Hypergraph Data analysis with PAOHVis

Paolo Buono, Miguel Ceriani and Maria Costabile

University of Bari, via Orabona, 4, 70125 Bari, Italy

Abstract

Many data analysis activities exploit graphs to model complex relationships among data. Very often such relationships are better described using hypergraphs, whose edges can connect more than two nodes. Hypergraphs are suitable to model networks of business partners or co-authorship networks with multiple authors per article. This paper is about a technique that visualizes dynamic hypergraphs called PAOH (Parallel Aggregated ordered Hypergraph). It represents vertices as parallel horizontal bars and hyperedges as parallel vertical lines and shows the evolution of hypergraphs representing them in discrete Time Slots. PAOH is described in details in an article published in 2021 in IEEE Transaction on Visualization and Computer Graphics. It has been applied in several domains, such as ego-networks, co-authorship, as well as digital humanities, it is well-suited for medium-sized hypergraphs (50-500 vertices). PAOHVis is a tool, available online, that implements this technique. In this paper we briefly describe the tool and show the application of PAOH to the soccer domain, to support the analysis of players' contributions to the goals of a team.

Keywords

Data visualization, Network analysis, Sport

1. Introduction

A graph is a flexible data structure that is able to describe multifarious complex real world data and relationships. The visual representation of graphs is a challenging topic. By representing graphs as node-link diagrams there are several problems to face, in particular when the relationships involve more than two entities. An example is co-authorship networks, where relations are articles written by more than two authors. These kinds of networks can be more accurately modeled as hypergraphs where each relation may involve several nodes.

This paper illustrates a technique that visualizes dynamic hypergraphs, called PAOH (Parallel Aggregated ordered Hypergraph), which has been developed in the last few years. The article that describes all details of PAOH was first published online on August 2019, while its printed version appeared in January 2021 in IEEE Transaction on Visualization and Computer Graphics journal [1]. PAOH has been applied to various domains. An early version of the technique visualized co-authorships of EuroVis conferences from 2000 to 2015 [2]. An application to Digital Humanities visualized the business network of a 17th century French woman merchant [3]. An

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy
paolo.buono@uniba.it (P. Buono); paolo.buono@uniba.it (M. Ceriani); paolo.buono@uniba.it (M. Costabile)
http://ivu.di.uniba.it/people/buono.htm (P. Buono); http://ivu.di.uniba.it/people/ceriani.htm (M. Ceriani); http://ivu.di.uniba.it/people/costabile.htm (M. Costabile)
0000-0002-1421-3686 (P. Buono); 0000-0002-5074-2112 (M. Ceriani); 0000-0001-8554-0273 (M. Costabile)

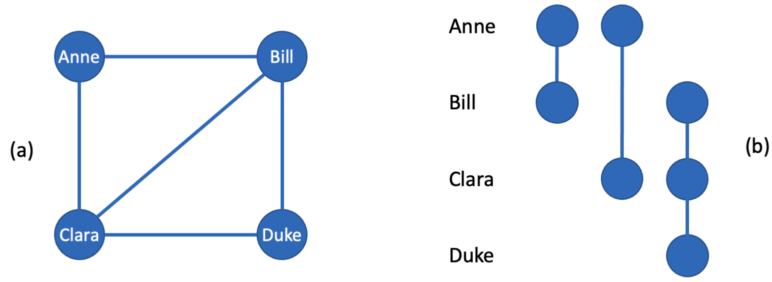

Figure 1: Co-authorships among four people shown with Node-Link and PAOH

extension of PAOH also considers ensemble clustering, analyzing a dataset related to authors' lineage in the VAST conference publication dataset [4].

A tool that implements this technique is called PAOHVis[1], and it is described in this paper; it is available online and its source code is open and maintained in a public repository[2]. After providing a simple example to illustrate the PAOH technique in Section 2, Section 3 describes some main features of PAOHVis using as case study the application of PAOH to the soccer domain, in order to support the analysis of players' contributions to the goals of a team. Section 4 provides conclusions and future works.

2. PAOH in short

PAOH was initially inspired by Biofabric [5], a technique that addresses the scalability of the hypergraph visualization but does not considers the network topology evolution over time, which in many cases is required. A technique that takes time into account is DyNetVis [6]; however, it considers time points instead of Time Slots as PAOH does, this do not allow to show different periods, which can be useful for the analysis. Another hypergraph visualization is Hyper-matrix [7]; in principle it is scalable but, when the number of nodes increases, the user loose details about the nodes. Even if there are other ways of visualizing hypergraphs, they are not scalable and thus they are not very significant (see for example [8]).

The PAOH technique is characterized by parallel hyperedges that are visualized along a time axis [1]. Vertices are ordered vertically according to a criterion. In order to give a simple example of the visualization, let us consider data about co-authorship. People are related if they are co-authors of a same publication. Since co-authors are often more than two, the usual node-link representation is not adequate. In Figure 1 (a), the authors are related because they share some publications, but it is not evident how many they share and which are the authors of each publication. Figure 1 (b) shows the PAOH visualization of the same co-authorship network. Each node is positioned vertically, in alphabetical order of the author name. All nodes involved in a co-authorship are connected by a vertical parallel line; this avoids edge crossing and well expresses the relationship that involve more than two people. It is evident that they

[1]https://www.aviz.fr/Research/Paohvis, last access: April 19th, 2021
[2]https://gitlab.inria.fr/aviz/paohvis

share three publications and who are the co-authors. The publications are visualized from left to right, chronologically.

Several definitions of dynamic graphs exist in the literature. The definition from the survey on dynamic graphs [9], is that a *dynamic graph* is a sequence of graphs $\{G_1, \cdots, G_n\} | G_i = (V, E_i), 1 \leq i \leq n$ sharing the same vertices but with a topology varying over time, in order to show how the relations evolve during time. A *hypergraph* G is a pair $G = (V, H)$, where V is the set of *vertices* and $H \subseteq \wp(V) \setminus \{\emptyset\}$ is the set of *hyperedges* (each hyperedge is a non-empty set of vertices). A *temporal hypergraph* is a sequence of hypergraphs $\{G_1, \cdots, G_n\} | G_i = (V, H_i), 1 \leq i \leq n$ that share the vertices and have the topology that varies over time. Each set of hyperedges H_i refers to a given Time Slot.

3. PAOHVis tool

In order to describe the PAOHVis tool that implements the PAOH technique, we refer to the soccer domain and more specifically, the challenge of visualizing discrete data about goal-leading shots and ball passes between players of the same soccer team. PAOH supports the study of the performance of players and teams in multiple contexts of varying granularity: multiple matches of the same team; all the teams and matches in a tournament; multiple tournaments at once. Soccer is today the most popular sport, with approximately 250 million players and 4 billion fans worldwide [10, 11]. As the concept of passes and scoring is common to many team sports, the described method is applicable to other team sports.

The dataset used for this application contains data from the first 26 weeks of the 2019-2020 season of the Italian Major league (Serie A)[3]. The data has been obtained by manual annotation from the *match highlights* video of each match, with the help of external sources for the players in the match. There are 20 teams in Serie A, 10 matches each week (each team plays once a week) during 26 weeks, for a total of 260 matches. In these matches 710 goals were scored and 353 players contributed at least a goal. The dataset also contains players' attributes: the team they are part of and their role in that team.

A *match phase* is composed of several passes among players of the same team, and is defined as a segment of a match in which a team keeps the possession of the ball. The phases of interest for this case study are those leading to successful scoring in the opponents' goal, called *goal-leading phases*. PAOH can be used also to show all phases of a match. To each phase, we associate information of the involved players and the match in which it happened. A phase is modeled as a hyperedge of a hypergraph and vertices are the players. In PAOHVis, all matches played in a week are grouped in a Time Slot, which is a hypergraph. The tournament is composed of the temporal sequence of all hypergraphs.

Figure 2 shows the user interface of PAOHVis. At the top of the app there is a green toolbar whose items refer to different sections of the tool; the orange item "View" is the one currently selected. At the center of the toolbar the name of the current dataset ("SerieA_1920_2.1.0-v2.4.json") is shown. The main area shows the dynamic hypergraphs. In Figure 2 players of two teams, i.e. Inter and Juventus, have been selected and shown ordered by the number of goal-leading phases they participated in. The players in the dataset can be filtered by team,

[3]Dataset available at: http://paovis.ddns.net/data/ItalianMajorLeague19-20.json, last visited: June 11th

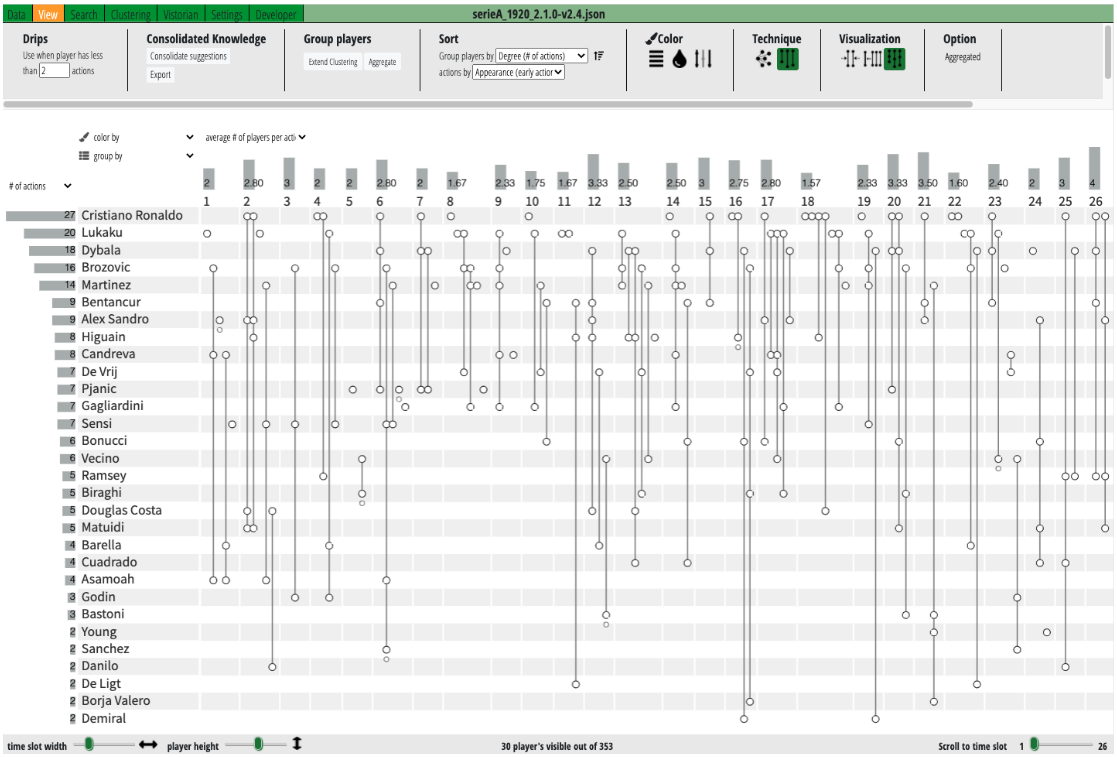

Figure 2: PAOHVis UI showing players of Inter and Juventus, two of the best teams in 2020 of "Serie A", the Major Italian league. The players are ordered by the number of goal-leading phases they participate.

player role, week. At the bottom of the user interface there are several controls through which the user can modify the visualized elements.

Among the several options *color by* and *group by* , visible on top left of the main area, revealed very useful. *color by* colors the vertices and their background while *group by* groups data according to a specific attribute selected by the user. For Example, the vertices in Figure 3 are colored by team and grouped by team. The updated screen now shows the attribute *team* chosen by the user. When the grouping is activated another column appears on the left of the players' names showing a label containing the group name.

Various sorting criteria are available, such as barycentric (the most connected players are vertically closer), degree (players are ordered vertically from the most to the less connected, as it is shown in Figure 2), group (players belonging to the same team appears together, as it is shown in Figure 3).

Figure 3 shows data about Inter and Juventus team. One player for each team has been highlighted by the user, respectively, Brozovic and Dybala. Both players did not participate much in goal-leading phases with the other principal attacker of the same team, respectively Lukaku and Cristiano Ronaldo. Indeed, Brozovic shares with Lukaku only 7 out of 20 goal-leading phases, similarly, Dybala shares only 7 out of 27 goal-leading phases. The two players are equivalent in terms of offensive capacity, both participated in a similar amount of goal-leading

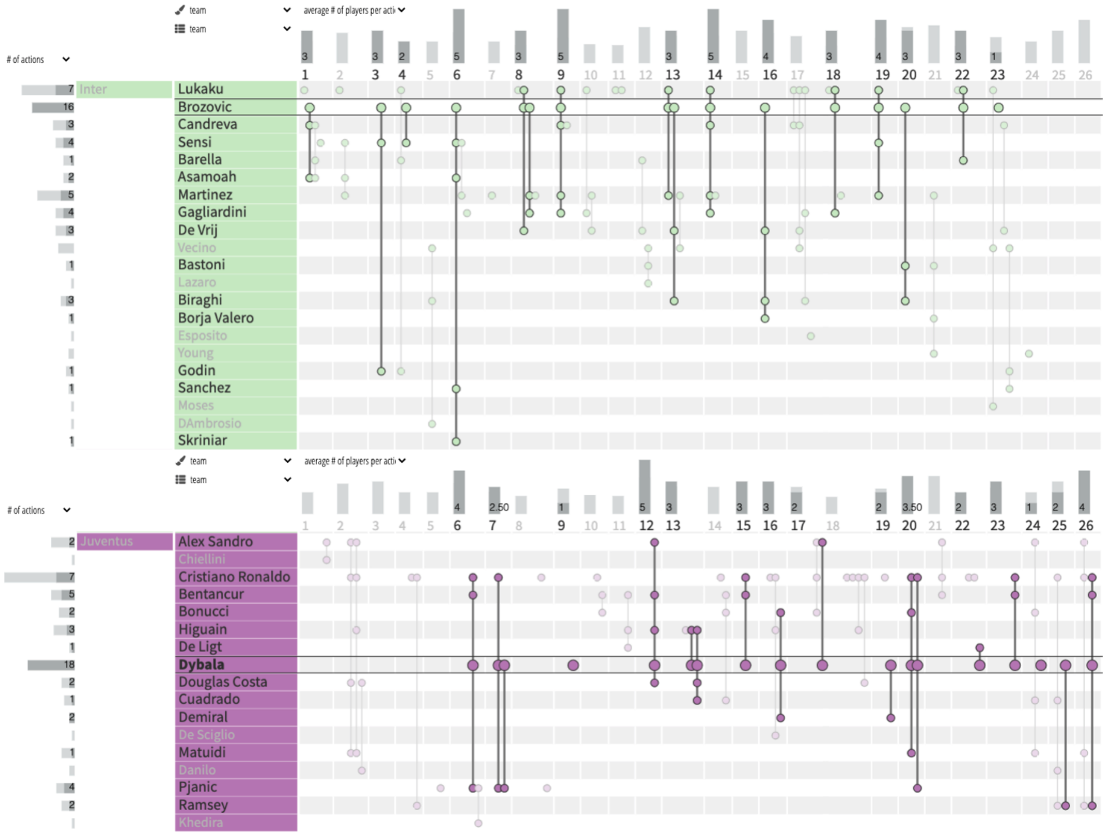

Figure 3: Comparison between two players of different teams: Brozovich (Inter) vs Dybala (Juventus).

phases (Brozovic 16 and Dybala 18) and both collaborated with most of their teammates. The single dots (one for Brozovic and two for Dybala) indicate free kicks. Grayed players never participate in goal-leading phases with the selected player. Brozovic does not pass the ball with 6 players (i.e. Vecino, Lazaro, Esposito, Young, Moses, D'Ambrosio), while Dybala with 4 players (i.e. Chiellini, De Sciglio, Danilo, Khedira). The total number of players participating in goal-leading phases is higher in the Inter team (21) than Juventus (17). This reveals that the Juventus attack uses globally fewer players that score more than Inter players. Finally, Brozovic is active until week 23, while Dybala plays since week 6 and intensifies his presence until the last week.

Figure 4 shows data of Hellas Verona team, which had three periods: week [1–9], the team scores low, week [10–23] the team is very active (Lazovic distinguished in most of the phases) and week [24–26] the team does not score at all. The horizontal histogram on the left (also visible in Figure 2 and Figure 3) reports the number of phases in which a player participated, while the histogram at the top reports, for each week, the number of players participating in goal-leading phases in the week.

PAOHVis offers also an aggregated view in which the representation is compacted by

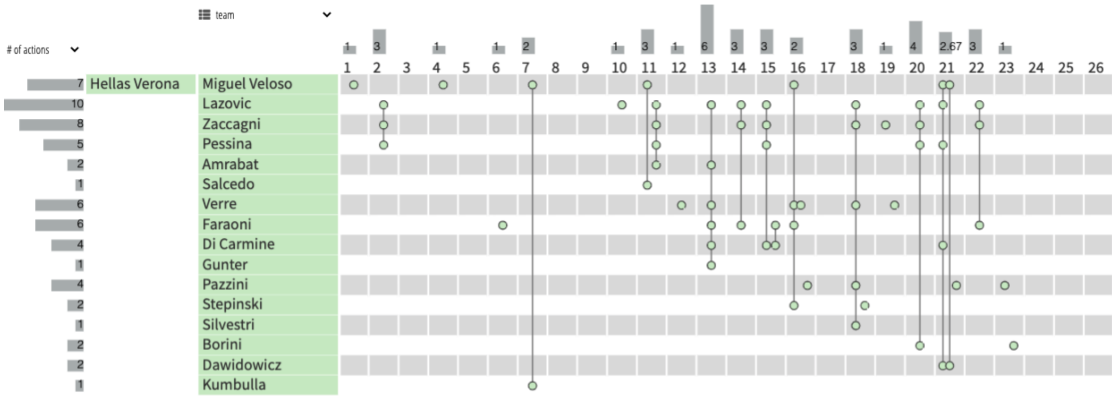

Figure 4: Goal-leading phases of Hellas Verona team. Three periods are visible: 1-9; 10-23; 24-26.

removing the horizontal constraint of weeks. This option leaves room for more data and it is useful when it is less important to know when a match occurs than to know who is involved in a goal-leading action. Figure 5 shows the aggregate view of Hellas Verona team. Three different moments of the interaction are reported, each has a different player highlighted: on the left, Miguel Veloso is highlighted; he participated in about a third of the goal-leading phases. In the middle and at the right Lazovic and Zaccagni are highlighted, respectively. These two players overlap a lot, which means they pass the ball a lot between them; on the contrary, Miguel Veloso is a more individualist player. More than a half of the players contributed to less than four goal-leading phases each. Such a group represent a third of the goal-leading phases, shown on the right of each of the three areas in Figure 5. Further information about the soccer domain and the use of PAOHVis in this domain is reported in [12].

PAOHVis offers many other features; we briefly described some of them. Time Slots are always displayed by their natural chronological order; however, within each Time Slot, the edges can be sorted differently. The edges can also be sorted according to their length. A packing algorithm optimizes the horizontal space by reorganizing hyperedges in each Time slot independently. It reuses the same horizontal position for hyperedges that do not overlap

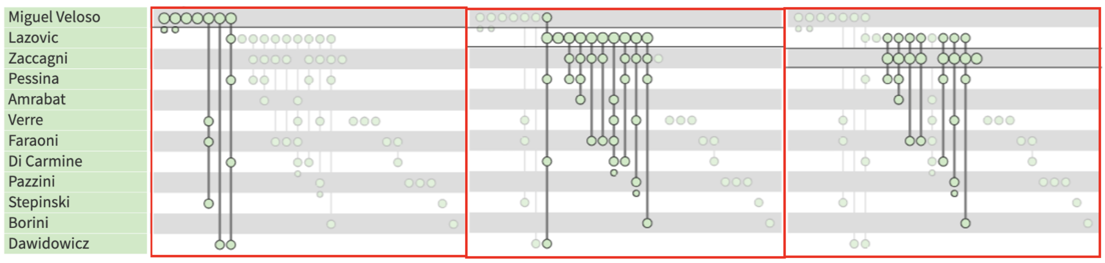

Figure 5: Aggregated view of Hellas Verona team. Each red box shows a player highlighted, from left to right: Veloso, Lazovic and Zaccagni

vertically. Optimizing packing is an NP-hard problem; PAOHVis uses the *first fit* bin packing approximation algorithm that inserts each edge in turn, from left to right, where it can fit without overlap [13] . The basic layout is monochrome so that the color can be used for specific needs. For example, the vertices background and the color of the dots can be used to indicate the vertices that belong to a category. Hovering over a hyperedge, similar hyperedges (those that share a similar set of vertices) are highlighted, in order to reveal recurrent relationships. Filtering helps reducing the size of the dataset. Double clicking on a vertex or searching a vertex name in the search box filters out all the vertices that are not connected to it. Faded dots reveal that visible vertices have relations with other vertices that have been filtered out. A specific filtering section has been added in the tool. It allows the user to apply various filters on attributes of the dataset.

4. Conclusions and future work

PAOH is a recent technique that is already appreciated in the InfoVis community. It is a unique tool that aims at helping people to analyze dynamic hypergraphs. It has successfully been adopted in several domains domains like the sport domain reported in this paper, in order to visualize goal-leading phases in soccer matches. As future work, the research will investigate how to improve the scalability of the PAOH technique. One direction is improving the clustering capabilities of PAOHVis to suggest aggregations and grouping similar teams or players. The challenge here is to find proper visualizations of the results and proper user interaction.

Acknowledgments

This work is partially supported by the Italian Ministry of University and Research, PON R&I 2014–2020 – Attraction and International Mobility (AIM) – project n. 1852414. The authors thank Alberto Rutigliano for his contribution in gathering the data analyzed in this article.

References

[1] P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, J.-D. Fekete, Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization, IEEE TVCG 27 (2021) 1–13. doi:10.1109/TVCG.2019.2933196, date of publication: (Aug 2019).

[2] P. Valdivia, P. Buono, J.-D. Fekete, Hypenet: Visualizing dynamic hypergraphs, in: Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Posters, EuroVis '17, Eurographics Association, Goslar, DEU, 2017, p. 33–35. URL: https://doi.org/10.2312/eurp.20171162. doi:10.2312/eurp.20171162.

[3] P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, J.-D. Fekete, Using Dynamic Hypergraphs to Reveal the Evolution of the Business Network of a 17th Century French Woman Merchant, in: VIS 2018 - 3rd Workshop on Visualization for the Digital Humanities, Berlin, Germany, 2018. URL: https://hal.inria.fr/hal-01855675.

[4] A. Pister, P. Buono, J. D. Fekete, C. Plaisant, P. Valdivia, Integrating prior knowledge in mixed-initiative social network clustering, IEEE Transactions on Visualization and Computer Graphics 27 (2021) 1775–1785. doi:10.1109/TVCG.2020.3030347.

[5] W. J. Longabaugh, Combing the hairball with BioFabric: a new approach for visualization of large networks, BMC Bioinformatics 13 (2012) 275. URL: http://dx.doi.org/10.1186/1471-2105-13-275. doi:10.1186/1471-2105-13-275.

[6] C. D. G. Linhares, B. A. N. Travençolo, J. G. S. Paiva, L. E. C. Rocha, Dynetvis: A system for visualization of dynamic networks, in: Proceedings of the Symposium on Applied Computing, SAC '17, ACM, New York, NY, USA, 2017, pp. 187–194. URL: http://doi.acm.org/10.1145/3019612.3019686. doi:10.1145/3019612.3019686.

[7] M. T. Fischer, D. Arya, D. Streeb, D. Seebacher, D. A. Keim, M. Worring, Visual analytics for temporal hypergraph model exploration, IEEE Transactions on Visualization and Computer Graphics 27 (2021) 550–560. doi:10.1109/TVCG.2020.3030408.

[8] N. H. Riche, T. Dwyer, Untangling euler diagrams, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1090–1099. doi:10.1109/TVCG.2010.210.

[9] F. Beck, M. Burch, S. Diehl, D. Weiskopf, A Taxonomy and Survey of Dynamic Graph Visualization, Computer Graphics Forum 36 (2017) 133–159. URL: http://dx.doi.org/10.1111/cgf.12791. doi:10.1111/cgf.12791.

[10] F. I. de Football Association (FIFA), Fifa big count 2006: 270 million people active in football, ????. URL: https://resources.fifa.com/image/upload/big-count-estadisticas-520058.pdf?cloudid=mzid0qmguixkcmruvema.

[11] F. I. de Football Association (FIFA), 2018 fifa world cup russia - global broadcast and audience summary, ????. URL: https://resources.fifa.com/image/upload/2018-fifa-world-cup-russia-global-broadcast-and-audience-executive-summary.pdf?cloudid=njqsntrvdvqv8ho1dag5.

[12] B. Paolo, M. Ceriani, M. Costabile, P. Valdivia, Analyzing goal-leading actions in soccer with a visual tool, in: Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter: Frontiers of HCI, CHItaly '21, accepted., 2021.

[13] E. G. Coffman, Jr., M. R. Garey, D. S. Johnson, Approximation Algorithms for Bin Packing: A Survey, in: D. S. Hochbaum (Ed.), Approximation Algorithms for NP-hard Problems, PWS Publishing Co., Boston, MA, USA, 1997, pp. 46–93. URL: http://dl.acm.org/citation.cfm?id=241938.241940.