A Statistical Analysis of Cricket One Day Internationals

Aishwarya Pratap Singh (1211162781) Ayushi Jain (1204841348)
Rahul Aakunuru (1209394573) Nikhil Lohia (1211168085)

Abstract—Cricket is a game played in over a hundred countries worldwide and One Day International (ODI) is one of the most popular forms played. In this paper, we aim to find ways to effectively visualize the data generated from ODI matches. We have created four visualizations in which we explore team performance, player performance, and the relationships between the two. Our visualizations include a bar chart, a scatter plot, a histogram, and a stacked area chart. A combination of all the visualizations forms a system which aims to provide a better insight to the games by showing what kinds of players are on teams, how they contribute to the success of their teams, and how teams perform again each other.

Index Terms—Cricket, Performance, Visualization types, scatter plot, histogram

1 Introduction

Cricket is an extremely popular game around the world. To put this in perspective, there were an estimated 2.2 billion viewers watching the game during the ICC Cricket World Cup 2011. With cricket becoming extremely fast paced and increasingly competitive, performance analysis is steadily making its way into the teams dressing rooms. Data visualization and data analysis in cricket has increasingly become a critical part of the game. For instance, players can try to become technically better by analyzing their own performances. Second, teams can make plans for their upcoming matches by analyzing the performance of their competitors. Third, Selectors can analyze team performance to select appropriate teams. Finally, the International Cricket Council can analyze trends to maintain the level of the games taking place. Our project aims to further investigate many such metrics that may affect the performance of a team or the popularity of the game. We present our visualizations in a way that they tell a story to the end user and let them understand how the game has progressed over the past years, in addition to how the teams have gone about looking at their performance. We try to present a visual narrative that is "an account of a series of events, facts, etc., given in order and with the establishing of connections between them". [**?**]

2 Related Works

Several papers exist that relate to what we have studied. We will examine some in the following sections.

2.1 Baseline Bar Display for Scores and Margin

Andy Cox and John Stasko have created a baseline bar display in their sportsViz [**?**] system to visualize the scores in a baseball match and the margin by which a team has won or lost. In their system, they have functionality to sort the data by runs or margin. In cricket, we have similar statistics called runs and margin of runs by which the team can win or lose. We have used similar design to visualize the runs scored by a team. Since we have multiple visualizations in our systems, we allow user to select each bar and the other visualizations will be updated accordingly.

2.2 Scatter Plot:

Fengbo et al. designed and built an application called NBAViz [**?**] in which they visualize the performance of NBA players. In their visualization, they have used scatter plot to show the impact of players on the performance of team. For each player played in a game, they have plotted number of minutes that the player played vs number of points that the player scored. In our cricket visualization, we have similar metric to measure batting performance of team, which is the number of runs scored vs number of balls faced by a batsman. So, we have used this idea to visualize how a batsman is impacting the outcome of the game.

2.3 Brushing and plotting a Bar Graph

The Scatter plot tells how good of a batsman a team has, but it doesn't allow user to zoom in to area and filter the values. We have used Brushing and linking, which is a well-known technique in infoViz, to allow user to select and zoom an area. The principle of brushing and linking was used by Becker and Cleveland in their paper Brushing Scatter-plots [**?**]. Brushing is a process of selecting a portion of graph or plot. Often brushing is followed by linking to some other visualization in the system. We have used brushing to allow user to select a part of scatter plot and consider the players involved in that area. This brushing is linked to a bar graph in which we have a stacked bar for each player who are present in the brushed area and height of the bar represents the number of times an inning played by the player is present in the brushed area.

2.4 Stacked Area Graph

Along with team's performance and player's impact on team, it is important to look at the player's performance

over the years. We could use histogram for each player where length of each bar in histogram is proportional to the runs scored by the player and show the histogram of different players in multiple views. This kind of visualization will give the player's performance over the years, but it will be difficult for the user to compare the performance between players. Theme-river [**?**] works better for these kind of visualizations. The width of the theme is consistent with the number of runs scored by a batsman and it connects the value with runs scored in before year with a smooth curve which will be visually pleasing. The horizontal flow represents the flow of time and the color currents within the river represents the player's data. But theme river doesn't give us the percentage of runs scored a player in total runs scored by top 10 players because theme-river expands on top and bottom part of x-axis. And hence we have choose Stacked area chart [**?**] which reduces the variation is graph making it easy to read and compare the values

3 Design Principles

The amount of information that can be represented on a unit area of screen is very limited and with increasing it data becomes harder to fetch relevant insights from it. Showing all the data at once can cause confusions and make the judgment even harder. The main goal of our design is to compare team vs team performance and how each player makes an impact on the outcome. To achieve this we visualize one team at a time and then add the option to compare it with another. The process can be summarized as:

We start with raw data obtained from Cricsheet and transform it to a csv format. Since data loss is inevitable, we choose to abstract attributes explicitly based upon the target audience. Our visualization caters to 2 types of audiences

- A sports journalist who wants to analyze a teams performance
- A fan who wishes to discover trends of his/her favorite team.

Keeping it in mind, we discover the relevant insights they would be interested in. Some of them could be

- Analyze a players contribution to the teams success.
- Identify players in a certain bracket.
- Team vs Team metrics
- Compare the performance of players against different oppositions.

A good visualization is one which has interactivity and makes the comparison engaging. We present the viewer with data controls and let them uncover things on their own. Based upon these reasonings we follow 2 main design principles. Firstly, the user should be able to select the teams to compare. Secondly, the user should be able to interact with the visualization to get additional insights in his area of interest.

4 System

4.1 Data

A single game of Cricket generates a huge amount of data. One game of two teams consists of three hundred deliveries divided into fifty overs each. Each delivery can result into scoring runs ranging from zero to six scored by eleven different batsmen. We can aggregate this data at many levels along this hierarchy and create several interactive visualizations. The user is served into exploring data with respect to a particular team against an opponent. For example, we can explore the insights when Australia plays against India. We leverage the ball by ball data obtained from Cricsheet. From this data we extract the outcome of each game, teams involved, and the runs scored by each player over the years. This creates a large collection of data for several players and teams. For example, we get a data on over $30,000$ different innings played over the time. Furthermore, this data is annotated by attributes like balls faced and runs scored.

Limited by the scope of this project, we focus only on the runs scored by a team and a batsmen as it has been proved to affect the outcome of a game. '*Strike Rate*' is a major factor which determines the performance of the player and is defined as:

$$Strike\ Rate = \frac{Runs\ Scored}{Balls\ Faced}$$

4.2 Layout

Our system is comprised of four modules that visualize team and player performance. We have primarily used the number of runs scored as the metric to measure the performances of the players as well as the impact it has on the overall performance of the team. In our system, we provide the user with two drop-down menus, that list the names of the various cricket playing nations. These drop downs provide an effective way of filtering the results based on a team and the opposition it has played against. The first drop-down allows users to select a country for which they would like to view the team and player performances, the second, optional drop-down allows users to narrow results by selecting a single country that has played against the first country. By default the second drop down is set to show the results against all the oppositions.

4.3 Views

4.3.1 Bar charts

The first visualization contains two parallel bar charts (positioned one below another) which can be used to analyze a teams performance using their runs score. The first chart shows the number of runs scored by the team in a time-line fashion. The x-axis contains an ordinal scale with match dates while the y-axis contains the scores (runs). The second chart shows the margin by which the team has won or lost in each match. The length of the each bar in both charts represents the score or margin respectively. In case the team won the match while batting second, the margin is taken to be one. The color of each bar shows whether the match was won or lost.

The visualization also has a sorting feature which allows

users to sort matches by date, runs score, or margin of runs scored. The sorting mechanism is useful and really effective in arriving at the conclusions such as, what is a safe score for a team if batting second, what is the probability that a team would successfully defend or chase a target set by the opposition.

Filtering based on the margin of runs helps us understand whether the team plays too many close games. This could be due to the fact that the team loses out in the critical points in the game. South Africa for instance has always had the tag of being "Chokers" as they have been known to be a team that loses important matches at in-spite of being one of the best teams in the world.

Another valuable insight that can be derived by sorting the data based on the margin is to gaze whether the team's performance has been increasing, if the margins have been decreasing, or vice versa.

Using these sorting features along with the second drop down to select the opposition, can help us understand how the team has been performing against different oppositions.

The next two visualizations in our system combine to make a multiple linked view. Using the data visualization technique of brushing, we linked a scatter-plot with a histogram.

4.3.2 Scatter-plot

A scatter-plot is used to depict the relationship between the number of balls faced by a batsman in innings and the runs they scored. The x-axis contains the number of balls in an inning faced by the batsmen and the y-axis contains the number of runs scored. The relationship between these two values is known as a "strike rate" in cricket. Scoring a lot of runs in the game is as important as the number of balls faced in scoring those runs. Thus, the players who score a lot at an extremely fast rate are considered to be more valuable players when compared to players who make a lot of runs at a slower rate. Essentially, each point in the scatter-plot will represent an inning played by player in terms of their strike rate.

The points in the scatter-plot are color-coded; the color of a point represents the outcome of the match in which the inning (i.e. the point of interest) was played. Blue is used to represent a win while red is used to represent a loss. For example, a point depicting the strike rate of Player A will be blue if the team had won the particular match. Color-coding the scatter-plot help us find an underlying relationship between a players performance and the outcome of the match.

On each point in the scatter-plot, we show a tool-tip which gives additional information such as the corresponding player name, the players runs scored, and the number of balls faced by the player.

The scatter-plot also has an option to select a range of points (brushing). When a set of points is selected in the scatter-plot, a dataset is generated with all the points in the selected area and passed to a histogram. that would visualize a stacked bar graph for each player for whom a point is found in the selection. The stacked bar will represent the number of points in the selected area for a lost cause vs winning cause for a player.

4.3.3 Stacked Bar

The stacked bar chart is designed to visualize player frequencies that will be retrieved from the dataset mentioned above. For example, if two points are selected in the scatter-plot and both correspond with strike rates of player "X", the histogram will contain a bar for player "X" with a height of two. This visualization is aimed at understanding how a particular player's performance affects the success of his team. A stacked bar in the chart represents the count of innings found in the selection made in the data visualized in the scatter plot. Each bar will be stacked based on the number of innings that were played in a lost cause against the number of innings played where the team won the game.

'Finishers' is a term often used to describe a player who has the ability to stay 'not out' till the end of the game and thus himself lead the team to a win. These players often have to face the pressure situation in a game. Players like the Indian Cricket team captain *Mr M.S. Dhoni* and *Suresh Raina* have made a name for themselves to be one of the best finishers in the game. The visualization not only helps us verify these facts but also help us find other such players who may not have received the due credit.

When used with the opposition filter, this visualization also helps us identify how certain players play against different oppositions. For instance, even though the legend *Sachin Tendulkar* has higher number of runs against 'Pakistan', it is actually *Yuvraj Singh* whose innings have helped India win 100% of games whenever he has scored a score of 80 or more in a game. Interestingly, *Yuvraj Singh* has had just one inning of 80 or more against another opposition 'New Zealand'.

4.3.4 Stacked Area Chart

Stacked area chart is an extremely effective visualization in visualizing the performance of the team based on the sum of the number of runs, the players of the team, have been scoring each year. Stacked area chart also provides an effective means of analyzing the contribution made by a key batsman in the total number of runs scored. This visualization is useful in understanding when the team has peaked and coinciding this with the important tournaments played over the years gives us an understanding of the results of a lot them.

For instance, team India performed really well in the years 2011 and 2013, the same time that they won the Cricket World cup and the ICC Champions trophy, two of the most prestigious trophies in cricket. New Zealand, who had an extraordinary year as a team in the year 2015, reached the finals of the Cricket World Cup for the first time ever in the history of the tournament.

Not only this, the visualization also helps us understand the effect that the retirement of the important players has on a team. In the year 2016, when three of the most prolific batsmen of 'Sri Lanka', *Mr. M. Jayawardene*, *Mr K. Sangakara* and *Mr T. Dilshan* retired, their performance as a team went to an all time low in the last 10 years.

This visualization also helps analyze how the careers of individual players have evolved over the years. From the debuts and retirements of the players to injuries, that led to

a temporary lapse in the performance of a player, can all be visualized using this visualization.

5 Results

The results go here.

6 Conclusion

The conclusion goes here.