Discovering Progression Stages in Trillion-Scale Behavior Logs

Kijung Shin¹, Mahdi Shafiei², Myunghwan Kim², Aastha Jain², Hema Raghavan²
¹Carnegie Mellon University, Pittsburgh, PA, USA
²LinkedIn Corporation, Mountain View, CA, USA
kijungs@cs.cmu.edu, {mashafiei, mukim, asjain, hraghavan}@linkedin.com

ABSTRACT

User engagement is a key factor for the success of web services. Studying the following questions will help establishing business strategies leading to their success: How do the behaviors of users in a web service evolve over time? To reach a certain engagement level, what are the common stages that many users go through? How can we represent the stage that each individual user lies in?

To answer these questions, we propose a behavior model that discovers the progressions of users’ behaviors from a given starting point – such as a new subscription or first experience of certain features – to a particular target stage such as a predefined engagement level of interest. Under our model, transitions over stages represent progression of users where each stage in our model is characterized by probability distributions over types of actions, frequencies of actions, and next stages to move. Each user performs actions and moves to a next stage following the probability distributions characterizing the current stage.

We also develop a fast and memory-efficient algorithm that fits our model to trillions of behavioral logs. Our algorithm scales linearly with the size of data. Especially, its distributed version implemented in the MapReduce framework successfully handles petabyte-scale data with one trillion actions.

Lastly, we show the effectiveness of our model and algorithm by applying them to real-world data from LinkedIn. We discover meaningful stages that LinkedIn users go through leading to predefined target goals. In addition, our trained models are shown to be useful for downstream tasks such as prediction of future actions.

ACM Reference Format:

1 INTRODUCTION

The behaviors of users of web-sites and apps change over time for reasons including temporal trends [14] and shift of personal interests [17]. When these behavioral changes are aligned with a certain direction, we observe the progression of user behavior. For example, in RateBeer, a beer review site, new users have similar tastes, but they start reviewing different types of beers as users gain experience and develop their own preferences [21, 29]. Another example is Wikipedia, where users utilize different navigation strategies throughout the information seeking process in order to reach the target information [28].

Understanding such progressions is of paramount importance to providing more personalized experiences, which can lead to greater engagement and potentially increasing revenue. To revisit the RateBeer example, the site may want to advertise beers to users based on their current tastes or recommend new flavors of beers considering the next stages in the progression of the users’ tastes.

While general understanding of those progression patterns is important, businesses are often interested in behavior progressions only from one state to another state due to strategic or organizational interests. For instance, a company may want to strategically focus on helping onboarding users to reach a certain level of engagement, or it may need to study users who eventually bring revenue. In another example, at LinkedIn, a professional online social network providing services including news feeds and job pages, a product manager for news feeds needs to study the progressions of users who eventually become engaged with the news feed. Another product manager for job pages may be interested in navigational patterns of users who are looking for jobs. Hence, insights about progressions of users with respect to a specific target state is very helpful for establishing practical strategies.

To summarize and better understand such progressions toward a target state, we need to identify common stages that many users go through while interacting with a web service. For thorough summarization, these stages should capture changes in three different aspects that we describe below.

The first aspect is the change in the types of actions performed or features frequently used by users. For example, in an online social network service like LinkedIn, new users focus on making connections, while more established users with enough connections spend more time on consuming content or interacting with their
connections. Changes in the types of actions performed by a user will then be considered as transitions between these stages.

Another aspect to consider is how often users visit, perform an action, or use a feature in a web service. Users who visit a web service every day and those who visit it once a year may not be in the same stage even though they perform similar actions or use similar features. This distinction is particularly important because service providers typically distinguish users by their level of activity and often aim to promote user activity and engagement.

The last aspect is the direction of changes toward a given target state. New users who are getting familiar with the web service and more established users who are becoming less active may be in different stages independent of how similar the other aspects of their behavior are. That is, stages should describe not only current behavior but also transitions to future stages.

In this work, we propose a behavior model with stages characterizing all the aforementioned aspects of changes, while most existing models target only one of these aspects (e.g., only types of actions [21, 29]). Specifically, each stage in our model is characterized by a probability distribution over types of actions, a probability distribution over frequencies of actions, and transition probabilities of moving to another stage as the next stage. Progressions of users’ behaviors can then be defined as transitions between these stages. That is, each user performs actions and moves to the next stage following the probability distributions describing the current stage.

Our algorithmic contribution is a fast and memory-efficient algorithm for training the parameters (i.e., probability distributions) of our model. It aims to find the parameters that best describe a behavior model. In this section, we describe our behavior model for capturing progressive changes in users’ behaviors. The symbols used to describe our model are listed in Table 1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}$</td>
<td>set of all users</td>
</tr>
<tr>
<td>$\mathcal{A}$</td>
<td>set of all types of actions</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>set of all binned time gaps</td>
</tr>
<tr>
<td>$\mathcal{S} = {s_1, ..., s_k}$</td>
<td>set of all stages</td>
</tr>
<tr>
<td>$n_u$</td>
<td>number of actions done by user $u \in \mathcal{U}$</td>
</tr>
<tr>
<td>$g_u \in {0, 1}$</td>
<td>whether user $u \in \mathcal{U}$ reaches the target stage after doing $n_u$ actions</td>
</tr>
<tr>
<td>$a_{u,j} \in \mathcal{A}$</td>
<td>type of the $j$-th action done by user $u \in \mathcal{U}$</td>
</tr>
<tr>
<td>$t_{u,j}$</td>
<td>timestamp of action $a_{u,j}$</td>
</tr>
<tr>
<td>$\delta_{u,j} \in \Delta$</td>
<td>binned time gap between $t_{u,j}$ and $t_{u,j-1}$</td>
</tr>
<tr>
<td>$a_u$</td>
<td>sequence of actions performed by user $u \in \mathcal{U}$</td>
</tr>
<tr>
<td>$\delta_u$</td>
<td>sequence of time gaps in $a_u$</td>
</tr>
<tr>
<td>$\mathcal{S}_{-k}$</td>
<td>set of all stages excluding the target stage $s_k$</td>
</tr>
<tr>
<td>$z_{u,j} \in \mathcal{S}_{-k}$</td>
<td>stage of user $u$ at $t_{u,j}$</td>
</tr>
<tr>
<td>$\bar{z}_u$</td>
<td>sequence of stages assigned to the actions in $a_u$</td>
</tr>
<tr>
<td>$\theta_{s_i}, \phi_{s_i}, \psi_{s_i}$</td>
<td>probability distributions of actions, time-gaps, and transitions in stage $s_i \in \mathcal{S}_{-k}$</td>
</tr>
<tr>
<td>$\lambda_\theta, \lambda_\phi, \lambda_\psi$</td>
<td>hyperparameters regarding the prior distributions of $\theta_{s_i}, \phi_{s_i}, \psi_{s_i}$</td>
</tr>
<tr>
<td>$\xi$</td>
<td>hyperparameter for the initialization step</td>
</tr>
</tbody>
</table>

## 2 BEHAVIOR MODEL

In this section, we describe our behavior model for capturing progressive changes in users’ behaviors. The symbols used to describe our model are listed in Table 1.

### 2.1 Notations and Model Description

Consider a set of users doing a sequence of actions. Let $\mathcal{U}$ be the set of users and $\mathcal{A}$ be the set of types of actions that can be done by the users. We bin the time gap between each two consecutive actions and use $\Delta$ to indicate the set of potential gap sizes.

Let $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ be the set of $K$ stages that the users may go through between the starting stage $s_1$ and the target stage $s_k$. Two stages $s_1$ and $s_k$ are treated as special cases. That is, we have predefined conditions that determine whether each user has reached them (e.g., users reach the starting stage $s_1$ if they join LinkedIn and reach the target stage $s_k$ as soon as they reach a certain number of connections).

We assume the monotonicity of the stages in $\mathcal{S}$ to model progression ‘towards’ the goal stage $s_k$. That is, for any $i$ and $j$ satisfying $1 \leq i < j \leq k$, users can transit from stage $s_i$ to stage $s_j$ but not in the opposite direction. This constraint, however, does not enforce that all users follow the same path towards the goal stage. Under our model, users are allowed to skip any intermediate stages.

Let $\mathcal{S}_{-k} = \{s_1, ..., s_{k-1}\}$ be the set of non-target stages. Each non-target stage $s_i \in \mathcal{S}_{-k}$ is characterized by probability distributions over types of actions, time gaps, and transitions from the stage, which are defined as follows:

- $\theta_{s_i} \in \mathbb{R}^{|\mathcal{A}|}$: probability distribution over types of actions performed by users in stage $s_i$.
- $\phi_{s_i} \in \mathbb{R}^{|\mathcal{A}|}$: probability distribution over time gaps between two consecutive actions performed by users in stage $s_i$.
- $\psi_{s_i} \in \mathbb{R}^{k-i+1}$: transition probability distribution over next stages moving from stage $s_i$ before performing each action.
We assume a symmetric Dirichlet prior \([4]\) over \(z\) (i.e., \(\phi\) distributions characterizing each stage (i.e., each \(s\)) are colored white. The type of each action (i.e., each \(a\)) and the time-gap for her first action is ignored since there can be no previous action. Each user repeats this process until she reaches the target stage.

As seen in Figure 2, the following are observable from log data: (a) whether each user had reached the target stage before the log data were collected (i.e., \(\{a_u\}_{u \in \mathcal{U}}\)), (b) the number of actions performed by each user between the starting stage (inclusive) and the target stage (exclusive) before the log data were collected (i.e., \(\{n_u\}_{u \in \mathcal{U}}\)), (c) the type of each action (i.e., \(\{\theta_u\}_{u \in \mathcal{U}}\)), and (d) the time gap between each two consecutive actions (i.e., \(\{\delta_u\}_{u \in \mathcal{U}}\)).

3 MODEL TRAINING

In this section, we discuss training the parameters of our model. We present the objective function in Section 3.1 and our optimization algorithm in Section 3.2. We extend the algorithm to external-memory, multi-core, and distributed settings in Section 3.3 and discuss its time and space complexities in Section 3.4.

3.1 Training Objective

Given an input dataset consisting of sequences of actions \(\{a_u\}_{u \in \mathcal{U}}\), sequences of time-gaps \(\{\delta_u\}_{u \in \mathcal{U}}\) and whether users reach the target stage \(\{\nu_u\}_{u \in \mathcal{U}}\), our objective is to find the most probable sequences of stages \(\{\hat{s}_u\}_{u \in \mathcal{U}}\) as well as probability distributions of actions \(\{\theta_s\}_{s_i \in \mathcal{S}_k}\), time-gaps \(\{\phi_s\}_{s_i \in \mathcal{S}_k}\), and transitions \(\{\psi_s\}_{s_i \in \mathcal{S}_k}\). Maximizing the posterior probability of the parameters of our model given the observed states, written as

\[
p(\{\theta_s\},s_i \in \mathcal{S}_k,\{\phi_s\},s_i \in \mathcal{S}_k,\{\psi_s\},s_i \in \mathcal{S}_k,\{\hat{s}_u\}_{u \in \mathcal{U}} | \{a_u\}_{u \in \mathcal{U}},\{\delta_u\}_{u \in \mathcal{U}},\{\nu_u\}_{u \in \mathcal{U}},\lambda_{\theta},\lambda_{\phi},\lambda_{\psi}) \tag{1}
\]

is equivalent to maximizing the following objective function \(f\):

\[
f(\{\hat{s}_u\}_{u \in \mathcal{U}}; \{\theta_s\},s_i \in \mathcal{S}_k; \{\phi_s\},s_i \in \mathcal{S}_k; \{\psi_s\},s_i \in \mathcal{S}_k) := \prod_{u \in \mathcal{U}} f_u(z_u; \{\theta_s\},s_i \in \mathcal{S}_k; \{\phi_s\},s_i \in \mathcal{S}_k; \{\psi_s\},s_i \in \mathcal{S}_k) \times \prod_{s_i \in \mathcal{S}_k} p(\{\theta_s\},s_i \in \mathcal{S}_k; \{\phi_s\},s_i \in \mathcal{S}_k; \{\psi_s\},s_i \in \mathcal{S}_k) \tag{2}
\]

where \(f_u(z_u; \{\theta_s\},s_i \in \mathcal{S}_k; \{\phi_s\},s_i \in \mathcal{S}_k; \{\psi_s\},s_i \in \mathcal{S}_k) := p(a_u | (\theta_s),s_i \in \mathcal{S}_k; \nu_u) \times p(\delta_u | (\phi_s),s_i \in \mathcal{S}_k; \nu_u) \times p(z_u | (\psi_s),s_i \in \mathcal{S}_k; \nu_u)\)

\[
= \left(\prod_{j=1}^{n_u} p(a_{u,j} | \hat{s}_{u,j}, \theta_{\hat{s}_{u,j}})\right) \times \left(\prod_{j=2}^{n_u} p(\delta_{u,j} | \hat{s}_{u,j-1}, \nu_{\hat{s}_{u,j-1}})\right) \times p(\nu_u | s_{u,n_u}, \nu_{s_{u,n_u}}). \tag{3}
\]

Notice that our objective function is non-convex and may have multiple local optima.

3.2 Optimization Algorithm

We propose an iterative refinement algorithm for optimizing our objective function \(f\) (i.e., Eq. (2)). Different from general optimization algorithms for any graphical model (e.g., EM [11]), our algorithm fully utilizes the dependency structure of our model for fast, memory-efficient, and parallel computation.

Our algorithm consists of the following three steps:

- **Initialization step** (Section 3.2.3): We initialize the probability distributions \(\theta_s, \phi_s, \text{ and } \psi_s\) in every stage \(s_i \in \mathcal{S}_k\).
- **Assignment step** (Section 3.2.1): Given the current probability distributions \(\theta_s, \phi_s, \text{ and } \psi_s\) in every stage \(s_i \in \mathcal{S}_k\), we update the stage assignments \(s_{u,i}\) of every user \(u \in \mathcal{U}\) so that our objective function \(f\) is maximized.

Figure 2: Plate notation [8] for our behavior model. Observed variables are colored grey, and unobserved variables are colored white. The type of each action (i.e., each \(a_{u,j}\)) and the time-gap between each two consecutive actions (i.e., each \(\delta_{u,j}\)) depend on the current stage (i.e., each \(z_{u,j}\)) and the probability distributions characterizing each stage (i.e., each \(\theta_{s_i}, \phi_{s_i}, \text{ and } \psi_{s_i}\)). Transitions between stages depend on the transition probability distribution in each stage (i.e., each \(\psi_{s_i}\)).
We observe that Algorithm 1 is guaranteed to find a local optima. Specifically, for each stage $s_l \in \mathcal{S}_k$, the assignment step and update step are guaranteed to converge. Each assignment step and each update step are guaranteed to improve the objective function. Therefore, our algorithm is guaranteed to find a local optimum.

We describe each step in detail in the following subsections. For ease of explanation, we first present the assignment and update steps then present the initialization step.

### 3.2.1 Assignment Step

In this step, we maximize the objective function $f$ by updating the stage assignments $\{z_u\}_{u \in \mathcal{U}}$. While fixing the probability distributions $\{\theta_{s_l}\}_{s_l \in \mathcal{S}_k}$, $\{\phi_{s_l}\}_{s_l \in \mathcal{S}_k}$, and $\{\psi_{s_l}\}_{s_l \in \mathcal{S}_k}$ to their current values. Once the probability distributions are fixed, the stage assignment $z_u$ of each user $u \in \mathcal{U}$ can be optimized independently by maximizing $f_u$ (i.e., Eq. (5)).

For each user $u \in \mathcal{U}$, we use dynamic programming [26] to update $z_u$, as described in detail in Algorithm 1. In the algorithm, $f_u(s_l)$ denotes the maximum posterior probability of the $j$-th or later actions and time gaps for user $u$ given that $z_u = s_l$. That is, $f_u(s_l)$ is defined as follows:

$$f_u(s_l) := \max_{(z_u,l)\geq j} \{ p(\{a_u^{l,j}\}_{j \in \mathcal{L}}, z_u) \}_{s_l \in \mathcal{S}_k} \times p(a_{u,j}=a \wedge z_u=s_l) \times p(z_u=s_l\mid a_{u,j}=a) \times p\left(\{z_u, l\}_{j \in \mathcal{L}}\mid \{\psi_{s_l}\}_{s_l \in \mathcal{S}_k}, z_u=s_l\right) \times p\left(\{a_{u,j}\}_{j \in \mathcal{L}}\mid \{\theta_{s_l}\}_{s_l \in \mathcal{S}_k}, z_u=s_l\right) \times p\left(\{\psi_{s_l}\}_{s_l \in \mathcal{S}_k}\right) \times p\left(\{\theta_{s_l}\}_{s_l \in \mathcal{S}_k}\right).$$

We observe that $f_u(s_l)$ is computed easily from $\{f_{u,j}(s_l)\}_{s_l \in \mathcal{S}_k}$ (lines 9-13). Our algorithm exploits this observation by computing $\{f_{u,j}(s_l)\}_{s_l \in \mathcal{S}_k}$ in the decreasing order of $j$ from $n_u$ to 1 (lines 2-14). Note that the stage assignments maximizing each $f_u(s_l)$ is stored in $h_u(s_l)$ (line 14). Specifically, $h_u(s_l) = s_l$ means that the stage assignments maximizing $f_u(s_l)$ are $z_u,j = s_l$ and those maximizing $f_u,j+1(s_l)$.

Once we have computed $\{f_{u,j}(s_l)\}_{s_l \in \mathcal{S}_k}$, which indicates the maximum posterior probabilities of everything except the initial transition, we can easily maximize $f_u$ by finding $z_u$ maximizing $\psi_{u}(z_{u,1}) \cdot f_u(z_{u,1})$ (line 15). The stage assignments $\{z_u\}_{u \in \mathcal{U}}$ maximizing $f_u$ can be obtained from $\{h_u(s_l)\}_{s_l \in \mathcal{S}_k}$ by following the path backward starting from $z_u,1$ (lines 16-17).

### 3.2.2 Update Step

In this step, we maximize $f$ by updating the probability distributions $\{\theta_{s_l}\}_{s_l \in \mathcal{S}_k}$, $\{\phi_{s_l}\}_{s_l \in \mathcal{S}_k}$, and $\{\psi_{s_l}\}_{s_l \in \mathcal{S}_k}$, while fixing the stage assignments $\{z_u\}_{u \in \mathcal{U}}$ to their current values.

To this end, we decompose our objective function based on the probability distribution that each term depends on as follows:

$$f = \prod_{u \in \mathcal{U}} \left( f_0(\theta_u) \times f_\phi(\phi_u) \times f_\psi(\psi_u) \right).$$

Specifically, for each stage $s_l \in \mathcal{S}_k$, we update the action type distribution $\theta_{s_l}$ so that $f_0(\theta_{s_l})$ (i.e., Eq. (5)) is maximized. For each type of action $a$, the probability $\theta_{s_l}(a)$ is updated as follows:

$$\theta_{s_l}(a) \leftarrow \frac{\lambda^{a} + c_\theta(s_l,a)}{\lambda + \sum_{a' \in \mathcal{A}} c_\theta(s_l,a')},$$

where $c_\theta(s_l,a) := \sum_{u \in \mathcal{U}} \{ 1 \leq j \leq n_u : a_{u,j} = a \wedge z_{u,j} = s_l \}$.

Likewise, for each stage $s_l \in \mathcal{S}_k$, we update the time gap distribution $\phi_{s_l}$ so that $f_\phi(\phi_{s_l})$ (i.e., Eq. (6)) is maximized. For each time gap $\delta \in \Delta$, the probability $\phi_{s_l}(\delta)$ is updated as follows:

$$\phi_{s_l}(\delta) \leftarrow \frac{\lambda^{\delta} + c_\phi(s_l,\delta)}{\lambda + \sum_{\delta' \in \Delta} c_\phi(s_l,\delta')}.$$
of the transition from \( s_l \) to \( s_l \) as follows:

\[
\psi_l(s_l) \leftarrow (1 - \psi_l(s_{l-1})) \times \frac{\lambda_l + c_G(s_l, s_l)}{(k-1)\lambda_l + \sum_{m=1}^{k-1} c_G(s_l, s_m)},
\]

where \( c_G(s_l, s_l) := \sum_{u \in U} |1 \leq j \leq n_u : z_{u, j-1} = s_l \land z_{u, j} = s_l| \).

### 3.3.3 Initialize Step
Since our objective function \( f \) (i.e., Eq. (2)) is non-convex, the solution found by our optimization algorithm and its speed of convergence (i.e., the number of iterations required for convergence) depend on initial parameter values. In this section, we present an initialization method that works well in practice, as we experimentally show in Section 4.5 and Section 4.6.

First, choose a subset of users and decide their stage variable in a simple way. Specifically, let \( U' \) be the set of users that has performed at least \( k \) actions (i.e., \( \xi \geq k-1 \)), where \( \xi \) is a hyperparameter. For each user \( u \in U' \), we divide her \( n_u \) stage variables into \( (k-1) \) continuous segments of equal length. Then, we assign stage \( s_i \) to the variables in each \( i \)-th segment. In this process, the users with a small number of actions (i.e., \( U - U' \)) are ignored since they are less likely to have gone through all \( (k-1) \) stages.

Once the stage variables of \( U' \) are set, the action-type probability distributions \( \{ \delta_u \}_{u \in U} \) and the time-gap probability distributions \( \{ \phi_u \}_{u \in S_k} \) are initialized by Eq. (8) and Eq. (9) with \( U' \) instead of \( U \). We initialize transition probability distributions \( \{ \psi_l(u) \}_{u \in S_k} \) so that the transition probability decreases exponentially with the distance from the current stage.2

### 3.3 Extensions to External-memory, Multi-core, and Distributed Settings
In this section, we extend our optimization algorithm, described in Section 3.2, to the external-memory, multi-core, and distributed settings without affecting its outputs. We focus on extending the assignment and update steps. The ideas are also applicable to the initialization step due to its similarity to the other steps.

#### 3.3.1 External-memory Settings
Our optimization algorithm does not require loading the entire input data (i.e., \( \{ \delta_u \}_{u \in U} \), \( \{ \phi_u \}_{u \in S_k} \), and \( \{ \psi_l(u) \}_{u \in S_k} \) in main memory at once. Instead, it can be done by simply shuffling the data by user ids. Each machine sums up the contributions of the assigned users to the counts (e.g., \( c_g(s_l, a) \) and \( c_G(s_l, s_l) \) in Eq. (8)-(Eq. (11)) by sequentially reading the assigned input data, as in Section 3.3.1. Then, the contributions are gathered and summed up in one machine, which then updates all probability distributions by Eq. (8)-(Eq. (11) and broadcasts them to all other machines so that they can be used in the next iteration.

#### 3.3.2 Multi-core Settings
Our optimization algorithm is easily parallelized in multi-core settings. In the assignment step, the stage assignment for one user (lines 2-17 of Algorithm 1) does not depend on that for the other users. Thus, stage assignments for different users can be run in parallel using multiple threads. Likewise, in the update step, the contributions of one user to the counts (e.g., \( c_g(s_l, a) \) and \( c_G(s_l, s_l) \) in Eq. (8)-(Eq. (11)) do not depend on those of the other users. Thus, computing the contributions of different users also can be run in parallel using multiple threads.

#### 3.3.3 Distributed Settings
We combine the extensions in the previous sections for distributed settings. We assume that the input data are distributed across machines so that (a) all data for the same user are stored in one machine and (b) the data in each machine are sorted by user id. In MapReduce [10], for example, this can be done by simply shuffling the data by user ids. Each machine sums up the contributions of the assigned users to the counts (e.g., \( c_g(s_l, a) \) and \( c_G(s_l, s_l) \) in Eq. (8)-(Eq. (11)) by sequentially reading the assigned input data, as in Section 3.3.1. Then, the contributions are gathered and summed up in one machine, which then updates all probability distributions by Eq. (8)-(Eq. (11) and broadcasts them to all other machines so that they can be used in the next iteration.

In Section 4.4, we experimentally show that the MapReduce implementation of our optimization algorithm successfully handles petabyte-scale data with one trillion actions.

### 3.4 Complexity Analyses
In this section, we analyze the time complexity and memory requirement of our optimization algorithm. We first show that the time complexity of our algorithm is linear in the total number of actions. Then, we show that its memory requirement is linear in the maximum number of actions among all users.

**Theorem 3.1 (Time Complexity).** Let \( N = \sum_{u \in U} n_u \) be the total number of actions and \( T \) be the number of iterations. If \( N = \Omega(|A|/k + |A|/k) \), then the time complexity of our optimization algorithm is \( O(TN^2) \).

**Sketch of Proof:** The time complexity of the assignment step (i.e., Algorithm 1) is \( O(NK^2) \) because line 9, which takes \( O(k) \), is executed \( O(\sum_{u \in U} n_u k) = O(NK) \) times.

The time complexity of the update step is \( O(N + |A|/k + |A|/k^2) \). For each of \( \{ (u, j) : u \in U, 0 \leq j \leq n_u \} \), whose size is \( N \), we need to increase a constant number of counts (e.g., \( c_g(s_l, a) \) and \( c_G(s_l, s_l) \) in Eq. (8)-(Eq. (11)) updating the probability distributions from the counts by Eq. (8)-(Eq. (11)) takes \( O(|A|/k + |A|/k^2) \).

Thus, one iteration (i.e., running the assignment and update steps once) takes \( O(NK^2 + |A|/k + |A|/k^2) \), which is \( O(NK^2) \) by our assumption. Hence, the total time complexity is \( O(TN^2) \).
We also verify that our optimization algorithm is effective by answering the following questions:

- **Q1. Effectiveness**: Do our model and algorithm discover meaningful progression stages in real-world data?
- **Q2. Applicability**: Are our trained models useful for downstream tasks such as prediction of future actions?
- **Q3. Scalability**: Does our algorithm scale linearly with the size of the input data? Can our algorithm handle a dataset with trillions of actions?
- **Q4. Convergence**: How rapidly does our algorithm converge?
- **Q5. Identifiability**: How accurately does our algorithm estimate ground-truth parameters?

### 4 EXPERIMENTS

#### 4.1 Experimental Settings

**Data**: We used a dataset provided by LinkedIn. The dataset is the log of the actions performed before mid June of 2017 by a subset of LinkedIn members who joined LinkedIn after mid April of 2017. The dataset has 22 types of actions including the followings:

- **visit**: visit LinkedIn.com.
- **profile-edit**: edit one’s profile.
- **profile-view**: view another member’s profile.
- **invite-abook**: invite a non-member to LinkedIn on a page that lists some emails imported from one’s address book.
- **conn-abook**: send a connection request to a member on a page that lists some members imported from one’s address book.
- **conn-ins**: send a connection request to a member on a page that lists some members in the same institution.
- **conn-rec**: send a connection request to a member on a page that lists some members recommended by LinkedIn.
- **conn-search**: send a connection request to a member on a page that shows the results of one’s search.
- **conn-profile**: send a connection request to a member on the member’s profile page.
- **conn-other**: send a connection request to a member by means other than those explained above.
- **conn-accept**: accept a connection request that one received.
- **job-view**: view a job posting.
- **message**: send a message to a member.

The time gaps between each consecutive actions performed by the same user are binned into second (within few seconds), minute, hour, day, week, month, and over-a-month (over a month).

### Starting and Target Stages

- **Connected**: The starting stage is defined as joining LinkedIn, and the target stage is defined as reaching 30 connections. The number of actions between the stages is about 500 millions.
- **Engaged**: The starting stage is defined as having 30 or more connections, and the target stage is defined as visiting LinkedIn 4 days in a week for 5 (not necessarily consecutive) weeks. The number of actions between the stages is about 150 millions.

#### 4.2 Q1. Effectiveness: Descriptive Results

We present results demonstrating intuitive patterns extracted by our model for both settings of ‘connected’ and ‘engaged’.

As shown in Table 2, our model extracted the following reasonable latent stages for the connected setting:

- **profile (stage 1)**: The first learned stage shows the behaviors of creating new profiles right after new members join the service.
- **on-boarding (stages 2-4)**: The next three stages describe the optional on-boarding process provided to new members. During the process, LinkedIn provides new members with a list of other members to connect to and non-members to invite. New members can rapidly (notice that the most probable time gap is second) send connection requests and invitations.
- **poke (stage 5)**: Members start poking other services and typically enjoying viewing other members’ profiles.
- **grow (stage 6)**: Members grow their networks by searching for other members and using connection recommendation services.
- **explore (stage 7)**: In addition to growing their networks, members start consuming content. They visit LinkedIn for browsing content (notice that visit is the most probable action). They also start seeing job postings. Bursts of actions are reduced.

Notice that not every user goes through every stage. For example, many members in the profile stage skip the optional on-boarding process and jump directly to the poke or explore stage.

For the engaged setting, the grow and explore stages (i.e., the last two stages extracted for the connected setting) were subdivided into multiple stages by our model as follows:

- **active-growth (stages 1-4)**: In the first four stages, members actively grow their networks by different means including connecting from profiles (ag-1), connecting from search (ag-2), importing address books (ag-3), and others (ag-4).
- **passive-growth and explore (stages 5-7)**: In the stage pg-1, members passively grow their networks by relying on connection recommendation services. The next stage, pg-2, captures the transition to the explore stage.

Once stages and progression of each user is learned, we can answer simple queries like “What is the stage that most members skip?” or perform post-processing to gain more detailed insights.

\[
\text{With } k = 8, \lambda_\theta = \lambda_\phi = 0.1, \text{ and } \xi = 30.
\]

\[
\text{With } k = 8, \lambda_\theta = \lambda_\phi = 0.1, \text{ and } \xi = 50.
\]
Table 2: Latent stages extracted by our model and algorithm from the LinkedIn dataset. The first table presents the stages after joining the service and before reaching 30 connections. Members (a) create their profiles, (b) go through on-boarding processes, (c) poke the service, (d) grow their networks, and finally (e) explore the service. The second table presents the stages after reaching 30 connections and before being engaged. Members (a) actively grow their networks by various means, (b) passively grow their networks relying on recommendation services, and finally (c) explore the service. For each stage, we list the three most probable types of actions, time gaps, and transitions, with their probabilities (****: 80%-100%, ****: 60%-80%, ***: 30%-60%, *: 10%-30%). See Section 4.1 for descriptions of the types of actions, and see Section 4.2 for explanation of the stages.

<table>
<thead>
<tr>
<th>Stages</th>
<th>Profile (s_1)</th>
<th>Ob-1 (s_2)</th>
<th>Ob-2 (s_3)</th>
<th>Ob-3 (s_4)</th>
<th>Poke (s_5)</th>
<th>Grow (s_6)</th>
<th>Explore (s_7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions (θ_s)</td>
<td>Profile-edit (**+)</td>
<td>Conn-ins (**+)</td>
<td>Invite-abook (***)</td>
<td>Conn-abook (****)</td>
<td>Conn-ins (*)</td>
<td>Invite-abook</td>
<td>Profile-view (**+)</td>
</tr>
<tr>
<td>Time-gaps (ϕ_s)</td>
<td>Hour (+**)</td>
<td>Minute (+**)</td>
<td>Second (+**)</td>
<td>Minute (+**)</td>
<td>Second (+***)</td>
<td>Second (+**)</td>
<td>Minute (+)</td>
</tr>
<tr>
<td>Transitions (ψ_s)</td>
<td>Profile (**+*)</td>
<td>Explore</td>
<td>Poke (****)</td>
<td>Ob-1 (****)</td>
<td>Explore</td>
<td>Ob-3 (****)</td>
<td>Explore</td>
</tr>
</tbody>
</table>

For example, we can analyze the most discriminative sequence of stages for a particular cohort to achieve a given target. Here, we use a cohort of the members who pass through the explore stage (s_7). To extract such paths, for a set of stages S' \subseteq S = \{s_2, \ldots, s_6\}, we define the discriminative score as:

\[
P(\text{reach target} | \text{pass through } s_7 \text{ and all of } S') \\frac{P(\text{reach target} | \text{pass through } s_7 \text{ but not all of } S')}{}
\]

We find the top 3 discriminative sequences of stages for a given target and present the results in Table 3. Interestingly, they are very different depending on targets. In the ‘connected’ setting, passing through either the grow or poke stage is important, while progressing step by step is more crucial for the ‘engaged’ setting. The connected setting results are intuitive because exploring other content does not necessarily help reaching the target stage without continuous engagement with networking components. On the other hand, the results from the engaged setting imply that exposure to various networking channels is helpful for longer engagement.

4.3 Q2. Applicability to Prediction Tasks

To show that our model is useful for downstream tasks, we use our model to predict (a) the type of each user’s next action, (b) the time gap between the current and next actions of each user, and (c) whether each user reaches the target stage within 100 actions. For each task, we compare the following approaches:

• Random: use a randomly chosen label.

Table 3: Top-3 discriminative paths for those who pass through the explore stage.

<table>
<thead>
<tr>
<th>Target setting</th>
<th>Top-3 paths (discriminative score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected</td>
<td>Grow (11.3)</td>
</tr>
<tr>
<td></td>
<td>Poke \rightarrow Grow (11.2)</td>
</tr>
<tr>
<td></td>
<td>Poke (8.7)</td>
</tr>
<tr>
<td>Engaged</td>
<td>Ag-3 \rightarrow Ag-4 \rightarrow Pg-1 \rightarrow Pg-2 (1.73)</td>
</tr>
<tr>
<td></td>
<td>Ag-2 \rightarrow Ag-3 \rightarrow Ag-4 \rightarrow Pg-1 \rightarrow Pg-2 (1.71)</td>
</tr>
<tr>
<td></td>
<td>Ag-3 \rightarrow Ag-4 \rightarrow Pg-1 \rightarrow Pg-2 (1.61)</td>
</tr>
</tbody>
</table>

- Frequent: use the label most frequent in the training set.
- LR: use logistic regression [22].
- Model: use the label most probable (according to our model) in the current stage of each user.
- Model + LR: divide users depending on their current stages (inferred by our model) and use logistic regression separately for each stage based on the actions in the stage.

For logistic regression and k-means++, we used (|A| + |Δ|) features corresponding to the frequencies of the action types and time gaps for each user. For action-type and time-gap predictions, we used relative frequencies rather than absolute ones, which led to
higher accuracy. We trained our model and logistic regression using randomly chosen half of the users in each dataset and tested using the others. For evaluation, we used the proportion of correct predictions (Accuracy) and a weighted sum \(^6\) of F1 scores (F1) [23].

As shown in Table 4, for all the tasks, prediction purely based on our model (Model), which is unsupervised, shows similar accuracy to logistic regression (LR), which is supervised. More importantly, their combination (Model + LR) was more accurate than combining k-means++ and logistic regression (K-Means + LR) as well as the individual methods (Model and LR) for all the tasks.

4.4 Q3. Scalability

We show the scalability of our optimization algorithm, described in Section 3. We consider the following implementations:

- In-Memory (Multi-Core): in-memory implementation where all data are loaded into memory. It can run in a parallel way in multi-core environments (see Section 3.3.2).
- Out-of-Core: out-of-core implementation using disk as the external memory (see Section 3.3.1).
- Distributed: MapReduce [10] implementation on Hadoop 2.6.1 (see Section 3.3.3).

First, we show that all implementations scale linearly with the size of the input data. To this end, we used synthetic datasets with different numbers of users while fixing \(|A| = |\Delta| = k = 10\) and \(n_u = 1000\) for every user \(u\). As seen in Figure 1 in Section 1, the per-iteration running times of all implementations increased linearly with the total number of actions. Especially, 'Distributed' processed a dataset with one trillion actions with per-iteration time less than 2 hours. We obtained similar results when we increased the number of actions per user, while fixing the number of users. These results are consistent with our theoretical analysis in Section 3.4.

Second, we show the near-linear speed-ups of our parallel and distributed implementations. To this end, we measured the speed-up \(^7\) of ‘Multi-Core’ with different numbers of threads and that of ‘Distributed’ with different number of machines. For ‘Multi-Core’, we used the LinkedIn datasets with \(k = 10\), and for ‘Distributed’, we used a larger synthetic dataset where \(|A| = |\Delta| = k = 10, |U| = 10\) millions, and \(n_u = 1000\) for every user \(u\). As seen in Figure 3, both implementations showed near-linear speed-ups.

Lastly, we show significant reductions in main memory requirements by out-of-core processing using external memory. To this end, we compared the amount of main memory space required by ‘In-Memory’ and ‘Out-of-Core’ for processing the LinkedIn datasets with \(k = 10\). As seen in Figure 4, ‘Out-of-Core’ required up to 1, 472\(x\) less main memory space than ‘In-Memory’. However, the increase in the running time was at most 50%.

4.5 Q4. Convergence

We show that our optimization algorithm converges within a small number of iterations. Figure 5 shows the value of our objective function (i.e., Eq. (2)) in each iteration of our optimization algorithm in the LinkedIn datasets. The number of iterations required for convergence increased with the number of stages (i.e., \(k\)). However, even with 20 stages, our algorithm converged within 20 iterations.

\(^{6}\)each weight is the proportion to the number of the corresponding label in the test set

\(^{7}\)relative speed compared to when a single thread (or a single machine) is used.
4.6 Q5. Identifiability

We show that our optimization algorithm learns parameters reasonably close to ground-truth parameters. We first created synthetic datasets by (a) choosing random probability distributions \( \{ \theta_i \}_{i \in S_{2,4}} \), \( \{ \phi_i \}_{i \in S_{2,4}} \), \( \{ \psi_i \}_{i \in S_{2,4}} \) \( \vdots \) and (b) generating action sequences of 100,000 users by following the generative process of our model (see Section 2.2). Then, we compared the probability distributions learned by our optimization algorithm with the ground-truth distributions in terms of cosine similarity [25]. We also compared the learned stage assignments with the ground-truth assignments in terms of accuracy (i.e., the proportion of correct assignments). Figure 6 shows the results averaged over 1,000 trials for each setting. As the number of stages \( k \) increased, similarity between trained and ground-truth parameters decreased. However, the increase in the number of action types \( (|A|) \) and time gap bins \( (|\Delta|) \) increased the similarity by making different stages more likely to have distinct probability distributions. In every setting, the learned parameters were reasonably close to the true values. Specifically, the cosine similarity was higher than 0.93 and the accuracy was higher than 0.82 in every setting. These results were not sensitive to the values of the hyperparameters.

5 RELATED WORK

Modeling user behavior has been extensively studied to tackle various kinds of tasks. One line of work aims to predict events in the future by learning the hidden patterns from historical behavior. At a microscopic level, for predicting each event, many models and algorithms have been proposed, including frequency-based pattern mining [5, 15], mechanistic model approaches [7], generative models using latent cluster variables [13, 19], and LSTM-based approaches [16, 30]. In terms of modeling approaches, our work is closest to the generative latent variable models. However, our model is designed to infer more interpretable, coarse-grained patterns of event sequences. Such macroscopic progressions cannot be directly extracted through the use of microscopic event models.

On the other hand, another line of literature presents methods for capturing more macroscopic views of user behaviors through clustering event sequences. Those methods can typically visualize the overall view of user behavior from a certain aspect — such as navigation patterns on websites [9], diagram of user activity transitions [6], clusters of user types [27], and topics of event streams [12, 20]. Our work is related to this line of work in the sense that it provides high-level insights. However, our model presents multi-dimensional insights on each individual user’s change, clusters of user actions, and transitions between these clusters.

Some recent work has focused on such multi-dimensional insights by grouping events and representing each entity’s progression over coarse event groups. This approach is thus able to account for relationships between groups of events [24], regime shifts in event streams [18], and evolution of users [21].

In particular, modeling the progression of users over the latent stages has been proposed to distinguish different patterns of progression including development of various diseases [29]. Our work is closely related to this work in the sense that both assume latent stages that given event sequences progress through as well as each observed event in the sequences depends on the current latent stage. However, while the previous work differently models the step-by-step progression through each latent class, our work does not address different classes but model different patterns of progression using transition probabilities between stages. Furthermore, our model defines starting moments and goal stages to clearly illustrate the multiple paths in a particular progression region of interest, whereas the previous work does not contain a component triggering the start or the goal.

6 CONCLUSIONS

In this work, we propose a behavior model where progressions of users’ behaviors from a starting state to a goal state are modeled as transitions over latent stages. These latent stages capture the progressions in three different aspects: types of actions, frequencies of actions, and directions of changes.

To fit our model to web-scale behavior logs, we propose a fast and memory-efficient optimization algorithm and extend it to multi-core, external-memory, and distributed settings. We theoretically and empirically demonstrate that our algorithm scales linearly with the size of the input data. Especially, the distributed version of our algorithm, implemented in the MapReduce framework, successfully handles a petabyte-scale dataset with one trillion actions.

We demonstrate the effectiveness of our model using a dataset from LinkedIn. Our model, however, can be applied to any dataset with a series of actions by different users over time. Our model discovers meaningful stages summarizing the progressions of LinkedIn users towards certain target states. We also show that stage information inferred by our model is useful for downstream tasks including prediction of next actions.
REFERENCES


[16] Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. 2017. NEMO: Next Career Move Prediction with Contextual Embedding. In WWW.

