"This is why we play": Characterizing Online Fan Communities of the NBA Teams

JASON SHUO ZHANG, CHENHAO TAN, AND QIN LV, University of Colorado Boulder, USA

Professional sports constitute an important part of people's modern life. People spend substantial amounts of time and money supporting their favorite players and teams, and sometimes even riot after games. However, how team performance affects fan behavior remains understudied at a large scale. As almost every notable professional team has its own online fan community, these communities provide great opportunities for investigating this research question. In this work, we provide the first large-scale characterization of online fan communities of professional sports teams.

Since user behavior in these online fan communities is inherently connected to game events and team performance, we construct a unique dataset that combines 1.5M posts and 43M comments in NBA-related communities on Reddit with statistics that document team performance in the NBA. We analyze the impact of team performance on fan behavior both at the game level and the season level. First, we study how team performance in a game relates to user activity during that game. We find that surprise plays an important role: the fans of the top teams are more active when their teams lose and so are the fans of the bottom teams in an unexpected win. Second, we study fan behavior over consecutive seasons and show that strong team performance is associated with fans of low loyalty, likely due to "bandwagon fans." Fans of the bottom teams tend to discuss their team's future such as young talents in the roster, which may help them stay optimistic during adversity. Our results not only contribute to understanding the interplay between online sports communities and offline context but also provide significant insights into sports management.

CCS Concepts: • **Applied computing → Law, social and behavioral sciences**;

Additional Key Words and Phrases: online fan communities, team performance, professional sports, NBA

ACM Reference Format:
Jason Shuo Zhang, Chenhao Tan, and Qin Lv. 2018. "This is why we play": Characterizing Online Fan Communities of the NBA Teams. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 197 (November 2018), 25 pages. https://doi.org/10.1145/3274466

1 INTRODUCTION

Our [the Los Angeles Lakers'] collective success having forged some kind of unity in this huge and normally fragmented metropolis, it cuts across cultural and class lines.

— Kareem Abdul-Jabbar, an NBA Hall of Famer.

We thank anonymous reviewers for helpful comments and discussions. We thank Jason Baumgartner and Jack Hessel for sharing the dataset that enabled this research. We thank Scott Holman from the CU Boulder Writing Center for the very detailed proofreading. This work is supported in part by the US National Science Foundation (NSF) through grant CNS 1528138.

Author's address: Jason Shuo Zhang, Chenhao Tan, and Qin Lv, University of Colorado Boulder, 1111 Engineering Dr, Boulder, CO, 80309, USA.

2573-0142/2018/11-ART197 $15.00
https://doi.org/10.1145/3274466

Professional sports not only involve competitions among athletes, but also attract fans to attend the games, watch broadcasts, and consume related products [66]. For instance, the 2017 final game of the National Basketball Association (NBA) attracted 20 million TV viewers [50]; a 30-second commercial during the Super Bowl cost around 4.5 million dollars in 2015 and these commercials have become an integral part of American culture [70].[1] Fans of sports teams can be very emotionally invested and treat fans of rival teams almost as enemies, which can even lead to violence [42].

Such excitement towards professional sports extends to online communities. A notable example is /r/NBA on Reddit, which attracts over a million subscribers and has become one of the most active subreddits on Reddit, a popular community-driven website [40]. Goldschein, a sports writer, has suggested that online fan communities are gradually replacing the need for sports blogs and even larger media outlets altogether [17]. The growth of online fan communities thus provides exciting opportunities for studying fan behavior in professional sports at a large scale.

It is important to recognize that fan behavior is driven by sports events, including sports games, player transfers between teams, and even a comment from a team manager. The dynamic nature of sports games indicates that discussions in online fan communities may echo the development in games, analogous to the waves of excitement in a stadium. Therefore, our goal in this paper is to characterize *online* fan communities in the context of *offline* games and team performance.

To do that, we build a large-scale dataset of online fan communities from Reddit with 479K users, 1.5M posts, and 43M comments, as well as statistics that document offline games and team performance.[2] We choose Reddit as a testbed because 1) Reddit has explicit communities for every NBA team, which allows us to compare the differences between winning teams and losing teams; and 2) Reddit is driven by fan communities, e.g., the ranking of posts is determined by upvotes and downvotes of community members. In comparison, team officials can have a great impact on a team's official Twitter account and Facebook page.

Organization and highlights. We first summarize related work (Section 2) and then provide an overview of the NBA fan communities on Reddit as well as necessary background knowledge regarding the NBA games (Section 3). We demonstrate the seasonal patterns in online fan communities and how they align with the NBA season in the offline world. We further characterize the discussions in these NBA fan communities using topic modeling.

We investigate three research questions in the rest of the paper. First, we study the relation between team performance in a game and this game's associated fan activity in online fan communities. We are able to identify game threads that are posted to facilitate discussions during NBA games. These game threads allow us to examine the short-term impact of team performance on fan behavior. We demonstrate intriguing contrasts between top teams and bottom teams: user activity increases when top teams lose and bottom teams win. Furthermore, close games with small point differences are associated with higher user activity levels.

Second, we examine how team performance influences fan loyalty in online communities beyond a single game. It is important for professional teams to acquire and maintain a strong fan base that provides consistent support and consumes team-related products. Understanding fan loyalty is thus a central research question in the literature of sports management [12, 13, 51, 72]. For instance, "bandwagon fan" refers to a person who follows the tide and supports teams with recent success. Top teams may have lower fan loyalty due to the existence of many bandwagon fans. Our results validate this hypothesis by using user retention to measure fan loyalty. We also find that a team's fan loyalty is correlated with the team's improvement over a season and with the average age of the roster.

[1] Most influential Super Bowl commercials: http://time.com/4653281/super-bowl-ads-commercials-most-influential-time/.
[2] The dataset is available at http://jasondarkblue.com/papers/CSCW2018NBADataset_README.txt.

Third, we turn to the content in online fan communities to understand the impact of team performance on the topics of discussion. Prior studies show that a strong fan base can minimize the effect of a team's short-term (poor) performance on its long-term success [46, 53]. To foster fan identification in teams with poor performance, fans may shift the focus from current failure to future success and "*trust the process*"[3] [6, 12, 24]. Discussions in online fan communities enable quantitative analysis of such a hypothesis. We show that fans of the top teams are more likely to discuss *"season prospects,"* whiles fans of the bottom teams are more likely to discuss *"future."* Here *"future"* refers to the assets that a team has, including talented young players, draft picks, and salary space, which can potentially prepare the team for future success in the following seasons.

We offer concluding discussions in Section 7. Our work develops the first step towards studying fan behavior in professional sports using online fan communities and provides implications for online communities and sports management. For online communities, our results highlight the importance of understanding online behavior in the offline context. Such offline context can influence the topics of discussion, the activity patterns, and users' decisions to stay or leave. For sports management, our work reveals strategies for developing a strong fan base such as shifting the topics of discussion and leveraging unexpected wins and potential future success.

2 RELATED WORK

In this section, we survey prior research mainly in two areas related to the work presented in this paper: online communities and sports fan behavior.

2.1 Online Communities

The proliferation of online communities has enabled a rich body of research in understanding group formation and community dynamics [3, 25, 29, 41]. Most relevant to our work are studies that investigate how external factors affect user behavior in online communities [37, 43, 49, 73]. Palen and Anderson [37] provide an overview of studies on social media behavior in response to natural disasters and point out limits of social media data for addressing emergency management. Romero et al. [43] find that communication networks between traders "turtle" up during shocks in stock price and reveal relations between social network structure and collective behavior. Other offline events studies include the dynamics of breaking news [26, 27, 31], celebrity death [15, 28], and Black Lives Matter [52, 59]. This literature illustrates that online communities do not only exist in the virtual world. They are usually deeply embedded in the offline context in our daily life.

Another relevant line of work examines user engagement in multiple communities and in particular, user loyalty [19, 55, 74]. Hamilton et al. [19] operationalize loyalty in the context of multi-community engagement and consider users loyal to a community if they consistently prefer the community over all others. They show that loyal users employ language that signals collective identity and their loyalty can be predicted from their first interactions.

Reddit has attracted significant interest from researchers in the past few years due to its growing importance. Many aspects and properties of Reddit have been extensively studied, including user and subreddit lifecycle in online platforms [36, 55], hate speech [7, 8, 45], interaction and conflict between subreddits [30, 54], and its relationship with other web sources [36, 60]. Studies have also explored the impacts of certain Reddit evolutions and policy changes on user behaviors. Notable events include pre-default subreddit [33] and Reddit unrest [7, 35, 36].

Our work examines a special set of online communities that derived from professional sports teams. As a result, regular sports games and team performance are central for understanding these

[3] A mantra that reflects Philadelphia 76ers' identity [39] 76ers went through a streak of losing seasons to get top talents in draft-lottery and rebuild the team.

	/r/NBA	Average of team subreddits (std)
#users	400K	13K (8K)
#posts	847K	24K (16K)
#comments	33M	328K (282K)

Table 1. Dataset Statistics. There are in total 30 teams in the NBA league. #users refers to the number of unique users who posted/commented in the subreddit.

communities and user loyalty in these communities. Different from prior studies, we focus on the impact of team performance on user behavior in online fan communities.

2.2 Sports Fan Behavior

As it is crucial for a sports team to foster a healthy and strong fan base, extensive studies in sports management have studied fan behavior. Researchers have studied factors that affect purchasing behavior of sports fans [48, 58, 63], including psychometric properties and fan motivation. A few studies also build predictive models of fan loyalty [4, 72]. Bee and Havitz [4] suggest that fan attraction, involvement, psychological commitment, and resistance can be predictors of fan behavioral loyalty. Dolton and MacKerron [11] estimate that the happiness that fans feel when their team wins is outweighted by the sadness that strikes when their team loses by a factor or two. Yoshida et al. [72] build regression models based on attitudinal processes to predict behavioral loyalty. The potential influence of mobile technology on sports spectators is also examined from different angles [23, 34, 57]. Torrez Riley [57] describes survey results that suggest the current usage of mobile technology among college sports fans. The work by Ludvigsen and Veerasawmy [34] examines the potential of interactive technologies for active spectating at sporting events. Most relevant to our work are studies related to fan identification [6, 12, 13, 20, 22, 24, 51, 53, 64] and we have discussed them to formulate our hypotheses. These studies usually employ qualitative methods through interviews or small-scale surveys.

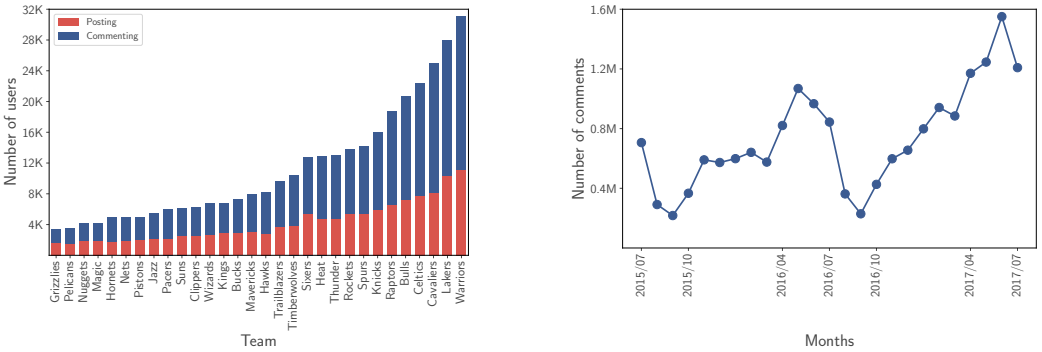
It is worth noting that fan behavior can differ depending on the environment. Cottingham [9] demonstrates the difference in emotional energy between fans in sports bars and those attending the game in the stadium. In our work, we focus on online communities, which are an increasingly important platform for sports fans. These online fan communities also allow us to study team performance and fan behavior at a much larger scale than all existing studies.

3 AN OVERVIEW OF NBA FAN COMMUNITIES ON REDDIT

Our main dataset is derived from NBA-related communities on Reddit, a popular website organized by communities where users can submit posts and make comments. A community on Reddit is referred to as a *subreddit*. We use community and subreddit interchangeably in this paper. We first introduce the history of NBA-related subreddits and then provide an overview of activity levels and discussions in these subreddits.

3.1 NBA-related Subreddits

On Reddit, the league subreddit /r/NBA is for NBA fans to discuss anything that happened in the entire league, ranging from a game to gossip related to a player. There are 30 teams in total in the NBA league, and each team's subreddit is for fans to discuss team-specific topics. Each subreddit has multiple moderators to make sure posts are relevant to the subreddit's theme. We collected posts and comments in these 31 subreddits (/r/NBA + 30 NBA team subreddits) from pushshift.io [65]



(a) Number of users in team subreddits.

(b) User activity in /r/NBA by month.

Fig. 1. Figure 1a shows the number of users that post and comment in team subreddits. /r/Warriors, /r/Lakers, /r/Cavaliers are the top three subreddits in both posting and commenting. The average number of users across all teams is 4,166 for posting and 11,320 for commenting. Figure 1b shows user activity level in /r/NBA by month. During the off season (July-mid October), user activity decreases sharply, as no games are played during this period. Then in the regular season (late October to next March), user activity increases steadily. The activity of /r/NBA peaks in May and June, as the championship games happen in these two months.

from the beginning of each subreddit until October 2017. [4] The overall descriptive statistics of our dataset are shown in Table 1.

A brief history of the NBA-related subreddits on Reddit. NBA-related subreddits have thrived since January 2008, when Reddit released a new policy to allow users to create their own subreddit. The Lakers' and the Celtics' subreddits were created by fans in 2008, and they are the first two NBA teams to have their team subreddits. These two teams are also widely acknowledged as the most successful franchises in the history of the NBA league [18]. It is also worth noting that these two teams' subreddits were created even before /r/NBA, which was created at the end of 2008. For the remaining 28 teams, 14 of their subreddits were created by users in 2010, and the other 14 were created in 2011. Moreover, three teams' subreddit names have changed. The Pelicans' subreddit changed their subreddit name from /r/Hornets to /r/Nolapelicans and the Hornets' subreddit from /r/Charlottebobcats to /r/Charlottehornets because these two teams changed their official team names. Additionally, the Rockets' subreddit shortened its name from /r/Houstonrockets to /r/Rockets. To rebuild each team's complete subreddit history, we combined posts and comments in these three teams' old and new subreddits. Figure 1a presents the number of users that post and comment in each team subreddit.

3.2 NBA Season Structure Reflected in Reddit Activity

As discussed in the introduction, fan behavior in NBA-related subreddits is influenced by offline events. In particular, the NBA runs in seasons, and seasonal patterns are reflected on Reddit. To show that, we start with a brief introduction of the NBA. 30 teams in the NBA are divided into two conferences (East and West). In each season, teams play with each other to compete for the final championship. The following three time segments make up a complete NBA season: [5]

[4] A small amount of data is missing due to scraping errors and other unknown reasons with this dataset [16]. We checked the sensitivity of our results to missing posts with a dataset provided by J.Hessel; our results in this paper do not change after accounting for them.

[5] Please see the NBA's official description for more details [2].

- **Off season**: from July to mid October. There are no games in this period.[6] Every team is allowed to draft young players, sign free agents, and trade players with other teams. The bottom teams in the last season get the top positions when drafting young players, which hopefully leads to a long-term balance between teams in the league. The goal of the off season for each team is to improve its overall competitiveness for the coming season.
- **Regular season**: from late October to middle of April. Regular season games occur in this period. Every team has 82 games scheduled during this time, 41 home games and 41 away games. A team's regular season record is used for playoff qualification and seeding.
- **Playoff season**: from the end of regular season to June. 16 teams (the top 8 from the Western conference and the top 8 from the Eastern conference) play knockouts in each conference and compete for the conference championship. The champion of the Western Conference and the Eastern Conference play the final games to win the final championship.

Given the structure, a complete NBA season spans two calendar years. In this paper, for simplicity and clarity, we refer to a specific season by the calendar year when it ends. For instance, the official 2016-2017 NBA season is referred to as *the 2017 season* throughout the paper.

User activity in NBA-related subreddits is driven by the structure of the NBA season. As an example, Figure 1b shows user activity in /r/NBA by month in the 2016 and 2017 season. From July to September, user activity decreases sharply as there are no games during this period. Then from October to the next March, the number of comments increases steadily as the regular season unfolds. According to the NBA rules, every game in the regular season carries the same weight for playoff qualification, the games in October should be equally important as the games in March. However, fans are much more active on Reddit as it gets closer to the end of the regular season because they deem these games "more critical." This may be explained by the "deadline pressure" phenomenon in psychology [1]. This circumstance has also been observed in other sports. For instance, Paton and Cooke [38] illustrate that the attendance of domestic cricket leagues in England and Wales is much higher in the later segment of the season than the earlier segment. Hogan et al. [21] find that the possibility of the home team reaching the knock-out stage had a significant positive impact on attendance in the European Rugby Cup. We also find that user activity drops a little bit in April in both the 2016 and 2017 season. One possible explanation is that as the regular season is ending, fans of the bottom teams that clearly cannot make the playoffs reduce their activity during this period. After that, the volume of comments increases dramatically during the playoff games. The activity of /r/NBA peaks in May and June, when the conference championship and final championship games happen.

3.3 Topic analysis

To understand what fans are generally talking about in NBA-related subreddits, we use Latent Dirichlet Allocation (LDA) [5], a widely used topic modeling method, to analyze user comments. We treat each comment as a document and use all the comments in /r/NBA to train a LDA model with the Stanford Topic Modeling Toolbox [14]. We choose the number of topics based on perplexity scores [61]. The perplexity score drops significantly when the topic number increases from 5 to 15, but does not change much from 15 to 50, all within 1370-1380 range. Therefore, we use 15 topics in this paper. Table 2 shows the top five topics with the greatest average topic weight and the top ten weighted words in each topic. Two authors, who are NBA fans and active users on /r/NBA, manually assigned a label to each of the five most frequent topics based on the top words in each topic. Each label in Table 2 summarizes the topic's gist, and the five labels are *"personal*

[6]There are Summer League games and preseason games played in this period, but the results don't count in season record.

LDA topic	top words	average topic weight
"personal opinion"	opinion, fact, reason, agree, understand, medium, argument, talking, making, decision	0.083
"game strategy"	defense, offense, defender, defensive, shooting, offensive, shoot, open, guard, post	0.082
"season prospects"	final, playoff, series, won, championship, beat, winning, west, east, title	0.078
"future"	pick, trade, star, top, chance, young, future, move, round, potential	0.075
"game stats"	top, number, league, stats, mvp, average, career, assist, put, shooting	0.075

Table 2. The top five topics by LDA using all the comments in /r/NBA. The top ten weighted words are presented for each topic. In preprocessing, all team and player names are removed. The remaining words are converted to lower case and lemmatized before training the LDA model.

opinion," "game strategy," "season prospects," "future," and *"game stats."* We describe our preprocessing procedure and present the other ten topics in Section A.1.

4 RESEARCH QUESTIONS AND HYPOTHESES

We study three research questions to understand how team performance affects fan behavior in online fan communities. The first one is concerned with team performance in a single game and that game's associated user activity, while the other two questions are about team performance in a season and community properties (fan loyalty and the topics of discussion).

4.1 Team Performance and Game-level Activity

An important feature of NBA-related subreddits is to support game related discussion. In practice, each game has a game thread in the home-team subreddit, the away-team subreddit, and the overall /r/NBA. Team performance in each game can have a short-term impact on fans' behavior. For instance, Leung et al. [32] show that losing games has a negative impact on the contributions to the corresponding team's Wikipedia page, but winning games does not have a significant effect. However, it remains an open question how team performance in games relate to user activity in *online sports fan communities*.

Previous studies find that fans react differently to top teams than to bottom teams based on interviews and surveys [12, 53, 72]. In particular, Doyle et al. [12] find that fans of teams with an overwhelming loss to win ratio can be insensitive to losses through interviews. In contrast, fans that support top teams may be used to winning. The hype created by the media and other fans elevates the expectation in the fan community. As a result, losing can surprise fans of the top teams and lead to a heated discussion. Therefore, we formulate our first hypothesis as follows:

H1: In subreddits of the top teams, fans are more active on losing days; in subreddits of the bottom teams, fans are more active on winning days.

4.2 Team Performance and Fan Loyalty

Researchers in sports management show that a team's recent success does not necessarily lead to a loyal fan base [4, 6, 51]. For instance, "bandwagon fan" refers to individuals who become fans of a team simply because of their recent success. These fans tend to have a weak attachment to the

team and are ready to switch to a different team when the team starts to perform poorly. On the contrary, in the bottom teams, active fans that stay during adversity are probably loyal due to their deep attachment to the team [6, 12]. They are able to endure current stumbles and treat them as a necessary process for future success. Our second hypothesis explores the relation between team performance and fan loyalty:

H2: Top team subreddits have lower fan loyalty and bottom team subreddits have higher fan loyalty.

4.3 Team Performance and Topics of Discussion

In addition to whether fans stay loyal, our final question examines what fans talk about in an online fan community. As a popular sports quote says, "*Winning isn't everything, it's the only thing.*"[7] A possible hypothesis is that that the discussion concentrates on winning and team success. However, we recognize the diversity across teams depending on team performance. Several studies find that fans of teams with poor performance may shift the focus from current failure to future success: staying optimistic can help fans endure adversity and maintain a positive group identity [6, 12, 24]. This is in clear contrast with the focus on winning the final championship of the top teams [6]. As a result, we formulate our third hypothesis as follows:

H3: The topics of discussion in team subreddits vary depending on team performance. Top team subreddits focus more on *"season prospects"*, while bottom team subreddits focus more on *"future"*.

5 METHOD

In this section, we first provide an overview of independent variables and then discuss dependent variables and formulate hierarchical regression analyses to test our hypothesis in each research question.

5.1 Independent Variables

To understand how team performance affects fan behavior in online fan communities, we need to control for factors such as a team's market value and average player age. We collect statistics of the NBA teams from the following websites: Fivethirtyeight, [8] Basketball-Reference, [9] Forbes. [10] and Wikipedia. [11] We standardize all independent variables for linear regression models. Table 3 provides a full list of all variables used in this paper. In addition to control variables that capture the differences between seasons and months, the variables can be grouped in three categories: performance, game information, and team information.

5.1.1 Performance. Since our research questions include both team performance in a single game and team performance over a season, we consider performance variables both for a game and for a season. First, to measure a team's game performance, we simply use whether this team wins or loses. Second, to measure a team's performance over a season, we use elo ratings of the NBA teams. The elo rating system was originally invented as a chess rating system for calculating the relative skill levels of players. The popular forecasting website FiveThirtyEight developed an elo rating system to measure the skill levels of different NBA teams [47]. These elo ratings are used

[7] Usually attributed to UCLA football coach Henry Russel Sanders.
[8] http://fivethirtyeight.com/.
[9] https://www.basketball-reference.com/.
[10] https://www.forbes.com/.
[11] https://www.wikipedia.org/.

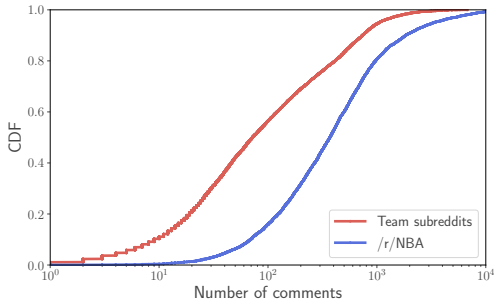
Variable	Definition	Source
Performance		
winning	Win or lose a game.	FiveThirtyEight
season elo	A team's elo rating at the end of a season.	FiveThirtyEight
season elo difference	A team's elo rating difference between the end of a season and its last season.	FiveThirtyEight
month elo	A team's elo rating at the end of that month.	FiveThirtyEight
month elo difference	A team's elo rating difference between the end of a month and its last month.	FiveThirtyEight
Game information		
team elo	A team's elo rating before the game.	FiveThirtyEight
opponent elo	The opponent's elo rating before the game.	FiveThirtyEight
point difference	Absolute point difference of the game.	FiveThirtyEight
rivalry or not	If the opponent team is a rivalry.	Wikipedia
top team	If a team is with the five highest elo ratings at the end of a season.	FiveThirtyEight
bottom team	If a team is with the five lowest elo ratings at the end of a season.	FiveThirtyEight
Team information		
market value	Transfer fee estimated to buy a team on the market.	Forbes
average age	The average age of the roster.	Basketball-Reference
#star players	The number of players selected to play the NBA All-Star Game.	Basketball-Reference
#unique users	The number of users that made at least one post/comment in the team's subreddit.	N/A
offense	The average points scored per game.	Basketball-Reference
defense	The average points allowed per game.	Basketball-Reference
turnovers	The average turnovers per game.	Basketball-Reference
Temporal information		
season	A categorical variable to indicate the season.	N/A (control variable)
month	A categorical variable to indicate the month.	N/A (control variable)

Table 3. List of variables and their corresponding definitions and sources. Measurements of team performance are in bold.

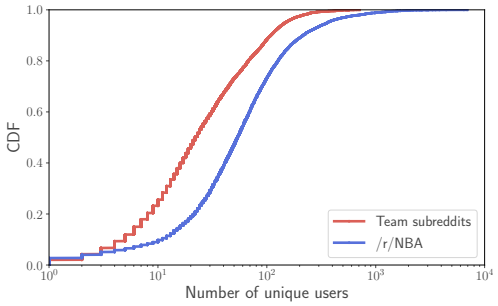
to predict game results on FiveThirtyEight and are well received by major sports media, such as ESPN [12] and CBS Sports. [13] The FiveThirtyEight elo ratings satisfy the following properties:

- A team's elo rating is represented by a number that increases or decreases depending on the outcome of a game. After a game, the winning team takes elo points from the losing one, so the system is zero-sum.
- The number of elo points exchanged after a game depends on the elo rating difference between two teams prior to the game, final basketball points, and home-court advantage. Teams gain more elo points for unexpected wins, great basketball point differences, and winning away games.
- The long-term average elo rating of all the teams is 1500.

[12] http://www.espn.com/.
[13] https://www.cbssports.com/.



(a) The CDF of #comments in game threads.



(b) The CDF of #unique users in game threads.

Fig. 2. Figure 2a shows the cumulative distribution of the number of comments in all game threads in both /r/NBA and team subreddits. Figure 2b shows the cumulative distribution of the number of unique users in all game threads in both /r/NBA and team subreddits.

To measure team performance, we use a team's elo rating at the end of a season as well as the elo difference between the end of this season and last season. A high elo rating at the end indicates an absolute sense of strong performance; a great elo rating difference suggests that a team has been improving. In addition to studying how team performance over a season affects fan loyalty, we also include team performance over a month to check the robustness of the results.

5.1.2 Game Information. To test **H1**, we need to use the interaction between game performance and top (bottom) team so that we can capture whether a top team loses or a bottom team wins. We define *top team* as teams with the highest five elo ratings at the end of a season and *bottom team* as teams with the lowest five elo ratings at the end of a season. We also include the following variables to characterize a single game: 1) Two team's elo ratings, which can partly measure the importance of a game; 2) (Basketball) point difference, which captures how close a game is; 3) Rivalry game: which indicates known rivalry relations in the NBA, such as the Lakers and the Celtics. We collect all pairs of the NBA rivalry teams from Wikipedia [67].

5.1.3 Team Information. To capture team properties, we include a team's market value, average age of players, and the number of star players. Market value estimates the value of a team on the current market. There are three key factors that impact a team's market value, including market size, recent performance and history [10]. We collect market values of all NBA teams from Forbes. We scrape the average age of players on the roster from Basketball-Reference, which computes the average age of players at the start of Feb 1st of that season. The website chooses to calculate average age on Feb 1st because it is near the player trade deadline [2], and every team has a relatively stable roster at that time. The number of star players measures the number of players selected to play in the NBA All-Star Game [68] of that season and this information is collected from Basketball-Reference. We further include variables that characterize a team's playing style: 1) Offense: the average points scored per game; 2) Defense: the average points allowed per game; 3) Turnovers: the average number of turnovers per game. Teams' playing style information by season is collected from Basketball-Reference.

5.2 Analysis for H1

Online fan communities provide a platform for fans to discuss sports games in real time and make the game watching experience interactive with other people on the Internet. Accordingly, every

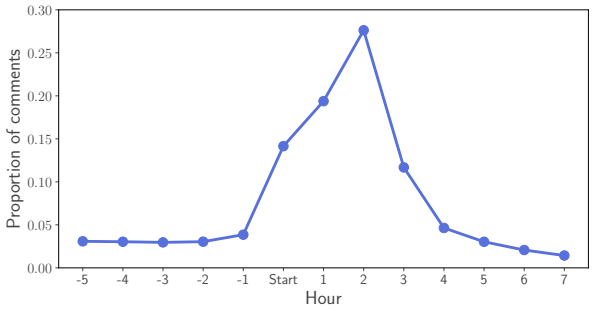


Fig. 3. The average proportion of comments made in each team subreddit by hour on the game day during the 2017 season (normalized based on game's starting hour). Error bars represent standard errors and are too small to see in the figure. Comment activity increases and peaks at the second hour after the game starts, as a typical NBA game takes around 2.5 hours.

team subreddit posts a "Game Thread" before the start of a game. Fans are encouraged to make comments related to a game in its game thread. Figure 2 shows the cumulative distributions of the number of comments and the number of unique users. A game thread can accumulate hundreds or thousands of comments. The number of comments is usually significantly higher during game time than other time periods. Figure 3 shows the average proportion of comments made in each team subreddit by hour on the game day of the 2017 season (normalized based on games' starting hour). The number of comments peaked around the game time.

We use the number of comments in game threads to capture the fan activity level for a game. Most game threads used titles that are similar to this format: "[Game Thread]: team 1 @ team 2".[14] We detected 8,596 game threads in team subreddits and 6,277 game threads in /r/NBA based on regular expression matching. Since NBA-related subreddits allow any fan to create game threads, titles of game threads do not follow the same pattern, especially in the earlier times of team subreddits. A detailed explanation and sanity check is presented in Section A.2.

Hierarchical regression analysis was used to analyze the effect of team performance in a single game on fan activity. Our full linear regression model to test **H1** is shown below:

$$\begin{aligned} \textit{\#comments in game thread} \sim & \beta_0 + \beta_s\ \textit{season} + \beta_m\ \textit{month} + \beta_t\ \textit{top team} + \beta_b\ \textit{bottom team} \\ & + \beta_1\ \textit{winning} + \beta_2\ \textit{top team winning} + \beta_3\ \textit{top team losing} \\ & + \beta_4\ \textit{bottom team winning} + \beta_5\ \textit{bottom team losing} \\ & + \beta_6\ \textit{team elo} + \beta_7\ \textit{opponent elo} + \beta_8\ \textit{rivalry or not} + \beta_9\ \textit{point difference} \\ & + \beta_{10}\ \textit{market value} + \beta_{11}\ \textit{average age} + \beta_{12}\ \textit{\#star players} + \beta_{13}\ \textit{\#unique users} \\ & + \beta_{14}\ \textit{offense} + \beta_{15}\ \textit{defense} + \beta_{16}\ \textit{turnovers}. \end{aligned} \tag{1}$$

To test our hypothesis in team subreddits, all the variables in Equation 1 are included. Unlike game threads in team subreddits, game threads in /r/NBA involve two teams and the following variables are ill-defined: "winning," "offense," "defense," and "turnovers." Therefore, these variables are removed when testing our hypothesis on game threads in /r/NBA.

[14] If more than one game thread is created for the same game, only the first one is kept, and the others are deleted by the moderator.

5.3 Analysis for H2

Fan loyalty refers to people displaying recurring behavior and a strong positive attitude towards a team [13]. To examine the relationship between team performance and fan loyalty in team subreddits, we first define active users as those that post or comment in a team subreddit during a time period. We then define two measurements of fan loyalty: *seasonly user retention* and *monthly user retention*. Seasonly user retention refers to the proportion of users that remain active in season $s + 1$ among all users that are active in season s. Monthly user retention refers to the proportion of users that remain active in month $m + 1$ among all users that are active in month m. The full linear regression models to test **H2** are shown below:

$$\begin{aligned} \textit{seasonly user retention} \sim & \beta_0 + \beta_s \textit{ season} \\ & + \beta_1 \textit{ season elo} + \beta_2 \textit{ season elo difference} \\ & + \beta_3 \textit{ market value} + \beta_4 \textit{ average age} + \beta_5 \textit{ \#star players} + \beta_6 \textit{ \#unique players} \\ & + \beta_7 \textit{ offense} + \beta_8 \textit{ defense} + \beta_9 \textit{ turnovers}. \end{aligned} \tag{2}$$

$$\begin{aligned} \textit{monthly user retention} \sim & \beta_0 + \beta_s \textit{ season} + \beta_m \textit{ month} \\ & + \beta_1 \textit{ month elo} + \beta_2 \textit{ month elo difference} \\ & + \beta_3 \textit{ market value} + \beta_4 \textit{ average age} + \beta_5 \textit{ \#star players} + \beta_6 \textit{ \#unique players} \\ & + \beta_7 \textit{ offense} + \beta_8 \textit{ defense} + \beta_9 \textit{ turnovers}. \end{aligned} \tag{3}$$

5.4 Analysis for H3

Among the five topics listed in Table 2, *"season prospects"* and *"future"* topics are closely related to our hypotheses about fans talking about winning and framing the future. By applying the trained LDA model to comments in each team subreddit, we are able to estimate the average topic distribution of each team subreddit by season. Our full linear regression model to test **H3** is shown below:

$$\begin{aligned} \textit{topic weight} \sim & \beta_0 + \beta_s \textit{ season} \\ & + \beta_1 \textit{ season elo} + \beta_2 \textit{ season elo difference} \\ & + \beta_3 \textit{ market value} + \beta_4 \textit{ average age} + \beta_5 \textit{ \#star players} + \beta_6 \textit{ \#unique players} \\ & + \beta_7 \textit{ offense} + \beta_8 \textit{ defense} + \beta_9 \textit{ turnovers}, \end{aligned} \tag{4}$$

where *topic weight* can be the average topic weight of either *"season prospects"* or *"future."*

6 RESULTS

Based on the above variables, our results from hierarchical regression analyses by and large validate our hypotheses. Furthermore, we find that the average age of players on the roster consistently plays an important role in fan behavior, while it is not the case for market value and the number of star players.

Variable	Team subreddits			/r/NBA		
	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3
Control: season						
2014	0.011***	0.012***	0.013***	0.007***	0.007***	0.005***
2015	0.032***	0.032***	0.045***	0.010***	0.010***	0.006***
2016	0.046***	0.046***	0.062***	0.014***	0.015***	0.012***
2017	0.067***	0.068***	0.081***	0.018***	0.019***	0.016***
Control: top/bottom team						
top team		0.012***	0.012***		0.012***	0.020***
bottom team		−0.012***	−0.007***		−0.009***	−0.007**
Performance						
winning			0.003**			–
top team winning			−0.006***			−0.018***
top team losing			0.006***			0.018***
bottom team winning			0.007***			0.006***
bottom team losing			−0.011***			−0.003*
Game information						
team elo			0.083***			0.091***
opponent elo			0.070***			0.111***
rivalry or not			0.010***			0.008***
point difference			−0.017***			−0.010***
Team information						
market value			0.051***			0.013***
average age			−0.067***			−0.015*
#star players			0.040***			0.020***
#unique users			0.058***			0.017***
offense			0.023**			–
defense			−0.012**			–
turnovers			0.084***			–
intercept	−0.010**	−0.011**	0.085***	0.001	−0.004**	−0.156***
Adjusted R^2	0.236	0.286	0.440	0.302	0.338	0.644
Intraclass Correlation (Season) [71]	0.087	–	–	0.021	–	–

Table 4. Hierarchical regression analyses for game-level activity in team subreddits and /r/NBA. Month is also added as a control variable for each model. **Throughout this paper, the number of stars indicate p-values, ***: $p < 0.001$, **: $p < 0.01$ *: $p < 0.05$.** We report p-values without the Bonferroni correction in all the regression tables. In Section A.3, we report F-test results with the null hypothesis that adding team performance variables does not provide a significantly better fit and reject the null hypothesis after the Bonferroni correction.

6.1 How does Team Performance Affect Game-level Activity? (H1)

Consistent with **H1**, regression results show that the top team losing and the bottom team winning correlate with higher levels of fan activity in both team subreddits and /r/NBA. Table 4 presents the results of our hierarchical regression analyses. The R^2 value is 0.40 for team subreddits and 0.63 for /r/NBA, suggesting that our linear variables can reasonably recover fan activity in game threads. Overall, fans are more active when their team wins in team subreddits (remember that the notion of one's team does not hold in /r/NBA). The interaction with the top team and the bottom

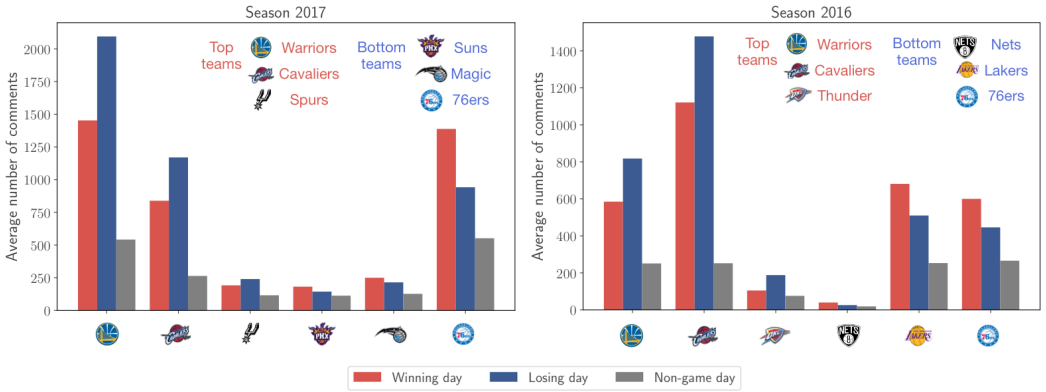


Fig. 4. Average number of comments on winning, losing and non-game days for the top three and the bottom three teams in the 2017 (left) and 2016 (right) regular season. In all the top and bottom teams, the average number of comments on game days is significantly higher than non-game days. In all top teams, the average number of comments on losing days is higher than winning days, while bottom teams show the opposite trend.

team show that surprise can stimulate fan activity: both the top team losing and the bottom team winning have significantly positive coefficients. To put this into context, in the 2017 season, the average winning percentage of the top five teams is 69%. Fans of the top teams may get used to their teams winning games, in which case losing becomes a surprise. On the contrary, the average winning percentage of the bottom five teams is 31%. It is invigorating for these fans to watch their team winning. The extra excitement can stimulate more comments in the game threads in both team subreddits and /r/NBA. In comparison, when top teams win or bottom teams lose, fans are less active, evidenced by the negative coefficient in team subreddits (not as statistically significant in /r/NBA).

To further illustrate this contrast, Figure 4 shows the average number of comments on winning, losing, and non-game days for the top three and the bottom three teams in the 2017 and 2016 regular season. Consistent patterns arise: 1) In all top and bottom teams, the average number of comments on game days is significantly higher than non-game days; 2) In all top teams, the average number of comments on losing days is higher than winning days, but bottom teams show exactly the opposite trend. Our results differ from that of Leung et al. [32], which finds that unexpected winning does not have a significant impact on Wikipedia page edits. One of the primary differences between our method and theirs is that they did not specifically control the effect of top/bottom team. It may also be explained by the fact that Wikipedia page edits do not capture the behavior of most fans and are much more sparse than comments in online fan communities. Online fan communities provide rich behavioral data for understanding how team performance affects fan behavior. The number of fans involved in our dataset is much higher than that in their Wikipedia dataset. A comparison between fans' behavior on Reddit and Wikipedia could be an interesting direction for future research. In addition, game information and team information also serve as important factors. Among variables about game information, point difference is negatively correlated with game-level user activity, as the game intensity tends to be higher when the point difference is small (a close game). Better teams (with higher elo ratings) playing against better teams or rivalry teams correlates with higher user activity levels. As for team information, a team's market value, the number of unique users, and the number of star players are positively correlated with the number of comments, since these

Variable	Seasonly user retention		Monthly user retention	
	Reg. 1	Reg. 2	Reg. 1	Reg. 2
Control: season				
2014	0.237***	0.237***	0.086***	0.055***
2015	0.184***	0.187***	0.126***	0.126***
2016	0.088***	0.116***	0.130***	0.139***
2017	0.073***	0.090***	0.137***	0.141***
Performance				
season elo		−0.370**		–
season elo difference		0.229***		–
month elo		–		−0.170**
month elo difference		–		0.032**
Team information				
market value		0.068*		0.051**
average age		−0.105*		−0.021
#star players		−0.038		0.041
#unique users		0.181***		0.111***
offense		−0.168		0.004
defense		−0.077		−0.053
turnovers		−0.037		−0.041
intercept	0.583***	0.629***	0.478***	0.460***
Adjusted R^2	0.286	0.503	0.155	0.232
Intraclass Correlation (Season) [71]	0.396	–	0.029	–

Table 5. Hierarchical regression analyses for seasonly user retention rate and monthly user retention rate in team subreddits. Month is also added as a control variable for the monthly user retention analysis. *#unique users* is counted every season for the seasonly user retention analysis and every month for the monthly user retention analysis. For both dependent variables, team's overall performance has a negative coefficient while short-term performance and market value has a positive coefficient.

two factors are closely related to the number of fans. Younger teams with more average points scored and less points allowed per game stimulate more discussion in team subreddits.

6.2 How does Team Performance Relate to Fan Loyalty in Team Subreddits? (H2)

Our findings confirm **H2**, that top teams tend to have lower fan loyalty and bottom teams tend to have higher fan loyalty, measured by both seasonly user retention and monthly user retention. Table 5 shows the hierarchical regression results. In both regression analyses, elo rating, which measures a team's absolute performance, has a statistically significant negative impact on user retention rate. The coefficient of elo rating also has the greatest absolute value among all variables (except intercept). Meanwhile, improved performance reflected by elo difference positively correlates with user retention.

Figure 5 presents the seasonly user retention rate and average monthly user retention rate of the top 3 and bottom 3 teams in the 2017 (left) and 2016 (right) season. It is consistent that in these two seasons, bottom teams have higher user retention rate than top teams, both seasonly and monthly. This may be explained by the famous "bandwagon" phenomenon in professional sports [62]: Fans may "jump on the bandwagon" by starting to follow the current top teams, which provides a short cut to achievement and success for them. In comparison, terrible team performance can serve as a loyalty filter. After a period of poor performance, only die-hard fans stay active and optimistic

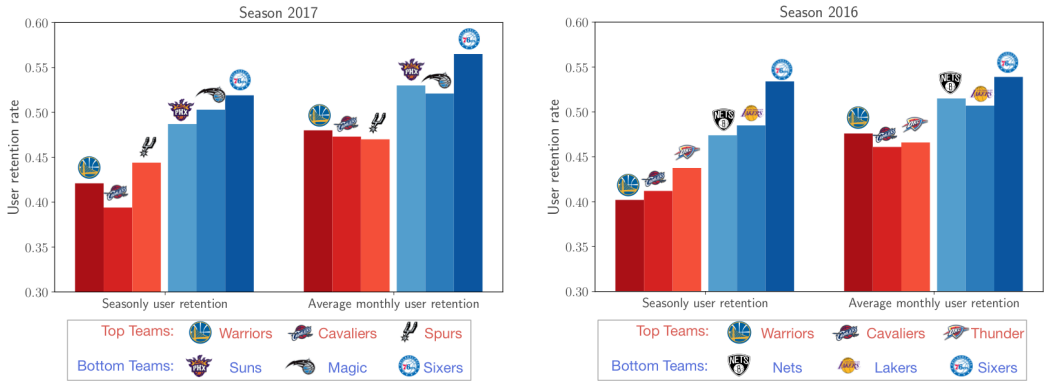


Fig. 5. Seasonly user retention rate and average monthly user retention rate of the top three and bottom three teams in the 2017 (left) and 2016 (right) season. Bottom teams consistently have higher user retention than top teams.

in the team subreddits. Our results echo the finding by Hirt et al. [20]: after developing strong allegiances with a sports team, fans find it difficult to disassociate from the team, even when the team is unsuccessful. It is worth noting that the low fan loyalty of the top teams cannot simply be explained by the fact that they tend to have more fans. In fact, teams with higher market value and more unique users (more fans) tend to have a higher user retention rate, partly because their success depends on a healthy and strong fan community.

Similar to game-level activity, fans are more loyal to younger teams, at least in seasonly user retention (the coefficient is also negative for monthly user retention and $p-$value is 0.07). Surprisingly, according to our hierarchical regression results, a team's number of star players and playing style (offense, defense, and turnovers) have no significant impact on user retention.

6.3 How does Team Performance Affect Topics of Discussion in Team Subreddits? (H3)

Our final question is concerned with the relation between team performance and topics of discussion in online fan communities. Our results validate **H3**, that better teams have more discussions on *"season prospects"* and worse teams tend to discuss *"future."* Table 6 presents the results of hierarchical regression analyses on *"future"* topic weight and *"season prospects"* topic weight computed with our LDA model. In both regressions, only team performance (season elo) and average age have statistically significant coefficients. Both team performance and average age are positively correlated with *"season prospects"* and negatively correlated with *"future"*. Moreover, the number of star players has a negative correlation with *"future"* but has no significant effect on *"season prospects."* Despite having only two or three variables (except control variables and intercept) with significant coefficients, both regression analyses achieve strong correlation with R^2 above 0.57. Note that the improvement in team performance (season elo difference) does not have a significant effect.

As an example, Figure 6 further shows topic weights of *"future"* and *"season prospects"* for all the teams in the 2017 and 2016 season. The top 3 teams and bottom 3 teams in each season are highlighted using team logos. The top teams are consistently in the lower right corner (high *"season prospects"*, low *"future"*), while the bottom teams are in the upper left corner (low *"season prospects"*, high *"future"*). Our results echo the finding in Doyle et al. [12]: framing the future is an important strategy for fans of teams with poor performance to maintain a positive identity in the absence of success.

Variable	*"season prospects"* Reg. 1	*"season prospects"* Reg. 2	*"future"* Reg. 1	*"future"* Reg. 2
Control: season				
2014	0.096***	0.056**	0.059*	0.137***
2015	0.067**	0.049*	0.054*	0.129***
2016	0.069**	0.053*	0.109***	0.175***
2017	0.061*	0.048*	0.065*	0.160***
Performance				
season elo		0.410***		−0.415**
season elo difference		−0.130		−0.099
Team information				
market value		−0.065		0.051
average age		0.149***		−0.189***
#star players		0.018		−0.187**
#unique users		0.071		−0.104
offense		−0.036		0.037
defense		−0.120		−0.115
turnovers		0.059		−0.092
intercept	0.398***	0.286***	0.478***	0.814***
Adjusted R^2	0.013	0.578	0.002	0.619
Intraclass Correlation (Season) [71]	0.059	–	0.003	–

Table 6. Hierarchical regression analyses for *"season prospects"* topic weight and *"future"* topic weight in team subreddits. Team performance has positive correlation with *"season prospects"* topic and negative correlation with *"future"* topic.

The effect of average age reflects the promise that young talents hold for NBA teams. Although it takes time for talented rookies that just come out of college to develop physical and mental strength to compete in the NBA, fans can see great potential in them and remain positive about their team's future, despite the team's short-term poor performance. In contrast, veteran players are expected to bring immediate benefits to the team and compete for playoff positions and even championships. For example, Rothstein [44] lists a number of veteran players who either took a pay cut or accepted a smaller role in top teams to chase a championship ring at the end of their career.

A team's playing style, including offense points, defense points, and turnovers, doesn't seem to influence the topic weights of these two topics. We also run regression for the other three top topics in Table 2 and present the results in Section A.4. Team performance plays a limited role for the other three topics, while average age is consistently significant for all three discussion topics.

7 CONCLUDING DISCUSSION

In this work, we provide the first large-scale characterization of online fan communities of the NBA teams. We build a unique dataset that combines user behavior in NBA-related subreddits and statistics of team performance. We demonstrate how team performance affects fan behavior both at the game level and at the season level. Fans are more active when top teams lose and bottom teams win, which suggests that in addition to simply winning or losing, surprise plays an important role in driving fan activity. Furthermore, a team's strong performance doesn't necessarily make the fan community more loyal. It may attract "bandwagon fans" and result in a low user retention rate. We find that the bottom teams generally have higher user retention rate than the top teams. Finally,

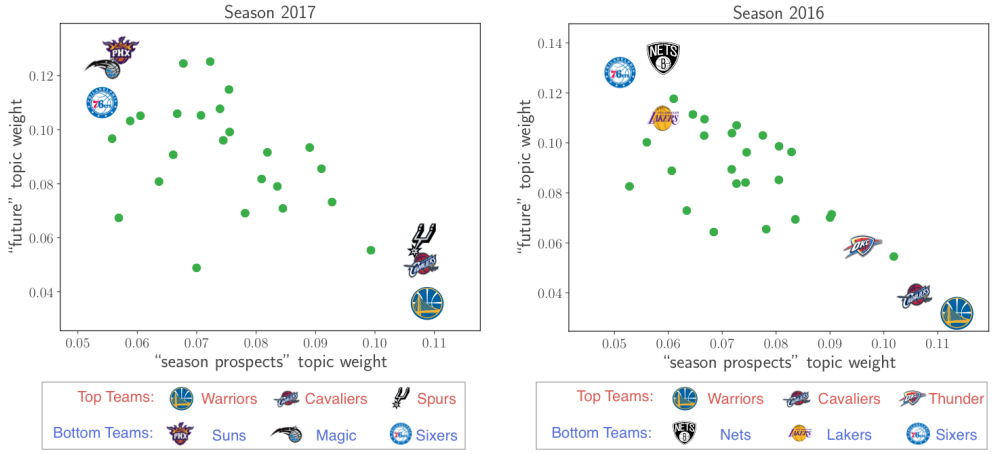


Fig. 6. Scatterplot of *"season prospects"* topic weight and *"future"* topic weight in all the team subreddits in the 2017 (left) and 2016 (right) season. The top three teams and bottom three teams are represented by team logos instead of points. Teams are ranked by elo rating at the end of each season. Fans of the top teams tend to discuss much more *"season prospects"* topics (lower right corner) and fans of the bottom teams tend to discuss much more *"future"* topics (upper left corner).

fans of the top teams and bottom teams focus on different topics of discussion. Fans of the top teams talk more about season records, playoff seeds, and winning the championship, while fans of the bottom teams spend more time framing the future to compensate for the lack of recent success.

Limitations. One key limitation of our work is the representativeness of our dataset. First, although our study uses a dataset that spans five years, our period coincides with the rapid growth of the entire Reddit community. We use *season* and *month* to try our best to account for temporal differences, but our sample could still be based upon fans with a mindset of growth. Second, although Goldschein [17] suggests that /r/NBA is now playing an important role among fans, the NBA fan communities on Reddit may not be representative of the Internet and the whole offline population.

Another limitation of our work lies in our measurement. For game-level activity, we only consider the number of comments in the game threads. This measurement provides a nice way to make sure that the comments are about the game, but we may have missed related comments in other threads. We do not consider other aspects of the comments such as sentiment and passion. In addition, our fan loyalty metric is entirely based on user retention. A user who posts on a team subreddit certainly supports the team to a different extent from those who do not. Our metric may fail to capture lurkers who silently support their teams. Finally, our topics of discussion are derived from topic modeling, an unsupervised approach. Supervised approaches could provide more accurate identification of topics, although the deduction approach would limit us to a specific set of topics independent of the dataset.

Implications for online communities. First, our work clearly demonstrates that online communities do not only exist in the virtual world; they are usually embedded in the offline context and attract people with similar offline interests. It is an important research question to understand to what extent and how online communities relate to offline contexts as well as what fraction of online communities are entirely virtual. Professional sports provide an interesting case, because these online fan communities, in a way, only exist as a result of the offline sports teams and games. Such connections highlight the necessity to combine multiple data sources to understand how fans' usage

of social media correlates with the on-going events of the topic of their interests. Our study has the potential to serve as a window into the relationship between online social behavior and offline professional sports. We show that subreddit activity has significant correlations with game results and team properties. Exploring the factors that motivate users of interest-based communities to communicate with social media is also an important and rich area for future research. For example, a promising future direction is to study the reasons behind fans departing a team subreddit. Possible reasons include being disappointed by the team performance or playing style, favorite players being traded, and being attacked by other fans in the team subreddit or /r/NBA.

Second, our results show that teams with strong performance correlate with low fan loyalty. These results relate to the multi-community perspective in online community research [19, 55, 74, 75]. One future direction is to examine where fans migrate to and whether fans leave the NBA or the Reddit altogether, and more importantly, what factors determine such migration decisions.

Third, our findings reveal strategies for the design of sports-related online platforms. Our results clearly demonstrate that teams in under-performing periods are more likely to develop a more loyal fan base that discusses more about their team's *"future."* Recognizing these loyal fans and acknowledging their contributions within the fan community can be critical for facilitating attraction and retention of these fans. For example, team subreddits' moderators may reward a unique flair to the users who have been active in the community for a long time, especially during the difficult times.

Implications for sports management. Our findings suggest that winning is not everything. In fact, unexpected losses can stimulate fan activity. The increase of fan activity does not necessarily happen in a good way. For example, the fans of the Cavaliers, which won the Eastern Championship of the 2017 season, started to discuss firing the team's head coach Tyronn Lue after losing three of the first six games in the following season. Managers may try to understand the role of expectation in fan behavior and guide the increased activity and attention towards improving the team and building a strong fan base.

We also find that the average age of the roster consistently plays an important role in fan behavior: younger teams tend to bring more fan activity on game days and develop a more loyal fan base that discusses about *"future."* These results contribute to existing literature on the effect of age in sports management. Timmerman [56] finds that the average age is positively correlated with team performance, while the age diversity is negatively correlated (in other words, veterans improve team performance but are not necessarily compatible with young players). The tradeoff between veteran players and young talents requires more research from the perspective of both team performance and fan engagement.

Finally, it is crucial for teams to maintain a strong fan base that can support them during unsuccessful times because it is difficult for sports teams to sustain winning for a long time. This is especially true in the NBA since the draft lottery mechanism is designed to give bottom teams opportunities to improve and compete. Consistent with Doyle et al. [12], we find that framing the *"future"* can be an important strategy for teams with poor performance to maintain a positive group identity. The absence of success can be a great opportunity to develop a deep attachment with loyal fans. Prior studies show that certain fan group would like to persevere with their supported team through almost anything, including years of defeat, to recognize themselves as die-hard fans. By doing this, they feel that they would reap more affective significance among the fan community when the team becomes successful in the future [22, 62]. It is important for managers to recognize these loyal fans and create ways to acknowledge and leverage their positions within the fan community. For instance, teams may host "Open Day" and invite these loyal fans to visit facilities and interact with star players and coaching staff. Hosting Ask Me Anything (AMA) [69] interviews is another strategy to engage with online fan communities.

REFERENCES

[1] Dan Ariely. 2008. *Predictably irrational.* HarperCollins New York.

[2] National Basketball Association. 2017. *NBA Official Rules 2017-18.* New York, NY, USA. https://turnernbahangtime.files.wordpress.com/2017/10/2017-18-nba-rule-book.pdf

[3] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of KDD.* ACM, 44–54.

[4] Colleen C Bee and Mark E Havitz. 2010. Exploring the relationship between involvement, fan attraction, psychological commitment and behavioural loyalty in a sports spectator context. *International Journal of Sports Marketing and Sponsorship* 11, 2 (2010), 37–54.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[6] Richard M Campbell Jr, Damon Aiken, and Aubrey Kent. 2004. Beyond BIRGing and CORFing: Continuing the Exploration of Fan Behavior. *Sport Marketing Quarterly* 13, 3 (2004).

[7] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages.

[8] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: identifying abusive behavior online with preexisting internet data. In *Proceedings of CHI.* ACM, 3175–3187.

[9] Marci D Cottingham. 2012. Interaction ritual theory and sports fans: Emotion, symbols, and solidarity. *Sociology of Sport Journal* 29, 2 (2012), 168–185.

[10] Jeff Desjardins and Visual Capitalist. 2017. The world's 50 most valuable sports teams. http://www.businessinsider.com/the-worlds-50-most-valuable-sports-teams-2017-8. (2017). [Online; accessed 15-April-2018].

[11] Peter Dolton and George MacKerron. 2018. Is football a matter of life and death–or is it more Important than that? *NIESR Discussion Paper* (2018).

[12] Jason P Doyle, Daniel Lock, Daniel C Funk, Kevin Filo, and Heath McDonald. 2017. 'I was there from the start': The identity-maintenance strategies used by fans to combat the threat of losing. *Sport Management Review* 20, 2 (2017), 184–197.

[13] Brendan Dwyer. 2011. Divided loyalty? An analysis of fantasy football involvement and fan loyalty to individual National Football League (NFL) teams. *Journal of Sport Management* 25, 5 (2011), 445–457.

[14] Bleacher Report Adam Fromal. 2017. Ranking Every NBA Franchise on Historical Success. http://bleacherreport.com/articles/2733544-ranking-every-nba-franchise-on-historical-success. (2017). [Online; accessed 15-April-2018].

[15] Katie Z. Gach, Casey Fiesler, and Jed R. Brubaker. 2017. "Control Your Emotions, Potter": An Analysis of Grief Policing on Facebook in Response to Celebrity Death. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 47 (Dec. 2017), 18 pages.

[16] Devin Gaffney and J. Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE* 13, 7 (07 2018), 1–13.

[17] Eric Goldschein. 2015. It's Time to Give /r/nba the Respect it Deserves. https://www.sportsgrid.com/as-seen-on-tv/media/its-time-to-give-rnba-the-respect-it-deserves/. (2015). [Online; accessed 15-April-2018].

[18] The Stanford Natural Language Processing Group. 2010. Stanford Topic Modeling Toolbox. https://nlp.stanford.edu/software/tmt/tmt-0.4/. (2010). [Online; accessed 15-April-2018].

[19] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in Online Communities. In *Proceedings of ICWSM.* AAAI, 377–386.

[20] Edward R Hirt, Dolf Zillmann, Grant A Erickson, and Chris Kennedy. 1992. Costs and benefits of allegiance: Changes in fans' self-ascribed competencies after team victory versus defeat. *Journal of personality and social psychology* 63, 5 (1992), 724.

[21] Vincent Hogan, Patrick Massey, and Shane Massey. 2017. Analysing match attendance in the European Rugby Cup: Does uncertainty of outcome matter in a multinational tournament? *European Sport Management Quarterly* 17, 3 (2017), 312–330.

[22] Craig G Hyatt and William M Foster. 2015. Using identity work theory to understand the de-escalation of fandom: A study of former fans of National Hockey League teams. *Journal of Sport Management* 29, 4 (2015), 443–460.

[23] Giulio Jacucci, Antti Oulasvirta, Antti Salovaara, and Risto Sarvas. 2005. Supporting the shared experience of spectators through mobile group media. In *Proceedings of SIGGROUP.* ACM, 207–216.

[24] Ian Jones. 2000. A model of serious leisure identification: The case of football fandom. *Leisure Studies* 19, 4 (2000), 283–298.

[25] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. 2012. The life and death of online groups: Predicting Group Growth and Longevity. In *Proceedings of WSDM.* ACM, 673–682.

[26] Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Staying in the loop: structure and dynamics of Wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM, 1.
[27] Brian Keegan, Darren Gergle, and Noshir Contractor. 2013. Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist* 57, 5 (2013), 595–622.
[28] Brian C Keegan and Jed R Brubaker. 2015. 'Is' to'Was': Coordination and Commemoration in Posthumous Activity on Wikipedia Biographies. In *Proceedings of CSCW*. ACM, 533–546.
[29] Amy Jo Kim. 2000. *Community Building on the Web: Secret Strategies for Successful Online Communities* (1st ed.). Addison-Wesley Longman Publishing Co., Inc.
[30] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of WWW*. International World Wide Web Conferences Steering Committee, 933–943.
[31] Alex Leavitt and Joshua A Clark. 2014. Upvoting hurricane Sandy: event-based news production processes on a social news site. In *Proceedings of CHI*. ACM, 1495–1504.
[32] Weiwen Leung, Haiyi Zhu, and Joseph A. Konstan. 2017. The Effect of Emotional Cues from the NFL on Wikipedia Contributions. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 66 (Dec. 2017), 21 pages.
[33] Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael S Bernstein. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth.. In *Proceedings of ICWSM*. AAAI, 132–141.
[34] Martin Ludvigsen and Rune Veerasawmy. 2010. Designing technology for active spectator experiences at sporting events. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. ACM, 96–103.
[35] J Nathan Matias. 2016. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of CHI*. ACM, 1138–1151.
[36] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proceedings of ICWSM*. AAAI, 279–288.
[37] Leysia Palen and Kenneth M Anderson. 2016. Crisis informatics: New data for extraordinary times. *Science* 353, 6296 (2016), 224–225.
[38] David Paton and Andrew Cooke. 2005. Attendance at county cricket: An economic analysis. *Journal of Sports Economics* 6, 1 (2005), 24–45.
[39] Max Rappaport. 2018. The Definitive History of 'Trust the Process'. http://bleacherreport.com/articles/2729018-the-definitive-history-of-trust-the-process. (2018). [Online; accessed 15-April-2018].
[40] Redditlist. 2018. Most Active Subreddits. http://redditlist.com/all. (2018). [Online; accessed 15-April-2018].
[41] Yuqing Ren, Robert Kraut, and Sara Kiesler. 2007. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies* 28, 3 (2007), 377–408.
[42] Alan Roadburg. 1980. Factors precipitating fan violence: A comparison of professional soccer in Britain and North America. *British Journal of Sociology* (1980), 265–276.
[43] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. 2016. Social networks under stress. In *Proceedings of WWW*. International World Wide Web Conferences Steering Committee, 9–20.
[44] Matthew Rothstein. 2017. For Aging Veterans, One Last Shot At An NBA Title Is All They Can Think About. https://uproxx.com/dimemag/ring-chasing-richard-jefferson-shane-battier/. (2017). [Online; accessed 15-April-2018].
[45] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In *First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)*.
[46] Matthew D Shank and Mark R Lyberger. 2014. *Sports marketing: A strategic perspective*. Routledge.
[47] Nate Silver and Reuben Fischer-Baum. 2018. How We Calculate NBA Elo Ratings. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/. (2018). [Online; accessed 15-April-2018].
[48] Aaron CT Smith and Bob Stewart. 2007. The travelling fan: Understanding the mechanisms of sport fan consumption in a sport tourism setting. *Journal of sport & tourism* 12, 3-4 (2007), 155–181.
[49] Kate Starbird, Leysia Palen, Amanda L Hughes, and Sarah Vieweg. 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of CSCW*. ACM, 241–250.
[50] Statista. 2018. Average TV viewership of NBA Finals games in the United States from 2002 to 2017 (in millions). https://www.statista.com/statistics/240377/nba-finals-tv-viewership-in-the-united-states/. (2018). [Online; accessed 15-April-2018].
[51] Shawn Stevens and Philip J Rosenberger. 2012. The influence of involvement, following sport and fan identification on fan loyalty: An Australian perspective. *International Journal of Sports Marketing and Sponsorship* 13, 3 (2012), 57–71.
[52] Leo Graiden Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. Drawing the Lines of Contention: Networked Frame Contests Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 96 (Dec. 2017), 23 pages.

[53] William A Sutton, Mark A McDonald, George R Milne, and John Cimperman. 1997. Creating and fostering fan identification in professional sports. *Sport Marketing Quarterly* 6 (1997), 15–22.
[54] Chenhao Tan. 2018. Tracing Community Genealogy: How New Communities Emerge from the Old. In *Proceedings of ICWSM*. AAAI, 395–404.
[55] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*. International World Wide Web Conferences Steering Committee, 1056–1066.
[56] Thomas A Timmerman. 2000. Racial diversity, age diversity, interdependence, and team performance. *Small Group Research* 31, 5 (2000), 592–606.
[57] Jessica Torrez Riley. 2012. A look at spectator technology: location-based services and mobile habits of collegiate sports fans. In *Proceedings of MobileHCI Companion*. ACM, 41–46.
[58] Galen T Trail and Jeffrey D James. 2001. The motivation scale for sport consumption: Assessment of the scale's psychometric properties. *Journal of sport behavior* 24, 1 (2001), 108.
[59] Marlon Twyman, Brian C Keegan, and Aaron Shaw. 2017. Black Lives Matter in Wikipedia: Collective Memory and Collaboration around Online Social Movements. In *Proceedings of CSCW*. ACM, 1400–1412.
[60] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. In *Proceedings of CHI*. ACM, 566.
[61] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of ICML*. 1105–1112.
[62] Daniel L Wann and Nyla R Branscombe. 1990. Die-hard and fair-weather fans: Effects of identification on BIRGing and CORFing tendencies. *Journal of Sport and Social issues* 14, 2 (1990), 103–117.
[63] Daniel L Wann, Frederick G Grieve, Ryan K Zapalac, and Dale G Pease. 2008. Motivational profiles of sport fans of different sports. *Sport Marketing Quarterly* 17, 1 (2008), 6.
[64] Daniel L Wann, Joel L Royalty, and Al R Rochelle. 2002. Using motivation and team identfication to predict sport fans' emotional responses to team performance. *Journal of Sport Behavior* 25, 2 (2002), 207.
[65] Pushshift Website. 2018. Reddit Dataset. https://files.pushshift.io/reddit/. (2018). [Online; accessed 15-April-2018].
[66] Lawrence A Wenner. 1989. *Media, sports, and society*. sage.
[67] Wikipedia. 2018. List of National Basketball Association rivalries. https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_rivalries. (2018). [Online; accessed 15-April-2018].
[68] Wikipedia. 2018. NBA All-Star Game — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=NBA%20All-Star%20Game&oldid=839843474. (2018). [Online; accessed 8-July-2018].
[69] Wikipedia. 2018. /r/IAmA. https://en.wikipedia.org/wiki//r/IAmA. (2018). [Online; accessed 15-April-2018].
[70] Wikipedia. 2018. Super Bowl commercials. https://en.wikipedia.org/wiki/Super_Bowl_commercials. (2018). [Online; accessed 15-April-2018].
[71] Matthew Wolak. 2018. Package 'ICC', Facilitating Estimation of the Intraclass Correlation Coefficient. https://cran.r-project.org/web/packages/ICC/ICC.pdf. (2018). [Online; accessed 17-August-2018].
[72] Masayuki Yoshida, Bob Heere, and Brian Gordon. 2015. Predicting behavioral loyalty through community: Why other fans are more important than our own intentions, our satisfaction, and the team itself. *Journal of Sport Management* 29, 3 (2015), 318–333.
[73] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. 2017. Shocking the Crowd: The Effect of Censorship Shocks on Chinese Wikipedia. In *Proceedings of ICWSM*. AAAI, 367–376.
[74] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of ICWSM*. AAAI, 377–386.
[75] Haiyi Zhu, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E Kraut. 2014. Selecting an Effective Niche: An Ecological View of the Success of Online Communities. In *Proceedings of CHI*. ACM, 301–310.

A APPENDIX

A.1 Topic Modeling and Additional Topics

The following pre-processing procedures are used to clean data before training topic models:

- Converting all the words to lower case.
- Removing all the HTML links in the comments.
- Removing all the player names and nicknames, such as Lebron, Kobe.
- Removing all the team names, such as Lakers, Celtics.
- Removing common stopwords.
- Lemmatizing all words.

LDA topic	top words	average topic weight
Topic 00	big, level, work, hard, league, talent, basketball, long, skill, playing	0.071
Topic 01	fucking, dude, day, kid, life, friend, face, talking, bitch, court	0.068
Topic 02	injury, minute, played, playing, half, quarter, end, night, ago, start	0.068
Topic 03	watch, watching, basketball, feel, fun, damn, hope, god, fucking, honestly, suck	0.067
Topic 04	call, foul, ref, throw, free, called, hand, rule, hit, foot	0.065
Topic 05	contract, money, deal, cap, million, sign, free, pay, max, salary	0.063
Topic 06	post, comment, thread, edit, read, friend, face, talking, bitch, court	0.055
Topic 07	sport, basketball, school, city, jersey, high, black, world, college, white	0.054
Topic 08	coach, bench, starting, role, system, fit, coaching, front, starter, minute	0.054
Topic 09	prime, greatest, goat, time, career, all, star, history, seasons, era	0.043

Table 7. The remaining 10 extracted topics by LDA using all comments in /r/NBA. The top ten weighted words are presented for each topic.

The remaining 10 topics generated after training are shown in Table 7.

A.2 Game Thread Matching

The percentage of game threads detected among all games by season is shown in Table 8. The percentages of game threads detected in /r/NBA are high in all seasons, with an average of 95%. The percentage of game threads detected in team subreddits are high in the 2016 and 2017 seasons, all above 80%. Significant amount of game threads are missing for the 2013 and 2014 seasons.

We further investigate the reasons why game threads are not detected in team subreddits. We randomly sampled 50 games in season 2014. If home-team game thread and away-team game thread exist for these games, 100 game threads should have been detected in total. We manually checked them in our dataset and found that 40 game threads are successfully detected, 54 game threads were not created and 6 game threads had special title formats. Based on this limited sample, our detection rate is in fact 87% (40/46). There are two major reasons to explain missing game threads: 1) Some team subreddits were relatively small and fans only created game threads for important games (e.g., games against rivalry teams and strong teams, or games that are critical for playoff spots); 2) Some Game Threads do not follow the standard format. For example, a game thread in /r/Timberwolves is titled “Last regular season game, boys travel to Houston!”.

	Team subreddits					/r/NBA				
year	2013	2014	2015	2016	2017	2013	2014	2015	2016	2017
#game threads detected	503	1176	2131	2362	2424	1106	1230	1311	1314	1305
#games	1314	1319	1311	1316	1309	1314	1319	1311	1316	1309
Percentage	19%	45%	81%	90%	93%	84%	93%	100%	100%	100%

Table 8. Percentage of game threads detected from team subreddits and /r/NBA by season. For each game, it is supposed to have one game thread in home-team subreddit, one game thread in away-team subreddit and one game thread in /r/NBA. The detected game thread is high for /r/NBA. For team subreddits, certain percentage of Game Thread are missing because: 1) Team subreddits only create game threads for important games when the community is relatively small; 2) Some game threads do not follow standard format.

	H1		H2		H3	
	Team subreddits	/r/NBA	Seasonly	Monthly	*"season prospects"*	*"future"*
F-value	26.9	51.6	15.2	12.3	16.3	20.1
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 9. F-test results. All the p-values are less than 0.0001 after the Bonferroni correction.

A.3 F-tests with Bonferroni correction

Table 9 summarizes the results of F-tests where the null hypothesis is that adding team performance variables does not provide a significantly better fit. All the p-values are less than 0.0001 after the Bonferroni correction, so we reject the null hypothesis. In all the regressions in Table 9, adding team performance variables provides a significantly better fit.

A.4 Additional Linear Regression Models for Topic Weights

Table 10 presents the results of OLS linear regression models for average topic weights of top five topics other than *"future"* and *"season prospects,"* i.e., *"personal opinion," "game strategy,"* and *"game stats."*

Received April 2018; revised July 2018; accepted September 2018

variables	*"personal opinion"* Reg. 1	*"personal opinion"* Reg. 2	*"game strategy"* Reg. 1	*"game strategy"* Reg. 2	*"game stats"* Reg. 1	*"game stats"* Reg. 2
Control: season						
2014	0.110***	0.125***	0.079***	0.082***	0.059***	0.059***
2015	0.146***	0.196***	0.096***	0.095***	0.083***	0.082***
2016	0.063**	0.107***	0.101***	0.090***	0.050***	0.088***
2017	0.152***	0.211***	0.089***	0.081***	0.104***	0.103***
Performance						
season elo		−0.380**		−0.126		−0.099
season elo difference		−0.125*		−0.024		0.044
Team information						
market value		0.080		0.015		0.014
average age		−0.121*		−0.069*		−0.053*
#star players		0.018		−0.068		−0.043
#unique users		−0.105		0.068		0.146***
offense points		0.158		0.112		−0.100
defense points		−0.213*		0.038		0.053
turnovers		−0.120		0.032		−0.079
intercept		0.781***		0.414***	0.043	0.414***
Adjusted R^2	0.048	0.208	0.008	0.251	0.025	0.334

Table 10. Hierarchical regression analyses for the other top 5 topics: *"personal opinion," "game strategy,"* and *"game stats."*