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

By Trambak Banerjee[‡], Peng Liu[§], Gourab Mukherjee[†,¶] Shantanu Dutta[¶] and Hai Che[‖]

University of Kansas[‡], Santa Clara University[§], University of Southern California[¶] and University of California, Riverside[‖]

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 usually 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.

[†]The research here was partially supported by NSF DMS-1811866.

[†]Corresponding author: gmukherj@marshall.usc.edu

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35

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.

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 were playing video games and represented around 40% of the world population [12]. The COVID-19 pandemic in particular has contributed to an unprecedented surge in this sector, both in terms of increased traffic from new subscribers as well as in substantial increases in the per-capita time spent in these games. For instance, Verizon reported an 82% increase in video game traffic over pre-COVID levels which was bigger than the increases in traffic related to virtual private network connections and associated collaboration tools [47]. In the U.S. alone, 35% of the gamers have reported a higher average time spent on gaming while until July 2020 nearly 3 out of every 4 people in the U.S. were playing video games, an increase of 32 million new subscribers to the game since 2018 [32]. This dual growth, in the number of new subscribers and on in-game time spent, has led to a remarkable increase in consumer spending in video games which reached 18.6 billion U.S Dollars in the fourth quarter (Q4) of 2020, an increase of 26% compared to Q4 2019 [33].

A particular genre of video games that have seen prodigious 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 [10]. These multiplayer games usually attract millions of active subscribers and along with the high graphics processing capacity of modern computers and game consoles, create an alternate fantasy world that provides a vibrant platform for forging social connections with other players ([44], p. 886). While the technological features of modern video games are quite appealing, players prefer MMORPGs over single player games for the social experience that they offer [25]. For example, players with different avatars in a MMORPG can make friends, form teams, cooperate and combat with other players in quests and battles. 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. Facilitated by the in-game chatting and video systems, a player, in her various avatars, can form teams to play the game and to share game items with other guild members. Consequently, players belonging to the same guild are expected to have correlated playing behavior. Moreover, the guilds themselves evolve over time as guild leaders recruit new members, existing members switch guilds, and the in-game activity and spending of guild members dynamically change

over time (see figure 2b 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 analysts rely on predicting key player responses, such as daily duration of play (playing time) and purchases, to develop strategies for monetizing social networks [36] and generating in-game advertising revenue [45]. For analyzing such multivariate responses, the modeling framework must first address the complex inter-dependencies between (1) a player's decision to play, (2) her time spent playing the game and (3) her 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 [4, 50, 36]. 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 her purchase propensity in MMORPGs. Existing joint modeling frameworks, such as CEZIJ [1] and APLES [23], can tackle such multivariate player responses in the setting of single player games and, consequently, assume that players play the game as independent entities. CREJM, in contrast, relies on a system of Cross Classified Random Effect Models [39] that is flexible enough to allow players in MMORPGs to be nested in guilds, thus accounting for the fact that players belonging to the same guild have correlated responses. Moreover, CREJM incorporates time varying idiosyncratic guild random effects that capture the dynamic influence of the guild on its member's playing behavior. Our proposed framework not only addresses the co-dependencies between daily duration of play and purchase propensity, but also provides a systematic understanding of how guilds influence playing behavior in MMORPGs. Besides being able to predict the future duration and purchase propensity of the individual players, the CREJM framework can be used to predict the time varying player correlations within each guild, both with respect to their daily duration of play and purchase activities. The ability to forecast such trajectories of player correlations is valuable

for managers, as this enables them to better assess the effectiveness of their promotional activities in engaging players and for tuning their monetization policies that are based on in-game advertising. 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 needs 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. We summarize the key features of the CREJM framework below:

Joint modeling of daily playing time and purchase decisions of players in MMORPGs – We propose a unified approach, CREJM, for jointly modeling a player's daily duration of play as well as her purchase decisions 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 [1]. However, it is well known in the marketing literature [50, 36, 49] that a player's activities in MMORPGs are deeply influenced by her friends. Thus, in addition to a player's individual playing and purchase history that are natural predictors of her future behavior, the CREJM framework also relies on the past activities of the focal player's friends for modeling her daily duration of play and purchase propensity. While MMORPGs 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 [11, 4]. As such, a player's motivation for playing MMORPGs and making purchases of premium features can be broadly attributed to two factors: (a) her individual achievements in the game, and (b) social interactions with peers. Here, controlling for these varied effects we jointly model player responses pertaining to daily duration of play and her purchase propensity using the MMORPG data described in Section 2.

A flexible framework that incorporates the dynamic influence of online communities such as guilds on the playing behavior – 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. While team mates greatly vary across games, each player can belong to only one guild at any time point. A guild

would typically have its unique objectives, discussion forums and characteristics [29]. 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 her guild's influences (see equation (3.1) in Section 3.1). Additionally, we model the dynamic influences of the guilds by extending (3.1) through time-varying random intercepts (see equation (3.2) in Section 3.2). Thus, our proposed CREJM relies on 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. Estimation in such large scale cross-classified designs involve several fundamental statistical challenges and is a topic of vibrant current research [17, 16, 15, 35, 18]. 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.

Simultaneous selection of fixed and random effects in high-dimensional penalized multivariate mixed models – The MMORPG data discussed in Section 2 involves longitudinal data on several daily player and guild characteristics. Existing literature [50, 36, 49, 20] 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 [21]. Following [22, 21, 1], 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 (4.1)). We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency in Section 4.1.

Prediction of the daily duration of play and purchase propensities of players in MMORPGs – 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 of individual player duration of play and purchase propensity, game managers may develop personalized promotional and improved in-game advertising policies. In Section 6.3 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 figure 8).

The rest of the paper is organized as follows. In Section 2 we describe our data from a popular MMORPG. In Section 3 we discuss our CREJM framework in details. Section 4 describes the variable selection procedure in CREJM and the associated asymptotic properties while Section 5 describes the estimation procedure. In Section 6, we use the CREJM framework to analyze the MMORPG dataset introduced in Section 2. The paper concludes with a discussion in Section 7. Proofs and other technical details are relegated to the supplementary materials.

2. Motivating Data. In this paper we consider the daily player level gaming information from a popular MMORPG wherein 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 [27] 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 her 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,

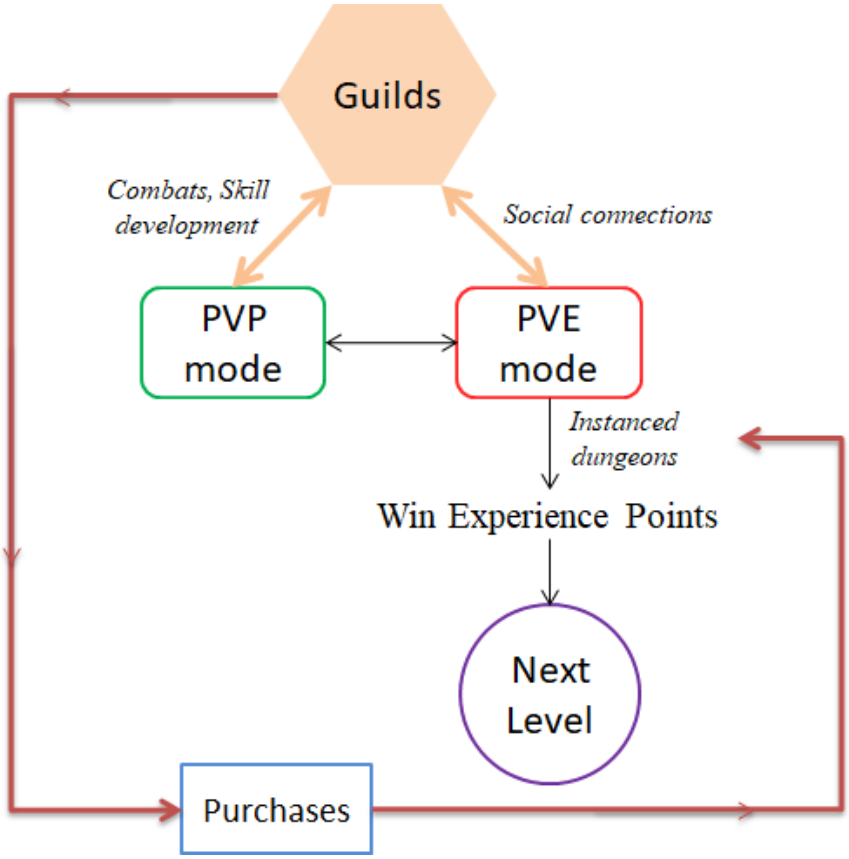
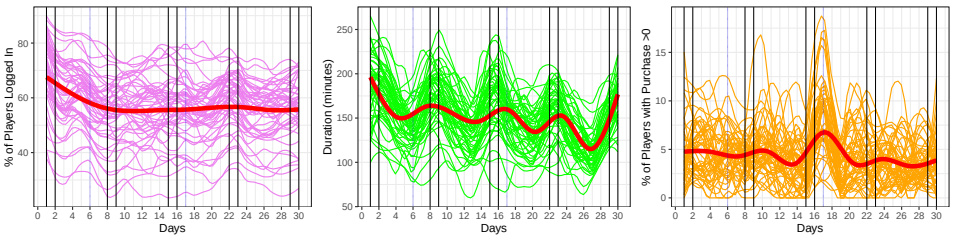


Fig 1: Game play flowchart.

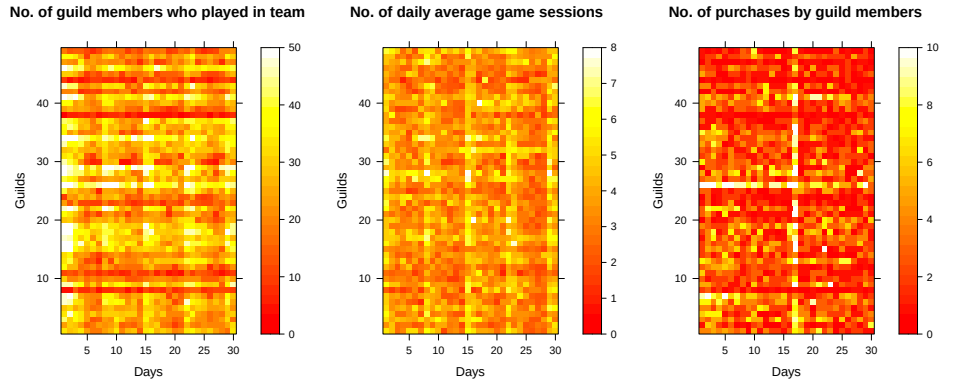
such as weapons and costumes, to perform better in the PVE mode and thus complete their tasks more efficiently.

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 her 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 her guild. This information is available for $K = 50$ guilds

that the players have been part of in those 30 days (See table 1 and summary table 2 in Appendix C of the supplementary material for details).



(a)



(b)

Fig 2: 2a - Left: Percentage of players who logged into the game by each guild over 30 days. The thick red curve represents the overall login percentages in the data over the 30 days while each pink curve is a guild specific representation of the login percentages across time. The black vertical lines indicate weekends while the blue 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 red 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 red curve represents the overall purchase percentages in the data across the 30 days.
2b - 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 figure 2a left, we present, for each guild, the percentage of players who logged into the game over the 30 days. The thick red curve represents the overall login percentages across time while each pink curve is a guild specific representation of the login percentages across the 30 days. The black vertical lines indicate weekends while the blue dotted lines represent, respectively, the Chinese New Year (day 6) and Valentine's Day (day 17). The remaining two charts in figure 2a 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 red 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 2a right) when all guilds seem to exhibit a spike in their purchase activities. This is related to an ongoing promotion at that time that coincided with Valentine's Day. In figure 2b 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, figure 2b presents a heatmap of their temporal evolution in each guild. It is interesting to note that with respect to the first two characteristics (figure 2b left and center), the temporal profiles of the guilds are relatively more dynamic than their temporal profiles for the number of purchases made (2b 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 figure 2b. 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.

3. Cross Classified Random Effects Joint Modeling framework. In this section 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_{ijk} which may denote the log duration of play for player i on day j in guild k. Here $i = 1, \ldots, n$, $j = 1, \ldots, m$ and $k = 1, \ldots, K$. A cross classified random effects model [39, 40] for K guilds may be specified as follows

$$y_{ijk} = x_{ij}\beta + b_i + c_k + g_{jk}\gamma + \epsilon_{ijk}, \tag{3.1}$$

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, $b_i \sim N(0, \sigma_1^2)$, $c_k \sim N(0, \sigma_2^2)$ are, respectively, the player and guild specific random intercepts that are independent of each other. We will assume that $\epsilon_{ijk} \sim N(0, \sigma_0^2)$ are independent of each other and the random intercepts. Note that under model (3.1) the correlation between the log duration of play $(Y_{ijk}, Y_{i'jk})$ of two players (i, i') belonging to the same guild k at time j is non zero. Furthermore, in model (3.1) 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_{ijk} 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 (3.1) may be modified to include time varying guild random effects as follows:

$$y_{ijk} = x_{ij}\beta + b_i + c_{jk} + g_{jk}\gamma + \epsilon_{ijk} \tag{3.2}$$

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 [7, 6]). Figure 3 presents a schematic representation of model (3.2) 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 blue boxes. The corresponding guild random intercepts $\{c_{j-1,k}, c_{j,k}, c_{j+1,k}\}$ are correlated which is shown via orange dotted lines in figure 3. 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

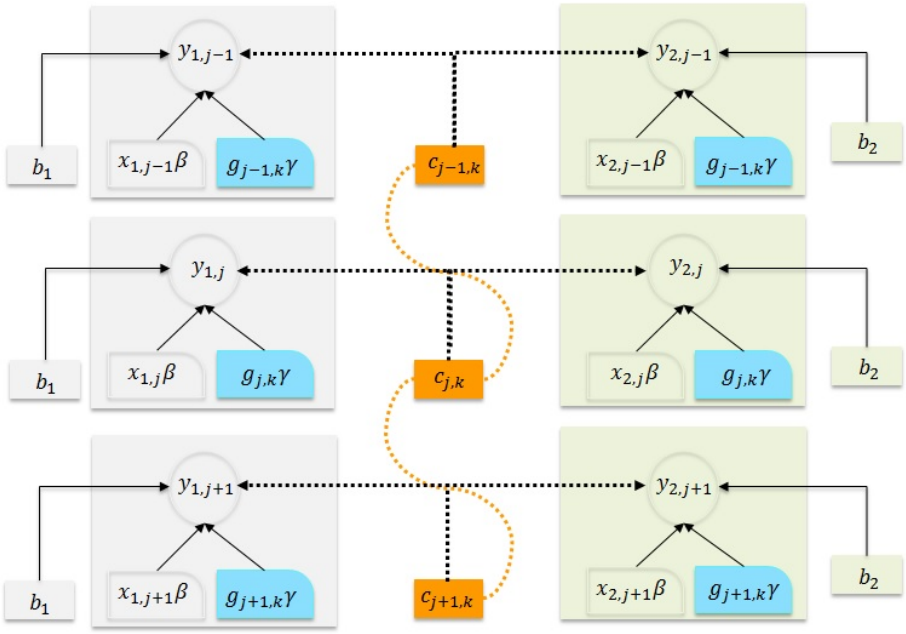


Fig 3: Schematic representation of model (3.2) for two players $(1, 2)$ who are part of guild k across the three time points $\{j-1, j, j+1\}$. The orange dotted lines indicate that the guild random effects $\{c_{j-1,k}, c_{j,k}, c_{j+1,k}\}$ are correlated. The blue boxes are the guild specific predictors $\{g_{j-1,k}, g_{j,k}, g_{j+1,k}\}$ that the players share. The black dotted arrows indicate that these guild random effects are common for both the players.

duration of play. In Section 3.2 below we present our proposed joint modeling framework CREJM which extends model (3.2) 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. At time j, let $d_{ijk} = 1$ if player i belongs to guild $k \in \{1, \ldots, K\}$ and 0 otherwise, and denote Y_{ij} as the duration of play for player i on day j with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})$. 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 her 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 players daily login probability whereas q_{ij} corresponds to her 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 combination of fixed and random effects. 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 (3.3). Thus, Y_{ij} has a mixture distribution with pdf,

$$g_1(\alpha_{ij}, y_{ij} | \boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}, \boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}) = (1 - \pi_{ij}) \mathbb{I}\{\alpha_{ij} = 0\} + \pi_{ij} (\sigma y_{ij})^{-1} \phi\Big(\frac{\log y_{ij} - \mu_{ij}}{\sigma}\Big) \mathbb{I}\{\alpha_{ij} = 1\}, \tag{3.3}$$

where

$$\mathsf{logit}(\pi_{ij}) = \boldsymbol{x}_{ij}^{(1)'} \boldsymbol{\beta}^{(1)} + \boldsymbol{z}_{ij}^{(1)'} \boldsymbol{b}_i^{(1)} + \sum_{k=1}^{K} d_{ijk} \Big(c_{jk}^{(1)} + \boldsymbol{g}_{jk}^{(1)'} \boldsymbol{\gamma}^{(1)} \Big), \tag{3.4}$$

$$\mu_{ij} = \boldsymbol{x}_{ij}^{(2)'} \boldsymbol{\beta}^{(2)} + \boldsymbol{z}_{ij}^{(2)'} \boldsymbol{b}_i^{(2)} + \sum_{k=1}^{K} d_{ijk} \Big(c_{jk}^{(2)} + \boldsymbol{g}_{jk}^{(2)'} \boldsymbol{\gamma}^{(2)} \Big). \tag{3.5}$$

In equation (3.4) 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 (3.5). Now, a player can potentially purchase ($\xi_{ij} = 1$) only if she logs into the game on day j ($\alpha_{ij} = 1$) and, even if the player logs in, she 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 (3.6), (3.7).

$$g_2(\alpha_{ij},\xi_{ij}|\boldsymbol{b}_i^{(1)},\boldsymbol{b}_i^{(3)},\boldsymbol{c}^{(1)},\boldsymbol{c}^{(3)}) = (1-\pi_{ij})\mathbb{I}\{\alpha_{ij}=0\}+ \pi_{ij}\Big[(1-q_{ij})\mathbb{I}\{\xi_{ij}=0\}+q_{ij}\mathbb{I}\{\xi_{ij}=1\}\Big]\mathbb{I}\{\alpha_{ij}=1\} \tag{3.6}$$

where

$$\mathsf{logit}(q_{ij}) = \boldsymbol{x}_{ij}^{(3)'}\boldsymbol{\beta}^{(3)}+\boldsymbol{z}_{ij}^{(3)'}\boldsymbol{b}_i^{(3)}+\sum_{k=1}^{K} d_{ijk}\Big(c_{jk}^{(3)}+\boldsymbol{g}_{jk}^{(3)'}\boldsymbol{\gamma}^{(3)}\Big) \tag{3.7}$$

In equations (3.3) and (3.6) 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 (3.4), (3.5) and (3.7) 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 $(s,s')\in\{1,2,3\}$, $\mathrm{Cov}(c_{kj}^{(s)},c_{kj'}^{(s')})=0$ 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.

4. Variable Selection in CREJM. The daily data generated by a MMORPG usually hold several player and guild level characteristics. 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 (3.4), (3.5) and (3.7). Under such a framework selection of fixed and random effect components has received considerable attention. For instance, [3] and [24] proposed penalized likelihood procedures to simultaneously select fixed and random effect components under the special case of a linear mixed effect model, while [13], [37] and [28] 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 [34, 23] and

the references therein. Recently, proposals for hierarchical variable selection in GLMMs have been introduced [21, 1] 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 hierarchical selection of fixed and random effects components in multivariate mixed models.

Let

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

denote the vector of all parameters to be estimated. Here $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}^{(s)} : s = 1, 2, 3)$ and $\boldsymbol{\Gamma}^{(s)} = (\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)}) \coloneqq \{\Gamma_{sr} : r \in \mathcal{I}\}$. 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 $\boldsymbol{\Sigma}_{rr}^{(s)}$ to be the variance of $b_{ir}^{(s)}$ for $r \in \mathcal{I}_c$, $s \in \{1, 2, 3\}$ and for any matrix $\boldsymbol{A}$, denote $\|\boldsymbol{A}\|_1 \coloneqq \sum_{i,j} |\boldsymbol{A}_{ij}|$. We solve the following maximization problem involving a penalized log-likelihood function $\ell_n(\boldsymbol{\Theta})$ for variable selection in the CREJM framework:

$$\max_{\boldsymbol{\Theta}, \boldsymbol{\Sigma} \succ 0, \boldsymbol{\Lambda} \succ 0} \ell_n(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}} w_{sr} \Big(|\Gamma_{sr}| + d_{sr} \boldsymbol{\Sigma}_{rr}^{(s)} \mathbb{I}\{r \in \mathcal{I}_c\} \Big) - n\lambda_2 \|\boldsymbol{P} * \boldsymbol{\Lambda}\|_1. \tag{4.1}$$

Here $(\lambda_1, \lambda_2) \in \mathbb{R}_+^2$ are the regularization parameters and $*$ denotes element-wise multiplication. In equation (4.1), 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 [8, 51, 30, 1] for details on these kind of approaches).

The penalty $\|\boldsymbol{P} * \boldsymbol{\Lambda}\|_1$, originally proposed in Bien and Tibshirani (2011) [2], enforces a banded structure on the covariance matrix $\boldsymbol{\Lambda}$ of the guild specific random effects $\boldsymbol{c}_k = (\boldsymbol{c}_k^{(1)}, \boldsymbol{c}_k^{(2)}, \boldsymbol{c}_k^{(3)})$ such that for $(s, s') \in \{1, 2, 3\}$, $\text{Cov}(c_{kj}^{(s)}, c_{kj'}^{(s')}) = 0$ if $|j - j'| > t'$. This is achieved through the $3m \times 3m$ symmetric matrix $\boldsymbol{P}$ where, for $(u, v) \in \{1, \ldots, 3m\}$,

(4.2)
$$\boldsymbol{P}(u,v) = \begin{cases} \mathbb{I}(|u-v| > t'), \text{ if } (l-1)m+1 \leq u \leq v \leq lm, \ l = 1,2,3 \\ \mathbb{I}(|u-v(\text{mod } m)| > t'), \text{ if } 1 \leq u \leq m, \ m+1 \leq v \leq 2m \\ \mathbb{I}(|m-u-v(\text{mod } m)| > t'), \text{ if } m+1 \leq u \leq 2m, \ 2m+1 \leq v \leq 3m. \end{cases}$$

In figure 4 we provide a representation of $\boldsymbol{P}$ using equation (4.2) for three different choices of t' and with $m = 5$. Here the entries with $\boldsymbol{P}(u, v) = 1$ are shaded in black while those with $\boldsymbol{P}(u, v) = 0$ are in gray. For a sufficiently

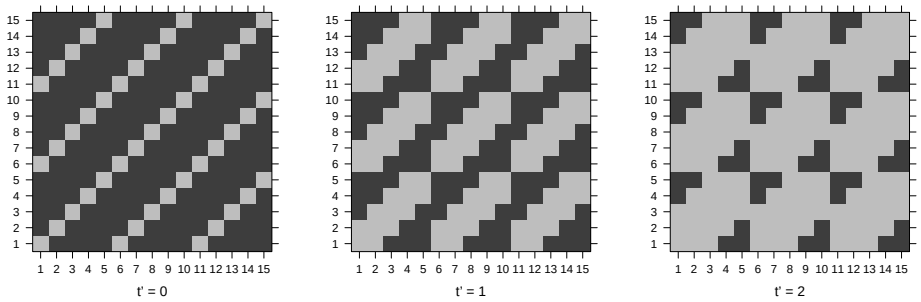


Fig 4: Representation of $\boldsymbol{P}$ using equation (4.2) for three different choices of t' and with $m = 5$. Here the entries with $\boldsymbol{P}(u, v) = 1$ are shaded in black while those with $\boldsymbol{P}(u, v) = 0$ are in gray.

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. We end this section with the remark that the maximization problem based on criterion (4.1) can be augmented with linear inequality constraints $\boldsymbol{A}^{(s)}\boldsymbol{\Gamma}^{(s)} \leq \boldsymbol{a}^{(s)}$ that may incorporate domain expertise and impose monotonicity, sign or other structural constraints on the components of $\boldsymbol{\Gamma}^{(s)}$.

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 (4.1) be denoted by $\ell_n^{pen}(\boldsymbol{\Theta})$ where

(4.3)
$$\ell_n^{pen}(\boldsymbol{\Theta}) = \ell_n(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}} w_{sr} \Big(|\Gamma_{sr}| + d_{sr} \boldsymbol{\Sigma}_{rr}^{(s)} \mathbb{I}\{r \in \mathcal{I}_c\} \Big) - n\lambda_2 \|\boldsymbol{P} * \boldsymbol{\Lambda}\|_1.$$

Denote $\boldsymbol{\Theta}_0 = (\boldsymbol{\Gamma}_0, \mathrm{vec}(\boldsymbol{\Sigma}_0), \mathrm{vec}(\boldsymbol{\Lambda}_0))$ 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 $\boldsymbol{\Theta}_2 = \mathbf{0}$. Similarly, for a local maximizer $\widehat{\boldsymbol{\Theta}}_n$ of equation (4.1), 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 $\lambda_{\mathsf{min}}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0))$ and $\lambda_{\mathsf{max}}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0))$ 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 (4.1) 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 $0 < c_1 < \lambda_{\mathsf{min}}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0)) < \lambda_{\mathsf{max}}(\boldsymbol{H}_n(\boldsymbol{\Theta}_0)) < 1/c_1 < \infty$ for some constant c_1.

(A2) For every $\epsilon > 0$, there exists a $\delta > 0$ such that for n large, $(1 - \epsilon)c_1 < \lambda_{\mathsf{min}}(\boldsymbol{H}_n(\boldsymbol{\Theta})) < \lambda_{\mathsf{max}}(\boldsymbol{H}_n(\boldsymbol{\Theta})) < (1 + \epsilon)/c_1$ 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 [9]. Conditions (A3) and (A4) are similar to assumptions (C5) and (C6) in [21]. 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 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 1, proved in Appendix A of the supplementary material, 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 (4.1). Here the suffix n will be implicit in our notations.

5.1. *Solving the maximization problem.* The marginal likelihood $\ell(\boldsymbol{\Theta})$ in equation (4.1) 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 [46], adaptive quadrature approximations [38], penalized quasi likelihood (PQL)[5] and EM algorithm [31], 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) [48]. 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

$\mathbb{Y}_i \coloneqq (\boldsymbol{\alpha}_i, \mathbf{Y}_i, \boldsymbol{\xi}_i)$ 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 (4.1) using data driven adaptive weights $(w_{sr}^{(t)}, d_{sr}^{(t)})$:

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta},\boldsymbol{\Sigma}\succ 0,\boldsymbol{\Lambda}\succ 0}{\arg\max}\quad \ell_{(t)}^{\mathsf{Q}}(\boldsymbol{\Theta}) - \\ n\lambda_1\sum_{s=1}^{3}\sum_{r\in\mathcal{I}} w_{sr}^{(t)}\Big(|\Gamma_{sr}| + d_{sr}^{(t)}\ \boldsymbol{\Sigma}_{rr}^{(s)}\mathbb{I}\{r\in\mathcal{I}_c\}\Big) - n\lambda_2\|\boldsymbol{P}*\boldsymbol{\Lambda}\|_1. \tag{5.1}$$

5.2. *Data Driven Adaptive Weights.* The data driven adaptive weights $(w_{sr}^{(t)}, d_{sr}^{(t)})$ are updated at the end of every iteration of the MCEM and their construction is designed to maintain the hierarchy in selecting the fixed and random effects (see [8, 51, 30] for details on these kind of approaches). In what follows, we use the approach described in Banerjee et. al (2020) [1] to construct these weights. Let $\boldsymbol{\Theta}^{(t)}$ denote the solution to the maximization problem in equation (4.1) at iteration t. Then in equation (5.1) we set $w_{sr}^{(t)} = \min\left(|\Gamma_{sr}^{(t)}|^{-\nu}, \epsilon_1^{-1}\right)$ and $d_{sr}^{(t)} = \min\left(|\boldsymbol{\Sigma}_{rr}^{(s,t)}|^{-\nu}|\Gamma_{sr}^{(t)}|^{-\nu}, \epsilon_2^{-1}\right)$ with $\nu = 2$. 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 $\epsilon_1 = 10^{-2}$ [8]. 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.

5.3. *Further details on the M-Step of the EM.* We now provide details around solving the maximization problem in equation (5.1). First note that the objective function in equation (5.1) is separable in $\{\mathbf{\Gamma}^{(s)} : s = 1, 2, 3\}$, $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$. To solve the three convex problems involving $\mathbf{\Gamma}^{(s)}$ we use a proximal gradient descent algorithm while to estimate $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$, which are non-convex problems, we adopt the majorization scheme of Bien and Tibshirani (2011) [2]. In particular, for estimating $\mathbf{\Gamma}^{(s)}$ at iteration t of the MCEM, the maximization problem in equation (5.1) can be represented as

$$\min_{\mathbf{\Gamma}^{(s)}} f_{(t)}(\mathbf{\Gamma}^{(s)}) + h_{(t)}(\mathbf{\Gamma}^{(s)}) \tag{5.2}$$

where $f_{(t)}(\mathbf{\Gamma}^{(s)})$ is convex and differentiable in $\mathbf{\Gamma}^{(s)}$, and

$$h_{(t)}(\mathbf{\Gamma}^{(s)}) = n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}} w_{sr}^{(t)} |\Gamma_{sr}|,$$

is convex but non-differentiable. Here $f_{(t)}(\mathbf{\Gamma}^{(s)})$ is the contribution of $\mathbf{\Gamma}^{(s)}$ to the negative of the Q function $\ell_{(t)}^{\mathsf{Q}}(\mathbf{\Theta})$ in iteration $(t+1)$. For instance when $s = 1$,

$$f_{(t)}(\mathbf{\Gamma}^{(1)}) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{d=1}^{D} \log p(\alpha_{ij} | \boldsymbol{b}_i^d, \boldsymbol{c}^d, \mathbf{\Gamma}^{(1)}) \omega_{id}^{(t)}.$$

where $\omega_{id}^{(t)}$ are the weights $p(\mathbb{Y}_i |\ \mathbf{\Theta}^{(t)}, \boldsymbol{b}_i^d, \boldsymbol{c}^d) / \sum_{d=1}^{D} p(\mathbb{Y}_i |\ \mathbf{\Theta}^{(t)}, \boldsymbol{b}_i^d, \boldsymbol{c}^d)$. To solve equation (5.2), we use the proximal gradient method that updates (5.2) in iteration $v = 1, 2, \ldots,$ as

$$\mathbf{\Gamma}_v^{(s)} = \mathsf{prox}_{t_v, h}\Big(\mathbf{\Gamma}_{v-1}^{(s)} - t_v \nabla f_{(t)}(\mathbf{\Gamma}_{v-1}^{(s)})\Big)$$

where

$$\mathsf{prox}_{t_v, h}(\boldsymbol{u}) = \arg\min_{\mathbf{\Gamma}^{(s)}} \Big(\frac{1}{2t_v} ||\mathbf{\Gamma}^{(s)} - \boldsymbol{u}||_2^2 + h(\mathbf{\Gamma}^{(s)})\Big), \tag{5.3}$$

$\boldsymbol{u} = \mathbf{\Gamma}_{v-1}^{(s)} - t_v \nabla f_{(t)}(\mathbf{\Gamma}_{v-1}^{(s)})$ and $\nabla f_{(t)}(\mathbf{\Gamma}_{v-1}^{(s)})$ is the derivative of $f_{(t)}(\mathbf{\Gamma}^{(s)})$ evaluated at $\mathbf{\Gamma}_{v-1}^{(s)}$. We solve equation (5.3) in CVXR [14] where the step size $t_v > 0$ is determined by backtracking line search.

In iteration (t) of the MCEM the optimization problem for estimating $\mathbf{\Sigma}$ is of the form

$$\min_{\mathbf{\Sigma} \succ 0} \log |\mathbf{\Sigma}| + \text{trace}(\boldsymbol{Q}^{(t)} \mathbf{\Sigma}^{-1}) + 2\lambda_1 \|\boldsymbol{V}^{(t)} * \mathbf{\Sigma}\|_1, \tag{5.4}$$

where $\boldsymbol{Q}^{(t)} = n^{-1}\sum_{i=1}^{n}\sum_{d=1}^{D}\boldsymbol{b}_i^d\boldsymbol{b}_i^{dT}\omega_{id}^{(t)}$ and $\boldsymbol{V}^{(t)} = \mathsf{diag}(d_{sr}^{(t)} : 1 \le s \le 3, 1 \le r \le p_c)$. The objective function in equation (5.4) is the sum of a convex function and a concave function $h(\boldsymbol{\Sigma}) = \log|\boldsymbol{\Sigma}|$. We use a majorization scheme that replaces $h(\boldsymbol{\Sigma})$ with its tangent $g(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_0) = \log|\boldsymbol{\Sigma}_0| + \mathrm{trace}\{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0)\}$. Equation (5.4) is then approximately solved using an iterative scheme that solves the following convex problem in iteration $v = 1, 2, \ldots,$

$$\hat{\boldsymbol{\Sigma}}^{(v)} = \underset{\boldsymbol{\Sigma}\succ 0}{\arg\min}\,\mathrm{trace}\{(\hat{\boldsymbol{\Sigma}}^{(v-1)})^{-1}\boldsymbol{\Sigma}\} + \mathrm{trace}(\boldsymbol{Q}^{(t)}\boldsymbol{\Sigma}^{-1}) + 2\lambda_1\|\boldsymbol{V}^{(t)} * \boldsymbol{\Sigma}\|_1. \tag{5.5}$$

We use the R-package `spcov` [2] that implements this iterative scheme involving (5.5) using a generalized gradient descent algorithm and initialize this algorithm with $\hat{\boldsymbol{\Sigma}}^{(0)} = \boldsymbol{Q}^{(0)}$. The aforementioned approach is also used to estimate $\boldsymbol{\Lambda}$ with $\boldsymbol{Q}^{(t)} = (nK)^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{d=1}^{D}\boldsymbol{c}_k^d\boldsymbol{c}_k^{dT}w_{id}^{(t)}$, $\boldsymbol{V}^{(t)} = \boldsymbol{P}$ and λ_2 in place of λ_1 in equation (5.4).

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. This data hold 20 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, covariates that capture the focal player's interaction with her friends, the in-game activities of those friends, and covariates that are related to the focal player's interaction with her guild. Out of these 20 player specific predictors, 17 predictors are treated as candidate composite effect predictors. The description of these predictors are provided in Table 1 of Appendix C in the supplementary material. Along with the player information, the data also hold 5 time varying guild characteristics for $K = 50$ guilds. We treat these guild characteristics as potential fixed effects with no corresponding random effect counterparts. 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 (5.1) 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 $(s, s') \in \{1, 2, 3\}$, $\text{Cov}(c_{kj}^{(s)}, c_{kj'}^{(s')}) = 0$ 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 $\mathsf{BIC}_\lambda = -2\ell^{\mathsf{Q}}(\widehat{\mathbf{\Theta}}) + \log(n)\text{dim}(\widehat{\mathbf{\Theta}})$ [3, 28, 21] and $\text{dim}(\widehat{\mathbf{\Theta}})$ is the number of non-zero components in $\widehat{\mathbf{\Theta}}$.

The following GitHub repository holds the R code for reproducing all the analysis in this paper: https://github.com/trambakbanerjee/crejm-code.

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. 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 she 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 [19], which has suggested that a larger guild size usually reduces a guild members' satisfaction and dilutes their social identity, thus, reducing her 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 1 in Appendix C of the supplementary material 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 her 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 her 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 her loyalty to the game [26], while from the login indicator model an increase in the guild size reduces the player's satisfaction and her 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 her 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 [36], 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`) her likelihood for future purchase is lower.

We now discuss the estimated covariance matrix $\hat{\mathbf{\Sigma}}$ of the player specific random effects. In figure 5 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 5 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`

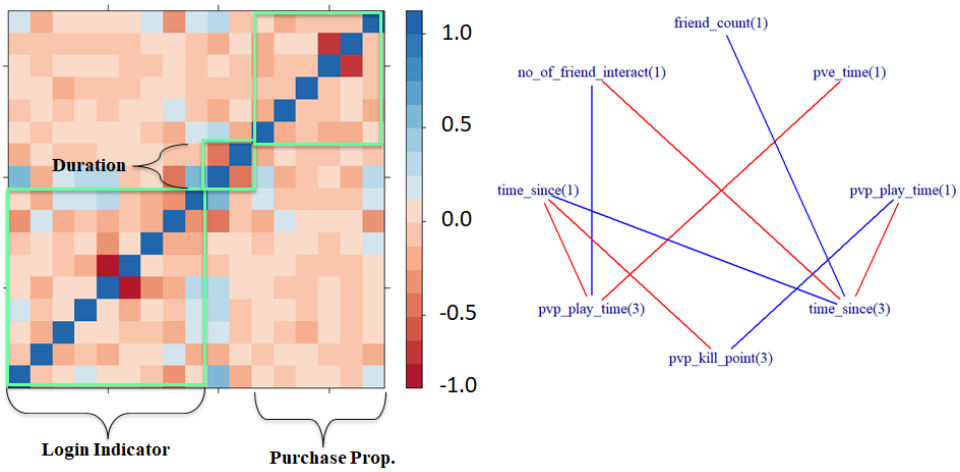


Fig 5: Left: Heatmap of the 17×17 correlation matrix obtained from $\hat{\boldsymbol{\Sigma}}$. On the horizontal axis are the selected composite effects for the three sub-models: Login Indicator, Duration and Purchase Propensity. The horizontal axis begins with the intercept from the Login Indicator model and ends with `time_since` from the Purchase Propensity model. Right: A network that demonstrate several cross correlations across the models. Blue lines represent positive correlations and red lines represent negative correlations. The model numbers are inside the parenthesis next to the predictor names.

(days since last login) in the Purchase Propensity model. Similarly the random effects associated with `pvp_kill_point(3)` and `pvp_play_time(2)` 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 discuss 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 [43] for variable selection. In particular, Bench-

mark 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 [22] 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 [5] 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) [42] and section 3 of Rizopoulos (2011) [41], and estimate the expected time j responses given the observed responses until time $j-1$ (details provided in Appendix B of the supplementary material).

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 (6.1) 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

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 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 $\mathrm{PE}_j^{\mathcal{M}}$ at time $j = 17, \ldots, 30$ as

$$\mathrm{PE}_j^{\mathcal{M}}(\mathbb{Y}^*, \widehat{\mathbb{Y}}^*) = \sum_{i=1}^{n} \left| \log Y_{ij}^* - \log \widehat{Y}_{ij}^* \right| \tag{6.1}$$

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 $\mathrm{PE}_j^{\mathcal{M}}$ 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.

6.3. *Time varying guild random effects.* In this section we discuss the estimates $\hat{c}_{jk}^{(s)}$ of the time varying guild random effects $c_{jk}^{(s)}$ that constitute a critical component of the CREJM framework. Recall that these guild random effects allow the players to be nested in guilds (see equations (3.4), (3.5) and (3.7)) and incorporate (1) the dynamic effect of a guild on its member's playing behavior, such as her duration of play, and (2) correlated playing behavior for players that are part of the same guild. Figure 6 presents the estimated guild random effects $\hat{c}_{jk}^{(s)}$ for each of the $K = 50$ guilds that were used for dynamically predicting the three responses over the next 14 days ($j = 17, \ldots, 30$) in Section 6.2 (see Appendix B of the supplementary material for details on estimating $c_{jk}^{(s)}$). These figures plot the temporal evolution of $\hat{c}_{jk}^{(s)}$ and highlight, in particular, the trajectory for three randomly chosen guilds: $4, 13$ and 42. It is interesting to note that in figure 6, $\hat{c}_{jk}^{(s)}$ for the Duration of Play model (center panel) exhibits an overall increasing trend over time while guilds 4 and 13 exhibit substantially different trajectories for

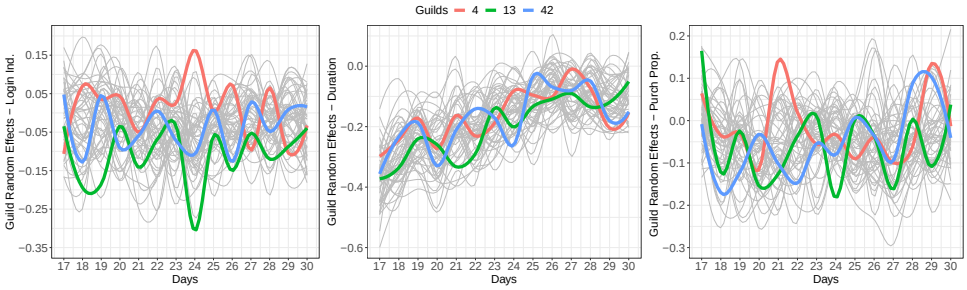


Fig 6: Temporal evolution of the estimated guild random effects $\hat{c}_{jk}^{(s)}$ where $s = 1$ for the Login Ind. model (left), $s = 2$ for the Duration of Play model (center) and $s = 3$ for the Purchase propensity model (right). For each of the 50 guilds, these estimated guild random effects were used for dynamically predicting the three responses over the next 14 days ($j = 17, \ldots, 30$) in Section 6.2. The highlighted trajectories represent $\hat{c}_{jk}$ for $k \in \{4, 13, 42\}$.

the login indicator (left panel) and purchase propensity (right panel) models. This dynamic nature of the estimated guild random effects translates into superior player specific predictions of future purchase propensities and duration of play under the CREJM framework as seen in table 2 of Section 6.2. Moreover, $\hat{c}_{jk}^{(s)}$ form a key component for predicting the temporal trajectories of player correlations within each guild and with respect to each of the three responses. These correlations are predicted in a similar fashion as discussed in Section 6.2 and Appendix B of the supplementary material, where for two players $(i, i^{'})$ that belong to guild k at time $j-1$, their correlation in duration of play at time j is predicted conditional on the estimated parameters and the observed responses until time $j-1$.

Figure 7 presents 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 $k = 1, \ldots, 50$. From the three panels in figure 7 we make several remarks.

- First, players who are part of a guild at any time j are correlated in terms of their probability of future login (left panel), the amount of time that they will spend in the game (center panel) and their probability of making future purchases (right panel). This is a direct consequence of the CREJM framework that incorporates time varying guild random effects into the joint modeling framework and allows for the possibility that players in guild k at time j may potentially exhibit correlated playing behavior.

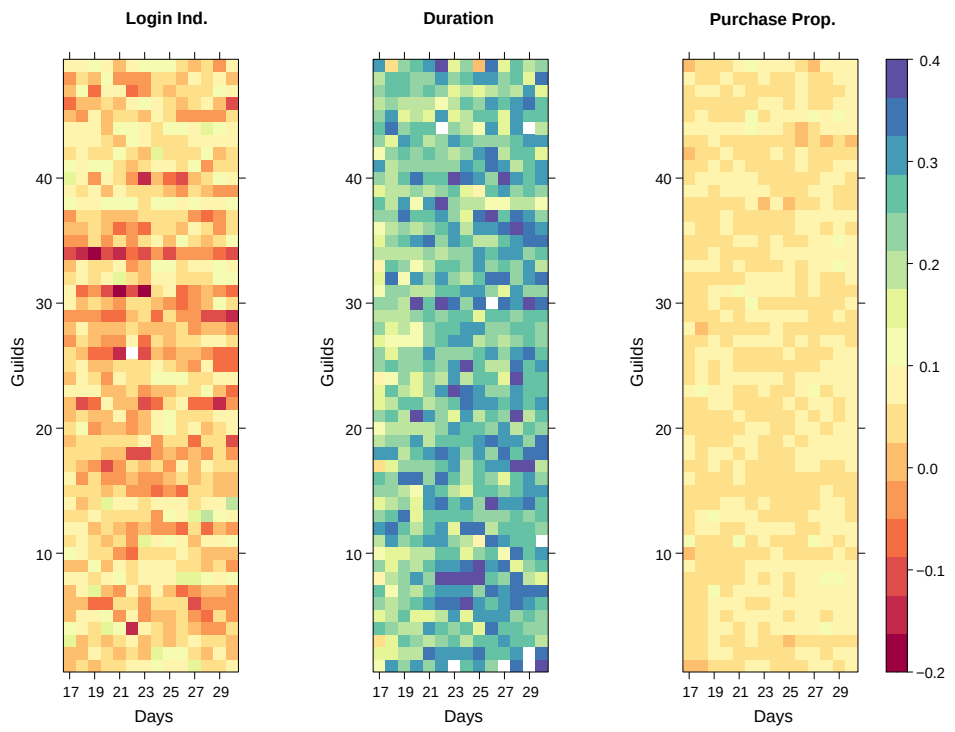


Fig 7: Mean player correlation for the three responses in 50 guilds across 14 days ($j = 17, \ldots, 30$). Left: Probability of Login, Center: Duration of Play and Right: Probability of Purchase. Details of the estimation provided in Appendix B of the supplementary material.

- Second, from the center panel of figure 7, the magnitude of correlations between the player duration of play are relatively larger than those estimated for the probabilities of login and purchase. This indicates that for a player her future purchase and login decisions are not as influenced by other members of the guild as her in-game time is. This is not surprising for purchases are rare in our data and login decisions may potentially be influenced by factors such as time zone differences. However, conditional on login, members of a guild play the game together as part of a team and thus spend similar amounts of in-game time.
- Third, 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. For instance, the 50×14 matrix of mean

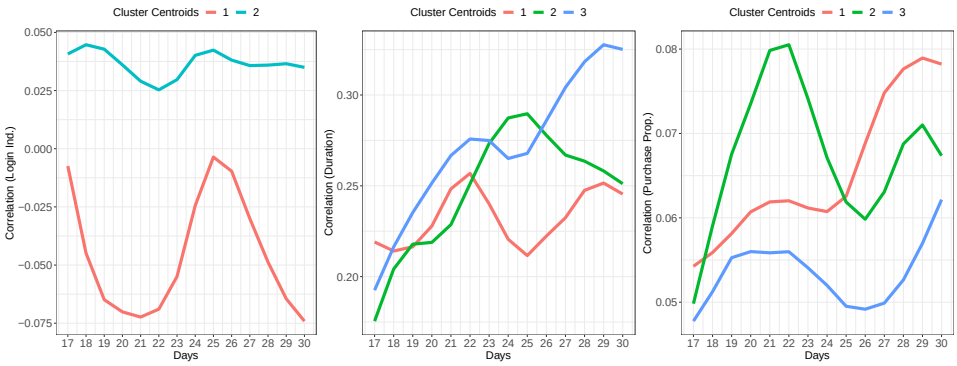


Fig 8: Functional cluster analysis of the mean guild correlations in Login Indicator (left), Duration of Play (center) and Purchase Propensity (right) models over 14 days ($j = 17, \ldots, 30$). The plot presents three cluster centroids. The sizes of these clusters are as follows: Login Indicator: $(7, 43)$; Activity Time: $(15, 19, 16)$; Purchase Propensity: $(21, 13, 16)$.

guild correlations of Duration of Play (figure 7 center panel) has been grouped into three clusters in figure 8 center panel. Similarly, the left and right panels of figure 8 show the cluster centroids when the corresponding matrices of mean guild correlations of Login Indicator and Purchase Propensity are grouped into two and three clusters, respectively. To determine these clusters, we use the `R` package `fda.usc` to cluster the rows of the 50×14 matrix of mean guild correlations using functional K-means clustering. The cluster centroids in figure 8 segment the 50 guilds into groups which demonstrate distinct correlation profiles in terms of their magnitudes and temporal trajectories. For Duration of Play, Cluster 3 holds 16 of the 50 guilds and exhibits an overall increasing trend over time except for days $23, 24$ and 25. Cluster 1, on the other hand, exhibits some of the smallest magnitude of correlations except for the first 5 days when it demonstrates relatively higher correlations than Cluster 2. For game managers and marketers, the guilds in Cluster 1 are of significance as players in these guilds are relatively less engaged with other guild members as far as their predicted duration of play is concerned. Future promotional and retention strategies may be developed to increase player engagement in the guilds represented in Cluster 1, while for the guilds in Cluster 3 promotional strategies may include loyalty rewards that further encourage player engagement in these guilds. For Login Indicator, there

are two clusters with Cluster 2 holding 43 of the 50 guilds and exhibits a relatively stable profile over time when compared to Cluster 1. For Purchase Propensity, Cluster 1 exhibits an overall increasing trend and has a similar trajectory to that of Cluster 3 from the Duration model. Clusters 2 and 3, on the other hand, show a decreasing trend between days 22 to 26. For platform managers, the temporal profiles of Clusters 1, 2 and 3 offer marketing insights on how to optimize price promotion across these clusters and such differential emphasis across these segments can increase the efficiency of future marketing campaigns.

7. Discussion. We propose a GLMM based joint modeling framework CREJM that provides a unified approach for jointly modeling and predicting a player's daily duration of play and her purchase propensity in MMORPGs. The key features of our framework that distinguish it from existing approaches is that (1) CREJM relies on a system of Cross Classified Random Effect Models that allow the players to be nested in guilds, thus accounting for the fact that players belonging to the same guild have correlated responses, and (2) CREJM incorporates time varying idiosyncratic guild random effects that capture the dynamic influence of the guild on its member's playing behavior. On a large scale data from a popular MMORPG, CREJM conducts hierarchical variable selection of the fixed and random effects and produces models with interpretable composite effects. We exhibit the superior performance of CREJM in dynamically predicting player responses. These predictions provide valuable insights to the game managers for developing personalized promotional strategies. Moreover, we use the estimates of the time varying guild random effects to generate predictions of the temporal trajectories of player correlations for the three responses within each guild. These correlation profiles have substantial business implications for platform monetization and enhancing the effectiveness of existing promotional and reward policies.

Our joint modeling framework can be used in a variety of applications that need analyzing multiple longitudinal outcomes wherein the individual subjects, such as patients or firms, are crossed with a dynamically evolving group such as hospitals or firm size, respectively. In this article, we have developed the CREJM framework for analyzing the dynamic effect of guilds on player responses in MMORPGs. Our current framework can, in principle, include additional time varying random intercepts to model the dynamic effect of multiple groups with which the players may be crossed, such as friendship networks or teams. However, variable selection and estimation in such multiple membership mixed models is challenging and we envision

pursuing this extension as part of future research.

References.

[1] Banerjee, T., G. Mukherjee, S. Dutta, and P. Ghosh (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*(530), 538–554.

[2] Bien, J. and R. J. Tibshirani (2011). Sparse estimation of a covariance matrix. *Biometrika 98*(4), 807–820.

[3] Bondell, H. D., A. Krishna, and S. K. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics 66*(4), 1069–1077.

[4] Borbora, Z., J. Srivastava, K.-W. Hsu, and D. Williams (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*, pp. 157–164. IEEE.

[5] Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*(421), 9–25.

[6] Cafri, G. and J. Fan (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*(1), 312–319.

[7] Cafri, G., D. Hedeker, and G. A. Aarons (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological methods 20*(4), 407.

[8] Candes, E. J., M. B. Wakin, and S. P. Boyd (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications 14*(5-6), 877–905.

[9] Chen, J. and Z. Chen (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 555–574.

[10] 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.

[11] 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.

[12] DFCIntelligence (2020). Global video game consumer segmentation. Available at https://www.dfcint.com/product/video-game-consumer-segmentation-2/.

[13] Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *Annals of statistics 40*(4), 2043.

[14] Fu, A., B. Narasimhan, and S. Boyd (2017). Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*.

[15] Gao, K. (2017). *Scalable Estimation and Inference for Massive Linear Mixed Models With Crossed Random Effects*. Ph. D. thesis, Stanford University.

[16] Gao, K., A. Owen, et al. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics 11*(1), 1235–1296.

[17] Gao, K. and A. B. Owen (2020). Estimation and inference for very large linear mixed effects models. *Statistica Sinica, arXiv:1610.08088*.

[18] Ghosh, S., T. Hastie, and A. B. Owen (2020). Backfitting for large scale crossed random effects regressions. *arXiv preprint arXiv:2007.10612*.

[19] Hackman, J. R. and N. Vidmar (1970). Effects of size and task type on group performance and member reactions. *Sociometry*, 37–54.

[20] Huang, Y., S. Jasin, and P. Manchanda (2019). level up: Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research 30*(3), 927–947.

[21] Hui, F. K., S. Müller, and A. Welsh (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica 27*(2).

[22] Hui, F. K., S. Müller, and A. Welsh (2017b). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association 112*(519), 1323–1333.

[23] Hui, F. K., S. Müller, and A. Welsh (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association 113*(524), 1759–1769.

[24] Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011). Fixed and random effects selection in mixed effects models. *Biometrics 67*(2), 495–503.

[25] Jin, W. and Y. Sun (2015). Understanding the antecedents of virtual product purchase in mmorpg: An integrative perspective of social presence and user engagement. In *PACIS*, pp. 191.

[26] Kang, J., I. Ko, and Y. Ko (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*, pp. 1–9. IEEE.

[27] Kumar, V. (2014). Making "freemium" work. *Harvard business review 92*(5), 27–29.

[28] Lin, B., Z. Pang, and J. Jiang (2013). Fixed and random effects selection by reml and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics 22*(2), 341–355.

[29] Liu, P., T. Chan, and H. Che (2020). The pursuit of leadership in a multiplayer online role-playing game and its effect on player spending. available from https://www.scu.edu/business/marketing/faculty/pliu/.

[30] Lu, C., Z. Lin, and S. Yan (2015). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing 24*(2), 646–654.

[31] McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association 92*(437), 162–170.

[32] NPD (2020). More people are gaming in the u.s., and theyre doing so across more platforms. *Press Release: July 21, 2020*. Available at https://www.npd.com/wps/portal/npd/us/news/press-releases/2020/more-people-are-gaming-in-the-us/.

[33] NPD (2021). Q4 2020 games market dynamics: U.s. *Press Release: February 1, 2021*. Available at https://www.npd.com/wps/portal/npd/us/news/press-releases/.

[34] Pan, J. and C. Huang (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing 24*(5), 725–738.

[35] Papaspiliopoulos, O., G. O. Roberts, and G. Zanella (2020). Scalable inference for crossed random effects models. *Biometrika 107*(1), 25–40.

[36] Park, E., R. Rishika, R. Janakiraman, M. B. Houston, and B. Yoo (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing 82*(1), 93–114.

[37] Peng, H. and Y. Lu (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis 109*, 109–129.

[38] Rabe-Hesketh, S., A. Skrondal, A. Pickles, et al. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal 2*(1), 1–21.

[39] 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*(4), 321–349.

[40] Raudenbush, S. W. and A. S. Bryk (2002). *Hierarchical linear models: Applications*

and data analysis methods, Volume 1. Sage.
[41] Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics 67*(3), 819–829.
[42] Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.
[43] Schelldorfer, J., L. Meier, and P. Bühlmann (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Journal of Computational and Graphical Statistics 23*(2), 460–477.
[44] Steinkuehler, C. A. and D. Williams (2006). Where everybody knows your (screen) name: Online games as third places. *Journal of computer-mediated communication 11*(4), 885–909.
[45] Terlutter, R. and M. L. Capella (2013). The gamification of advertising: analysis and research directions of in-game advertising, advergames, and advertising in social network games. *Journal of advertising 42*(2-3), 95–112.
[46] Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association 81*(393), 82–86.
[47] Verizon (2020). Verizons covid-19 network reliability report. *May 21 update*. Available at https://www.verizon.com/about/news/how-americans-are-spending-their-time-temporary-new-normal.
[48] Wei, G. C. and M. A. Tanner (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85*(411), 699–704.
[49] Wei, Y., W. Zhang, S. Yang, and X. Chen (2019). Online communities and social network structure. *Available at SSRN 3420525*.
[50] Zhang, C., C. W. Phang, Q. Wu, and X. Luo (2017). Nonlinear effects of social connections and interactions on individual goal attainment and spending: Evidences from online gaming markets. *Journal of Marketing 81*(6), 132–155.
[51] Zhao, Y.-B. and M. Kočvara (2015). A new computational method for the sparsest solutions to systems of linear equations. *SIAM Journal on Optimization 25*(2), 1110–1134.

Analytics, Information and Operations Management
School of Business
University of Kansas
Lawrence, KS 66045
E-mail: trambak@ku.edu

Department of Marketing
Santa Clara University
Santa Clara, CA 95053
E-mail: pliu2@scu.edu

Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
E-mail: gmukherj@marshall.usc.edu

Department of Marketing
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
E-mail: shantanu@marshall.usc.edu

Department of Marketing
University of California, Riverside
Riverside, CA 92521
E-mail: chehai@ucr.edu

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

By Trambak Banerjee[‡], Peng Liu[§], Gourab Mukherjee[†,¶] Shantanu Dutta[¶] and Hai Che[‖]

University of Kansas[‡], Santa Clara University[§], University of Southern California[¶] and University of California, Riverside[‖]

APPENDIX A: PROOF OF THEOREM 1

We can rewrite $\ell_n^{pen}(\boldsymbol{\Theta})$ in equation (4.3) as

$$\ell_n^{pen}(\boldsymbol{\Theta}) = \ell_n(\boldsymbol{\Theta}) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}_f \cup \mathcal{I}_g} w_{sr} |\Gamma_{sr}| - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}_c} w_{sr} (|\Gamma_{sr}| + d_{sr} \boldsymbol{\Sigma}_{rr}^{(s)}) - n\lambda_2 \sum_{i=1}^{3m} \sum_{j=1}^{3m} \boldsymbol{P}(i,j) \boldsymbol{\Lambda}_{ij}.$$

Let $a_n = \sqrt{p/n}$ and $D_n(\boldsymbol{u}) = \ell_n^{pen}(\boldsymbol{\Theta}_0 + a_n \boldsymbol{u}) - \ell_n^{pen}(\boldsymbol{\Theta}_0)$ where $\boldsymbol{u}$ is such that $\boldsymbol{\Theta}_0 + a_n \boldsymbol{u} \in \mathcal{F}_n$. We will first show that given any $\epsilon > 0$, there exists a constant M such that for n large

$$\mathbb{P}\Big(\sup_{\boldsymbol{u}: \|\boldsymbol{u}\|_2 = M, \boldsymbol{\Theta}_0 + a_n \boldsymbol{u} \in \mathcal{F}_n} D_n(\boldsymbol{u}) < 0 \Big) \geq 1 - \epsilon. \tag{A.1}$$

When the above holds, it implies that there exists a local maximizer $\widehat{\boldsymbol{\Theta}}_n$ in $\{\boldsymbol{\Theta}_0 + a_n \boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq M,\ \boldsymbol{\Theta}_0 + a_n \boldsymbol{u} \in \mathcal{F}_n\}$ such that $\|\widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}_0\|_2 = O_p(a_n)$ (see [1]). To prove equation (A.1) we begin by noting that for any $r \in \mathcal{I}_f \cup \mathcal{I}_g$, $\Gamma_{0,sr} = 0$ implies $|\Gamma_{0,sr} + a_n u_{sr}| - |\Gamma_{0,sr}| \geq 0$. Similarly, for any $r \in \mathcal{I}_c$, $\Gamma_{0,sr} = \boldsymbol{\Sigma}_{0,rr}^{(s)} = 0$ implies $\{(|\Gamma_{0,sr} + a_n u_{sr}| + d_{sr} |\boldsymbol{\Sigma}_{0,rr}^{(s)} + a_n v_{1,rr}^{(s)}|) - (|\Gamma_{0,sr}| + d_{sr} |\boldsymbol{\Sigma}_{0,rr}^{(s)}|)\} \geq 0$. Finally, for any $(i,j) \in \{1, \ldots, 3m\}$, $\boldsymbol{\Lambda}_{0,ij} = 0$ implies $|\boldsymbol{\Lambda}_{0,ij} + a_n v_{2,ij}| - |\boldsymbol{\Lambda}_{0,ij}| \geq 0$. So we have,

$$\begin{aligned}
D_n(\boldsymbol{u}) =& \ell_n^{pen}(\boldsymbol{\Theta}_0 + a_n \boldsymbol{u}) - \ell_n^{pen}(\boldsymbol{\Theta}_0) \\
&\leq \ell_n(\boldsymbol{\Theta}_0 + a_n \boldsymbol{u}) - \ell_n(\boldsymbol{\Theta}_0) - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}_{0f} \cup \mathcal{I}_{0g}} w_{sr} \Big(|\Gamma_{0,sr} + a_n u_{sr}| - |\Gamma_{0,sr}| \Big) \\
&\quad - n\lambda_1 \sum_{s=1}^{3} \sum_{r \in \mathcal{I}_{0c}} w_{sr} \Big\{ \Big(|\Gamma_{0,sr} + a_n u_{sr}| + d_{sr} |\boldsymbol{\Sigma}_{0,rr}^{(s)} + a_n v_{1,rr}^{(s)}| \Big) - \Big(|\Gamma_{0,sr}| + d_{sr} |\boldsymbol{\Sigma}_{0,rr}^{(s)}| \Big) \Big\} \\
&\coloneqq (I) - (II) - (III),
\end{aligned}$$

where $\mathcal{I}_{0f}, \mathcal{I}_{0g}$ and $\mathcal{I}_{0c}$ are, respectively, the true non-zero player specific fixed effects, guild specific fixed effects and composite effects. Note that in the above display the term involving $n\lambda_2$ is 0 since for any $(i,j) \in \{1, \ldots, 3m\}$, $\boldsymbol{\Lambda}_{0,ij} \neq 0$ implies $\boldsymbol{P}(i,j) = 0$.

[†]The research here was partially supported by NSF DMS-1811866.

[†]Corresponding author: gmukherj@marshall.usc.edu

For the term $(I) \coloneqq \ell_n(\boldsymbol{\Theta}_0 + a_n\boldsymbol{u}) - \ell_n(\boldsymbol{\Theta}_0)$, a Taylor's expansion gives

$$(I) \coloneqq a_n\boldsymbol{u}^T\nabla\ell_n(\boldsymbol{\Theta}_0) + \frac{1}{2}a_n^2\boldsymbol{u}^T\nabla^2\ell_n(\bar{\boldsymbol{\Theta}}_n)\boldsymbol{u} \;\coloneqq\; I_1 + I_2$$

where $\bar{\boldsymbol{\Theta}}_n$ lies on the line segment joining $\boldsymbol{\Theta}_0 + a_n\boldsymbol{u}$ and $\boldsymbol{\Theta}_0$. Now by Chebychev's inequality and assumption (A1), $|I_1| = |a_n\boldsymbol{u}^T\nabla\ell_n(\boldsymbol{\Theta}_0)| \leq a_n\|\nabla\ell_n(\boldsymbol{\Theta}_0)\|\|\boldsymbol{u}\| = O_p(a_n\sqrt{np})\|\boldsymbol{u}\| = O_p(na_n^2)\|\boldsymbol{u}\|$. Next we consider I_2 where

$$I_2 = -\frac{1}{2}na_n^2\boldsymbol{u}^T\Big(-\frac{1}{n}\nabla^2\ell_n(\bar{\boldsymbol{\Theta}}_n)\Big)\boldsymbol{u} = -\frac{1}{2}na_n^2\boldsymbol{u}^T H_n(\bar{\boldsymbol{\Theta}}_n)\boldsymbol{u}.$$

By Cauchy Schwartz inequality and assumptions (A1), (A2), $I_2 \leq -(1/2)na_n^2\|\boldsymbol{u}\|^2(1-\epsilon)c_1 < 0$.

For the term $(II) \coloneqq n\lambda_1\sum_{s=1}^3\sum_{r\in\mathcal{I}_{0f}\cup\mathcal{I}_{0g}} w_{sr}\Big(|\Gamma_{0,sr} + a_nu_{sr}| - |\Gamma_{0,sr}|\Big)$, we have

$$\begin{aligned}(II) \;&\coloneqq\; n\lambda_1 a_n\sum_{s=1}^3\sum_{r\in\mathcal{I}_{0f}\cup\mathcal{I}_{0g}} w_{sr}u_{sr}\text{sign}(\Gamma_{0,sr})\\ &\leq\; n\lambda_1 a_n\sum_{s=1}^3\sum_{r\in\mathcal{I}_{0f}\cup\mathcal{I}_{0g}} w_{sr}|u_{sr}|.\end{aligned}$$

From the display above and by Cauchy-Schwartz inequality, assumptions (A3), (A4) part (a), we have term (II) is $O_p(n\lambda_1 a_n\sqrt{\tilde{p}}) = o_p(na_n^2)$. Similarly for the term (III), Cauchy-Schwartz inequality, assumptions (A3) and (A4) part (a) imply $(III) = o_p(na_n^2)$. Thus for sufficiently large $\|\boldsymbol{u}\| = M$ such that $\boldsymbol{\Theta}_0 + a_n\boldsymbol{u} \in \mathcal{F}_n$, $D_n(\boldsymbol{u})$ is dominated by the term I_2 which is negative. This proves the statement in equation (A.1) and the first part of Theorem 1 on the estimation consistency of $\widehat{\boldsymbol{\Theta}}_n$.

We will now prove the selection consistency of $\widehat{\boldsymbol{\Theta}}_n$ following an approach similar to that of [2]. To do so we will consider four cases and in each case we will show that if the true parameter is 0 then the corresponding estimate must also be zero with probability tending to 1.

Case 1 - Suppose that for some $r \in \mathcal{I}_f \cup \mathcal{I}_g$ and $s \in \{1,2,3\}$, $\Gamma_{0,sr} = 0$ but $\widehat{\Gamma}_{sr} \neq 0$. Now by the KKT optimality conditions

$$0 = \frac{\partial\ell_n^{pen}(\boldsymbol{\Theta})}{\partial\Gamma_{sr}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} = \frac{\partial\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} - n\lambda_1 w_{sr}\text{sign}(\widehat{\Gamma}_{sr}). \tag{A.2}$$

Consider the first term on the right hand side of equation (A.2). A Taylor's expansion gives,

$$\begin{aligned}\frac{\partial\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} &= \frac{\partial\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}}\Big|_{\boldsymbol{\Theta}_0} + \sum_{k\in\boldsymbol{\Theta}}\frac{\partial^2\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}\partial\Theta_k}\Big|_{\bar{\boldsymbol{\Theta}}_n}(\widehat{\Theta}_{nk} - \Theta_{0k})\\ &\leq \frac{\partial\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}}\Big|_{\boldsymbol{\Theta}_0} + n\|\widehat{\boldsymbol{\Theta}}_n - \boldsymbol{\Theta}_0\|_2\Big\{\sum_{k\in\boldsymbol{\Theta}}\Big(\frac{1}{n}\frac{\partial^2\ell_n(\boldsymbol{\Theta})}{\partial\Gamma_{sr}\partial\Theta_k}\Big|_{\bar{\boldsymbol{\Theta}}_n}\Big)^2\Big\}^{1/2}\\ &\coloneqq I_1 + I_2,\end{aligned}$$

where the inequality in the second line of the display above follows from the Cauchy-Schwartz inequality. We know that $I_1 = O_p(\sqrt{np})$ and for n large, assumptions $(A1)$, $(A2)$ and the consistency

of $\widehat{\boldsymbol{\Theta}}_n$ imply $I_2 = O_p(na_n) = O_p(\sqrt{np})$. Thus, the first term on the right hand side of equation (A.2) is $O_p(\sqrt{np})$. For the second term, we use assumption (A3) with $\nu \geq 1$ to note that $n\lambda_1 w_{sr}/\sqrt{np} = O_p\{\lambda_1(n/p)^{(\nu+1)/2}\} \geq O_p\{\lambda_1(n/p)^{(\nu+3)/4}\}$ where $\lambda_1(n/p)^{(\nu+3)/4} \to \infty$ as $n \to \infty$ by assumption (A4) part (b). Combining these results, we have a contradiction since the right hand side of equation (A.2) is asymptotically dominated by the second term which, with probability tending to 1, cannot equal 0. Thus for all $r \in \mathcal{I}_f \cup \mathcal{I}_g$ with $\Gamma_{0,sr} = 0$, we have $\mathbb{P}(\widehat{\Gamma}_{sr} = 0) \to 1$ as $n \to \infty$.

Case 2 - Suppose that for some $r \in \mathcal{I}_c$ and $s \in \{1, 2, 3\}$, $\Gamma_{0,sr} = 0$ but $\widehat{\Gamma}_{sr} \neq 0$. Note that in this case since $r \in \mathcal{I}_c$, $\Gamma_{0,sr} = 0$ implies $\boldsymbol{\Sigma}^{(s)}_{0,rr} = 0$. The KKT optimality conditions in this scenario are again given by equation (A.2) and it follows from the preceding discussion on Case 1 that for all $r \in \mathcal{I}_c$ with $\Gamma_{0,sr} = 0$, we have $\mathbb{P}(\widehat{\Gamma}_{sr} = 0) \to 1$ as $n \to \infty$.

Case 3 - Suppose that for some $r \in \mathcal{I}_c$ and $s \in \{1, 2, 3\}$, $\boldsymbol{\Sigma}^{(s)}_{0,rr} = 0$ but $\widehat{\boldsymbol{\Sigma}}^{(s)}_{rr} \neq 0$. From the KKT optimality conditions

$$0 = \frac{\partial \ell_n^{pen}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Sigma}^{(s)}_{rr}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} = \frac{\partial \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Sigma}^{(s)}_{rr}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} - n\lambda_1 w_{sr} d_{sr}. \tag{A.3}$$

Assumptions $(A1)$, $(A2)$ and the consistency of $\widehat{\boldsymbol{\Theta}}_n$ imply that the first term on the right hand side of equation (A.3) is $O_p(\sqrt{np})$. For the second term, we again use assumption (A3) with $\nu \geq 1$ to note that $n\lambda_1 w_{sr} d_{sr}/\sqrt{np} = O_p\{\lambda_1(n/p)^{\nu+1/2}\} \geq O_p\{\lambda_1(n/p)^{(\nu+3)/4}\}$ where $\lambda_1(n/p)^{(\nu+3)/4} \to \infty$ as $n \to \infty$ by assumption (A4) part (b). Combining these results, we have a contradiction since the right hand side of equation (A.3) is asymptotically dominated by the second term which, with probability tending to 1, cannot equal 0. Thus for all $r \in \mathcal{I}_c$ and $s \in \{1, 2, 3\}$, $\boldsymbol{\Sigma}^{(s)}_{0,rr} = 0$ implies $\mathbb{P}(\widehat{\boldsymbol{\Sigma}}^{(s)}_{rr} = 0) \to 1$ as $n \to \infty$.

Case 4 - Suppose that for some $(i, j) \in \{1, \ldots, 3m\}$ with $i \neq j$, $\boldsymbol{\Lambda}_{0,ij} = 0$ but $\widehat{\boldsymbol{\Lambda}}_{ij} \neq 0$. Since $\boldsymbol{\Lambda}_{0,ij} = 0$ we have $\boldsymbol{P}(i, j) = 1$. From the KKT optimality conditions

$$0 = \frac{\partial \ell_n^{pen}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{ij}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} = \frac{\partial \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Lambda}_{ij}}\Big|_{\widehat{\boldsymbol{\Theta}}_n} - n\lambda_2 \boldsymbol{P}(i, j). \tag{A.4}$$

Assumptions $(A1)$, $(A2)$ and the consistency of $\widehat{\boldsymbol{\Theta}}_n$ imply that the first term on the right hand side of equation (A.4) is $O_p(\sqrt{np})$. Also since $\lambda_2 > 0$ is fixed, the second term in equation (A.4) is $O_p(\sqrt{n/p})$ which cannot equal 0 with probability 1. This leads to a contradiction since the right hand side of equation (A.4) is asymptotically dominated by the second term. Thus for all $(i, j) \in \{1, \ldots, 3m\}$ with $i \neq j$, $\boldsymbol{\Lambda}_{0,ij} = 0$ implies $\mathbb{P}(\widehat{\boldsymbol{\Lambda}}_{ij} = 0) \to 1$ as $n \to \infty$.

The aforementioned four cases suffice to prove the desired result $\mathbb{P}(\widehat{\boldsymbol{\Theta}}_{n2} = \mathbf{0}) \to 1$ as $n \to \infty$.

APPENDIX B: PREDICTION EQUATIONS

We will first discuss the prediction problem of Section 6.2 where we are interested in predicting the time $u > t$ expected longitudinal outcomes of Login Indicator, Duration, and Purchase Propensity given the observed responses $\mathcal{Y}_i(t) = \{\alpha_{ij}, Y_{ij}, \xi_{ij} : 0 \leq j \leq t\}$ for player i. We consider the case of predicting $w_i(u \mid t) \coloneqq \mathbb{E}\{Y_{iu} \mid \mathcal{Y}_i(t); \boldsymbol{\Theta}\}$ as an example as the rest follow along similar lines. Let $\widehat{\alpha}_{iu}$ be the predicted Login Indicator for player i at time u conditional on $\mathcal{Y}_i(t)$. Then note that

$$\mathbb{E}\{Y_{iu} \mid \mathcal{Y}_i(t); \boldsymbol{\Theta}\} = \int \mathbb{E}\{Y_{iu} \mid \boldsymbol{b}_i, \boldsymbol{c}; \boldsymbol{\Theta}\} p(\boldsymbol{b}_i, \boldsymbol{c} \mid \mathcal{Y}_i(t); \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{b}_i \mathrm{d}\boldsymbol{c}.$$

From Section 7.2 of [3] an estimate of $w_i(u \mid t)$ is given by

$$\widehat{w}_i(u \mid t) = \begin{cases} 0, \text{ if } \widehat{\alpha}_{iu} = 0 \\ \exp\Big\{ \boldsymbol{x}_{it}^{(2)'} \widehat{\boldsymbol{\beta}}^{(2)} + \boldsymbol{z}_{it}^{(2)'} \widehat{\boldsymbol{b}}_i^{(2)} + \sum_{k=1}^{K} d_{itk} \Big(\widehat{c}_{tk}^{(2)} + \boldsymbol{g}_{tk}^{(2)'} \widehat{\boldsymbol{\gamma}}^{(2)} \Big) + \dfrac{\widehat{\sigma}^2}{2} \Big\}, \text{ otherwise} \end{cases}$$

where $\widehat{\boldsymbol{b}}_i = (\widehat{\boldsymbol{b}}_i^{(s)} : 1 \leq s \leq 3) = \arg\max_{\boldsymbol{b}} \log p(\boldsymbol{b} \mid \mathcal{Y}_i(t); \widehat{\boldsymbol{\Theta}})$ and, recalling from Sections 3.2 and 4 that $\boldsymbol{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K)$ and $\boldsymbol{c}_k = (\boldsymbol{c}_k^{(1)}, \boldsymbol{c}_k^{(2)}, \boldsymbol{c}_k^{(3)}) \overset{i.i.d}{\sim} N(\boldsymbol{0}, \boldsymbol{\Lambda})$, $\widehat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \log p(\boldsymbol{c} \mid \mathcal{Y}_i(t); \widehat{\boldsymbol{\Theta}})$.

We will now turn to the discussion in Section 6.3 and present our approach for predicting the temporal trajectories of player correlations within each guild and with respect to each of the three responses. Let $u = t+1$. For two players (i, i') that belong to the same guild at time t, we consider the case of predicting $\rho(Y_{iu}, Y_{i'u} | \mathcal{Y}_{it}, \mathcal{Y}_{i't}; \boldsymbol{\Theta})$ which is the Pearson's correlation coefficient between the Durations $(Y_{iu}, Y_{i'u})$ at time u given the observed responses $\{\mathcal{Y}_i(t), \mathcal{Y}_{i'}(t)\}$ until time t. Note that the preceding discussion can be used to estimate $w_{ii}(u|t) \coloneqq \mathbb{E}\{Y_{iu}^2 \mid \mathcal{Y}_i(t); \boldsymbol{\Theta}\}$ and $w_{i'i'}(u|t) \coloneqq \mathbb{E}\{Y_{i'u}^2 \mid \mathcal{Y}_{i'}(t); \boldsymbol{\Theta}\}$, and so we focus only on estimating $w_{ii'}(u|t) \coloneqq \mathbb{E}\{Y_{iu} Y_{i'u} \mid \mathcal{Y}_i(t), \mathcal{Y}_{i'}(t); \boldsymbol{\Theta}\}$. Now, we have

$$w_{ii'}(u|t) = \int \mathbb{E}\{Y_{iu} Y_{i'u} \mid \boldsymbol{b}_i, \boldsymbol{b}_{i'}, \boldsymbol{c}; \boldsymbol{\Theta}\} p(\boldsymbol{b}_i \mid \mathcal{Y}_i(t); \boldsymbol{\Theta}) p(\boldsymbol{b}_{i'} \mid \mathcal{Y}_{i'}(t); \boldsymbol{\Theta}) p(\boldsymbol{c} \mid \mathcal{Y}_i(t), \mathcal{Y}_{i'}(t); \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{b}_i \mathrm{d}\boldsymbol{b}_{i'} \mathrm{d}\boldsymbol{c},$$

where

$$\mathbb{E}\{Y_{iu} Y_{i'u} \mid \boldsymbol{b}_i, \boldsymbol{b}_{i'}, \boldsymbol{c}; \boldsymbol{\Theta}\} = \mathbb{E}\{Y_{iu} \mid \boldsymbol{b}_i, \boldsymbol{c}; \boldsymbol{\Theta}\} \mathbb{E}\{Y_{i'u} \mid \boldsymbol{b}_{i'}, \boldsymbol{c}; \boldsymbol{\Theta}\}.$$

An estimate of $w_{ii'}(u \mid t)$ is given by $\widehat{w}_{ii'}(u \mid t)$ so that $\widehat{w}_{ii'}(u \mid t) = 0$ if $\min(\widehat{\alpha}_{iu}, \widehat{\alpha}_{i'u}) = 0$ and

$$\widehat{w}_{ii'}(u \mid t) = \exp\Big\{ \Big(\boldsymbol{x}_{it}^{(2)'} + \boldsymbol{x}_{i't}^{(2)'} \Big) \widehat{\boldsymbol{\beta}}^{(2)} + \boldsymbol{z}_{it}^{(2)'} \widehat{\boldsymbol{b}}_i^{(2)} + \boldsymbol{z}_{i't}^{(2)'} \widehat{\boldsymbol{b}}_{i'}^{(2)} + 2 \Big(\widehat{c}_{tk}^{(2)} + \boldsymbol{g}_{tk}^{(2)'} \widehat{\boldsymbol{\gamma}}^{(2)} \Big) + \widehat{\sigma}^2 \Big\}, \text{ otherwise,}$$

where $\widehat{\boldsymbol{b}}_i = (\widehat{\boldsymbol{b}}_i^{(s)} : 1 \leq s \leq 3) = \arg\max_{\boldsymbol{b}} \log p(\boldsymbol{b} \mid \mathcal{Y}_i(t); \widehat{\boldsymbol{\Theta}})$ and $\widehat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \log p(\boldsymbol{c} \mid \mathcal{Y}_i(t), \mathcal{Y}_{i'}(t); \widehat{\boldsymbol{\Theta}})$.

APPENDIX C: DATA DESCRIPTION

We list the raw covariates along with their description that were available in the data in Table 1 below. The column 'FE/CE' indicates whether the covariate is a candidate fixed effect (FE) or a composite effect (CE). Table 2 presents a descriptive summary of the raw covariates.

TABLE 1
List of covariates and the three responses.

Type	FE/CE	Covariates	Description
Focal player's in-game characteristics	CE	level	players character level in the game
	CE	pvp_play_time	players daily gaming time of playing PVP mode in seconds
	CE	pvp_kill_point	kill points a player achieves by playing PVP mode
	CE	quest_count	no. of quest a player accomplished when PVE mode
	CE	mission_count	no. of missions a player accomplished when playing PVE mode
	CE	pve_time	players daily gaming time of playing PVE mode in seconds
	CE	no_of_game	daily total number of PVE game rounds a player played
Focal player's interaction with her friends and the in-game activities of these friends	CE	friend_count	no. of friends a player has
	CE	friend_mean_level	mean character level of a focal players all friends
	CE	no_of_friend_purch	no. of times of a focal players friends made purchases
	CE	total_friend_buy	monetary value of all purchases made by a focal players friends
	CE	no_of_friend_interact	no. of friends a player played with in teams during game sessions
	CE	game_round_play_with_friends	no. of game sessions a player played in team with her friends
Focal player's interaction with her guild	CE	guild_tenure	no. of days a player has been associated with a guild
	CE	no_of_guildmem_interact	no. of guild member a player played with in teams during game sessions
	CE	no_of_game_with_guildmem	no. of game sessions a player played with guild members
Guild characteristics	FE	guildmem_interact	no. of guild members played in teams
	FE	avg_game_with_guildmem	average no. of game sessions a guild member played as part of a team
	FE	guild_total_purch	monetary value of all purchases made by guild members
	FE	no_of_guildmem_purch	no. of times guild members made purchases
	FE	guildmem_count	no. of players associated with a guild
Other characteristics	FE	gender	dummy indicator for players virtual gender in the game
	FE	weekend	dummy indicator for a weekend
	FE	holiday	dummy indicator for Chinese New Year and Valentine's day
	CE	time_since	days since last login
	Sl No	Response	Description
	1	login_ind	whether active in a day (0 - No, 1 - Yes)
	2	duration	total time played in a day in minutes
	3	purch_ind	whether positive purchase from the player in a day (0 - No, 1 - Yes)

TABLE 2
Summary statistics of the covariates reporting % of 0, mean, the $25^{th}, 50^{th}, 75^{th}$, 95^{th} percentiles and the standard deviation of all players who logged-in ($\alpha_{ij} = 1$) across the $m = 30$ days. For the guild specific characteristics and time_since, however, the statistics are reported for all players and not just the ones who logged-in.

Type	Covariates	% of 0	Mean	25^{th}	50^{th}	75^{th}	95^{th}	Std.
Focal player's in-game characteristics	level	0.00	31.28	28.00	33.00	37.00	39.00	7.94
	pvp_play_time	99.78	8.85	0.00	0.00	0.00	0.00	252.02
	pvp_kill_point	99.78	7.66	0.00	0.00	0.00	0.00	248.17
	quest_count	71.29	0.97	0.00	0.00	1.00	5.00	2.79
	mission_count	69.49	1.15	0.00	0.00	1.00	6.00	2.95
	pve_time	19.49	4410.57	707.00	3254.00	6693.00	13313.00	4488.43
	number_of_game	19.46	5.46	1.00	4.00	8.00	17.00	5.92
Focal player's interaction with her friends and the in-game activities of these friends	friend_count	0.13	37.00	17.00	30.00	49.00	90.00	34.82
	friend_mean_level	0.00	30.25	27.57	31.56	34.50	37.41	6.16
	no_of_friend_purch	28.22	2.36	0.00	1.00	3.00	8.00	3.02
	total_friend_buy	28.56	219.17	0.00	74.95	249.30	910.00	436.06
	no_of_friend_interact	53.86	1.09	0.00	0.00	2.00	4.00	1.63
	game_round_play_with_friends	53.86	2.08	0.00	0.00	3.00	9.00	3.57
Focal player's interaction with her guild	guild_tenure	0.00	15.43	8.00	15.00	23.00	29.00	8.72
	no_of_guildmem_interact	57.84	1.23	0.00	0.00	2.00	6.00	2.06
	no_of_game_with_guildmem	57.84	1.62	0.00	0.00	2.00	8.00	3.03
Guild characteristics	guildmem_interact	0.13	25.78	18.00	25.00	33.00	46.00	11.59
	avg_game_with_guildmem	0.13	3.64	2.74	3.46	4.44	5.85	1.22
	guild_total_purch	14.77	213.07	27.43	109.00	253.88	805.58	333.85
	no_of_guildmem_purch	14.77	2.79	1.00	2.00	4.00	8.00	2.53
	guildmem_count	0.06	104.16	92.00	107.00	116.00	121.00	18.47
Other characteristics	gender	54.48	0.46	0.00	0.00	1.00	1.00	0.50
	weekend	66.72	0.33	0.00	0.00	1.00	1.00	0.47
	holiday	93.56	0.06	0.00	0.00	0.00	1.00	0.25
	time_since	59.68	2.93	0.00	0.00	3.00	17.00	5.85

REFERENCES

[1] Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*(3), 928–961.
[2] Hui, F. K., S. Müller, and A. Welsh (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica 27*(2).
[3] Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Analytics, Information and Operations Management
School of Business
University of Kansas
Lawrence, KS 66045
E-mail: trambak@ku.edu

Department of Marketing
Santa Clara University
Santa Clara, CA 95053
E-mail: pliu2@scu.edu

Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
E-mail: gmukherj@marshall.usc.edu

Department of Marketing
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
E-mail: shantanu@marshall.usc.edu

Department of Marketing
University of California, Riverside
Riverside, CA 92521
E-mail: chehai@ucr.edu