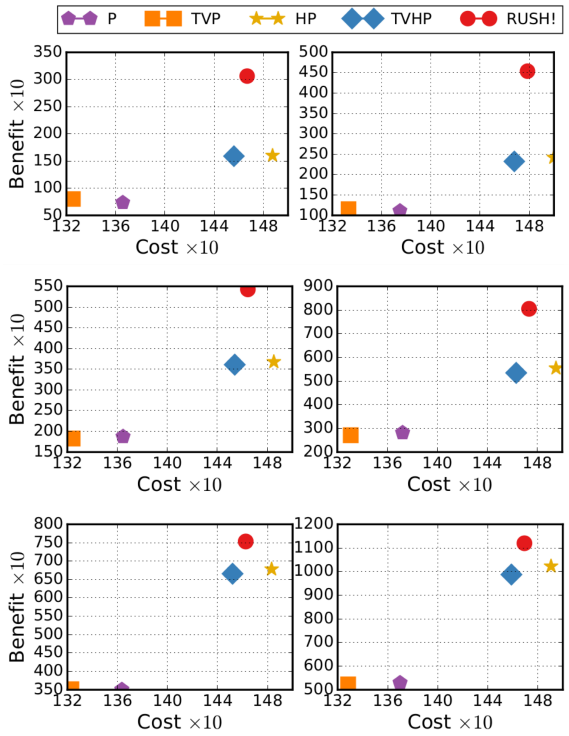


Figure 10: Total benefit vs. cost on test data by different models for coupon duration (left) $\tau = 1hr$ and (right) $\tau = 2hrs$ and for number of offers per coupon (top to bottom) $k = \{1, 2, 3\}$.

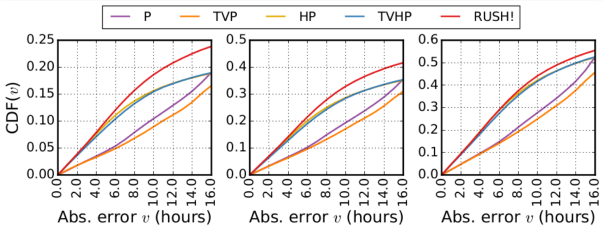


Figure 11: Fraction of correct predictions within absolute error of x hrs (i.e., CDF curve) captured with (from left to right) $k = \{1, 2, 3\}$ offers per coupon.

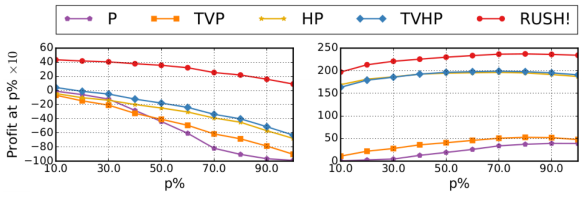
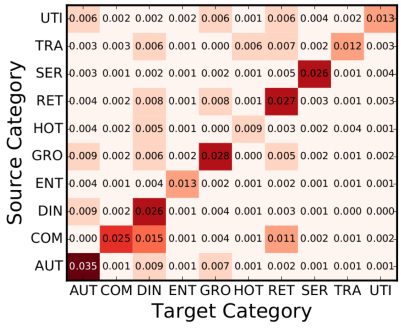


Figure 12: Total profit with the $p\%$ most confident predictions, coupon duration = 1 hour, (left) $k = 1$ category and (right) $k = 3$ categories. $k = 2$ is shown in Fig. 1.

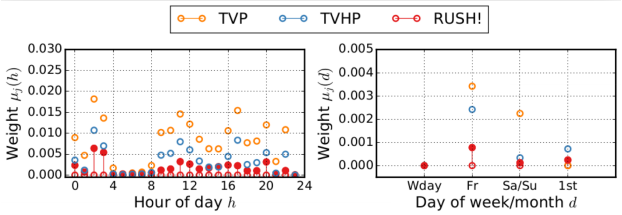
coupons. The confidence of each prediction $\hat{t}$ is the value of its posterior density $\hat{h}^*(t)$. Observe that the total profit peaks at 10% for $k = 1$ and 80% for $k = 2$. For all values of $p\%$ RUSH! dominates competitors in total profit.

Model	λ_0^{-1}	$\lambda_0^{-1}\ln(2)$	$\min(\frac{\beta+\lambda_0}{\lambda_0}, \beta)$	$\alpha^{-1}\ln(2)$
P	23.64	16.39	—	—
TVP	∞	∞	—	—
HP	98.51	68.28	3.18	24.00
TVHP	∞	∞	0.005	24.00
RUSH!	∞	∞	see (b)	24.00

(a) Model parameters interpreted as the mean inter-event time λ_0^{-1}, the median inter-event time $\lambda_0^{-1}\ln(2)$ ($\lambda_0 = \sum_m \lambda_{m0}$ for RUSH!), the relative intensity increase from self-excitation $(\beta+\lambda_0)/\lambda_0$ (reported as β if $\lambda_0 = 0$) and the self-excitation half-life $\alpha^{-1}\ln(2)$. All times are in hours.



(b) RUSH! $\boldsymbol{\beta} \times 10^3$ excitation matrix



(c) Temporal feature weights. "Wday" implies Mon/Tue/Wed/Thu and "1st" implies the first of the month. For RUSH!, we show two markers for each feature; the max. and min. weight across dimensions.

Figure 13: Exploratory analysis of our model parameters.

Scalability. Parameter inference of the proposed model is linear in number of events, as shown empirically in Fig. 2. The runtime is around 60 minutes to fit RUSH! to ≈2.8 million transactions. All experiments were performed on an Intel Xeon E7-8860 v3 at 2.2Ghz with 4 physical CPUS and 4 cores per CPU.

6 MODEL INTERPRETATION

An advantage of our modeling approach is the interpretability of the parameters, which provides useful insights into consumer purchase behavior. We now focus on each parameter of our fitted models. Fig 13(a) lists the base rates interpreted as the mean inter-event times λ_0^{-1} and median inter-event times $\ln(2)\lambda_0^{-1}$ for each model. For RUSH! we list the total base rate across dimensions. We observe that incorporating memory (P to HP) reduces the base-rate: this is intuitive, since part of what triggers purchases is now attributed to self-excitation. We also observe that on incorporating time-variance (P to TVP, RUSH!), the base-rate drops close to zero leading to mean and median inter-event times of ∞. This indicates that base purchase rates are entirely driven by the wall-clock time, with no common rate across times. The excitation decay half-life remains the same across models, indicating its independence from the other parameters.

We also consider the relative boost in purchase rates due to self/mutual-excitation, $\min(\frac{\beta+\lambda_0}{\beta}, \beta)$. When $\lambda_0 \approx 0$, the boost is simply β. We observe that incorporating time-variance into memory-driven models reduces the excitation boost, since part of what triggers purchases is now explained by time-variation. This observation also carries over to RUSH!, with its excitation matrix $\boldsymbol{\beta}$ shown in Fig 13(b).

The patterns of excitation revealed by $\boldsymbol{\beta}$ are of particular interest to marketers. The number in each cell denotes the magnitude of excitation of the source category on the target category. We observe the following based on the excitation matrix of RUSH! in Fig. 13(b), where category names have been abbreviated to 3 characters (see Fig. 4(c) for full names):

(1) RETAIL (shopping) is a strong trigger for DINING. In fact all purchases trigger dining purchases to varying degrees. Thus, marketers looking to predict dining purchases can rely on the occurrence of prior purchases with varying levels of confidence, depending on the prior purchase category.
(2) RETAIL purchases are highly self-excited, but also excited by COMMUTE (public transport), TRAVEL (airlines, train) and UTILITIES (gas, electricity) purchases. In turn, RETAIL purchases excite GROCERY purchases (but not vice-versa). A common ordered pattern is indicated, of paying bills and/or commuting, shopping at department stores and then buying groceries.
(3) HOTEL (long-term stay) purchases are triggered by TRAVEL (airlines, train), which is intuitive. However, TRAVEL is only weakly triggered by all other purchase categories, indicating its inherent unpredictability.

Finally, we consider the effect of time on purchase rates in Fig 13(c). We observe that for all models, 4-8AM is the least likely purchase period, and 11-12PM, 5-6PM and 2-3AM are the most likely purchase periods. This agrees with our empirical findings in Fig. 6. The effect of the day of the week is similar. While the learned parameters of all models follow similar trends, their magnitudes reduce when incorporating self/cross-excitation, since the total effect is now shared across time-variation and memory.

7 RELATED WORK

Time-limited Promotions. There is much work in the marketing literature on understanding the factors that affect coupon redemption [3], with recent focus on mobile coupons in particular [9] employing randomized experiments to study the influence of coupon face value, timing and duration on redemption. A recent economic model directly relates coupon redemption to various factors based on behavioral theory [15]. While the model is calibrated on real-world data and provides managerial insights into the effect of various factors on redemption rates, it is used primarily as an exploratory technique and is not intended or evaluated to function in a predictive manner.

Event Forecasting. Recommender systems are a typical application domain of forecasting future events. Most such systems

focus on rating prediction that can help identify likely movies a user will watch [10], or songs they will hear [7], etc. Several work also leverage temporal dynamics for prediction and modeling, including the award-winning work by Koren [18] that showed how incorporating time effects helps improve predictive accuracy in recommender systems, and Benson et al. [4] that showed the impact of time in consumer reconsumption behavior. This group of work, however, does not focus on predicting the *occurrence time* of future events, with the exception of recent work by Du et al. [12] that leveraged point process models for making time-sensitive recommendations.

A particular family of point process models, called Hawkes processes, has also been used to forecast events such as earthquakes [21], financial contagion [2] and information diffusion [13]. Their defining characteristic is 'self-excitation', where each event triggers (i.e., increases the rate of) future events. Self-excitation has also been used for modeling terrorist activity [23] and tweet popularity [27]. Simple point process models are effective in capturing occurrence time of future events, however for events of the same *type* or category. Most recently, Minor et al. also addressed the problem of activity prediction [22]. Despite observing different types of activities, their goal is to predict the time until the next occurrence of a *given* activity from sensor data.

To capture multiple types of events occurring over time, where an event of a certain type can trigger events of other types, multi-dimensional point process models have been used. Applications include modeling topic diffusion [26], social network user interactions [28], stock market transactions [25], TV program views [20], and most recently spatial trajectory prediction [11] where the event types respectively are topics, users, transaction types, TV program types, and locations.

8 CONCLUSION

We proposed RUSH! to deliver time-limited coupons via purchase forecasts, based on a continuous-time point-process model estimated from millions of real-world customer transactions. We unify both temporal dynamics and mutual/self-excitation behavior into a single, interpretable model that scales linearly to datasets with millions of transactions. The interpretability of our model parameters provide a number of valuable insights into consumer purchase behavior that are valuable for marketers. In addition, we present a cost-benefit framework motivated by business insights that more closely mirrors the expectations from the model when deployed in the real-world. We demonstrate that our method outperforms competing baselines within this framework.

Given the promise demonstrated by our cost-benefit analysis, our next step is to pursue A/B testing and randomized experiments with predictions from our model driving the delivery of smartphone coupons. Such experiments may reveal numerous further factors affecting coupon redemption, such as the aesthetic design or text content of a coupon, or its connection with a timely event (such as a football game). Extending our model to incorporate new phenomena revealed by such experiments is an important direction of future work.

ACKNOWLEDGMENTS

This research is sponsored by the DARPA Transparent Computing Program under Contract No. FA8650-15-C-7561, ARO Young Investigator Program under Contract No. W911NF-14-1-0029, NSF CAREER 1452425 and IIS 1408287. Any conclusions expressed in this material are of the authors and do not necessarily reflect the views, expressed or implied, of the funding parties.

REFERENCES

[1] Praveen Aggarwal and Rajiv Vaidyanathan. 2003. Use it or lose it: Purchase acceleration effects of time-limited promotions. *J. of Consumer Behaviour* 2, 4 (2003), 393–403.
[2] Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger JA Laeven. 2015. Modeling financial contagion using mutually exciting jump processes. *J. of Financial Economics* (2015).
[3] Michelle Andrews, Jody Goehring, Sam Hui, Joseph Pancras, and Lance Thornswood. 2016. Mobile promotions: A framework and research priorities. *Journal of Interactive Marketing* 34 (2016), 15–24.
[4] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. 2016. Modeling User Consumption Sequences.. In *WWW*. 519–529.
[5] BIA/Kelsey. 2011. BIA/Kelsey Revises Deals Forecast Upward Slightly, Due to More Entrants, Rapid Market Expansion and Growing Consumer Adoption.
[6] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16, 5 (1995), 1190–1208.
[7] O. Celma, M. Ramirez, and P. Herrera. 2005. Foafing the music: A music recommendation system based on RSS feeds and user preferences. In *Inter. Conf. on Music Info. Retrieval*.
[8] Daryl J Daley and David Vere-Jones. 2003. *An introduction to the theory of point processes*. Springer Science & Business Media.
[9] Peter J Danaher, Michael S Smith, Kulan Ranasinghe, and Tracey S Danaher. 2015. Where, when, and how long: Factors that influence the redemption of mobile phone coupons. *Journal of Marketing Research* 52, 5 (2015), 710–725.
[10] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *KDD*. ACM, 193–202.
[11] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent Marked Temporal Point Processes: Embedding Event History to Vector.. In *KDD*. 1555–1564.
[12] Nan Du, Yichen Wang, Niao He, Jimeng Sun, and Le Song. 2015. Time-Sensitive Recommendation From Recurrent User Activities.. In *NIPS*. 3492–3500.
[13] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. 2015. Coevolve: A joint point process model for information diffusion and network co-evolution. In *NIPS*. 1954–1962.
[14] Federal Deposit Insurance Corportion. 2016. 2015 FDIC National Survey of Unbanked and Underbanked Households.
[15] Richard C Hanna, Scott D Swain, and Paul D Berger. 2016. Optimizing time-limited price promotions. *J. of Marketing Analytics* 4, 2-3 (2016), 77–92.
[16] J Jeffrey Inman and Leigh McAlister. 1994. Do coupon expiration dates affect consumer behavior? *J. of Marketing Research* (1994), 423–428.
[17] Inmar. 2016. Promotion Industry Trends: A Year in Review.
[18] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics.. In *KDD*. ACM, 447–456.
[19] Thomas Josef Liniger. 2009. *Multivariate hawkes processes*. Ph.D. Dissertation. ETH Zurich.
[20] Dixin Luo, Hongteng Xu, Yi Zhen, Xia Ning, Hongyuan Zha, Xiaokang Yang, and Wenjun Zhang. 2015. Multi-Task Multi-Dimensional Hawkes Processes for Modeling Event Sequences.. In *IJCAI*. 3685–3691.
[21] David Marsan and Olivier Lengline. 2008. Extending earthquakes' reach through cascading. *Science* 319, 5866 (2008), 1076–1079.
[22] Bryan Minor, Janardhan Rao Doppa, and Diane J Cook. 2015. Data-driven activity prediction: Algorithms, evaluation methodology, and applications. In *KDD*. 805–814.
[23] Michael D. Porter and Gentry White. 2012. Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* 6, 1 (03 2012), 106–124.
[24] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".
[25] J.A. McGill V. Chavez-Demoulin. 2012. High-frequency financial data modeling using Hawkes processes. *J. of Banking & Finance* 36, 12 (2012), 3415–3426.
[26] Shuang-Hong Yang and Hongyuan Zha. 2013. Mixture of Mutually Exciting Processes for Viral Diffusion.. In *ICML*. 1–9.
[27] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *KDD*. 1513–1522.
[28] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes.. In *AISTATS*. 641–649.
[29] Su Zhou, Alan Montgomery, and Geoff Gordon. 2016. Exploring Customer Spending Behavior and Payday Effect using Prepaid Cards Transaction Data. *Preprint* (2016).