Submitted to the Annals of Applied Statistics

JOINT MODELING OF PLAYING TIME AND PURCHASE PROPENSITY IN MASSIVELY MULTIPLAYER ONLINE ROLE PLAYING GAMES USING CROSSED RANDOM EFFECTS

BY TRAMBAK BANERJEE[1,a], PENG LIU[2,b], GOURAB MUKHERJEE[3,c], SHANTANU DUTTA[4,d] AND HAI CHE[5,e]

[1]*Analytics, Information and Operations Management, University of Kansas,* [a]*trambak@ku.edu*

[2]*Department of Marketing, Santa Clara University,* [b]*pliu2@scu.edu*

[3]*Department of Data Sciences and Operations, University of Southern California,* [c]*gmukherj@marshall.usc.edu*

[4]*Department of Marketing, University of Southern California,* [d]*shantanu@marshall.usc.edu*

[5]*Department of Marketing, University of California, Riverside,* [e]*chehai@ucr.edu*

Massively Multiplayer Online Role Playing Games (MMORPGs) offer a unique blend of a personalized gaming experience and a platform for forging social connections. Managers of these digital products rely on predictions of key player responses, such as playing time and purchase propensity, to design timely interventions for promoting, engaging and monetizing their playing base. However, the longitudinal data associated with these MMORPGs not only exhibit a large set of potential predictors to choose from but often present several other distinctive characteristics that pose significant challenges in developing flexible statistical algorithms that can generate efficient predictions of future player activities. For instance, the existence of virtual communities or 'guilds' in these games complicate prediction since players who are part of the same guild have correlated behaviors and the guilds themselves evolve over time and, thus, have a dynamic effect on the future playing behavior of its members. In this paper, we develop a *Crossed Random Effects Joint Modeling* (CREJM) framework for analyzing correlated player responses in MMORPGs. Contrary to existing methods that assume player independence, CREJM is flexible enough to incorporate both player dependence as well as time varying guild effects on the future playing behavior of the guild members. On a large-scale data from a popular MMORPG, CREJM conducts simultaneous selection of fixed and random effects in high-dimensional penalized multivariate mixed models. We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency. Besides providing superior predictions of daily playing time and purchase propensity over competing methods, CREJM also predicts player correlations within each guild which are valuable for optimizing future promotional and reward policies for these virtual communities.

1. Introduction. The online video game industry is revolutionizing the space of modern entertainment and social networking. As of August 2020, an estimated 3.1 billion people, which constitutes around 40% of the world population were playing video games (DFCIntelligence, 2020). A particular genre of video games that have seen very high levels of interest from users in recent years are the Massively Multiplayer Online Role Playing Games (MMORPGs). MMORPGs have been one of the biggest drivers of growth in the video games sector and are projected to grow at a historic rate of 9.22% between 2019 - 2023 versus 6.84% between 2015 -2019 (CISION, 2020). Players prefer MMORPGs over single player games

Keywords and phrases: Large-scale longitudinal data analysis, massively multiplayer online role playing games, monetization of digital products, online communities, guilds, cross classified random effect models.

for the social experience that they offer (Jin and Sun, 2015) and facilitated by the in-game chatting and video systems, players can make friends, form teams and collaborate on gaming tasks in these MMORPGs. Apart from the friendship networks, social connections in these games are achieved through guilds which are groups of players that have a shared interest. Guild, also known as 'clan', is a virtual community with hierarchical ranks that allow players to interact with each other and each player is a member of only one guild at any time point. Consequently, players nested within the same guild have correlated playing behavior. Moreover, the guilds themselves are dynamic communities that evolve over time as guild leaders recruit new members, existing members switch guilds, and the in-game activity and spending of guild members change over time (see panel (b) of Figure 3 in Section 2).

These phenomena create a highly dynamic environment which poses significant challenges in developing personalized promotional and monetization strategies for MMORPGs. For instance, game managers rely on predicting key player responses, such as daily duration of play (playing time) and purchases, to develop strategies for monetizing social networks (Park et al., 2018) and generating in-game advertising revenue (Terlutter and Capella, 2013). For analyzing such multivariate responses, the modeling framework must first address the complex inter-dependencies between (1) a player's decision to play, (2) their time spent playing the game and (3) their propensity to make an in-game purchase. Additionally, it must incorporate the two key structural features of MMORPGs wherein (1) players who are members of a guild have correlated playing behavior and (2) guilds have a dynamic effect on their member's playing behavior, such as their duration of play or purchase decisions. However, for modeling such multivariate player responses in MMORPGs, the statistical tools employed in contemporary research either assume that the player responses are not correlated or players in MMORPGs play as independent entities (Borbora et al., 2011; Zhang et al., 2017; Park et al., 2018). The first approach fails to uncover the positive, negative, or zero co-dependencies among the various responses of a player while the second approach ignores the dynamic influence of the guild and its members on a player's game behavior.

In this article we develop a *Crossed Random Effect Joint Modeling* (CREJM) framework for jointly modeling a player's daily duration of play and their purchase propensity in MMORPGs. In the context of single player mobile games, joint modeling of such player responses have been shown to be of significant importance for developing efficient marketing policies and for improved prediction of future playing behavior (Banerjee et al., 2020). Moreover, existing joint modeling frameworks, such as CEZIJ (Banerjee et al., 2020) and APLES (Hui, Müller and Welsh, 2018), can tackle multivariate player responses in the setting of single player games and, consequently, assume that players play the game as independent entities. CREJM, in contrast, is flexible enough to incorporate both player dependence as well as time varying guild effects on the future playing behavior of the guild members. We summarize the key features of the CREJM framework below:

- In MMORPGs the social influence of fellow gamers on a player can be from (a) team mates in the game-play, such as combat team mates, and (b) affiliations to online communities such as guilds. We incorporate the effects of fellow team mates (whose number can be very large in a player's lifetime) using global parameters in a Generalized Linear Mixed Models (GLMM) based joint estimation framework. To incorporate the effects of guilds we use guild specific random intercepts. Thus, to model a player's characteristics we use a cross-classified set-up with the crossing being a player's individual characteristics and their guild's influences (see Equation (1) in Section 3.1). Additionally, we model the dynamic influences of the guilds by extending Equation (1) through time-varying random intercepts (see Equation (2) in Section 3.2). Thus, our proposed CREJM framework is a system of cross classified random effect models that incorporate the key structural

features of MMORPGs wherein guilds have a dynamic effect on their member's playing behavior and players who are members of a guild have correlated playing behavior. While cross classified random effect models have been used in developing large-scale recommender systems (Koren, Bell and Volinsky, 2009; Khanna et al., 2013), estimation in such large-scale cross-classified designs involve several fundamental statistical challenges and is a topic of vibrant current research (Gao and Owen, 2020; Gao et al., 2017; Gao, 2017; Papaspiliopoulos, Roberts and Zanella, 2020; Ghosh, Hastie and Owen, 2022). To the best of our knowledge, the use of cross classified models as analytical tools for studying MMORPGs is new and we develop a disciplined algorithm for estimating the parameters in CREJM.

- The MMORPG data discussed in Section 2 involves longitudinal data on several daily player and guild characteristics. Existing literature (Zhang et al., 2017; Park et al., 2018; Wei et al., 2019; Huang, Jasin and Manchanda, 2019) judiciously uses a subset of these available attributes in a regression model. It is desirable to use all the available features and to choose the relevant set of gaming characteristics that provides best predictive performance. Our proposed GLMM based CREJM framework conducts simultaneous selection of fixed and random effects. It imposes a hierarchical structure on the selection mechanism and includes covariates either as fixed effects or composite effects where the latter are those covariates that have both fixed and random effects (Hui, Müller and Welsh, 2017a). Following Hui, Müller and Welsh (2017b,a); Banerjee et al. (2020), we use data-driven weighted ℓ_1 penalties on the fixed effects as well as on the diagonal entries of the covariance matrix of the player specific random effects (see Section 4). However, compared to the aforementioned works, CREJM involves an additional penalty for estimating the covariance matrix of the time varying guild specific random effects (see Equation (8)). We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency in Section 4.1.
- We conduct prediction of daily duration of play and purchase propensities of players conditional on the observed longitudinal information (see Section 6). Based on these dynamic predictions, game managers may develop personalized promotional and improved in-game advertising policies. In Section 5 of the supplement (Banerjee et al., 2023a), we use the CREJM framework for predicting the temporal trajectories of player correlations within each guild and with respect to their daily duration of play and purchase activity. Guilds with similar predicted correlation profiles over time provide valuable insights into the future playing behavior of their members and can be used to design and optimize promotional or reward policies specifically targeting those guild members (see figures 4 and 5 in Section 5 of the supplement (Banerjee et al., 2023a)).

2. Motivating Data. In this paper we consider the daily player level gaming information from a popular MMORPG where the players use one of the following four avatars; warrior, archer, sorceress and cleric, to play. The game is typically played on personal computers and is a "freemium" game (Kumar, 2014) as any player can download and play the game for free without paying any subscription fee. Figure 1 provides the game play wherein our MMORPG involves two main playing modes: player-versus-environment (PVE) mode and player-versus-player (PVP) mode. In the PVE mode, players accumulate experience points by completing missions and fighting monsters and villains in instanced dungeons. In the PVP mode, players practice and improve game skills in one-on-one or group combats. The main goal is character level progression and a player can elevate their game level by accumulating experience points, mainly through accomplishing missions and killing monsters in PVE combats. Social connections with other players are forged through friendship networks and guilds, and combat teams with guild members, friends and random players are formed to

complete adventure missions. Moreover, within a guild members are ranked hierarchically, from a leader at the top to associate leaders, senior members, junior members and finally new members at the bottom. Purchases constitute one of the primary revenue streams for the game managers and players purchase in-game items, such as weapons and costumes, to perform better in the PVE mode and thus complete their tasks more efficiently. Figure 2 provides

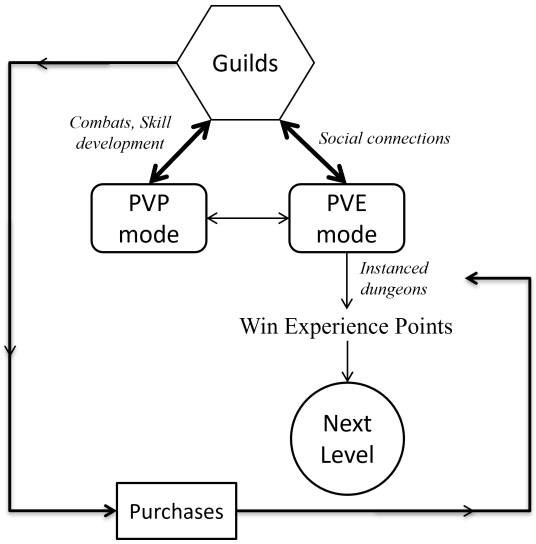


Fig 1: Game play flowchart.

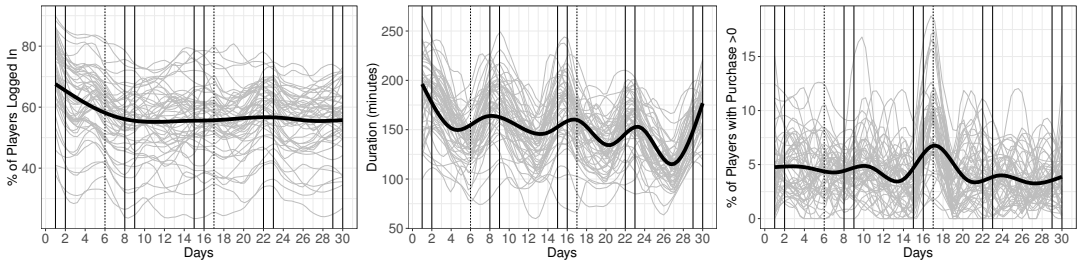
two screenshots of the web version of our MMORPG. The screenshot on the left represents a

Fig 2: Game screenshots from the web. Left: screenshot represents a Player Versus Environment (PVE) game and shows the avatar's character information on the top left, an online chat box in the middle and a control panel for items, missions, game points, and maps in the right. Right: screenshot shows the control panel of guild members.

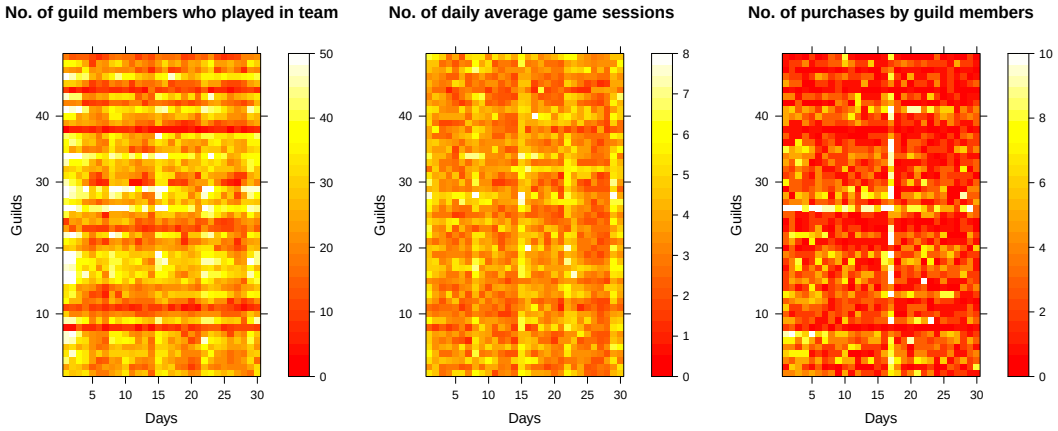
Player Versus Environment (PVE) game and shows the avatar's character information on the top left, an online chat box in the middle and a control panel for items, missions, game points, and maps in the right. The screenshot on the right shows the control panel of guild members who can use the online voice chatting system to collaborate with other guild members to complete missions together.

There are 5,188 players in our database that stores daily player level activity and their real money purchases for 30 consecutive days. We use a part of the data for estimation and the

other part as the hold out set for prediction (see details in Section 6). For each player the database holds a host of time dependent covariates that are generated through the game-play and include the focal player's in-game characteristics, characteristics that capture the focal player's interaction with their friends and the in-game activities of those friends. Additionally, on any one of the 30 days every player in our data has been part of a guild and so our data also hold time varying guild characteristics and covariates that capture the focal player's interaction with their guild. This information is available for $K = 50$ guilds that the players have been part of in those 30 days (See Tables 2 and 3 in Section 6 of the supplement for details).



(a)



(b)

Fig 3: Panel (a): Percentage of players who logged into the game by each guild over 30 days. The thick curve represents the overall login percentages in the data over the 30 days while each thin curve is a guild specific representation of the login percentages across time. The solid vertical lines indicate weekends while the dotted lines represent, respectively, the Chinese New Year (day 6) and Valentine's Day (day 17). Center: Average duration of play in minutes for each guild and conditional on login. The thick curve represents the overall average duration in the data across the 30 days. Right: Percentage of players with purchases > 0 by each guild and conditional on login. The thick curve represents the overall purchase percentages in the data across the 30 days. Panel (b): Heatmaps of the temporal evolution of three characteristics in each guild. Left: the number of guild members who played in a team. Center: the average game sessions played in the guild. Right: the number of purchases made in the guild.

In panel (a) of Figure 3, the left plot presents, for each guild, the percentage of players who logged into the game over the 30 days. The thick curve represents the overall login percentages across time while each thin curve is a guild specific representation of the login percentages across the 30 days. The solid vertical lines indicate weekends while the dotted lines represent, respectively, the Chinese New Year (day 6) and Valentine's Day (day 17). The

remaining two charts in panel (a) of Figure 3 are guild specific representation of the average duration of play in minutes (center plot) and the percentage of players with positive purchase (right plot), both conditional on login. The thick curve in these two plots are respectively, the average duration and the overall percentage of players with positive purchase over the 30 days. These charts indicate that playing behavior, in terms of login, duration of play and purchase, is substantially different across guilds. For instance, the observed login percentages across the 50 guilds range from 40% to 80% on day 1 while conditional on login the range of duration across guilds is atleast 80 minutes on any day. The purchase activities also vary considerably across the guilds with a notable exception on day 17 (Figure 3 panel (a) right) when all guilds seem to exhibit a spike in their purchase activities. This is related to an ongoing promotion at that time which coincided with Valentine's Day. In panel (b) of Figure 3 we further demonstrate that guilds are dynamic groups that evolve over time. We consider the following three characteristics that represent player engagement within a guild: the number of guild members who played in a team, the average game sessions played in the guild and the number of purchases made in the guild. For each of these three guild characteristics, panel (b) of Figure 3 presents a heatmap of their temporal evolution in each guild. It is interesting to note that with respect to the first two characteristics (panel (b) left and center plots), the temporal profiles of the guilds are relatively more dynamic than their temporal profiles for the number of purchases made (panel (b) right). This is expected since purchases are rare in our data and on an average less than 5% of the players who login make a purchase.

Our CREJM framework captures the heterogeneity in playing behavior across guilds by incorporating guild specific random effects. These guild specific random effects are time varying to account for the dynamic nature of the guilds as seen in panel (b) of Figure 3. Together, they incorporate two key structural features of MMORPGs into our joint modeling framework wherein (1) members of a guild have correlated playing behavior and (2) guilds have a dynamic effect on their member's playing behavior. In the following section we formally introduce the CREJM framework and discuss its key features. Additional details regarding the data are provided in Section 6 of the supplement (Banerjee et al., 2023a).

3. Cross Classified Random Effects Joint Modeling framework. Here we first introduce a generic cross classified random effect model and then present our proposed joint modeling framework CREJM.

3.1. *Cross Classified Random Effect Models.* Suppose we are interested in predicting a single longitudinal outcome y_{ij} which may denote the log duration of play for player i on day j. Here $i = 1, \ldots, n$ and $j = 1, \ldots, m$. At time j, let $d_{ijk} = 1$ if player i belongs to guild $k \in \{1, \ldots, K\}$ and 0 otherwise. A cross classified random effects model (Raudenbush, 1993; Raudenbush and Bryk, 2002) for K guilds may be specified as follows:

$$Y_{ij} = x_{ij}\beta + b_i + \sum_{k=1}^{K} d_{ijk}\Big(c_k + g_{jk}\gamma\Big) + \epsilon_{ij}, \tag{1}$$

where x_{ij}, g_{jk} are some player and guild specific predictors at time j. Here (β, γ) represent the vector of unknown fixed effect coefficients, $b_i \overset{i.i.d}{\sim} N(0, \sigma_1^2)$, $c_k \overset{i.i.d}{\sim} N(0, \sigma_2^2)$ are, respectively, the player and guild specific random intercepts that are independent of each other. We will assume that (i) $\epsilon_{ij} \sim N(0, \sigma_0^2)$ are independent of each other and the random intercepts, and (ii) the number of guilds K is fixed so that players may change guilds across time but no new guilds are formed, and existing guilds are not dissolved. Note that under model (1) the correlation between the log duration of play $(Y_{ij}, Y_{i'j})$ of two players (i, i') belonging to the same guild k at time j is non zero. Furthermore, in model (1) the guild random effects

$(c_k : 1 \leq k \leq K)$ do not vary over time which indicates that the guilds are static and exert the same effect on a player's duration of play Y_{ij} over time. However, as discussed in Section 2, guilds are dynamic entities and their effect on the playing behavior, duration of play in this example, changes over time. To address this possibility, Equation (1) may be modified to include time varying guild random effects as follows:

$$Y_{ij} = x_{ij}\beta + b_i + \sum_{k=1}^{K} d_{ijk}\Big(c_{jk} + g_{jk}\gamma\Big) + \epsilon_{ij} \tag{2}$$

where c_{jk} now depends on time and one may assume $\boldsymbol{c}_k = (c_{1k}, \ldots, c_{mk}) \sim N_m(\mathbf{0}, \mathbf{\Lambda})$ to emphasize the dependence between $c_{jk}, c_{j'k}$ through the covariance matrix $\mathbf{\Lambda}$ which can be unstructured, banded or first-order autoregressive (see for example Cafri, Hedeker and Aarons (2015); Cafri and Fan (2018)). Figure 4 presents a schematic representation

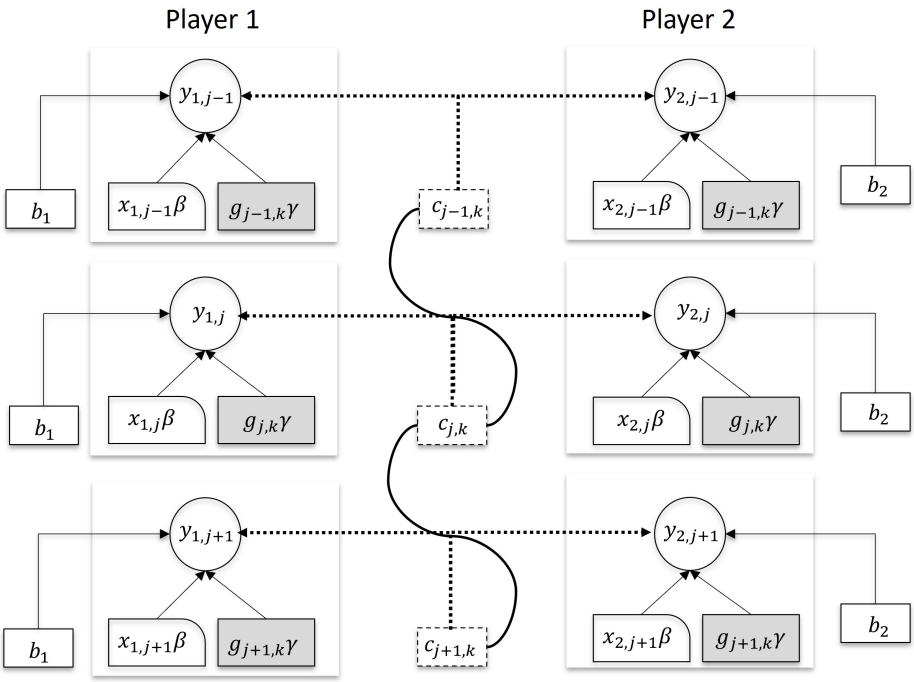


Fig 4: Schematic representation of model (2) for two players $(1, 2)$ who are part of guild k across the three time points $\{j-1, j, j+1\}$. The curved lines indicate that the guild random effects $\{c_{j-1,k}, c_{j,k}, c_{j+1,k}\}$ are correlated. The dotted arrows indicate that these guild random effects are common for both the players. The shaded rectangular boxes are the guild specific predictors $\{g_{j-1,k}, g_{j,k}, g_{j+1,k}\}$ that the players share.

of model (2) for two players $(1, 2)$ who are part of guild k across the three time points $\{j-1, j, j+1\}$. These players share the same guild specific predictors $\{g_{j-1,k}, g_{j,k}, g_{j+1,k}\}$ that are represented in the shaded rectangular boxes. The corresponding guild random intercepts $\{c_{j-1,k}, c_{j,k}, c_{j+1,k}\}$ are correlated which is shown via curved lines in Figure 4. The black dotted arrows indicate that these guild random effects are common for both the players and play the dual role of introducing dependence between the log duration of play for players 1 and 2 in guild k as well as exerting a dynamic effect on their log duration of play. In Section 3.2 we present our proposed joint modeling framework CREJM which extends model (2) to the case of a vector of longitudinal responses that are modeled jointly.

3.2. *CREJM framework.* We consider data from n players where every player $i = 1, \ldots, n$ is observed over m time points. Denote Y_{ij} as the duration of play for player i on day j with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})$ and recall from Section 3.1 that at time j, $d_{ijk} = 1$ if player i belongs to guild $k \in \{1, \ldots, K\}$, and 0 otherwise. Suppose α_{ij} is the indicator of the event that player i logs into the game on day j ($Y_{ij} > 0$) and ξ_{ij} is the indicator of their purchase activity with $\pi_{ij} = \mathbb{P}(\alpha_{ij} = 1)$, $q_{ij} = \mathbb{P}(\xi_{ij} = 1 | \alpha_{ij} = 1)$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{im})$ and $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{im})$. Thus, π_{ij} here represents a player's daily login probability whereas q_{ij} corresponds to their purchase propensity conditional on the event that the player has logged into the game on day j. We jointly model the three components $[\boldsymbol{\alpha}_i, \mathbf{Y}_i, \boldsymbol{\xi}_i]$ given the observations. Denote $\mathcal{I}$ to be the full set of p_0 predictors in the data with $\mathcal{I}_f \subset \mathcal{I}$ as the set of player specific fixed effects (time varying or not) predictors and $\mathcal{I}_c \subset \mathcal{I}$, with $\mathcal{I}_c \cap \mathcal{I}_f = \emptyset$, as the set of time varying player specific predictors which are modeled by both fixed and random effect coefficients. Such predictors that have both fixed and random effect components in the model are called composite effect predictors (Hui, Müller and Welsh, 2017a). Finally, let $\mathcal{I}_g = \mathcal{I} \backslash \{\mathcal{I}_c \cup \mathcal{I}_f\}$ be the set of guild specific time varying fixed effect predictors. Let $p_f = |\mathcal{I}_f|$, $p_c = |\mathcal{I}_c|$, $p_g = |\mathcal{I}_g|$ and so, $p_g + p_c + p_f = p_0$. For each of the three models, $s = 1, 2, 3$, let $\boldsymbol{x}_{ij}^{(s)} = (x_{ijr}^{(s)} : r \in \mathcal{I}_f \cup \mathcal{I}_c)$, $\boldsymbol{z}_{ij}^{(s)} = (z_{ijr}^{(s)} : r \in \mathcal{I}_c)$ denote, respectively, the set of player specific fixed and random effect predictors in the s^{th} model and let $\boldsymbol{g}_{jk}^{(s)} = (g_{jkr}^{(s)} : r \in \mathcal{I}_g)$ be the corresponding set of guild specific fixed effect predictors. We denote player i specific random effects by $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \boldsymbol{b}_i^{(3)})$ and the time varying guild specific random intercepts for guild k are denoted $\boldsymbol{c}_k = (\boldsymbol{c}_k^{(1)}, \boldsymbol{c}_k^{(2)}, \boldsymbol{c}_k^{(3)})$ with $\boldsymbol{c}_k^{(s)} = (c_{jk}^{(s)} : 1 \le j \le m)$ and $\boldsymbol{c}^{(s)} = (\boldsymbol{c}_k^{(s)} : 1 \le k \le K)$.

We now discuss the models for duration of play and purchase propensity. First note that player i logs into the game only at some time points, and so the observed duration of play $\mathbf{Y}_i$ has a mix of zeros and positive observations. To capture both the prevalence of these zeros and potential large values observed in the support of Y_{ij}, we consider a zero inflated Log Normal model for Y_{ij} in Equation (3). Thus, Y_{ij} has a mixture distribution with pdf,

$$g_1(\alpha_{ij}, y_{ij} | \boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}) = (1 - \pi_{ij}) \mathbb{I}\{\alpha_{ij} = 0\} + \pi_{ij} (\sigma y_{ij})^{-1} \phi\Big(\frac{\log y_{ij} - \mu_{ij}}{\sigma} \Big) \mathbb{I}\{\alpha_{ij} = 1\}, \tag{3}$$

where

$$\text{logit}(\pi_{ij}) = \boldsymbol{x}_{ij}^{(1)'} \boldsymbol{\beta}^{(1)} + \boldsymbol{z}_{ij}^{(1)'} \boldsymbol{b}_i^{(1)} + \sum_{k=1}^{K} d_{ijk} \Big(c_{jk}^{(1)} + \boldsymbol{g}_{jk}^{(1)'} \boldsymbol{\gamma}^{(1)} \Big), \tag{4}$$

$$\mu_{ij} = \boldsymbol{x}_{ij}^{(2)'} \boldsymbol{\beta}^{(2)} + \boldsymbol{z}_{ij}^{(2)'} \boldsymbol{b}_i^{(2)} + \sum_{k=1}^{K} d_{ijk} \Big(c_{jk}^{(2)} + \boldsymbol{g}_{jk}^{(2)'} \boldsymbol{\gamma}^{(2)} \Big). \tag{5}$$

In Equation (4) we use a logistic regression model with player specific and guild specific random effects to model the login indicator α_{ij}, while an identity link is used to model expected log duration of play in Equation (5). Now, a player can potentially purchase ($\xi_{ij} = 1$) only if they log into the game on day j ($\alpha_{ij} = 1$) and, even if the player logs in, they may not exhibit a positive purchase. Thus, conditional on the player logging into the game, we model the binary response $\xi_{ij} | \alpha_{ij} = 1$ with the covariates and the random effects through a logit link in equations (6), (7).

$$g_2(\alpha_{ij}, \xi_{ij} | \boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(3)}, \boldsymbol{c}^{(1)}, \boldsymbol{c}^{(3)}) = (1 - \pi_{ij}) \mathbb{I}\{\alpha_{ij} = 0\} + \tag{6}$$

$$\pi_{ij}\Big[(1-q_{ij})\mathbb{I}\{\xi_{ij}=0\}+q_{ij}\mathbb{I}\{\xi_{ij}=1\}\Big]\mathbb{I}\{\alpha_{ij}=1\}$$

where

$$\text{logit}(q_{ij}) = \boldsymbol{x}_{ij}^{(3)'}\boldsymbol{\beta}^{(3)} + \boldsymbol{z}_{ij}^{(3)'}\boldsymbol{b}_i^{(3)} + \sum_{k=1}^{K} d_{ijk}\left(c_{jk}^{(3)} + \boldsymbol{g}_{jk}^{(3)'}\boldsymbol{\gamma}^{(3)}\right) \tag{7}$$

In equations (3) and (6) the dependence on the fixed effects and the covariates are kept implicit in the notations and only the involved random effects are explicitly demonstrated. The three responses modeled in equations (4), (5) and (7) are interrelated as they carry information about the playing behavior of individuals as well as the guilds. To model the association between these responses we correlate the random heterogeneous effects from each of the responses. Specifically, we let $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \boldsymbol{b}_i^{(3)}) \overset{i.i.d}{\sim} N_{3p_c}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{c}_k = (\boldsymbol{c}_k^{(1)}, \boldsymbol{c}_k^{(2)}, \boldsymbol{c}_k^{(3)}) \overset{i.i.d}{\sim} N_{3m}(\mathbf{0}, \boldsymbol{\Lambda})$. Moreover, we assume that (a) for any $\{i, j, k\}$, b_{iu} is uncorrelated with $c_{jk}^{(s)}$ for all $u = 1, \ldots, 3p_c$, and (b) $\boldsymbol{\Lambda}$ is such that for $(s, s') \in \{1, 2, 3\}$, $\text{Cov}(c_{kj}^{(s)}, c_{kj'}^{(s')}) = 0$ if $|j - j'| > t'$ which indicates that although the guild specific effects are dynamic, the persistence of past effects vanish after a gap of t' time points. In the context of our MMORPG data that we analyze in Section 6, such a banded structure on $\boldsymbol{\Lambda}$ is natural since players do not login to the game daily and so the persistence of past guild effects is limited. The aforementioned banded structure on $\boldsymbol{\Lambda}$ not only captures an important aspect of the nature of guild effects but it also drastically reduces the number of non-zero elements of $\boldsymbol{\Lambda}$ which are finally estimated under the CREJM framework.

4. Variable Selection in CREJM. The daily data generated by a MMORPG usually hold several player and guild level characteristics. Recall that the proposed CREJM framework involves modeling three responses: Login indicator (4), duration of play conditional on login (5) and purchase propensity conditional on login (7). For the login indicator model, which forms the base of our joint model, it is possible to judiciously choose, based on some prior knowledge or heuristics, a subset of predictors that may predict a player's future login probability (see for example Zhang et al. (2017); Park et al. (2018); Wei et al. (2019); Huang, Jasin and Manchanda (2019)). However for the remaining two conditional models, choosing such relevant predictors from a large list of potential predictors only based on prior knowledge is difficult as the dynamics of the conditional response variable may render conventional expert judgment incorrect. Thus, to identify important characteristics that may help predict player duration of play and purchase propensity in these games, we conduct automated variable selection in the mixed model framework of equations (4), (5) and (7). Under such a framework selection of fixed and random effect components has received considerable attention. For instance, Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) proposed penalized likelihood procedures to simultaneously select fixed and random effect components under the special case of a linear mixed effect model, while Fan and Li (2012), Peng and Lu (2012) and Lin, Pang and Jiang (2013) conduct selection of fixed and random effects using a two stage approach. Several procedures to select only the fixed effects or the random effects have also been proposed under a GLMM framework; see Pan and Huang (2014); Hui, Müller and Welsh (2018) and the references therein. Recently, proposals for hierarchical variable selection in GLMMs have been introduced (Hui, Müller and Welsh, 2017a; Banerjee et al., 2020) wherein the selection mechanism conducts joint selection of fixed and random effects in a hierarchical manner such that a candidate random effect is included into the model only if the corresponding fixed effect is in the model. In this section, we discuss the variable selection mechanism in CREJM that conducts such hierarchical selection of fixed and random

effects components in multivariate mixed models and ensures that non-zero random effects in the model are accompanied by their corresponding nonzero fixed effects.

Let

$$\boldsymbol{\Theta} \;=\; (\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\boldsymbol{\beta}^{(3)},\boldsymbol{\gamma}^{(1)},\boldsymbol{\gamma}^{(2)},\boldsymbol{\gamma}^{(3)},\sigma,\mathrm{vec}(\boldsymbol{\Sigma}),\mathrm{vec}(\boldsymbol{\Lambda})) \;:=\; (\boldsymbol{\Gamma},\sigma,\mathrm{vec}(\boldsymbol{\Sigma}),\mathrm{vec}(\boldsymbol{\Lambda})),$$

denote the vector of all parameters to be estimated. Here $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}^{(s)} : s = 1,2,3)$ and $\boldsymbol{\Gamma}^{(s)} = (\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}) := \{\Gamma_{sr} : r \in \mathcal{I}\}$. The marginal log-likelihood of the observed data under the joint model is:

$$\ell_n(\boldsymbol{\Theta}) = \log \int \Big\{ \prod_{i=1}^{n}\prod_{j=1}^{m} p\Big(\alpha_{ij}, y_{ij}, \xi_{ij} | \boldsymbol{b}_i, \boldsymbol{c}, \boldsymbol{\Gamma}, \sigma\Big)\Big\} p(\boldsymbol{b}|\boldsymbol{\Sigma}) p(\boldsymbol{c}|\boldsymbol{\Lambda}) \mathrm{d}\boldsymbol{b}\mathrm{d}\boldsymbol{c},$$

where $\boldsymbol{b} = \{\boldsymbol{b}_i : 1 \le i \le n\}$ and $\boldsymbol{c} = (\boldsymbol{c}_1,\ldots,\boldsymbol{c}_K)$. Let $\boldsymbol{\Sigma}^{(s)}_{rr}$ to be the variance of $b^{(s)}_{ir}$ for $r \in \mathcal{I}_c,\ s \in \{1,2,3\}$ and for any matrix $\boldsymbol{A}$, denote $\|\boldsymbol{A}\|_1 := \sum_{i,j}|\boldsymbol{A}_{ij}|$. We solve the following maximization problem involving a penalized log-likelihood function $\ell_n(\boldsymbol{\Theta})$ for variable selection in the CREJM framework:

$$\max_{\boldsymbol{\Theta},\boldsymbol{\Sigma}\succ 0,\boldsymbol{\Lambda}\succ 0} \ell_n(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3}\sum_{r\in\mathcal{I}} w_{sr}\Big(|\Gamma_{sr}| + d_{sr}\boldsymbol{\Sigma}^{(s)}_{rr}\mathbb{I}\{r\in\mathcal{I}_c\}\Big) - n\lambda_2\|\boldsymbol{P}*\boldsymbol{\Lambda}\|_1. \tag{8}$$

Here $(\lambda_1,\lambda_2) \in \mathbb{R}^2_+$ are the regularization parameters and $*$ denotes element-wise multiplication. We now discuss the two penalties associated with λ_1 and λ_2 in Equation (8). The penalty $\|\boldsymbol{P}*\boldsymbol{\Lambda}\|_1$, originally proposed in Bien and Tibshirani (2011), enforces a banded structure on the covariance matrix $\boldsymbol{\Lambda}$ of the guild specific random effects $\boldsymbol{c}_k = (\boldsymbol{c}^{(1)}_k,\boldsymbol{c}^{(2)}_k,\boldsymbol{c}^{(3)}_k)$ such that for $(s,s') \in \{1,2,3\}$, $\mathrm{Cov}(c^{(s)}_{jk}, c^{(s')}_{j'k}) = 0$ if $|j-j'| > t'$. This is achieved through the $3m \times 3m$ symmetric matrix $\boldsymbol{P}$ where, for $(u,v) \in \{1,\ldots,3m\}$,

$$\boldsymbol{P}(u,v) = \begin{cases} \mathbb{I}(|u-v| > t'), \text{ if } (l-1)m+1 \le u \le v \le lm,\ l = 1,2,3 \\ \mathbb{I}(|u-v(\mathrm{mod}\ m)| > t'), \text{ if } 1 \le u \le m,\ m+1 \le v \le 2m \\ \mathbb{I}(|m-u-v(\mathrm{mod}\ m)| > t'), \text{ if } m+1 \le u \le 2m,\ 2m+1 \le v \le 3m. \end{cases} \tag{9}$$

In Figure 5 we provide a representation of $\boldsymbol{P}$ using Equation (9) for three different choices of t' and with $m = 5$. Here the entries with $\boldsymbol{P}(u,v) = 1$ are in a darker shade while those with $\boldsymbol{P}(u,v) = 0$ are in a lighter shade. For a sufficiently large λ_2, the entries of $\boldsymbol{\Lambda}$ that correspond to the non-zero entries of $\boldsymbol{P}$ are shrunk towards 0. Moreover, the resulting covariance matrix $\boldsymbol{\Lambda}$ is denser for larger t' indicating that the guild effects persist longer. In Section 6 we discuss the choice of t' for our application involving the MMORPG data of Section 2.

In Equation (8), the penalty associated with λ_1 is designed to maintain the hierarchy in selecting the fixed and random effects. For instance, when $r \in \mathcal{I}_c$ the penalty ensures that either the corresponding fixed and random effect is shrunk to zero or only the random effect is shrunk to zero. The adaptive weights $(w_{sr}, d_{sr}) \in \mathbb{R}^2_+$ play a crucial role in this hierarchical selection mechanism. In Section 5 we discuss the construction of these weights and present an iterative algorithm that alternates between estimating $\boldsymbol{\Theta}$ and redefining the data-driven weights (w_{sr}, d_{sr}) such that the weights used in any iteration are computed from the solutions of the previous iteration (see Candes, Wakin and Boyd (2008); Zhao and Kočvara (2015); Lu, Lin and Yan (2015); Banerjee et al. (2020) for details on these kind of approaches).

We end this section with the remark that the maximization problem based on criterion (8) can be augmented with linear inequality constraints $\boldsymbol{A}^{(s)}\boldsymbol{\Gamma}^{(s)} \le \boldsymbol{a}^{(s)}$ that may incorporate domain expertise and impose monotonicity, sign or other structural constraints on the components of $\boldsymbol{\Gamma}^{(s)}$.

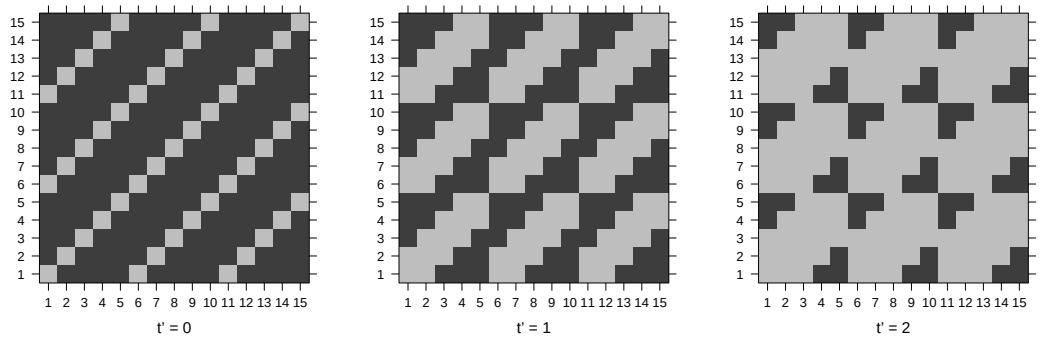


Fig 5: Representation of $\boldsymbol{P}$ using Equation (9) for three different choices of t' and with $m = 5$. Here the entries with $\boldsymbol{P}(u, v) = 1$ have a darker shade while those with $\boldsymbol{P}(u, v) = 0$ are in a lighter shade.

4.1. *Asymptotic Properties.* In this section we study the asymptotic properties of the variable selection procedure in CREJM. Our analysis will keep p_c fixed and allow $p = 3(p_f + p_g)$ to grow at a slower rate than n. We first introduce some notations where the dependence on p will be implicit and then state our main result.

Let the penalized likelihood criteria in Equation (8) be denoted by $\ell_n^{pen}(\boldsymbol{\Theta})$ where

$$\ell_n^{pen}(\boldsymbol{\Theta}) = \ell_n(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}} w_{sr} \Big(|\Gamma_{sr}| + d_{sr} \boldsymbol{\Sigma}_{rr}^{(s)} \mathbb{I}\{r \in \mathcal{I}_c\} \Big) - n\lambda_2 \|\boldsymbol{P} * \boldsymbol{\Lambda}\|_1. \tag{10}$$

Denote $\boldsymbol{\Theta}_0 = (\boldsymbol{\Gamma}_0, \text{vec}(\boldsymbol{\Sigma}_0), \text{vec}(\boldsymbol{\Lambda}_0))$ to be the true parameter values and $\tilde{p} = |\{\Gamma_{0r} : r \in \mathcal{I}_f \cup \mathcal{I}_g\}|$ be the number of true non-zero fixed effects in $\boldsymbol{\Gamma}_0$. Let $\boldsymbol{\Theta}_1$ denote the non-zero elements of $\boldsymbol{\Theta}_0$ and, without loss of generality, let $\boldsymbol{\Theta}_0 = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ where $\boldsymbol{\Theta}_2 = \boldsymbol{0}$. Similarly, for a local maximizer $\widehat{\boldsymbol{\Theta}}_n$ of Equation (8), we write $\widehat{\boldsymbol{\Theta}}_n = (\widehat{\boldsymbol{\Theta}}_{n1}, \widehat{\boldsymbol{\Theta}}_{n2})$. Denote $\boldsymbol{H}_n(\boldsymbol{\Theta}_0) = -n^{-1}\partial^2 \ell_n(\boldsymbol{\Theta}) / \partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^T |_{\boldsymbol{\Theta}_0}$ to be the observed Fisher Information matrix at $\boldsymbol{\Theta}_0$ with $\lambda_{\min}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0))$ and $\lambda_{\max}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0))$ being its minimum and maximum eigenvalues. We denote $\mathcal{F}_n = \{\boldsymbol{\Theta} : \boldsymbol{\Sigma} \succ 0, \boldsymbol{\Lambda} \succ 0\}$ to be the parameter space over which the maximization problem in Equation (8) is defined and impose the following regularity conditions that are needed in our technical analysis.

(A1) For all n, $\boldsymbol{H}_n(\boldsymbol{\Theta}_0)$ satisfies $0 < c_1 < \lambda_{\min}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0)) < \lambda_{\max}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0)) < 1/c_1 < \infty$ for some constant c_1.

(A2) For every $\epsilon > 0$, there exists a $\delta > 0$ such that for n large, $(1 - \epsilon)c_1 < \lambda_{\min}(\boldsymbol{H}_n(\boldsymbol{\Theta})) < \lambda_{\max}(\boldsymbol{H}_n(\boldsymbol{\Theta})) < (1 + \epsilon)/c_1$ for all $\boldsymbol{\Theta}$ satisfying $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\|_2 < \delta$.

(A3) The weights satisfy $w_{sr} = O_p(1)$, $d_{sr} = O_p(1)$ whenever $r \in \boldsymbol{\Theta}_1$, and for $\nu > 0$, $w_{sr} = O_p\{(n/p)^{\nu/2}\}$, $d_{sr} = O_p\{(n/p)^{\nu/2}\}$ whenever $r \in \boldsymbol{\Theta}_2$.

(A4) As $n \to \infty$, (a) $\lambda_1(n\tilde{p})^{1/2} \to 0$ (b) $\lambda_1(n/p)^{(\nu+3)/4} \to \infty$.

Condition (A1) ensures that at the true parameter value $\boldsymbol{\Theta}_0$ the observed Fisher information matrix is positive definite and its eigenvalues are uniformly bounded while condition (A2) extends this to a small neighborhood of $\boldsymbol{\Theta}_0$. These conditions are similar to assumptions A4 and A5 in Chen and Chen (2012). Conditions (A3) and (A4) are similar to assumptions (C5) and (C6) in Hui, Müller and Welsh (2017a). In particular (A3) requires that the data-driven adaptive weights exhibit different asymptotic behavior for the true zero and true non-zero

parameters while condition (A4) restricts the rate of growth of the regularization parameter λ_1 and allows p to grow with n such that $(p/n)^{(\nu+3)/4}(n\tilde{p})^{1/2} \to 0$ as $n \to \infty$.

THEOREM 4.1. *Under assumptions $A1 - A4$, there exists a local maximizer $\widehat{\boldsymbol{\Theta}}_n = (\widehat{\boldsymbol{\Theta}}_{n1}, \widehat{\boldsymbol{\Theta}}_{n2})$ of $\ell_n^{pen}(\boldsymbol{\Theta})$ such that $\|\widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}_0\|_2 = O_p(\sqrt{p/n})$ and $\mathbb{P}(\widehat{\boldsymbol{\Theta}}_{n2} = \mathbf{0}) \to 1$ as $n \to \infty$.*

Theorem 4.1, proved in Section 1 of the supplement (Banerjee et al., 2023a), establishes the selection consistency of the variable selection procedure under the CREJM framework in the sense that there exists a $\sqrt{n/p}$ consistent maximizer $\widehat{\boldsymbol{\Theta}}_n$ of $\ell_n^{pen}(\boldsymbol{\Theta})$ that identifies the true non-zero elements of $\boldsymbol{\Theta}_0$ with high probability as $n \to \infty$.

5. Estimation Procedure. In this section we discuss our estimation procedure that involves solving the maximization problem of Equation (8). Here the suffix n will be implicit in our notations.

The marginal likelihood $\ell(\boldsymbol{\Theta})$ in Equation (8) involves a high dimensional integral with respect to the random effects. In GLMMs these integrals often have no analytical form and several approaches, such as Laplacian approximations (Tierney and Kadane, 1986), adaptive quadrature approximations (Rabe-Hesketh et al., 2002), penalized quasi likelihood (PQL)(Breslow and Clayton, 1993) and EM algorithm (McCulloch, 1997), have been proposed to tackle this computational hurdle. We use an iterative algorithm which is similar to the Monte Carlo EM (MCEM) algorithm of Wei and Tanner (1990). In the context of crossed-classified random effects model, the MCEM algorithm has recently been used for large-scale (Koren, Bell and Volinsky, 2009) and parallel (Khanna et al., 2013) matrix factorization problems in machine learning.

The MCEM algorithm treats the random effects $(\boldsymbol{b}_i, \boldsymbol{c})$ as 'missing data' and obtains $\widehat{\boldsymbol{\Theta}}$, an estimate of $\boldsymbol{\Theta}$, by maximizing the expected value of the complete data likelihood $\ell^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}, \boldsymbol{c})$ where,

$$\begin{aligned}\ell^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}, \boldsymbol{c}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \log p(\alpha_{ij}, y_{ij}, \xi_{ij} | \boldsymbol{b}_i, \boldsymbol{c}, \boldsymbol{\Gamma}, \sigma) + \sum_{i=1}^{n} \log p(\boldsymbol{b}_i | \boldsymbol{\Sigma}) + \sum_{k=1}^{K} \log p(\boldsymbol{c}_k | \boldsymbol{\Lambda}) \\ &= \sum_{i=1}^{n} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i, \boldsymbol{c}).\end{aligned}$$

Denote the Q-function $\ell^{\mathsf{Q}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathbb{E}\ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i, \boldsymbol{c})$ where the expectation is over the conditional distribution of $(\boldsymbol{b}_i, \boldsymbol{c})$ given the observations at the current parameter estimates. Let $\boldsymbol{\Theta}^{(t)}$ denote the parameter estimates at iteration t. In iteration $t+1$, the MCEM algorithm performs the following two steps until convergence:

E-step - Evaluate $\ell_{(t)}^{\mathsf{Q}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{b}_i, \boldsymbol{c} | \boldsymbol{\Theta}^{(t)}, \mathbb{Y}_i} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i, \boldsymbol{c})$ where the expectation is over the conditional distribution of $(\boldsymbol{b}_i, \boldsymbol{c})$ given the observations $\mathbb{Y}_i := (\boldsymbol{\alpha}_i, \mathbf{Y}_i, \boldsymbol{\xi}_i)$ at the current estimates $\boldsymbol{\Theta}^{(t)}$. Now,

$$\mathrm{E}_{\boldsymbol{b}_i, \boldsymbol{c} | \boldsymbol{\Theta}^{(t)}, \mathbb{Y}_i} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i, \boldsymbol{c}) = \int \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i, \boldsymbol{c})\, p(\boldsymbol{b}_i, \boldsymbol{c} \mid \mathbb{Y}_i, \boldsymbol{\Theta}^{(t)})\, \mathrm{d}\boldsymbol{b}_i \mathrm{d}\boldsymbol{c}$$

and

$$p(\boldsymbol{b}_i, \boldsymbol{c} \mid \mathbb{Y}_i, \boldsymbol{\Theta}^{(t)}) = \exp\{-\ell_i(\boldsymbol{\Theta}^{(t)})\}\, p(\mathbb{Y}_i |\, \boldsymbol{\Theta}^{(t)}, \boldsymbol{b}_i, \boldsymbol{c})\, \phi_{3p_c}(\boldsymbol{b}_i | \mathbf{0}, \boldsymbol{\Sigma}^{(t)}) \phi_{3mK}(\boldsymbol{c} | \mathbf{0}, \boldsymbol{I}_{3mK} \otimes \boldsymbol{\Lambda}^{(t)}),$$

where, $\phi_q(\cdot\,|\,\mathbf{0}, \boldsymbol{\Sigma}^{(t)})$ is q dimensional normal density with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}^{(t)}$. In the display above, the expectation involves a multivariate integration with respect to the random effects $\boldsymbol{b}_i, \boldsymbol{c}$ which is evaluated by Monte Carlo integration. We approximate it as:

$$\left(\sum_{d=1}^{D} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i^d, \boldsymbol{c}^d)\; p(\mathbb{Y}_i|\; \boldsymbol{\Theta}^{(t)}, \boldsymbol{b}_i^d, \boldsymbol{c}^d)\right) \Big/ \left(\sum_{d=1}^{D} p(\mathbb{Y}_i|\; \boldsymbol{\Theta}^{(t)}, \boldsymbol{b}_i^d, \boldsymbol{c}^d)\right)$$

where $\boldsymbol{b}_i^d, \boldsymbol{c}^d$ are random samples from $\phi_{3p_c}(\cdot\,|\,\mathbf{0}, \boldsymbol{\Sigma}^{(t)})$, $\phi_{3mK}(\cdot\,|\mathbf{0}, \boldsymbol{I}_{3mK} \otimes \boldsymbol{\Lambda}^{(t)})$ respectively and $D = 2000$ is the number of monte carlo samples.

M-step - Solve the maximization problem in Equation (8):

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}, \boldsymbol{\Sigma} \succ 0, \boldsymbol{\Lambda} \succ 0}{\arg\max}\; \ell_{(t)}^{\mathsf{Q}}(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}} w_{sr}^{(t)} \left(|\Gamma_{sr}| + d_{sr}^{(t)}\; \boldsymbol{\Sigma}_{rr}^{(s)} \mathbb{I}\{r \in \mathcal{I}_c\}\right) - n\lambda_2 \|\boldsymbol{P} * \boldsymbol{\Lambda}\|_1, \tag{11}$$

where $w_{sr}^{(t)} = \min\left(|\Gamma_{sr}^{(t)}|^{-\nu}, \epsilon_1^{-1}\right)$ and $d_{sr}^{(t)} = \min\left(|\boldsymbol{\Sigma}_{rr}^{(s,t)}|^{-\nu} |\Gamma_{sr}^{(t)}|^{-\nu}, \epsilon_2^{-1}\right)$ with $\nu = 2$, are data driven adaptive weights that are updated at the end of every iteration of the MCEM. The construction of these weights is designed to maintain the hierarchy in selecting the fixed and random effects (see Candes, Wakin and Boyd (2008); Zhao and Kočvara (2015); Lu, Lin and Yan (2015) for details on these kind of approaches) and follows the approach described in Banerjee et al. (2020). For numerical stability and to allow a non-zero estimate in the next iteration given a zero valued estimate in the current iteration, we fix $\epsilon_1 = 10^{-2}$ (Candes, Wakin and Boyd, 2008). Moreover, whenever $|\Gamma_{sr}^{(t)}| = 0$ we enforce a large penalty on the corresponding diagonal element of $\boldsymbol{\Sigma}$ in iteration $(t+1)$ by setting $\epsilon_2 = 10^{-4}$. So if $r \in \mathcal{I}_c$, the penalty $w_{sr} d_{sr}$ on the diagonal elements of $\boldsymbol{\Sigma}$ encourages hierarchical selection of random effects. Further details on solving Equation (11) is provided in Section 2 of the supplement (Banerjee et al., 2023a).

6. Analysis of MMORPG Data. In this section we analyze the MMORPG data discussed in Section 2 and use the CREJM framework for modeling the three responses: Login Indicator, Duration of Play and Purchase Propensity. The data hold 18 player level gaming characteristics across $n = 5,188$ players observed over a period of 30 days and include the focal player's in-game characteristics, their virtual gender in the game, covariates that capture the focal player's interaction with their friends, the in-game activities of those friends, and covariates that are related to the focal player's interaction with their guild. It is well known in the marketing literature (Zhang et al., 2017; Park et al., 2018; Wei et al., 2019) that a player's activities in MMORPGs are deeply influenced by their friends. Thus, in addition to a player's individual playing and purchase history that are natural predictors of their future behavior, the CREJM framework also relies on the past activities of the focal player's friends for modeling their daily duration of play and purchase propensity. Along with these player specific covariates, the data hold 5 time varying guild characteristics for $K = 50$ guilds. These guild characteristics capture valuable information pertaining to the size of the guilds, the number of members that played as part of a team within the guild and other covariates measuring social contagion within the guilds. Finally, there are 2 indicator variables in the data that identify weekends and holidays in our panel. The description of these 25 predictors and their distribution are provided in Section 6 of the supplement (Banerjee et al., 2023a).

In the absence of any prior knowledge regarding which of these 25 predictors may appear in the three models, we conduct variable selection in CREJM. As discussed in Section 4,

CREJM conducts joint selection of fixed and random effects in a hierarchical manner and ensures that non-zero random effects in the model are accompanied by their corresponding fixed effects. We treat the 17 player specific covariates, excluding the indicator variable that captures the player's virtual gender, as potential composite effect predictors and allow the CREJM variable selection procedure to determine which predictors enter the model as just fixed effects or as fixed and random effect predictors. This is the most flexible representation of these player specific covariates because the selection procedure includes these covariates into the model either as fixed or composite effects in a data driven fashion. Finally, the 5 guild characteristics and the 3 indicator variables are treated as potential fixed effects with no corresponding random effect counterparts. In principle, it is possible to treat the guild characteristics as composite effects and let the selection mechanism select them as either fixed or composite effect predictors in the model. However, in such a scenario the CREJM variable selection algorithm would require an additional penalty in the optimization problem of Equation (8). Solving such an optimization problem is computationally challenging as it requires careful tuning of multiple regularization parameters and we do not pursue that extension in this article.

Overall, the CREJM selection mechanism must select random effects from a set of 54 potential random effects (17 for each of the 3 sub-models and their 3 intercepts) and select fixed effects from a set of 78 potential fixed effects (25 for each of the three sub-models and their 3 intercepts). For variable selection and estimation (Section 6.1), the CREJM framework relies on the first 15 days worth of data while the remaining 15 days are used for assessing its prediction performance (Section 6.2). Furthermore, CREJM considers time $j-1$ values of the predictors for modeling the three responses at time point j because at time j these player and guild characteristics are known only upto the previous time point. We initialize the CREJM algorithm and the adaptive weights $(w_{sr},\ d_{sr})$ in Equation (11) by fitting a saturated model on a subset of 500 players. As discussed in Section 3.2, the guild specific random effect covariance matrix $\mathbf{\Lambda}$ is such that for any $(s,s') \in \{1,2,3\}$, $\text{Cov}(c_{jk}^{(s)}, c_{j'k}^{(s')}) = 0$ if $|j-j'| > t'$, which indicates that the persistence of past guild effects vanish after a gap of t' time points. In this application we take $t'=3$ which allows the CREJM framework to capture guild effects from the previous 3 days. This is reasonable since the daily average time since last login has a mean of about 2 days across the first 15 days in our data. Finally, the regularization parameter λ_2 is fixed at 0.5 while λ_1 is chosen as that value of $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$ which minimizes BIC_λ where $\mathsf{BIC}_\lambda = -2\ell^{\mathsf{Q}}(\widehat{\mathbf{\Theta}}) + \log(n)\text{dim}(\widehat{\mathbf{\Theta}})$ (Bondell, Krishna and Ghosh, 2010; Lin, Pang and Jiang, 2013; Hui, Müller and Welsh, 2017a) and $\text{dim}(\widehat{\mathbf{\Theta}})$ is the number of non-zero components in $\widehat{\mathbf{\Theta}}$.

The R code for reproducing all the analysis in this paper is available in Banerjee et al. (2023b).

6.1. *The fitted joint model and its interpretations.* The analysis presented in this section relies on the first 15 days worth of data for variable selection and estimation using CREJM. To obtain the estimates of the fixed effect coefficients, and the random effect covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$, we rely on the selected variables for each of the three models and re-fit the selected model on the entire data using the unconstrained MCEM algorithm of Section 5. The list of selected predictors and their estimated fixed effect coefficients for the submodels of Login Indicator, Duration and Purchase Propensity are presented in Table 1 where 'PVP' stands for player versus player and 'PVE' stands for player versus environment. The column 'FE/CE' indicates whether the covariate is a candidate fixed effect (FE) or a composite effect (CE). The selected composite effects are those predictors that exhibit a (*) over their coefficient estimates in Table 1. All the selected fixed and random effects obey the hierarchical

structure discussed in Section 4. In what follows, we discuss the fitted coefficients and their interpretation for each of the three sub-models.

Login Indicator - The CREJM selection mechanism selects 9 player specific composite effect predictors and 1 guild specific predictor. The coefficient sign on the variables `pvp_play_time, pvp_kill_point, no_of_games` and `time_since` indicate that, all other things remaining constant, a more active player has a higher odds of logging into the game the next day. In particular, `pvp_play_time` and `no_of_games` increase the odds of login by almost 49% and 39% respectively. Moreover, higher a player's achievement level (`pvp_kill_point, quest_count`), the higher is the likelihood that they will login the next day. We find that the social contagion factors like experience with friends and guild membership also influence a player's likelihood of login. For example, a higher degree centrality as measured by the number of friends (`friend_count`), increases the odds of login by a factor of 2. Interestingly, a larger guild size (`guildmem_count`) reduces the odds of login by more than 30%. This negative relationship is consistent with previous literature (Hackman and Vidmar, 1970), which has suggested that a larger guild size usually reduces a guild members' satisfaction and dilutes their social identity, thus, reducing their likelihood of logging into the game.

TABLE 1

Selected fixed effect coefficients and their estimates under the submodels Login Indicator, Duration of Play and Purchase Propensity. The selected composite effects are those predictors that exhibit a (*) over their coefficient estimates in Table 1. See Table 2 in Section 6 of the supplement (Banerjee et al., 2023a) for a detailed description of the predictors.

Type	FE/CE	Predictors	Login Ind.	Duration	Purch Prop.
	CE	intercept	-1.043*	0.548*	-1.103*
Focal player's in-game characteristics	CE	level	-	-	-
	CE	pvp_play_time	0.397*	-	0.873*
	CE	pvp_kill_point	0.403*	-	-1.016*
	CE	quest_count	0.098*	-	-
	CE	mission_count	-	-	-
	CE	pve_time	-0.021*	0.082	-
	CE	no_of_game	0.330*	0.055	-
Focal player's interaction with their friends and the in-game activities of these friends	CE	friend_count	0.967*	0.405*	0.437*
	CE	friend_mean_level	-	-	-0.258*
	CE	no_of_friend_purch	-	-	-
	CE	total_friend_buy	-	-	-
	CE	no_of_friend_interact	0.360*	-	-
	CE	game_round_play_with_friends	-	-	-
Focal player's interaction with their guild	CE	guild_tenure	-	-	-
	CE	no_of_guildmem_interact	-	-	-
	CE	no_of_game_with_guildmem	-	-	-
Guild characteristics	FE	guildmem_interact	-	0.106	-
	FE	avg_game_with_guildmem	-	-	-
	FE	guild_total_purch	-	-	-
	FE	no_of_guildmem_purch	-	-	-
	FE	guildmem_count	-0.389	-	-0.655
Other characteristics	FE	gender	-	-	-
	FE	weekend	-	0.184	-
	FE	holiday	-	-	-
	CE	time_since	-0.558*	-	-0.503*

Duration of play - CREJM selects a relatively sparser model for Duration of play which is conditioned on the event of login. The selected model has 6 predictors of which two are player specific composite effect predictors and one is a guild specific predictor. For this model, we find that the coefficient signs on the two selected social contagion predictors (`friend_count`, `guildmem_interact`) are positive. This indicates that conditional on login and all other things remaining constant, a higher degree centrality as measured by

the number of friends (`friend_count`), leads to an overall increase in the future game time. Moreover, being part of a guild that has a higher `guildmem_interact`, which measures total number of guild members that played as part of a team within the guild, predicts a longer gaming time on the next day. This positive effect of guild interaction highlights the importance of social connections in enhancing a player's engagement in MMORPGs and is of managerial importance to the platforms. Note that a guild in which members form teams to play the game has a positive effect on a player's social identity and their loyalty to the game (Kang, Ko and Ko, 2009), while from the login indicator model an increase in the guild size reduces the player's satisfaction and their social identity. Thus different aspects of a player's guild experience can have different impact on the playing behavior. From Table 1, we also find evidence of a weekend effect (`weekend`) which predicts a relatively longer duration of play on weekends.

Purchase Propensity - CREJM selects 7 predictors of which six are player specific composite effect predictors and one is a guild specific predictor. We note that conditional on login, both individual experience and social contagion factors impact a player's future purchase propensity. For instance, the longer a player is engaged in the PVP mode (`pvp_play_time`), the higher is their odds of future purchases. We also observe that when a player has higher PVP killing points (`pvp_kill_point`) from the last login, it reduces their odds of future purchases. This indicates that more active but less skilled players are more likely to make future purchases, perhaps to increase their skill. The social contagion experience from friends (`friend_mean_level`, `friend_count`, `guildmem_count`) have different effects on the propensity to purchase. Consistent with prior literature on social contagion (Park et al., 2018), we observe that when degree centrality (`friend_count`) increases, it also increases the odds of future purchase, all other things remaining constant. However, the odds of future purchases are also impacted by the nature of friends a player plays with. When a player plays with friends who have a higher average skill level (`friend_mean_level`), it reduces the focal player's odds of future purchase. As in the case of login indicator model, we find that the coefficient sign on `time_since` (days since last login) is negative and when a player belongs to a guild with a larger guild size (`guildmem_count`) their likelihood for future purchase is lower.

We now discuss the estimated covariance matrix $\hat{\mathbf{\Sigma}}$ of the player specific random effects. In Figure 6 left, we present a heatmap of the 17×17 correlation matrix obtained from $\hat{\mathbf{\Sigma}}$. Within the three sub-models that were modeled jointly, we note that the random effects of the selected composite effect predictors are correlated. This indicates that players exhibit idiosyncratic playing profiles over time. Furthermore, we notice several instances of cross correlations across the three sub-models. For example from Figure 6 right, the random effect for the predictor `no_of_friend_interact` (no. of friends a player played with in teams during game sessions) in the Login Indicator model has a negative correlation with the random effect for `time_since` (days since last login) in the Purchase Propensity model. Similarly the random effects associated with `pvp_kill_point(3)` and `pvp_play_time(1)` demonstrate a positive correlation. These cross correlations suggest that the modeled responses are correlated for a player and the CREJM joint modeling framework allows such information pooling across the related responses which ultimately aids game managers in better predicting future player responses over time as discussed in Section 6.2.

6.2. *Prediction performance.* Here we present the prediction performance of the fitted joint model of Section 6.1 in dynamically predicting the three responses over the next 14 days ($j = 17, \ldots, 30$). For predicting the three responses, we consider two competing models-Benchmark I and Benchmark II, which we discuss below.

For Benchmark I we adopt a generalized linear model (GLM) setup and use the R-package glmmLasso (Schelldorfer, Meier and Bühlmann, 2014) for variable selection. In particular,

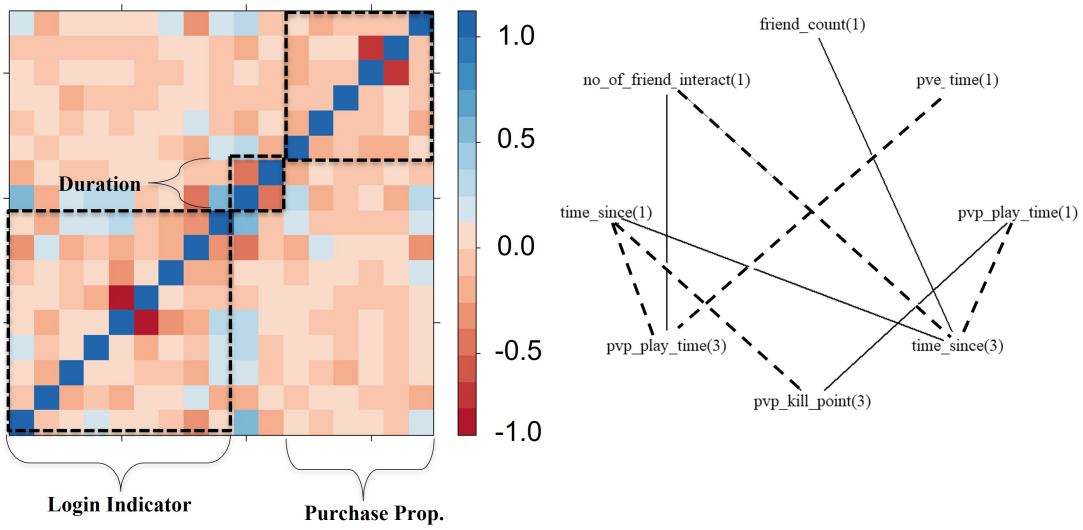


Fig 6: Left: Heatmap of the 17×17 correlation matrix obtained from $\hat{\boldsymbol{\Sigma}}$. On the horizontal axis are the selected composite effects for the three sub-models: Login Indicator, Duration and Purchase Propensity. The horizontal axis begins with the intercept from the Login Indicator model and ends with `time_since` from the Purchase Propensity model. Right: A network that demonstrate several cross correlations across the models. Solid lines represent positive correlations and dotted lines represent negative correlations. The model numbers are inside the parenthesis next to the predictor names.

Benchmark I does not model the three outcomes jointly, has no player or guild specific random effects and relies on logit links for Login Indicator and Purchase Propensity, and an identity link for log of positive Duration of play. In case of Benchmark II we consider the GLMM setup and use the R-package rpql (Hui, Müller and Welsh, 2017b) to perform joint selection of fixed and random effects with similar link functions as used in Benchmark I. The `rpql` package uses a regularized PQL (Breslow and Clayton, 1993) to perform simultaneous selection of fixed and player specific random effects but unlike CREJM it does not model the responses jointly and ignores the guild specific random effects. Predictions from Benchmark I are obtained by evaluating the fitted model on the validation data. However, since Benchmark II and CREJM are both mixed models, the prediction process must, respectively, estimate the latent random effects $(b_i^{(s)}, c_{jk}^{(s)})$ and appropriately account for the endogenous nature of the three responses. To do that we use the simulation scheme in Section 7.2 of Rizopoulos (2012) Rizopoulos (2012) and Section 3 of Rizopoulos (2011) Rizopoulos (2011), and estimate the expected time j responses given the observed responses until time $j-1$ (details provided in Section 3 of the supplement (Banerjee et al., 2023a)).

Table 2 summarizes the results of predictive performance of CREJM and the two benchmark models. For the binary responses of Login Indicator and Purchase Propensity, Table 2 presents the false positive (FP) rate and the false negative (FN) rate respectively averaged over the 14 time points. The FP rate measures the percentage of cases where the model incorrectly predicted login (or positive purchase) whereas the FN rate measures the percentage of cases where the model incorrectly predicted no login (or no purchase). For the login indicator model, Benchmark II exhibits the highest FP rate (Table 2) while Benchmark I has the lowest FN rate. CREJM, on the other hand, has the lowest FP rate and

TABLE 2

Results of predictive performance of CREJM and Benchmarks I, II. For Login Indicator and Purchase Propensity, the false positive (FP) rate / the false negative (FN) rate averaged over 14 time points are reported. For Duration of Play, the ratio of prediction errors (12) of Benchmarks I, II to CREJM averaged over the 14 time points are reported.

Submodels	Benchmark I	Benchmark II	CREJM
Login Indicator (FP / FN)	13.19 / 12.43	13.45 / 13.32	8.68 / 15.76
Duration of Play	3.07	1.83	1
Purchase Propensity (FP / FN)	0.00 / 2.55	0.3 / 2.13	0.07 / 2.17

its FN rate is relatively larger than the two benchmarks. However, for predicting the zero inflated response of Purchase Propensity, CREJM demonstrates a relatively superior performance over the Benchmark models. To assess the relative prediction performance for positive Duration of Play, we adopt a different approach and first calculate the time j prediction errors PE_j for the Benchmark models and CREJM as follows. For any model $\mathcal{M} \in \{\text{Benchmark I, Benchmark II, CREJM}\}$, we define $\text{PE}_j^{\mathcal{M}}$ at time $j = 17, \ldots, 30$ as

$$\text{PE}_j^{\mathcal{M}}(\mathbb{Y}^*, \widehat{\mathbb{Y}}^*) = \sum_{i=1}^{n} \left| \log Y_{ij}^* - \log \widehat{Y}_{ij}^* \right| \tag{12}$$

where $Y_{ij}^* = Y_{ij}$ if $\alpha_{ij} = 1$ and 1 otherwise, and $\widehat{Y}_{ij}^* = \widehat{Y}_{ij}$ if $\widehat{\alpha}_{ij} = 1$ and 1 otherwise, where $\widehat{Y}_{ij}$, $\widehat{\alpha}_{ij}$ are model $\mathcal{M}$ predictions of Duration and Login, respectively, for player i at time j. Note that $\text{PE}_j^{\mathcal{M}}$ measures the total absolute deviation of the prediction from the truth at any time j and for notational convenience its dependence on α_{ij}, $\widehat{\alpha}_{ij}$ has been suppressed. For the Duration model, Table 2 presents the ratio of the prediction errors of the Benchmarks to the CREJM model averaged over the 14 time points. So, a ratio in excess of 1 indicates a larger absolute deviation of the prediction from the truth when compared to CREJM. We note that the two Benchmark models exhibit prediction error ratios bigger than 1 with Benchmark I being the worse. Benchmark II exhibits a relatively better prediction error ratio than Benchmark I as it benefits from using the GLMM framework. However, unlike CREJM, it is unable to account for the dependencies between the responses which is reflected in its prediction error ratios being still bigger than 1.

Section 4 of the supplement (Banerjee et al., 2023a) includes additional analyses for comparing the predictive performance of CREJM and the two Benchmark models in dynamically predicting the three responses. In particular, we demonstrate that CREJM improves the average median prediction error on duration of play from about 4 minutes for Benchmark II to approximately 1.5 minutes. This is a substantial improvement considering the importance of accurate and reliable predictions of duration of play is concerned for developing personalized promotional and monetization strategies for MMORPGs.

6.3. *Time varying guild random effects and predicted player correlations.* Due to the paucity of space, we discuss the estimates $\hat{c}_{jk}^{(s)}$ of the time varying guild random effects $c_{jk}^{(s)}$ in Section 5 of the supplement (Banerjee et al., 2023a). Additionally, in Figure 4 of Section 5, we present three heat-maps, one for each of the three responses, that plot the mean predicted correlation over time of all players that are members of guild k where $k = 1, \ldots, 50$. Finally, we discuss how guilds with similar predicted correlation profiles over time provide valuable insights into the future playing behavior of their members and can be used to design promotion or reward policies specifically targeting those guild members.

7. Discussion. MMORPGs are dynamic environments that focus on the development of a player's in-game virtual character through persistent exploration of the gaming environment. One of their key distinguishing features from single player games is that they not only develop a player's gaming skill but also aim to provide an enhanced experience by involving collaboration and team building in game-play (Clements, 2012; Borbora et al., 2011). For instance, all other things remaining constant, we find that (Section 6.1) the future purchase potential is higher for a player with more friends whereas their future duration of play is predicted to be lower when they are part of a community that has many members as this leads to a loss in the player's social identity. Understanding the factors that influence such dichotomous future playing and purchase behavior is valuable for managers, as this enables them to better assess the effectiveness of their promotional activities in engaging players. Furthermore, a player's motivation for playing these games and making purchases of premium features can be broadly attributed to three factors: (a) their individual achievements in the game, (b) the social influence of their peers, and (c) their social interactions with online communities such as guilds. Therefore, a joint modeling framework that conducts prediction of correlated player attributes, such as their duration of play and purchase propensity, must rely not only on the player's past individual performance in the game but also on the influence of their friends and the communities that they are part of. Moreover, the ability of such a framework to incorporate player dependence as well as time varying network effects, such as guild effects, on the future playing behavior of its members can have a substantial impact on its prediction performance as demonstrated in Section 6.2.

The proposed CREJM framework is a GLMM based joint modeling framework that provides a unified approach for jointly modeling and predicting a player's daily duration of play and their purchase propensity in MMORPGs. Variable selection in CREJM relies on a flexible representation where initially all player specific covariates are set as composite effects, and then a data-driven procedure selects these covariates into the model as either fixed effects by removing the corresponding random effect or as composite effects. This approach avoids subjective usages of random effects which are very difficult to comprehend for the conditional models used for purchase propensity and duration of play. Moreover, variable selection in mixed effects models involve substantial computational burden when the relative importance of the candidate covariates are determined solely based on prior knowledge, information criteria or the fence (Jiang et al., 2008). CREJM uses a novel MCEM algorithm with iteratively adjusting weights to significantly reduce computational costs and enhance scalability. Thus, CREJM can be used for analyzing large gaming datasets.

We apply the CREJM framework on a large-scale data from a popular MMORPG and exhibit its superior performance in dynamically predicting player responses. Our analysis reveals that modeling the player responses jointly and allowing time-varying guild effects improves prediction accuracy relative to other benchmark models and such improved predictions provide valuable insights to the game managers for developing personalized promotional strategies. Additionally, CREJM provides data-driven evidence that (i) players exhibit idiosyncratic playing profiles, (ii) the player responses are correlated, and (iii) the guild effects are time-varying. Such insights are unavailable from simpler models that do not incorporate player specific random effects, fail to account for the correlations across a player's responses and ignore the time-varying nature of the guild effects. We use the estimates of the time-varying guild random effects from the CREJM framework to predict the temporal trajectories of player correlations within each guild and with respect to each of the three responses. These correlation profiles have substantial business implications for platform monetization and enhancing the effectiveness of existing promotional and reward policies. For instance, future promotional and retention strategies may be developed to increase player engagement in those guilds that are part of a cluster that represents the smallest magnitude of correlations.

Similarly, a guild cluster that exhibits relatively higher correlations may include loyalty rewards that further encourage player engagement in these guilds.

While this article demonstrates the applicability of the CREJM framework for the disciplined study of MMORPGs, it can be used in a wide range of other applications that need analyzing multiple longitudinal outcomes where the subjects, such as patients or firms, are nested within a dynamically evolving group structure, such as hospitals or firm size, and within those groups the subjects are not necessarily independent of each other. An interesting avenue for future research will be to study the dynamic effect of multiple groups with which the players may be crossed, such as friendship networks and guilds. We envision incorporating such multiple networks into our GLMM based joint modeling framework in the following two ways: First, we may assume that the three intercepts, one for each of the models, are player specific and time varying. Then, following Li, Levina and Zhu (2019); Le and Li (2020), we may incorporate a fusion penalty on these intercepts for each time point such that two players who are in the same friendship network at time j will have similar intercepts. This can be achieved through some constraint matrix that operates on $\mathbf{\Gamma}^{(s)}$ in Equation (8). However, variable selection and estimation in this model will be challenging, both theoretically and computationally, and will require further investigation. The second approach may instead include additional time varying random intercepts to model the dynamic effect of multiple groups with which the players may be crossed. Variable selection in such multiple membership mixed models is extremely interesting to us and we will pursue these extensions as part of future research.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. T. Banerjee was partially supported by the University of Kansas General Research Fund allocation #2302216. G. Mukherjee was partially supported by NSF DMS-1811866.

SUPPLEMENTARY MATERIAL

Supplement A: Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games using Crossed Random Effects
This supplement provides the proof of the theoretical result and additional numerical results.

Supplement B: Source code for "Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games using Crossed Random Effects"
This supplement holds the R code for reproducing all analyses in this paper.

REFERENCES

BANERJEE, T., MUKHERJEE, G., DUTTA, S. and GHOSH, P. (2020). A Large-Scale Constrained Joint Modeling Approach for Predicting User Activity, Engagement, and Churn With Application to Freemium Mobile Games. *Journal of the American Statistical Association* **115** 538–554.

BANERJEE, T., LIU, P., MUKHERJEE, G., DUTTA, S. and CHE, H. (2023a). Supplement to Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games using Crossed Random Effects.

BANERJEE, T., LIU, P., MUKHERJEE, G., DUTTA, S. and CHE, H. (2023b). Source code for Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role Playing Games using Crossed Random Effects.

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820.

BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics* **66** 1069–1077.

BORBORA, Z., SRIVASTAVA, J., HSU, K.-W. and WILLIAMS, D. (2011). Churn prediction in mmorpgs using player motivation theories and an ensemble approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* 157–164. IEEE.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88** 9–25.

CAFRI, G. and FAN, J. (2018). Between-within effects in survival models with cross-classified clustering: Application to the evaluation of the effectiveness of medical devices. *Statistical methods in medical research* **27** 312–319.

CAFRI, G., HEDEKER, D. and AARONS, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological methods* **20** 407.

CANDES, E. J., WAKIN, M. B. and BOYD, S. P. (2008). Enhancing sparsity by reweighted L1 minimization. *Journal of Fourier analysis and applications* **14** 877–905.

CHEN, J. and CHEN, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* 555–574.

CISION (2020). Implications of COVID-19 on the Global Role Playing Games Market. **News: September 2020**. Available at https://www.prnewswire.com/news-releases/implications-of-covid-19-on-the-global-role-playing-games-market-301139710.html.

CLEMENTS, R. (2012). RPGs Took Over Every Video Game Genre. Available at https://www.ign.com/articles/2012/12/12/rpgs-took-over-every-video-game-genre.

DFCINTELLIGENCE (2020). Global Video Game Consumer Segmentation. Available at https://www.dfcint.com/product/video-game-consumer-segmentation-2/.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics* **40** 2043.

GAO, K. (2017). Scalable Estimation and Inference for Massive Linear Mixed Models With Crossed Random Effects, PhD thesis, Stanford University.

GAO, K., OWEN, A. et al. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics* **11** 1235–1296.

GAO, K. and OWEN, A. B. (2020). Estimation and inference for very large linear mixed effects models. *Statistica Sinica, arXiv:1610.08088*.

GHOSH, S., HASTIE, T. and OWEN, A. B. (2022). Backfitting for large scale crossed random effects regressions. *The Annals of Statistics* **50** 560–583.

HACKMAN, J. R. and VIDMAR, N. (1970). Effects of size and task type on group performance and member reactions. *Sociometry* 37–54.

HUANG, Y., JASIN, S. and MANCHANDA, P. (2019). "Level Up": Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research* **30** 927–947.

HUI, F. K., MÜLLER, S. and WELSH, A. (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica* **27**.

HUI, F. K., MÜLLER, S. and WELSH, A. (2017b). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association* **112** 1323–1333.

HUI, F. K., MÜLLER, S. and WELSH, A. (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association* **113** 1759–1769.

IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503.

JIANG, J., RAO, J. S., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36** 1669–1692.

JIN, W. and SUN, Y. (2015). Understanding the Antecedents of Virtual Product Purchase in MMORPG: An Integrative Perspective of Social Presence and User Engagement. In *PACIS* 191.

KANG, J., KO, I. and KO, Y. (2009). The impact of social support of guild members and psychological factors on flow and game loyalty in MMORPG. In *2009 42nd Hawaii International Conference on System Sciences* 1–9. IEEE.

KHANNA, R., ZHANG, L., AGARWAL, D. and CHEN, B.-C. (2013). Parallel matrix factorization for binary response. In *2013 IEEE International Conference on Big Data* 430–438. IEEE.

KOREN, Y., BELL, R. and VOLINSKY, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30–37.

KUMAR, V. (2014). Making "freemium" work. *Harvard business review* **92** 27–29.

LE, C. M. and LI, T. (2020). Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*.

LI, T., LEVINA, E. and ZHU, J. (2019). Prediction models for network-linked data. *The Annals of Applied Statistics* **13** 132–164.

LIN, B., PANG, Z. and JIANG, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics* **22** 341–355.

LU, C., LIN, Z. and YAN, S. (2015). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing* **24** 646–654.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association* **92** 162–170.

PAN, J. and HUANG, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing* **24** 725–738.

PAPASPILIOPOULOS, O., ROBERTS, G. O. and ZANELLA, G. (2020). Scalable inference for crossed random effects models. *Biometrika* **107** 25–40.

PARK, E., RISHIKA, R., JANAKIRAMAN, R., HOUSTON, M. B. and YOO, B. (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing* **82** 93–114.

PENG, H. and LU, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis* **109** 109–129.

RABE-HESKETH, S., SKRONDAL, A., PICKLES, A. et al. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* **2** 1–21.

RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics* **18** 321–349.

RAUDENBUSH, S. W. and BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* **1**. Sage.

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829.

RIZOPOULOS, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

SCHELLDORFER, J., MEIER, L. and BÜHLMANN, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using L1-penalization. *Journal of Computational and Graphical Statistics* **23** 460–477.

TERLUTTER, R. and CAPELLA, M. L. (2013). The gamification of advertising: analysis and research directions of in-game advertising, advergames, and advertising in social network games. *Journal of advertising* **42** 95–112.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association* **81** 82–86.

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association* **85** 699–704.

WEI, Y., ZHANG, W., YANG, S. and CHEN, X. (2019). Online Communities and Social Network Structure. *Available at SSRN 3420525*.

ZHANG, C., PHANG, C. W., WU, Q. and LUO, X. (2017). Nonlinear effects of social connections and interactions on individual goal attainment and spending: Evidences from online gaming markets. *Journal of Marketing* **81** 132–155.

ZHAO, Y.-B. and KOČVARA, M. (2015). A new computational method for the sparsest solutions to systems of linear equations. *SIAM Journal on Optimization* **25** 1110–1134.