Title: Network-based Team Formation: A Case Study of Pakistan Super League

Members:

Abeeha Zawar, SDP, az06225@st.habib.edu.pk

Abeer Khan, CS, ak05419@st.habib.edu.pk

Maria Hunaid Samiwala, SDP, ms05686@st.habib.edu.pk

Section 1: Introduction

Social network analysis gives information and insights into the network's individuals' complex interactions and network dynamics. For our research, we created a small world network of PSL cricket players from 2016 to 2022. The nodes represented the players, while the number of matches played between them served as the edges. The objective of this research was to use social network analysis methodologies such as centrality metrics to evaluate an individual player's belongingness within the team as well as their individual performances. In conclusion, the players' value and performance as team members were reflected in team formation, which is critical in team sports. The important players can be chosen for team building based on that characterization.

Cricket is a bat-and-ball sport played on a pitch between two teams of eleven players each. The batting team strives to score as many runs as they can, meanwhile, their competitors are the bowling team, which includes fielding and bowling and aims to obtain as many wickets as possible while limiting the batting side's runs. Additionally, wicket-keepers guard the wicket stumps, and all-rounders are players that perform well as both batters and bowlers. The duration of each inning in short-term cricket ranges from 20 overs (T-20 cricket) to 50 overs (T-50 cricket).

According to ESPN, the Pakistan Super League (PSL) is the popular domestic cricket league of Pakistan. It is played every year in Spring by franchise teams representing Pakistani cities from various regions. There are now six teams competing in this league, and each player of the team is chosen through a player auction conducted by the franchises. Additionally, a player can be bought by other franchises before a new season begins, thus, some players can switch teams multiple times in their PSL career, or never at all. Each franchise can pick a roster of 20 or over players, with 11 of them participating on the day of the match (called the 'Playing XI'), and the rest serving as extra members or reserves.

This paper proposes a ranking structure based on player performance evaluation with respect to their belongingness. Centrality measures (degree centrality, betweenness centrality, and closeness centrality) and local clustering coefficients are employed to assess a player's belongingness. The statistics for this analysis are derived from PSL T-20 cricket matches played between 2016 and 2022.

Our approach is based on the belief cricket is essentially a team-sport, rather than an individual one, hence the performance of players with respect to their peers is critical in their selection in a team. We selected players who have the ability of playing a team sport, which is what results in the success of the entire team.

The objective is that using this method, a group of players can be generated from which a playing XI team may be chosen based on their performance as team players. The selected team based on our approach are then assessed with the official squad chosen by the Pakistan Cricket Board (PCB; the country's official cricketing body) for the recent ICC T20 Men's World Cup 2022. Since the official PCB squad for the World Cup contains a 15 member squad with 3 reserves, our approach will also consider 18 players, based on their centrality rankings, local clustering coefficients, and roles (ICC, 2022). Additionally, we have assessed the small-world characteristic of the PSL network, comparison with seminal models, as well as a temporal analysis of each season.

The key findings of our paper:

- 11 of the 18 players in the World Cup squad appear in the teams selected based on our approach;
- The team formed using the local clustering coefficient measure is completely different from the other 3 centrality-based teams;
- Most of the 10 players that do not appear in any of our teams, but were selected for the World Cup, are still high up in our rankings;

– Some of the discordance betweens players performances and selection can be attributed to the nuances that come with the sport of cricket;

– The data depends on attributes such as the role of the players, whether they were injured depend on whether a player has switched teams; these can be the reason for some of the discordance in teams;

– The distribution of degree and betweenness centrality of the entire PSL network followed the Power Law distribution, whereas the closeness centrality followed the Normal distribution, and the local clustering coefficient was a mix of both

Section 2: Research Questions

1. Does the analysis of the PSL T20 Network inherit characteristics of a small world network?
2. How does it compare to the seminal models, i.e.the Erdős-Renyi (E-R) model, the Watts-Strogatz (W-S) model, and the Barabási-Albert (B-A) model?
3. What can we infer about the PSL network(s) over time, across all seasons?
4. Can the application of local clustering coefficient and centrality measures (betweenness centrality, degree centrality, closeness centrality), help quantify the quality of team belongingness and efficiency of each player, with respect to forming an ideal team?
5. How does our proposed approach and outcome compare with the recent ICC T20 World Cup 2022 team selection and performance?

Section 3: Related Work

In 2016, Dey, Paramita, Ganguly, Maitreyee & Roy, and Sarbani studied a creative use of small-world network features to construct a team of players with the greatest performance and belongingness within a cricket team network in their paper, "*Network Centrality Based Team Formation: A Case Study on T20 Cricket*". They used the small-world network approach on T-20 cricket teams to validate their hypothesis by representing players as network nodes, and the frequency of interactions between team members as the edges connecting those nodes. The network analysis revealed that the T-20 cricket network had all of the features of a small-world network. This research paper presents an updated technique based on social networking to assess and determine the quality of team belongingness and each player's performance, such as factors that include success against a formidable opponent or performance as a productive and impactful team member like fielding, running between the wickets, and solid partnership. The use of Social Network Analysis (SNA) is considered in order to evaluate and analyze player performance and rank. This paper builds a bidirectional weighted network of players and employs it for network analysis obtaining data from T20 cricket (2014-2016). Therefore, the team was constructed based on that rating and their 2016 IPL (Indian Premier League) results.

In the paper, "*Analyzing passing networks in association football based on the difficulty, risk, and potential of passes*", Wiig et al. (2019) examine the use of network analysis in reference to football to determine important players in teams and identify patterns and trends of movement and distribution within teams. Networks are built using player passes, and different centrality metrics are studied in relation to three alternative approaches for rating player passes. Four seasons of data from Norway's top division are employed to determine crucial players and evaluate matches from a selected team. The weights assigned to the passes in this study are based on three separate elements of the passes: their likelihood of completion, the likelihood that the team retains possession after the successful pass, and the likelihood that the pass is part of a sequence leading to a shooting. The findings of this paper indicated that utilizing multiple measures and network weights leads to the discovery of crucial passes in various phases of play and places on the field.

In the paper, "*Quantifying individual performance in cricket - a network analysis of batsmen and Bowlers*", Mukherjee et al. (2013) used players' performance and ability in cricket in order to create teams and selected them for international matches. The total number of runs scored by batsmen (average runs) and wickets obtained by bowlers (average wickets) is a logical means of assessing and evaluating a cricketer's efficiency and performance. Batsmen and bowlers are conventionally ranked based on their batting and bowling averages. Nevertheless, in a sport like Cricket, the technique and method in which one gets runs or takes a wicket are also extremely essential, which Mukherjee included in her paper. She gave more recognition to players who score runs against a good bowling attack or put in a stellar performance against a team with a formidable batting order. A player's average does not capture this component of the game. The authors explored the use of Social Network Analysis (SNA) to assess the performance of individual team members. Using the player-vs-player data available for Test and ODI cricket, they created a directed and weighted network of batsmen-bowlers. Their findings revealed that M. Muralitharan was the most outstanding bowler in international cricket.

In the paper, *"Complex network analysis in cricket:Community Structure, player's role and performance index"*, Mukherjee (2012) uses network approaches to study the interactions between players of sports teams. He examines the interaction and relationship of batsmen in International Cricket matches by creating a batting partnership network (BPN) for each team and calculating the exact values of the networks' clustering coefficient, average degree, and average shortest path length and comparing them to the Erdös-Rényi framework. In his study, he notices that the networks exhibit small-world behaviour. The most connected batsman is not always the most important, influential or central, and the most central players are not always the ones with the highest batting averages. Mukherjee studied the BPNs' community structure to determine each player's function based on inter- and intra-community relationships. He noted that Sir DG Bradman, widely recognized as the greatest batsman in cricket history, did not hold the network's central position, the so-called connecting hub. Mukherjee expands his methodology to quantify the efficiency, performance, relevance, and impact of removing a member from the team using different centrality values.

In the paper, *"Networks as a novel tool for studying team ball sports as complex social systems"*, Davids et al. (2010) discuss and assess the innovative value of network approaches for explaining players' interpersonal interactions within societal neurobiological systems such as sports teams. They demonstrate how the accumulation of interpersonal contacts that occur from system agent activity supports collaborative system networks (such as players in a sports team). To put this theory to the test, the authors used the approach to examine intra-team clusters behaviours in the behaviours of water polo. The number of interactions between team members culminated in diverse intra-team synchronization and coordination patterns of play, distinguishing between successful and failed performance results.

Section 4: Methodology

4.1: Data Acquisition

The data was taken from cricsheet.org which has ball-by-ball information for a variety of cricket matches. Cricsheet.org provides a registry containing entries of 12,423 people that have participated in the cricket industry. This data allows a person to be correctly identified by their ids on various sites, such as CricketArchive, Cricinfo, CricHQ, Pulse, etc. From this dataset, we found all the players that participated in PSL seasonally and overall, the teams they played in respectively, and the number of matches they played. After building this data, we created a performance pool of only Pakistani players that are eligible to participate in the 2022 T20 World Cup. For this, we analysed the data and developed a set of constraints to perform performance pool selection.

4.2: Network Formation

When focusing on network formation, we used the dataset to extract the complete list of players that have participated in PSL. A python script was built to extract this information. Then the information extracted was manually checked to locate any duplicates or errors. The file that was generated contained a list of players and their respective unique identifiers as given by the Cricsheet.org register. Upon inspection, we found some duplicates and errors which we removed. Using this cleaned file we were able to identify which player played for which teams in each season, and the seasonal edge lists. The network formation for all analyses includes retired and international players, from which we made a performance pool of Pakistani eligible players.

The edge list contains a list of pairs of players that have participated in a match together, the ratio of common matches they've appeared in, and the total number of matches a player has played. All players were taken into consideration for this when building a network (using these edge lists). RStudio was used to construct the network using the libraries igraph, sqldf, DirectedClustering, reshape2, and dplyr. A bidirectional network was built.

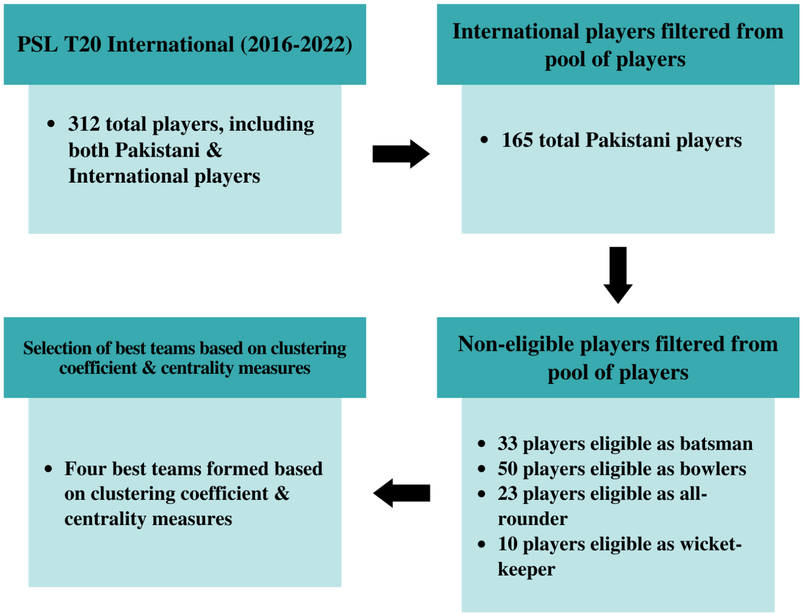
In this network, we did not take into account 4 missing matches, as the data was not available on Cricinfo.org. For team selection, a lot of the analysis of players' performance and form over the years was based on our personal observations of the T20 matches that we have been watching, as there was no literature and articles available on these players' performances. While calculating the clustering coefficients, we noticed that the values for an undirected and directed network were slightly different. Due to this, we used the DirectedClustering library which allowed us to calculate the clustering coefficients of directed networks.

4.3: Team selection

When selecting our team, we created a performance pool selection that involved dividing the Pakistani players into two categories; eligible, and non-eligible. Firstly, we filtered the international players from Pakistani players from the players' pool. A player is only considered eligible if he has not been prohibited by the Pakistan Cricket Board (PCB) or ICC from participating in T20 cricket, is less than 40 years old, and has not retired from international T20 cricket. Whereas, a player is only considered non-eligible if he has recently or is currently prohibited by the PCB or ICC from participating in T20 cricket, is older than 40 years old, is above 35 and has never played T20 cricket internationally (or last international t20 match was played five years ago), or has officially retired from international T20 cricket.

We have created 4 18-player teams ranked according to the 3 centrality measures and local clustering coefficient in decreasing order. Since the official PCB squad for the World Cup contains a 15-member squad with 3 reserves, our approach was to consider 18 players, based on their roles and centrality. The ICC team selected had 6 batsmen, 3 all-rounders, 2 wicket-keepers, and 7 bowlers; the same ranking that we used when selecting players for the team formation. The performance of the players in the World Cup (selected and not selected by our approach) were analyzed. Our entire approach is outlined in Figure 4.1 in detail.

Figure 4.1: Flow diagram of the proposed approach.


PSL T20 International (2016-2022)
• 312 total players, including both Pakistani & International players
International players filtered from pool of players
• 165 total Pakistani players
Non-eligible players filtered from pool of players
• 33 players eligible as batsman
• 50 players eligible as bowlers
• 23 players eligible as all-rounder
• 10 players eligible as wicket-keeper
Selection of best teams based on clustering coefficient & centrality measures
• Four best teams formed based on clustering coefficient & centrality measures

There were some differences the roles of 3 players in regards to their roles in the official ICC T20 team and their roles in their PSL career i) Iftikhar Ahmed performed as a batsman in his PSL career, but performed as an all-rounder in the WC; ii) Mohammad Haris performed as a batsman in his PSL career, but was taken as an wicket-keeper reserve in the WC; and iii) Mohammad Wasim performed as an all-rounder in his PSL career, but performed as a bowler in the WC. We considered their original roles from their PSL career for the ease of our analysis.

Section 5: Network Analysis

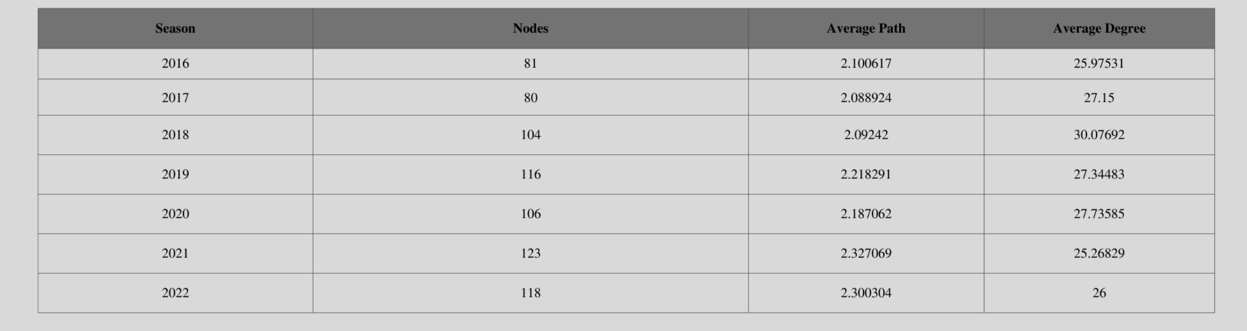
Our network analysis takes on various forms: firstly we have briefly discussed the seasonal analysis of all 7 PSL seasons and their trends; secondly, we checked whether our single generated network (overall, of all seasons) exhibits any small world characteristics. We additionally, compared and contrasted it with the three seminal networks of the Erdős-Renyi (E-R) model, the Watts-Strogatz (W-S) model, and the Barabási-Albert (B-A) model. Finally, we analysed the local clustering coefficient and the centrality measures (degree, betweenness, and closeness) as well as their distribution and what it means for the network.

5.1: Seasonal Analysis

Temporal analysis of networks is the study of how networks change over time. This can include examining how the structure of a network changes, how the nodes in a network change, and how the relationships between nodes change. Temporal analysis can be used to understand the dynamic processes that occur within networks, and to identify patterns and trends in how networks evolve over time. It is a valuable tool for studying networks in many different fields, including sociology, biology, and computer science. One example of temporal analysis in networks is the study of the spread of a disease through a network of people. By examining how the disease spreads from person to person over time, researchers can gain valuable insights into the dynamics of the spread of the disease, and can identify patterns and trends in how the disease moves through the network. This information can be used to develop strategies for controlling the spread of the disease and protecting people from becoming infected.

The first season of PSL was played in 2016. Since then, there have been a total of seven consecutive seasons played. For this section, the total number of players that have participated in matches, average path length, and average degree values are calculated for each season. which are then closely examined to identify any particular patterns that emerge from this data, and formulate the reasons for their occurrence.

Table 5.1: PSL seasons and their respective number of players, average path length, and average degree.



Season	Nodes	Average Path	Average Degree
2016	81	2.100617	25.97531
2017	80	2.088924	27.15
2018	104	2.09242	30.07692
2019	116	2.218291	27.34483
2020	106	2.187062	27.73585
2021	123	2.327069	25.26829
2022	118	2.300304	26

5.1.1: Nodes

An overall increase in the number of participating players can be observed from Table 5.1, indicating an increase in the size of the network, with a decline being observed for the 2017, 2020, and 2022 seasons. This means that the number of players that participated in this season was lower than the previous one. However, in 2020, the decline is probable due to the travelling restrictions and social distancing regulations that were imposed throughout that year due to the COVID-19 outbreak. The PSL 2021 season has the largest number of participating players which is found to be 123, and PSL 2017 season has the lowest number of participating players which is found to be 80. The sharp increase in the nodal count from 2018 onwards can be attributed to the inclusion of a sixth team from the third season onwards.

5.1.2: Average Path Length

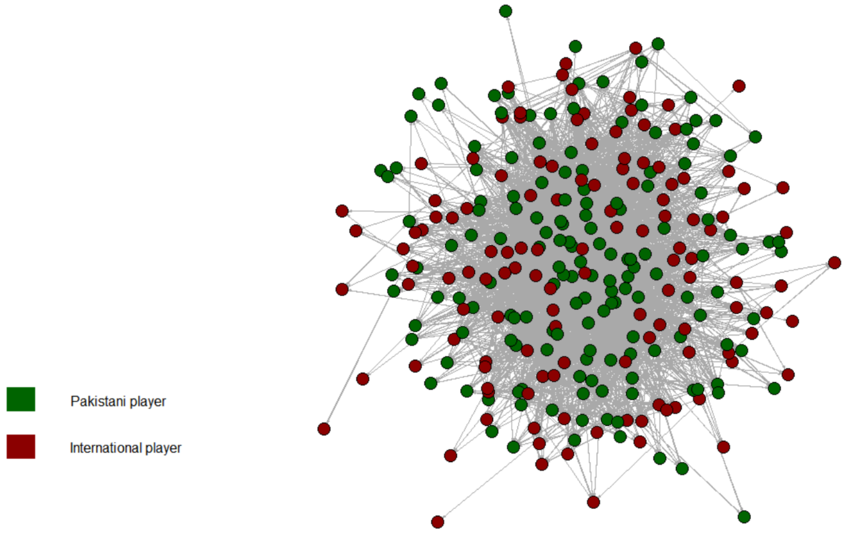
With the increase in number of nodes, an overall increase in average path length is also observed from Table 5.1, with a decline being observed for the 2017, 2020, and 2022 seasons. A slight decrease in the 2020 season could be probable due to the COVID-19 outbreak. The PSL 2022 season has the largest average path length of 2.300304, and the PSL 2017 season has the lowest average path length of 2.088924. An increase in node count typically leads to an increase in average path length because there are more possible paths that must be traversed in order to reach a destination node.

5.1.3: Average Degree

The average degree can be noted to stay more or less the same throughout the seasons, according to Table 5.1. The PSL 2018 season has the highest average degree of 30.07692, and the PSL 2021 season has the lowest average degree of 25.26829. In the first three seasons, the average degree is increasing in proportion to the increasing number of nodes. An increase in the total number of nodes in a network will typically lead to an increase in the average degree of the nodes. This is because, as the number of nodes increases, there are more opportunities for each node to form connections with other nodes, leading the network to become more connected. However, a decline can be noted from the fourth season onwards and the average degree is no longer increasing in proportion to the increasing number of nodes. This is possible because of the inclusion of new players that haven’t formed as many connections as the existing players that have participated in the previous seasons.

5.2: Small world characteristic & Seminal Models

Fig 5.2.1: Entire PSL T20 Network



The nodes in a small world network are not directly connected, rather the number of hops (i.e. distance) between any two nodes is significantly less. For our centrality analysis, these nodes connecting other nodes play a significant role in the network, hence, assessing the small world characteristic of the PSL network is essential. The measures used for this assessment were global clustering, average path length, network diameter, edge density, and degree distribution.

To adequately determine the characteristic of our entire PSL network (all seasons compiled), we also compared it to the three seminal models discussed in the course i.e. the Erdős-Renyi (E-R) model, the Watts-Strogatz (W-S) model, and the Barabási-Albert (B-A) model. The three networks were randomly generated on RStudio, using the same number of nodes and edges as the PSL network.

The average path length of a network shows how well the players (nodes) are connected to the other players in the player network. For a small world network, this means that average path length should be really low. As we know for small world networks, the average path length should be close to the natural logarithm of its number of nodes. This is shown as ln(N), where N is the number of nodes.

For the PSL network, the number of players (nodes) are 284, meaning its average path length should be close to 5.65 to be considered a small world. The actual average path length of our network is 2.16, which is much less than the required, positively indicating that the network is a very small world. It means the average number of hops between any two players in the network are less than three, indicating it is a highly-connected network. Additionally, the diameter is the maximum shortest path length of a network. The diameter for the PSL network is 5, meaning that the maximum number of hops separating two players is 5.

When compared to the seminal models, it is greater than the 1.96 average path length of the E-R model, but lesser than the 4.02 of the W-S network, and close to the 2.08 average path lengths of the B-A network. Random interactions are characteristic in the E-R network, so the PSL network being larger in path length and diameter confirms that the connections in the PSL network are not random, because the players only have direct interactions with each other when they play matches together. It is most comparable to the B-A network, which can mean that players might have preferentially attached to other players.

Fig 5.2.5: E-R Degree Distribution of PSL Network

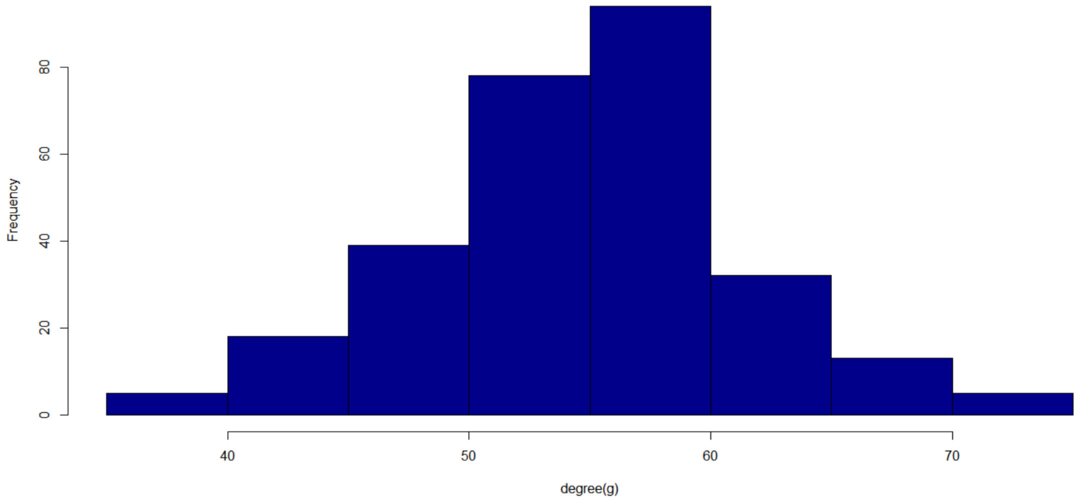


Fig 5.2.6: W-S Degree Distribution of PSL Network

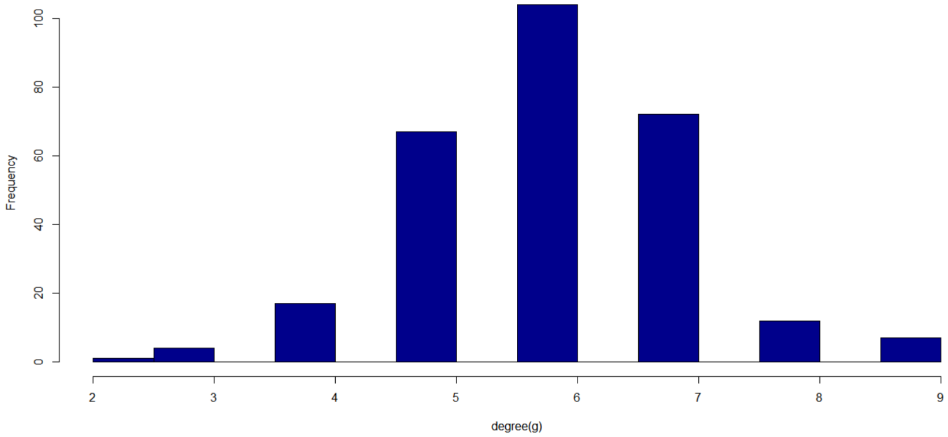
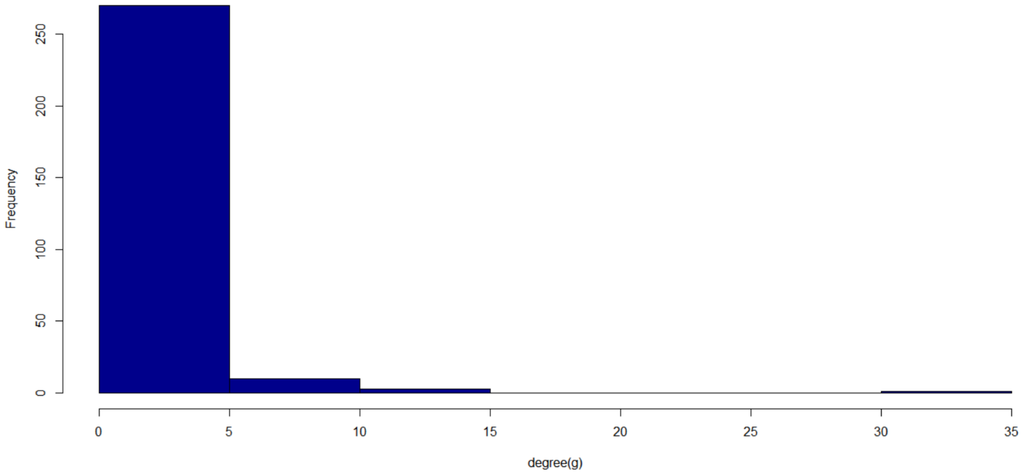


Fig 5.2.7: B-A Degree Distribution of PSL Network



The preferential attachment can also be confirmed through the degree distribution of all the networks. The PSL network has a skewed degree distribution (Figure 5.3.1), which follows the Power Law distribution and the B-A model, unlike the E-R and W-S models (Figures 5.2.5, 5.2.6, and 5.2.7). This indicates the presence of hubs in the network and Barabasi-Albert's theory that new nodes in a network preferentially or consciously attach to existing nodes with the highest degrees. It supports their claim of human networks being inherently inegalitarian, where the rich get richer. For the PSL network, it means that there are very few players that are extremely well-connected, while most are not (discussed more elaborately in the 'Degree Centrality' section).

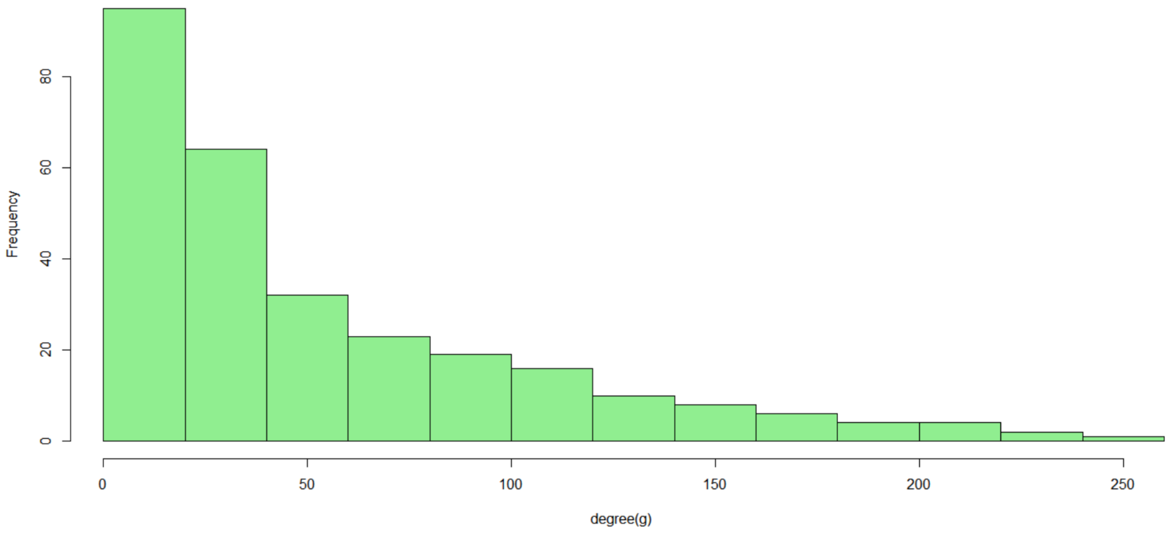
Additionally, in optimal human networks, we know that the average clustering should be above 20% or 0.20. The global clustering of our PSL network was found to be 0.2536755 or 25.4%, which confirms it is an ideal human network. However, it is close to the 33.0% average clustering for the W-S network, and much larger than the 9.74% clustering for the E-R network and the 0% for the B-A network. This is contrary to the preferential attachment depicted in the path length and degree distribution of the PSL network, which is closely comparable to the B-A model. This can be explained by understanding that all the seminal models are randomly generated for the same number of nodes as the PSL match, meaning that the players in our actual network have deliberate interactions with each other that are dependent on the matches they play together.

Hence, the presence of clusters can indicate players that have played multiple matches together, and hubs can indicate players that were almost always in the starting eleven of their team, or seldom got injured, or switched teams. Furthermore, we understand that human social networks tend to be sparse i.e. having low density because ties are weak. For the PSL network, the edge density is 0.0975 or 9.75%, which is the same as 9.75% of the E-R network, and greater than the 2.12% and 0.35% of the W-S and B-A networks, respectively. This confirms the nuances of the interactions in the PSL network (discussed more elaborately in ‘Centrality measures’ section).

5.3: Centrality measures & Clustering Coefficient

5.3.1: Degree centrality

Fig 5.3.1: Degree Distribution of PSL Network



Degree centrality assigns an importance score based simply on the number of links held by each node (player). It shows how many direct, or ‘one hop’, connections each node has to other nodes in the network. It indicates the number of matches that node (player) has shared with its direct neighbours. We used degree centrality to analyze the connected players, popular players, players who are likely to hold most information or players who can quickly connect with the wider network, and also to look at in-degree (number of inbound links) and out-degree (number of outbound links).

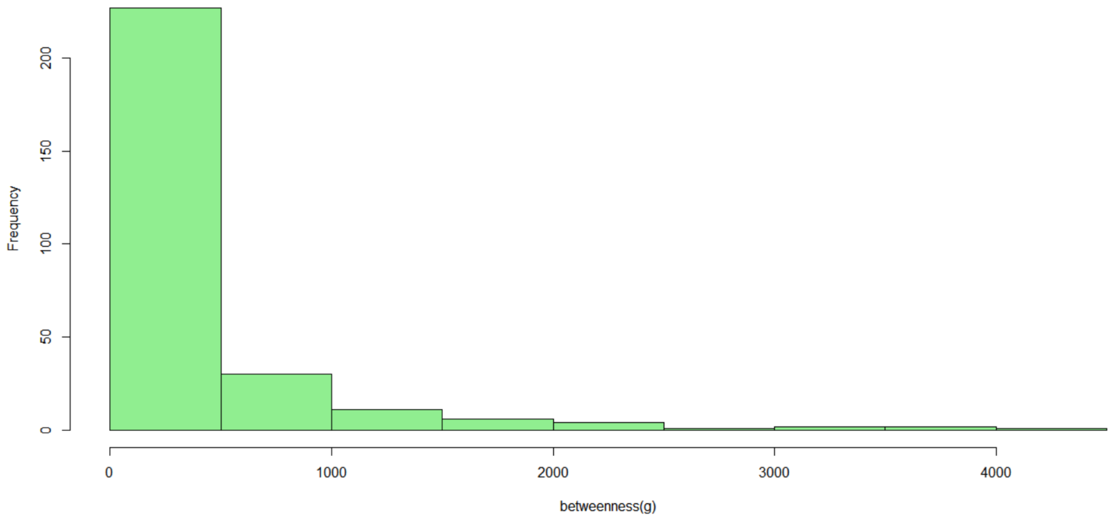
The maximum degree centrality in our entire PSL network (across all seasons) is 252 indicating that the player (Wahab Riaz) has directly been in contact (or played matches) with 252 other players in his PSL career. The average degree centrality is 55.21, meaning on average, a player ‘n’ has played direct matches with 55 other players in his PSL career.

The degree distribution of the PSL network is skewed and in accordance with the Barabasi-Albert model or Power Law. 95 nodes have a degree centrality of 20 or less (between 0-20), whereas only 1 node has a degree above 240. This indicates the presence of hub i.e. superconnectors in the PSL network, meaning that there are very few players that are extremely well-connected, while most are not.

While degree centrality is a good measure of the total connections a player has, it will not, however, necessarily indicate the importance of a node in connecting others or how central it is to the main group. High degree centrality means the player has played a lot of matches, but this is not a complete reflection of their ability, as some players might have been injured during their PSL career or they might have been in the reserved players (i.e. not in the starting 11). It is important to note that the degree centrality of a node, and other centrality measures, depend on whether a player has switched teams. Furthermore, these figures cannot account for the benefits based on roles. Batters are only able to perform in the batting innings of a match if their preceding peers lose their wicket and their turn comes up, whereas all bowlers in a starting 11 get their 5 overs to perform in the bowling innings. On the other hand, wicket-keepers and all-rounders can perform in both innings. Thus, these are nuances that are unaccounted for in the data. Thus, the hubs can indicate players that were frequently in the starting eleven of their team, which can be an acknowledgement of their ability and their experience.

5.3.2: Betweenness centrality

Fig 5.3.2: Betweenness Distribution of PSL Network



Betweenness centrality for any node 'n' in a network can be defined as the proportionality of total shortest paths passing through that node 'n' to all possible shortest paths present in the network. A player with a high value of betweenness centrality has large influence on the other players, as they are dependent on passing through him to interact with other nodes he is connected to.

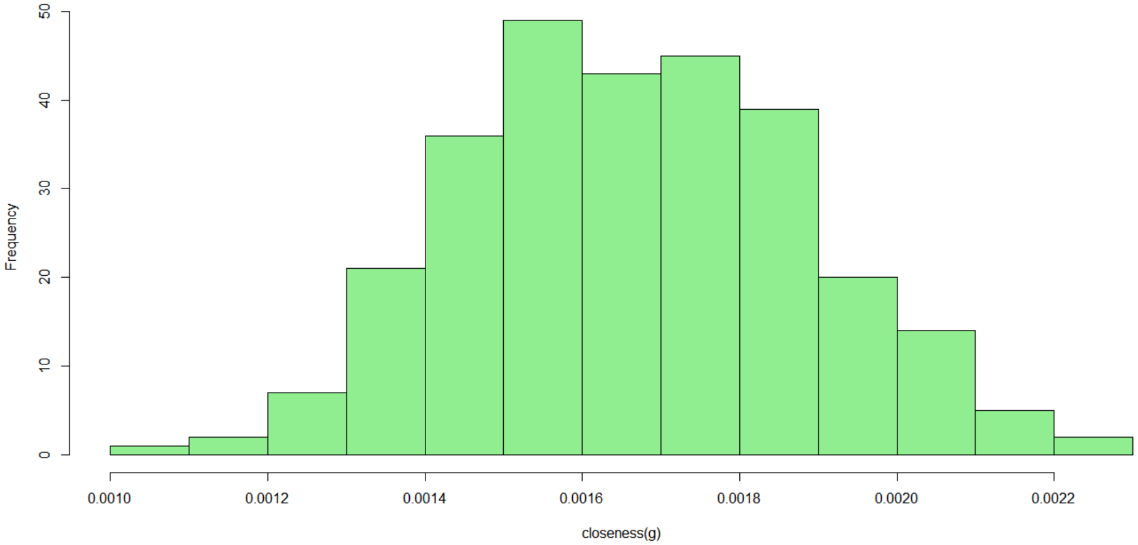
The average betweenness in our entire PSL network (across all seasons) is 328.0, meaning that on average, for a player 'n' in the PSL network, there are 328 shortest paths that include 'n' in-between, when it comes to other players trying to interact w each other. Similarly, the maximum betweenness of a player is 4423.4, and the minimum is 0.

Additionally, the distribution of the betweenness centrality of the network is heavily skewed, following the Power Law. This means only a handful of players have high betweenness while most do not. Only 10 players have a betweenness that is above 2000, while 227 players have a betweenness between 0-500. Hence, the mean might not be an accurate measurement of the betweenness in the PSL network.

While betweenness centrality is a good measure of the influence a player has on his neighbours, we iterate that these figures depend on whether the player was injured, or in the starting 11 of a match, their role, or if they switched teams in his PSL career. Hence, high between can mean the players were frequently in the starting eleven of their team, which can be an acknowledgement of their ability and their experience.

5.3.3: Closeness centrality

Fig 5.3.3: Closeness Distribution of PSL Network



In a graph representing a small world network, closeness centrality measures how close a node is to others in the network. In other words, the closeness centrality of a node in a network is the inverse of distance between two nodes in that network. A high closeness centrality value indicates that a node in a graph is well connected to other nodes in the graph. This means that the node is able to reach other nodes in the graph quickly and efficiently, which makes it an important node in the network. The clustering coefficient measures the local connectivity of a node, while the closeness centrality measures its global reach.

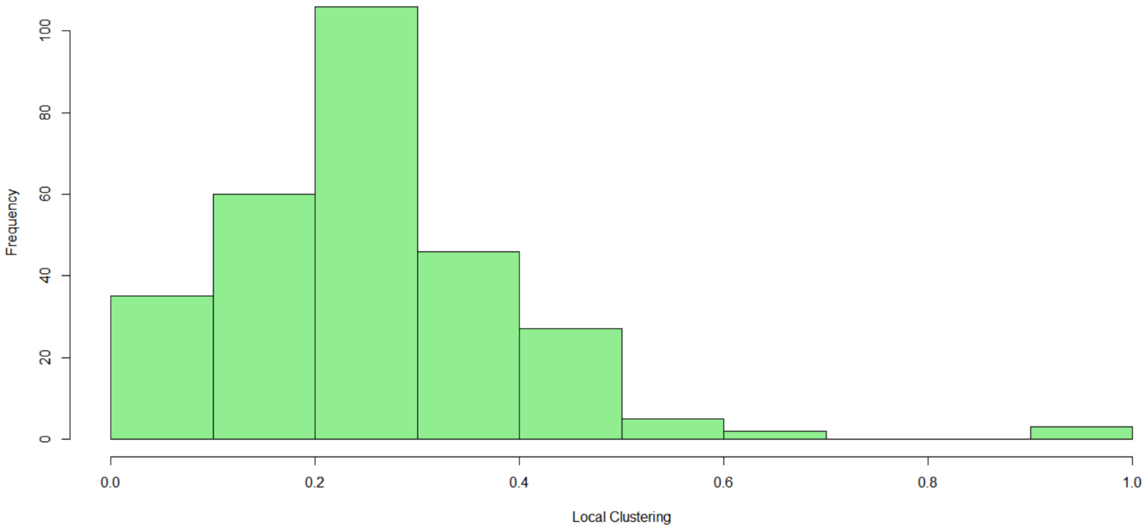
The average closeness centrality in our entire PSL network (across all seasons) is 0.00166, meaning that the nodes in our network are very close to each other, which is an acknowledgment of the PSL network being a small-world network.

The distribution of the closeness centrality represents a normal distribution. At the tail ends, 3 nodes have closeness centrality between 0.0010-0.0012, and 7 nodes have a closeness centrality between 0.0021-0.0023, meaning there are only a few nodes that are able to reach many other nodes. Additionally 49 nodes have a closeness centrality between 0.0015-0.0016.

While closeness centrality is a good measure of how influential a player is because they are able to quickly and easily reach many other nodes in the network, we reiterate that figures depend on whether a player switched teams, was injured, their role, or was on the starting 11 of their team i.e. not a reserved player. Thus, players with high closeness centrality are those that were more often than not in the starting eleven of their team, which can be an acknowledgement of their ability and their experience.

5.3.4: Clustering coefficient

Fig 5.3.4: Clustering Coefficient Distribution of PSL Network



Clustering coefficient of a node (player) in the network signifies characteristics of that player forming local clusters, which is numbers of players that are influenced by that particular player. The dense local cluster signifies that the node has great influence on other nodes.

Our global clustering for our entire PSL network (across all seasons) came as 0.2537, or 25% clustering, showing characteristics for small-world human social networks. It means that on average, 25% of the neighbours of a player 'n' are connected to each other. This is good because it means that the neighbours are mostly dependent on player 'n' to interact with each other, making 'n' a central player in the PSL network. This also means that a cluster forms around 'n', where it plays a central role for its neighbouring players to interact with each other.

The distribution of the local clustering coefficient is normal. Only 10 players have clustering coefficients above 0.5 or 50%, meaning that these players and their neighbours are equally dependent on each other. At the tail ends, only 3 nodes have a local clustering coefficient between 0.9-1.0, whereas 35 nodes have a local clustering coefficient between 0-0.1, and only 10 nodes have local clustering coefficient above 0.5. This shows that while the distribution is normal overall, there is a disparity in the tail ends of the distribution, hence it is a mix between a normal and a Power distribution.

While local clustering coefficient is a good measure of the influence of a player on his neighbours, a high clustering coefficient can mean that the neighbours of player 'n' are not dependent on n to interact with each other, as they are well-connected amongst each other without 'n'. Additionally, as discussed in degree centrality, these figures depend on if the player was in the starting 11 of a match, their role, if they were injured, or switched teams. Hence, the presence of clusters in the PSL network can indicate players that have played multiple matches together, and were frequently in the starting eleven of their team, which can be an acknowledgement of their ability and their experience.

Section 6: Team selection

From the pool of players of total PSL players who are Pakistani, we divided them into two categories: eligible and non-eligible. As mentioned in Section 4, we considered a player as eligible if he is less than 40 years old, has not retired from international T20 cricket, and has not been prohibited by the PCB or ICC from participating in T20 cricket. From this pool we created

4 teams ranked according to their centrality measures (degree, betweenness centrality, and closeness centrality) and local clustering coefficient in descending order. Our entire team selection approach has been laid out in Figure 4.1.

We created four 18-player teams in accordance to the official ICC T20 World Cup 2022 Team that was selected by the PCB before the start of the World Cup (WC) in October of 2022. The official PCB squad for the World Cup contained a 15-member squad with 3 reserves: it had 6 batsmen, 3 all-rounders, 2 wicket-keepers, and 7 bowlers, so our ranking has taken these exact numbers and roles into our own selection consideration. We noticed there were some difference in the roles of 3 players in regards to their roles in the official ICC T20 team and their roles in their PSL career–i) Iftikhar Ahmed performed as a batsman in his PSL career, but performed as an all-rounder in the WC; ii) Mohammad Haris performed as a batsman in his PSL career, but was taken as an wicket-keeper reserve in the WC; and iii) Mohammad Wasim performed as an all-rounder in his PSL career, but performed as a bowler in the WC. For the ease of our analysis, we considered their original roles from their PSL career. The four teams as shown in Table 6.1, the performances of the WC players in PSL in Table 6.3, and the performance of players that appear in our teams (and not in the WC) in Table 6.2.1 and 6.2.2.

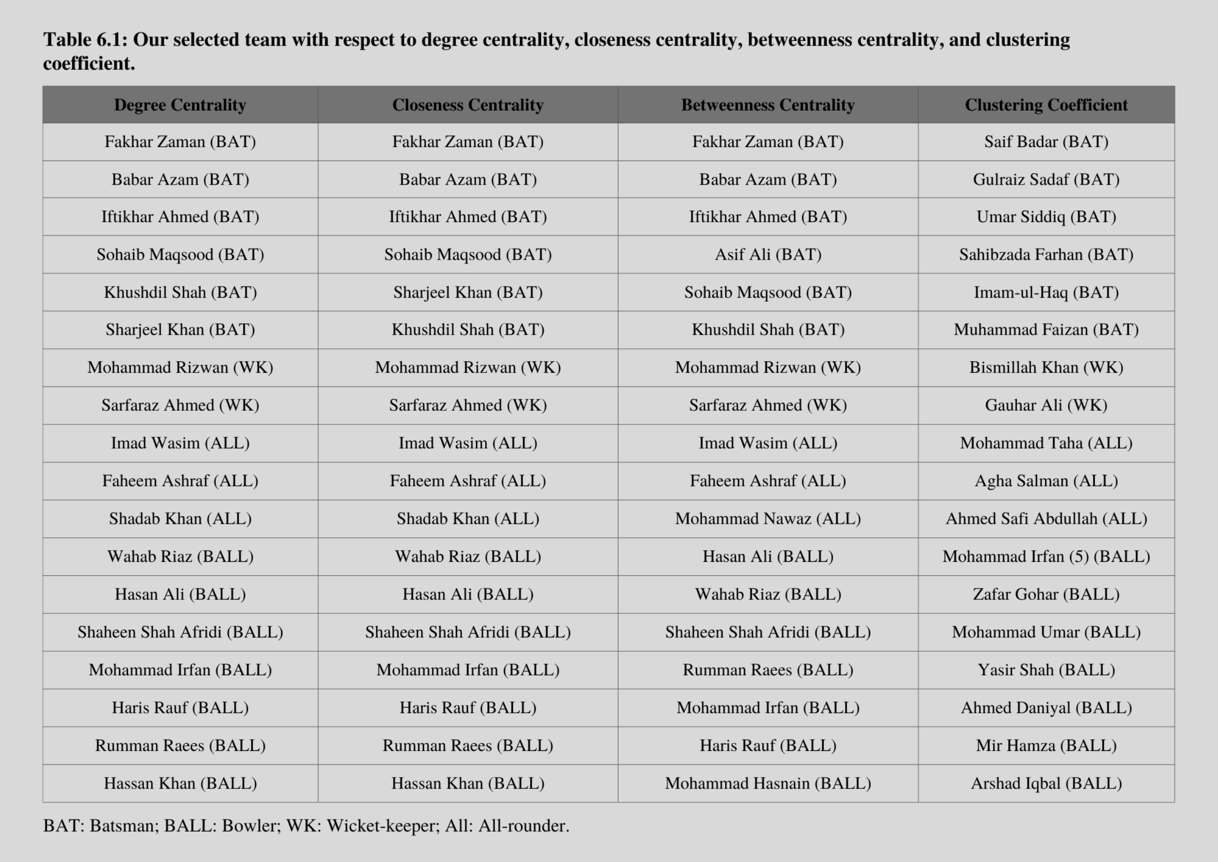


Table 6.1: Our selected team with respect to degree centrality, closeness centrality, betweenness centrality, and clustering coefficient.

Degree Centrality	Closeness Centrality	Betweenness Centrality	Clustering Coefficient
Fakhar Zaman (BAT)	Fakhar Zaman (BAT)	Fakhar Zaman (BAT)	Saif Badar (BAT)
Babar Azam (BAT)	Babar Azam (BAT)	Babar Azam (BAT)	Gulraiz Sadaf (BAT)
Iftikhar Ahmed (BAT)	Iftikhar Ahmed (BAT)	Iftikhar Ahmed (BAT)	Umar Siddiq (BAT)
Sohaib Maqsood (BAT)	Sohaib Maqsood (BAT)	Asif Ali (BAT)	Sahibzada Farhan (BAT)
Khushdil Shah (BAT)	Sharjeel Khan (BAT)	Sohaib Maqsood (BAT)	Imam-ul-Haq (BAT)
Sharjeel Khan (BAT)	Khushdil Shah (BAT)	Khushdil Shah (BAT)	Muhammad Faizan (BAT)
Mohammad Rizwan (WK)	Mohammad Rizwan (WK)	Mohammad Rizwan (WK)	Bismillah Khan (WK)
Sarfaraz Ahmed (WK)	Sarfaraz Ahmed (WK)	Sarfaraz Ahmed (WK)	Gauhar Ali (WK)
Imad Wasim (ALL)	Imad Wasim (ALL)	Imad Wasim (ALL)	Mohammad Taha (ALL)
Faheem Ashraf (ALL)	Faheem Ashraf (ALL)	Faheem Ashraf (ALL)	Agha Salman (ALL)
Shadab Khan (ALL)	Shadab Khan (ALL)	Mohammad Nawaz (ALL)	Ahmed Safi Abdullah (ALL)
Wahab Riaz (BALL)	Wahab Riaz (BALL)	Hasan Ali (BALL)	Mohammad Irfan (5) (BALL)
Hasan Ali (BALL)	Hasan Ali (BALL)	Wahab Riaz (BALL)	Zafar Gohar (BALL)
Shaheen Shah Afridi (BALL)	Shaheen Shah Afridi (BALL)	Shaheen Shah Afridi (BALL)	Mohammad Umar (BALL)
Mohammad Irfan (BALL)	Mohammad Irfan (BALL)	Rumman Raees (BALL)	Yasir Shah (BALL)
Haris Rauf (BALL)	Haris Rauf (BALL)	Mohammad Irfan (BALL)	Ahmed Daniyal (BALL)
Rumman Raees (BALL)	Rumman Raees (BALL)	Haris Rauf (BALL)	Mir Hamza (BALL)
Hassan Khan (BALL)	Hassan Khan (BALL)	Mohammad Hasnain (BALL)	Arshad Iqbal (BALL)

BAT: Batsman; BALL: Bowler; WK: Wicket-keeper; All: All-rounder.

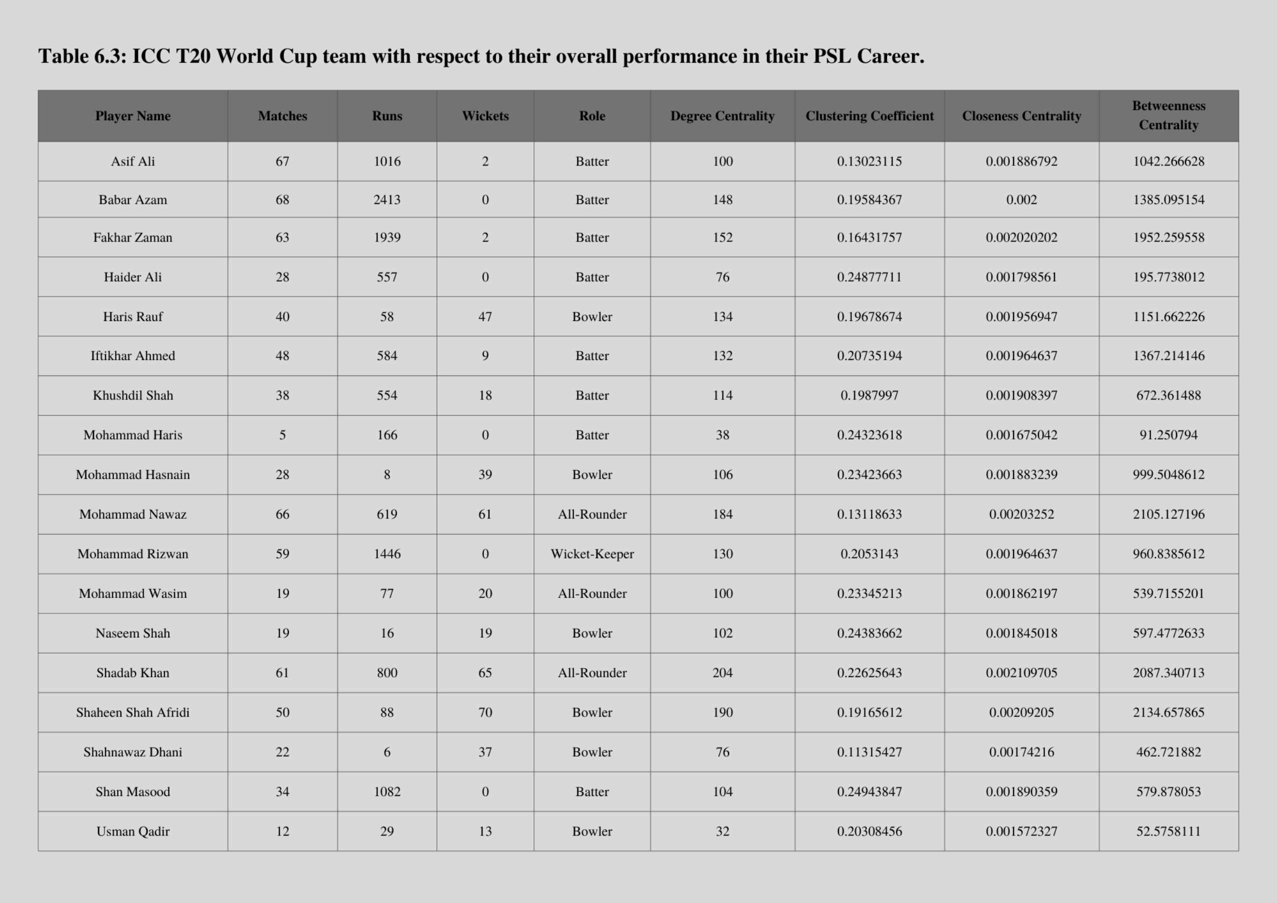


Table 6.3: ICC T20 World Cup team with respect to their overall performance in their PSL Career.

Player Name	Matches	Runs	Wickets	Role	Degree Centrality	Clustering Coefficient	Closeness Centrality	Betweenness Centrality
Asif Ali	67	1016	2	Batter	100	0.13023115	0.001886792	1042.266628
Babar Azam	68	2413	0	Batter	148	0.19584367	0.002	1385.095154
Fakhar Zaman	63	1939	2	Batter	152	0.16431757	0.002020202	1952.259558
Haider Ali	28	557	0	Batter	76	0.24877711	0.001798561	195.7738012
Haris Rauf	40	58	47	Bowler	134	0.19678674	0.001956947	1151.662226
Iftikhar Ahmed	48	584	9	Batter	132	0.20735194	0.001964637	1367.214146
Khushdil Shah	38	554	18	Batter	114	0.1987997	0.001908397	672.361488
Mohammad Haris	5	166	0	Batter	38	0.24323618	0.001675042	91.250794
Mohammad Hasnain	28	8	39	Bowler	106	0.23423663	0.001883239	999.5048612
Mohammad Nawaz	66	619	61	All-Rounder	184	0.13118633	0.00203252	2105.127196
Mohammad Rizwan	59	1446	0	Wicket-Keeper	130	0.2053143	0.001964637	960.8385612
Mohammad Wasim	19	77	20	All-Rounder	100	0.23345213	0.001862197	539.7155201
Naseem Shah	19	16	19	Bowler	102	0.24383662	0.001845018	597.4772633
Shadab Khan	61	800	65	All-Rounder	204	0.22625643	0.002109705	2087.340713
Shaheen Shah Afridi	50	88	70	Bowler	190	0.19165612	0.00209205	2134.657865
Shahnawaz Dhani	22	6	37	Bowler	76	0.11315427	0.00174216	462.721882
Shan Masood	34	1082	0	Batter	104	0.24943847	0.001890359	579.878053
Usman Qadir	12	29	13	Bowler	32	0.20308456	0.001572327	52.5758111

As apparent from Table 6.1, the team formed using the local clustering coefficient measure is completely different from the other 3 centrality-based teams, all of which have considerable overlap of members. Additionally, none of these 18 players were selected for the ICC World Cup. To reiterate our concerns from section 5.3, we believe that clustering coefficient is not an accurate reflection of a player's belongingness, because high clustering indicates all the player's neighbours are connected to each other, thus, it can mean that they are not dependent on said player to interact with each other. Hence, we believe this is a reflection of an error in the clustering criteria, but comprehending it is beyond the scope of our abilities.

We will then proceed to analyze the 3 teams formed from the centrality measures of degree, betweenness, and closeness, and compare them to the players from the WC team. From the 21 players that occur in these 3 teams, 11 players appear in the ICC World Cup team, while 10 players do not–namely batters Sohaib Maqsood and Sharjeel Khan; wicket-keeper Sarfaraz

Ahmed; all-rounders Imad Wasim and Faheem Ashraf; bowlers Wahab Riaz, Hasan Ali, Mohammad Irfan, Rumman Raees, and Hassan Khan.

Firstly, considering the matter of age, while our eligibility criteria eliminated players that were above 40 or above 35 and domestic players, in cricket the retirement age is an unsaid understanding. It is widely accepted that above the age of 30 makes you easily replaceable, and the younger the age, the better the chances of selection. The mean age of the members of the WC squad is 26 (as of December 2022, obtained from ESPNcricinfo). Considering this, 9 of the 10 players are above the age of 26, and 6 of which are above the age of 30. However, there are 4 players in the WC team above 30–Shan Masood (33), Iftikhar Ahmed (32), Fakhar Zaman (32), and Asif Ali (31). Fakhar Zaman was benched early in the WC because of an injury and was replaced, while the rest did not perform particularly well as batsmen.

Additionally, while our eligibility criteria eliminated players that have not played international cricket for 5 years, some players like Mohammad Irfan and Rumman Raees have not played T20 international cricket since 2019 and 2018, respectively, hence it might have been a consideration for not selecting them in the World Cup.

Furthermore, when you remove the 10 players from the rankings, then a lot of the immediate players that replace them are the same ones selected in the WC team. For example, all-rounders Shadab Khan ranks fourth in betweenness centrality and Mohammad Nawaz was fourth in both closeness and degree centrality; batsmen Shan Masood and Asif Ali were respectively ranked seventh and eighth in both closeness and degree centrality. For the bowlers, Mohammad Hasnain ranked eighth in both closeness and degree measures, and Naseem Shah ranked tenth in betweenness and closeness centrality and ninth in degree centrality.

There were seven players in the World Cup team that did not appear on any of our lists–batters Haider Ali, Shan Masood, and Mohammad Haris; bowlers Naseem Shah, Mohammad Wasim, Shahnawaz Dahani, and Usman Qadir. As aforementioned, Shan Masood and Naseem Shah were high on the rankings and Mohammad Wasim ranked in the top ten for all three teams categories. Mohammad Haris was a star performer during the World Cup in terms of performance, although he was on-boarded as a reserve and replaced Fakhar Zaman in the third match. Despite it being his first WC, his daring approach scored much-needed runs for his team, and he was such a good asset that they added him in the starting 11 for all the remaining matches. Hence, despite his low rankings in PSL, he proved to be a valuable team asset. Even in his PSL career, he appeared in only 5 matches, but scored over a 100 runs, unlike many of his senior peers. The rest, Haider Ali, Usman Qadir and Shahnawaz Dahani were low on our rankings and did not perform in the starting 11 of a lot of matches in the WC either, thus their selection can be considered questionable.

Additionally, even though almost all of the batsmen were in our rankings, the overall performance of the batting side in the WC was poor, where they were frequently unable to set and chase the target runs. The two openers, Mohammad Rizwan and captain Babar Azam, were high in our rankings but performed poorly in most of the matches, losing their wicket early on and not scoring big, and the same can be said for the middle-order of Khushdil Shah, Shan Masood, and Iftikhar Ahmed. Shan Masood had an incredible PSL career, scoring over a 1000 runs in only 34 matches (unlike his peers), hence his selection can be understood, but the rest can be considered questionable. The exception from the batting side was newcomer Mohammad Haris, whose daring approach of scoring big paid off, as opposed to the cautious form of the other senior batsmen, and all-rounder Shadab Khan, who was high in our rankings and also a two-time Man of the Match winner.

On the other hand, the bowling side performed incredibly well throughout the WC. Shaheen Afridi and Haris Rauf were high on our rankings and took a lot of crucial wickets in the WC, and the same can be said for Mohammad Wasim, Naseem Shah, and Shadab Khan. Mohammad Hasnain was low on our rankings and was not in the starting 11 of a lot of the matches, but his few appearances were good. Newcomer Naseem Shah played an incredible WC tournament, as well as in PSL, despite it being his first WC, taking crucial wickets and frequent dot balls (no runs), hence he was another valuable asset for the team.

It is important to reestablish the limitations of our data set and thus, selection, as the rankings depend on whether a player has switched teams, and their role in the teams. Hence, these figures cannot account for the nuances of the cricket sport. Additionally, our eligibility idea was arbitrary and is another limitation in our selection. These attributes can be said to reflect and discordance with the performance of the members in the teams selected from our approach, and the actual performance of the 2022 ICC T20 World Cup players.

Section 7: Outlook and Concluding Remarks

The theme of this study is team selection and creation utilizing the attributes of small-world networks found in T-20 PSL Cricket. This research employs four centrality metrics: betweenness centrality, closeness centrality, degree centrality, and clustering coefficient to evaluate the players. Few prominent players with greater centrality values or clustering coefficients may have an influential impact on many other players and play an important role in team building. We focused on measuring player performance using previous years' data on their batting and bowling performances, as well as taking a qualitative estimate based on clustering coefficient and centrality metrics generated from the network of players. Both traits were applied as a technique for team creation, and a role-based team was established. Some of the limitations of the paper were:

The dataset that was downloaded and used did not contain ball-by-ball information for four PSL matches: i) 11th March 2020 - Quetta Gladiators vs Multan Sultans; ii) 29th Feb 2020 - Islamabad United vs Peshawar Zalmi; iii) 2nd March 2018 - Karachi Kings vs Multan Sultans; and iv) 11th Feb 2016 - Peshawar Zalmi vs Karachi Kings. While we tried to fill this blank, it proved to be very time consuming and heavily prone to manual error. Ultimately, we decided to gather the ball by ball information for one of the missing matches and disregarded the other three, due to time constraints. This ball-by-ball information was collected from the PCB website. In addition, thorough checking of the player information was required as many discrepancies were identified within the acquired dataset. Using the player registry dataset, we manually checked for any player name discrepancies with cricinfoespn. We found that there were some instances of duplicates and multiple players with same names which we then attempted to correct.

For building the performance pool, Dey et. al (2016) set a minimum benchmark for performance evaluation when selecting players. Our pool of Pakistani players was already limited to a small amount and we had observed that there was a large inclusion of new young players in the later PSL seasons. This made it difficult to devise a performance benchmark as these new players did not have enough experience when compared to their more experienced team members who had been participating in the seasons much longer than them, and thus, had a greater overall performance. Furthermore, they also sorted clustering coefficient values in descending order, but we believe that this is not an accurate measurement because high clustering values can indicate low influence. Hence, the clustering-based selection proved to be very different from the other three generated teams and we believe this is a reflection of an error in the clustering criteria, but comprehending or explaining it is beyond the scope of our abilities.

We also observed that there were some differences in the roles of three players in their selection in the official ICC T20 team and their roles in their PSL teams, which can affect the accuracy of our analysis and team selection process. Additionally, the 7 season data does not account for nuances in player interactions, or take into account players that were injured and unable to play, in the reserves and not starting 11, the nature of team switching. Furthermore, these figures cannot account for the benefits based on roles. Batters are only able to perform in the batting innings of a match if their preceding peers lose their wicket and their turn comes up, whereas all bowlers in a starting 11 get their 5 overs to perform in the bowling innings. On the other hand, wicket-keepers and all-rounders can perform in both innings. Thus, these are nuances that are unaccounted for in the data.

The eligibility and non-eligiblity criteria for the performance pool of players had to be self-determined because some players do not formally announce their retirement and there is no maximum age limit established for participation in ICC T20 World Cup. We also believe that a lot of older players do not get chosen because it is widely accepted that younger players are more preferred over players that are above thirty-five. Hence, the boundaries for age and when players were last active has been set arbitrarily.

Furthermore, collecting and analyzing data on how a network changes over time can be a complex and challenging task, and it may require a significant amount of time and effort. Additionally, temporal analysis can be limited by the availability and quality of data on the network. If the data is incomplete or unreliable, it can affect the accuracy and usefulness of the analysis. Due to this, the seasonal analysis is quite limited. There is room for more potential introspection to understand the reasoning behind the emergence of the patterns identified in the PSL networks and include other measures such as centrality.

As a result, for future work, a combination of temporal analysis and network centrality measures can be explored to understand the factors behind a cricket team's victory. Changes in team squads over a period of time can potentially indicate which players are the most important as their inclusion can potentially guarantee a team's positive performance.

Bibliography

1. Davids, K., Araújoa, D., Pazc, N., Minguénsd, J., & Mendesd, J. (2010, December 9). Networks as a novel tool for studying team ball sports as complex social systems. Journal of Science and Medicine in Sport. Retrieved November 28, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S1440244010006602
2. Dey, Paramita & Ganguly, Maitreyee & Roy, Sarbani. (2016). Network Centrality Based Team Formation: A Case Study on T20 Cricket. Applied Computing and Informatics. https://www.researchgate.net/publication/311730381_Network_Centrality_Based_Team_Formation_A_Case_Study_on_T20_Cricket
3. ICC. (2022, October 14). Pakistan bolster batting stocks with change to T20 World Cup squad. ICC-Cricket. Retrieved from: https://www.icc-cricket.com/news/2850499
4. Mukherjee, S. (2012, June 21). Complex network analysis in cricket : Community Structure, player's role and performance index. Complex Network Analysis in Cricket : Community structure, player's role and performance index . Retrieved November 28, 2022, from https://www.arxiv-vanity.com/papers/1206.4835 /
5. Mukherjee, S., Easley, D., Lusher, D., Saavedra, S., Wilson, R., Iyer, S. R., Bergstrom, C. T., Watts, D. J., Albert, R., Tadic, B., Freeman, L. C., Castellano, C., Newman, M. E. J., Pan, R. K., Price, D. J. de S., Chen, P., West, J., Radicchi, F., Ben-Naim, E., … Sire, C. (2013, September 13). Quantifying individual performance in cricket - a network analysis of batsmen and Bowlers. Physica A: Statistical Mechanics and its Applications. Retrieved November 28, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0378437113008819
6. Pakistan Cricket Board. Retrieved from: https://www.pcb.com.pk/
7. Wiig, A. S., Håland, E. M., Stålhane, M., & Hvattum, L. M. (2019, December 1). Analyzing passing networks in association football based on the difficulty, risk, and potential of passes. International Journal of Computer Science in Sport. Retrieved November 28, 2022, from https://sciendo.com/de/article/10.2478/ijcss-2019-0017