SMOGS: Social Network Metrics of Game Success

Fan Bu **Sonia Xu** **Katherine Heller** **Alexander Volfovsky**

Department of Statistical Science, Duke University

Abstract

In this paper we propose a novel metric of basketball game success, derived from a team's dynamic social network of game play. We combine ideas from random effects models for network links with taking a multi-resolution stochastic process approach to model passes between teammates. These passes can be viewed as directed dynamic relational links in a network. Multiplicative latent factors are introduced to study higher-order patterns in players' interactions that distinguish a successful game from a loss. Parameters are estimated using a Markov chain Monte Carlo sampler. Results in simulation experiments suggest that the sampling scheme is effective in recovering the parameters. We also apply the model to the first high-resolution optical tracking data set collected in college basketball games. The learned latent factors demonstrate significant differences between players' passing and receiving patterns in a loss, as opposed to a win. Our model is applicable to team sports other than basketball, as well as other time-varying network observations.

1 INTRODUCTION

Basketball is a sport played between two teams of five players, in which a game is won by seeing which team can score the most points from baskets in the time alloted. Starting from the beginning, a basketball game continuously evolves in time and space and is comprised of a constant flow of player movements, interactions, and decision making that contribute to the game's outcome. As a team sport, basketball requires collaboration of the players to successfully bring the ball to the basket, and such collaboration relies heavily on passing the ball between teammates. Understanding and evaluating the decisions made by players on whether to pass, when to pass, and whom to pass to is an ongoing challenge in sports analytics.

Passes between teammates can be modeled as interactions within a network, and a variety of previous methods have taken a network approach to studying passing sequences. Fewell et al. (2012) focus on exploratory network analysis to explain key facets of the game like key players and styles of team play, where the roles of different players on a team are explained through weighted graphs of passing frequencies. Gudmundsson and Horton (2017) calculate rebound probability with spatial coordinates to measure team and player performance in a graph theoretical framework. Xin et al. (2017) implement a continuous-time stochastic block model to cluster players based on passing networks.

Although basketball games were traditionally analyzed in a discretized manner based on "box score" statistics for forecasting (Hollinger, 2005) and player evaluation (Omidiran, 2011), the installment of optical tracking systems in professional basketball arenas in 2013 has allowed for more detailed statistical analyses. High-resolution spatial and temporal information has been leveraged to model the temporal sequences of games and characterize basketball strategies that were overlooked in low-resolution analyses. Miller et al. (2014) summarize player shot locations as low dimensional spatial bases by smoothing empirical shot locations through non-negative matrix factorization (NMF). The spatial bases for each player translate well in determining a player's position on the team. Pelechrinis and Papalexakis (2017) use tensor decomposition to create a weighted shot chart from a 12-dimensional representation of each player. Franks et al. (2015) characterize the spatial structure of defensive basketball play and quantitatively evaluate the guarding choices and movements of defending players. Cervone et al. (2016) develop a multi-resolution stochastic process model to calculate the expected points the offense will score in a possession conditioned on the evolution of the game up to a time point.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

In recent years, an NCAA Division I basketball team partnered with SportsVu to install optical tracking systems in their home stadium, collecting the first high-resolution spatio-temporal dataset of college-level basketball games (a detailed description of this dataset is given in section 4.2). This paper aspires to evaluate player interactions and develop metrics for game success that are applicable to (but not restricted to) collegiate basketball settings. We build on the work conducted by Cervone et al. (2016) by modeling passes from a network perspective and introducing multiplicative latent factors to study players' passing choices and preferences. These multiplicative effects provide a novel assessment of the efficacy and effectiveness of the passing game of a team.

Treating a pass from player i to player j at time t as a dynamic relational link from i to j, the observations of whether and between whom a pass is made in a basketball possession conditional on the spatio-temporal evolution of the possession up to time t are analogous to repetitive observations of relational ties on networks. In the last two decades, some authors have used non-additive random effects on top of fixed covariate effects to model nodal links on networks. Nowicki and Snijders (2001) assume that the probability of a link between two nodes depends on the shared membership in a collection of latent classes. The more general class of latent space models maps nodal characteristics onto an unobserved social space with ties depending on the similarity between actors within the latent space (Hoff et al., 2002; Hoff, 2005). This class of models has been extended to allow heterogeneous additive and multiplicative sender- and receiver-effects (Hoff, 2009; Hoff et al., 2013; Hoff, 2018) as well as to dynamic networks (Durante and Dunson, 2014; Sewell and Chen, 2015). This work extends the class of latent factor network models by modeling link occurrences as non-homogeneous Poisson processes on a dynamic network in the complex spatio-temporal setting of basketball games.

The novelty of this work is as follows:

- Our model is built upon the stochastic process model proposed by Cervone et al. (2016), but includes additional multiplicative latent factors from a network perspective that distinguish the effects of making a pass and receiving a pass, thus capturing additional information on pair-wise interactions among team players;
- Instead of modeling probabilities of binary network edges (Hoff, 2005, 2009), we model the intensity function of a non-homogeneous spatio-temporal Poisson process, while adjusting for game-related and player-specific covariates;
- We present the first analysis of high-resolution optical tracking data for college basketball, which differs from professional basketball (like NBA) in rules, court conditions, and player characteristics.

The remainder of the paper is organized as follows. In the next section we provide an overview of the key aspects of both the stochastic model by Cervone et al. (2016) and a latent factor social network model, in a basketball setting. In section 3, we present our novel latent factor stochastic passing model in detail and discuss our parameter estimation procedure. Section 4 presents experimental results on synthetic datasets and real optical basketball tracking data, and lastly, the conclusion is in section 5.

2 BACKGROUND

2.1 Multiresolution Stochastic Process Model

The model introduced by Cervone et al. (2016) begins with a coarsened view of a basketball possession. At any time point, the going-ons in a game fall into one of the three types of states: a possession state, a transition state, and an end state. The ball does not change hands in a possession state, and thus this state can be modeled by the micro movements of players. An end state, as suggested by the name, simply represents the end of the possession via points (0, 2, or 3) earned by the offense[1]. It is within a transition state that the dynamics of a basketball game changes qualitatively: a transition can be a pass, a shot attempt, or a turnover, after which either the ball carrier changes or the possession ends. Based on the assumption that, given the ending state of a transition, a future possession is conditionally independent of the history up to the beginning of that transition, modeling the occurrence and end state of a transition is essential to predict the outcome of a possession well.

Among the three transitions (pass, shot attempt, turnover), a shot attempt results in either a made shot or a failed shot, a turnover leads to a change of possession, but a pass has four potential outcomes corresponding to the four other teammates as potential receivers, which depend on various spatio-temporal factors. More specifically, assuming player i possesses the ball at time t, let $\theta_{i,j}(t)$ be the hazard for the occurrence of a pass to teammate j in $(t, t+\epsilon]$,

$$\theta_{i,j}(t) = \lim_{\epsilon \to 0+} \frac{\mathbb{P}(\{i \text{ passes to } j \text{ in } (t, t+\epsilon]\} | \mathcal{H}(t))}{\epsilon}, \tag{1}$$

[1] Possessions with fouls are not considered, so free throws (which result in 1 point) are not included.

where $\mathcal{H}(t)$ denotes the history of the game up to time t. Further assume the hazard is log-linear,

$$\log(\theta_{i,j}(t)) = W_{i,j}(t)^T \eta_{i,j} + \xi_{i,j}(\mathbf{s}_i(t)) + \zeta_{i,j}(\mathbf{s}_j(t)), \tag{2}$$

where $W_{i,j}(t)$ is a time-varying covariate vector, $\eta_{i,j}$ is the corresponding coefficient vector, and $\xi_{i,j}(\mathbf{s}_i(t))$ and $\zeta_{i,j}(\mathbf{s}_j(t))$ are functions that map the ball carrier's location and the potential receiver's location on the court to additive spatial effects (a detailed description follows in section 3).

2.2 Multiplicative Latent Factor Model

Observations on a social network can be characterized by an $n \times n$ matrix $\mathbf{Y} = \{y_{ij}\}$ where y_{ij} is a binary variable representing the existence of a link from node i to node j. Given a set of covariate vectors $\mathbf{X} = \{\mathbf{x}_{ij}\}$, a common approach to modeling the association between $\mathbf{Y}$ and $\mathbf{X}$ that also accounts for unobserved dependencies is to use random effects models,

$$y_{ij} = \mathbf{1}[\beta' \mathbf{x}_{ij} + z_{ij} > 0], \tag{3}$$

where $\mathbf{1}[x > 0]$ is an indicator function that returns 1 if $x > 0$ and 0 otherwise. z_{ij} is the unobserved random effect of pair (i, j), and the z_{ij}'s are not necessarily independent so as to capture potential dependencies in the relational observations. Hoff (2005) and Hoff (2009) motivate multiplicative effects on top of additive effects to model z_{ij} to represent higher-order network structure,

$$z_{ij} = a_i + b_j + u_i' v_j + \epsilon_{ij}, \tag{4}$$

where a_i and b_j are additive row-specific and column-specific effects that respectively represent the proclivity of player i to pass the ball and the popularity of player j as a receiver of the ball, u_i and v_j are multiplicative latent space factors, and the ϵ_{ij}'s are independent random errors.

A variation of such model, an additive and multiplicative effects (AME) model was previously fit on aggregated passing networks in possessions using a subset of the college basketball optical tracking data to model the passing structure of the Division I basketball team. Let y_{ij} be the indicator of whether player i passes to player j in a possession, and assume that

$$y_{ij} = \mathbf{1}[\beta_d x_d + r_i + s_j + u_i^T v_j + \epsilon_{ij} > 0], \tag{5}$$

where $s_j = \beta_j x_j + b_j$ represent additive sender effects and $r_i = \beta_i x_i + a_i$ represent additive receiver effects. The dyadic features are represented by $\beta_d x_d$, and the multiplicative sender and receiver effects by $u_i^T v_j$. The additive sender and receiver effects include row- and column- specific effects (a_i, b_j) and row- and column-specific nodal features (β_i, β_j).

For this model, the dyadic features include indicators for shared basketball position, shared height, shared weight, and shared college class between players. Nodal features include binary variables of whether a player was in a previous possession, whether a player is in a current possession,and points earned per game by a player.

In direct contrast to modeling the risk of a pass at a particular time, this setup is inherently non-temporal. As such, it cannot capture the variability in any given play but instead provides a high level overview of the structure of the passing network for each individual play. The results presented below provide a strong motivation for integrating the multiplicative latent effect framework into the temporal modeling of basketball possessions.

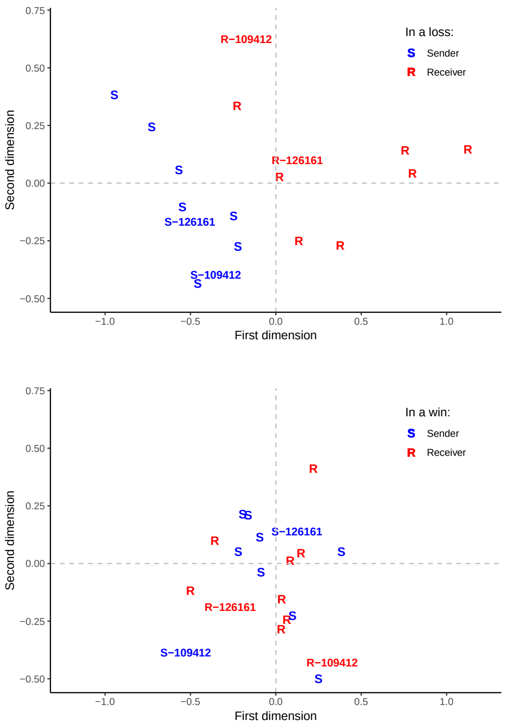


Figure 1: Learned multiplicative sender-specific effects ("s" in blue) and receiver-specific effects ("r" in red) by the AME model in a lost game (top) and a won game (bottom). Effects for players 126161 and 109412 are represented by letter "s" or "r" together with their id codes for demonstration purposes.

Parameters are inferred for each game separately using the Markov chain Monte Carlo sampling algorithm suggested by Hoff (2008) implemented in Hoff et al. (2014). Figure 1 presents the posterior means of the multiplicative sender/passer (blue) and receiver (red) effects for a win and a loss.

Players whose multiplicative effects are in the *same quadrant and differing colors are more likely to interact*. Operationally, to form an effective attack, we would expect a forward (like player 126161) to receive the ball at a reasonable rate from his teammates (like player 109412, a guard and frequent passer), but we only see this phenomenon in the win plot (in the bottom left quadrant) but not in the loss plot. And in fact, player 126161's receiver effect is separated from all his teammates' sender effects, indicating very limited assists to 126161 from his teammates in the lost game. We can also evaluate the overall passing in the game using these plots. In the win the players are clustered together, suggesting that they move the ball among themselves with approximately the same probability. Such behavior is not observed in the loss plot where the multiplicative effects are more spread out and the ball movement is more fragmented: the overlap between the sender and receiver effects is restricted in the upper left quadrant, where only two players are likely to receive passes from their teammates. The products of the multiplicative effects for the remaining players are negative suggesting a reduction in pass rates among them compared to baseline.

3 MODEL

The findings of the AME model presented in section 2.2 demonstrate that the learned latent factors are directly translatable to players' passing patterns which are distinctive between a win and a loss, and that latent factor models help uncover the network characteristics in passing which are predictive of basketball game outcomes. In this section we introduce latent factors in spatio-temporal stochastic modeling of basketball passes and state our full model.

3.1 Model Formulation

Let $Y_{i,j,g}(t)$ denote the event that the ball carrier i passes the ball to teammate j during the time period $(t, t+\delta]$ in game g, and let $\theta_{i,j,g}(t)$ be the log-risk of $Y_{i,j,g}(t)$ given $\mathcal{H}_g(t)$, the history up to time t in game g,

$$\exp(\theta_{i,j,g}(t)) = \lim_{\delta \to 0+} \frac{\mathbb{P}(Y_{i,j,g}(t)|\mathcal{H}_g(t))}{\delta}. \qquad (6)$$

Assume that

$$\begin{aligned}&\theta_{i,j,g}(t)\\ =&W_{i,j,g}(t)^T\eta_{i,j} + \xi_{i,j}(\mathbf{s}_{i,g}(t)) + \zeta_{i,j}(\mathbf{s}_{j,g}(t)) + z_{i,j,g}(t),\end{aligned} \qquad (7)$$

and that

$$z_{i,j,g}(t) = u_{i,g}^T v_{j,g} + \epsilon_{i,j,g}(t). \qquad (8)$$

In Eq (7) $W_{i,j,g}(t)$ is a 5-dimensional covariate vector w.r.t. players i, j at time t in game g, including a constant 1 representing the baseline passing rate, an indicator of whether player i has started dribbling, the log-transformed distance between player i and his nearest defender, j's rank of closeness to i (from 1 to 4, with 1 indicating the closest teammate), a numeric evaluation of how open the passing route is from i to j (a metric introduced by (Cervone et al., 2016)), while $\xi_{i,j}$ maps player i's location on the half court to an additive spatial effect of passing off the ball to j, and $\zeta_{i,j}$ maps a player j's location on the half court to an additive spatial effect of receiving a pass from i based on j's basketball position, with $\epsilon_{i,j,g}(t)$ as an independent standard normal errors. In Eq (8), $u_{i,g}$ and $v_{j,g}$ are R-dimensional vectors representing sender-specific and receiver-specific attributes of ballcarrier i and teammate j mapped onto an R-dimensional latent space. The subscript g indicates that these latent spaces are allowed to vary across games.

Furthermore, let

$$\xi_{i,j}(\mathbf{s}) = \gamma_{i,j}\bar{\xi}_i(\mathbf{s}), \qquad \zeta_{i,j}(\mathbf{s}) = \tilde{\gamma}_{i,j}\bar{\zeta}_{i,\text{pos}(j)}(\mathbf{s}), \qquad (9)$$

where pos(j) denotes player j's basketball position, and $\int_{\mathcal{S}} \bar{\xi}_{i,j}(\mathbf{s})d\mathbf{s} = \int_{\mathcal{S}} \bar{\zeta}_{i,j}(\mathbf{s})d\mathbf{s} = 1$, with $\mathcal{S}$ as the half court and $\mathbf{s}$ as a pair of coordinates corresponding to a location on $\mathcal{S}$. Setting $X_{i,j,g}(t) = (W_{i,j,g}(t)^T, \bar{\xi}_{i,j}(\mathbf{s}_{i,g}(t)), \bar{\zeta}_{i,j}(\mathbf{s}_{j,g}(t)))^T$ and $\beta_{i,j} = (\eta_{i,j}^T, \gamma_{i,j}, \tilde{\gamma}_{i,j})^T$, Eq (7) becomes

$$\log(\theta_{i,j,g}(t)) = X_{i,j,g}(t)^T\beta_{i,j} + u_{i,g}^T v_{j,g} + \epsilon_{i,j,g}(t). \qquad (10)$$

We would like to emphasize that the model formulation is not only an extension of Eq (2), but also an extension of Eq (3) and (4). A hierarchical structure in the multiplicative latent factors is introduced to allow *differing sender-specific and receiver-specific effects in different games*, and here we model the time-varying risk (intensity function) of a *non-homogeneous spatio-temporal Poisson process* rather than the probability of binary links using logistic (or probit) regression.

3.2 Parameter Estimation

Conditioning on all the covariate vectors $X_{i,j,g}(t)$ regarding n players in G games in total and the latent space dimentionality R, the unknown quantities of the model are

- $\Theta = \{\theta_{i,j,g}(t)\}$, the set of log-risks;
- $\beta = \{\beta_{i,j}\}$, the set of coefficients for all player pairs;
- $\mathbf{U} = \{U_g\}$, the set of $n \times R$ matrices, with each row representing the sender-specific effect of a player in

a game, and $\mathbf{V} = \{V_g\}$, the set of $n \times R$ matrices, with each row representing the receiver-specific effect of a player in a game.

Given priors for parameters $\beta_{i,j}, u_{i,g}$ and $v_{i,g}$, a fully Bayesian inference procedure is deployed that estimates the full posterior distribution of the parameters via Markov Chain Monte Carlo. We adopt an efficient Metropolis-within-Gibbs sampling scheme that allows us to adequately address the complex dependence issues that are visible in network models and are typically challenging to address with variational inference or a more complicated sampling scheme.

Given all the observations $\mathbf{Y} = \{Y_{i,j,g}(t)\}$ and a set of initial values $\phi^{(0)} = \{\Theta^{(0)}, \beta^{(0)}, \mathbf{U}^{(0)}, \mathbf{V}^{(0)}\}$, a sequence of parameter samples $\{\phi^{(s)}\}$ can be iteratively generated from the full conditional distributions of the parameters. In each iteration, based on the latest parameter sample $\phi^{(s-1)}$, a new sample $\phi^{(s)}$ is acquired through the following steps:

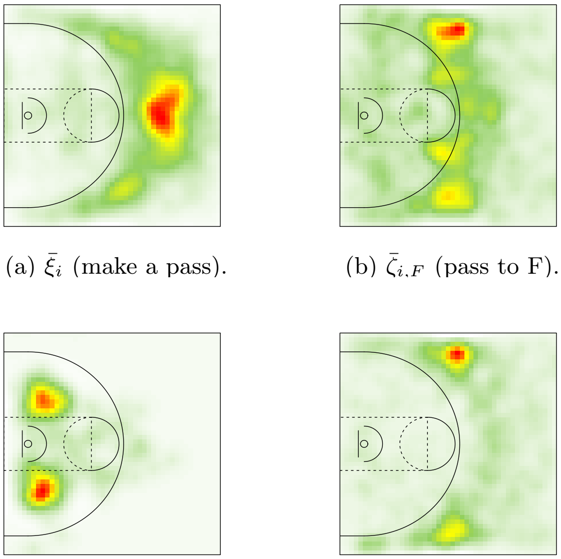
1. For each pair (i,j), $1 \leq i \neq j \leq n$, sample $\beta^{(s)}_{i,j}$ from $p(\beta_{i,j}|\Theta^{(s-1)}, \mathbf{U}^{(s-1)}, \mathbf{V}^{(s-1)})$;
2. For game $g = 1, \ldots, G$ and $i = 1, \ldots, n$,
   (a) sample $U_g[i,]^{(s)}$ from $p(U_g[i,]|\Theta^{(s-1)}, \beta^{(s-1)}, V_g^{(s-1)})$;
   (b) sample $V_g[i,]^{(s)}$ from $p(V_g[i,]|\Theta^{(s-1)}, \beta^{(s-1)}, U_g^{(s-1)})$;
3. For each pair (i,j) and time point t in game g, set $\theta^*_{i,j,g}(t) = X_{i,j,g}(t)^T \beta^{(s)}_{i,j} + (U_g[i,]^{(s)})^T V_g[j,]^{(s)} + \epsilon^*_{i,j,g}(t)$, where $\epsilon^*_{i,j,g}(t)$ is a standard normal error; set $\theta^{(s)}_{i,j,g}(t) = \theta^{(s-1)}_{i,j,g}(t)$ first, and then replace it by $\theta^*_{i,j,g}(t)$ with probability $\min\left(\frac{p(Y_{i,j,g}(t)|\theta^*_{i,j,g}(t))}{p(Y_{i,j,g}(t)|\theta^{(s-1)}_{i,j,g}(t))}, 1\right)$.

We choose the priors for $\beta_{i,j}, u_{i,g}$ and $v_{i,g}$ to be non-informative and independent multivariate normal distributions. This allows us to derive the full conditional distributions for all the parameters. Details are provided in the Supplement.

3.3 Spatial Effect Estimation

Although treated as known covariates in parameter estimation, the normalized additive spatial effect functions in Eq (9), $\bar{\xi}_i$ and $\bar{\zeta}_{i,\text{pos}(j)}$, are not readily available in the optical tracking data and thus need to be estimated. We employ a *different, simpler and more efficient* method than the Gaussian Markov random field approximation used by Cervone et al. (2016). Beyond reducing computation time, this method also does not require a massive volume of data to produce reasonable estimates of the additive spatial effect functions.

We first divide the half court $\mathcal{S}$ (47 feet by 50 feet) into 1 ft × 1 ft tiles and use thin plate splines regression (Duchon, 1977) to estimate smooth 2-d functions based on empirical counts on the tiles. For each player i, the estimation of $\bar{\xi}_i$ is based on the total number of times player i stood on each tile when he made a pass, and for each possible basketball position of his teammate, $\text{pos}(j) \in \{\text{forward (F)}, \text{center (C)}, \text{guard (G)}\}$, the estimation of $\bar{\zeta}_{i,\text{pos}(j)}$ is based on the total number of times any teammate playing position pos(j) stood on each tile when he received the ball from the player i.



(a) $\bar{\xi}_i$ (make a pass).

(b) $\bar{\zeta}_{i,F}$ (pass to F).

(c) $\bar{\zeta}_{i,C}$ (pass to C).

(d) $\bar{\zeta}_{i,C}$ (pass to G).

Figure 2: Estimated spatial effects $\bar{\xi}$ and $\bar{\zeta}$ for a player i (id code 601140). A redder/darker color corresponds to a higher log-risk of making a pass. For example, this player most frequently passes off the ball when he is outside the center of the three-point line, and he tends to pass the ball to a center who is approaching the restricted area from either side.

Take the spatial effect function $\bar{\xi}_i$ for a certain player i for example. Suppose $\{n_k\}_{k=1}^K$ are the empirical counts of i's passing location on tiles $k = 1, \ldots, K$ centered at $\{\mathbf{c}_k\}_{k=1}^K$. Set $\tilde{n}_k = \frac{n_k}{\sum_{k=1}^K n_k}$, and the function $\bar{\xi}_i : \mathcal{S} \to \mathbb{R}$ is obtained by minimizing

$$\sum_{k=1}^{K} \|\tilde{n}_k - \bar{\xi}_i(\mathbf{c}_k)\|^2 + \lambda \int_{\mathcal{S}} \|\frac{\partial^2 \bar{\xi}_i(\mathbf{s})}{\partial \mathbf{s}^2}\|_F^2 d\mathbf{s}, \qquad (11)$$

where $\|\cdot\|_F$ is the Frobenius norm of matrices, and the smoothness parameter λ is chosen by generalized cross validation, as introduced by Green and Silverman (1993).

Figure 2 visualizes the estimated spatial effect functions for one player 601140, who plays as a guard. There are distinct spatial patterns in his passing choices and preferences, where he tends to pass off the ball outside the center three-point line (subplot (a)), and his forward teammate(s) is more likely to receive the ball from him when this teammate is at a corner of the three-point line (subplot (b)).

4 EXPERIMENTS

4.1 Synthetic Data

A synthetic dataset including records in 2 games involving 8 players in total is generated from our model. Around 10,000 observations are generated in total. For each game, the first 90% of observations are used for model training, and the last 10% observations are held out as the testing data. We evaluate the log-likelihoods of the training data and the held-out data for each of these three models:

1. The full model with 2-dimensional multiplicative latent factors (labeled as "**Latent**");

2. A subset model with only the time-varying covariates and spatial effects $X_{i,j,g}(t)$ (labeled as "**Covariate**");

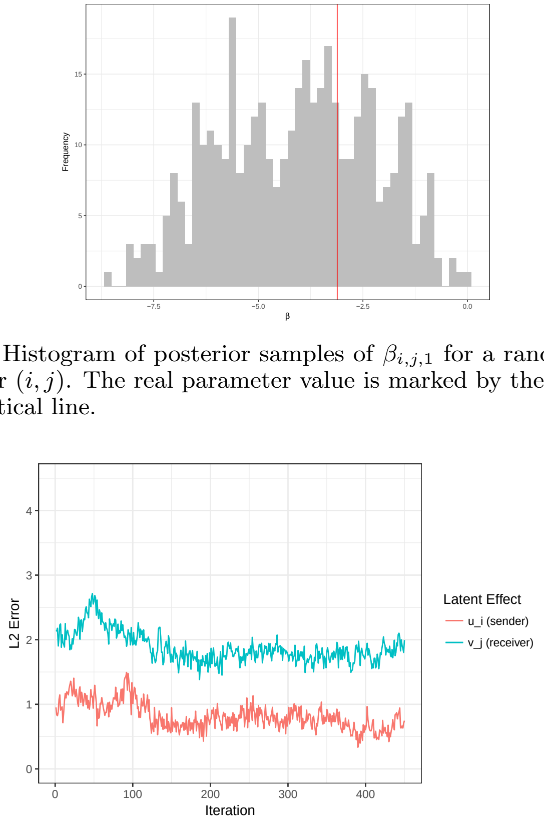
3. A further subset model with only the spatial effects (that is, without the time-varying covariates $W_{i,j,g}(t)$) (labeled as "**Spatial**").

Quality of sampling algorithm: The plots in Figure 3 validate the effectiveness of the sampling scheme (in section 3.2) in recovering the parameters. Plot (a) shows the histogram of the posterior samples of $\beta_{i,j,1}$, the baseline log-risk for player i to pass to j, with i and j randomly selected. The posterior samples are approximately centered around the true parameter value (denoted by the red vertical line). Plot (b) visualizes the squared errors (measured by squared euclidean distances) between the samples of sender-specific effect vector and receiver-specific effect vector between the respective true vector values for a random player pair i, j in the first game. The squared errors fluctuate in the early iterations due to the random nature of the sampling algorithm, but drop down and stabilize at the end.

Model performance: Numerical results based on repeated experiments are presented in Table 1. The full model can accommodate both spatial, temporal and multiplicative effects and fits the synthetic data better than competitors. The good performance on the held-out likelihood demonstrates that this is not an artifact of overfitting by a larger model.

Table 1: Log-likehoods of the threes models in simulation experiments. The full model constantly outperforms the subset models.

Model	Training log-likelihood	Held-out log-likelihood
Latent	-10025.93 ± 189.31	-1219.80 ± 57.62
Covariates	-10719.68 ± 120.26	-1314.00 ± 43.43
Spatial	-14255.32 ± 92.88	-1689.61 ± 16.20



(a) Histogram of posterior samples of $\beta_{i,j,1}$ for a random pair (i, j). The real parameter value is marked by the red vertical line.

(b) Squared errors of $u_{i,g}$ samples and $v_{j,g}$ samples with respect to the actual latent space vector values for game 1 and a random player pair i, j.

Figure 3: Parameter recovery checking plots in simulation experiments.

4.2 Real Data

4.2.1 Data Description

The real dataset is collected by SportsVu optical tracking systems from the home arena of an NCAA Division I basketball program. This is the first time a SportsVu dataset for college basketball has been made available for analysis. Previous NBA SportsVu datasets have been released because every basketball stadium is mandated to retain a tracking video camera. Features were created by taking snapshots of the game every 1/25th of a second and recording each player's location and action, the location of the ball, and general identification information about the game, the teams, and the players[2]. All of the observations were automatically translated and stored into data files by the SportsVu software, originally as 3 different types of files:

1. **Boxscore**: The overall player statistics (assists, points, rebounds) for each game. It is used as a reference for evaluating player performance.

2. **Play by Play**: The event summary (dribble, foul, pass, etc.) for each observation in each game. It is used to divide each game into basketball possessions and to extract passing networks.

3. **Sequence Optical**: The locational summary of each player and of the ball for each game at a 25Hz resolution. It is used to calculate relevant spatio-temporal covariates and map the passing order for each possession.

The final dataset was created by merging the three datasets together by time, player id, and game id. Each game was divided into possessions, which ended on made shots, missed shots, and turnovers. Although the end of a play may not have ended after a non-turnover violation, the locations of the ball and players were reset after these events. For this, non-turnover possessions (i.e. kick ball violation) were also noted as the end of a possession. Possessions that ended with fouls were removed from the dataset to reduce the number of transition states in the model, similar to Cervone et al. (2016).

4.2.2 Model Fitting and Results

We fit the full model (with the dimension of the latent space $R = 2$) on the records for all the games from the beginning of December 2014 to the end of January 2015. For model validation, only 90% of the records in each game are used for model fitting, with the rest 10% held-out to test model predicting capabilities. Same as in section 4.1, the full model ("**Latent**") is compared with two subset models (i.e. "**Covariate**" and "**Spatial**"). The log-likelihoods on training data and testing data in Table 2 indicate that the addition of multiplicative latent factors yields better explanation of the passing patterns as well as better out-of-sample prediction of passing occurrences in real-time basketball games.

Figure 4 visualizes the key results of our model: the sender-specific effects U_g and receiver-specific effects V_g for a game of interest g. Each player i's sender-specific effect vector $u_{i,g}$ corresponds to a 2-dimensional coordinate marked in red, and his receiver-specific effect vector $v_{i,g}$ corresponds to a 2-dimensional coordinate marked in blue. For player i and j, if $u_{i,g}$ and $v_{j,g}$ reside in the same quadrant, then i tends to pass to j more frequently in game g. Here we present the sender- and receiver-specific effects for two games, g_1 and g_2, where the home team lost the former (plot (a)) and won the latter (plot (b)). Players' passing behaviors are apparently different in a loss than in a win. We demonstrate this by highlighting two players, 601140 and 842297: In a loss (plot (a)), 842297 passes frequently to 601140, but hardly receives the ball back. On the other hand, in a win (plot (b)), 601140 and 842297 pass to each other with almost the same frequency.

Looking at all the sender- and receiver- effects, we can see that in a game the team lost (plot (a)), the latent effects are farther away from the origin than those in a game the team won (plot (b)). This indicates that in a loss player passing behavior is extremely variable

Table 2: Log-likehoods of the full model and two subset models on training data and testing data.

Model	Training log-likelihood	Held-out log-likelihood
Latent	-679.33 ± 114.51	-58.52 ± 11.20
Covariates	-917.89 ± 220.41	-64.68 ± 12.59
Spaitial	-904.76 ± 151.81	-67.54 ± 7.91

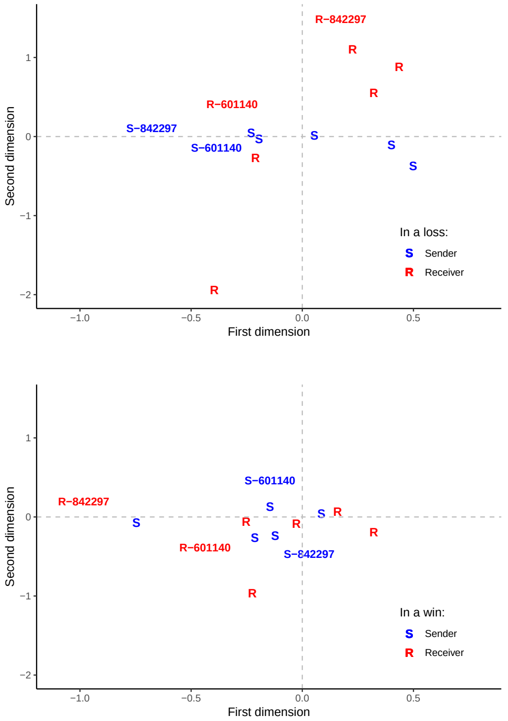
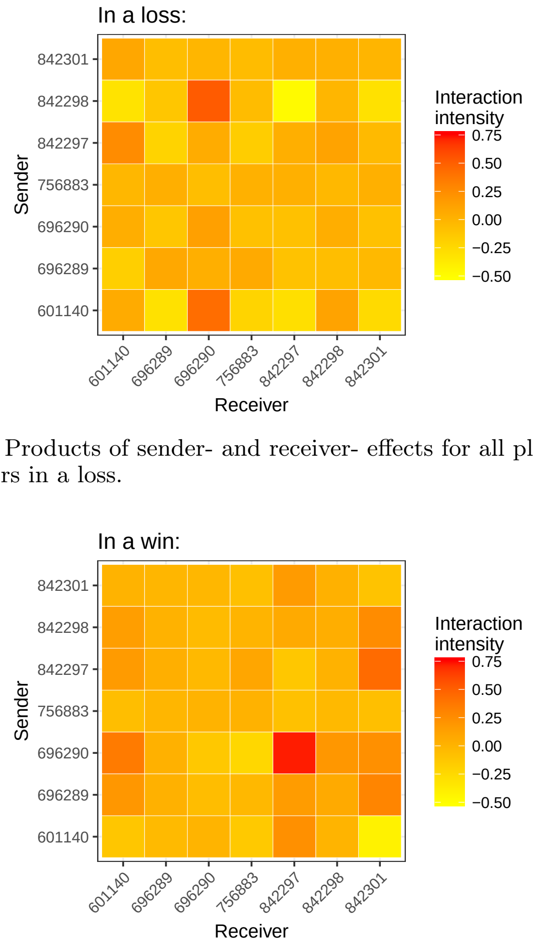


Figure 4: Passing decision multiplicative latent factors in a loss (top) versus in a victory (bottom). Sender- and receiver-specific effects are marked with "S" (in blue) and "R" (in red) , respectively. The latent factors for player 601140 and 842297 are also represented by their player id codes for later demonstration.

[2] In the dataset, every player and every team is identified by a randomly assigned id, and every game is identified by the date of the game.

and depends on who they are passing to, whereas in a victory the players are more measured in terms of passing and receiving behavior. That is, our results provide a measurable indicator of consistent ball movement between all the players and we see that this is associated with positive game outcomes.



(a) Products of sender- and receiver- effects for all player pairs in a loss.

(b) Products of sender- and receiver- effects for all player pairs in a win.

Figure 5: Player interaction intensity matrix in a loss versus in a victory. Darker color indicates stronger tendency of passing.

The differences in player behaviors between a win and a loss are more obvious when we directly examine the inner product of the multiplicative latent effects, $u_{i,g}^T v_{j,g}$, for each player pair (i, j) in game g. According to Eq (7) and (8), the higher the value of $u_{i,g}^T v_{j,g}$ is, the more likely i passes to j in game g. In other words, the matrix $U_g V_g^T$ is a player interaction intensity matrix in game g, *after adjusting for the spatio-temporal factors*[3]. We visualize the matrix entries in a loss (plot (a)) and in a win (plot (b)) in Figure 5. If we revisit the aforementioned example and take $i = 601140$ and $j = 842297$, we observe that $u_{i,g}^T v_{j,g}$ is noticeably larger than $u_{j,g}^T v_{i,g}$ in the lost game (plot (a)), while the two quantities are approximately equal in the win (plot (b)), which implies that the two players' interactions are somewhat one-sided in a loss but more balanced in a victory. Further, this new set of plots allows us to evaluate the overall passing tendencies in a game and detect that they are higher for the majority of player pairs in a successful game, which supports the idea that more interactions and more active teamwork lead to better outcomes. Furthermore, we can identify other passing anomalies in these plots: in a loss, player 842298 holds both the highest and lowest passing intensities of all players, a distinction that no player holds in the win–this type of preferential passing behavior can be extremely detrimental to game outcomes as it might lead to under- and over-utilization of certain players.

5 CONCLUSION

We propose a novel social network metric of basketball game success based on latent factors that capture higher-order patterns in players' passing choices and preferences in basketball games. Our method expands on both the state-of-the-art spatio-temporal stochastic process model and latent factor models for binary relational links. Parameter inference is carried out by a Markov chain Monte Carlo sampler, which is effective in estimating the parameters of interest values, as suggested by experiments on synthetic data. Experiments on the very first high-resolution optical tracking dataset in college basketball show that our model outperforms current state-of-the-art models in both goodness-of-fit and out-of-sample prediction, and that the learned latent sender-specific and receiver-specific effects offer direct interpretation of the interactions among players on the same team and provide an interpretable differentiation between wins and losses. While this model is applicable to basketball and other team sports, it can also be translated into modeling longitudinal social networks observations, such as email conversations, international trades, and regional conflicts.

In the future we plan to model interactions between defenders and offenders as network links and scale up the inference algorithm via sparse tensor techniques.

Acknowledgement

We thank the reviewers and area chair as well as Dan Cervone for helpful discussions. The work was partially supported by the U.S. Army Research Institute for the Behavioral & Social Sciences (grant W911NF1810233).

[3] Therefore, this matrix carries different (and deeper) information than a simple frequency table of passes between players in a game (see Figure 4 in Supplement for reference).

References

D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514):585–599, 2016.

J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.

D. Durante and D. B. Dunson. Nonparametric bayes dynamic modelling of relational data. *Biometrika*, 101(4): 883–898, 2014.

J. H. Fewell, D. Armbruster, J. Ingraham, A. Petersen, and J. S. Waters. Basketball teams as strategic networks. *PloS one*, 7(11):e47445, 2012.

A. Franks, A. Miller, L. Bornn, K. Goldsberry, et al. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9 (1):94–121, 2015.

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach.* CRC Press, 1993.

J. Gudmundsson and M. Horton. Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, 50 (2):22, 2017.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664, 2008.

P. Hoff, B. Fosdick, A. Volfovsky, and K. Stovel. Likelihoods for fixed rank nomination networks. *Network Science*, 1 (3):253–277, 2013.

P. Hoff, B. Fosdick, A. Volfovsky, and K. Stovel. amen: Additive and multiplicative effects modeling of networks and relational data. *R package version 0.999. URL: http://CRAN. R-project. org/package= amen*, 2014.

P. D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, 100(469): 286–295, 2005.

P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261, 2009.

P. D. Hoff. Additive and multiplicative effects network models. *arXiv preprint arXiv:1807.08038*, 2018.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

J. Hollinger. Pro basketball forecast: 2005-2006. *Dulles, VA: Potomac*, 2005.

A. Miller, L. Bornn, R. Adams, and K. Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *International Conference on Machine Learning*, pages 235–243, 2014.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

D. Omidiran. A new look at adjusted plus/minus for basketball analysis. In *MIT Sloan Sports Analytics Conference [online]*, 2011.

K. Pelechrinis and E. Papalexakis. thoops: A multi-aspect analytical framework spatio-temporal basketball data using tensor decomposition. *arXiv preprint arXiv:1712.01199*, 2017.

D. K. Sewell and Y. Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

L. Xin, M. Zhu, H. Chipman, et al. A continuous-time stochastic block model for basketball networks. *The Annals of Applied Statistics*, 11(2):553–597, 2017.