

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Graph-based visualization for time-evolving data: visualization of the temporal evolution of players in national teams and clubs

Patrícia do Carmo Nunes Oliveira



Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Alexandre Miguel Barbosa Valle de Carvalho

July 16, 2024

**Graph-based visualization for time-evolving data:
visualization of the temporal evolution of players in
national teams and clubs**

Patrícia do Carmo Nunes Oliveira

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. António Augusto de Sousa

Examiner: Prof. Paulo Miguel de Jesus Dias

Supervisor: Prof. Alexandre Miguel Barbosa Valle de Carvalho

July 16, 2024

Resumo

O ser humano tende a compreender os dados de forma mais eficaz quando estes são apresentados de maneira visual e memorável, especialmente no contexto do universo *online*, dadas as diferentes características dos utilizadores e os níveis de literacia visual. A ZOS, Lda. (ZOS) tem vindo a informar adeptos e profissionais de desporto desde 2003 em diversos meios de comunicação.

A empresa é atualmente a proprietária da maior base de dados de futebol do mundo, abrangendo informações sobre jogadores, equipas, seleções nacionais, competições e jogos. A complexidade, conectividade e dimensões temporais das informações disponíveis representam um desafio significativo: como capturar e apresentar entidades dinâmicas com relações multifacetadas e em evolução de uma forma facilmente compreensível.

É necessário apresentar essa informação aos utilizadores de uma maneira acessível, apelativa, simples e intuitiva, garantindo clareza sem sobrecarregar os utilizadores. Além disso, a informação deve manter-se visualmente consistente ao longo do processo de narrativa, proporcionando *insights* valiosos para analistas, treinadores e entusiastas do desporto.

O principal objetivo deste trabalho é estudar técnicas para representar visualmente dados altamente interligados e em constante evolução temporal. Este trabalho testa a hipótese de que visualizações baseadas em grafos são adequadas para apresentar dados em evolução temporal e que as dependências temporais são eficazmente transmitidas através de transições animadas entre os grafos, resultando num método de visualização adaptado à natureza complexa e dinâmica dos dados enriquecido com uma dimensão temporal adicional. Para este propósito, é proposta uma solução de visualização animada baseada em grafo circular de pesos dinâmicos.

Para validação, é especificado um procedimento experimental, envolvendo um protótipo funcional desenvolvido usando a biblioteca *D3.js*. São realizados testes com dados de futebol em evolução temporal (evolução temporal de jogadores em seleções nacionais e equipas em Campeonatos do Mundo, com ênfase em transições de jogadores, trajetórias de carreira e dinâmicas de equipa ao longo do tempo) para avaliar a usabilidade, eficácia e eficiência da solução proposta na transmissão visual de informações.

Os resultados demonstram que as visualizações baseadas em grafos para dados em evolução temporal podem representar e transmitir corretamente as informações aos espectadores, validando assim a hipótese proposta e destacando o potencial para melhorar a apresentação de dados intrincados e dinâmicos no contexto das estatísticas de futebol e noutros. No entanto, há espaço para melhorias, particularmente na usabilidade em dispositivos táteis.

Palavras-chave: Visualização de Informação, Informação altamente Conectada, Informação Histórica, Infografias, Relações em Evolução

Classificação ACM: Computação centrada no ser humano → Visualização → Domínios de aplicação de visualização → Visualização de informação

Abstract

Humans tend to understand data better when presented visually and memorably, especially in the context of the online universe, given the different user profiles and visual literacy levels. ZOS, Lda. (ZOS) has been informing sports fans and professionals since 2003 in several media outlets.

The company is currently the home of the world's largest soccer database, featuring comprehensive information on players, clubs, national teams, competitions, and matches. The complexity, connectivity, and temporal dimensions of the available information pose a significant challenge: how to capture and present dynamic entities with multifaceted and evolving relationships in a way that is easy to understand.

It is necessary to present such information to users in an accessible, appealing, simple, and intuitive way that ensures clarity without overwhelming users. Furthermore, the information must remain visually consistent throughout the storytelling process, providing valuable insights to analysts, coaches, and sports enthusiasts.

The main goal of this work is to study techniques for visually representing highly connected and time-evolving data. This work tests the hypothesis that graph-based visualizations are well suited for visually presenting time-evolving data and that temporal dependencies are effectively conveyed through animated transitions between the graphs, resulting in a visualization method tailored to the complex and dynamic nature of the data enriched with an additional temporal dimension. A solution, animated-based dynamic weight circular graph visualization, is proposed for this purpose.

For validation, an experimental procedure is specified, involving a functional prototype developed using *D3.js* library. Tests with time-evolving soccer data (temporal evolution of players in national teams and clubs in World Cups, emphasizing player transitions, career trajectories, and team dynamics over time) are conducted to assess the proposed solution's usability, effectiveness, and efficiency in visually conveying information.

The results show that graph-based visualizations for time-evolving data can correctly represent and convey the information to viewers, therefore validating the proposed hypothesis and underscoring the potential for enhancing the presentation of intricate and dynamic data in the context of soccer statistics and beyond. However, there is room for improvement, particularly in usability in touch devices.

Keywords: Information Visualization, High-connected Information, Historical Information, Infographics, Evolving Relationships

ACM Classification: Human-centered computing → Visualization → Visualization application domains → Information visualization

Acknowledgements

I am deeply grateful to everyone who helped and supported me throughout my journey as a student. Your kind words and advice during the toughest moments were invaluable in keeping me moving forward. The challenges you provided that pushed me out of my comfort zone helped me grow both personally and professionally.

Firstly, I would like to thank my supervisor, Alexandre Valle de Carvalho, who closely supported and guided me throughout this work. The weekly meetings provided essential feedback, cleared the next steps, and kept me motivated. I sincerely appreciate the trust shown in my work while continually challenging me.

I extend my gratitude to all the ZOS collaborators for their warm welcome, kindness, and friendship. Their feedback and involvement were crucial to the success of my work.

A special word to the ZOS owners, Pedro Dias ("boss Pedro") and Marco Sousa ("boss Marco"), for their constant concern and readiness to assist with any matter. They ensured I had all the necessary resources to complete my work, from technological equipment to full access to the facilities. The open environment for free speech and the biweekly meetings provided genuine feedback, direction, and motivation.

I am also thankful to my friends who provided moments of fun and relaxation during this stressful period. Particularly, to Margarida Ferreira and Henrique Nunes for their fellowship during college. Their individuality, knowledge, and approach to life enriched my learning experience.

I want to express my heartfelt gratitude to my mother and brother for their support and for calling me out when needed. Their sacrifices were what made everything possible, and much of who I am, I owe to them.

I also want to thank my boyfriend, Pedro Garcia, for his love, immense patience, and support. From walks to shared crepes, your presence has completed my life and brought me happiness every day. Thank you for always believing in me.

Lastly, I thank the members of Stand 4 Good, notably Mafalda Bastos and Susana Prata, for their mentorship and guidance, providing opportunities for personal and professional growth.

Through my experiences and the support of all these people, I have learned the importance of striving to reach our true potential, being honest and supportive, and never giving up even when the world seems to be falling apart. There are always people who care about us, and it's okay to ask for help. Our ability to overcome challenges reveals our true strength.

Patrícia Oliveira

*“There’s nothing more I could ever want other
than for it to last forever but... It can’t.”*

The Vampire Diaries

Contents

1	Introduction	1
1.1	Context and Motivation	1
1.2	Problem Identification	2
1.3	Hypothesis	2
1.4	Research Questions	3
1.5	Methodology	3
1.6	Document Structure	4
2	Literature Review	5
2.1	Research Domain	5
2.2	Identification Phase	6
2.2.1	Refinement Process	7
2.2.2	Search Criteria	9
2.3	Screening Phase	9
2.3.1	Inclusion and Exclusion Criteria	10
2.3.2	Remove Duplicates	10
2.3.3	Screened Records	11
2.4	Eligibility Phase	11
2.4.1	Full-text Access	11
2.4.2	Data Collection	11
2.4.3	Selected Records	12
2.5	Results and Discussion	16
2.6	Beyond PRISMA	17
2.7	Summary	17
3	State of the Art	18
3.1	Information Visualization	18
3.1.1	Sports Data Visualization	19
3.1.2	Data-oriented Visualization Techniques	20
3.1.3	Interaction and Animation in Information Visualization	22
3.1.4	Storytelling in Information Visualization	25
3.1.5	Web Environment in Information Visualization	26
3.2	Graph-Based Visualization	26
3.2.1	Graph Mining	28
3.2.2	Dynamic Graph Visualization	29
3.2.3	Technologies	37
3.2.4	Applications	40
3.3	Summary	41

4	Problem Statement	43
4.1	Problem	43
4.2	Research Questions	44
4.3	Solution Requirements	44
4.4	Scope	45
4.5	Summary	46
5	Solution	47
5.1	Proposed solution	47
5.2	Prototype Implementation	49
5.2.1	Data Gathering	50
5.2.2	Technology Selection	51
5.2.3	Graphs Layout	53
5.2.4	Timeline Integration	59
5.2.5	Additional Controls	60
5.2.6	Components Organization	61
5.2.7	Contextual Elements	62
5.2.8	Documentation and Flexibility	63
5.3	Summary	63
6	Experimental Protocol	64
6.1	Test Methodology	64
6.1.1	General Description	64
6.1.2	Gather Participants	65
6.1.3	Experiment Procedure	66
6.2	Assessment Methods	67
6.2.1	Formative Questionnaire	67
6.2.2	Internal Analytics	72
6.3	Pilot Tests	75
6.4	Summary	77
7	Results	78
7.1	Formative Questionnaire	78
7.1.1	Outliers Identification	78
7.1.2	Demographic Statistics	80
7.1.3	Task Statistics	80
7.1.4	Usability Statistics	90
7.2	Internal Analytics	94
7.2.1	Google Analytics	94
7.2.2	Hotjar	107
7.3	Discussion	110
7.3.1	Outliers Identification Criteria	110
7.3.2	Demographics and Recruitment Strategies	111
7.3.3	Tasks Performance and Demographics Influence	111
7.3.4	System Usability and Participant Feedback	112
7.3.5	Internal Analytics and User Engagement	113
7.3.6	Final Remarks	113
7.4	Summary	114

8	Conclusions	115
8.1	Final Remarks	115
8.2	Future Work	118
	References	120
A	Prototype Code and Resources	125
B	Formative Questionnaire	126
C	Supplementary Results	131

List of Figures

2.1	PRISMA flowchart.	6
3.1	ForVizor: Interface of the system [24].	19
3.2	Visualization of the overview of the player attributes [52].	20
3.3	Complementary views over multivariate data using Kohonen’s SOM (a), Parallel Coordinates (b), and t-SNE methods (c) [4].	21
3.4	Helical graphs representing the variation from 2000 to 2021 of the tropical (A) and the boreal (B) forests [19].	22
3.5	CroP’s interactive user interface of a temporal dataset: options sidebar and three different panels [14].	23
3.6	Animated Vega-Lite specification to encode Gapminder animation [67].	24
3.7	The story synthesis framework illustrated with social media data [21].	25
3.8	Layouts of the same static weighted relational data: (a) node-link radial layout. (b) node-link bipartite layout. (c) and (d) adjacency matrix and the corresponding adjacency list [7].	27
3.9	Network visualization of twitter data with Indonesian cyberbullying classification module [46].	28
3.10	Dynamic network dataset (a) represented by a Massive Sequence View (b) [25].	30
3.11	Pattern enhancement workflow in flow maps visualization [66].	30
3.12	Small multiples example showing timeslices 0, 2, 4, 6, and 8 side-be-side [12].	31
3.13	Graph structural and temporal encodings: Juxtaposition (A, D), Superimposition (B, E), and Animation with Playback Controls (C, F) [28].	31
3.14	Hierarchical structures in a bubble layout [20].	32
3.15	Node-link diagrams of real-world datasets: examples of large and dense graph structures [8].	33
3.16	Adjacency matrices of a dense dataset reordered using distinct ordering strategies [8].	33
3.17	Combination of node-link diagrams and adjacency matrices to ease the identification of visual patterns [8].	33
3.18	Solution of a problem in the geometry of position by Leonhard Euler: <i>Seven Bridges of Königsberg</i> [7].	34
3.19	E-MCBN sampling for MSV in real-world scenarios: students’ phone call dataset [25].	35
3.20	Multiscale Snapshots temporal and structural aggregates approach [11].	35
3.21	Network topology: Tree (a), Star (b), Circular (c), Low (d), and Clique (e) [39].	36
3.22	Dynamic network layout in 4 consecutive timesteps [64].	36
	(a) Layouts by Gorochowski method.	36
	(b) Layouts by Wei method.	36

3.23	Overview of <i>DiffSeer</i> on explicitly visualizing the differences between adjacent timeslices to support the analysis of the dynamic weighted graph evolution over a long time [23].	37
3.24	<i>Collaboration Spotting X</i> detail network and detail schema [6].	38
3.25	A framework for spatial-temporal cluster analysis. [48].	38
3.26	<i>Multiscale Snapshots</i> visualization (A) and the query interface (B) [11].	39
3.27	Correlation of topics in self-presentations, grouped by age (a), education (b), and sex (c) [51].	39
3.28	<i>Ecoxight</i> platform UI. [5].	40
3.29	Visualizations in soccer-related applications.	41
	(a) Team performance comparison visualization: Real Madrid FC vs. Stade Rennais FC [52].	41
	(b) Passes network in the Huskies team visualization [37].	41
5.1	Proposed Solution - Visualization Structure Schema.	48
5.2	Prototype's data flow and architecture.	49
5.3	Graph layout experiments using D3 and <i>d3-force</i>	52
	(a) Experiment 1.	52
	(b) Experiment 2.	52
	(c) Experiment 3.	52
	(d) Experiment 4.	52
5.4	Graph animation experiment using D3 in three consecutive timeslices.	53
	(a) Timeslice 1.	53
	(b) Timeslice 2.	53
	(c) Timeslice 3.	53
5.5	Exploration of the visual variable shape to represent club nodes in the visualization.	54
	(a) Represented only by circles.	54
	(b) Represented only by logos.	54
	(c) Represented with circles placed behind logos.	54
	(d) Represented either by a circle or logo.	54
5.6	Exploration of the visual variable color to represent club nodes in the visualization.	55
	(a) Represented only by colored circles.	55
	(b) Represented with colored circles placed behind logos.	55
	(c) Represented with colored circles with lower opacity placed behind logos.	55
	(d) Represented either by a colored circle or logo.	55
5.7	Exploration of the club logos for distinguishing player clusters in the visualization.	56
	(a) Club logos placed behind clusters with spatially separated groups.	56
	(b) Club logos placed behind clusters with groups close together.	56
5.8	Resulting layouts from the applied forces across the visualization graphs.	58
	(a) Overview graph.	58
	(b) Club details graph.	58
	(c) National team details graph.	58
5.9	Timeline design and organization resembling a music player.	59
5.10	Overview graph for the Germany 2006 World Cup sorted by distinct criteria.	60
	(a) Outer nodes sorted by name.	60
	(b) Outer nodes sorted by continent.	60
5.11	Nodes outlined in the visualization for Qatar 2022 when searching for <i>Cristiano Ronaldo</i>	61
	(a) Results in the overview graph.	61

(b) Results in the club details graph.	61
5.12 Organization of components in the visualization depending on screen size.	62
(a) Layout on larger screens.	62
(b) Layout on smaller screens.	62
5.13 Header integrated into the visualization page.	62
5.14 Footer integrated into the visualization page.	63
6.1 Custom Dimensions configured in Google Analytics.	73
6.2 Comparison of search result layout improvements from the pilot tests.	76
(a) Original layout.	76
(b) Updated layout.	76
7.1 Bar chart of the distribution of the number of incorrect answers (errors) to the task-specific section of the questionnaire among participants.	83
7.2 Box-plot of time spent per number of incorrect answers among participants.	84
7.3 Bubble chart of the number of incorrect answers (errors) among participants with the same levels of expertise.	86
7.4 Heatmap of the mean of incorrect answers (errors) among participants with the same levels of expertise.	86
7.5 Heatmap of the mean of time spent among participants with the same levels of expertise.	87
7.6 Box-plot of ratings per SUS question among participants.	91
7.7 Overview of visits and user activity during the experiment period.	94
7.8 Demographic details of users during the experiment period.	95
7.9 Technology details of users during the experiment period.	96
7.10 Engagement details of users during the experiment period.	96
7.11 Total events recorded during the experiment period.	97
7.12 Breakdown of <i>node_highlight</i> events by category and graph type.	97
7.13 Most highlighted nodes of <i>node_highlight</i> events by graph type.	98
(a) Nodes in the cluster graph.	98
(b) Nodes in the detail graph.	98
7.14 Breakdown of <i>node_highlight</i> events by timeslice.	99
7.15 Breakdown of <i>node_in_detail</i> events by category and graph type.	99
7.16 Most viewed nodes of <i>node_in_detail</i> events by graph type.	100
(a) Nodes in the cluster graph.	100
(b) Nodes in the detail graph.	100
7.17 Breakdown of <i>node_in_detail</i> events by timeslice.	101
7.18 Breakdown of <i>control_timeline</i> events by category.	101
7.19 Breakdown of <i>change_sort</i> events by value.	102
7.20 Breakdown of <i>change_filter</i> events by category.	103
7.21 Most selected values of <i>change_filter</i> events by category.	104
(a) Values in filter national teams by continent.	104
(b) Values in filter clubs by continent.	104
(c) Values in filter clubs by country.	104
7.21 Most selected values of <i>change_filter</i> events by category.	105
(d) Values in filter players by position.	105
7.22 Breakdown of <i>change_search</i> events by value.	105
7.23 Breakdown of node events by graph type segmented by device category.	106
(a) Segmentation in <i>node_highlight</i> events.	106

(b) Segmentation in <i>node_in_detail</i> events.	106
7.24 Hotjar demographics and technology overview during the experiment period. . .	108
7.25 Hotjar heatmap over desktop clicks during the experiment period.	109
7.26 Hotjar heatmap over desktop mouse movement during the experiment period. . .	109
7.27 Hotjar heatmap over mobile clicks during the experiment period.	110
B.1 Evaluation survey - Informed Consent Form.	126
B.2 Evaluation survey - Pre-test Questionnaire.	127
B.3 Evaluation survey - Task-specific Questionnaire (Part 1).	128
B.4 Evaluation survey - Task-specific Questionnaire (Part 2).	129
B.5 Evaluation survey - Post-test Questionnaire.	130
C.1 Histogram of the distribution of age among participants.	131
C.2 Bar chart of the distribution of information visualization knowledge among participants.	132
C.3 Bar chart of the distribution of sports knowledge among participants.	132
C.4 Bar chart of the distribution of responses to question T1 (Task 1) among participants.	133
C.5 Bar chart of the distribution of responses to question T2 (Task 2) among participants.	133
C.6 Bar chart of the distribution of responses to question T3 (Task 3) among participants.	134
C.7 Bar chart of the distribution of responses to question T4 (Task 4) among participants.	134
C.8 Bar chart of the distribution of responses to question T5 (Task 5) among participants.	135
C.9 Bar chart of the distribution of responses to question T6 (Task 6) among participants.	135
C.10 Bar chart of the distribution of responses to question T7 (Task 7) among participants.	136
C.11 Bar chart of the distribution of responses to question T8 (Task 8) among participants.	136
C.12 Bar chart of the distribution of responses to question T9 (Task 9) among participants.	137
C.13 Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T1 (Task 1) among participants.	138
(a) Correct and incorrect answers per information visualization knowledge to question T1 (Task 1).	138
(b) Correct and incorrect answers per sports knowledge to question T1 (Task 1).	138
C.14 Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T2 (Task 2) among participants.	139
(a) Correct and incorrect answers per information visualization knowledge to question T2 (Task 2).	139
(b) Correct and incorrect answers per sports knowledge to question T2 (Task 2).	139
C.15 Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T3 (Task 3) among participants.	140
(a) Correct and incorrect answers per information visualization knowledge to question T3 (Task 3).	140
(b) Correct and incorrect answers per sports knowledge to question T3 (Task 3).	140
C.16 Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T4 (Task 4) among participants.	141
(a) Correct and incorrect answers per information visualization knowledge to question T4 (Task 4).	141
(b) Correct and incorrect answers per sports knowledge to question T4 (Task 4).	141
C.17 Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T5 (Task 5) among participants.	142
(a) Correct and incorrect answers per information visualization knowledge to question T5 (Task 5).	142

(b)	Correct and incorrect answers per sports knowledge to question T5 (Task 5).	142
C.18	Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T6 (Task 6) among participants.	143
(a)	Correct and incorrect answers per information visualization knowledge to question T6 (Task 6).	143
(b)	Correct and incorrect answers per sports knowledge to question T6 (Task 6).	143
C.19	Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T7 (Task 7) among participants.	144
(a)	Correct and incorrect answers per information visualization knowledge to question T7 (Task 7).	144
(b)	Correct and incorrect answers per sports knowledge to question T7 (Task 7).	144
C.20	Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T8 (Task 8) among participants.	145
(a)	Correct and incorrect answers per information visualization knowledge to question T8 (Task 8).	145
(b)	Correct and incorrect answers per sports knowledge to question T8 (Task 8).	145
C.21	Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T9 (Task 9) among participants.	146
(a)	Correct and incorrect answers per information visualization knowledge to question T9 (Task 9).	146
(b)	Correct and incorrect answers per sports knowledge to question T9 (Task 9).	146
C.22	Box-plot of the distribution of the number of incorrect answers (errors) per interaction type among participants.	147
C.23	Bar chart of the distribution of the number of incorrect answers (errors) per interaction type among participants.	147
C.24	Values of <i>control_timeline</i> events in <i>next</i> category.	148
C.25	Values of <i>control_timeline</i> events in <i>prev</i> category.	148
C.26	Values of <i>control_timeline</i> events in <i>change</i> category.	149
C.27	Values of <i>control_timeline</i> events in <i>play</i> category.	149
C.28	Values of <i>control_timeline</i> events in <i>pause</i> category.	150
C.29	Values of <i>control_timeline</i> events in <i>goto</i> category.	150

List of Tables

2.1	Initial search strings definition.	7
2.2	Search strings results per digital library.	9
2.3	Inclusion and exclusion criteria.	10
2.4	Extracted information from included records.	11
2.5	Research purpose description for the studies published in 2023.	12
2.6	Research purpose description for the studies published in 2022.	13
2.7	Research purpose description for the studies published in 2021.	13
2.8	Research purpose description for the studies published in 2020.	14
2.9	Research purpose description for the studies published in 2019.	14
2.10	Research purpose description for the studies published in 2018.	15
3.1	Techniques common to interaction and animation taxonomies [67].	23
6.1	Reference of custom events for Google Analytics tracking.	74
7.1	Quartiles of the time spent and the task error rate of the whole population.	79
7.2	Distribution of age and knowledge in information visualization and sports among participants.	80
7.3	Task Accuracy per Question.	81
7.4	Distribution of time spent per number of incorrect answers among participants.	84
7.5	Spearman correlation results for information visualization and sports knowledge.	87
7.6	Two-proportion Z-test results comparing task accuracy between Mouse Movement and Touch Interaction groups per question.	89
7.7	Distribution of the number of incorrect answers per type of interaction among participants.	90
7.8	Distribution of ratings per SUS question among participants.	91

Listings

5.1	Example of the JSON object retrieved from the endpoint.	50
6.1	Google Analytics Installation Tag.	72
6.2	Hotjar Tracking Code.	74

Abbreviations and Symbols

InfoViz	Information Visualization
UI	User Interface
WoS	Web of Science
MSV	Massive Sequence View
PCP	Parallel Coordinates Plot
SOM	Self-Organizing Maps
t-SNE	t-distributed Stochastic Neighbor Embedding
H-graphs	Helical Graphs
NL	Node-Link
M	Matrix
SI	Superimposition
JP	Juxtaposition
ANC	Animation with Playback Control
AN	Auto Animation
API	Application Programming Interface
JSON	JavaScript Object Notation
PHP	Hypertext Preprocessor
HTML	HyperText Markup Language
CSS	Cascading Style Sheets
JS	JavaScript
SVG	Scalable Vector Graphics
ID	Identifier
SUS	System Usability Scale

Chapter 1

Introduction

The exponential growth of information emphasizes the necessity to present it in a user-friendly format, ensuring clarity and comprehension. Many companies share this concern, prompting the question of how to present information to their users in a more appealing and effective way. This task is a daily challenge facing ZOS, Lda., a company that manages a substantial volume of intricate and historical data that must be presented in an attractive and understandable manner. Numerous innovative techniques have been developed globally to convey data through diverse visual formats, making selecting the most fitting approach based on the context and specific data characteristics imperative.

1.1 Context and Motivation

ZOS, Lda. (ZOS) has been informing sports fans and professionals since 2003 [68]. The company has an extensive data collection spanning various sports over the years. This repository is constantly updated and enriched, ensuring its accuracy and relevance. In the realm of soccer, ZOS is the home of the world's largest soccer database [68], featuring comprehensive information on the history of soccer, national teams, major league clubs, and players, as well as those in minor or regional leagues. This wealth of detailed information is maintained through ongoing collaboration with numerous institutions and the contributions of enthusiastic supporters, establishing the company as a leading and trusted authority in the industry.

Considering all available information, it is crucial to present and communicate the multidimensional and multivariate data efficiently. The company maintains a dedicated focus on employing innovative strategies to properly convey information to its users. One effective visual tool used in the sports domain is infographics, which presents complex data in an attractive and easily understandable manner, allowing viewers to identify and analyze the content shown. ZOS aims to present its substantial volumes of information, mainly historical and highly interconnected data, in a more comprehensive and dynamic format capable of attracting and engaging its users.

1.2 Problem Identification

The vast and complex realm of sports data presents a unique challenge in making informative, clear, and comprehensive visualizations. The different user profiles and visual literacy levels in the online world further exacerbate this challenge. The complexity lies in sports data's intricate and dynamic nature and its usage in several media outlets. This work focuses on visualizing the temporal evolution of players in national teams and clubs. This problem involves the following dimensions:

- The players may or may not belong to a national team at a specific time. Therefore, one must consider both the players who have affiliations with both national teams and clubs simultaneously, as well as those who exclusively participate at the club level;
- The relationships between players and clubs can change over time, whether through transfers, loans, or retirements.

Capturing and presenting inherently dynamic data as a player, characterized by complex time-evolving relationships, is still not trivial. The challenge is not only to accurately represent this complexity but also to maintain visual consistency throughout the storytelling process [21], providing valuable insights to analysts, coaches, and sports enthusiasts about player transitions, career trajectories, and team dynamics.

Highly connected data refers to complex networks with numerous interdependencies [6] while time-evolving data pertains to data that changes over time [28]. In the context of highly connected and time-evolving information visualization is essential to ensure that the visual representation remains comprehensible and coherent without overwhelming the user. Balancing the need for detail with the imperative of clarity poses a significant challenge that forms the core of the problem addressed in this thesis.

1.3 Hypothesis

Considering the problem previously identified, this work formulates the following hypothesis:

Graph-based visualizations for time-evolving data are well suited for presenting highly connected entities with complex relationships while preserving the story within the visualization.

Graphs are renowned for their capacity to capture complex relationships and temporal dependencies within data [28]. The historical information and dynamic changes over time can be represented using animations [11]. This results in a visualization method tailored to the intricate and dynamic nature of complex, hyper-connected data enriched with an additional temporal dimension.

1.4 Research Questions

Given the hypothesis outlined in the preceding section, it is imperative to identify and understand visualization techniques and tools that align with the specific characteristics of the data under consideration (highly interconnected data with temporal evolution). It involves exploring graph-based approaches, encompassing their capabilities, limitations, and applications. These objectives can be translated into the following set of research questions:

- Q1:** Which visualization techniques are commonly used to present highly connected data?
- Q2:** Which visualization techniques are commonly used to present temporal data?
- Q3:** Which visualization techniques are commonly used to present multidimensional and multi-variate data?
- Q4:** What are the limitations and applications of graph-based visualizations for time-evolving data?

1.5 Methodology

To prove the hypothesis outlined in Section 1.3, this dissertation follows the structured methodology described in the next paragraphs.

Firstly, a **Systematic Literature Review** is conducted using the PRISMA methodology [49]. This comprehensive review aims to delve into existing visualization techniques, highlighting their advantages and disadvantages, particularly regarding the representation of interconnected and time-evolving data, prioritizing graph-based approaches. Additionally, the review extends its scope to assess and recognize best practices in data visualization, with a specific focus on usability and animation strategies that have proven successful in practical applications.

Secondly, a **State-of-the-Art** is written. This step involves the construction of a cohesive narrative that encapsulates the outcomes of the Systematic Literature Review. This narrative will reflect the field's current knowledge, providing a panoramic view of existing visualization techniques. The clarity gained in this phase by extracting meaningful insights from the selected literature ensures a well-informed selection of visualization methods aligned with the defined problem.

Thirdly, a precise **solution definition** to address the identified problem is articulated based on the acquired insights from the literature review and state-of-the-art analysis.

Subsequently, a **prototype** mirroring the proposed solution is **implemented**. This phase includes writing comprehensive documentation on the intricacies of the development process, detailing the specific technologies chosen and their application in creating the prototype.

Finally, the **functional prototype** is **validated** through the specification and application of an experiment procedure, whose primary goal is to assess the proposed solution's usability, effectiveness, and efficiency in visually conveying information. This procedure encompasses defining the parameters for the experiment, conducting the test, and collecting the results, followed by the analysis and conclusions from the experiment's results.

1.6 Document Structure

This dissertation comprises seven chapters alongside the introduction. This section briefly describes the primary purpose of each chapter.

Chapter 2 details the methodology employed for the literature review on Information Visualization of interconnected and time-evolving data.

Chapter 3 presents the state-of-the-art, providing a comprehensive overview of existing developments based on the pertinent literature retrieved during the literature review phase.

Chapter 4 elaborates on the problem previously introduced, including the clarification of the research questions, solution requirements, and problem scope.

Chapter 5 specifies the proposed solution, delving into the technological and design aspects of the prototype implementation.

Chapter 6 explains the experimental protocol defined to validate the developed prototype, outlining the test methodology and the assessment methods employed.

Chapter 7 presents the results obtained from the experiments conducted, followed by their thorough analysis and discussion.

Chapter 8 summarizes the main findings and outlines future work on the addressed topic.

Chapter 2

Literature Review

A systematic literature review is undertaken using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology [49, 27]. This methodology offers a rigorous, structured, and comprehensive approach to reviewing existing literature. By meticulously synthesizing previous research, it gives a deeper understanding of the current knowledge in the field from multiple perspectives.

2.1 Research Domain

Over the years, numerous approaches have evolved to address emerging challenges in data visualization. These challenges stem from the exponential growth of information, its increasing complexity, and growing user demands. The objective has been to define or refine visual metaphors that can simplify and adapt to these new requirements. In the context of visualizing highly interconnected and time-evolving data, the current study focuses on the research questions listed in Section 1.4.

The goal is to identify visualization techniques and tools that align with the specific characteristics of the data under consideration: highly connected and time-evolving data. Particularly graph-based visualization techniques: capabilities, limitations, and applications. Fig. 2.1 visually encompasses the complete set of steps performed according to the PRISMA methodology, which will be presented in detail in the following sections.

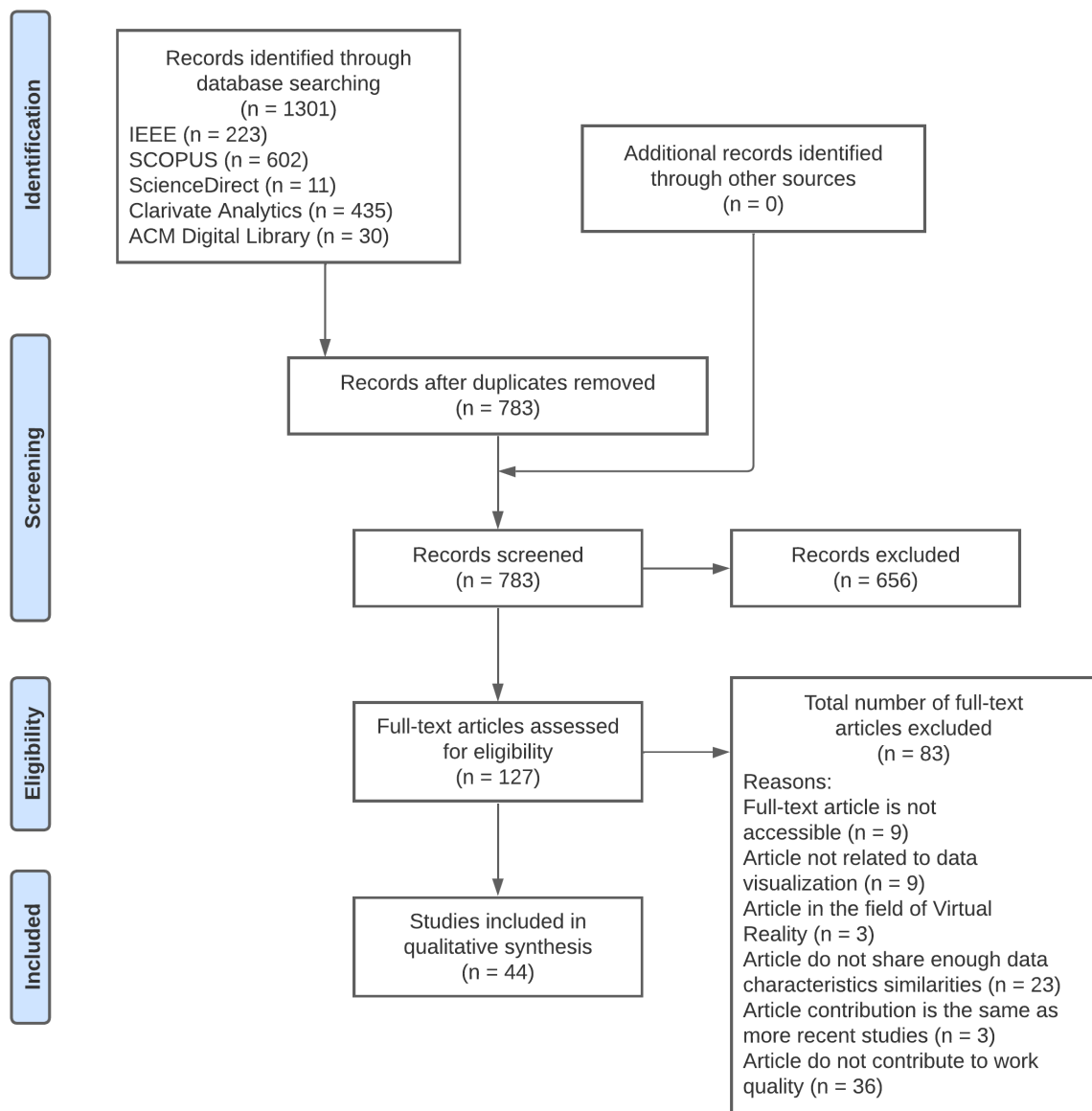


Figure 2.1: PRISMA flowchart.

2.2 Identification Phase

An extensive search is conducted in the digital libraries of scientific literature to gather knowledge on the domain identified. Journal articles, workshop articles, conference papers, and books are included in the search process.

Firstly, five digital libraries are selected as platforms for the search process. These libraries are IEEE Xplore, SCOPUS, ScienceDirect, Clarivate Web of Science, and ACM Digital Library. Next, the keywords for the search process are defined and refined. As the use of general keywords results in a large number of research papers, specific keywords are formulated to find the most relevant studies. The initial search strings are shown in Table 2.1.

Table 2.1: Initial search strings definition.

QS1	("high connected" OR hyperconnected) AND (data OR information) AND visuali?ation
QS2	temporal AND (data OR information) AND visuali?ation
QS3	(multidimensional AND multivariate) AND (data OR information) AND visual?ation
QS4	(graph OR network) AND ((time NEAR/2 evolving NEAR/2 data) OR (temporal NEAR/2 data)) AND visuali?ation

The search strings consider the different spellings of the word “visualization” in British and American versions of English by using the wildcard operator. Furthermore, as the terms “data” or “information” and “visualization” by themselves appear in various contexts and domains, the motivation of the first, second, and third search strings is to narrow down the search results by individually specifying the data characteristics inherent to the review (not excluding the possibility of obtaining articles that simultaneously present data with both traits). Meanwhile, the motivation of the fourth search string is to direct the search to results more related to the specific hypothesis described in Section 1.3.

2.2.1 Refinement Process

The refinement process of the search strings shown in Table 2.1 was performed using only the IEEE Xplore as reference. Inspecting the search results of each search string helps to understand terms that may be too broad or missing and adjust accordingly.

2.2.1.1 QS1 Refinement Process

The search string (*“high connected” OR hyperconnected*) AND (*data OR information*) AND *visuali?ation* results in 1 poorly record. The terms used for referring to the data characterized by relationships were shown to be not the most appropriate ones. In the next version of the query, the terms “highly connected” and “interconnected” were used instead, resulting in the search string (*“highly connected” OR interconnected*) AND (*data OR information*) AND *visuali?ation*. This refinement increases the results to 49, which is more suited to the reviewed topic.

2.2.1.2 QS2 Refinement Process

The search string *temporal AND (data OR information) AND visuali?ation* retrieves close to 5300 results without filters. However, when filtering the type, year, and topic of the publication according to the criteria defined in Section 2.2.2, the number of results decreases to 699. Still, this is a very large number, probably derived from how generic the term “temporal” is. By analyzing the results titles, it is trying to achieve a more relevant dataset by specifying the term to refer to data characterized by change along time, resulting in the search string *time-evolving AND (data OR information) AND visuali?ation*. This modification reduces the results to 9. After an analysis

of this list, some of the results seem relevant for the search, so it does not make to replace the term but extend the query to include results with synonymous terms. Combining with the term “dynamic” increases the results to 845, while integrating with the term “evolving” increases the results to 224. However, both increases do not directly imply higher relevance for the review by quickly analyzing the titles. In the final version, the term “time-evolving” was combined with the term “time-varying”, resulting in the query *(time-evolving OR time-varying) AND (data OR information) AND visuali?ation* that retrieves 96 results.

2.2.1.3 QS3 Refinement Process

The search string *(multidimensional AND multivariate) AND (data OR information) AND visuali?ation* retrieves 7 results. In this refinement, it is considered if it would be beneficial to use the operator OR instead of AND between the terms used to characterize the data. Using the OR operator increases the number of results to 453, which is not beneficial. This broadens the results too much, yielding mostly irrelevant records. Furthermore, the review is primarily interested in data that has both characteristics so opting for the AND operator is more advantageous.

2.2.1.4 QS4 Refinement Process

The search string *(graph OR network) AND ((time NEAR/2 evolving NEAR/2 data) OR (temporal NEAR/2 data)) AND visuali?ation* retrieves 63 results. After inspecting the results’ titles, the conclusion is that this query does not need extra refinement. However, with this analysis, it is identified that the term “dynamic graph visualization” is commonly used to refer to representations of network evolutions over time. So, to capture these techniques, a new query is considered - *dynamic NEAR/1 graph NEAR/1 visuali?ation* - instead of modifying this one, this allows us to inspect a more significant range of possibilities (even though some overlaps exist between the two, they are not seen as mutually exclusive).

2.2.1.5 Refinement Process Result

After the refinement of each search string (QS1 to QS4), the final versions used in the PRISMA methodology are the following:

QS1. (“highly connected” OR interconnected) AND (data OR information) AND visuali?ation

QS2. (time-evolving OR time-varying) AND (data OR information) AND visuali?ation

QS3. (multidimensional AND multivariate) AND (data OR information) AND visuali?ation

QS4. (graph OR network) AND ((time NEAR/2 evolving NEAR/2 data) OR (temporal NEAR/2 data)) AND visuali?ation

QS5. dynamic NEAR/1 graph NEAR/1 visuali?ation

From this final set of search queries, considering all the digital libraries, a total of 1301 papers were retrieved. The results were exported and imported into the references management system being used (EndNote). The number of results for each library is schematized in Table 2.2.

Table 2.2: Search strings results per digital library.

	QS1	QS2	QS3	QS4	QS5	Total
IEEE	49	96	7	63	8	223
Scopus	164	232	39	133	34	602
ScienceDirect	1	7	0	2	1	11
Clarivate WoS	121	184	42	70	18	435
ACM Digital Library	18	10	2	no proximity operator		30

2.2.2 Search Criteria

Data and visualization concepts have been widely used in various contexts for a long time. Therefore, the search results are filtered to the ones published in the last five years (from 2018 or later) in the area of “Human-centered computing”, specifically “Information/Data Visualization” and “Graph Theory”. These search criteria are adapted according to the capabilities of each search engine as follows:

- IEEE: 2018+ journal articles, conference papers, and books on the topic of “data visualization” or “graph theory”;
- Scopus: 2018+ English conference papers, articles, and books in the subject area of “Computer Science” or “Engineering” (search within Article title, Abstract, Keywords);
- ScienceDirect: 2018+ articles in the subject area of “Engineering” or “Computer Science”;
- Clarivate Web of Science: 2018+ English article or proceeding/conference paper in the research areas of “Computer Science” or “Engineering” (search within Topic - title, abstract, and keywords);
- ACM Digital Library: 2018+ publications (search within Title, Abstract).

2.3 Screening Phase

The screening phase in the PRISMA methodology encompasses defining inclusion and exclusion criteria, eliminating duplicate results, and meticulously examining each unique record. The goal is to filter out records that do not align with the specified criteria, facilitating an initial selection of potentially relevant literature on the topic.

2.3.1 Inclusion and Exclusion Criteria

A total of 1301 papers are found in the identification phase. Hence, the definition of inclusion criteria is crucial to finding the papers more relevant to this research. The collected papers have undergone twelve phases of selection through the use of inclusion and exclusion criteria, as outlined in Table 2.3.

Table 2.3: Inclusion and exclusion criteria.

Phases	Inclusion/Exclusion criteria
1	Search engine results after executing a search query
2	Date of the publication since 2018 (inclusive)
3	Full-text access
4	Published in the English language
5	Published in the area of “Information Visualization” or “Graph-based Visualization”
6	Not in the field of Virtual Reality (VR), Augmented Reality (AR), or Mixed Reality
7	Visualization of interconnected, temporal dependent, multidimensional, and/or multivariate information (share similar characteristics with the problem: time-evolving relationships)
8	Not duplicated
9	Not retracted publication
10	Not work-in-progress publication
11	Article, Conference Paper or Book
12	Not summaries of workshops, seminars, lectures, or tutorials

The defined criteria ensure that the literature assessed as relevant to the research aligns with the goal of the research: existing data visualization techniques and tools for interconnected data with temporal evolution, particularly graph-based approaches. Therefore, there is a need for more specific criteria based on similarities between data characteristics and environment (web-based only).

2.3.2 Remove Duplicates

The process of finding the duplicates is performed using the EndNote abilities and adjusting the duplicate criteria in 3 phases: match of authors, year, title, and reference type; exact match of DOI; and partial match of the title. This considers the differences between the title and/or author notation in the several search engines’ export functionality to accurately get the unique papers. The process results in a dataset of 783 unique papers and 518 duplicates.

2.3.3 Screened Records

The 783 unique records were individually screened by inspecting their title and abstract to exclude those that do not match some criteria defined in Table 2.3, like topic, data characteristics, format, etc. This process results in the exclusion of 656 records, remaining 127 papers.

2.4 Eligibility Phase

In the subsequent phase of the PRISMA methodology, the focus turns to the eligibility phase for the remaining 127 papers. This pivotal step involves a meticulous evaluation, applying predefined inclusion and exclusion criteria to assess each paper's relevance to the research objectives. The process ensures a systematic and transparent approach to selecting papers that align with the study's goals, maintaining the methodological rigor outlined by PRISMA.

2.4.1 Full-text Access

The first stage of this phase consists of the search for the PDFs related to the included records with the auxiliary EndNote capability of "full-text search". This functionality located 26 PDFs and 70 URLs and could not find 31 PDFs. The remaining PDFs were located manually. Already in this stage, it was possible to find 8 records without full-text access and 1 in which the year or paper type was mislabeled and did not match the required inclusion criteria presented in Table 2.3.

2.4.2 Data Collection

At this stage, the relevant information to extract from the selected papers was defined to facilitate subsequent data analysis. Table 2.4 details the specific information to be extracted.

Table 2.4: Extracted information from included records.

Category	Data
Basic Information	Title Author(s)
Publication Information	Source Year Type
Research information	Application Domain Research Purpose Used visualization techniques and tools Relevance and Contribution to the proposed solution

2.4.3 Selected Records

The following stage is based on the complete analysis of the full text of the records, descending order based on the year, summarizing its main goals and contributions, and deciding its relevance to the work being performed.

Although there were 1301 papers retrieved from the identification phase, only 44 papers are considered relevant to this study after applying the exclusion criteria. Tables 2.5, 2.6, 2.7, 2.8, 2.9, and 2.10 succinctly describe the research purpose of the selected papers.

Table 2.5: Research purpose description for the studies published in 2023.

Reference	Research Purpose	Approach/domain
[6]	To introduce Collaboration Spotting X, an open-source web-based visual analytics tool for retrieving, modeling, and exploring highly connected data as networks.	Network visualization and exploration tool
[28]	To present a study that assesses the accuracy and preferences of participants on structural (NL or M) and temporal (SI, JP, ANC, or AN) encoding combinations.	Dynamic network visualization
[39]	To introduce <i>LargeNetVis</i> , a web-based tool for exploring large temporal (or time-evolving) networks.	Large temporal network visualization tool
[44]	To propose an algorithm that relies on a temporal RDF graph to calculate disease propagation.	Dynamic graph visualization
[25]	To propose an edge sampling algorithm (E-MCBN) that introduces MCBN sampling into dynamic graph sampling to reduce visual clutter in Massive Sequence View (MSV).	Dynamic graph visualization
[58]	To investigate the effects of spatial visualization (SV) and domain differences (educational background, professional profile, motivation) on visualization task performance.	Case study on visualization performance
[19]	To explore the use of helical graphs for visualizing the NDVI temporal variation of tropical and boreal forest vegetation trend.	Ecological data visualization using helical graphs
[23]	To propose a novel difference-based approach, <i>Diff-Seer</i> , for dynamic weighted graph visualization, which employs color differences to explicitly visualize graph structure details differences between adjacent time-slices.	Dynamic weighted graph visualization
[65]	To develop a visual analytics framework, <i>Gaze-GraphVis</i> , to investigate how individuals explore graph visualization, namely gaze behaviors, during path tracing tasks and to provide layout guidelines to generate more expressive and readable graph visualizations.	Eye behavior analytics in static graph visualization
[67]	To introduce <i>Animated Vega-Lite</i> (Vega-Lite extension), a grammar that provides a unified abstraction for static, interaction, and animation modalities of a visualization.	Interactive and animated data visualization toolkits

Table 2.6: Research purpose description for the studies published in 2022.

Reference	Research Purpose	Approach/domain
[14]	To introduce CroP, an interactive and customizable visualization tool that uses a coordinated multiple views framework to represent and explore relational, temporal, and multivariate datasets at different levels of detail.	Visualization tool for high-dimensional data
[36]	To introduce a web-based visual tool to the management and visualization of environmental data with spatial and temporal components.	Visual tool for environmental data
[37]	To design and analyze team performance and cooperation from a network of player passes.	Network visualization in soccer teams
[48]	To discuss clustering in graphs opportunities in visualization, graph mining, classification, and compression (WIP approach).	Graph-based spatial-temporal cluster analysis
[63]	To introduce <i>Angular-based Edge Bundled Parallel Coordinates Plot</i> (APCP), a visualization simplification of PCP to present large ensemble simulation datasets, focusing on meteorological data.	High-dimensional data visualization
[66]	To describe a workflow to enhance the pattern recognition in temporal evolving set data.	Pattern visualization in temporal set data

Table 2.7: Research purpose description for the studies published in 2021.

Reference	Research Purpose	Approach/domain
[8]	To develop a web-based tool for visualizing and exploring dynamic graphs through interactive visualization techniques.	Dynamic graph visualization
[9]	“To propose a visual analytics tool for explore a graph sequence generated from algorithm dynamics.”	Algorithms for dynamic graphs
[11]	To introduce <i>Multiscale Snapshots</i> , a visual analytics tool to present structural and temporal changes in dynamic graphs.	Dynamic graph visualization
[12]	To propose a systematic approach to design and implement online force-directed (OFD) algorithms for visualization of dynamic graphs.	Algorithms for dynamic graphs
[29]	To present two models for the representation of connected components of a dynamic graph in the presence of edge changes over time.	Dynamic graph visualization

Table 2.8: Research purpose description for the studies published in 2020.

Reference	Research Purpose	Approach/domain
[1]	To introduce time-aligned edge plots (TEPs), a representation that provides a “time-scalable overview of dynamic graphs that preserves the users’ mental map and does not suffer from visual clutter”.	Dynamic graphs Visualization
[21]	To propose a framework for story synthesis that addresses the challenge of effectively communicating complex visual analytics findings to non-expert audiences.	Visual Storytelling
[22]	To introduce <i>TimeSum</i> , an approach to present the complex topologies of dynamic networks through abstractions, particularly graph differences and graph summarization.	Dynamic Network Visualization
[51]	To explore the design and features of a visual tool to present interconnected data using circular graphs.	Interconnected Data Visualization
[60]	To present an approach to visualize and analyze complex interconnected objects using static graphs.	Multidimensional Data Visualization

Table 2.9: Research purpose description for the studies published in 2019.

Reference	Research Purpose	Approach/domain
[2]	To introduce <i>TimeCluster</i> , a visual system that combines sliding window approach with dimension reduction techniques to enhance pattern identification, labeling, and outlier detection of time-series data.	Time-series data visualization
[16]	To introduce <i>TEFNET</i> , a system that used graph visualization techniques for fiscal data analysis over multiple fiscal periods, adopting a timeline representation to map the temporal dimension to the space dimension.	Temporal Network Visualization tool for fiscal data
[38]	To analyze and compare visualization techniques for time-series data.	Time-series visualization taxonomy
[54]	To explore the importance of data visualization and the effectiveness of several chart types in representing different types of data.	Data visualization
[26]	To introduce a unified optimization framework for generating structure-aware fisheye views that enable efficient exploration of large graphs while preserving the graph’s structural integrity and improving readability.	Large graphs visualization and exploration
[62]	To propose a method for representing the proximity of communities along with their temporal changes using a deep neural network on top of the graph convolution and dimension reduction techniques in the visualization phase.	Dynamic graphs visualization
[24]	To introduce <i>ForVizor</i> , an interactive visual analytic system to present formation changes in-game of a soccer team over time through visualizations.	Visualization Tool to soccer formations

Table 2.10: Research purpose description for the studies published in 2018.

Reference	Research Purpose	Approach/domain
[3]	To propose a visualization method for efficiently observing temporal and periodic features of events, integrating glyphs, <i>ChronoView</i> , and pseudo multidimensional scaling (MDS) to represent similarities among event groups.	Visualization of Temporal Data, particularly events.
[4]	To explore and compare three multivariate visualization approaches - Self-organizing maps, Parallel coordinate plots, and t-distributed Stochastic neighbor embedding - when presenting vibration signal data.	Multivariate data visualization
[5]	To describe the design and development of <i>ecoxight</i> , a Web-based ecosystem visualization tool (including visual encodings and interactions).	Visualization of Ecosystems using dynamic networks
[7]	To present a dynamic graph visualization technique that, by computing the density regions of node-link graph layouts, provides an abstraction of the global dynamic behavior of relational data over time (time-varying graph data).	Dynamic graph visualization
[20]	To present a web-based interactive visual tool to explore data from a movie database using two views: movie cosmos and career lines.	Hierarchical data visualization
[17]	To explore and evaluate existing multidimensional visualization tools in supporting decision-making tasks.	Multidimensional visualization tools
[18]	To introduce the <i>Dynamic Influence Network</i> (DIN), an approach for presenting and analyzing the dynamics of rules influences in rule-based complex systems, particularly protein-protein interaction networks, using dynamic networks.	Visual approach for rule-based influence
[46]	To propose a “transformation tool which can be used for transforming Twitter raw data format into a web-based graph visualization”.	Graph Visual tool of Twitter data
[52]	To propose a visualization method for presenting and comparing strong and weak points of individual players and teams in soccer.	Visual tool for soccer data
[56]	To explore how aesthetic and mental map criteria of a graph visualization can influence each other.	Graph Layouts analysis
[64]	To propose an algorithm for the spatial rearrangement of graphs when visualizing a dynamic network evolution process.	Algorithms for dynamic graphs

Alongside these summary tables, the reasoning behind the exclusion of the irrelevant papers is annotated and visually represented in Fig. 2.1. Moreover, citations of the relevant papers are highlighted to identify possible statements helpful to the construction of the State-of-the-Art.

2.5 Results and Discussion

Establishing a coherent connection between the selected records, outlined in the preceding section, and the study's research questions in Section 1.4 provides a solid foundation for understanding their relevance.

The dataset reveals a significant trend focused on graph-based visualization techniques. Specifically, 18 out of the 44 records delve into this area [44, 25, 23, 65, 37, 8, 9, 11, 12, 29, 22, 51, 26, 5, 7, 18, 46, 64], forming a robust resource for addressing Research Questions 1, 3, and 4. Simultaneously, the examination of temporal data visualization, evident in 11 records of the 44 [39, 19, 36, 66, 2, 16, 38, 62, 24, 3, 52], directly aligns with the exploration of Research Question 2. Interestingly, the intersection of these two areas highlights a predominant emphasis on dynamic graph visualization, aligning with the hypothesis presented in Section 1.3.

Pertinent insights for Research Question 3 emerge from the examination of multidimensional and multivariate data visualization within 9 of the 44 records [6, 14, 63, 51, 60, 54, 4, 20, 17]. Furthermore, the exploration of limitations and applications of graph-based visualizations for time-evolving data, discussed in 4 records [28, 48, 1, 56], closely aligns with Research Question 4. Additionally, the nuanced consideration of incorporating animations in an online scenario is addressed in two records [67, 21].

Notably, the results include two specific instances of information visualization in the realm of soccer-related challenges [24, 52].

Overall, the selected records not only provide insights into various visualization techniques and tools for presenting complex, highly connected, and evolving data but also seamlessly align with the research questions outlined in the study (refer to Section 1.4). Moreover, the importance and applicability of the information visualization domain in diverse contexts are underscored.

It is crucial to emphasize that, beyond applying exclusion criteria, a discerning approach was employed during the eligibility phase, leveraging critical thinking and quality criteria. The aim was to ensure that the selected records resonated with the context, problem, and solution and significantly contributed to future applications. Key considerations included evaluating whether the study enhanced the overall quality of ongoing work, if it explored methods and approaches related to the visualization of data, and whether it introduced novel and pertinent information. Furthermore, if an older paper closely resembled a more recent one already selected, the newer one was chosen.

This method aimed to capture recent progress and viewpoints, creating a current representation of data visualization's status in the specified fields of interest.

2.6 Beyond PRISMA

The Systematic Literature Review process using the PRISMA methodology was rigorously executed, following all defined steps and criteria. However, in addition to the records selected with PRISMA, there arose the need to include an additional document in the context of this review. Specifically, an older article that does not meet the established publication year criteria but serves as a foundation for areas of interest related to the theme should be considered. The article by Shneiderman [55] is such a case, and its inclusion is crucial given its value in the field of visualization, defining the Mantra “Overview first, zoom and filter, then details-on-demand.” This article provides insights into the presentation of data on the screen to be most effective for users.

2.7 Summary

This chapter provides a comprehensive overview of the literature review undertaken, laying the foundation for the state-of-the-art depicted in the next chapter.

The Systematic Literature Review delves into graph-based visualizations for the effective communication of complex and dynamic data, more specifically on visually conveying intricate relationships within the data and their change over time. The review, conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, systematically selects and scrutinizes pertinent studies about hiper-connected time-dependent data and graph-based visualizations. Key findings highlight the challenges in representing highly connected and time-evolving data and the importance of maintaining visual consistency in information visualization. The review emphasizes the significance of visual storytelling in conveying complex data, particularly in domains like sports, and underscores the growing importance of infographics for presenting dynamic data. This review enriches the understanding of the prevailing knowledge of data visualization techniques and tools that align with the characteristics of the data under consideration, particularly graph-based approaches.

Chapter 3

State of the Art

This chapter presents a comprehensive overview of the State of the Art in Information Visualization (Section 3.1) and Graph-based Visualization (Section 3.2), particularly dynamic graph visualization (Section 3.2.2). The narrative results from summarizing and organizing the findings of the Systematic Literature Review performed in Chapter 2.

3.1 Information Visualization

In the current digital era, enterprises have access to many digitized data from various sources such as press releases, social media, and online databases [5]. This surge in digital data accessibility has empowered decision-makers with unprecedented access, allowing them to have a large volume of information at their fingertips [5]. However, with the increased availability of data, new challenges arise due to the size and variety of the collected data, presenting difficulties in extracting meaningful insights and drawing reliable conclusions [48, 6].

Acknowledging the transformative role of data in shaping our lives, from political, social, and physical aspects, information visualization emerges as a crucial tool, spread around diverse platforms such as magazines, news outlets, scientific papers, blogs, and online spaces to enhance engagement with data in ways that text alone cannot achieve [58]. To address the challenges posed by the volume and variety of data, information visualization has been a longstanding solution applied across various domains, including big data scenarios, to document, organize, and facilitate exploration and analysis of large datasets [48, 14].

Within information visualization, it is not merely a matter of translating data properties into visual form. Complex cognitive activities are at play, influenced by the visible structure itself [58]. This underscores the nuanced role of information visualization, extending beyond a simple representation of data to actively shaping the cognitive processes involved in comprehending and extracting insights from the presented information [58]. As a result, the application of information visualization remains crucial in various domains, namely in the sports domain, offering a powerful means not only to organize extensive datasets but also to facilitate exploration, analysis, and the extraction of new information from the available data [14].

3.1.1 Sports Data Visualization

Sports data visualization has become integral for understanding and analyzing various athletic disciplines. The integration of visualization techniques into sports analysis extends to multiple sports such as table tennis, tennis, basketball, and soccer [24]. Furthermore, specific visualization tools are designed to enhance the intuitiveness of data representations and simplify pattern recognition complexities, find application across the sports domain, particularly in the soccer field [24, 52]. Tableau Software *Sports + Data Viz Gallery* [57] showcases a diverse collection of exemplary data visualizations across various sports.

Wu *et al.* [24] states that "Soccer is one of the most popular sports in the world with millions of participants and more than tens of thousands of professionals". The popularity of soccer has contributed to the development of visualization tools to promote the understanding of soccer data [24], such as tools for visualizing game results or pass networks; *CourtVision* and *SnapShot* (designed for soccer, basketball, and hockey that introduced heatmaps tailored to ball and puck shots); *Soccer Scoop* (designed for player performance comparison); *SoccerStories* [24, 52].

In the concrete field of soccer data, more recent developments have been made to improve the analysis of the data, namely in-game spatio-temporal evolution of team formation [24] and players and teams performances [52]. Wu *et al.* presents *ForVizor* [24] a system that focuses on the analysis of the performance of in-game dynamics, particularly formation changes over time, whose interface is presented in Fig. 3.1. Ryoo *et al.* [52] introduces a new visualization approach for the analysis of player performance in a pixel-based representation and comparison with other players and even team performance, as illustrated in Fig. 3.2.

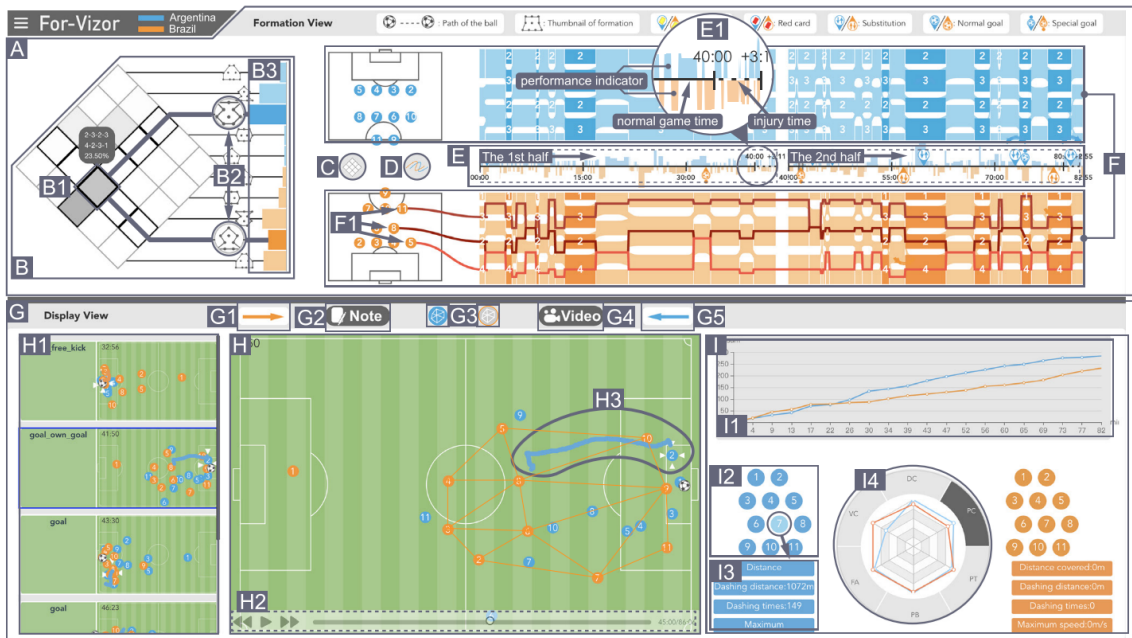


Figure 3.1: ForVizor: Interface of the system [24].

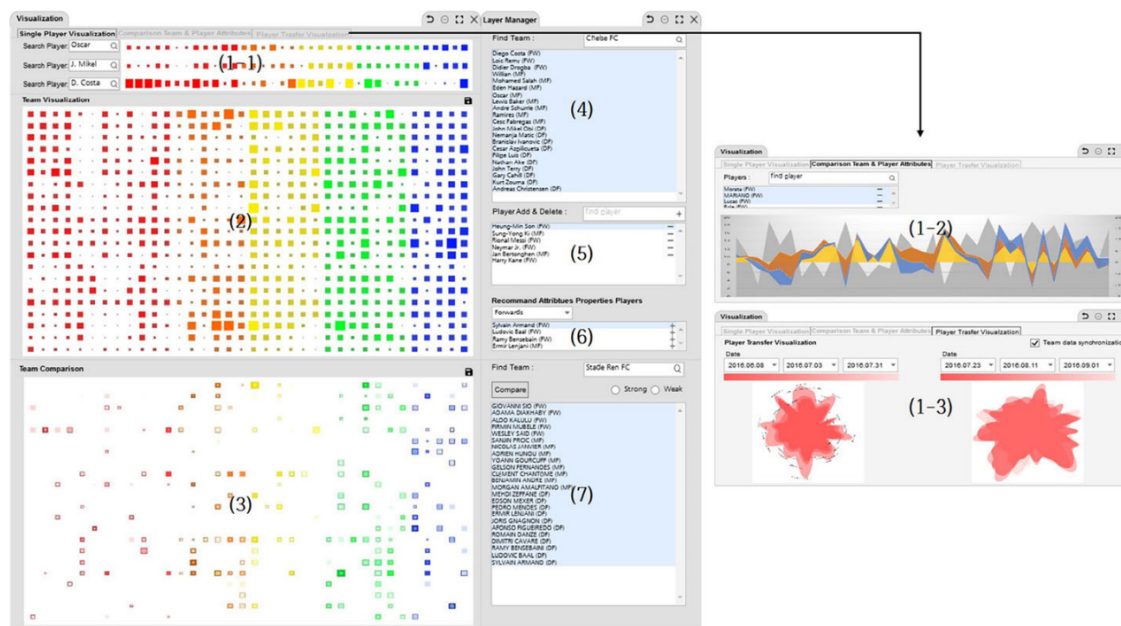


Figure 3.2: Visualization of the overview of the player attributes [52].

Furthermore, in soccer data, many examples specifically centered around data visualizations about the World Cup can be found [47, 53].

Our work is built on the foundation laid by Shneiderman [55], who identified seven tasks and seven types of visualizations at that time. The assumptions we hold regarding visualization, tasks, and interaction remain valid today, having been revisited by multiple authors. Consequently, the perspective identified in this state-of-the-art goes beyond Shneiderman’s work but encompasses the subsequent developments in the field.

A significant contribution from Shneiderman [55] is his Mantra on how to effectively present data to users on screens. This recognizes that all tasks require visualization, and the approach is heavily influenced by the nature of the data. The following sections will explore this in more detail.

3.1.2 Data-oriented Visualization Techniques

In the rapidly evolving landscape of data-driven decision-making, using data plays a pivotal role in informed solutions and choices [54]. Data visualization emerges as a cornerstone technique, facilitating a deeper understanding and analysis of extensive datasets. The effectiveness of data visualization lies in its ability to simplify data analysis and enhance readability, enabling rapid comprehension of substantial data volumes [54]. However, the success of data visualization relies on the appropriate choice of visualization techniques, a critical factor in conveying accurate insights about the data [54].

The decision-making process, as highlighted by Shneiderman [55], depends on the nature of the data. To address the unique characteristics of the problem data, it becomes imperative to delve deeper into four distinct types: high-dimensional data, multivariate data, time-evolving data, and

highly connected data. Multivariate data encompasses information with multiple variables or attributes [4], whereas high-dimensional data specifically denotes datasets with an exceptionally large number of variables or dimensions [63]. Additionally, time-evolving data implies information that changes over time, with each item linked to a specific time or interval [19]. Lastly, highly connected data refers to information where items are characterized by intricate relationships, forming complex interconnected networks [6].

For handling **high-dimensional data**, mainly meteorological ensemble simulation data [63], the *Parallel Coordinates Plot* (PCP) proves to be an effective visualization technique [63]. PCP is commonly employed for analyzing high-dimensional datasets, revealing outliers, identifying trends, and understanding variable interrelations. Despite its strengths, PCP encounters challenges, such as visual cluttering due to overplotting, significantly as the number of represented members increases. Various approaches have been explored to mitigate visual cluttering, recognizing its impact on the effectiveness of PCP in high-dimensional data analysis [63].

In the realm of **multivariate data** representation, three algorithms - *Self-Organizing Maps* (SOM), *Parallel Coordinate Plots* (PCP), and *t-distributed Stochastic Neighbor Embedding* (t-SNE) - are explored by Bartkowiak and Zimroz [4] (Fig. 3.3). These algorithms provide diverse methods for visually displaying multivariate data, catering to different analytical needs. Specifically, in their experience, the SOM and t-SNE algorithms demonstrated an ability to recognize and visualize the geometrical localization of data points. In contrast, PCP provided details about the groups' content.

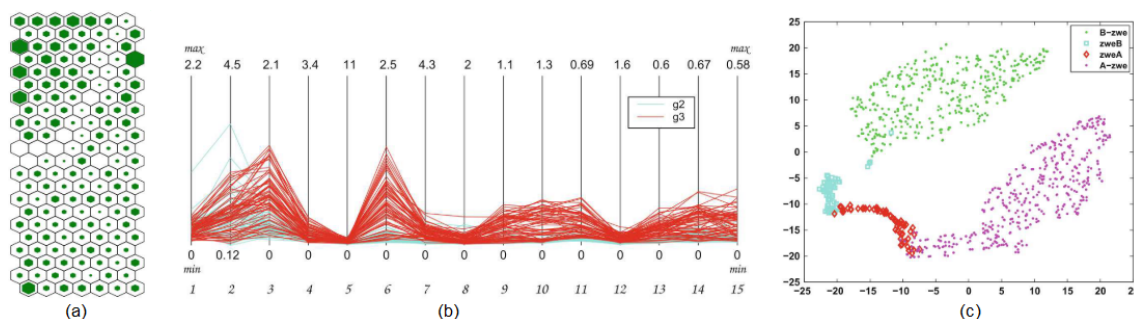


Figure 3.3: Complementary views over multivariate data using Kohonen's SOM (a), Parallel Coordinates (b), and t-SNE methods (c) [4].

The landscape of **time-evolving data** visualization techniques is diverse, encompassing various methods outlined by Komarek *et al.* [38], including line charts, area charts, radar charts, and pie charts. In temporal event variations, Anzai and Misue [3] emphasize the effectiveness of line charts, bar charts, and area charts, with spiral charts proving useful for capturing event periodicity. However, challenges arise in handling multiple event groups within a single chart. Introducing an innovative approach, Thouverai *et al.* [19] presents *Helical Graphs* (H-graphs) as a practical tool for visualizing generalized trends in multitemporal ecological studies (Fig. 3.4). H-graphs are designed to display ecosystem changes, depicting the change of a variable over time through moving averages and rates of change on the y-axis and x-axis, respectively. This novel visualization

method proves promising for revealing trends in time-evolving data, offering potential applications in ecological studies and beyond.

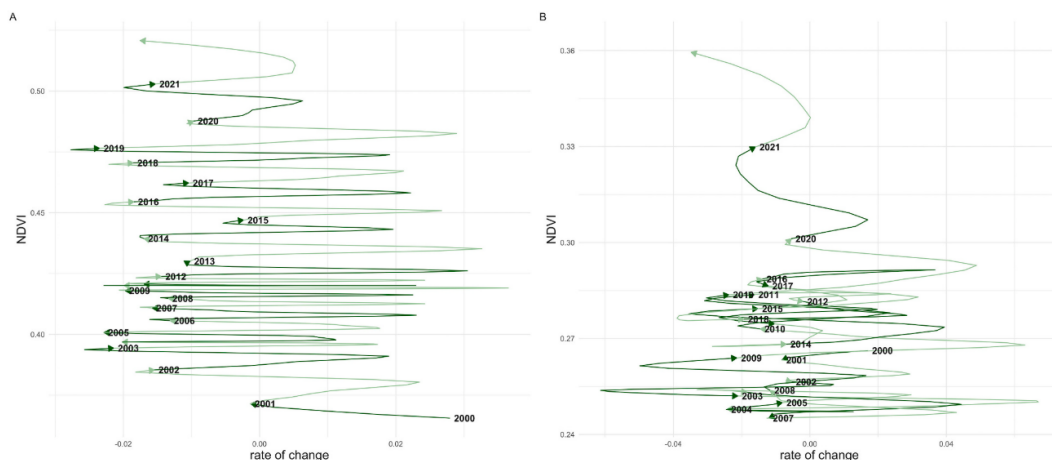


Figure 3.4: Helical graphs representing the variation from 2000 to 2021 of the tropical (A) and the boreal (B) forests [19].

The representation and visualization of data as networks have become widespread in fields like medicine, social sciences, and information retrieval [6]. Networks offer valuable insights into implicit connections within the **highly connected data**, but their large size poses challenges in visualization and interpretation. Bobic *et al.* [6] emphasizes the importance of filtering, aggregation tools, and accompanying visualizations to address these challenges and enhance the utility of network representations.

Considerations about scale, resolution, and the delicate balance between simplicity and detail in visualizations are highlighted in [19], underscoring the need for thoughtful selection of visualization techniques based on the specific characteristics of the data. In this context, Dimara *et al.* worked to develop guidelines to aid in deciding the most suitable system and visualization technique for a specific dataset [17].

3.1.3 Interaction and Animation in Information Visualization

Interaction and animation are two relevant techniques for performing tasks and effectively exploring the data, producing higher knowledge, and increasing engagement over a visualization [67].

Zong and Pollock [67] distinguish interaction from animation: interactions, driven by user input, transform data (e.g., filtering) and update visual properties (e.g., re-coloring marks), while animations perform the same changes in response to a timer. From their perspective, *"interactive and animated visualization techniques occupy a spectrum of dynamic, event-driven behaviors"* and several techniques are shared by interaction and animation taxonomies as described in Table 3.1. Furthermore, they highlight some applications of these techniques in the field of data visualization, such as the animation of static migration charts, where specific species can be highlighted with hover interactions, and the animation of scatter plots throughout the year, offering a more dynamic representation compared to multiple scatter plots for each year.

Table 3.1: Techniques common to interaction and animation taxonomies [67].

Example Technique	Interaction Intent	Animation Type
Conditional Encoding	Select	-
Panning	Explore	View transformation
Zooming	Abstract/Elaborate	View transformation
Axis re-scaling	Reconfigure	Substrate transformation
Axis sorting	Reconfigure	Ordering
Filtering	Filter	Filtering
Enter/exit	Explore	Timestep
Multi-view	Connect	-
Changing encoding	Encode	Visualization change, Data schema change

The significance of working in an interactive environment is emphasized in the context of data visualization [14]. Here, users are empowered to control the information on the screen, switch between visualization models, navigate different levels of detail, and filter data points to highlight significant ones. This interactive approach proves particularly valuable for the representation and analysis of temporal and relational data, as highlighted by Cruz et al. [14] with the development of the visual tool *CroP* (Fig. 3.5). Moreover, the effectiveness of interactive visualization in data analysis is underscored by its ability to represent high-dimensional datasets comprehensively and offer an environment for exploration, discovery, and knowledge extraction [14].

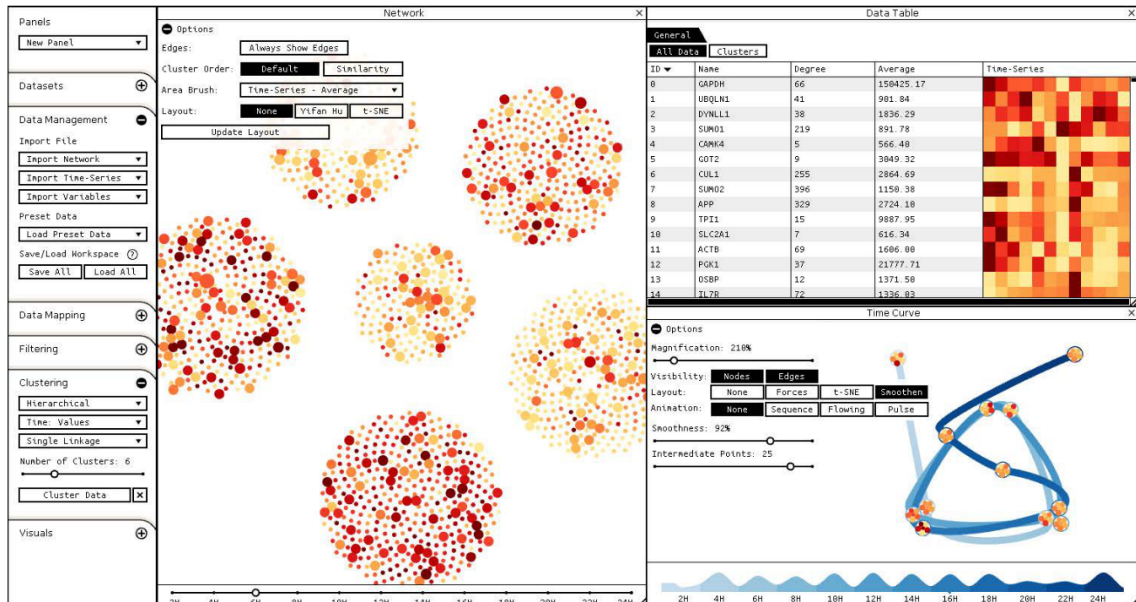


Figure 3.5: CroP's interactive user interface of a temporal dataset: options sidebar and three different panels [14].

However, the value of animation in visualization has been a subject of debate. While some studies have identified advantages, such as animating chart transitions or conveying uncertainty in data values, concerns have been raised about animations' potential complexity and speed, impacting accurate perception. Notably, Tversky *et al.* questioned the efficacy of animated graphics in a seminal 2002 paper, with subsequent studies offering optimistic and cautious perspectives [67].

Considering the support of interaction and animation in existing visualization toolkits, Zong and Pollock [67] discuss two classes of toolkits. The first class supports either interaction or animation but not both, while the second class supports both modalities via largely distinct abstractions. In the first class, the authors highlight systems such as *Vega* and *VegaLite* (Fig. 3.6) - "which offer interaction primitives in the form of signals and selections but do not provide abstractions for animation" - and *ganimate*, *Data Animator*, *Canis/CAST*, and *Gemini/Gemini2* systems - "which express animation in terms of transitions between discrete visualization states known as keyframes but do not provide treatment for interaction". In the second class, the authors include toolkits such as *D3* and *Plotly*, which support both modalities with transitions or frames for animation and event handlers or a typology of techniques for interaction.



Figure 3.6: Animated Vega-Lite specification to encode Gapminder animation [67].

The limitations of these toolkits are overcome by new extensions, such as that in [67], which add animation features into interactive toolkits by default. The extension, *Animated Vega-Lite*, offers a model for animated visualizations as time-varying data queries, aiming to facilitate future work on interactive and animated visualization [67].

In particular, graph exploration software, such as *Gephi*, commonly incorporates interactive features like zooming, panning, and node rearrangement [28].

Furthermore, in the realm of interaction and animation, understanding the target audience of a visualization is a crucial aspect of creating effective visualization designs. Users' cognitive abilities and background are essential in their ability to interpret and interact with visual elements. The research underscores the importance of considering cognitive and demographic factors, such as spatial visualization and educational background, to enhance the accessibility and engagement of visualizations, thereby empowering communities to comprehend and analyze data effectively [58]. Moreover, the domain's specificity can impact interaction dynamics and design requirements.

3.1.4 Storytelling in Information Visualization

Information communication through visualization plays a crucial role in conveying complex, multifaceted data. Recently, there has been a notable emphasis on visual storytelling within the information visualization community, as researchers have been developing authoring tools to create compelling narratives with visual support [21].

Regarding the organization of complex information for storytelling, the framework proposed by Chen et al. [21] suggests a structured approach after the visual analytics phase (Fig. 3.7). Analysts are encouraged to engage in a story synthesis phase, organizing and presenting their findings comprehensibly to non-expert audiences. This involves selecting crucial information, summarizing it, and transforming it into a suitable representation. The framework advocates arranging information based on inherent structural facets like categories, values, time, and space, thereby creating a multi-perspective view of the findings. Additionally, it recommends using annotations and comparative views to provide detail on demand and illustrate findings through examples.

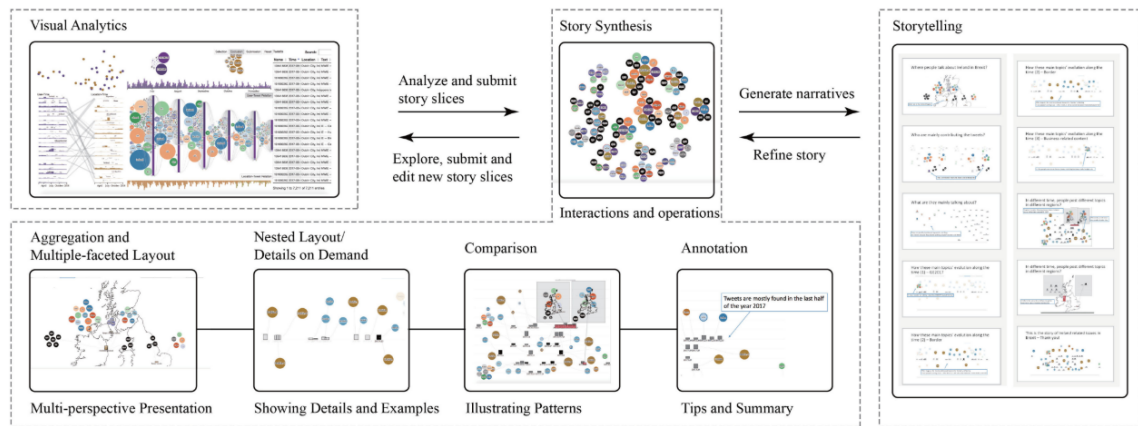


Figure 3.7: The story synthesis framework illustrated with social media data [21].

Furthermore, the proposed framework applies to any visual analytics, including those represented in graph-based visualizations. The emphasis on selecting and arranging information based on inherent structural facets can be translated to graph-based visualizations by organizing nodes and edges according to their attributes and relationships. Moreover, annotations and comparative views can provide detail and illustrate findings in graph-based visualizations, enhancing their communicative effectiveness.

In conclusion, the guidelines derived from the framework outlined in [21] offer valuable insights for analysts seeking to create compelling visual stories and communicate their findings to a broader audience. By incorporating these principles into the process of synthesizing and presenting information, analysts can enhance the clarity and impact of their visual storytelling efforts, ensuring a more accessible and informative communication of complex data.

3.1.5 Web Environment in Information Visualization

Traditionally, data management in several contexts relied on general-purpose software tools, like spreadsheets and specialized statistical tools, but faced challenges in processing subsets of data, returning analysis results, and overall efficiency [36]. Recognizing the need to target a broader audience, researchers have turned to developing interactive web-based applications. D3, a JavaScript library introduced by Bostock *et al.*, has gained popularity for its ability to manipulate documents based on data, providing powerful visualization components and offering compatibility, debugging, and performance advantages. Many researchers leverage D3 to construct interactive web-based analysis and visualization tools [46].

The shift from specialized desktop software to dynamic web pages, facilitated by application servers and web browser-based data visualization, transforms the entire process [36]. This transition allows for the emergence of tools designed to replace standard, repetitive, and time-consuming processes with visual tools that extract valuable insights from data. A similar perspective is echoed in another example [14] (refer to Fig. 3.5), where opting for an interactive environment with multiple views enhances data comprehension and exploration, including a network view.

Information visualization in a web environment involves critical aspects such as the user interface, data transfer to the client side, and user interactions [36]. A crucial consideration for successful web applications is the time required for data collection and scene rendering, emphasizing the need for efficiency to ensure smooth user experiences [36]. Relational databases, widely employed for general website data, including content management systems, are also preferred for spatio-temporal visualization due to their capacity for easy and rapid query and data extraction based on user queries [36].

Furthermore, in the context of web-based approaches, Basole and Srinivasan [5] outline best practices, emphasizing the need for visualization tools to provide multiple perspectives over the data. It underscores the importance of exploring relationships, understanding temporal dynamics, and supporting appropriate techniques to explore dynamic networks [5]. These considerations align with the idea that an ideal visualization tool in a web environment should mimic best practices in User Interface (UI) design found on the web, ensuring familiarity and user-friendly interactions [5].

3.2 Graph-Based Visualization

Graph-based visualization, often employed to represent relationships between entities, is a valuable tool for understanding the interconnected nature of data. In graph theory, a graph (G) is constructed from vertices (V) and edges (E) [46]. Beyond a mathematical abstraction, graphs serve as powerful models to map relationships, particularly in the realm of complex networks, where they enable the study of real-world events and the analysis of the structure and dynamics of complex systems [39, 46].

The exponential growth of interconnected data accentuates the importance of processing graph data, as it increasingly plays a critical role in various aspects of our lives [65]. In this context, graph visualization becomes crucial for efficiently understanding the interconnected properties of graph data. Despite its significance, the challenges posed by the exponential growth of interconnections necessitate careful consideration of efficient visual cognition in graph data [65].

Furthermore, exploring how individuals perceive and interact with graph visualizations is a valuable experiment that provides insights for constructing more efficient visualizations. The experiment sheds light on considerations that enhance the construction of graph visualizations, aiming for increased effectiveness in conveying topological information and facilitating knowledge extraction from graph data [65].

In static graph visualization, the primary visual metaphor is often the node-link diagram, a widely used representation for displaying relationships between entities [7]. This visual metaphor effectively communicates the graph's structure by depicting nodes as entities and links as connections between them (Fig. 3.8). However, in recent developments, alternative representations such as adjacency matrices and adjacency lists, or hybrid approaches combining matrices with node-link representations (Fig. 3.8), have gained particular interest in the visualization community [7].

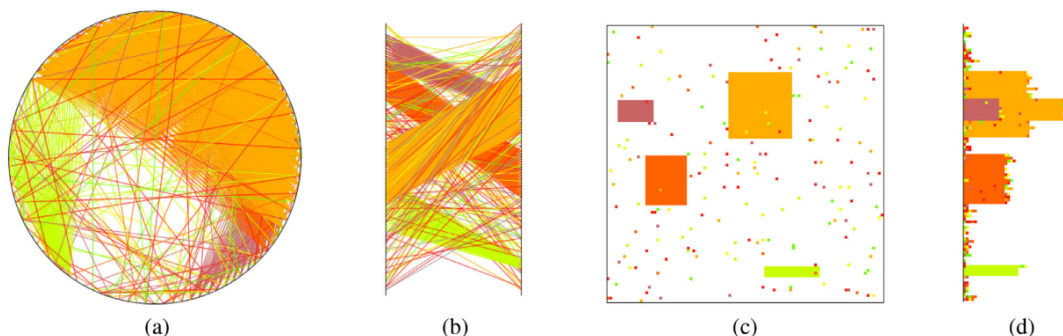


Figure 3.8: Layouts of the same static weighted relational data: (a) node-link radial layout. (b) node-link bipartite layout. (c) and (d) adjacency matrix and the corresponding adjacency list [7].

Another significant application of graph visualization is in the context of knowledge graphs, which have garnered attention for embedding domain or interdisciplinary knowledge into fact triplets. Knowledge graphs facilitate profound retrieval and reasoning based on relationships among entities, making them valuable in various domains. In educational scenarios, teachers leverage discipline knowledge graphs to achieve learning assessment, learning navigation, and personalized recommendation, showcasing the versatility of graph-based visualizations in supporting educational processes [65].

However, despite the effectiveness of various visual metaphors in static graph visualization, an additional layer of complexity arises when dealing with dynamic graph data. Unlike static graphs, many real-world scenarios involve graphs frequently changing over time, such as call graphs during software development or friend relationships in social networks [7]. In dynamic graph visualization, the number of vertices and edges is of interest, and the temporal dimension becomes a crucial factor for a visualizer. The dynamics in graph data introduce challenges in terms

of visual, perceptual, and algorithmic scalability, demanding specialized approaches to comprehend the evolving nature of relationships over time [7]. This area of dynamic graph visualization, characterized by the temporal evolution of graphs, is further explained in Section 3.2.2.

3.2.1 Graph Mining

Graph mining encompasses a broad spectrum of tasks to process, analyze, and extract meaningful information from graph data. Within this domain, several applications serve diverse purposes, including visualization, graph pattern mining, graph classification, and graph compression [48]:

- **Graph visualization:** is a crucial aspect of graph mining, focusing on techniques to represent graph data in readable, scalable, and effective diagrams. In social network research, for instance, graph visualization tools are frequently employed to represent network structures and derive insights from the data [48] (Fig. 3.9).
- **Graph pattern mining:** involves identifying patterns or subgraphs containing specific information, contributing to discovering recurring structures within the data [48].
- **Graph classification:** pertains to assigning labels to entire graphs or subgraphs based on existing data, facilitating categorization and analysis [48].
- **Graph compression:** is concerned with reducing the information in a graph without compromising its intrinsic value, aiding in efficient storage and analysis of large-scale graph datasets [48].

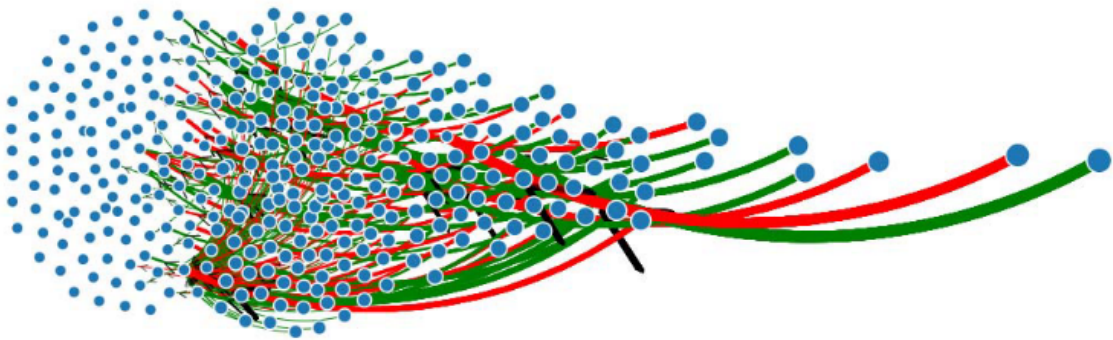


Figure 3.9: Network visualization of twitter data with Indonesian cyberbullying classification module [46].

Together, these diverse applications within the realm of graph mining underscore the multifaceted nature of this field, addressing various challenges and objectives related to the analysis and extraction of meaningful insights from graph data.

3.2.2 Dynamic Graph Visualization

Dynamic graph visualization is crucial for representing the evolving relationships within networks. It involves modeling the information as a sequence of static graphs linked with timestamps. This visualization approach is essential for uncovering hidden relationships and patterns in various applications, allowing analysts to observe temporal changes effectively [12]. In dynamic visualization, challenges arise due to the multidimensional nature of dynamic graph data, particularly the inclusion of the time dimension, making the visualization process complex [44]. The core objective of dynamic graph visualization is to provide readable, scalable diagrams that effectively capture the evolving relationships between entities. This approach facilitates the extraction of characteristics from nodes and edges, contributing to a deeper understanding of dynamic data [44].

Dynamic network visualization is a thriving research field, offering intuitive diagrams and extensive interactions to aid users in observing and analyzing the evolving nature of complex networks. Within this domain, according to the way how changes are conveyed to the user, existing methods can be broadly categorized into **animation-based** and **timeline-based** approaches [25, 29, 39] and hybrid visualizations [39]:

- **animation-based**: showcase the temporal evolution through animated transitions [23], utilizing node-and-link layouts, where each layout represents the snapshot of dynamic graphs at a given time, allowing users to observe temporal changes throughout the animation [12].
- **timeline-based**: visualization of dynamic graphs through a time-to-space mapping (mapping the time dimension to a spatial dimension). It involves displaying sequential graphs on a timeline, providing users with a sequential representation of graph evolution [23], often employing matrix-based layouts such as Adjacency Matrix Layout and Parallel Coordinates Layout [12].

Dynamic graph visualization encompasses various techniques for representing the temporal evolution of networks, typically focusing on nodes and edges. One prevalent method involves the representation of **dynamic node-link graphs**, which can take two primary forms: a time-space mapping using multiple static and small graphs in chronological order and a time map with animated graphics [44]. **Dynamic weighted graphs** are an extension of dynamic node-link graphs that, besides focusing on nodes and edges volatility, also concern edge weight changes over time. [23].

Among the timeline-based visualization forms, the **Massive Sequence View (MSV)** stands out as a classic approach widely employed for analyzing dynamic social networks (Fig. 3.10). MSV offers fine-grained visualization on a continuous timeline, leveraging fixed node locations to aid in retaining the viewer's mental map. However, challenges arise regarding readability due to mass overplotting or overlapping edges. Various techniques have been proposed to address this, categorized into abstractions in the visual or data space. These include interactions (e.g., zooming and brushing), reordering (optimizing node locations), and distortions (highlighting important patterns with visual encodings) [25].

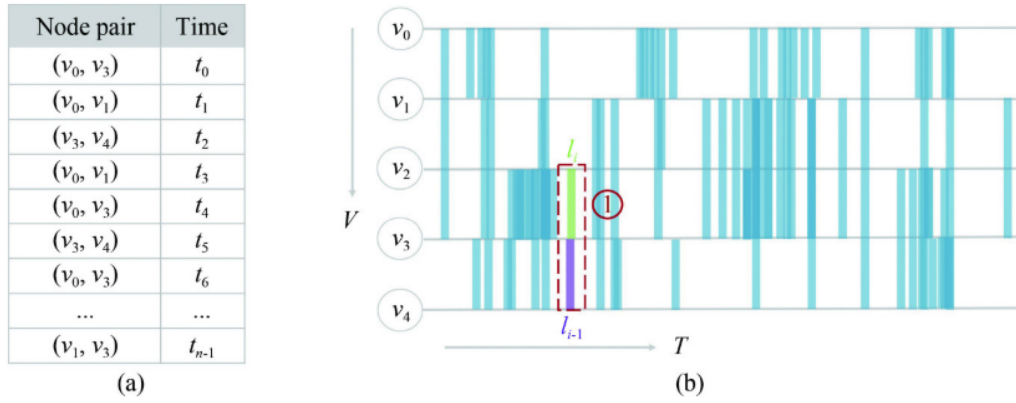


Figure 3.10: Dynamic network dataset (a) represented by a Massive Sequence View (b) [25].

Another timeline-based approach involves **flow maps** (Fig. 3.11), akin to storylines, frequently used for visualizing temporal data, including dynamic graphs. Key metrics such as line crossings and wiggles contribute to the aesthetic aspect of this visualization form, making it applicable to various scenarios, including social media data and dynamic graphs [66].

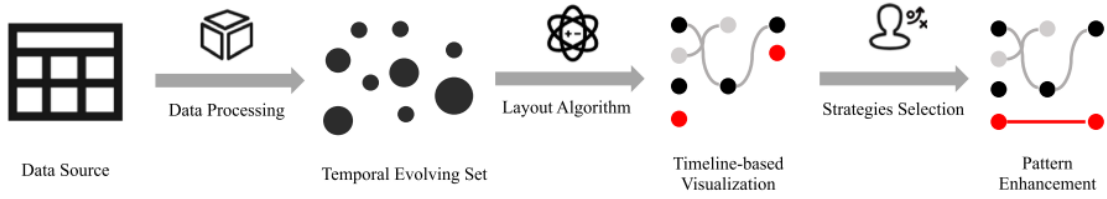


Figure 3.11: Pattern enhancement workflow in flow maps visualization [66].

This diversity in approaches highlights the ongoing efforts to enhance the effectiveness of dynamic graph visualization methods. However, the challenges inherent in visualizing dynamic graphs stem from their constant evolution over time, presenting difficulties in predicting when and how changes may occur, which is transversal to every approach described. The unpredictable nature of nodes and edges appearing or disappearing in an ad hoc manner, coupled with the inconsistent volume and arrival rate of incoming data, contributes to the complexity of dynamic graph visualization [12]. Moreover, in the quest to discern temporal evolution patterns in dynamic graphs, users are often required to engage in a mental comparison of changes or differences between adjacent timeslices, whether presented through animated graphs or a sequence of static graphs [23].

In visual analysis for dynamic networks, a key objective is to maximize drawing stability across a sequence of graph changes. This stability is closely tied to preserving the user's mental map, ultimately facilitating memorability and easing cognitive tasks [29]. The literature suggests various criteria to enhance the stability and readability of dynamic graph visualizations. Notably, minimizing changes in the placement of nodes and edges between adjacent timeslices is recommended in mental map preservation. Small multiples, an alternative approach, show multiple timeslices simultaneously on the screen (Fig. 3.12). However, challenges arise, particularly in

distinguishing the evolution of a graph due to the potential similarity of layouts presented on a screen. Moreover, mental map preservation may prove less suitable for active and large-scale dynamic graphs, where a rapid change in a substantial volume of nodes and edges occurs [12].

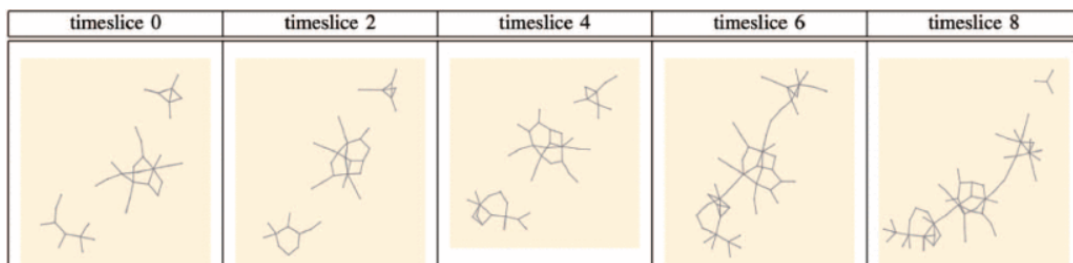


Figure 3.12: Small multiples example showing timeslices 0, 2, 4, 6, and 8 side-by-side [12].

Notably, this emphasis on mental map preservation becomes crucial in dynamic weighted graphs. The intricate nature of weight changes in these graphs necessitates a reliance on users' cognitive memory, posing challenges in effectively and quickly analyzing the temporal changes within the dynamic weighted graph structure [23]. Existing studies highlight the necessity of understanding edge weight differences and the original graph structures. The *DiffSeer* method [23] (refer to Fig. 3.23) addresses this challenge by effectively visualizing temporal changes in dynamic weighted graphs, mainly focusing on edge weight differences.

3.2.2.1 Structural Representations and Temporal Encodings

Dynamic network visualization has become a mature and thriving research field due to the increased availability of time-dependent datasets [28]. The design space for dynamic network visualization, as defined by Kerracher *et al.* [28] (Fig. 3.13), encompasses structural representation and temporal encoding, providing a comprehensive framework for evaluating different approaches.

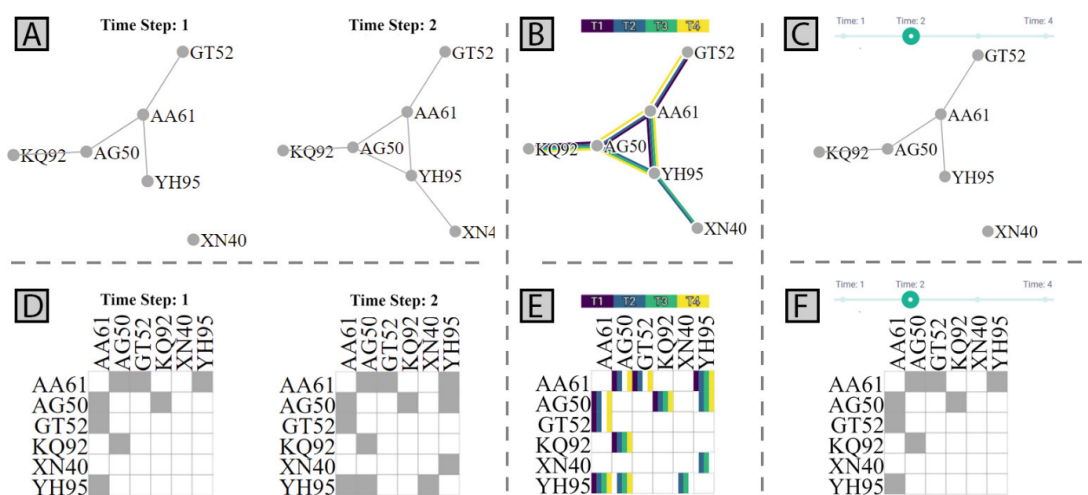


Figure 3.13: Graph structural and temporal encodings: Juxtaposition (A, D), Superimposition (B, E), and Animation with Playback Controls (C, F) [28].

On one hand, in terms of **structural representations**, two prominent methods are Node-Link (NL) diagrams and Adjacency Matrices (M) (Fig. 3.13). NL diagrams visually present the relational structure of the graph, connecting entities using lines and circles, while Adjacency Matrices visualize the network as an $n \times n$ table, indicating edge presence with non-zero values [28]. Studies suggest that NL diagrams excel in supporting memorability and connectivity tasks, while matrices are quicker and more accurate for tasks involving common neighbors and group-related activities [28].

The choice of structural representation also depends on the inherited structure and characteristics of the data. For instance, for hierarchical structures, a bubble layout, as demonstrated by Burch [20] (Fig. 3.14), enhances the recognition of the structure by replacing edges with bubbles and allowing for interactive exploration [28].

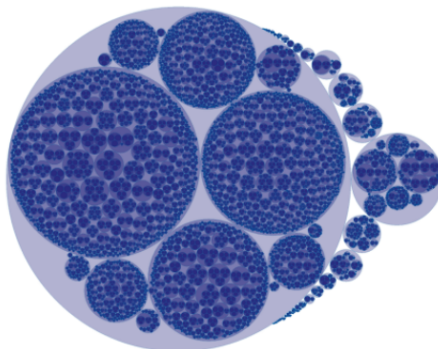


Figure 3.14: Hierarchical structures in a bubble layout [20].

On the other hand, **temporal encodings** play a crucial role in dynamic network visualization. Superimposition (SI), Juxtaposition (JP), Animation with Playback Control (ANC), and Auto Animation (AN) are different temporal encoding techniques [28]. SI overlays the temporal dimension on the same screen space, JP represents temporal dynamics as distinct layouts, ANC uses a time slider for animation control, and AN depicts smooth transitions between subsequent timeslices [28].

The study performed by Filipov [28] provides empirical evidence indicating that considering the structural representations, node-link diagrams are generally preferred over matrices in dynamic network visualization, regardless of the temporal encoding, but matrices scale better to larger graphs than node-link. Regarding the temporal encodings, it is concluded that ANC is the most favored temporal encoding when coupled with node-link base representation and that preferences between superimposition and juxtaposition encodings show no significant differences; however, superimposition is more confusing for node-link tasks compared to matrices tasks, and that Auto Animation and ANC support a more substantial number of timeslices compared to superimposition and juxtaposition. Finally, the study highlights that one of the more effective representations is combining ANC with Auto Animation coupled with a node-link representation.

Moreover, in the realm of structural representations, the comparative study of the two techniques, NL and M, performed by Burch *et al.* [8] concludes that both node-link diagrams and

adjacency matrices offer distinct advantages and limitations in visualizing dynamic graphs. Node-link diagrams are efficient for sparse and small graphs, clearly representing the graph structure and facilitating the identification of clusters and anomalies (Fig. 3.15). On the other hand, adjacency matrices are more suitable for dense networks, enabling the discovery of different grouping and clustering patterns based on the chosen reordering strategy (Fig. 3.16). By integrating both visualization techniques, the study emphasizes the potential for gaining comprehensive insights into dynamic graph data by leveraging the strengths of each approach. (Fig. 3.17).

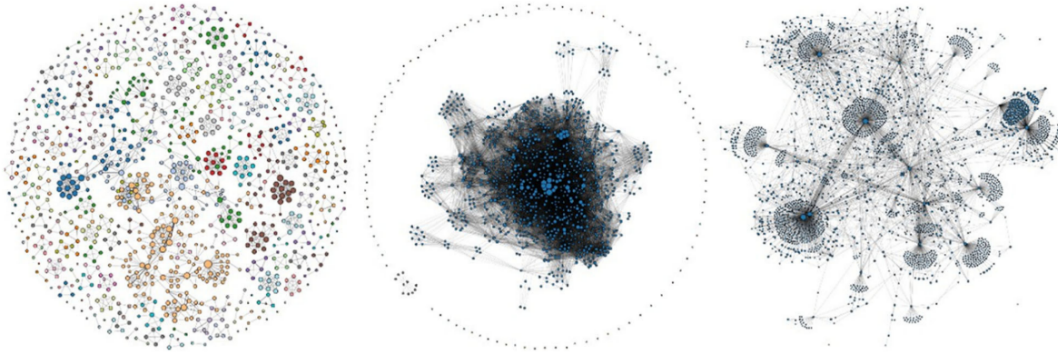


Figure 3.15: Node-link diagrams of real-world datasets: examples of large and dense graph structures [8].

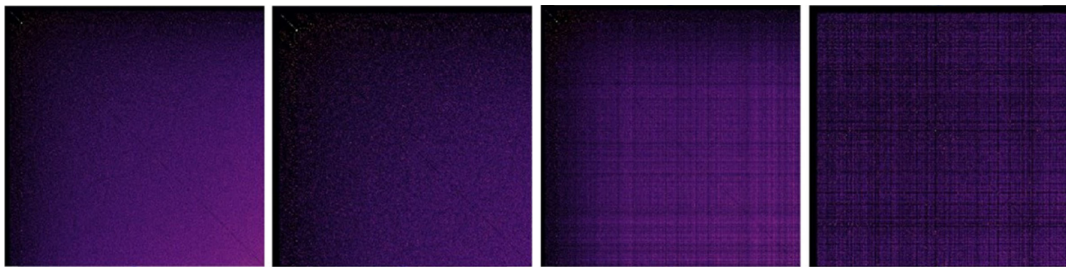


Figure 3.16: Adjacency matrices of a dense dataset reordered using distinct ordering strategies [8].

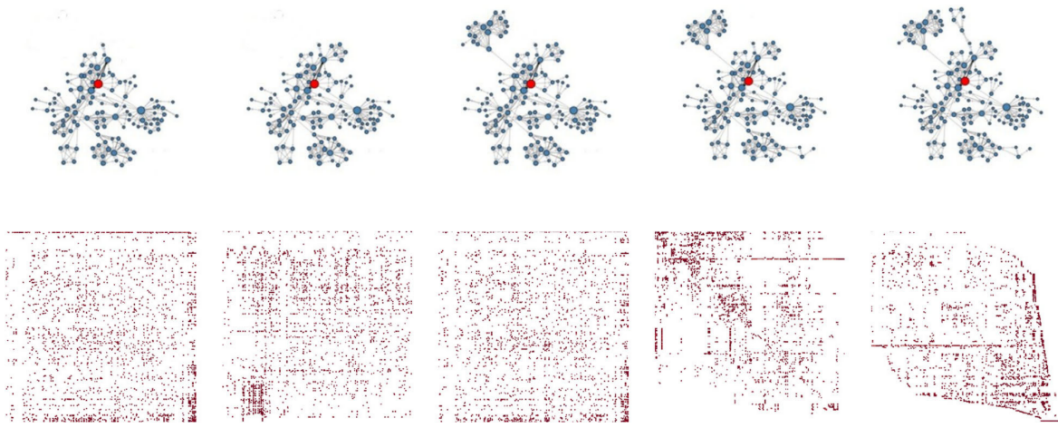


Figure 3.17: Combination of node-link diagrams and adjacency matrices to ease the identification of visual patterns [8].

In summary, the interplay between structural representations and temporal encodings in dynamic network visualization presents a nuanced landscape, where the choice of visualization method and encoding technique depends on the specific tasks, sizes of graphs, and user preferences [28].

3.2.2.2 Visual Abstraction

Graph drawing or visualization has a long history dating back to Leonhard Euler's work on the *Seven Bridges of Königsberg*, where he used a node-link diagram, a prevalent visual metaphor still used today [7] (Fig. 3.18). Over the years, numerous layout algorithms have been developed to create aesthetically pleasing representations of graphs, emphasizing readability and understanding [7, 65].

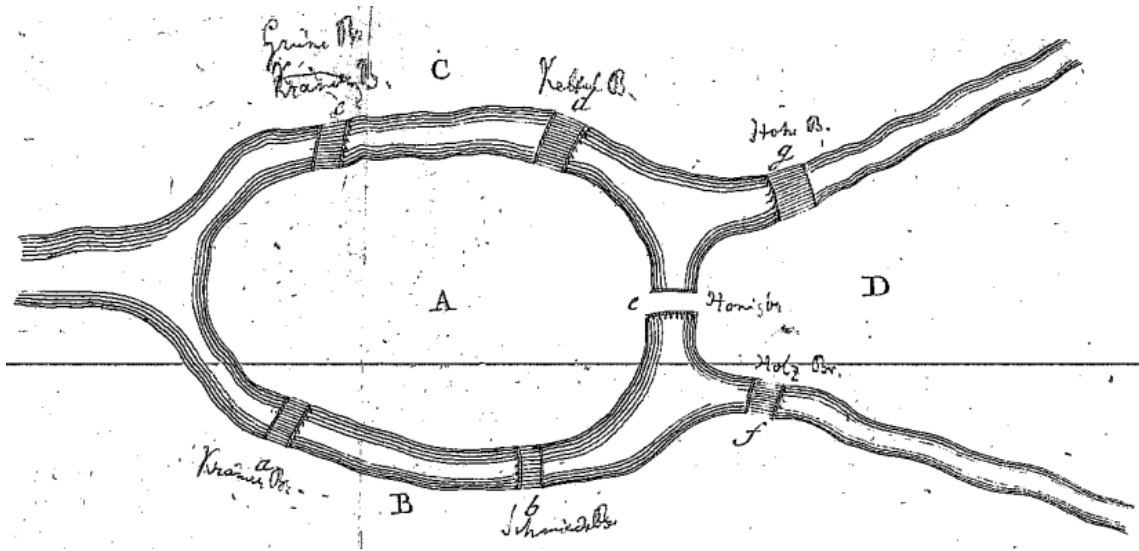


Figure 3.18: Solution of a problem in the geometry of position by Leonhard Euler: *Seven Bridges of Königsberg* [7].

Efficiency in topology-based search tasks is crucial in graph visualization, prompting the emphasis on reducing edge crossings and node-edge overlappings for improved effectiveness [65]. Visual abstraction, a form focused on conveying meaningful information intuitively, is widely used across various domains, such as trajectory analysis, sports events analysis, wireless communication analysis, and more [25]. However, in big data visualization, overdrawn and overlapping can lead to visual clutter, decreasing readability. To address this, techniques like edge-based multi-class blue noise have been introduced to reduce visual clutter effectively [25] (Fig. 3.19).

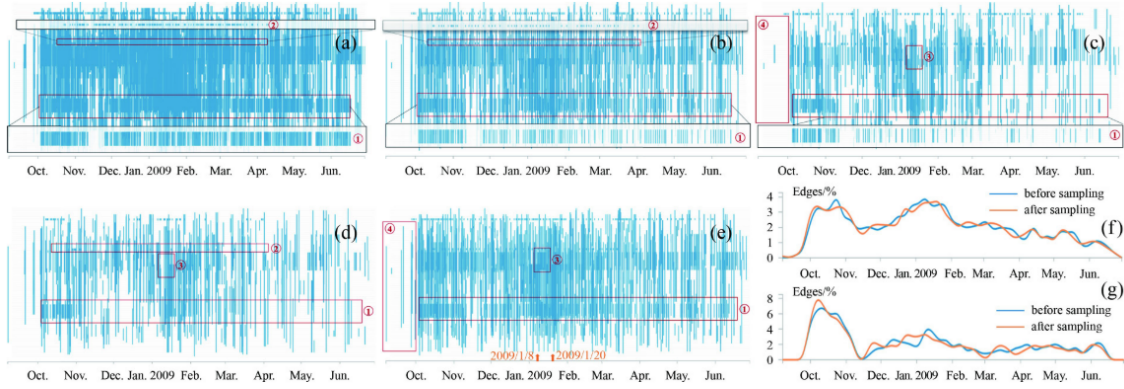


Figure 3.19: E-MCBN sampling for MSV in real-world scenarios: students' phone call dataset [25].

Temporal networks introduce increased complexity, particularly in dealing with large networks where visual clutter becomes a significant challenge. Traditional layouts like node-link diagrams, alluvial diagrams, *GraphFlow*, and *Massive Sequence View* (MSV) representations suffer from clutter, leading to difficulties in visual exploration [39]. Various methods have been proposed to address this issue, including node positioning, node and edge sampling, time-slicing, and representation simplification. Despite these advances, enabling compelling visual exploration of large temporal networks remains a challenging problem [39]. Visual analysis of dynamic graph data often incorporates temporal abstraction methods such as temporal aggregation and dimensionality reduction [11] (Fig. 3.20). The challenges posed by complex, time-varying graphs with many entities include visual clutter hindering exploration. Solutions involve partitioning large graphs into manageable sizes and collapsing similar nodes into meta-nodes. Interaction techniques then allow users to expand meta-nodes based on the desired level of detail, simplifying visual tracking of the evolution of local components [62].

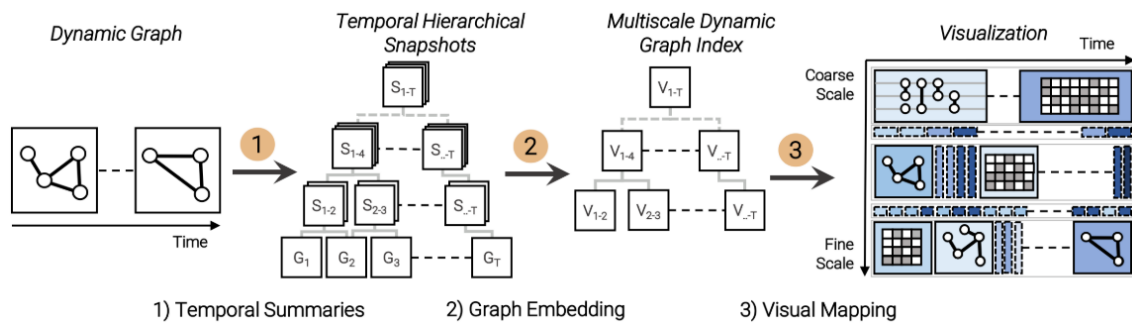


Figure 3.20: Multiscale Snapshots temporal and structural aggregates approach [11].

In summary, the evolution of graph visualization techniques underscores the importance of addressing issues related to visual clutter, emphasizing readability, and leveraging abstraction methods to facilitate compelling exploration across various applications and domains.

3.2.2.3 Force-direct algorithms

Visualizing the dynamics of a graph algorithm involves representing a sequence of graphs, either as an animation or as static representations side-by-side [9]. An effective approach to showcasing dynamic networks, where relationships change over time, is through node-link diagrams, particularly suited for tasks involving understanding network topology [64] (Fig. 3.21).

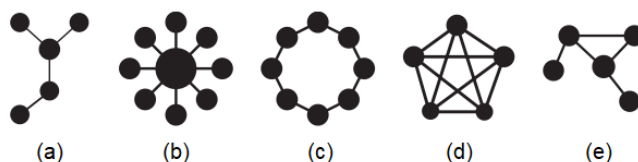


Figure 3.21: Network topology: Tree (a), Star (b), Circular (c), Low (d), and Clique (e) [39].

In dynamic networks, the force-directed method is a well-known approach to network layout and has found extensive use in information visualization [64]. This method proves valuable for capturing the changing nature of networks over time. Specifically, force-directed algorithms leverage attractive and repulsive forces between nodes to determine their positions in the layout. The resulting layout evolves over time, reflecting the underlying network dynamics.

Wei *et al.* [64] contributed a new algorithm tailored to handle the temporal evolution of graphs. This algorithm considers the changes in graph time slices, offering a nuanced perspective on dynamic network visualization. The force-directed nature of the algorithm ensures that the layout adapts to the evolving relationships within the network, providing an insightful representation of the changing structure (Fig. 3.22).

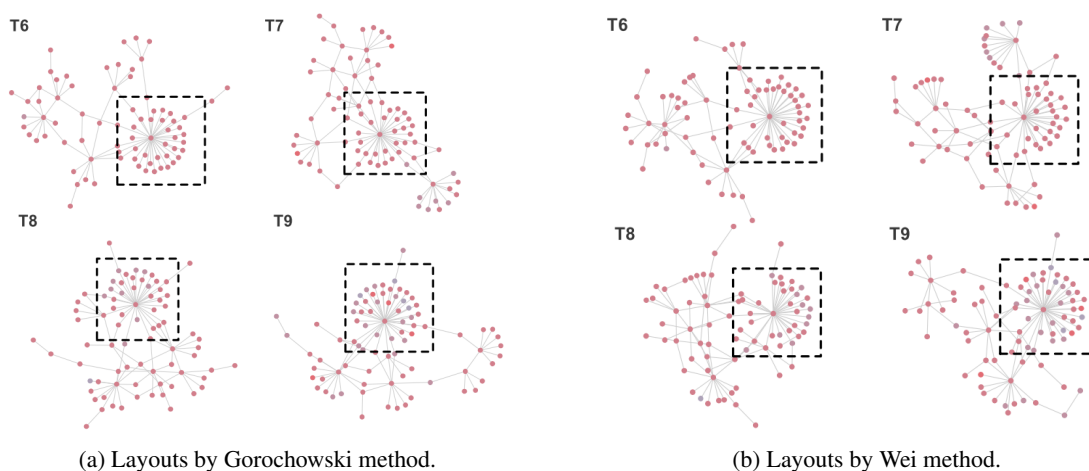


Figure 3.22: Dynamic network layout in 4 consecutive timesteps [64].

In summary, force-directed algorithms, such as the one proposed by Wei *et al.* [64], play a crucial role in visualizing dynamic networks. These algorithms offer a powerful means to represent evolving graph structures over time, enhancing our understanding of complex relationships in dynamic scenarios.

3.2.3 Technologies

Various technologies implement graph-based visualizations, each tailored to address specific challenges and requirements. These tools cater to multiple domains and analytical needs, from gaze behavior analysis to dynamic weighted graph visualization. Below is a list providing a brief overview of some noteworthy technologies in the field of graph-based visualization:

- **GazeGraphVis** [65]: is a visual analytics system designed for analyzing gaze behaviors in path tracing tasks. It offers an analytical framework to investigate how individuals explore graph visualizations across different domains, catering to knowledge graph constructors and analysts.
- **DiffSeer** [23]: introduces a novel approach to dynamic weighted graph visualization. It focuses on effectively visualizing changes in graph structures, such as edge weight differences, between adjacent time slices. This tool is particularly valuable for understanding evolving graph dynamics over time (Fig. 3.23).

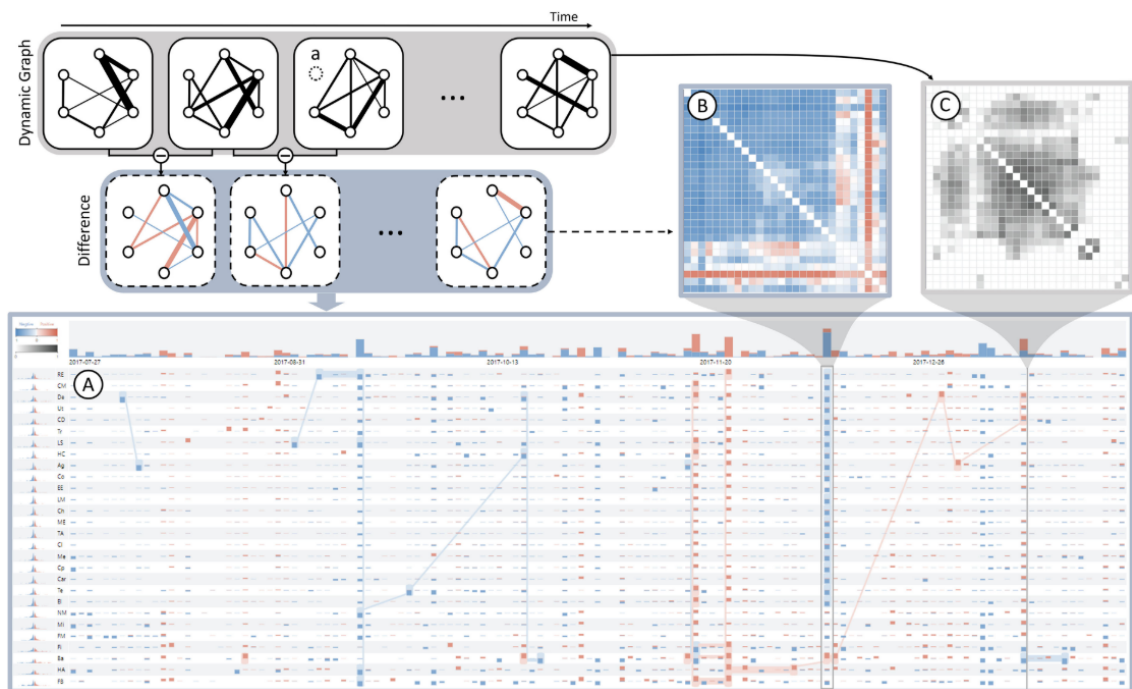


Figure 3.23: Overview of *DiffSeer* on explicitly visualizing the differences between adjacent timeslices to support the analysis of the dynamic weighted graph evolution over a long time [23].

- **LargeNetVis** [39]: an example of a tool employing multiple views for a detailed inspection of large-scale dynamic graphs. It addresses challenges associated with large datasets with temporal dimensions, providing valuable insights into complex interconnected data.
- **Collaboration Spotting X** [6]: is a system facilitating quick exploration of visualizations for highly connected data, primarily using network visualization as its main approach.

While lacking history and data annotation capabilities, it is an effective tool for overviewing large datasets (Fig. 3.24).

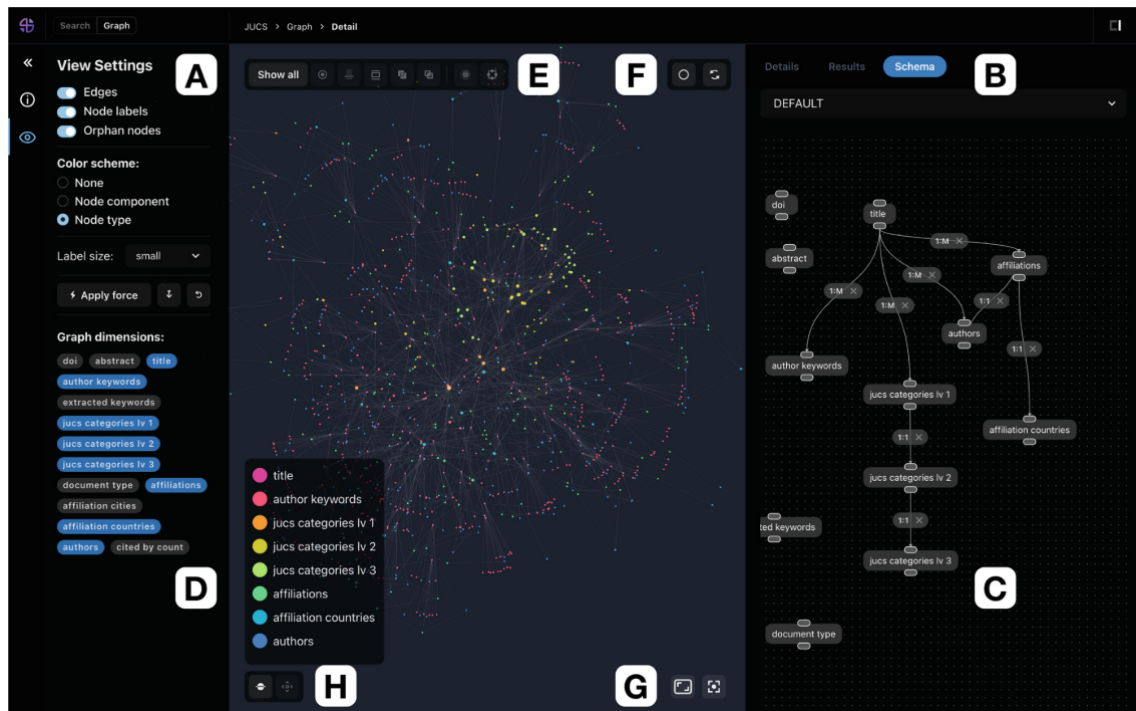


Figure 3.24: *Collaboration Spotting X* detail network and detail schema [6].

- **Cluster Framework** [48]: utilizes clustering to enhance graph visualizations (Fig. 3.25), enabling subsequent graph mining, classification, and compression. It provides a structured approach to managing complex graph structures.

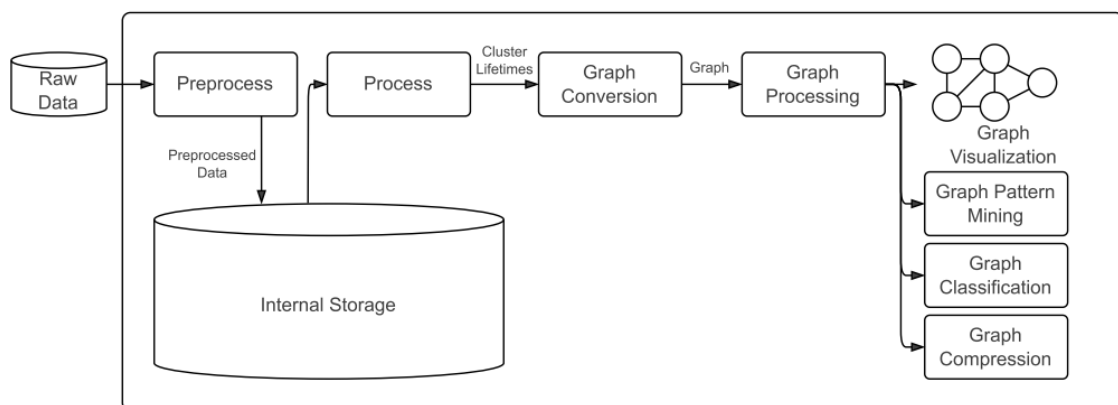


Figure 3.25: A framework for spatial-temporal cluster analysis. [48].

- **Multi Scale Snapshots** [11]: is an example of a tool offering an overview of large-scale dynamic graphs. It tackles the challenges of visualizing intricate networks, providing a valuable solution for exploring dynamic graph structures (Fig. 3.26).

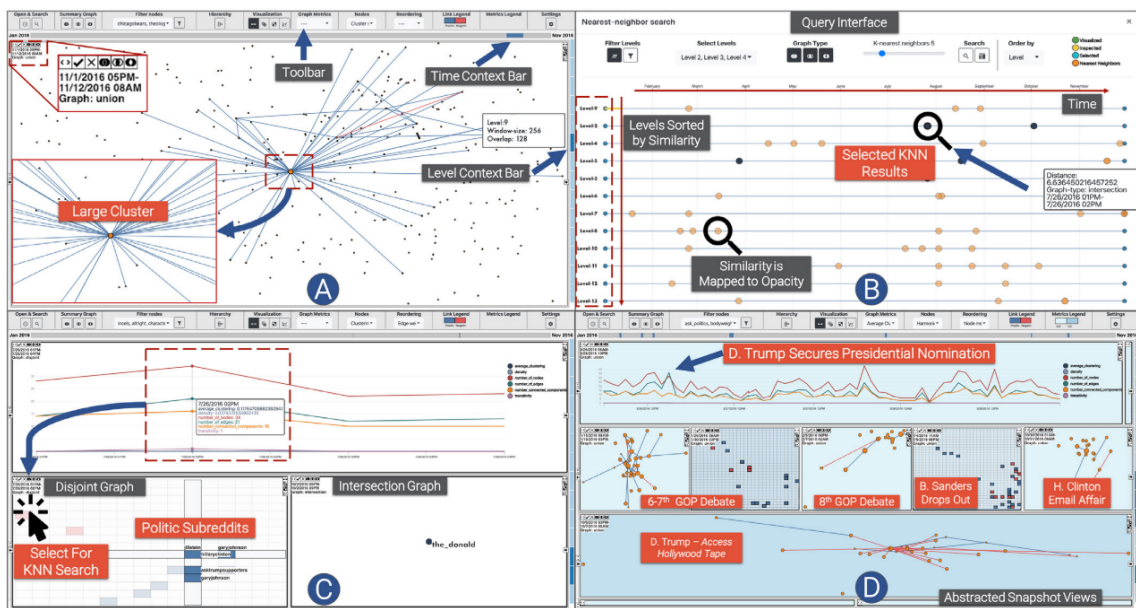


Figure 3.26: *Multiscale Snapshots* visualization (A) and the query interface (B) [11].

- **Multidimensional Objects** [60]: presents a working approach to constructing graphs for objects such as politicians. It serves as an excellent example of representing multidimensional objects through graph-based visualization. However, it lacks a temporal aspect, focusing on portraying complex objects in a graph structure.
- **Circular Graph** [51]: is an existing visual tool relying on circular graphs to represent interconnected data. While lacking a temporal dimension, it is an example of circular graph-based visualizations (Fig. 3.27).

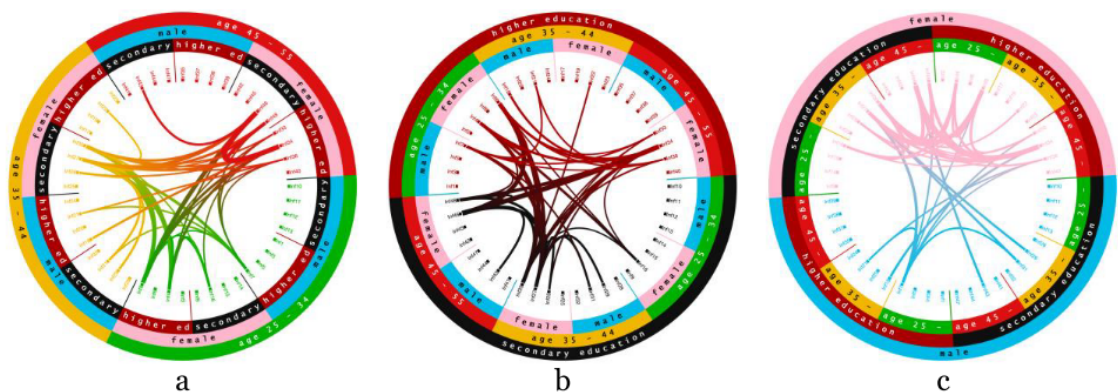


Figure 3.27: Correlation of topics in self-presentations, grouped by age (a), education (b), and sex (c) [51].

- **Edge Plots** [1]: is an approach that aligns with solution requirements, addressing the challenges of visualizing large temporal datasets while maintaining a mental map without succumbing to visual clutter.

- **TeFNet** [16]: utilizes graph visualization techniques to support tax officers in analyzing temporal data over multiple fiscal periods. It adopts a suitable timeline representation mapping the temporal dimension to the spatial dimension, aiding in analyzing temporal patterns.
- **Ecoxight** [5]: is a web-based system tailored to explore complex ecosystems visually. It handles multipartite, multiplex, multivariate, dynamic, and geospatial data. *Ecoxight* utilizes graph visualization conventions to represent ecosystems as networks and employs temporal visual encodings for users to understand evolutionary changes in the network over time (Fig. 3.28).

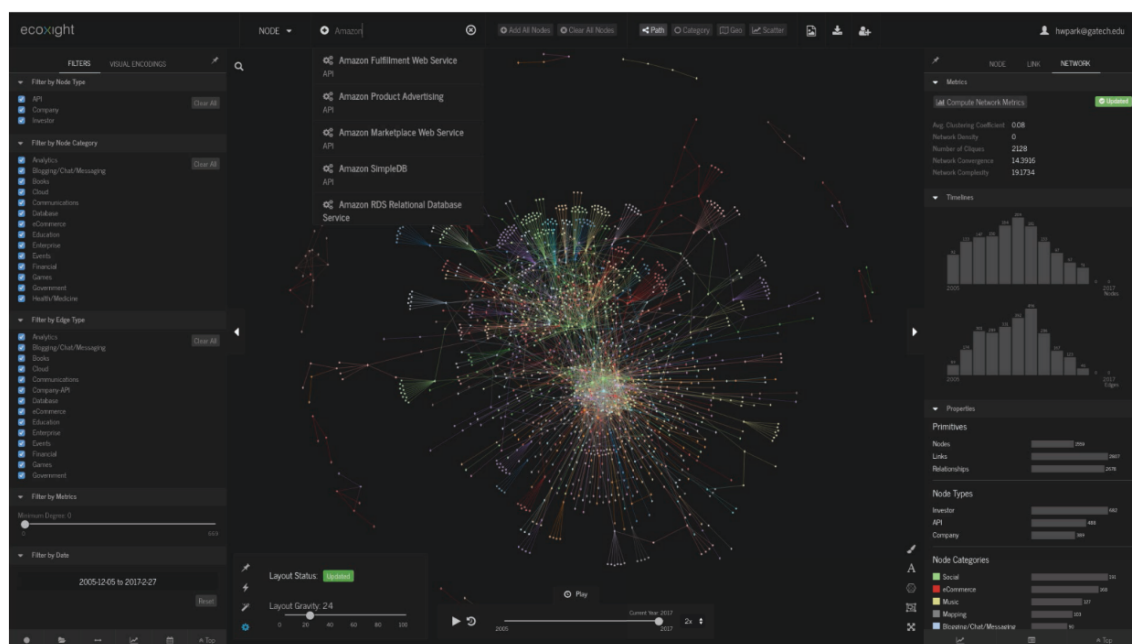


Figure 3.28: *Ecoxight* platform UI. [5].

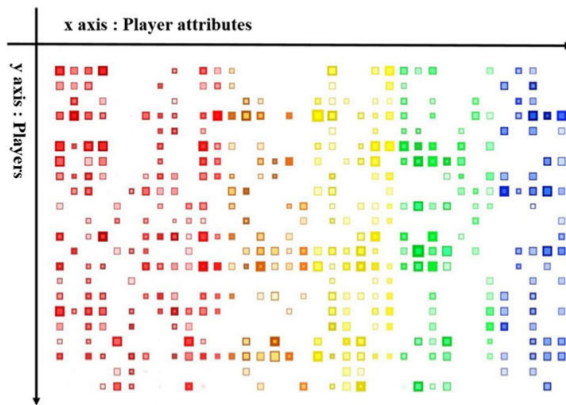
3.2.4 Applications

Graph visualization, particularly dynamic graph visualization, finds diverse applications across various domains. One prominent application lies in exploring the temporal evolution of intricate relationships between entities. This is observed in contexts such as social, financial, and communication networks, where understanding how relationships change over time is crucial for gaining insights [23]. In complex systems within the sciences, networks are powerful models for representing relationships between vertices. Examples range from the internet and the WWW to biological networks, highlighting the versatility of graph-based representations in capturing the interconnected nature of diverse systems [44]. Social networks, especially in highly connected contexts, benefit significantly from graph visualization techniques [46]. Moreover, ecosystem visualization and analysis present a challenging yet rewarding application. Ecosystems, defined as large, heterogeneous sets of interconnected entities, pose challenges due to their dynamic nature. Entities enter and leave, relationships form and dissolve, and attributes change continually. This

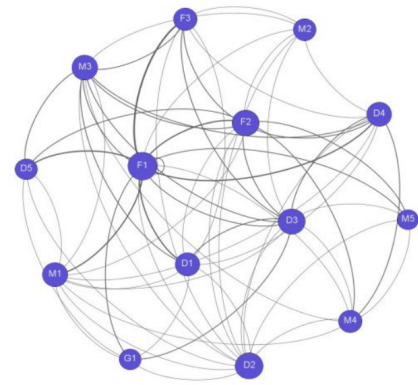
complexity requires graph visualization to address the dynamic characteristics of ecosystem data, providing valuable insights into the evolving nature of these systems [5].

3.2.4.1 Soccer-related applications

In the realm of soccer analytics, recent studies, such as [52] and [37] (Fig. 3.29), have made significant strides in scientifically collecting and visualizing soccer data to enhance performance analysis. While existing works focus on aspects like player movements, ball trajectories, and team cooperation, a notable gap emerges from the systematic literature review. The analysis highlights a lack of comprehensive studies that employ dynamic graph visualization to uncover the temporal evolution of players in clubs and national teams based on historical data. Notably, the reviewed literature indicates an untapped opportunity to delve into players' nuanced relationships, compositions, and performance metrics over time, paving the way for a more holistic understanding of soccer dynamics and strategy evolution.



(a) Team performance comparison visualization: Real Madrid FC vs. Stade Rennais FC [52].



(b) Passes network in the Huskies team visualization [37].

Figure 3.29: Visualizations in soccer-related applications.

3.3 Summary

The contemporary landscape of information and graph-based visualization, particularly dynamic graph visualization, underscores the increasing importance of visual storytelling for the effective communication of complex data. A primary challenge lies in presenting data in a user-friendly format, prompting the use of innovative techniques like infographics, which are especially crucial in domains such as sports with comprehensive and interconnected data requiring dynamic presentations.

The literature emphasizes data's transformative role in shaping decision-making processes across diverse domains. In this context, information visualization is recognized as pivotal for translating data into visual form and actively influencing cognitive processes in comprehending

and extracting insights. Chen *et al.*'s framework [21] advocates for structured organization in storytelling, extending its applicability to graph-based visualizations.

Acknowledging the potential of graph-based visualization for representing highly connected, temporal, and multidimensional data illuminates pathways to enhance understanding of intricate relationships and dynamic changes over time. Graphs prove valuable for unraveling interconnected data within complex networks and capturing relationships and temporal dependencies through animations. However, visualizing large graphs requires efficient consideration of visual cognition.

Despite the significance of graph data, challenges arising from its exponential growth necessitate exploration into user perceptions and interactions, offering insights for constructing more efficient visualizations and enhancing knowledge extraction. Identified challenges, including overwhelming users and the need for effective animation and data reduction techniques, highlight the importance of guidelines for organizing complex information for storytelling.

For node-link dynamic graph-based visualizations, integrating an animated timeline with controls emerges as a promising strategy, aligning with the need for effective animation and providing a dynamic framework for exploring temporal changes and intricate relationships.

The literature also identifies a notable gap in soccer data visualization, explicitly depicting player trajectories within clubs and national teams over time. Addressing this gap holds the potential to enrich the understanding of player dynamics and soccer data evolution.

In conclusion, this synthesis emphasizes the critical role of visual storytelling and the transformative impact of information visualization. While graph-based visualization offers valuable insights, challenges in visualization and interpretation, particularly with large and time-evolving datasets, necessitate efficient visual cognition. The exploration of user perceptions contributes to constructing more effective visualizations, with the identified soccer data gap offering a promising avenue for future research. Overall, this provides a comprehensive guide for researchers and practitioners, highlighting the potential of graph-based visualization techniques and the need for tailored strategies to convey complex information effectively.

Chapter 4

Problem Statement

This chapter consists of an expanded elucidation of the problem introduced in 1.2. Section 4.1 meticulously details the specific issue targeted for exploration in this work. Subsequently, Section 4.2 introduces and characterizes the research questions that will guide the entire investigative process. Moving forward, Section 4.3 outlines the essential requirements for any viable solution aimed at addressing the identified problem comprehensively. Finally, Section 4.4 succinctly presents the scope of this work, providing clear parameters for the study's focus.

4.1 Problem

In the current era marked by the exponential surge in digital data, decision-makers find themselves equipped with unprecedented information [5]. The contemporary landscape is transforming, driven by data-centric decision-making and innovative solutions. The correlation between effective data utilization and intelligent decision-making is evident [54]. This principle holds particularly true for entities like ZOS, a company that manages a substantial volume of intricate and historical sports data that must be presented in an attractive and understandable manner to provide valuable insights to analysts, coaches, and sports enthusiasts is not trivial.

However, the complexity, connectivity, and temporal dimensions of the available information pose a significant challenge: how to capture and present dynamic entities with multifaceted and evolving relationships in a way that is easy to understand?

To meet this challenge, it is necessary to present that information to the users in an accessible, appealing, simple, and intuitive way that ensures clarity without overwhelming users. The online landscape adds an extra layer of complexity, given the diverse user profiles and varying levels of visual literacy. Moreover, the information must remain visually consistent throughout the storytelling process to effectively convey the information.

The challenge for ZOS goes beyond showing data; it's about turning complexity into a clear, visually appealing story that connects with a diverse audience. This approach allows ZOS to unlock its data's true potential, giving stakeholders the tools to make informed decisions and uncover meaningful insights in our rapidly changing digital landscape.

4.2 Research Questions

The problem identified in the previous section, as presented in Section 1.4, can be translated into the following set of research questions:

Q1: Which visualization techniques are commonly used to present highly connected data?

Exploration of common visualization techniques tailored for presenting highly connected data. The objective is to pinpoint methods that allow viewers to intuitively grasp the intricate relationships inherent in the data.

Q2: Which visualization techniques are commonly used to present temporal data?

Identify prevalent visualization techniques for presenting temporal data. The emphasis is on ensuring that viewers can easily discern changes over time and comprehend the evolving nature of the presented information.

Q3: Which visualization techniques are commonly used to present multidimensional and multi-variate data?

Uncover the most commonly used methods for presenting multidimensional and multivariate data. The objective is to allow viewers to effortlessly understand multiple perspectives of the data simultaneously.

Q4: What are the limitations and applications of graph-based visualizations for time-evolving data?

Understanding the main applications and core limitations of dynamic graph-based visualizations. It is crucial to delineate scenarios where such visualizations excel and to comprehensively consider their usability, encompassing advantages and disadvantages, and comprehension for viewers with varying levels of data literacy.

4.3 Solution Requirements

The proposed solution must encompass the following key requirements to effectively address the challenges outlined earlier and respond to the research questions.

Dynamic Visualization of Players in Clubs and National Teams Over Time.

The solution should enable the visualization of players in both clubs and national teams over time, considering the changes in player affiliations through transfers, loans, and retirements. For this purpose, a mechanism should be implemented to allow users to inspect the network in specific years. This requires employing techniques that maintain visual consistency when transitioning between different years, ensuring a coherent and understandable representation of the evolving player-network dynamics.

Intuitive Representation of Player Transitions, Career Trajectories, and Team Dynamics.

The interface should provide a user-friendly means to track player transitions, visualize career trajectories, and understand team dynamics over time. Clear visual cues and interactive elements should be incorporated to highlight significant events, such as transfers or pivotal moments in a player's career, fostering an intuitive understanding of these aspects.

User-Friendly Interface, Intuitive Controls and Animations.

The interface design, manipulation controls, and animations must prioritize user-friendliness and intuitiveness. This consideration is crucial given the diversity of users and varying levels of data literacy in the online universe. Interactive controls should be easily accessible, allowing users to navigate the visualization seamlessly, with animations enhancing comprehension without causing confusion.

Compatibility with Web Browsers and Adherence to ZOS Project Standards.

The implementation must be compatible with commonly used web browsers, ensuring a wide accessibility range for users. The solution should align with the standards and architecture of the ZOS project, facilitating seamless integration into its website.

By incorporating these considerations into the solution design and implementation, ZOS can provide a robust and engaging platform for visualizing the intricate relationships between players, clubs, and national teams over time, catering to online users' diverse needs and literacy levels.

4.4 Scope

The solution focuses on how the users will consume the information through visual representation on the ZOS website. The primary objective is to create visually compelling infographics that effectively convey the intricate relationships between players, clubs, and national teams over time. An essential element of this endeavor is the integration of animations to attract and engage users, providing a dynamic and immersive experience. The design and implementation will prioritize user experience, ensuring an intuitive interface that is visually appealing and accessible to a diverse audience. User interactivity will be a central focus, allowing seamless navigation and exploration of player transitions, career trajectories, and team dynamics.

While the initial solution will not incorporate the capability for professionals to select specific data to display, being tailored for gathering only players within clubs and national teams' information over time from the ZOS databases, due to time constraints, this limitation is acknowledged, and the potential for future improvement is recognized.

In conclusion, the scope of the work being developed is to contribute a comprehensive solution that transforms data into dynamic infographics, offering an engaging platform for users interested in visually exploring the evolving landscape of sports dynamics on the ZOS website.

4.5 Summary

This chapter delineates in Section 4.1 the specific issue under exploration, emphasizing the exponential growth of digital data and the challenges in presenting intricate and historical sports data in an engaging and comprehensible manner and underscoring the link between effective data utilization and intelligent decision-making. Section 4.2 introduces and characterizes the research questions guiding the investigative process. Section 4.3 outlines essential requirements for a viable solution, emphasizing the need for accessibility, appeal, simplicity, and intuitiveness in presenting dynamic entities with complex and evolving relationships. Furthermore, Section 4.4 succinctly defines the scope of this work, providing clear parameters for the study's focus. In light of these definitions, the chapter transitions to the next, laying the groundwork for the proposed solution designed to address the challenges outlined in this chapter, "Problem Statement".

Chapter 5

Solution

This chapter introduces the envisioned solution for visualizing the evolution of players within national teams and clubs. In Section 5.1, the proposed solution is elaborated, providing insights into its anticipated features. Section 5.2 delves into the implementation and design details of the prototype mirroring the proposed solution.

5.1 Proposed solution

The literature review conducted in Chapter 2 and 3 has revealed a prominent category in the realm of visual representation techniques for highly connected time-evolving data: **dynamic graphs visualization**. Within this category, two primary strategies for representing graphs at different time frames have emerged: **timeline** and **animation**. The timeline approach offers simplicity and objectivity in representation. Still, it poses challenges in analyzing and manually inspecting differences between static network representations, especially with large networks that lack layout consistency between frames. On the other hand, animation can express the evolution of variables over time more clearly; however, it implies a crucial concern regarding graph layout and interactivity. Notably, when considering structural representation within graphs, the combination of animation with node-link representation outperforms other evaluated combinations.

The literature emphasizes the role of dynamic representation in enhancing user engagement and storytelling. However, it also acknowledges the complexity introduced by dynamics, raising the risk of overwhelming users. To address this, techniques such as clustering and multi-level views have been employed to minimize visual clutter.

The literature underscores the significance of selecting a technique aligned with data characteristics and context. In the context of ZOS, a sports domain in the online universe, ensuring the simplicity and intuitiveness of visualizations is essential. Simultaneously, the visualizations must be engaging and offer interaction mechanisms to control the displayed data, facilitating the distinction between player transactions and the comprehension of the story involving players, clubs, and national teams.

Based on the main considerations from the literature presented in the previous paragraphs and the New York Times visualization for the soccer clubs with players on at least two national in the Brazil 2014 World Cup [47], the proposed solution is an **animated-based dynamic weight circular graph visualization**. The primary concept behind this visualization is representing national teams as nodes in an outer circle and clubs as nodes in the inner circle (respectively, orange and blue nodes in the main network window in Fig. 5.1). Players are organized into club clusters to reduce visual clutter, with the club cluster node size representing its overall contribution to the national teams in the outer circle and the club-national team's edge weight indicating the number of players from a club participating in a specific national team (edge weight differences in Fig. 5.1). Its composition and edge strength determine a club's placement within the inner circle. In practical terms, clubs comprising players from various national teams from opposed positions in the outer circle in roughly equal numbers will find themselves positioned closer to the center. While clubs heavily dominated by players from a specific national team will be situated closer to that particular national team node in the outer circle of the visualization (mimicking the ChronoView [3] logic). Clicking on a specific club triggers a separate window above the visualization, revealing a detailed view of the cluster - the connection between club players and national teams. This mimic can be extended to the national teams, where clicking on a specific national team reveals the players that compose that national team. In both perspectives, the relative placement between the club and respective players in the two windows is maintained to increase visual consistency. Additionally, national teams without club or player affiliations can be visually distinguished, perhaps using a gray color or transparency percentage. This organizational structure is illustrated in Fig. 5.1.

As for player evolution, the visualization employs animation between network frames to visually convey changes, such as node translation and nodes appearing and disappearing. Users can navigate through the timeline, moving backward and forward between years, selecting specific years, and playing the entire animation for a comprehensive understanding (bottom of Fig. 5.1).

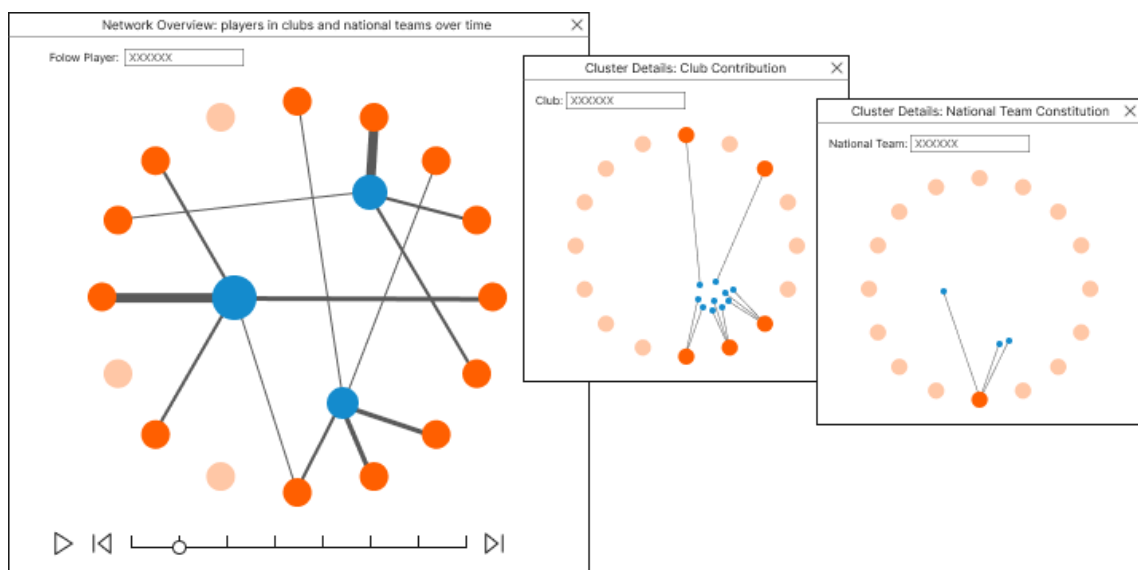


Figure 5.1: Proposed Solution - Visualization Structure Schema.

Additional interactions can be contemplated, such as sorting outer nodes to alter the data perspective and filtering inner and outer variables to ease data inspection and exploration. Even with these dynamic changes, the visualization structure's underlying logic remains consistent with the previously explained framework. Furthermore, users can search for a specific player to easily trace their trajectory, using visual variables to highlight its associated nodes when present in the active network frame.

The solution's features are in accordance with the requirements outlined in Section 4.3. Given the time constraints for solution development, it is imperative to establish priority levels for the implementation of features. In this context, the priority sequence is as follows: prioritizing the graph structure and organization, encompassing the detailed cluster view; subsequent focus on animation and controls for navigating through the year timeline; and the specific player highlight. If time permits, additional features such as sorting and filtering will be considered in the final stages of development.

5.2 Prototype Implementation

The prototype (Annex A) comprises two primary components: data gathering and data visualization. The first component involves communication with an external Application Programming Interface (API), specifically the zerozero API, to collect the data for representation. The response from the API is saved in a cache file for future access, reducing loading times upon the subsequent visualization access. The second component features a frontend developed in PHP. This interface retrieves the data from the cache file, transforming it into a format suitable for manipulation and consumption by the visualization. This initial parsing process enhances the performance of the visualization, minimizing the computational power needed during user interaction. The visualization is implemented using JavaScript and D3 [43]. Users can interact with the visualization to explore the information. The data flow and architecture are illustrated in Fig. 5.2. Further details on each component's implementation are provided in subsequent sections.

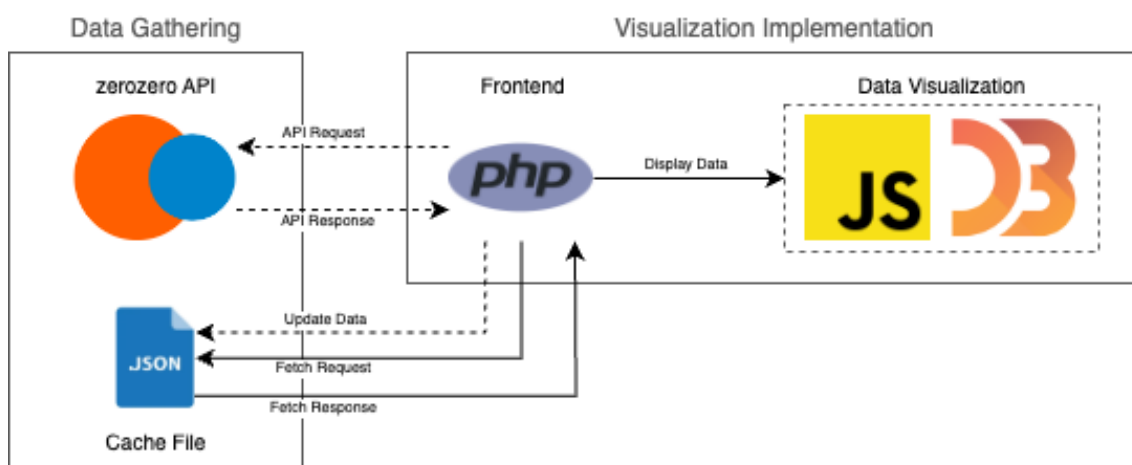


Figure 5.2: Prototype's data flow and architecture.

5.2.1 Data Gathering

The prototype focuses on tracking the **evolution of players in national teams and clubs during the World Cups**, from the first tournament in 1930 to the most recent in 2022.

Data collection involves communicating with an external Application Programming Interface (API), specifically the zerozero API, to gather the necessary data. The structure of the JSON object retrieved from this endpoint is presented in Listing 5.1.

```
1 {
2   "Name": "Campeonato do Mundo",
3   "Img": "https://www.zerozero.pt/img/logos/competicoes/",
4   "seasons": {
5     "Uruguai 1930": [
6       {
7         "id": "814",
8         "img": "https://www.zerozero.pt/img/logos/equipas/814_imgbank_flag.
          png",
9         "fk_continente": "4",
10        "name": "Argentina",
11        "teams": {
12          "1179": {
13            "color": "273E7C",
14            "logo": "https://www.zerozero.pt/img/logos/equipas/1179
              _imgbank_1686668524.png",
15            "name": "Boca Juniors",
16            "continente": "America do Sul",
17            "country": "Argentina",
18            "count": "4",
19            "players": {
20              "20750": {
21                "name": "Roberto Cherro",
22                "posicao": "Avancado",
23                "link": "/jogador/roberto-cherro/20750",
24              },
25              // other players
26            }
27          },
28          // other teams
29        }
30      },
31      // other national teams
32    ],
33    // other World Cup tournaments
34  }
35 }
```

Listing 5.1: Example of the JSON object retrieved from the endpoint.

The JSON object is organized into a hierarchical structure. The keys of the *seasons* property correspond to World Cup tournaments, with each value containing detailed information about that tournament, structured into three nested levels. The first level is the list of the national teams participating in that World Cup. For each national team, the *teams* property specifies the clubs that contributed with players to that national team. Within each club, the *players* property identifies the individual players involved in that relationship (national team to club).

Each entity (national team, club, and player) is accompanied by supplementary descriptive attributes used to construct other aspects of the visualization. These attributes include names, images, continents, countries, player positions, and links to detailed profiles.

5.2.1.1 Cache Mechanism

Due to the stale nature of the information and the time and resource-intensive nature of the API request - which involves complex queries across multiple tables - the API response is saved in a JSON file on the server side for future use. The content of this cached file is updated if it has been latest modified more than a month ago or upon a specific request (by adding the query parameter *clean* with the value *1* to the URL). If neither condition is met, the visualization is generated using the data from the cached file, bypassing the need for a new API request.

This caching mechanism significantly reduces the load time of the visualization, enhancing user experience.

5.2.2 Technology Selection

When evaluating various technologies for implementing the visualization, compatibility with ZOS standards and technologies was a primary consideration. Given that the zerozero website, where the visualization is expected to be integrated, is developed using PHP, HTML, CSS, and JavaScript, selecting a library that seamlessly integrates with this existing infrastructure was essential. Another critical factor was the flexibility and alignment of the technology with the solution requirements outlined in Section 4.3. Additionally, ensuring broad web browser support and compatibility was crucial to guarantee widespread accessibility across commonly used browsers.

Given these considerations, the best strategy was researching JavaScript visualization libraries suitable for implementing dynamic graph-based visualizations. In this context, D3.js [43] emerged as a leading choice. D3 is extensively used in data visualization, particularly in web environments, as presented in the state-of-the-art (Section 3.1.5). Although D3 requires more manual coding due to its raw nature and lack of user-friendly APIs and built-in features found in newer libraries (such as *vasturiano/force-graph* [61]), its flexibility in controlling the layout and interactions can be highly advantageous, especially for developing innovative and custom visualizations. Many specific libraries have default settings that cannot be overridden and do not align with the proposed solution (e.g., *vasturiano/force-graph* does not allow the specification of the drag event to specific nodes and stops receiving hover events when dragging a node).

No other library explored offered sufficient advantages to justify discarding D3.js as the primary technology. Furthermore, numerous examples of graph-based visualizations using D3 support its suitability for this work, including layout configuration with simulated forces¹ and timeline manipulation². Combining simulated forces to define node placement and organization with timeline animation and manipulation closely approximates the proposed solution's visualization.

5.2.2.1 Graph Layout Exploration

To become more familiar with the selected technology, the first step involved exploring the layout organization using D3's *d3-force* mechanism to assess its suitability for addressing node placement challenges. Fig. 5.3 corresponds to initial experiments implemented with D3 using mock data.

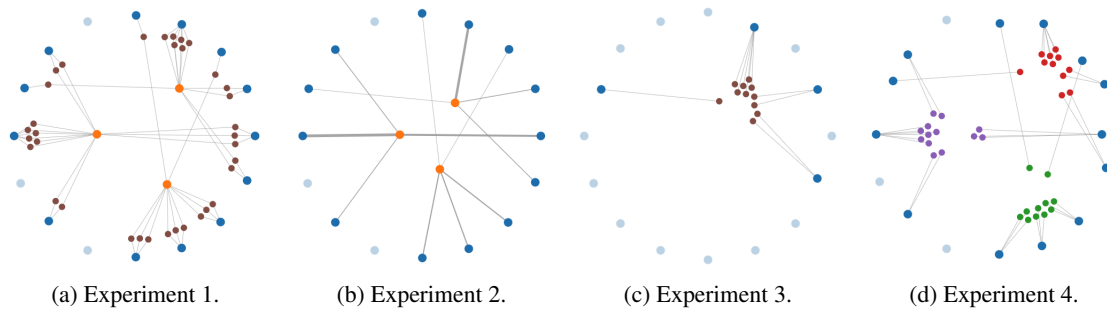


Figure 5.3: Graph layout experiments using D3 and *d3-force*.

Fig. 5.3a attempts to replicate the node organization scheme from the NYT example [47]. Fig. 5.3b approaches the organization of the main visualization in the proposed solution, while Fig. 5.3c explores the layout of the clustered details visualization. Additionally, Fig. 5.3d examines how the visualization would appear without clustering, providing a more granular view of the network where individual players are represented as nodes directly on the first level, distinguished by club colors.

These initial explorations into layout manipulation using D3 highlight its potential to replicate the desired organizational structure. A comparison between Fig. 5.3a and 5.3d, and Fig. 5.3b and 5.3c reveals differences in the number of nodes and links between approaches. Specifically, the former two show a higher density of nodes and links than the latter, demonstrating scalability and adaptability to dynamic environments for configurations with fewer nodes. This observation supports the decision-making process for the visualization idealized detailed in Section 5.1.

¹<https://observablehq.com/@d3/force-directed-graph/2>; <https://observablehq.com/@d3/disjoint-force-directed-graph/2>

²<https://observablehq.com/@d3/temporal-force-directed-graph>

5.2.2.2 Graph Animation Exploration

To explore implementing a timeline associated with a force-directed graph, an example was developed to test D3's capabilities in animating changes in graph nodes across a timeline using mock data. Fig. 5.4 illustrates the visualization at three consecutive timeslices.

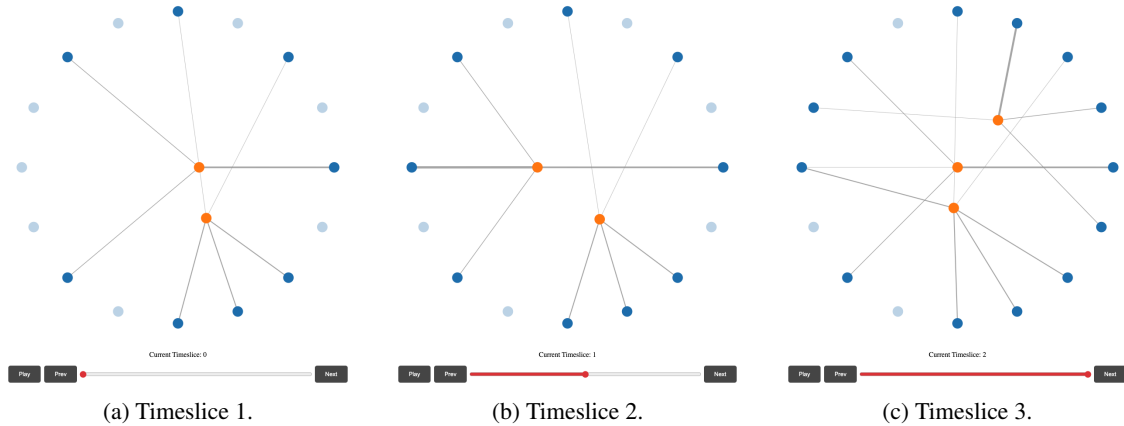


Figure 5.4: Graph animation experiment using D3 in three consecutive timeslices.

This experiment showed that D3's force simulation for node positioning automatically handles animations such as translations, insertions, and deletions between timeslices. By consistently assigning IDs to nodes throughout the timeline, D3 can differentiate between enter, update, and exit nodes. This capability allows for the codification of animations tailored to each scenario.

5.2.3 Graphs Layout

The visualization consists of two graphs that share the same organizational structure of nodes yet differ in implementation details. The main graph, also called the overview graph, illustrates the overall contribution of clubs to national teams. In contrast, the details graph focuses on individual players from a club and their affiliations with national teams or players from a national team distinguished by their clubs. Due to these distinctions, the layout rules for each graph also vary. Further details on the forces utilized, and visual variables employed within and between the graphs to maintain visual consistency are elaborated in the following paragraphs.

5.2.3.1 Design and Visual Variables

The visualization design leverages various visual variables to enhance clarity and interpretability across the two graphs, the overview graph and the details graph, providing users with a comprehensive view of the data. The visual variables depicted in the following paragraphs are implemented using D3's capabilities for manipulating SVG elements.

Size. Size emphasizes elements' importance based on their contribution value. In the overview graph, the size of a club node indicates its total contribution to the national teams, while the size of a link represents a club's specific contribution to a national team.

Shape. All nodes are represented by circles, except for club nodes contributing more than four players, represented by their respective logos. Initially, all clubs were represented by circles (Fig. 5.5a). Replacing all circles with logos was considered, but performance issues and poor visibility at smaller node sizes made this impractical (Fig. 5.5b). Another approach involved placing a circle behind each logo for visual consistency, resulting in increased visual clutter (Fig. 5.5c). Logos are used selectively to highlight clubs with higher contribution values (Fig. 5.5d). Using logos in this manner has a greater visual impact and conveys more meaning than labeled circles.



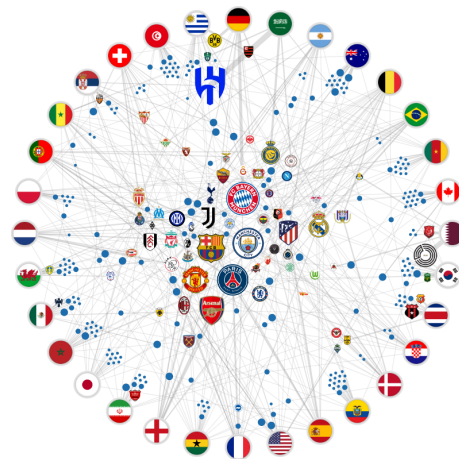
(a) Represented only by circles.



(b) Represented only by logos.



(c) Represented with circles placed behind logos.



(d) Represented either by a circle or logo.

Figure 5.5: Exploration of the visual variable shape to represent club nodes in the visualization.

Color and Fill. National team nodes are filled with their respective flag images. In contrast, club and player nodes are filled with respective club colors to facilitate visual matching between the overview and detailed graphs, besides maintaining their relative placement.

The impact of the color in club nodes was tested for the shape approaches presented in the previous paragraph. However, in cases where visual clutter was already evident, distinguishing the nodes by color only exacerbated the issue (Fig. 5.6a and 5.6b), even after reducing opacity (Fig. 5.6c). The final design selectively uses color in smaller circles, as the meaning in larger circles is typically conveyed by the club logos (Fig. 5.6d).



(a) Represented only by colored circles.



(b) Represented with colored circles placed behind logos.



(c) Represented with colored circles with lower opacity placed behind logos.



(d) Represented either by a colored circle or logo.

Figure 5.6: Exploration of the visual variable color to represent club nodes in the visualization.

On another note, displaying club logos within player node clusters was considered but discarded due to visual clutter when groups are close together (Fig. 5.7).

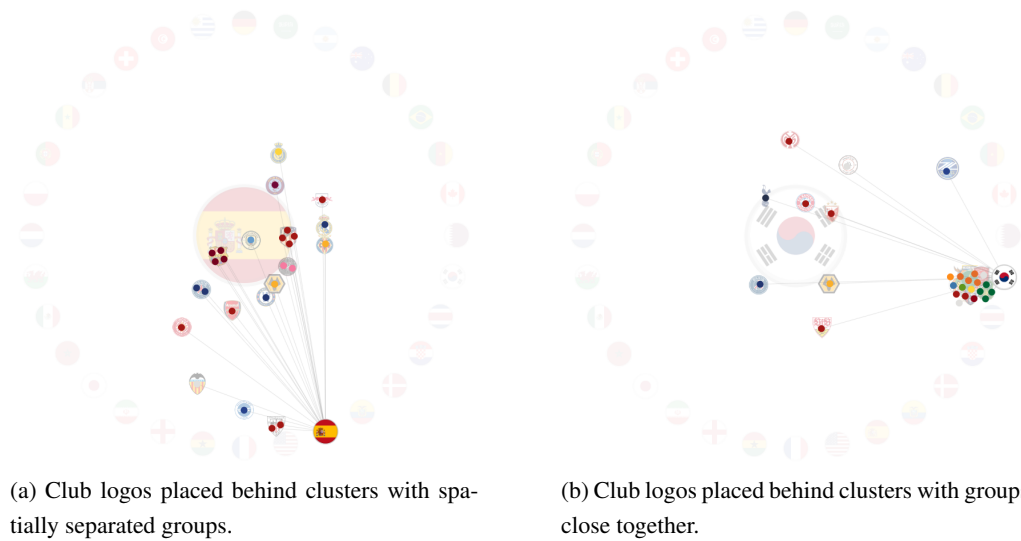


Figure 5.7: Exploration of the club logos for distinguishing player clusters in the visualization.

Pattern. A striped pattern is used to identify nodes representing *No club* players who played for a national team without being affiliated with any official club.

Opacity. Opacity reduces the visual importance of certain elements. This includes lowering the opacity of links compared to nodes, denoting national teams with no affiliations, and distinguishing nodes and links directly connected to a highlighted node by reducing the opacity of unconnected elements.

Outline. Outlines highlight nodes related to a searched player, including the player, their club, and their national team.

Additionally, to improve contextual awareness in the details graph, an image of the national team's flag or the club's logo (if defined) of the node being inspected is included at the center of the graph with lower opacity. This feature, although not initially envisioned in the proposed solution, enhances the design of the graph.

5.2.3.2 Forces and Nodes Placement

The construction and definition of the layout for both the overview and detail graphs utilize D3's *d3-force* mechanism. This mechanism simulates physical forces between nodes to determine their placement in the visualization.

While some forces are applied identically in both graphs, others have varying parameters depending on the graph type. The key forces used to replicate the layout specified in the proposed solution are as follows:

- **charge:** Simulates electrostatic charge between nodes. The strength is negative to ensure repulsion: `d3.forceManyBody().strength(-1)`.
- **collide:** Prevents nodes from overlapping based on their radius value: `d3.forceCollide(d => this.nodeRadius(d) + 2)`.
- **x and y (position):** Push nodes toward a desired position along the respective dimension: `d3.forceX().x(0).strength(value)` and `d3.forceY().y(0).strength(value)`. The strength value is 0.01 for the overview graph and 0.3 for the detail graph.
- **link:** Draws linked nodes together based on their strength value: `d3.forceLink().id(d => d.id).strength(value)`. The value is `d => d.value * 0.1` in the overview graph to reflect the number of players the link represents. In the detail graph, the value is 0.0 for detailing a national team or 0.2 for detailing a club to create a clearer visual separation between player clusters in both contexts.
- **outerXY:** Custom force to place outer nodes on the circumference: `this.outerXY.bind(this)`. This force moves nodes linearly to their positions if they were not previously on the outer circle or along the circumference to their new position. The expected positions of the national teams in the outer circle are precalculated based on the number of nodes, sorting criteria, and circle radius.
- **withinCircleBounds:** Custom force to keep inner nodes within the circumference: `this.withinCircleBounds.bind(this)`. It cancels drag movements that attempt to move nodes outside the circle and prevents other forces from placing nodes outside the circle during layout adjustments.

The strength values differ between the two graphs to achieve a balanced layout. Specifically, in the detail graph, it is crucial to maintain the relative placement of a player to their club, as seen in the overview graph. This involves updating the x and y forces to match to the club's coordinates whenever the club's position in the overview graph changes significantly (due to drag movements or timeline animations). This ensures that the spatial relationship between the player and their club is preserved in the club and national team details.

The nodes and links of the forces are updated based on the data of the current timeslice, and the simulation is restarted to adjust the nodes' positions accordingly. Fig. 5.8 illustrates the resulting layouts generated by these forces across the three distinct scenarios.

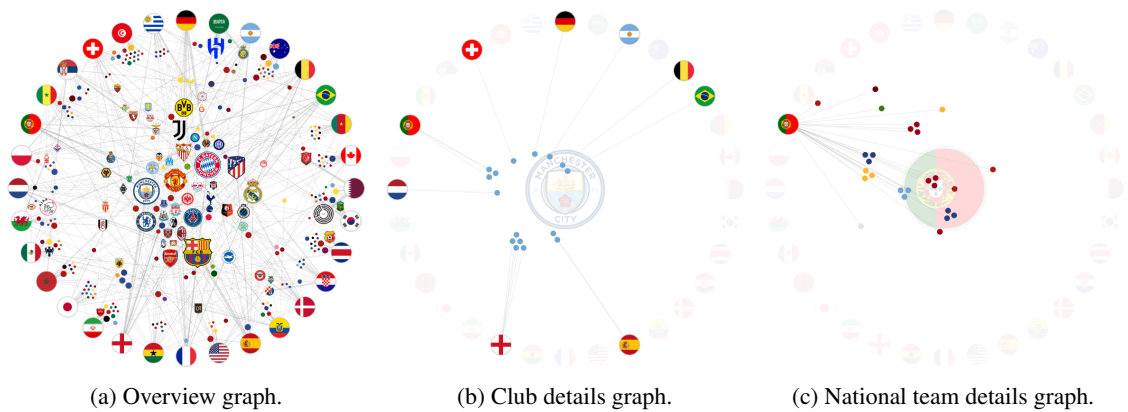


Figure 5.8: Resulting layouts from the applied forces across the visualization graphs.

5.2.3.3 Interaction Mechanisms

The interaction mechanisms for the overview and detail graphs align with the proposed solution specifications. They involve node manipulation through three main mechanisms: hover, click, and drag. The detailed descriptions and device-specific adaptations are outlined below.

In the **overview graph**, when a club or national team node is hovered over, its directly connected nodes and links are highlighted to emphasize connections and enhance clarity. Additionally, an info card appears in the bottom right corner, displaying the node's descriptive name and the number of players and relationships involved. For instance, highlighting a club shows the number of players it contributes and the number of national teams it supports (e.g., *Club X: n players -> m national teams*), while highlighting a national team shows the number of players it receives and the number of contributing clubs (e.g., *National Team Y: n players <- m clubs*). Clicking on a club or national team node opens a detailed view of that node, allowing users to explore the specific players associated with it.

In the **detail graph**, hovering over a national team or player node also highlights its directly connected nodes and links. The info card in this context displays the following information: for national teams, the number of players connected to them, and for players, the logo and name of the club they belong to. Clicking on a national team node switches the detail view to that national team, while clicking on a player node opens the player's page on the zerozero website, providing direct access to more detailed player information.

For both graphs, the label of simple colored circle nodes (those without images) in the highlighted network is also displayed. This involves showing the descriptive names of the nodes. To prevent label overlap and ensure readability, a force is added in this state using the *d3-bboxCollide* plugin³, which handles rectangular collisions, as D3's default forces only consider circular collisions. Adding labels above images was considered, but it only created more visual clutter.

³<https://github.com/emeeks/d3-bboxCollide>

All inner nodes in both graphs can be dragged. Due to implemented forces, nodes cannot be dragged outside the circle and will stabilize back into position when dragging stops, maintaining layout coherence and boundary constraints. Furthermore, dragging a club node in the overview graph updates the position of its associated players in the detailed graph to maintain spatial consistency. Outer nodes cannot be dragged.

Device-specific adjustments ensure smooth interaction across various device formats. On laptops, hovers are triggered by mouse events (*mouseenter* and *mouseleave*), and detailed inspections are triggered by click events. On touch devices, a double-tap mechanic is used: the first tap highlights the node (hover), and a second consecutive tap on the same node opens the detailed view (click). This approach was essential for enhancing the overall usability of the visualization, given the diverse target audience. No adjustments were needed for the drag interaction.

However, the logic implemented for distinguishing device types has some flaws. In rare instances, touch devices are misidentified as laptops, leading to incorrect consideration of the interaction type and usability issues. Despite testing, no clear pattern regarding device characteristics was found to explain these errors. This disclaimer acknowledges these issues, but their rarity suggests they do not significantly impact the results analysis.

5.2.4 Timeline Integration

With the layout and interactions defined for a static visualization graph, the next step was implementing its evolution over time. The timeline development draws inspiration from the scrubber used in the Temporal Force-Directed Graph example⁴ but simplified to meet the proposed solution's requirements: previous, next, jumping to specific points in time, and controlling animation playback. The visualization seamlessly updates its displayed data across timeslices, leveraging predefined forces (Section 5.2.3.2) and animations (subsequent subsection) to rearrange the layout dynamically. The timeline design adopts a style similar to that of a music player (Fig. 5.9), making the controls and usage more intuitive and recognizable to most ZOS users.

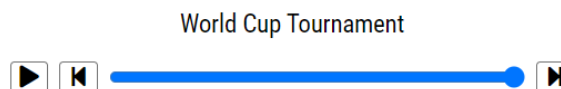


Figure 5.9: Timeline design and organization resembling a music player.

5.2.4.1 Animation between Timeslices

Nodes within a new timeslice can either exist from previous slices, be newly introduced, or be removed altogether. As discussed in previous D3 experiences (Section 5.2.2.2), D3 automatically identifies and manages these scenarios based on their configured settings.

In prototype implementation, animations for each scenario depend on the node type involved. When new **club nodes** appear, they start at coordinates (0,0) and stabilize into their final positions

⁴<https://observablehq.com/@d3/temporal-force-directed-graph>

due to applied forces. A scaling transition smoothly grows the node from size zero to its intended size. Opacity is set to visible from the start, requiring no additional opacity transition. Similarly, nodes in both timeslices undergo positional updates through force stabilization, opacity transitions, and size adjustments. When transitioning between circle and image to represent the club node, it was essential to manually ensure that the starting size of the element matched the old size of the previous element, as D3 does not handle this automatically. Nodes absent in the subsequent timeslice gradually transition their opacity to zero. Correspondingly, changes in stroke width are animated for links across all three scenarios. Balancing the duration of these transitions was crucial to maintain visual clarity and ensure smooth, perceptible animations.

In the details graph, a similar approach is employed to the **player nodes** with some adjustments. Instead of resizing nodes, as there is no distinction in size between timeslices, color transitions are used to emphasize player career trajectories and club changes. Furthermore, The positions of player nodes begin from the last coordinates of their corresponding club cluster if known; otherwise, they start from the average position derived from the existing club clusters at that timeslice.

Opacity transitions for **national team nodes** align with those of other nodes. However, positioning transitions for national team nodes are managed by a custom force, relying on fixed positions and manually coded animations between old and new node positions.

5.2.5 Additional Controls

All controls added to enhance the visualization are dynamically generated based on available data fields and values. These controls allow users to select and interact with specific data subsets within the visualization. As users adjust these controls, the graphs update dynamically to display the selected data, using predefined forces and transitions to rearrange the visualization accordingly. The following paragraphs provide detailed descriptions of each type of control.

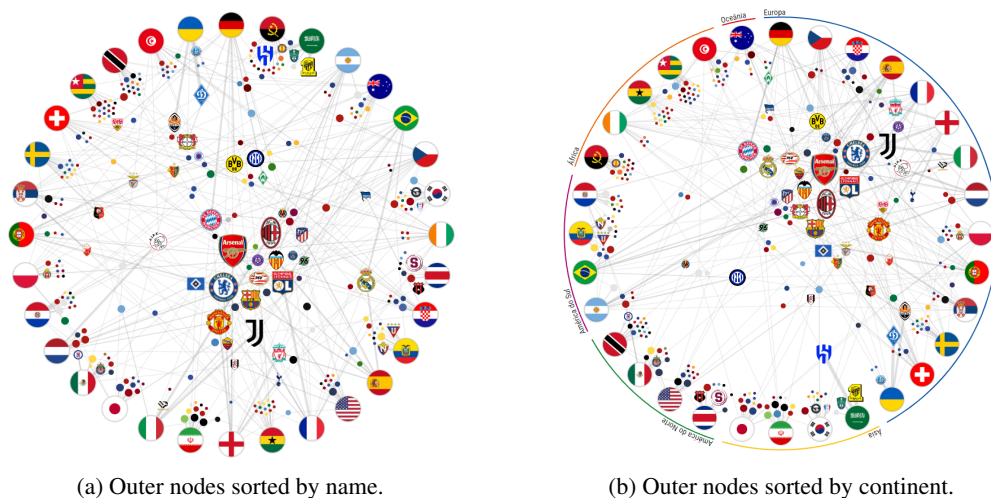


Figure 5.10: Overview graph for the Germany 2006 World Cup sorted by distinct criteria.

Sort. This feature enables sorting the visualization’s outer nodes (national teams) based on specific field values. Valid fields include any except *id*, *img*, or *teams*. When sorting by a field other than *name*, visual arcs are added to the visualization to visually group nodes with the same field value (Fig. 5.10). Each group’s color is defined in a configuration file, with continents following the same color logic as the world map. Nodes within each group are alphabetically ordered by default.

Filter. This function allows users to filter nodes in the visualization (national teams, clubs, or players). Filtering options for each node type are determined based on extracted fields and values from the data, considering a set of predefined fields to dismiss.

Additionally, data can be filtered directly within the visualization using URL query parameters to specify specific club IDs (*cluster* parameter) or player IDs (*detail* parameter) to include. This restricts the fields and values available in the controls of the filtered visualization, as the input data also changes.

Search. This tool enables users to search for specific players within the visualization. The list of possible player names is derived from the data. Below the search control, the list of World Cups in which the player participated is displayed, and the associated nodes are revealed (outlined) within the visualization for ease of identification (Fig. 5.11).

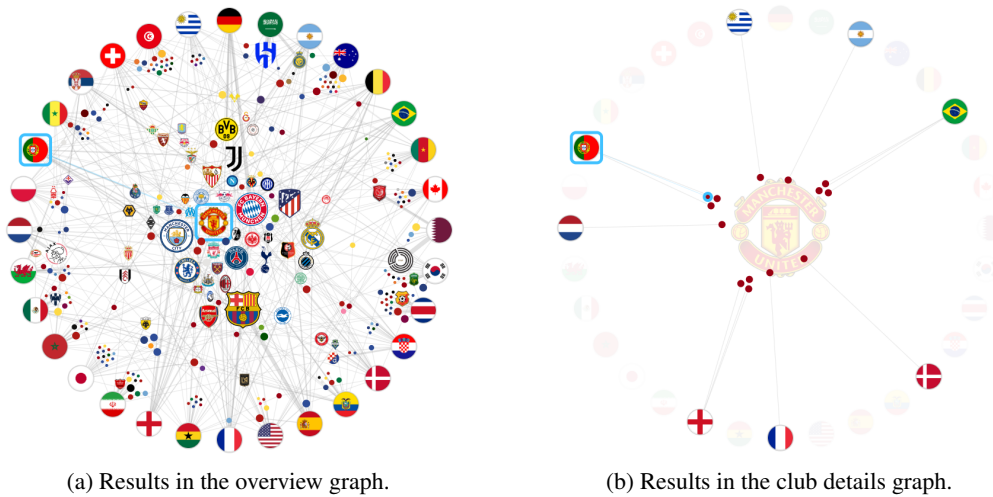


Figure 5.11: Nodes outlined in the visualization for Qatar 2022 when searching for *Cristiano Ronaldo*.

5.2.6 Components Organization

To accommodate the various components of the visualization (overview graph, details graph, timeline, and controls), the layout is organized into two columns: one occupying 1/3 and the other 2/3 of the available width or a defined maximum width, whichever is smaller. The first column at the bottom houses the visualization controls, while the top provides space for the secondary graph.

The second column features the main visualization graph, followed by the timeline. A toggle button in the upper left corner of this column allows users to switch between the focused graph, facilitating exploration and detailed inspection of the larger graph while maintaining an overview and storytelling aspect with the smaller one.

Adjustments were made for smaller screens: all columns now occupy 100% of the available width, and the space for the secondary graph is behind the main graph area. The toggle button allows users to switch between graphs.

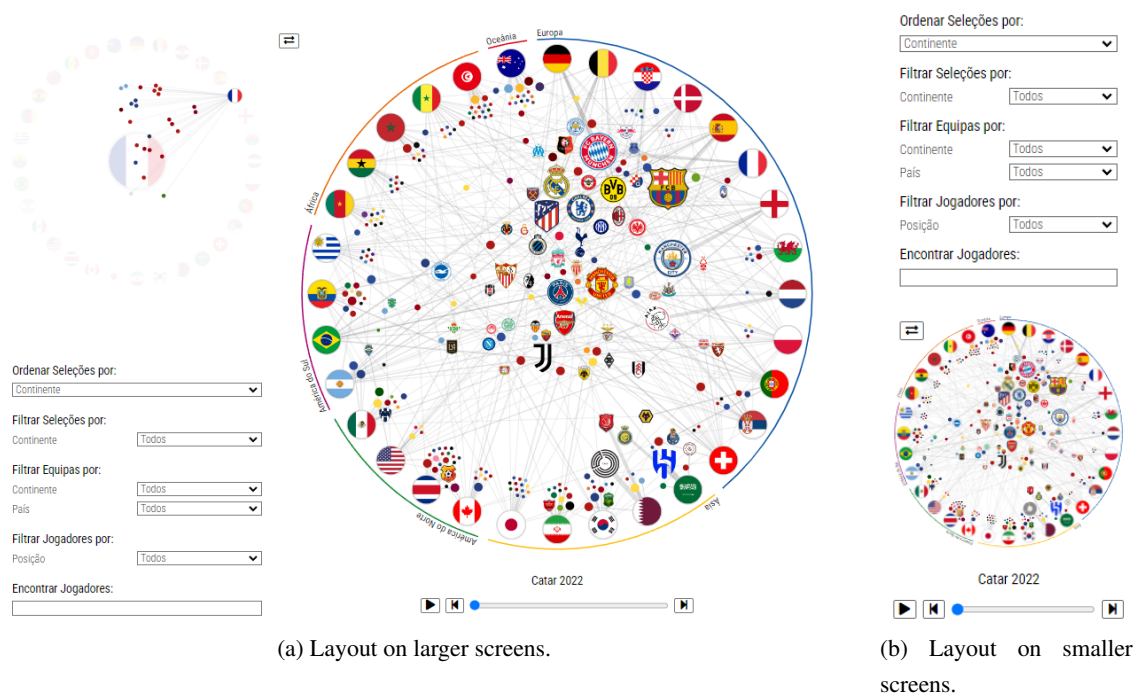


Figure 5.12: Organization of components in the visualization depending on screen size.

All text included in the various components is in Portuguese to align with the context of zerozero's users and is presented in a user-friendly format.

5.2.7 Contextual Elements

To align the visualization with the ZOS aesthetics, contextual elements such as a header (Fig. 5.13) and footer (Fig. 5.14) were incorporated into a standalone page dedicated to the visualization (refer to the link in Annex A). A ZOS designer designed these elements to offer users an explanation of the work's context.

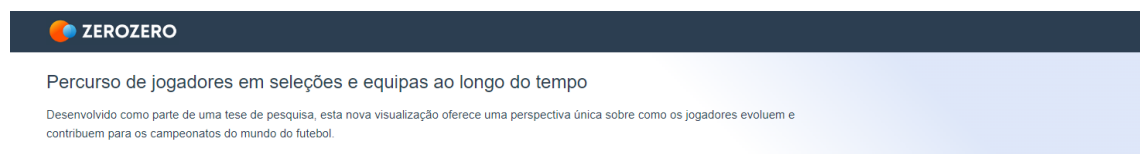


Figure 5.13: Header integrated into the visualization page.

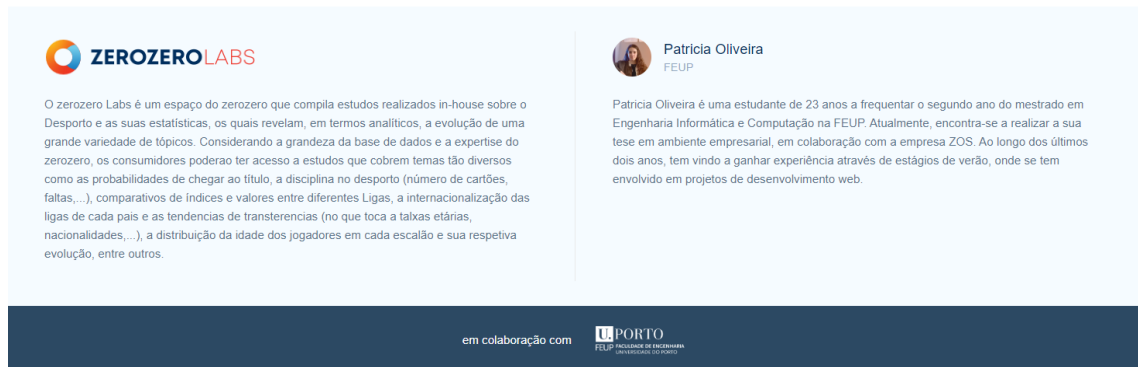


Figure 5.14: Footer integrated into the visualization page.

5.2.8 Documentation and Flexibility

The codebase is documented to support comprehension and maintenance by ZOS developers. Each component and complex functionality within the visualization is explained, facilitating efficient updates and modifications in the future.

Moreover, the visualization uses PHP "template" logic to enhance ease of use and adaptability across various use cases. By including the visualization PHP file with specific configuration parameters, it can be seamlessly integrated or embedded into different contexts without extensive modification, taking full advantage of PHP's flexible templating system.

During development, compatibility across distinct web browsers has been tested. The visualization is confirmed to work seamlessly on popular browsers such as Google Chrome, Opera, Firefox, Microsoft Edge, and Safari.

5.3 Summary

This chapter introduces the proposed solution tailored to address the challenges outlined in Chapter 4. In Section 5.1, the solution is meticulously detailed, encompassing its organization, features, and interactive aspects of the visualization. This section supports decision-making by drawing on insights from the outcomes of the literature review and state-of-the-art analysis. Section 5.2 delves into the implementation details of the prototype, discussing technology selection, visual variables usage, mechanisms for node placement, interaction, and animation challenges, controls employed, as well as visualization organization and page design.

Chapter 6

Experimental Protocol

This chapter provides a comprehensive overview of the experimental protocol designed to assess the accuracy, understandability, usability, and engagement of the multiview-weighted graph-based visualization for transmitting multivariate, highly interconnected data with temporal evolution. Section 6.1 elucidates the methodology employed in the experiments, while Section 6.2 delves into the assessment methods utilized. The pilot tests conducted and their implications on both the procedure and the prototype are outlined in Section 6.3. Finally, Section 6.4 consolidates the critical points of the developed experimental procedure.

6.1 Test Methodology

The main goal of this methodology is to test the understandability, intuitiveness, usability, and engagement of the visualization to test the hypothesis presented in Section 1.3: that dynamic graph-based visualizations effectively present highly connected and time-evolving data. In this context, assessing whether users understand, feel engaged, and can efficiently and effectively interact with the prototype implementing the visualization supporting the hypothesis is crucial.

6.1.1 General Description

The experiments aim to address questions such as:

Question 1: Can users clearly and unambiguously understand what is being displayed?

Question 2: Can users easily interact with the visualization and draw more insights from it?

Question 3: Do users feel engaged and spend time exploring the visualization?

Having detailed answers to these questions can also help to refine the proposed solution to meet the capabilities and expectations of the readers.

To ensure meaningful results, the prototype is validated among ZOS website users, who are the primary audience for the visualization. This approach ensures that participants with diverse data

literacy levels will be considered, being a faithful representation of the online world, particularly sports fans. With this emphasis, the following points are specified:

- **Where and When:** the prototype developed is accessible online for users to conduct their experiments from anywhere. The experiment spans approximately two weeks, from April 19th to May 5th.
- **Equipment and Software:** only a device with internet access and a web browser (computer, tablet, or mobile device) is required.
- **Coordinator, observer, and facilitator:** all experiments are conducted individually, without direct control or supervision. While this approach entails a degree of risk, it facilitates the contribution of a more significant number of participants. It resembles the discovery of a new visualization on a website without external assistance. Testing without auxiliaries ensures an authentic assessment of the visualization's intuitiveness and usability.
- **Participants:** the experiment welcomes as many participants as possible without restrictions, with participation being entirely voluntary. Strategies for reaching participants are explained in Section 6.1.2.
- **Assessment Methods:** the primary method involves formative assessment through questionnaires, covering understanding, accuracy, usability, and satisfaction of the visualization. Additionally, internal metrics are collected to evaluate engagement and interest. These two methods are detailed in Section 6.2.
- **Help Resources:** no external guidance or auxiliaries are provided during the procedure. Participants do not have access to an instruction manual, as it would contradict the objective of measuring ease of use and intuitiveness. The visualization should be self-explanatory.

It's worth noting that although the experiments extend over a long period, the prototype will remain unchanged during this time. This choice ensures consistency, with all participants interacting with the same version throughout the experiment. Any potential enhancements resulting from the experiment's findings will only be implemented after the conclusion of the experimental period.

6.1.2 Gather Participants

Ensuring a diverse and representative sample of participants is crucial for obtaining comprehensive insights into the usability and effectiveness of the visualization. To achieve this goal, three distinct strategies were employed, each tailored to engage different segments of the target audience, thereby reflecting the diverse characteristics of ZOS users regarding backgrounds, levels of expertise, and interests.

Firstly, direct communication was established with company employees through messages within the company's Slack workspace. Secondly, the core audience segment of the ZOS community, consisting of devoted soccer enthusiasts and regular visitors to the platform, was engaged through a dedicated publication about the visualization on the company's website¹ and LinkedIn². Lastly, a separate outreach effort targeted friends to involve individuals less knowledgeable or interested in the soccer world.

6.1.3 Experiment Procedure

As mentioned in the previous section, the experiment is voluntary and individual, without direct interaction between the experiment orchestrator and the participants to maintain the authenticity of the user experience. The procedure that each participant has to follow is outlined as follows:

1. Participants access a dedicated page within ZOS, where they are presented with the visualization in its initial state, reflecting the prototype's state upon webpage reload. This state includes the complete set of teams, players, and national teams from the latest World Cup, Catar 2022.
2. Alongside the visualization, participants were provided with a voluntary questionnaire to collect insights about the visualization's comprehensibility and usability. The questionnaire included sections covering consent terms, demographics, visualization questions, and the System Usability Scale (SUS).
3. If participants chose to respond to the questionnaire, they were required to consult the page with the visualization to provide their responses. They were free to consult the visualization at any point and were responsible for submitting the questionnaire upon completion.
4. The system automatically captures interaction analytics upon access to the page.

To encourage participation, the questionnaire (see Annex B) was designed to be concise and self-explanatory, containing all the necessary information to complete the experiment. Further details regarding the questionnaire's structure can be found in Section 6.2.1. As for interaction analytics, the tools utilized, specific events captured, and any encountered limitations are elaborated on in Section 6.2.2.

6.1.3.1 System Load

The prototype relies primarily on client-side processing, making the system's performance mainly dependent on the capabilities of the specific device used in the experiment. This approach minimizes the impact of participant numbers on system load. Nonetheless, occasional delays may

¹<https://www.zerozero.pt/noticias/clubs-in-nations-temos-uma-ferramenta-revolucionaria-/618832>

²https://www.linkedin.com/posts/zerozero-pt_zerozero-zerozerolabs-jornalismo-activity-7192203201989722113-miFX

occur due to high traffic on other ZOS services, particularly when fetching external resources such as images of national team flags and club logos from zerozero's website.

On another note, the displayed data is periodically updated under normal circumstances using an available endpoint that obtains the data from the ZOS database. However, to ensure consistency and prevent potential changes over time that could impact task accuracy, the data is fetched from a static file throughout the experiment period, guaranteeing standardized test conditions.

6.1.3.2 Completion Criteria

An experiment is complete once the participant acknowledges and accepts the Informed Consent Form, fills all mandatory fields in the questionnaire's sections (pre-test, in-test, and post-test), and submits it. However, comprehensive internal metrics are gathered throughout the experiment period. Therefore, the whole experiment is considered complete only at the end of this period, including all submissions and analytics collected during the experiment for results analysis.

6.1.3.3 Success Criteria

Regarding the experiment questionnaire, each task is binary - either correct or incorrect - with a score of 1 for correct answers and 0 for incorrect ones. Correct answers align with those previously determined by test conductors during preparation, discriminated in Section 6.2.1.3. Any response diverging from the predetermined correct answer, including empty responses, is considered incorrect. Given the inability to measure the completion time per task, success criteria rely solely on response accuracy - whether it matches the correct answer.

6.2 Assessment Methods

This evaluation encompasses measuring visualization accuracy, overall usability, and user experience through two methods: a formative questionnaire and internal analytics, both involving direct interaction with the visualization.

The questionnaire is designed using Google Forms [31], while Google Analytics [35] and Hotjar [13] are employed to gather analytics data, providing insights into user behavior and preferences. To facilitate correlating data between these methods, a custom session ID is used to link the records across the tools. Detailed explanations of specific configurations, challenges, and decisions are provided in the subsequent sections.

6.2.1 Formative Questionnaire

The primary method utilized to assess the effectiveness of the visualization involves gathering participant feedback through a formative questionnaire administered alongside the visualization. This questionnaire is a vital tool for perceiving participants' perspectives on accuracy, understanding, usability, and overall satisfaction.

When selecting the platform in which to develop the questionnaire, various options were considered, ultimately choosing Google Forms [31] for its user-friendly interface and widespread recognition. Despite exploring alternatives like Maze [42], Google Forms stood out for its simplicity and accessibility, aligning well with the diverse backgrounds of the target audience. Furthermore, given the context and background of ZOS, Portuguese is selected as the language of the questionnaire to ensure greater comprehension and engagement among participants.

However, it's essential to acknowledge the inherent risks associated with making the questionnaire widely available. The potential for biases and anomalies in the collected data underscores the need for a cautious approach during analysis. To mitigate this risk, the submissions are limited to one per email, utilizing Google Forms' capabilities.

The questionnaire (see Annex B) begins with an introductory section outlining its purpose, context, and terms (Informed Consent Form, Section 6.2.1.1) and proceeds to solicit demographic information such as age and experience (Pre-test Questionnaire, Section 6.2.1.2). Participants are then presented with tasks related to the visualization, aimed at evaluating user interaction and comprehension (Task-specific Questionnaire, Section 6.2.1.3). Additionally, participants are asked to rate the system's usability and provide open comments for further insights (Post-test Questionnaire, Section 6.2.1.4). The subsequent sections delve deeper into the specific structure of the questionnaire designed.

6.2.1.1 Informed Consent Form

In this section of the questionnaire, participants are provided with an introduction detailing the goal and purpose of the experiment. They are informed that participation is entirely voluntary, allowing them to withdraw at any point during the study. Participants are assured that the data collected will be used solely for research purposes.

Before moving forward, participants are asked to agree to the following terms to consent to participate and provide the necessary information:

- "Do you agree to provide demographic information and details about your experience in interpreting visualizations for the purpose of this experiment?"
- "Are you willing to participate in the experiment, understanding that your responses and interactions with the visualization will be recorded and analyzed for research purposes?"
- "Do you agree to share your subjective evaluation of the visualization's usability using the System Usability Scale (SUS), acknowledging that your feedback will be used for research and improvement purposes?"

6.2.1.2 Pre-test Questionnaire

The pre-test questionnaire collects essential participant information, encompassing demographics, domain-specific experience, and experimental conditions. All data gathered ensures participant

anonymity and remains exclusively utilized for statistical analysis, safeguarding against sharing or commercial use.

Demographic Information. Participants indicate their gender and age, offering valuable insights into the composition and characterization of the participant pool.

Experience Levels. Participants rate their familiarity with two key domains, information visualization and sports, namely soccer, on a 7-point Likert Scale, enabling the exploration of potential correlations between prior domain knowledge and experimental outcomes.

Experiment Conditions. Several system-generated fields and variables serve to enrich the comprehension and delineation of the experimental environment of each participant, compensating for the absence of an administrator or observer during the experiment:

- **Start Date and Time:** this records the precise moment the participant initiates the questionnaire. Utilizing this metric enables the estimation of response duration when correlated with submission timestamps.
- **Interaction Type:** this denotes the mode of interaction the system employs during the experiment, such as mouse events or double click, facilitating comparisons across user clusters. This data illuminates potential discrepancies in usability between interaction modes, providing insights for tailored improvements or adaptations.
- **Session ID:** this unique identifier links participant responses to their respective analytics, streamlining the examination of individual submissions if behavioral scrutiny becomes necessary.

While attempts were made to conceal these fields, practical solutions proved unfeasible. Extensions for Google Forms with free plans, such as Customize UI Addon [40], FormPrefill Addon [41], and Pretty Forms Designer Addon [15], had submission limits (e.g., a maximum of 10 submissions per month) that rendered them unsuitable. Custom scripts in Apps Script [33] for Google Forms were also not viable options. As a result, participants are directed to the questionnaire via a pre-filled link from the visualization page containing the start date and time, interaction type, and session identifier. Though these fields are visible, participants are instructed not to alter the pre-filled information.

6.2.1.3 Task-specific Questionnaire

The task-specific questionnaire evaluates participants' ability to independently interact with the visualization to accurately answer a series of questions. This approach assesses whether participants can comprehend the displayed information and use various visualization features effectively. Each question represents a distinct task, requiring specific actions and outputs. The tasks vary in

granularity and complexity, guiding users from an overview to a detailed inspection of the visualization concerning the evolution of players in national teams and clubs throughout soccer's world cups.

The questions are designed as direct queries focused on the "what" rather than the "how", with dropdown menus offering multiple response options, including the correct answer. This structure minimizes bias and response fluctuations, accommodating a diverse audience, simplifying the process, and reducing complexity. Additionally, avoiding indirect questions helps streamline analysis, mainly when dealing with a larger participant pool.

The questions are carefully formulated to evaluate specific aspects of the visualization, ensuring comprehensive coverage of its various features. These aspects include static layout interpretation (of both graphs), understanding of size significance, usage of filters, and recognition of temporal differences. The developed questions (translated to English), along with the aspects they intend to test and their expected answers, are as follows:

- T1:** In the 2022 Qatar World Cup, which club contributed the most players to the national teams? (Expected answer: *Barcelona*) - Tests the visualization's ability to represent club contributions through node size differences in a specific World Cup.
- T2:** In the 2022 Qatar World Cup, how many national teams did Real Madrid supply with players? (Expected answer: 7) - Tests the visualization's ability to represent the intricate relationships between clubs and national teams in a specific World Cup.
- T3:** In the 2022 Qatar World Cup, from which club did the German national team receive the most players? (Expected answer: *Bayern München*) - Tests the visualization's ability to represent club contributions to a specific national team through edge sizes in a specific World Cup.
- T4:** In the 2022 Qatar World Cup, for which national team did Manuel Akanji play, knowing he was playing for Manchester City? (Expected answer: *Switzerland*) - Tests the visualization's ability to represent and inspect the composition of a club in a specific World Cup.
- T5:** How many different clubs did Cafu play for, considering his participation in the World Cups? (Expected answer: 3) - Tests the visualization's ability to preserve and represent a player's trajectory throughout the World Cups.
- T6:** In the 1994 USA World Cup, how many forwards did the Spanish national team receive from Barcelona? (Expected answer: 3) - Tests the visualization's ability to represent the intricate relationships between players, clubs, and national teams.
- T7:** Which was the first World Cup to feature a national team from Oceania? (Expected answer: *Germany 1974*) - Tests the visualization's ability to navigate the World Cup timeline and represent temporal differences.

- T8:** Considering only the 13 World Cups in which a Portuguese club contributed players, which club was always present, always providing at least one player to a national team? (Expected answer: *Sporting*) - Tests the visualization's ability to represent changes over time through animations while maintaining visual consistency.
- T9:** Analyzing the journey of players in national teams and clubs over time, from the first World Cup (Uruguay 1930) to the last (Qatar 2022), what phenomenon can be observed? (Expected answer: *The globalization of soccer...*) - Tests the visualization's ability to preserve historical narratives and illustrate "migratory" patterns.

The participants' responses provide valuable insights into their understanding of the visualization. Additionally, the ease and comprehension of the visualization can be directly observed through the analysis of session recordings, if necessary, as detailed in Section 6.2.2.2. This observational data complements the questionnaire responses, offering a holistic view of user interaction and understanding.

By answering questions about the composition of national teams and clubs over the years, we can assess whether the graph-based visualization, designed using clusters and multiple views, effectively translates the information and is properly understood by the audience.

6.2.1.4 Post-test Questionnaire

The post-test questionnaire includes the System Usability Scale (SUS) form alongside a section for participants to provide open comments or feedback on the visualization. This combination aims to glean insights into the ease of use, comprehension, and engagement experienced by the participant. The SUS form adheres to a standard ten-item attitude Likert scale, offering a comprehensive overview of subjective usability assessments:

1. I think that I would like to use this system.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in the system were well integrated.
6. I thought there was too much inconsistency in the system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

6.2.2 Internal Analytics

Internal metrics from Google Analytics [35] and Hotjar [13], such as the duration users spend on the visualization, the visitors (both new and returning users) and their details over demographics and technology, and the clicks on prototype elements, are intermittently collected during the extension of the experiment period. This approach aims to infer the visualization's attraction and popularity based on user behavior. Furthermore, these metrics enable a nuanced understanding of user interactions with various controls, highlighting the most frequently utilized functionalities through common events and movement heatmaps provided by Google Analytics and Hotjar, respectively.

As mentioned in Section 6.2, a custom session ID is generated using PHP session capabilities to correlate user interactions with their questionnaire responses. The detailed implementation of how this ID is sent to each analytics tool is explained in their respective sections.

A significant limitation of this approach is users that use extensions that block third-party scripts, such as ad blockers or privacy tools. These extensions can prevent data collection from Google Analytics and Hotjar, leading to incomplete or biased metrics. This affects the accuracy of measuring user engagement and interactions, as the data may not fully represent all user behavior. However, a large and diverse user pool helps mitigate this limitation by providing a broader range of data.

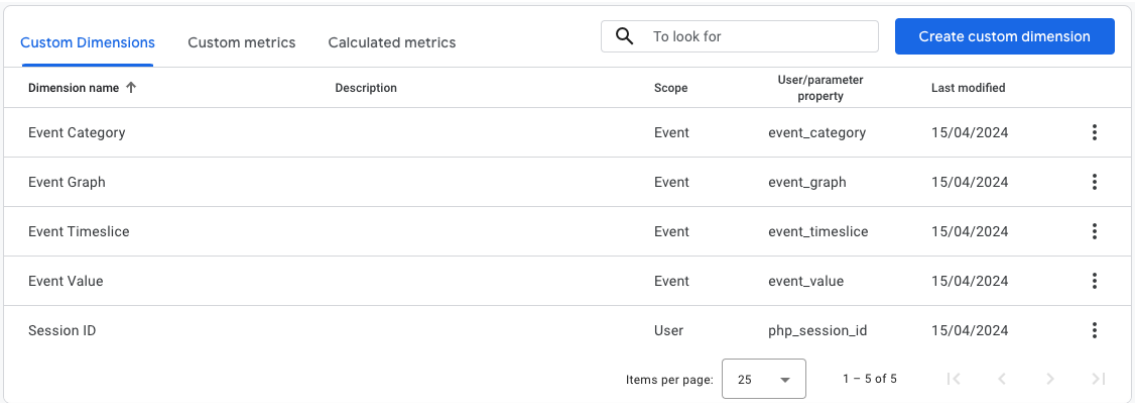
6.2.2.1 Google Analytics

The initial step to configure Google Analytics involves creating a Google Analytics account and a property to aggregate the data to collect from the prototype. To enable analytics gathering with Google Analytics, the Listing 6.1 must be included immediately after the `<head>` element of all pages intended to record analytics. Since the prototype is a single-page application, this listing is only included on one page.

```
1      <script async src="https://www.googletagmanager.com/gtag/js?id=G-XXXXXXXXXX"></  
      script>  
2      <script>  
3          window.dataLayer = window.dataLayer || [];  
4          function gtag(){dataLayer.push(arguments);}  
5          gtag("js", new Date());  
6  
7          gtag("set", { "user_properties": {  
8              "php_session_id": "<?php echo session_id(); ?>"  
9          }});  
10         gtag("config", "G-XXXXXXXXXX");  
11     </script>
```

Listing 6.1: Google Analytics Installation Tag.

The Listing 6.1 includes the essential code to link the user’s session ID with all subsequent records in lines 7-9. This ensures all subsequent events are associated with the `php_session_id` established here. However, a custom user dimension must be previously configured in Google Analytics to interpret this variable correctly. Fig. 6.1 displays all custom dimensions configured during the Google Analytics setup process, including the session ID.



The screenshot shows the 'Custom Dimensions' tab in the Google Analytics interface. It features a search bar, a 'Create custom dimension' button, and a table of existing dimensions. The table has columns for 'Dimension name', 'Description', 'Scope', 'User/parameter property', and 'Last modified'. Five dimensions are listed: Event Category, Event Graph, Event Timeslice, Event Value, and Session ID. The 'Session ID' dimension is highlighted, showing it is of 'User' scope and uses the 'php_session_id' property.

Dimension name ↑	Description	Scope	User/parameter property	Last modified
Event Category		Event	event_category	15/04/2024
Event Graph		Event	event_graph	15/04/2024
Event Timeslice		Event	event_timeslice	15/04/2024
Event Value		Event	event_value	15/04/2024
Session ID		User	php_session_id	15/04/2024

Figure 6.1: Custom Dimensions configured in Google Analytics.

In addition to the standard metrics collected by Google Analytics, such as page views (tracks the number of visits, offering insights into overall interest and traffic), time on page (measures the duration users spend interacting with the visualization, with longer times potentially indicating higher engagement), and bounce rate (analyzes the percentage of users who leave the page without further interaction, potentially indicating usability or relevance issues), specific interactions with the visualization’s features must be tracked:

- **Clicks/Hovers on Nodes:** tracks interactions with the nodes in the visualization, revealing which data points users explore.
- **Timeline Interactions:** monitors how users interact with the timeline to control the displayed year of data, including buttons clicked and range interactions. This provides insights into user interest across different periods.
- **Controls Usage:** measures the usage of filters and other controls to customize the visualization, including which controls are used and the specific values selected or entered. This offers insights into how users manipulate data to meet their needs.

To collect this information, custom dimensions have been integrated at the event level (see Fig. 6.1) to serve as parameters for custom events registered on the prototype side, as detailed in Table 6.1. These events and their parameters are transmitted to the analytics using the event command of the Google tag API [32] whenever a relevant action occurs, as described in the scenarios mentioned earlier.

Table 6.1: Reference of custom events for Google Analytics tracking.

Event	Parameters
<i>change_sort</i> Sent when changing the value of a sort control.	<i>event_category</i> : scope (e.g., "national-teams") <i>event_value</i> : criteria (e.g., "name")
<i>change_filter</i> Sent when changing the value of a filter control.	<i>event_category</i> : scope_criteria (e.g., "teams_country") <i>event_value</i> : value (e.g., "Portugal")
<i>change_search</i> Sent when changing the value of a search control.	<i>event_category</i> : scope (e.g., "players") <i>event_value</i> : query (e.g., "Cafu")
<i>control_timeline</i> Sent when clicking on a timeline element - buttons or range.	<i>event_category</i> : action (e.g., "next") <i>event_value</i> : timeslice (e.g., "EUA 1194")
<i>node_highlight</i> Sent when hovering a node.	<i>event_category</i> : type (e.g., "teams") <i>event_value</i> : (id) name (e.g., "(C-9) FC Porto") <i>event_graph</i> : graph (e.g., "cluster") <i>event_timeslice</i> : timeslice (e.g., "Catar 2022")
<i>node_in_detail</i> Sent when clicking a node.	<i>event_category</i> : type (e.g., "teams") <i>event_value</i> : (id) name (e.g., "(C-9) FC Porto") <i>event_graph</i> : graph (e.g., "cluster") <i>event_timeslice</i> : timeslice (e.g., "Catar 2022")

When configuring Google Analytics, references from Google for Developers [34, 32] and Google Analytics Help [30] were consulted to understand how to send custom events and incorporate custom dimensions.

6.2.2.2 Hotjar

To start collecting analytics from Hotjar [13], it is necessary to add the tracking code included in Listing 6.2 immediately after the <head> element of all pages intended to record analytics, similar to the setup process for Google Analytics.

```

1  <script>
2      (function(h,o,t,j,a,r){
3          h.hj=h.hj||function(){(h.hj.q=h.hj.q||[]).push(arguments)};
4          h._hjSettings={hjid:3930621,hjsv:6};
5          a=o.getElementsByTagName("head")[0];
6          r=o.createElement("script");r.async=1;
7          r.src=t+h._hjSettings.hjid+j+h._hjSettings.hjsv;
8          a.appendChild(r);

```

```

9      }) (window, document, "https://static.hotjar.com/c/hotjar-", ".js?sv=");
10    </script>

```

Listing 6.2: Hotjar Tracking Code.

The analytics and conclusions drawn from Hotjar are limited to the features available in the free version. The paid version would allow integration with Google Analytics and provide more comprehensive reports about trends and other metrics. However, upgrading to a higher-priced plan was deemed unnecessary since Google Analytics already covers most of the essential aspects not included in the free Hotjar plan.

The free version is limited to recording 35 sessions per day, making the quantitative metrics of Google Analytics a better fit for broader data collection. Nevertheless, the value of Hotjar emerges from its heatmaps and session recordings that offer a more visual perspective of user interactions:

- **Heatmaps:** visualizes where users click or interact most within the visualization, identifying popular areas or features.
- **Session Recordings:** observes how users navigate through the visualization, where they encounter difficulties, and how they use various controls. This is particularly important when analyzing the more unusual or exciting user behaviors identified in the questionnaire submissions.

A limitation of opting for the free plan is the inability to send additional data, such as the session ID or other custom data, to Hotjar via the User Attributes feature within the Identify API. This feature is only available on specific paid plans: Observe Business, Observe Scale, Ask Business, and Ask Scale³. As a workaround, the session ID is added directly to the page to be saved within the recording.

6.3 Pilot Tests

A pilot test was conducted on April 15th to evaluate the entire procedure, including the adequacy of the duration, the structure, and readability of the questionnaire, the clarity of instructions and tasks, the functionality of internal metrics mechanisms, and to identify any prototype errors or inconsistencies. This step is particularly crucial due to the expected sample size in the main experiment phase, aiming to ensure the reliability and insightfulness of the results by addressing any potential issues beforehand. This test involved three collaborators from ZOS. The structure of the pilot test is outlined in the following points:

1. The moderator directed participants to the visualization, following a pre-formulated guide.
2. Participants answered the available questionnaire.

³<https://help.hotjar.com/hc/en-us/articles/360038394053-How-to-Set-Up-User-Attributes>

3. The moderator observed the participants, noting common struggles without assisting.
4. At the end, the moderator conducted open-session interviews to gather direct feedback.

Based on the insights gathered from participant behavior and feedback during the pilot tests, the prototype and test procedures (questionnaire and analytics) were refined. The key improvements included:

- **Reordering Form Options:** participants had difficulty finding the response they sought in the dropdown of a question. To address this, numeric options were reordered in ascending order, and text options were reordered alphabetically.
- **Reordering Form Questions:** some questions' difficulty increased because previous questions involved using filters that influenced the displayed data. This was especially problematic when the filter excluded data needed for the next question, and users forgot they had the filters selected (e.g., looking for a player after filtering players by position that does not match that specific player). The questions were reordered to minimize the influence of filters selected in previous questions, allowing each question to be judged more independently.
- **Changing Timeline Direction:** some participants misinterpreted the orientation of the timeline, likely because it ran from the most recent to the oldest years, causing confusion in identifying the first World Cup by always choosing the leftmost side. The direction was adjusted from left (oldest) to right (most recent) to follow the more natural way of ordering timelines, similar to videos. The initial state remained on the most recent timeslice, and the play button was set to start from the beginning when clicked, having the most recent World Cup active, to potentially improve timeline interpretation during the experiment phase.
- **Updating Search Result Layout:** when searching for a specific player by name, participants had difficulty identifying the highlighted elements in the visualization that indicated the nodes where the player is present or even simply finding the relevant timeslices. To fix this, the search results layout was improved by transforming the simple text of the timeslices into links and adding a "hint" to the visual variable used to distinguish those nodes to improve usability (see Fig. 6.2 for a comparison of the layout before and after this change).

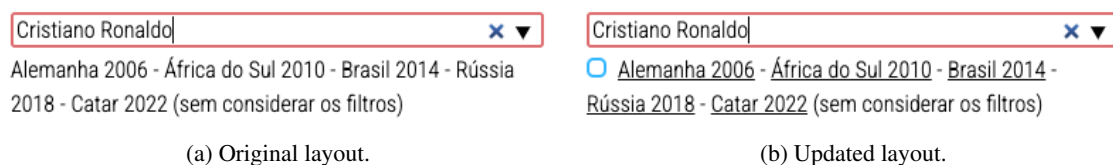


Figure 6.2: Comparison of search result layout improvements from the pilot tests.

Another issue was noted, but due to its complexity and the substantial work required, it was decided that addressing it would need a larger sample size to justify the effort. This issue involved the card with extra information about the node not being perceived by users.

6.4 Summary

The experiment is conducted online over approximately two weeks, allowing users to interact with the visualization remotely. Using formative questionnaires and internal analytics tools like Google Analytics and Hotjar, the procedure evaluates users' comprehension, interaction with the visualization, and overall engagement. Diverse recruitment strategies ensured a balanced participant pool, considering varying data literacy levels and sports familiarity, thereby enhancing the validity of the findings. Additionally, clear completion and success criteria were established as benchmarks for evaluation.

Pilot tests were conducted to refine the prototype and test procedures. Observations and feedback from these tests were crucial in identifying areas for improvement, leading to adjustments that enhanced user interaction, comprehension, and overall experience before the experimental period.

In conclusion, this chapter outlines the experimental protocol designed to evaluate the effectiveness of the multiview-weighted graph-based visualization for presenting complex, interconnected data with temporal evolution. The primary objectives are to assess the visualization's accuracy, understandability, usability, and engagement, validating its ability to convey the data in a visually engaging and understandable manner.

Chapter 7

Results

This chapter presents the results of the experiment detailed in Chapter 6. Section 7.1 covers the results from the formative questionnaire, including outlier identification, demographic statistics, task performance, and usability statistics. Section 7.2 focuses on the results obtained from the analytics tools: Google Analytics and Hotjar. Finally, Section 7.3 interprets and discusses these results.

7.1 Formative Questionnaire

The formative questionnaire is an assessment method to gather insights into the participant population's demographics, enabling a better understanding of the sample's characteristics. It also allows the evaluation of participants' performance on the given tasks and measures their overall satisfaction with the visualization. The following subsections delve into the statistics of these three aspects: demographics (Section 7.1.2), task performance (Section 7.1.3), and usability (Section 7.1.4).

A crucial first step is removing incoherent submissions to ensure the analysis is based on high-quality data and obtain more robust and credible findings. This step is particularly important due to the uncontrolled nature of participation. The process of identifying and discarding inconsistent responses is detailed in Section 7.1.1.

7.1.1 Outliers Identification

As mentioned in Section 6.1.1, the formative questionnaire is remotely accessible, with voluntary and individual participation. However, there is an inherent risk in making this assessment method widely available without direct control over how participants engage with the questionnaire. Several errors can occur: participants might treat the questionnaire as a test of cultural knowledge, neglecting to use the visualization for their responses, or they may provide random answers, thus distorting the results. These potential issues resonate with implementing a methodology for identifying outlier submissions.

Following an initial review of the complete set of submissions, two criteria are established for identifying outliers:

- Submissions that spend less than or equal to a defined threshold of time completing the entire questionnaire and have an error rate (number of wrong answers) equal to or greater than a specified threshold in the task-questionnaire section. The time spent is calculated by subtracting the start time from the end time recorded for each submission.
- Submissions with identical responses to all questions in the SUS section of the questionnaire (e.g., all ones or all fives).

The first criterion is based on the assumption that participants who spend very little time responding and make many errors are likely to be responding randomly. This does not solely consider the errors because the number of options per question varies, and some questions may inherently have a higher chance of success, even with random responses. Additionally, some participants may consistently make errors despite utilizing the visualization, which indicates a genuine misunderstanding. Likewise, time is not the sole variable used, as knowledge in both sports and information visualization domains may justify less time spent with fewer errors. Hence, combining both variables results in a more comprehensive identification strategy. The second criterion assumes that participants who have the same response to all SUS questions didn't take the time to read the statements thoroughly. This is due to the structured nature of the SUS scale, which alternates between positive and negative statements. As a result, their responses may not reflect their true sentiments regarding the usability and satisfaction with the visualization.

The thresholds mentioned in the first criteria are methodologically calculated based on the quartiles of the mentioned variables considering the entire population. Specifically, the 1st quartile for time spent and the 3rd quartile for the task error rate are used, as highlighted in Table 7.1.

Table 7.1: Quartiles of the time spent and the task error rate of the whole population.

	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum
Time Spent (<i>in minutes</i>)	2	4	8	16.25	61
Task Error Rate (<i>frequency</i>)	0	1	4	5	9

Using these thresholds, the Equation 7.1 is applied to each submission to identify outliers:

$$((Time \leq 4) \wedge (Errors \geq 5)) \vee (SUS_1 \equiv SUS_2 \equiv \dots \equiv SUS_{10}) \quad (7.1)$$

If a submission meets any of these two criteria (one or both), it is excluded from the analysis as a whole, independently of the side of the criteria it matches. This ensures that all statistics across various questionnaire sections are based on the same valid population, maintaining the statistical value of potential correlations.

This methodology identifies 30 out of 144 submissions as outliers, resulting in 114 valid submissions for the analysis detailed in the subsequent subsections.

7.1.2 Demographic Statistics

Among the 114 valid submissions identified after the outlier identification process detailed in the previous section, 99 participants identified as male (86.84%), 14 as female (12.28%), and 1 participant chose not to specify their gender (0.88%). Table 7.2 displays the distribution of age and knowledge in the domains of information visualization and sports among these participants.

Table 7.2: Distribution of age and knowledge in information visualization and sports among participants.

	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum	Standard Deviation
Age	18.00	23.00	28.00	41.00	53.00	10.35
InfoViz Knowledge	1.00	4.00	5.00	6.00	7.00	1.55
Sports Knowledge	1.00	4.00	6.00	6.00	7.00	1.46

From the table values and the histogram and bar charts illustrating the distribution of these variables (Annex C, Fig. C.1, C.2, and C.3), it's evident that most participants fall within the age range of 23-30, with a majority identifying as having expertise in the domain of sports, which aligns with expectations given the context of this work. However, such extremes are not as pronounced in the information visualization domain, which is an intriguing characteristic of the participant pool. Nevertheless, there is also a noticeable dispersion across the values of these three variables.

Furthermore, concerning the characteristics of their experimental environment, 84 participants (73.68%) engaged with the visualization through the mouse movement mechanism, while 30 (26.32%) interacted using the double-click touch mechanism (refer to Section 5.2.3.3 to a detailed explanation of the types of interaction).

7.1.3 Task Statistics

To evaluate task performance and assess visualization efficacy, success is defined as a participant's answer matching the correct answer determined by the experiment orchestrator, as detailed in Section 6.1.3.3. It is important to note that empty answers are also considered incorrect, and specific instances where this occurs are indicated in the analysis of each task. Table 7.3 presents the task accuracy per question listed in Section 6.2.1.3 for the 114 valid submissions.

Table 7.3: Task Accuracy per Question.

	Correct Answers	Task Accuracy
Task 1	63 / 114	55.26%
Task 2	73 / 114	64.04%
Task 3	102 / 114	89.47%
Task 4	101 / 114	88.60%
Task 5	74 / 114	64.91%
Task 6	81 / 114	71.05%
Task 7	62 / 114	54.39%
Task 8	57 / 114	50.00%
Task 9	104 / 114	91.23%

From the table above, it is evident that at least 50% of the participants answered each question correctly. Notably, for questions 3, 4, and 9, more than 100 participants (approximately 90%) provided correct answers. A more detailed analysis of the results of each question is presented in the following paragraphs.

Question T1. Out of 114 participants, 63 correctly identified *Barcelona*. The remaining answers were distributed as follows: 22 selected *Manchester City*, 8 chose *Real Madrid*, 7 picked *Bayern München*, 5 opted for *Al Hilal*, and 9 participants selected five other distinct options with 1 to 3 participants per option. No responses were left empty. A total of 10 unique values were selected from the 17 possible options (the clubs that contributed ten or more players to the given World Cup) presented in the dropdown for this question. Refer to the bar chart in Annex C (Fig. C.4) for a visual representation of this distribution.

Question T2. Out of 114 participants, 73 participants correctly answered 7. The other responses were distributed among the options as follows: 15 participants answered 13, 14 answered 9, 10 chose 5, and 1 participant selected 3. The option 0 was not selected by any participant, and there was 1 empty response. Refer to the bar chart in Annex C (Fig. C.5) for a visual representation of this distribution.

Question T3. Out of 114 participants, 102 participants correctly answered *Bayern München*. The remaining answers were distributed as follows: 5 participants chose *Barcelona*, 2 picked *Borussia Dortmund*, 2 selected *RB Leipzig*, and *Manchester City* and *Real Madrid* were each chosen by 1 participant. There was also 1 empty response. Excluding the empty response, 6 unique values were selected from the 13 possible choices (the clubs that contributed with at least one player to the German national team) in the dropdown for this question. Refer to the bar chart in Annex C (Fig. C.6) for a visual representation of this distribution.

Question T4. Out of 114 participants, 101 participants correctly answered *Switzerland*. Other responses included 3 for *Belgium*, 2 each for *Argentina*, *Cameroon*, and the *Netherlands*, and 1 each for *Mexico*, *England*, and *Ghana*. There was 1 empty response. Excluding the empty response, only 8 distinct values were chosen from the 32 possible options (all national teams present in that World Cup). Refer to the bar chart in Annex C (Fig. C.7) for a visual representation of this distribution.

Question T5. Out of 114 participants, 74 participants correctly answered 3. Other responses included 18 participants choosing 4, 9 selecting 2, 6 picking 5, and 3 opting for 1. There were 4 empty responses, making it the question with the highest number of empty responses. The option 0 was not chosen by any participant. Refer to the bar chart in Annex C (Fig. C.8) for a visual representation of this distribution.

Question T6. Out of 114 participants, 81 participants correctly answered 3. The other responses were distributed as follows: 14 participants answered 0, 10 chose 5, 5 picked 9, and 2 selected 7. There were 2 empty responses. The option 12 was not chosen by any participant. Refer to the bar chart in Annex C (Fig. C.9) for a visual representation of this distribution.

Question T7. Out of 114 participants, 62 participants correctly answered *Germany 1974*. Other responses included 14 for *Germany 2006*, 6 for *Korea/Japan 2002*, and 5 for *England 1966*. Additionally, 3 participants each chose *France 1998*, *Argentina 1978*, *USA 1994*, *Spain 1982*, and *South Africa 2010*. There were 7 other occurrences of 2 or 1 participant per option and 3 empty responses. Excluding the empty responses, the responses were distributed among 16 of the 22 possible choices (every World Cup ever played). Refer to the bar chart in Annex C (Fig. C.10) for a visual representation of this distribution.

Question T8. Out of 114 participants, 57 participants correctly answered *Sporting*. Other responses included 39 for *Benfica*, 16 for *FC Porto*, and 1 for *Barreirense*. There was also 1 empty response. Only 4 distinct values were selected as answers from the 25 possible options (all Portuguese clubs that ever contributed players to the World Cup). Refer to the bar chart in Annex C (Fig. C.11) for a visual representation of this distribution.

Question T9. Out of 114 participants, 104 participants correctly answered *globalization*. Other responses included 5 for *number of selections decreases*, 3 for *deglobalization*, and 1 for *it is not possible to observe anything*. There was 1 empty response. All 4 possibilities presented were selected by at least one participant. Refer to the bar chart in Annex C (Fig. C.12) for a visual representation of this distribution.

7.1.3.1 Task Performance per Participant

Examining task accuracy per participant among the 114 submissions reveals that 24 participants (21.05%) answered all questions correctly. Meanwhile, 19 participants (16.67%) made only 1 error, and 15 participants (13.16%) made 2 errors. On the other end of the spectrum, 4 participants (3.51%) made 7 errors, none made 8 errors, and 1 participant (0.88%) answered all questions incorrectly. The remaining 43 participants had errors distributed as follows: 11 participants (9.65%) made 3 errors, 18 participants (15.79%) made 4 errors, 14 participants (12.28%) made 5 errors, and 8 participants (7.02%) made 6 errors. Notably, more than 50% of the participants made between 0 and 2 errors, whereas 24% of the participants made more than 4 errors. Refer to the bar chart in Fig. 7.1 for a visual representation of this distribution.

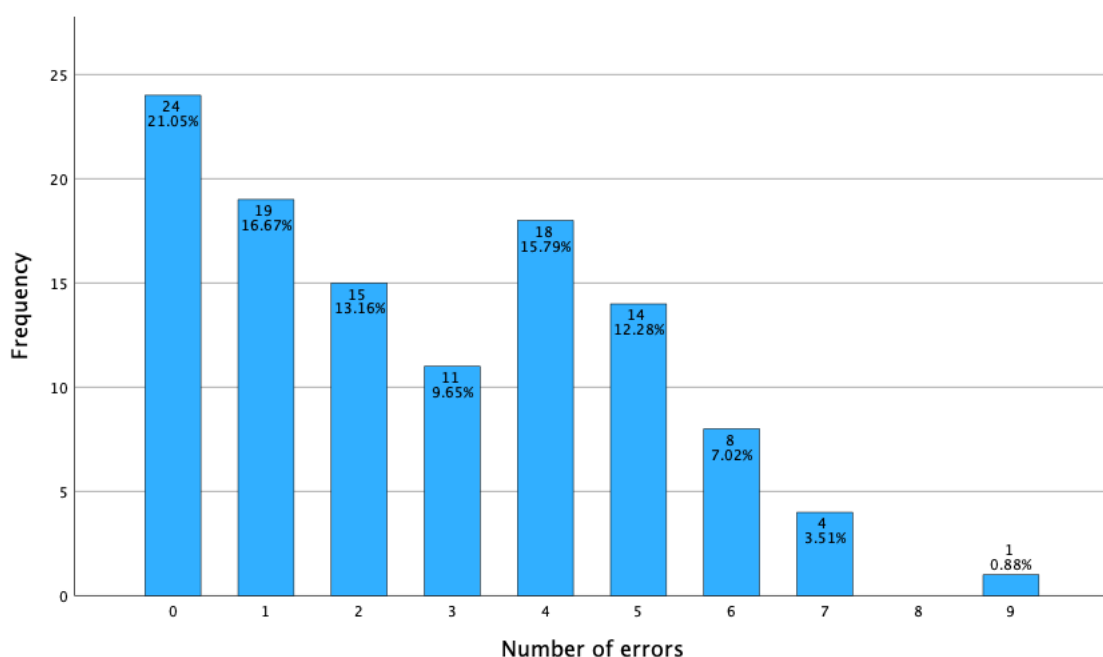


Figure 7.1: Bar chart of the distribution of the number of incorrect answers (errors) to the task-specific section of the questionnaire among participants.

7.1.3.2 Time Spent Influence in Task Performance

The field start time combined with the submission timestamp enables the estimation of the total time each participant takes to answer the entire questionnaire; it does not capture time spent per individual question. This approach allows for an overall assessment of the time spent by each participant and correlates it with the number of incorrect answers (task error rate). Table 7.4 and Fig. 7.2 provide insight into the relationship between the time spent by participants on the questionnaire and the number of errors they made. As a reference, the frequency of participants per number of mistakes is detailed in the preceding section.

Table 7.4: Distribution of time spent per number of incorrect answers among participants.

Errors	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum	Standard Deviation
0	6.00	11.50	14.50	20.50	61.00	11.62
1	7.00	12.50	17.00	19.50	49.00	10.92
2	2.00	17.00	19.00	28.50	48.00	10.60
3	2.00	3.50	4.00	13.00	38.00	11.05
4	2.00	4.00	5.50	9.75	29.00	6.27
5	5.00	5.25	6.50	7.75	13.00	2.22
6	5.00	6.75	8.50	13.25	23.00	5.93
7	5.00	5.00	5.00	7.00	13.00	4.00
8	–	–	–	–	–	–
9	5.00	5.00	5.00	5.00	5.00	–

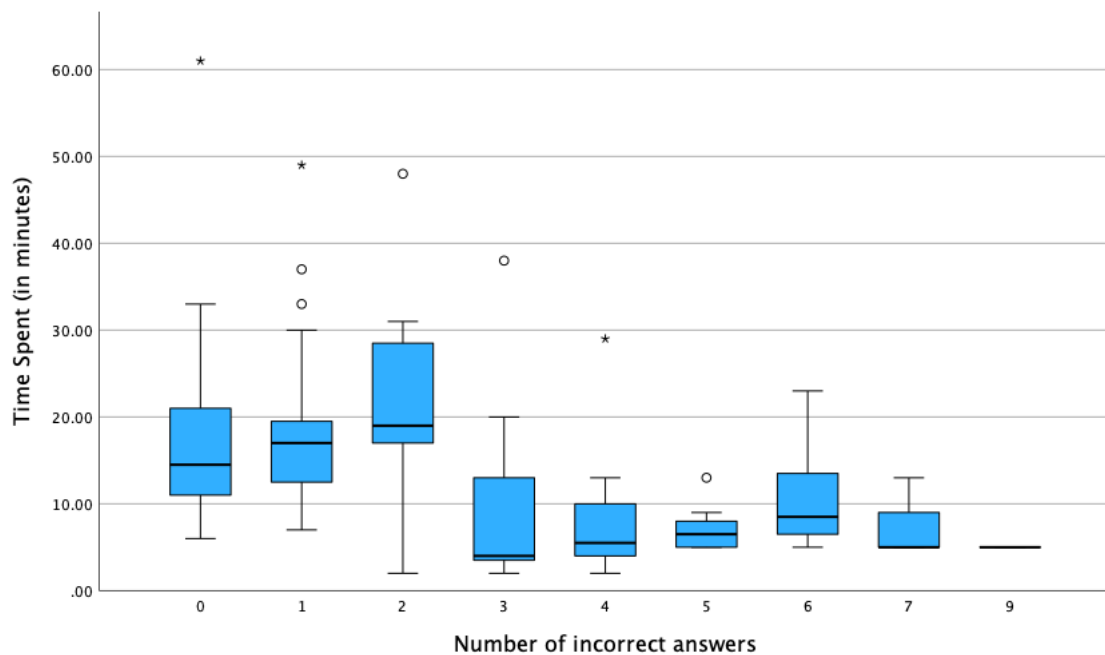


Figure 7.2: Box-plot of time spent per number of incorrect answers among participants.

Several notable trends are evident after inspecting the values of the table and the boxplot:

- **0 Errors:** participants who answered all questions correctly spent a considerable amount of time on the questionnaire, with a median time of 14.5 minutes. The range of time varies widely from 6 to 61 minutes, indicating that those who were most accurate tended to spend more time overall. The high standard deviation (11.62) supports this variability.
- **1-2 Errors:** participants with one error spent a median of 17 minutes, and those with two errors had a median time of 19 minutes. These participants generally took longer than those with perfect scores.

- **3-4 Errors:** participants who made 3 or 4 errors spent less time answering the questionnaire, with medians of 4 and 5.5 minutes, respectively. Notably, the minimum time drops significantly to 2 minutes for these groups, and the standard deviations (11.05 for 3 errors and 6.27 for 4 errors) reflect a wide range of time spent.
- **5-7 Errors:** participants making 5 to 7 errors spent even less time, with medians ranging from 5 to 8.5 minutes. These groups show relatively consistent low times, particularly those making 7 errors with a fixed minimum and median time of 5 minutes. The narrow interquartile ranges and lower standard deviations (2.22 for 5 errors, 5.93 for 6 errors, and 4.00 for 7 errors) support this statement.
- **9 Errors:** only one participant made 9 errors, spending precisely 5 minutes on the questionnaire.

The results indicate a clear trend: participants who spend less time on the questionnaire tend to make more errors, while those who invest more time tend to perform better. This trend is evident as the 1st quartile of time spent by participants with 0 to 2 errors is generally higher or very close (within 2 minutes) to the 3rd quartile of participants with more errors, especially when compared to participants with 7 to 9 errors.

Exceptions and variability suggest that differing levels of knowledge in information visualization and sports may also significantly influence performance. Additionally, the outlier identification strategy, which excluded more submissions with lower times and higher error rates, impacts the minimum time values observed. This strategy likely skews the data, reinforcing the observed trend that more time correlates with genuine and better performance.

7.1.3.3 Domain Knowledge Influence in Task Performance

Each participant characterized themselves based on their experience and knowledge in the fields of information visualization and sports. This data allows for an investigation of the potential correlation between prior knowledge in these domains and participant performance. As a first naive approach, a visual analysis is conducted using a bubble chart and heatmap (Fig. 7.3 and 7.4) to illustrate the frequency and mean error rate, respectively, per participant, depending on their level of expertise.

From the bubble chart and heatmap, it is evident that participants with higher data literacy generally perform better, regardless of their knowledge in the sports domain. However, most perfect scores (0 wrong answers) occur when participants have high levels of expertise in both domains. This conclusion is supported by the large number and size of bubbles with 0 errors (dark blue) in the upper right corner of the bubble chart.

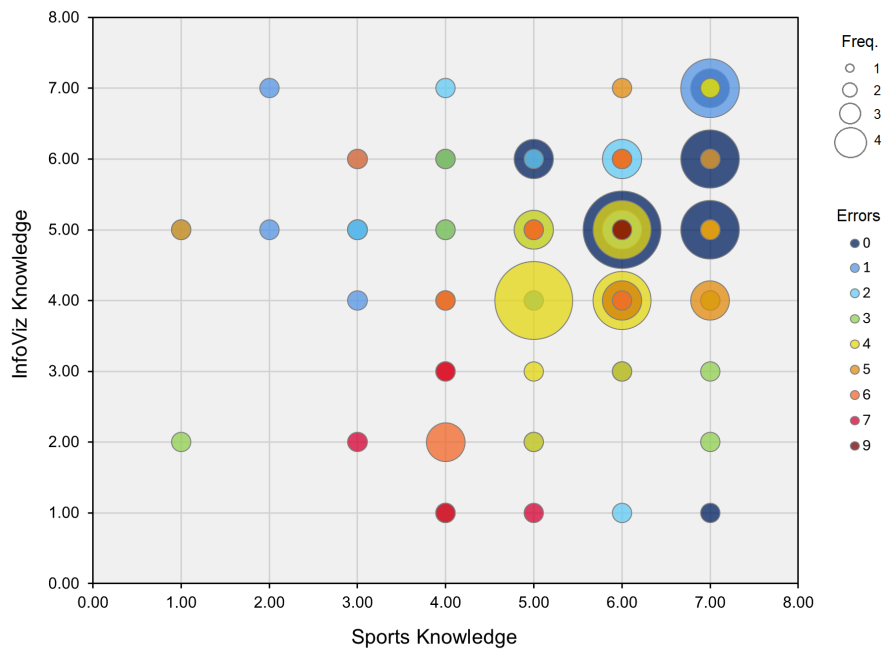


Figure 7.3: Bubble chart of the number of incorrect answers (errors) among participants with the same levels of expertise.

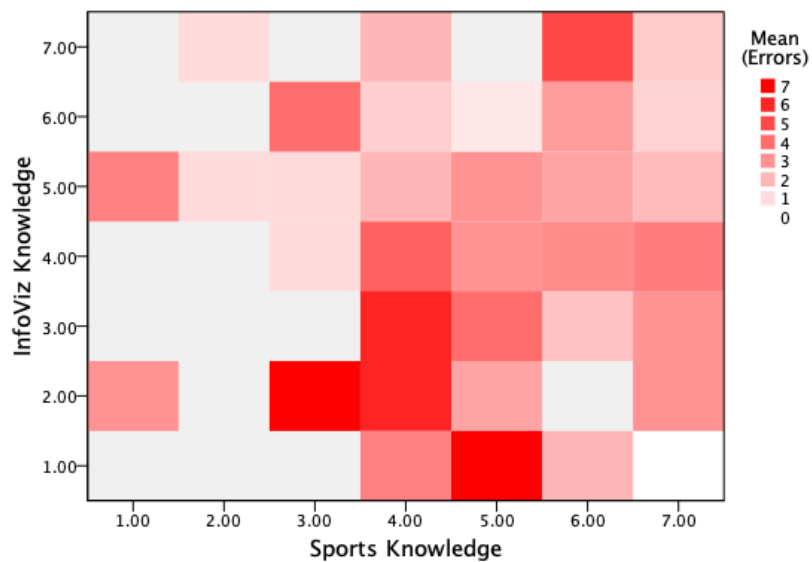


Figure 7.4: Heatmap of the mean of incorrect answers (errors) among participants with the same levels of expertise.

For a more rigorous statistical assessment, considering that domain knowledge variables are ordinal (ranging from 1 to 7) and the number of wrong answers is a count variable (ranging from 0 to 9), the Spearman rank correlation coefficient [10] is calculated. This method measures the

strength and direction of the association between two ranked variables. Table 7.5 presents the correlation coefficients and p-values between each domain knowledge and the number of wrong answers.

Table 7.5: Spearman correlation results for information visualization and sports knowledge.

	Correlation Coefficient	P-value
InfoViz Knowledge	-0.3069	0.0009
Sports Knowledge	-0.1965	0.0361

This method states that the correlation coefficients, ranging from -1 to 1, indicate the strength and direction of the association and that a significant p-value (< 0.05) suggests a meaningful relationship between the variables. The results reveal that sports knowledge and information visualization knowledge negatively correlate with the number of wrong answers, indicating that higher expertise in these domains is associated with fewer errors. However, the correlation appears to be stronger for information visualization knowledge compared to sports knowledge, as indicated by the larger magnitude of the Spearman correlation coefficient and the smaller p-value. These findings align with the conclusions drawn from the initial visual analysis.

Despite this, no clear trend emerges due to significant variation in error rates across different levels of domain knowledge (no extreme coefficients). Consequently, a further analysis combining domain knowledge, performance, and time spent would likely yield inconclusive results and offer little benefit. Additionally, the mean time participants spent responding to the questionnaire, considering their level of expertise in the domains, also exhibits considerable variation (Fig. 7.5).

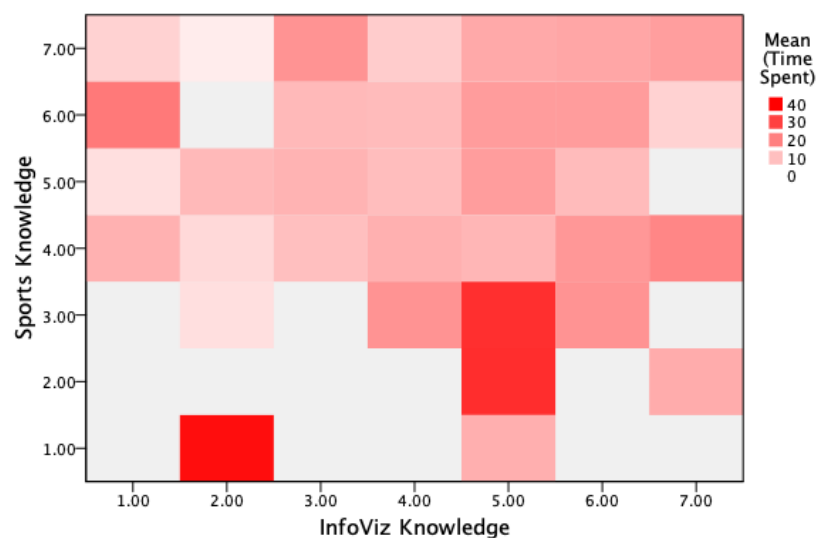


Figure 7.5: Heatmap of the mean of time spent among participants with the same levels of expertise.

In addition to analyzing the influence of previous domain knowledge on individual participant performance, this data allows for the assessment of its impact on individual question accuracy. This helps determine if specific questions are more likely to be answered correctly by participants with a particular level of expertise in either domain. Annex C (Fig. C.13 to C.21) includes bar charts for each task, visually presenting the distribution of correct and incorrect answers across the two domains of expertise.

Based on these charts, no extreme cases are observed where a level of expertise significantly impacted question accuracy, either positively or negatively. However, there are instances (e.g., task 4) where a higher level of sports expertise facilitates answering correctly, likely because participants who are more enthusiastic about sports already knew the answer without needing to refer to the visualization. This is an assumption, as there is no time measure per question. It is also important to note that the distribution of participants by domain expertise skews towards higher levels, as described in the demographic statistics (Section 7.1.2).

7.1.3.4 Interaction Type Influence in Task Performance

The interaction type is related to the testing conditions of each participant. This variable depends on the device used for the test, as the system employs one of two interaction modes with the visualization: mouse movement or touch interaction. This distinction allows for an investigation into the potential impact of these mechanisms, and implicitly the device, on question accuracy. The results can highlight possible issues with the usability and suitability of the visualization depending on the device used.

To compare the question accuracy between the two groups of participants (mouse movement vs. touch interaction) and check if the differences are statistically significant, the Two-proportion Z-test [50] is employed. Taking into consideration the different group sizes (84 vs. 30, as presented in the last paragraph of the Section 7.1.2), this test compares the proportions of correct answers in the two groups, posing two hypotheses:

- H_0 (null hypothesis): The proportion of correct answers is the same in both groups, $p_1 = p_2$.
- H_1 (alternative hypothesis): The proportion of correct answers is different in the two groups, $p_1 \neq p_2$.

Considering the overall accuracy presented in Table 7.3 and the group accuracy (Table 7.6) for each question, the procedure to calculate the result of this test is as follows:

1. **Calculate the Standard Error** (Std. Error): $\sqrt{p(1-p)(\frac{1}{84} + \frac{1}{30})}$, where p is the proportion of correct answers in the pooled population.
2. **Calculate the Z-score**: $\frac{p_1 - p_2}{Std.Error}$, where p_1 and p_2 are the proportions of correct answers in the mouse movement and touch interaction groups, respectively.

3. **Determine the P-value:** $2(1 - \text{normal.dist}(|z\text{-score}|))$, use the standard normal distribution to find the P-value corresponding to the calculated Z-score.
4. **Compare the P-value** to a significance level of 0.05:
 - If the P-value is less than 0.05, reject the null hypothesis, suggesting a significant difference in accuracy rates between the two groups.
 - If the P-value is greater than or equal to 0.05, fail to reject the null hypothesis, suggesting no significant difference between the groups.

Table 7.6 presents the number of correct answers and the accuracy percentage, as well as the results of the Two-proportion Z-test (standard error, Z-score, and p-value) for each question.

Table 7.6: Two-proportion Z-test results comparing task accuracy between Mouse Movement and Touch Interaction groups per question.

	Mouse Movement		Touch Interaction		Two-proportion Z-test		
	Correct	Accuracy	Correct	Accuracy	Std. Error	Z-score	P-value
Task 1	55 / 84	65.48% (+10%)	8 / 30	26.67% (-29%)	0.11	3.67	0.0002
Task 2	57 / 84	67.86% (+4%)	26 / 30	53.33% (-11%)	0.10	1.42	0.1548
Task 3	78 / 84	92.86% (+3%)	24 / 30	80.00% (-9%)	0.07	1.97	0.0489
Task 4	76 / 84	90.48% (+2%)	25 / 30	83.33% (-5%)	0.07	1.06	0.2907
Task 5	54 / 84	64.29% (-1%)	20 / 30	66.67% (+2%)	0.10	-0.23	0.8145
Task 6	65 / 84	77.38% (+6%)	16 / 30	53.33% (-18%)	0.10	2.49	0.0127
Task 7	52 / 84	61.90% (+8%)	10 / 30	33.33% (-21%)	0.11	2.70	0.0070
Task 8	50 / 84	59.52% (+10%)	7 / 30	23.33% (-27%)	0.11	3.40	0.0007
Task 9	80 / 84	95.24% (+4%)	24 / 30	80.00% (-11%)	0.06	2.53	0.0113

Based on the table, for most tasks, there is a significant difference in accuracy rates between the mouse movement and touch interaction groups ($p\text{-value} < 0.05$), with the mouse movement group consistently performing better. The exceptions are Tasks 2, 4, and 5, which do not show significant differences ($p\text{-value} > 0.05$). The influence of interaction type on accuracy is particularly evident in Task 1 ($p\text{-value} = 0.0002$) and Task 8 ($p\text{-value} = 0.0007$), where the touch interaction mechanism negatively impacts accuracy. For these tasks, fewer than 35% of the participants in the touch interaction group answered correctly. In contrast, for all tasks, the mouse movement group's accuracy is consistently at or above 60%.

Furthermore, the accuracy hierarchy between tasks remains consistent across both groups, indicating that the relative difficulty of tasks is similar regardless of interaction type.

Beyond evaluating the influence of interaction type on question accuracy, it is also insightful to examine the task performance from the participants' perspective. Table 7.7 presents the descriptive statistics (quartiles and standard deviation) for the number of mistakes per interaction type. Refer

to Annex C (Fig. C.22 and C.23) for the corresponding box plot and bar chart visualizations of this distribution.

Table 7.7: Distribution of the number of incorrect answers per type of interaction among participants.

Type of Interaction	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum	Standard Deviation
Mouse Movement	0	0	2	4	7	2.02
Touch Interaction	0	3	4	5	9	2.17

The results suggest that participants using the mouse movement interaction generally perform better, as indicated by lower values for the 1st quartile, median, and 3rd quartile compared to those using touch interaction. Additionally, the maximum number of errors is higher for the touch interaction group (9) compared to the mouse movement group (7), indicating greater variability in performance within the touch interaction group. However, both groups have similar standard deviations (2.02 for mouse movement and 2.17 for touch interaction), suggesting a comparable spread of errors around the mean in both groups.

To account for the different group sizes and the non-normal distribution of the data, the Mann-Whitney U test [45] was employed. This non-parametric test is appropriate for comparing two independent groups when the data are ordinal or continuous but not normally distributed. The hypotheses tested were:

- H_0 (null hypothesis): There is no difference in the number of incorrect answers between participants using mouse movement and those using touch interaction.
- H_1 (alternative hypothesis): There is a difference in the number of incorrect answers between the two groups.

The results of the Mann-Whitney U test yielded a test statistic of 710.5 and an associated p-value of approximately 0.0004. The null hypothesis is rejected since the p-value is less than the significance level (0.05). This indicates that the choice of interaction type (mouse movement vs. touch interaction) significantly influences the number of incorrect answers recorded by participants. This supports the conclusion, drawn from the descriptive analysis, that there is a statistically significant difference in the number of wrong answers between the two interaction types, with the mouse movement group demonstrating better performance.

7.1.4 Usability Statistics

To assess the overall satisfaction of the visualization, the usability statistics focus on the ratings from the System Usability Scale (Section 7.1.4.1) and the feedback and comments provided by participants (Section 7.1.4.2), considering the 114 valid submissions.

7.1.4.1 System Usability Scale (SUS)

As detailed in Section 6.2.1.4, participants rated the standard ten-item SUS using a 5-point Likert scale. These questions offer insights into the effectiveness, usability, and satisfaction of the visualization from the participants' perspective. Table 7.8 and Fig. 7.6 presents the calculated minimum, 1st quartile, median, 3rd quartile, maximum, and standard deviation (std. dev.) for each question.

Table 7.8: Distribution of ratings per SUS question among participants.

	Minimum	1 st Quartile	Median	3 rd Quartile	Maximum	Std. Dev.
Q01 (Appreciation)	1.00	4.00	4.00	5.00	5.00	0.88
Q02 (Complexity)	1.00	2.00	2.00	3.00	5.00	0.91
Q03 (Ease of use)	1.00	3.00	4.00	4.00	5.00	0.87
Q04 (Ease of use)	1.00	1.00	2.00	3.00	4.00	1.06
Q05 (Consistency)	1.00	4.00	4.00	5.00	5.00	0.84
Q06 (Consistency)	1.00	1.00	2.00	2.75	4.00	0.89
Q07 (Speed)	1.00	3.00	3.50	4.00	5.00	1.06
Q08 (Ease of use)	1.00	1.00	2.00	3.00	5.00	0.92
Q09 (Confidence)	1.00	3.00	4.00	4.00	5.00	0.96
Q10 (Speed)	1.00	1.00	2.00	2.00	5.00	0.98

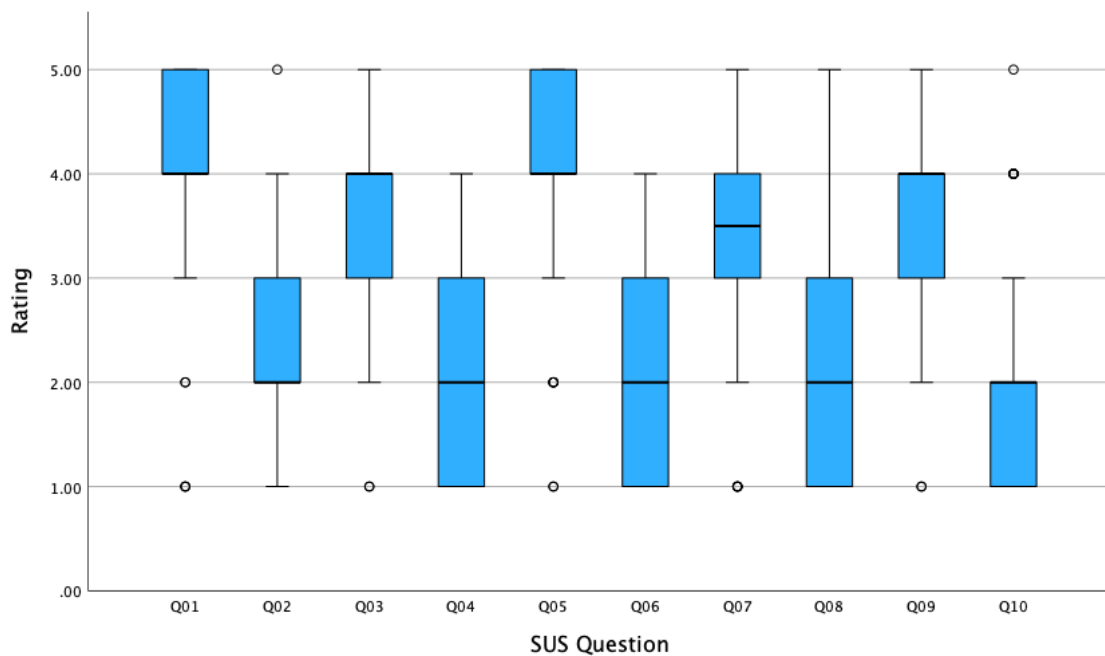


Figure 7.6: Box-plot of ratings per SUS question among participants.

Given that the odd-numbered questions are positively phrased (higher values indicate better responses) and the even-numbered questions are negatively phrased (lower values indicate better responses), the SUS results suggest that participants generally had a positive experience with the

visualization. The odd-numbered questions have relatively high median (mainly 4.00) and quartile values (1st quartile always equal or greater than 3.00). Conversely, the even-numbered questions exhibit lower median (2.00) and quartile values (3rd quartile always equal or lower than 3.00). Despite some response variability, the standard deviation values (typically closer to 1.00) indicate a relatively consistent range. Moreover, question 7, *I would imagine that most people would learn to use this system very quickly*, slightly deviates from this trend.

To gauge overall participant satisfaction with the visualization, the SUS Score [59] is computed for each questionnaire response according to Equation 7.2.

$$\left(\left(\sum_{i=1,3,5,7,9} Q_i - 5 \right) + \left(25 - \sum_{j=2,4,6,8,10} Q_j \right) \right) \times 2.5 \quad (7.2)$$

Each response results in a score on a scale from 0 to 100, with each question contributing 10 points. The average SUS score among the participants is 72.70, with a median score of 75.00. According to the interpretation table in [59], this corresponds to a grade of B. This indicates that, in general, the usability performance in terms of effectiveness, efficiency, and overall ease of use is good, aligning with the previously presented descriptive statistics.

7.1.4.2 Feedback and Comments

Out of 114 valid submissions, 40 participants provided feedback or comments on the utility of the tasks in revealing new functionalities or offered additional suggestions about the visualization. The feedback was positive and constructive, with some participants investing significant effort in their responses.

Regarding the utility of tasks, twelve participants stated that the tasks did not help, with one noting that the system's functionalities were already simple and explicit. Another mentioned that the tasks only enhanced their use of already discovered functionalities, as the interface was intuitive for their level of data analysis experience.

Conversely, several participants found the tasks useful in discovering new functionalities:

- Two participants recognized the information card on the corner of the visualization.
- Thirteen participants discovered the details graph and the ability to switch between the two graph perspectives, including four who appreciated clicking on clubs or national teams to see their players.
- One participant noted the feature of opening a player's information page by clicking on their node.
- Eight participants recognized and used the filters more extensively, such as player's position and clubs' continent, for more detailed analysis.

- Five participants perceived the highlight mechanism when searching for specific players, considering it not disturbing the overall visualization.
- Two participants found the *play* button for viewing the timeline progression through World Cups and the ability to step through each one with the other timeline buttons.

Eighth of the other participants responded *yes* without further details. Other mentions include understanding the significance of node size and edge thickness in comparing club contributions.

Regarding additional comments, ten participants praised the work with remarks like *Great job!*, *Interesting and well-executed tool*, *Useful and appealing tool*, and *Intuitive tool*. One participant, despite praising the work, noted difficulties due to their lack of soccer expertise. Another participant appreciated the intuitive organization of the information in a weighted circle layout.

Participants also provided constructive criticism and suggestions for improvements:

- One participant had trouble understanding why some clubs are identified by a logo or color.
- One participant suggested that when highlighting a national team, the sizes of the clubs should dynamically change based on their contribution to that national team instead of keeping their original size representing the total number of players supplied to all national teams.
- Three participants suggested adding a reset button for the filters.
- Two participants recommended improving the visibility of the highlighted nodes of the search player in each World Cup, suggesting visual animations to draw attention to them.
- One participant recommended that after filtering clubs by continent, the country filter should only display countries within the selected continent.
- One participant felt that the filters would work better if they automatically restricted the timeline dates to the not empty timeslices.
- One participant suggested displaying a summary title of the active filters to remind participants of the current filters applied.
- One participant suggested associating the visualization with statistics for each World Cup, such as adding tables or graphs for aggregated information.

Most suggestions involved the filter functionalities, with no significant issues reported related to the overall structure and organization of the visualization. Other feedback included reports of spelling errors in the questionnaire, complicated and unexpected questions, and a comment that the system was good for a quiz.

The feedback highlighted a generally positive reception of the visualization tool, emphasizing its ease of use, intuitive design, and effectiveness in revealing new insights over complex, intricate, and time-evolving soccer data. One comment particularly encapsulated the overall sentiment (translated from Portuguese):

Never seen anything like this regarding soccer. Besides seeing which clubs contribute the most to national teams, we can also see the players and easily select the year of the competition, the national team, the club, and even the player with a direct link to their page on zerozero. I liked it a lot. Congratulations.

7.2 Internal Analytics

Internal analytics is an assessment method to gather insights into the visualization's popularity and engagement through user behavior analysis. The metrics include the duration users spend on the visualization, the number of page visits, demographic details, and the technology used, as well as the tracked user's interactions with the prototype elements. These metrics help understand how users feel attracted to the visualization and how they interact with it. Users can also be segmented based on characteristics such as new vs. returning visitors, geographic location, or device type. Comparing engagement metrics and user behavior across these segments helps identify patterns and preferences.

Section 6.2.2 detailed the purpose and limitations of each tool used: Google Analytics and Hotjar. As mentioned in that section, a significant limitation is that the data presented in the following sections do not account for users who have extensions blocking data collection by these analytics tools.

7.2.1 Google Analytics

During the experiment period (from April 19 to May 5), the visualization page *Campeonatos do Mundo* (World Cups) received a total of 2k visits from approximately 1.5k distinct active users (see Fig. 7.7). The highest influx of new users occurred on April 23 and 24, coinciding with news about the visualization on the company's website, as mentioned in Section 6.1.2. Additionally, 152 returning users (those who initiated at least one previous session) were registered. The average views per user were 1.34 ($= 2007/1497$).

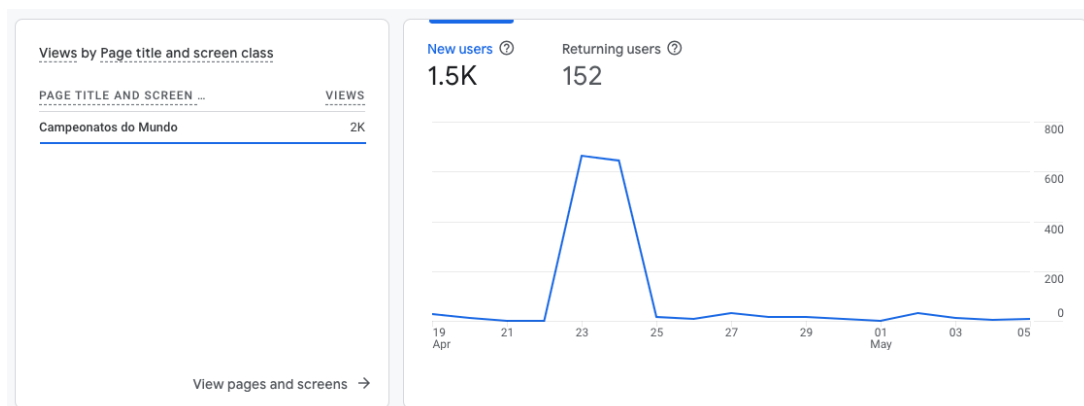


Figure 7.7: Overview of visits and user activity during the experiment period.

Users from 36 different countries accessed the visualization, primarily from Europe and America, with the majority being from Portugal (1.3k users). Specifically, in Portugal, most visits were from Lisbon. Additionally, the majority of users were Portuguese. It's important to note that active VPNs and system configurations can influence these values. The overview of user demographics is presented in Fig. 7.8.

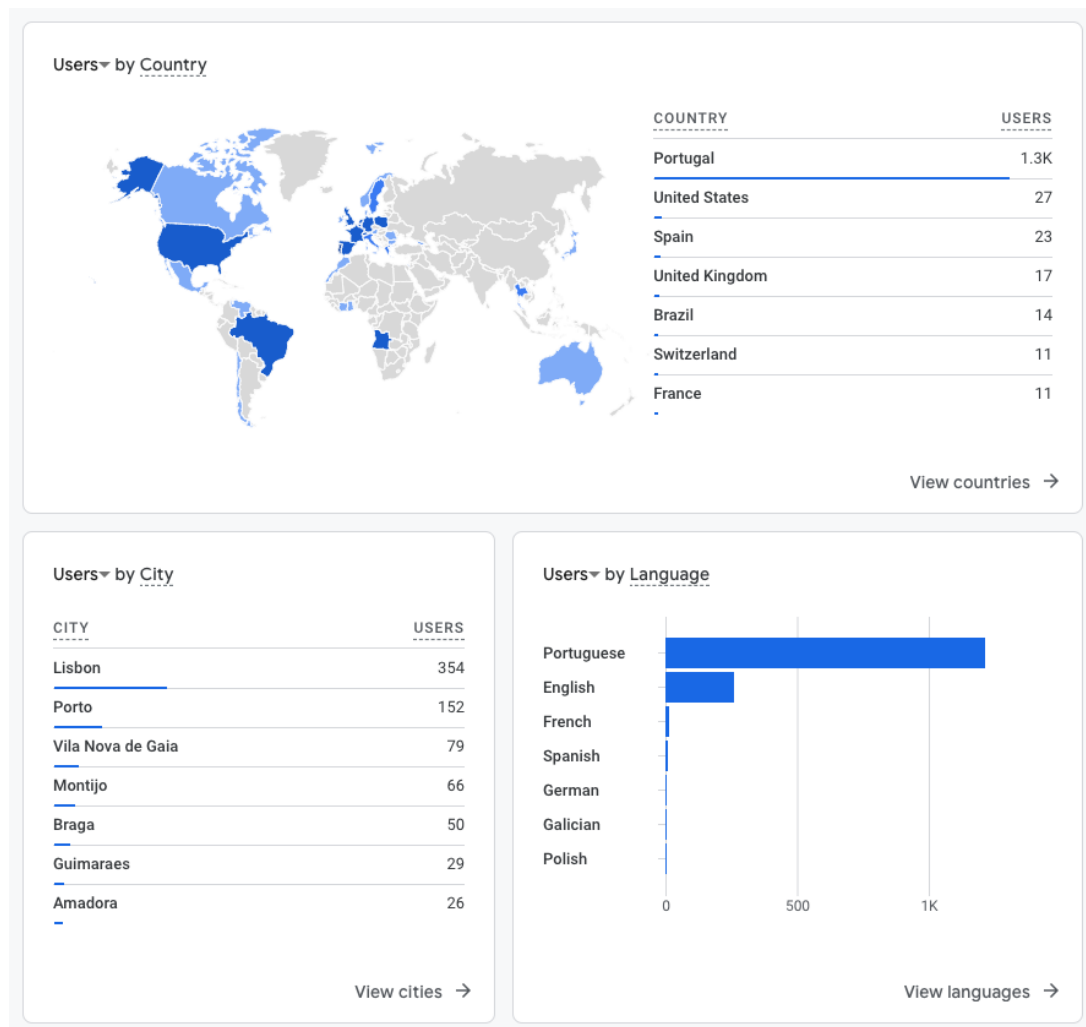


Figure 7.8: Demographic details of users during the experiment period.

Regarding technology (see Fig. 7.9), the number of users accessing via mobile or desktop was nearly equivalent: 745 on mobile, 743 on desktop, and 9 on tablets. This distribution supports adapting the visualization's interaction and layout for both predominant device categories. Six distinct operating systems were recorded. The diversity in screen resolutions (126 different values) highlights the need for a responsive layout. Additionally, the various browsers used (the 7 listed in Fig. 7.9 plus Opera) underscore the necessity for compatibility across different web browsers on both mobile and desktop.

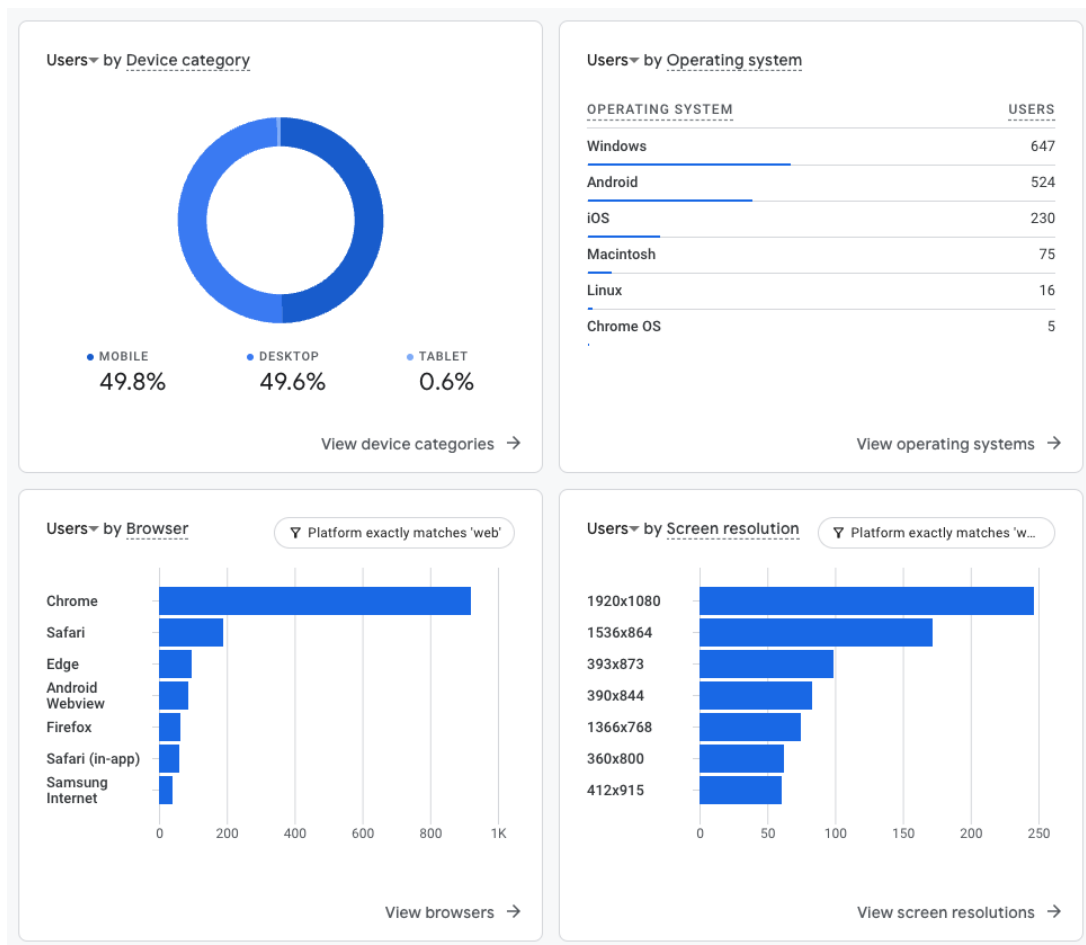


Figure 7.9: Technology details of users during the experiment period.

Based on Fig. 7.10, the average engagement time per user was 1 minute 50 seconds, and the average session count per active user during the experiment period was 0.92. The second metric indicates that some sessions did not involve further interaction with the visualization. Precisely, the percentage of sessions that were not engaged was 21.4%.

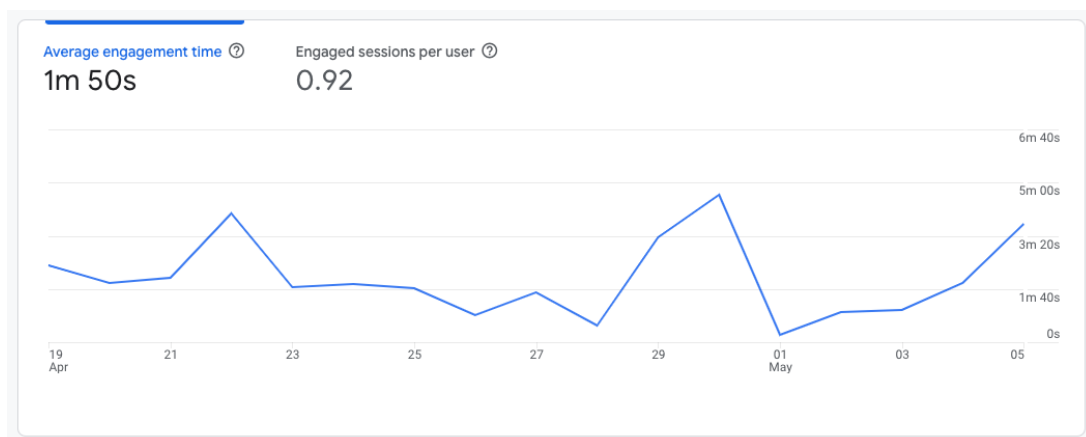


Figure 7.10: Engagement details of users during the experiment period.

7.2.1.1 Events Analytics

In addition to providing demographic, behavioral, and technological insights, the configured events (refer to Table 6.1) help highlight the most frequently used functionalities. During the experiment period, a total of 48k events were recorded (see Fig. 7.11), resulting in an average event count per user of 32.12 ($= 48083/1497$).

Event name		↓ Event count	Active users	Event count per user
Totals		48,083	1,497	32.12
1	node_highlight	20,833	994	20.96
2	control_timeline	11,995	738	16.25
3	node_in_detail	4,514	693	6.51
4	change_filter	2,359	521	4.53
5	page_view	2,007	1,496	1.34
6	session_start	1,752	1,496	1.17
7	first_visit	1,489	1,486	1
8	scroll	1,332	1,162	1.15
9	user_engagement	1,199	974	1.23
10	change_sort	339	194	1.75
11	change_search	225	149	1.51
12	click	39	37	1.05

Figure 7.11: Total events recorded during the experiment period.

node_highlight event On average, each of the 1k users who utilized the node highlight mechanism highlighted 21 nodes, which could be distinct or repeated. Users explored nodes in the cluster graph more frequently than in the detail graph (16k vs. 5k occurrences). In both graphs, inner nodes were the most explored: clubs in the cluster graph and players in the detail graph. Although national teams were less explored, the difference was only 1.8k occurrences for each graph type (see Fig. 7.12).

Event Graph		cluster			detail			Totals		
Event name	Event Category	Event count	Active users	Event count per user	Event count	Active users	Event count per user	↓ Event count	Active users	Event count per user
Totals		15,838 76.02% of total	990 99.6% of total	16 Avg -23.67%	4,995 23.98% of total	507 51.01% of total	9.85 Avg -52.99%	20,833 100% of total	994 100% of total	20.96 Avg 0%
1	node_highlight teams	8,837	848	10.42	0	0	0	8,837	848	10.42
2	node_highlight national teams	7,001	833	8.4	1,644	357	4.61	8,645	864	10.01
3	node_highlight players	0	0	0	3,351	401	8.36	3,351	401	8.36

Figure 7.12: Breakdown of *node_highlight* events by category and graph type.

Not surprisingly, when examining the top 10 specific nodes explored in each graph, the most explored node in the cluster graph was the national team of *Portugal* (2k occurrences), followed by three prominent Portuguese clubs: *Benfica* (842), *FC Porto* (754), and *Sporting* (707) (see Fig. 7.13a). In the detail graph, only the national teams of *Portugal* and *Brazil* stood out (see

Fig. 7.13b). Interestingly, in the cluster graph, the top nodes were mainly clubs, while in the detail graph, the top nodes were mainly national teams. The occurrence of the player *Cafu* among the top 10 most explored nodes in the detail graph can be attributed to a questionnaire question specifically targeting that player.

Event name	Event Value	↕ Event count	Active users	Event count per user
Totals		15,838	990	16
1	node_highlight (O-811) Portugal	2,129	629	3.38
2	node_highlight (C-4) Benfica	842	325	2.59
3	node_highlight (C-9) FC Porto	754	317	2.38
4	node_highlight (C-16) Sporting	707	244	2.9
5	node_highlight (C-40) Barcelona	645	248	2.6
6	node_highlight (C-86) Manchester City	450	243	1.85
7	node_highlight (O-812) Alemanha	397	172	2.31
8	node_highlight (O-816) Brasil	381	176	2.16
9	node_highlight (C-108) Bayern München	374	182	2.05
10	node_highlight (C-87) Manchester United	276	194	1.42

(a) Nodes in the cluster graph.

Event name	Event Value	↕ Event count	Active users	Event count per user
Totals		4,995	507	9.85
1	node_highlight (O-811) Portugal	523	213	2.46
2	node_highlight (O-816) Brasil	117	48	2.44
3	node_highlight (O-822) Espanha	85	50	1.7
4	node_highlight (O-814) Argentina	70	41	1.71
5	node_highlight (E-1946) Cafu	64	21	3.05
6	node_highlight (O-812) Alemanha	59	38	1.55
7	node_highlight (O-1018) Suíça	55	40	1.38
8	node_highlight (O-1016) Sérvia	51	38	1.34
9	node_highlight (O-824) França	48	39	1.23
10	node_highlight (E-3458) Marco Caneira	47	37	1.27

(b) Nodes in the detail graph.

Figure 7.13: Most highlighted nodes of *node_highlight* events by graph type.

Regardless of the graph type, the most explored nodes generally belonged to the World Cup of *Qatar 2022*, with over 10k occurrences (see Fig. 7.14). This discrepancy is also evident when comparing the number of active users.

	Event name	Event Timeslice	↓ Event count	Active users	Event count per user
	Totals		20,833	994	20.96
1	node_highlight	Catar 2022	10,563	908	11.63
2	node_highlight	Inglaterra 1966	1,042	178	5.85
3	node_highlight	EUA 1994	1,010	121	8.35
4	node_highlight	Coreia/Japão 2002	984	170	5.79
5	node_highlight	África do Sul 2010	932	154	6.05
6	node_highlight	Alemanha 2006	861	164	5.25
7	node_highlight	Rússia 2018	796	177	4.5
8	node_highlight	México 1986	770	140	5.5
9	node_highlight	Brasil 2014	673	154	4.37
10	node_highlight	França 1998	475	98	4.85

Figure 7.14: Breakdown of *node_highlight* events by timeslice.

node_in_detail event On average, each user views detailed information for 6 nodes out of the 700 users who engaged with this feature (less than half of all page visitors). Most of these interactions occurred in the cluster graph, which attracted more than double the number of active users compared to the detail graph. In the cluster graph, national teams and clubs were explored in roughly equal measure. In contrast, in the detail graph, national teams were significantly more explored than individual players (see Fig. 7.15).

Event Graph		cluster			detail			Totals		
Event name	Event Category	Event count	Active users	Event count per user	Event count	Active users	Event count per user	↓ Event count	Active users	Event count per user
Totals		3,293 72.95% of total	692 99.86% of total	4.76 Avg 26.94%	1,221 27.05% of total	339 48.92% of total	3.6 Avg 44.7%	4,514 100% of total	693 100% of total	6.51 Avg 0%
1	node_in_detail national teams	1,626	529	3.07	792	246	3.22	2,418	529	4.57
2	node_in_detail teams	1,667	503	3.31	0	0	0	1,667	503	3.31
3	node_in_detail players	0	0	0	429	198	2.17	429	198	2.17

Figure 7.15: Breakdown of *node_in_detail* events by category and graph type.

Examining the top 10 most viewed nodes, those in the cluster graph (Fig. 7.16a) closely mirror the top nodes highlighted in the *node_highlight* events (Fig. 7.13a), with the top 5 nodes being precisely the same. However, in the detail graph (Fig. 7.16b), the top 10 nodes consist exclusively of national teams.

Event name	Event Value	↓ Event count	Active users	Event count per user
Totals		3,293	692	4.76
1	node_in_detail (O-811) Portugal	607	328	1.85
2	node_in_detail (C-4) Benfica	242	142	1.7
3	node_in_detail (C-9) FC Porto	209	117	1.79
4	node_in_detail (C-16) Sporting	188	103	1.83
5	node_in_detail (C-40) Barcelona	141	90	1.57
6	node_in_detail (O-812) Alemanha	101	66	1.53
7	node_in_detail (O-816) Brasil	99	59	1.68
8	node_in_detail (C-86) Manchester City	94	77	1.22
9	node_in_detail (O-822) Espanha	81	51	1.59
10	node_in_detail (C-108) Bayern München	59	49	1.2

(a) Nodes in the cluster graph.

Event name	Event Value	↓ Event count	Active users	Event count per user
Totals		1,221	339	3.6
1	node_in_detail (O-811) Portugal	335	134	2.5
2	node_in_detail (O-816) Brasil	36	18	2
3	node_in_detail (O-812) Alemanha	33	21	1.57
4	node_in_detail (O-1016) Sérvia	29	14	2.07
5	node_in_detail (O-822) Espanha	21	13	1.62
6	node_in_detail (O-824) França	20	16	1.25
7	node_in_detail (O-815) Bélgica	18	13	1.38
8	node_in_detail (O-814) Argentina	17	13	1.31
9	node_in_detail (O-1018) Suíça	16	14	1.14
10	node_in_detail (O-835) Polónia	16	14	1.14

(b) Nodes in the detail graph.

Figure 7.16: Most viewed nodes of *node_in_detail* events by graph type.

Similarly, the most common timeslices for detailed inspections (Fig. 7.17) generally align with those for *node_highlight* events (Fig. 7.14). Most of these nodes belong to the World Cup of *Qatar 2022*, with user engagement dropping significantly for nodes outside this timeslice (from nearly 600 users to no more than 100).

	Event name	Event Timeslice	↓ Event count	Active users	Event count per user
	Totals		4,514	693	6.51
1	node_in_detail	Catar 2022	2,231	566	3.94
2	node_in_detail	EUA 1994	320	67	4.78
3	node_in_detail	Inglaterra 1966	273	88	3.1
4	node_in_detail	Coreia/Japão 2002	239	73	3.27
5	node_in_detail	Rússia 2018	231	64	3.61
6	node_in_detail	África do Sul 2010	191	59	3.24
7	node_in_detail	Alemanha 2006	170	63	2.7
8	node_in_detail	México 1986	165	62	2.66
9	node_in_detail	Brasil 2014	149	51	2.92
10	node_in_detail	Itália 1990	112	34	3.29

Figure 7.17: Breakdown of *node_in_detail* events by timeslice.

control_timeline event Approximately half of the active users interacted with the timeline component, engaging with either the timeline buttons or the range selector. The distribution of event occurrences per timeline action is illustrated in Fig. 7.18. Users predominantly utilized the next or previous buttons to navigate through the timeline. Directly selecting a timeslice using the range slider also emerged as a popular choice (*change* event category). However, automatic manipulation of timeslices using the play/pause button was less frequent, with only around 150 users pausing the timeline. Interestingly, the *goto* feature (changing to a timeslice via a search result link) was even less utilized.

	Event name	Event Category	↓ Event count	Active users	Event count per user
	Totals		11,995	738	16.25
1	control_timeline	next	5,029	342	14.7
2	control_timeline	prev	3,526	326	10.82
3	control_timeline	change	2,000	466	4.29
4	control_timeline	play	736	355	2.07
5	control_timeline	goto	380	88	4.32
6	control_timeline	pause	324	144	2.25

Figure 7.18: Breakdown of *control_timeline* events by category.

The specific values associated with each *control_timeline* event category are related to a timeslice. However, its meaning varies. For the *prev*, *next*, *play*, and *pause* categories, categories that involve clicking on a button, the value is related to the timeslice where the button was clicked. In contrast, for the *change* and *goto* categories, the value indicates the specific timeslice that was changed to or selected.

When analyzing the specific timeslices associated with each event category:

- *next*: The frequency of events and users ranged closely, with 200-250 event counts and 120-180 users (refer to Annex C, Fig. C.24).
- *prev*: The Qatar 2022 timeslice stood out slightly from others, showing higher event counts and active users (refer to Annex C, Fig. C.25).
- *change*: The first two options corresponded precisely to the timeline extremes, with the remaining values clustered closely together but distinctly different from the top two (refer to Annex C, Fig. C.26).
- *play*: Exhibited similar patterns to the *change* category, but with less representation at the lower extreme of the timeline and generally lower event counts (refer to Annex C, Fig. C.27).
- *pause*: No standout timeslices were observed, likely due to the automatic cessation of animation upon reaching the end of Qatar 2022 (refer to Annex C, Fig. C.28).
- *goto*: The values are highly influenced by the players searched for, as the displayed links correspond to the World Cups where those players were present. The first two timeslices, especially the first, align with the World Cups that featured the player targeted in the formative questionnaire users were asked to search for (refer to Annex C, Fig. C.29).

change_sort event In the prototype, sorting functionality is limited to nodes in the outer circle. This restriction resulted in 339 occurrences of the *change_sort* event (Fig. 7.11), specifically involving the sorting of national team nodes. By default, these nodes were ordered by continent. Out of the 1.5k users, 194 interacted with this control.

Most users changed the sort order to alphabetical by national team name, as depicted in Fig. 7.19. Interestingly, only a few users reverted to the original continent-based sorting after exploring other options.

	Event name	Event Value	↓ Event count	Active users	Event count per user
	Totals		339	194	1.75
1	change_sort	name	227	194	1.17
2	change_sort	fk_continente	112	96	1.17

Figure 7.19: Breakdown of *change_sort* events by value.

change_filter event Among the available interactive controls in the prototype, filters were the most frequently used, with around one-third of the active users engaging with them. These filters are categorized into four types: national teams by continent, clubs by continent, clubs by country, and players by position. The distribution of event occurrences per filter category is shown in Fig. 7.20. Notably, the number of users filtering players by position is relatively lower than the

other three categories. Most filter changes are related to filtering clubs by country, followed closely by filtering national teams by continent. Conversely, filtering players by position recorded the fewest occurrences among the categories.

Event name	Event Category	↓ Event count	Active users	Event count per user
Totals		2,359	521	4.53
1	change_filter teams_country	710	374	1.9
2	change_filter national-teams_fk_continente	694	393	1.77
3	change_filter teams_continente	546	313	1.74
4	change_filter players_posicao	409	222	1.84

Figure 7.20: Breakdown of *change_filter* events by category.

When inspecting the most commonly selected criteria within each filter category reveals the following trends:

- **National Teams by Continent:** The most frequently chosen option is *Europe*, followed by *all* and *Oceania*. Across the remaining options, it is not shown significant discrepancies in selection frequency (Fig. 7.21a).
- **Clubs by Continent:** Similar to the national teams filter, *Europe* emerges as the top choice, followed by *all*. The other options have relatively similar selection frequencies with no significant differences in event count or active user numbers per option (Fig. 7.21b).
- **Clubs by Country:** The most selected option is *Portugal*, followed by *all*. Notably, *Germany* and *Spain* are in the middle range with about 30 occurrences each (Fig. 7.21c). In total, there were change filter events related to 65 distinct values, but after the 10th place, the event count per user is mostly 1, with the number of users and event count not exceeding 4.
- **Players by Position:** *Avançado* (Forward) stands out as the preferred choice, aligning with a formative questionnaire question targeting this functionality. This is followed by the *all* option. For the remaining options, the event count per user is very close to 1, indicating that almost no user selects the same value twice (Fig. 7.21d).

Across all filter categories, the default option is consistently *all*, ranking as the second most selected value. This suggests that users maintain their chosen filter selection rather than frequently reverting to the default *all* option unless they actively change it back.

It's important to note that the analysis of filters is conducted independently, without data on simultaneous filter combinations, to derive specific insights.

Event name	Event Value	↓ Event count	Active users	Event count per user
Totals		694	393	1.77
1	change_filter 1 Europe	388	354	1.1
2	change_filter all	158	120	1.32
3	change_filter 6 Oceania	63	59	1.07
4	change_filter 4 South America	31	26	1.19
5	change_filter 2 Asia	24	24	1
6	change_filter 5 Africa	19	18	1.06
7	change_filter 3 North America	11	11	1

(a) Values in filter national teams by continent.

Event name	Event Value	↓ Event count	Active users	Event count per user
Totals		546	313	1.74
1	change_filter Europa	330	288	1.15
2	change_filter all	102	76	1.34
3	change_filter Oceania	27	26	1.04
4	change_filter América do Sul	22	21	1.05
5	change_filter América do Norte, América Central, Caraíbas	17	15	1.13
6	change_filter CEV	15	15	1
7	change_filter EHF	14	14	1
8	change_filter Ásia	10	9	1.11
9	change_filter África	9	9	1

(b) Values in filter clubs by continent.

Event name	Event Value	↓ Event count	Active users	Event count per user
Totals		710	374	1.9
1	change_filter Portugal	349	315	1.11
2	change_filter all	132	88	1.5
3	change_filter Alemanha	33	33	1
4	change_filter Espanha	32	30	1.07
5	change_filter Brasil	13	11	1.18
6	change_filter Inglaterra	13	13	1
7	change_filter Argentina	9	9	1
8	change_filter Polónia	7	7	1
9	change_filter Angola	6	6	1
10	change_filter Paraguai	5	5	1

(c) Values in filter clubs by country.

Figure 7.21: Most selected values of *change_filter* events by category.

	Event name	Event Value	↓ Event count	Active users	Event count per user
	Totals		409	222	1.84
1	change_filter	Avançado	165	154	1.07
2	change_filter	all	103	86	1.2
3	change_filter	Guarda Redes	56	55	1.02
4	change_filter	Defesa	38	36	1.06
5	change_filter	Médio	20	20	1
6	change_filter	Jogador de Campo	17	17	1
7	change_filter	Fixo	10	10	1

(d) Values in filter players by position.

Figure 7.21: Most selected values of *change_filter* events by category.

change_search event The prototype’s search functionality was exclusively available for player nodes, accounting for the 225 occurrences depicted in Fig.7.11. These events correspond to user searches specifically for player nodes. Notably, only 149 out of 1.5k users interacted with this feature during the experiment period.

Upon examining of the specific players searched, the data reveals that the most frequently searched names were *Cafu* and *Manuel Akanji*, both highlighted in the formative questionnaire. Furthermore, there was considerable interest in Portuguese soccer legends and globally recognized players, as indicated in Fig.7.22.

	Event name	Event Value	↓ Event count	Active users	Event count per user
	Totals		225	149	1.51
1	change_search	Cafu	55	49	1.12
2	change_search	Manuel Akanji	24	20	1.2
3	change_search	Cristiano Ronaldo	11	10	1.1
4	change_search	Cris	8	5	1.6
5	change_search	Aaron Ramsdale	7	7	1
6	change_search	Ronaldo	7	7	1
7	change_search	Lionel Messi	6	6	1
8	change_search	Diogo Costa	4	4	1
9	change_search	Eusébio	4	4	1
10	change_search	Luís Figo	4	4	1

Figure 7.22: Breakdown of *change_search* events by value.

7.2.1.2 Segmentation

Comparing user behavior across different segments based on characteristics such as geographic location, language, or device type helps identify differences in preferences.

Regarding geographic location and language, these characteristics slightly influence the specific top 10 nodes explored in each graph for both the *node_highlight* and *node_in_detail* events. Specifically, certain players appear more frequently in the top 10 nodes explored or inspected in these segments. However, these segments do not significantly affect other events (e.g., how controls are used and values selected for sorting, filters, search, and timeline).

Regarding technology, when considering segments by device category (see Fig. 7.23), there is a clear tendency for desktop users to interact more with nodes for both *node_highlight* and *node_in_detail* events, particularly in cluster graph. However, an exception exists: the event count is higher on mobile devices in the detail graph for the *node_in_detail* events. This may be due to layout differences. Both graphs are displayed side by side on desktops, so users likely prefer to interact with the main one. On mobile devices, only one graph is shown at a time, leading users to interact more with the graph in focus.

Event Graph			cluster		detail		Totals	
Event name		Device category	Event count	Event count per user	Event count	Event count per user	↓Event count	Event count per user
Totals			15,838 76.02% of total	16 Avg -23.67%	4,995 23.98% of total	9.85 Avg -52.99%	20,833 100% of total	20.96 Avg 0%
1	node_highlight	desktop	13,271	21	3,850	11.88	17,121	27.09
2	node_highlight	mobile	2,528	7.14	1,142	6.27	3,670	10.25
3	node_highlight	tablet	39	9.75	3	3	42	10.5

(a) Segmentation in *node_highlight* events.

Event Graph			cluster		detail		Totals	
Event name		Device category	Event count	Event count per user	Event count	Event count per user	↓ Event count	Event count per user
Totals			3,293 72.95% of total	4.76 Avg -26.94%	1,221 27.05% of total	3.6 Avg -44.7%	4,514 100% of total	6.51 Avg 0%
1	node_in_detail	desktop	2,396	5.36	440	2.49	2,836	6.34
2	node_in_detail	mobile	877	3.64	781	4.82	1,658	6.85
3	node_in_detail	tablet	20	5	0	0	20	5

(b) Segmentation in *node_in_detail* events.

Figure 7.23: Breakdown of node events by graph type segmented by device category.

The device category does not significantly affect the use of other events individually. However, looking at the bigger picture, the mobile segment shows less distinction in event counts between different event types, with *control_timeline* being the most frequent at nearly 5k events. In contrast, *node_highlight* drops to second place with around 4k occurrences.

7.2.2 Hotjar

The analysis using Hotjar data is constrained by the limitations of the free plan, which includes no advanced filters or segment insights, no trend analysis, a limit of 35 sessions per day (particularly affecting periods with higher traffic), and a restricted packet size for each session, sometimes resulting in incomplete session recordings. However, Hotjar ensures a representative sample by distributing the collection of session recordings throughout the day despite the daily limit.

During the experiment period, from April 19 to May 5, Hotjar recorded 240 sessions, with an average session duration of 2 minutes and 16 seconds and an average time on page of 1 minute and 47 seconds. The data showed 1.8k clicks, a 66.8% average scroll depth, a 79.2% abandonment rate, three rage clicks, and 19 sessions with console errors. However, the lack of filtering capabilities limited detailed inspection of the two later issues. Additionally, the data showed that the button to answer the questionnaire in the top alert was used more frequently than the one in the footer (64 sessions vs. 7 sessions).

The user acquisition, demographics, and technology overview of recorded sessions are presented in Fig. 7.24. Most users were new and accessed the visualization from Portugal. Approximately 41% accessed the visualization via the zerozero website. The percentage of mobile and desktop sessions was very similar, with only one tablet user recorded. Chrome was the most popular browser, but there was notable diversity in browsers and operating systems.

Inspection of the heatmaps for the initial state of the page (the cluster graph of Qatar 2022) during desktop sessions revealed several user interaction patterns. In the clicks heatmap (see Fig. 7.25), the next and previous timeline buttons were most frequently used, followed by other timeline components and various controls, with a notable preference for filters of national teams and clubs and sorting controls. Specific nodes with visible clicks included the Portugal and Germany national teams.

The mouse movement heatmap for desktop sessions in Fig. 7.26 showed Portugal as the most explored node and significant activity around Germany, Barcelona, Manchester City, Benfica, and Brazil. The timeline and control components mirrored the clicks heatmap patterns, with a more explicit preference for exploring years closer to the timeline extremes, especially the most recent timeslices.

The mobile clicks heatmap (Fig. 7.27) closely resembled the desktop clicks heatmap (Fig. 7.25) but with a stronger focus on using the filter by player positions and less emphasis on selecting the timeslice through direct range clicks. This suggests that the device category does not heavily influence visualization control usage and timeline manipulation.

Overall, users interacted with various aspects of the visualization, with some controls being less popular.

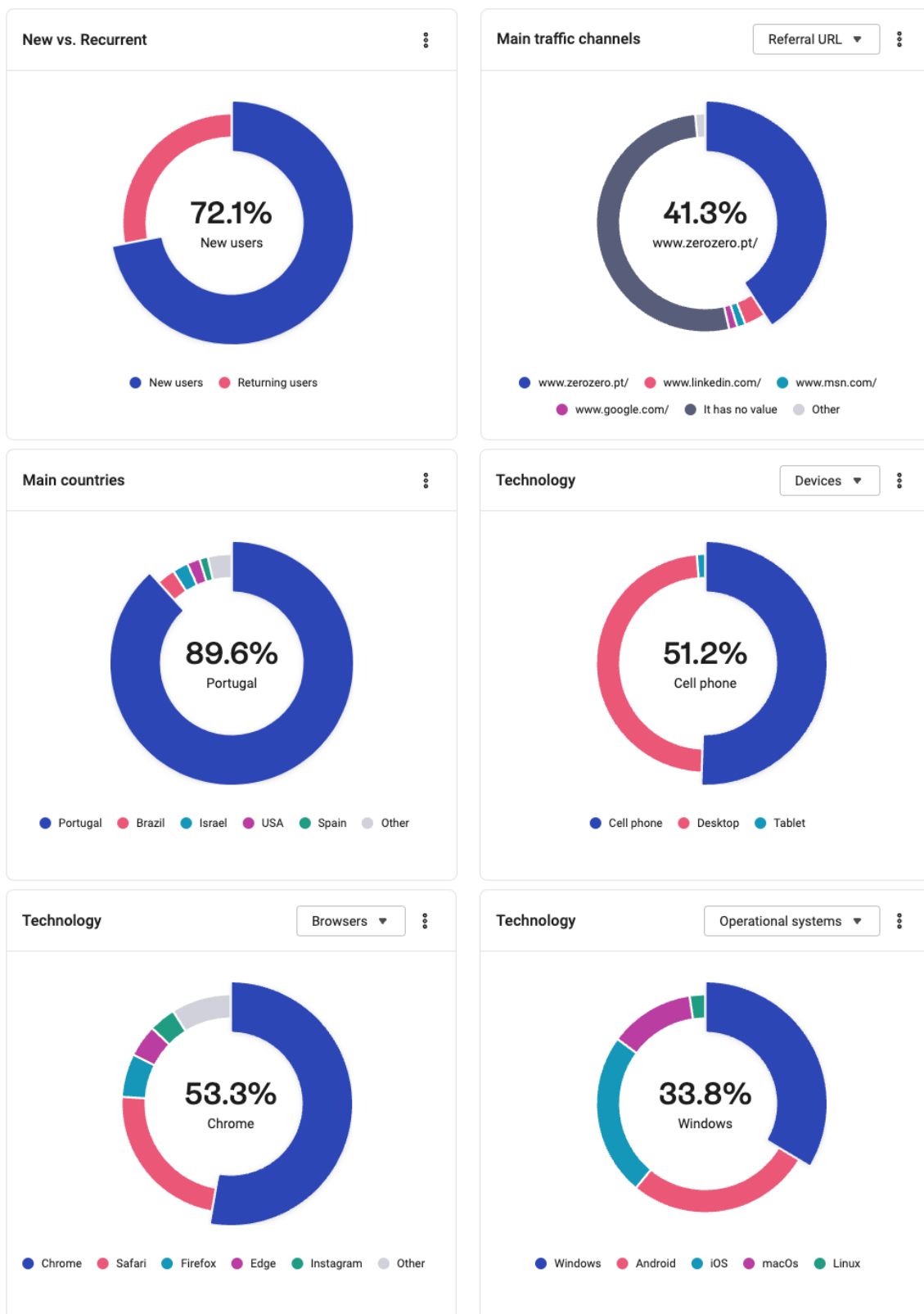


Figure 7.24: Hotjar demographics and technology overview during the experiment period.

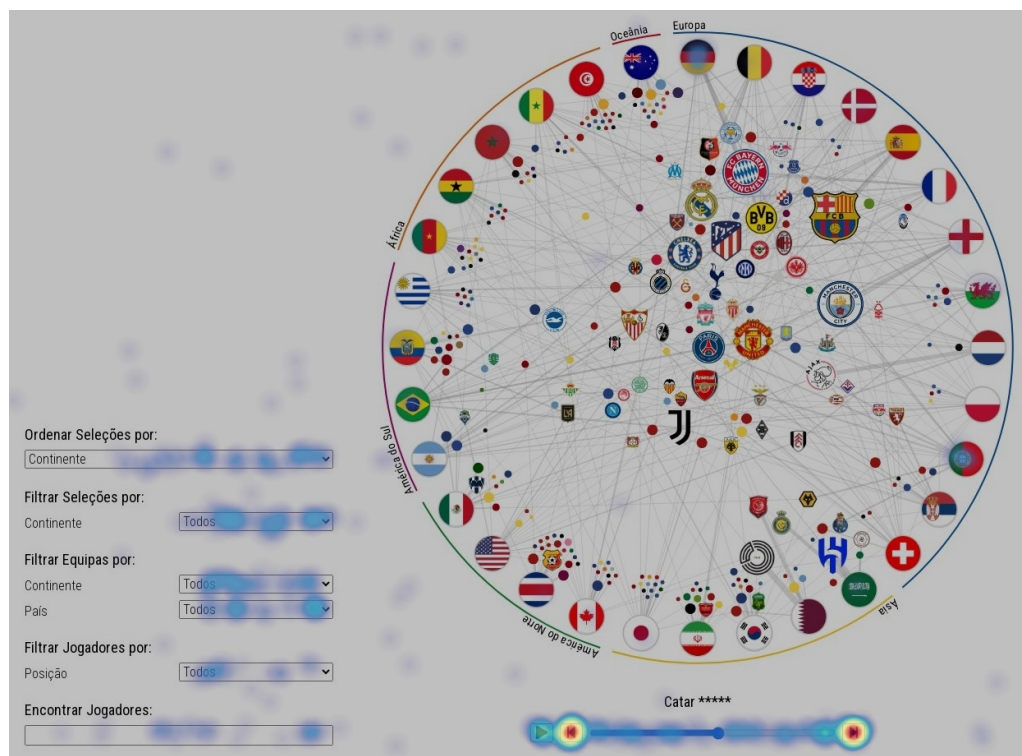


Figure 7.25: Hotjar heatmap over desktop clicks during the experiment period.

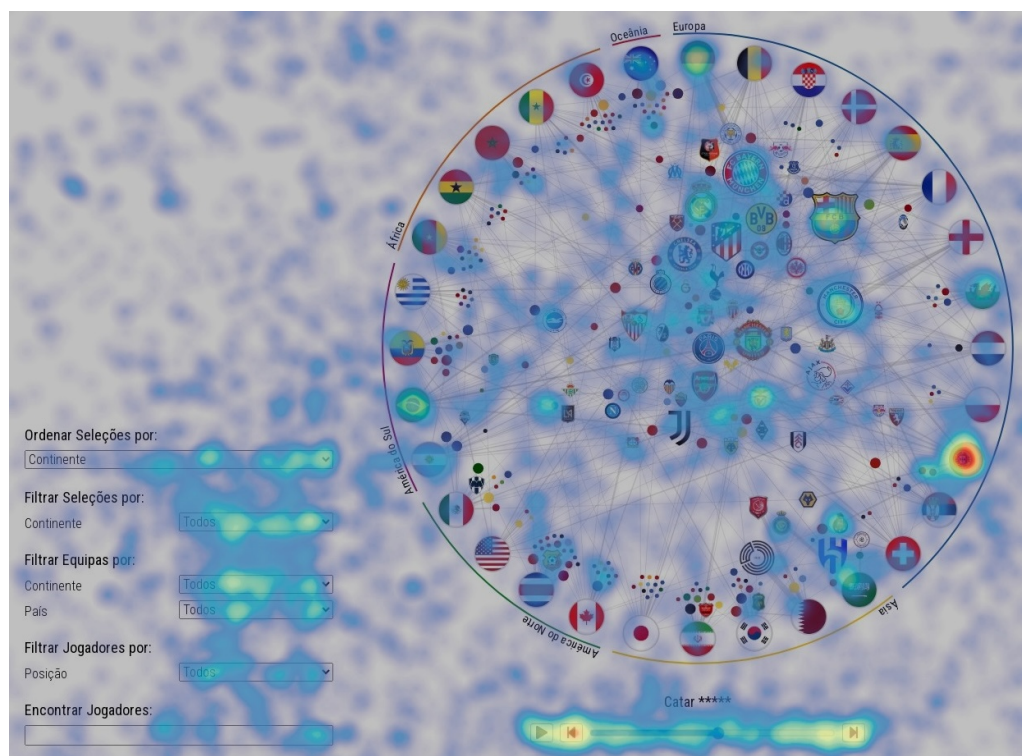


Figure 7.26: Hotjar heatmap over desktop mouse movement during the experiment period.



Figure 7.27: Hotjar heatmap over mobile clicks during the experiment period.

The movement heatmap for other timeslices was not inspected since Google Analytics already provided accurate data on explored nodes, offering no additional benefits. Similarly, nodes explored in the details graph were not separately analyzed.

Furthermore, no specific session recordings were inspected in detail, as no questionnaire responses indicated the necessity. The free Hotjar plan does not allow selection based on particular characteristics, such as errors or rage clicks.

7.3 Discussion

This section discusses the findings and implications of the results presented in the preceding sections, providing insights into some decision-making processes and suggesting potential improvements for the visualization.

7.3.1 Outliers Identification Criteria

The outlier identification strategy used in Section 7.1.1 proved to be a rigorous and unbiased method for filtering out random responses, which is crucial in an online, uncontrolled scenario. However, some results, particularly in individual task responses, may seem unusual because the criteria for identifying outliers were based on the entire submission rather than individual responses to each question. This latest approach might overlook smaller yet representative answers, potentially reflecting genuine misunderstandings of the visualization.

Furthermore, the average time recorded in Google Analytics was not used as the time spent for outlier identification due to the limitations of potentially not recorded sessions, errors in recording, and consideration of interaction time before and after the questionnaire. To avoid bias, the response time to the questionnaire registered in the questionnaire itself was used as the time metric for identifying and removing outlier submissions, as it better reflects the actual scenario of the time users spent interacting with the visualization across the questionnaire's multiple sections.

7.3.2 Demographics and Recruitment Strategies

The demographic statistics (Section 7.1.2) indicate that the varied recruitment strategies employed to gather participants described in Section 6.1.2 resulted in a balanced and inclusive participant pool. This diverse group, encompassing different backgrounds, expertise levels, and interests, enhances the validity of the findings and ensures that the visualization meets the needs and preferences of a broad user spectrum.

7.3.3 Tasks Performance and Demographics Influence

The high task accuracy per question (see Table 7.3) and overall positive task performance per participant (see Section 7.1.3.1) indicate that the visualization is effective for a diverse audience. However, incorporating a guide or tutorial could help users better understand how to interpret and use this innovative visualization, potentially improving performance.

Some lower accuracies in individual tasks can be explained. In Task 1, the difficulty arises from trying to visually infer the largest club, especially on smaller devices and among various distinct club logos. In Task 7, as observed in pilot tests, the interaction of navigating to the end of the timeline and moving forward to the first appearance of a club from a continent is not intuitive. Additionally, the term "first World Cup" used in this question is ambiguous, as not everyone associates it with the earliest tournament. In Task 8, some users might rely on their memory rather than exploring the visualization, leading to bias toward their favorite club. Furthermore, this task requires significant manipulation of the timeline and filters, contributing to the observed lower accuracy. Errors in matching the clubs or national teams mentioned in the questionnaire to their logo in the visualization also justify some inaccuracies, indicating that the placeholders with respective names are not clear enough. Conversely, in Task 4, the accuracy is high despite the numerous possible options and regardless of the interaction type or domain knowledge, suggesting an overall understanding of the organization and manipulation of information in the visualization.

Examining the influence of time spent and domain knowledge on task performance reveals that participants who spent more time on the questionnaire generally performed better, making fewer errors. Domain knowledge also played a role, with higher information visualization knowledge correlating generally with better performance, regardless of sports knowledge. However, participants with good knowledge in both domains achieved the most perfect scores. Participants with higher levels of sports knowledge independently of the InfoViz knowledge likely knew the

answers to some questions without needing to explore the visualization but might lack the expertise to navigate the visualization for more complex questions. Despite this, the statistical methods used to inspect the correlation between domains and task performance did not reveal a clear trend, likely due to the self-reported nature of domain knowledge, which might introduce bias.

Regarding the influence of interaction type on task performance, it is evident that touch interaction negatively affects some tasks. In Task 1, visually comparing different club node sizes is more challenging on smaller screens. In Task 8, the difficulty of individually checking club nodes in each timeline can be more tedious and confusing on a touch interface compared to using a hover mechanism with a mouse. For Task 7, similar to Task 1, perceiving nodes, text, and legends on a smaller screen can be more challenging, as the visualization's zoom is reduced to fit the screen. Participants using touch interfaces tended to make more errors, highlighting the need for a guide to explain these less familiar interaction types.

7.3.4 System Usability and Participant Feedback

The System Usability Scale (SUS) results (Section 7.1.4.1) show participants' positive inclination toward using the system, finding it user-friendly, consistent, well-integrated, and relatively simple. However, perceptions of system learnability varied based on participants' levels of expertise in information visualization and sports. Considering the diverse target audience, the learning curve may be more pronounced. Thus, while the system is generally perceived as easy to learn, factors such as audience diversity influence overall perceptions.

Based on participant feedback, several improvements could enhance the visualization. Integrating highlighted node information within the visualization could help avoid losing visual context. For example, adding tooltips with specific contribution values alongside each node would address the issue of the unnoticed information in the lower right corner. However, this approach might obscure the visual perception of the edges and node sizes. Additionally, incorporating a placeholder for the details graph in the desktop layout would guide users in exploring more information. Providing overlays or tooltips for timeline buttons would clarify their functions beyond the image icons used. To further improve usability, a title reflecting active filters could help users keep track of their settings. Adding a reset filters button and reducing the number of filters to those essential would simplify the interactions and interface for less experienced users.

Some suggested improvements were discarded, such as dynamically changing the nodes connected to a highlighted node, as the edge size already indicates the contribution value. Dynamically altering the timeline values based on filters was also not considered, as this would create gaps in the timeline, making it harder to understand. The goal is to maintain timeline comprehension by displaying all time slices, even those without nodes after specific filter combinations. Additionally, incorporating supplementary statistics was deemed unnecessary for a visualization that serves as an exploratory tool. Instead, users are already redirected to player profiles on the zerozero website for detailed statistics.

7.3.5 Internal Analytics and User Engagement

Internal analytics from Google Analytics and Hotjar (Section 7.2.1 and 7.2.2, respectively), particularly the number of users, page visits, and sessions, indicate strong user interest in the visualization among ZOS users. Both tools recorded an average interaction time of around two minutes, with enough scrolling activity to view the entire visualization. While the high number of visits and users suggests strong interest, the median average time on the page may indicate average engagement or potential issues, as some users leave the page without further interaction. However, the return visits (shown by the difference between visits and visitors in Google Analytics and the percentage of new vs. returning users recorded in Hotjar) suggest that some users found the visualization compelling enough to return.

The geographical dispersion of accesses, spanning various countries and cities, accurately represents the diverse user base of ZOS. This diversity extends to the technology used to access the site (user devices, browsers, and operational systems), emphasizing the necessity for a functional visualization across multiple platforms. The decision to use D3.js, a versatile and widely compatible web visualization library, supports this need by ensuring compatibility and responsiveness across different environments.

Regarding the most utilized features of the visualization, the nodes related to Qatar 2022 were the most explored, likely because it is the initial World Cup loaded and has recent relevance. This suggests that many users do not extensively manipulate the timeline, which is why the Hotjar heatmap analysis of desktop sessions was limited to this timeslice. Users tend to interact with the visualization in its initial state, as indicated by the high interaction with events in this timeslice. Still, they do not frequently navigate through time or use the controls. The sort and search controls are the least used, indicating that the default settings and overall perspective are sufficient for most users' needs.

Although internal analytics tools were configured to correlate metrics with specific questionnaire submissions using the session ID, no submissions required such detailed correlation.

7.3.6 Final Remarks

To conclude, this section addresses the questions posed in Section 6.1.1, summarizing the objectives of the experiments regarding the visualization. Based on the presented results, the following conclusions can be drawn:

Question 1: Can users clearly and unambiguously understand what is being displayed?

The positive results of task accuracy per question and task performance per participant among a diverse audience regarding age and knowledge in both InfoViz and Sports domains indicate that the visualization is intuitive. Generally, users can understand what is being displayed.

Question 2: Can users easily interact with the visualization and draw more insights from it?

Considering the SUS results, particularly on the questions focusing on ease of use, confidence, and speed, as well as an overall SUS score of 75, the visualization is perceived as having good usability. Users can easily interact with it to derive correct answers to various questions, reinforcing the findings of the first question.

Question 3: Do users feel engaged and spend time exploring the visualization?

Based on the number of visits and returning users, it is evident that users were at least curious about the visualization. The average time spent on the page, the number of distinct events registered, and the heatmaps generated during the experiment period indicate that users were engaged and explored various aspects of the visualization but not extensively. Adherence to the formative questionnaire and positive comments and feedback further support this conclusion.

In summary, the findings suggest that while the visualization effectively presents the information and engages a diverse audience, certain areas, such as providing interpretation guidance and optimizing interaction, particularly filter usage, require further enhancement to improve the overall user experience and system usability.

7.4 Summary

This chapter presents the results of the experiment conducted to evaluate the effectiveness of the animated dynamic weight graph visualization for representing the temporal evolution of players in national teams and clubs within the context of the World Cups. The formative questionnaire provided insights into participant demographics, task performance, and usability statistics. Outliers were identified and removed to ensure the analysis was based on genuine submissions. The analytics tools (Google Analytics and Hotjar) offered additional perspectives on user interactions.

The results indicate a positive reception of the visualization, with participants demonstrating satisfactory task performance and usability feedback. These findings support the hypothesis that graph-based visualizations are well-suited for presenting time-evolving data in a clear and engaging manner, effectively communicating complex sports data to a diverse audience.

Chapter 8

Conclusions

8.1 Final Remarks

This thesis addresses the challenge faced by ZOS, a company dealing with large volumes of complex sports data. The main issue is presenting this information attractively and understandably for sports enthusiasts online. The goal is to transform complex data into a clear story, considering various user profiles and visual literacy levels. The hypothesis suggests that graph-based visualizations are suitable for showcasing interconnected entities with complex relationships while preserving the temporal dimension. The research questions focus on identifying and understanding visualization techniques and tools that align with the specific characteristics of the data under consideration (highly interconnected data with temporal evolution). It also involves exploring graph-based approaches, encompassing their capabilities, limitations, and applications.

The Systematic Literature Review (SLR) critically examined existing approaches for data with similar characteristics, more specifically, graph-based visualizations for the effective communication of complex and dynamic data. The SLR followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology, thoroughly reviewing and synthesizing the most recent relevant documents about hyper-connected, time-dependent data and graph-based visualizations. Key findings emphasize the challenges of representing highly connected and time-evolving data, the importance of maintaining visual consistency, and the critical role of visual storytelling in conveying complex information. The review highlights the growing importance of infographics for presenting dynamic data, particularly in sports and web environments.

The State-of-the-Art analysis constructed a narrative based on the SLR outcomes, offering a comprehensive view of the field of information visualization for highly connected data with temporal evolution. The literature underscores the transformative role of data visualization in decision-making, emphasizing the need for structured storytelling in graph-based visualizations to enhance understanding of intricate relationships and temporal changes. Challenges include managing large datasets and ensuring adequate user cognition. Addressing user perceptions and interactions can lead to more efficient visualizations. A notable gap exists in visualizing soccer

data, particularly player trajectories, which could significantly enrich understanding player dynamics. Overall, the synthesis advocates for tailored visualization strategies to effectively convey complex information, offering a comprehensive guide for researchers.

After carefully analyzing and specifying the problem statement, detailing the solution requirements, and clarifying the scope of the problem, the work proposes an innovative solution: animated-based dynamic weight circular graph visualization. This visualization strategically represents the evolution of players within national teams and clubs by organizing the information using outer and inner circles to represent national teams and clubs, respectively, with edges depicting player connections. Moreover, the solution incorporates clustering and multi-level views, ensuring a user-friendly representation and allowing a detailed inspection of player evolution across different national teams and clubs. The solution design aligns with the insights gained from the Systematic Literature Review and the State-of-the-Art analysis. It also draws inspiration from practical examples, namely the New York Times [47] and ChronoView [3].

A prototype mirroring the proposed solution was implemented using D3, a JavaScript visualization library known for its flexibility in constructing and manipulating dynamic visualizations. This choice meets the concerns regarding integration with ZOS standards and web browser compatibility. The central concept used to employ the graph layouts was the *d3-force* mechanism, which included the manual coding of custom forces to implement boundary constraints and define outer node placement and animation. For timeline integration, D3 logic was essential to configure the transition of nodes based on three scenarios: entering, updating, and exiting nodes between timeslices. Regarding design and usability, the visualization leverages visual variables and additional custom controls to provide more meaningful information. Interaction and layout adaptability for various devices were carefully considered, leading to the development of two types of interaction and two organizational structures based on device characteristics (mouse or touch availability and screen size). Additionally, concerns regarding loading times, performance, and user experience were addressed by adding a cache mechanism to update the data only periodically based on endpoint responses. The visualization is live on a standalone page that integrates contextual elements to match ZOS's image.

An experimental protocol was designed to evaluate the effectiveness of the multiview-weighted graph-based visualization for presenting complex, interconnected data with temporal evolution. Considering the target audience, the experiment was conducted online over approximately two weeks, with participation being completely voluntary and unsupervised. The protocol encompasses two main components: a formative questionnaire and internal analytics. In the questionnaire, participants provided demographic information, including their knowledge in the information visualization and sports domains on a scale from 1 to 7, and answered questions about the graph-based visualization presented. They also completed a System Usability Scale (SUS) form and had the option to leave extra comments or feedback. Additional fields such as start date and time, interaction type, and session ID were included for analysis purposes. The internal analytics tools, Google Analytics and Hotjar, recorded data based on user behavior. Together, these two assessment methods allowed the evaluation of the visualization's comprehension, usability, and

engagement. The protocol and prototype were improved based on findings from a pilot test round conducted before the experiment period.

The results revealed that most users found the visualization easy to interact with and could derive insights from it, indicating good usability and interpretation of the visualization schema. This is supported by high response accuracy across both question and participant levels. Users demonstrated engagement by exploring various aspects of the visualization and frequently returning to the visualization page, as shown by internal analytics data. Furthermore, concrete feedback from the formative questionnaires and comments on social media aligns with a median SUS score of 75, confirming the usability and usefulness of the visualization. The large and diverse participant pool, regarding background and knowledge levels, further supports these findings. Despite these positive results, there are areas for improvement. Introducing a tutorial could help less experienced users better understand and utilize the visualization. Additionally, optimizing interactions by reducing and simplifying filters and enhancing layouts for smaller screens would further enhance the user experience and system usability.

Based on the work performed, the answers to the research questions are as follows:

Q1: Which visualization techniques are commonly used to present highly connected data?

Network visualizations, such as node-link diagrams and matrix-based representations, are commonly used to illustrate relationships and connections between entities effectively.

Q2: Which visualization techniques are commonly used to present temporal data?

Temporal data is often visualized using time series plots, Gantt charts, and animated visualizations. These methods typically rely on a timeline to showcase changes over time, allowing users to track the evolution of data and identify patterns or trends.

Q3: Which visualization techniques are commonly used to present multidimensional and multivariate data?

Multidimensional and multivariate data are frequently visualized through scatter plots, parallel coordinates, and multidimensional scaling. These techniques enable the representation of multiple variables and dimensions in a single visual display. Additionally, multi-level views present the same data from different perspectives in simpler formats, making it easier for diverse audiences to explore multiple viewpoints of the data simultaneously in different coherent and synchronized views.

Q4: What are the limitations and applications of graph-based visualizations for time-evolving data?

Graph-based visualizations are valuable for time-evolving data due to their ability to show dynamic relationships and changes over time. However, they face limitations such as potential visual clutter as the data grows, the complexity of visualizing large graphs, challenges in representing dynamic changes effectively and maintaining user comprehension with evolving data. Applications of graph-based visualizations include network analysis, social network visualization, and dynamic systems modeling. Graph-based visualizations

excel in showing interconnected relationships and capturing complex dependencies. Dynamic graph-based visualizations emerge using structural and temporal encoding to tackle the problem of representing dynamic changes. Techniques like force-directed graphs help maintain node placement consistency between timeslices, making identifying changes in dynamic visualizations easier. Additionally, multiviews, clustering techniques, and interactive graph manipulation allow users to select specific data, reducing visual clutter and enhancing comprehension of evolving data trends.

After addressing the research questions, the hypothesis that *"Graph-based visualizations for time-evolving data are well suited for presenting highly connected entities with complex relationships while preserving the story within the visualization"* can now be evaluated. The graph-based visualizations effectively present intricate dynamic information in an engaging and interpretable format, achieved through structural and temporal encodings. The premise is that presenting the data in simpler, consistent graphs enhances user comprehension and perception while maintaining the storytelling capability of the visualization when compared to presenting the data only in one more complex and difficult-to-read graph. Additionally, animations play a crucial role in enhancing the visualization's storytelling aspects, facilitating data comparisons across different time points.

The animated-based dynamic weight circular graph visualization, incorporating clustering and multi-level views, effectively captures the complexity, connectivity, and temporal dimensions of historical and highly connected sports data. This supports the validation of the hypothesis and underscores the ability of graph-based visualization for time-evolving data to convey comprehensive insights into complex dynamic data.

8.2 Future Work

Although the research questions have been addressed and the proposed hypotheses validated, there are several topics to be further explored and tested, particularly concerning the developed prototype. Insights gathered from the experiments suggest key areas for enhancing visualization, improving user experience and system usability.

One area of focus involves enhancing the visibility of less noticeable elements in the visualization. This involves redesigning information cards and using a placeholder to indicate the space reserved for the details graph to increase their appeal and visual emphasis. Moreover, to address the first issue, it could be beneficial to explore displaying the number of players involved in a highlighted network directly on the overview graph (e.g., as a badge over the highlighted nodes). However, caution is needed to avoid visual clutter that may hinder interpretation when incorporating additional elements.

Another critical aspect is optimizing filter usage. This involves reducing the number of controls available and implementing more intelligent filters that dynamically adjust choices based on previously selected values. For instance, filters related to countries and continents could be interconnected, as suggested by a participant. Additionally, incorporating a reset button for filters

and indicating active filters on the page, possibly through a prominent title, would significantly enhance usability.

Improving usability on mobile devices is also essential. This includes increasing the text size of node labels for better readability and exploring zooming options in the visualization for easier comparison of node sizes. Addressing device misidentification issues, where mobile devices are mistakenly identified as laptops, leading to the selection of the inappropriate interaction type, is also crucial for improving usability.

Further enhancements involve testing animations alongside visual outlines to highlight nodes during player searches. Moreover, enhancing the visualization's flexibility and adaptability to different use cases is crucial to improve its versatility and utility for ZOS. This encompasses the calculus of the node and link size based on a more complex mathematical expression to account for higher contribution values, adjusting force strength values to smooth transitions involving nodes or edges with significant contributions, and allowing players to be associated with multiple clusters and outer nodes within the same timeslice.

Incorporating a skippable tutorial to guide users through the visualization's features and extending language support to comprise more ZOS domains enhances accessibility and user understanding.

Future versions of the experimental protocol should consider using a smaller yet diverse participant pool in a controlled environment. This approach would allow for measuring task response times using the visualization and comparing them to the times required to complete the same tasks on the zerozero website. This method assesses the visualization's efficiency relative to existing frameworks. Considering participants' prior knowledge and familiarity with the framework is essential.

References

- [1] M. Abdelaal, A. Lhuillier, M. Hlawatsch, and D. Weiskopf. Time-Aligned Edge Plots for Dynamic Graph Visualization. In *Proc. 2020 24th International Conf. Information Visualization*, pages 248–257, 2020.
- [2] M. Ali, M. W. Jones, X. H. Xie, and M. Williams. TimeCluster: dimension reduction applied to temporal data for visual analytics. *Visual Computer*, 35(6-8):1013–1026, 2019.
- [3] Y. Anzai and K. Misue. Integration of ChronoView with pseudo MDS for Visualization of Temporal Data. IEEE, 2018.
- [4] A. M. Bartkowiak and R. Zimroz. Complementary View on Multivariate Data Structure Based on Kohonen’s SOM, Parallel Coordinates and t-SNE Methods. volume 9, pages 255–265, 2018.
- [5] R. C. Basole, A. Srinivasan, H. Park, and S. Patel. ecoxight: Discovery, Exploration, and Analysis of Business Ecosystems Using Interactive Visualization. *ACM Trans. Manage. Inf. Syst.*, 9(2):Article 6, 2018.
- [6] A. Bobic, J. M. Le Goff, and C. Gütl. Exploring Tabular Data Through Networks. In *Lecture Notes in Computer Science*, volume 13982 LNCS, pages 195–200, 2023.
- [7] M. Burch. Exploring density regions for analyzing dynamic graph data. *Journal of Visual Languages and Computing*, 44:133–144, 2018.
- [8] M. Burch, K. B. Ten Brinke, A. Castella, G. K. S. Peters, V. Shteriyarov, and R. Vlasvinkel. Dynamic graph exploration by interactively linked node-link diagrams and matrix visualizations. *Visual Computing for Industry Biomedicine and Art*, 4(1), 2021.
- [9] M. Burch, G. Wallner, H. Van De Wetering, F. Rooks, and O. Morra. Visual Analysis of Graph Algorithm Dynamics. In *ACM International Conference Proceeding Series*, 2021.
- [10] C. Spearman. Spearman Rank Correlation Coefficient. In *The Concise Encyclopedia of Statistics*, pages 502–505. Springer New York, 2008.
- [11] E. Cakmak, U. Schlegel, D. Jackle, D. Keim, and T. Schreck. Multiscale Snapshots: Visual Analysis of Temporal Summaries in Dynamic Graphs. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):517–527, 2021.
- [12] S. H. Cheong, Y. W. Si, and R. K. Wong. Online force-directed algorithms for visualization of dynamic graphs. *Information Sciences*, 556:223–255, 2021.
- [13] Contentsquare. Hotjar, 2024. Available at <https://www.hotjar.com/pt-BR/>, last accessed in May 2024.

- [14] A. Cruz, J. P. Arrais, and P. Machado. Multivariate Data Exploration Through Coordinated Views. *IEEE Access*, 10:78788–78817, 2022.
- [15] Design Beautiful Forms. Pretty Forms Designer, 2024. Available at <https://www.addxt.com/pretty-forms-designer>, last accessed in April 2024.
- [16] W. Didimo, L. Grilli, G. Liotta, F. Montecchiani, and D. Pagliuca. Visual querying and analysis of temporal fiscal networks. *Information Sciences*, 505:406–421, 2019.
- [17] E. Dimara, A. Bezerianos, and P. Dragicevic. Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):749–759, 2018.
- [18] A. G. Forbes et al. Dynamic Influence Networks for Rule-based Models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):184–194, 2018.
- [19] E. Thouverai et al. Helical graphs to visualize the NDVI temporal variation of forest vegetation in an open source space. *Ecological Informatics*, 74, 2023.
- [20] M. Burch et al. IMDb Explorer: Visual Exploration of a Movie Database. In *Proc. 11th International Symposium on Visual Information Communication and Interaction*, page 88–91, 2018.
- [21] S. M. Chen et al. Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling. *IEEE Transactions on Visualization and Computer Graphics*, 26(7):2499–2516, 2020.
- [22] S. Murugesan et al. State-based network similarity visualization. *Information Visualization*, 19(2):96–113, 2020.
- [23] X. L. Wen et al. DiffSeer: Difference-Based Dynamic Weighted Graph Visualization. *IEEE Computer Graphics and Applications*, 43(3):12–23, 2023.
- [24] Y. C. Wu et al. ForVizor: Visualizing Spatio-Temporal Team Formations in Soccer. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):65–75, 2019.
- [25] Y. Peng et al. Visual abstraction of dynamic network via improved multi-class blue noise sampling. *Frontiers of Computer Science*, 17(1), 2023.
- [26] Y. Wang et al. Structure-aware Fisheye Views for Efficient Large Graph Exploration. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):566–575, 2019.
- [27] M. Fatehah, V. Mezhuyev, and M. Al-Emran. A Systematic Review of Metamodelling in Software Engineering. *Recent Advances in Intelligent Systems and Smart Applications*, 295:3–27, 2020.
- [28] V. Filipov, A. Arleo, M. Bögl, and S. Miksch. On Time and Space: An Experimental Study on Graph Structural and Temporal Encodings. In *Graph Drawing and Network Visualization*, volume 13764, pages 271–288, 2023.
- [29] E. Di Giacomo, W. Didimo, M. Kaufmann, and G. Liotta. Stable visualization of connected components in dynamic graphs. *Information Visualization*, 20(1):3–19, 2021.

- [30] Google. Analytics Help, 2024. Available at <https://support.google.com/analytics/?hl=en&sjid=15034663517637837540-EU#topic=14090456>, last accessed in May 2024.
- [31] Google. Google Forms, 2024. Available at <https://www.google.com/forms/about/>, last accessed in May 2024.
- [32] Google for Developers. Google tag API Reference | Google tag (gtag.js), 2023. Available at <https://developers.google.com/tag-platform/gtagjs/reference>, last accessed in May 2024.
- [33] Google for Developers. Apps Script, 2024. Available at <https://developers.google.com/apps-script>, last accessed in April 2024.
- [34] Google for Developers. Introduction to Google Analytics 4, 2024. Available at <https://developers.google.com/analytics/devguides/collection/ga4>, last accessed in May 2024.
- [35] Google Marketing Platform. Google Analytics, 2024. Available at <https://marketingplatform.google.com/about/analytics/>, last accessed in May 2024.
- [36] D. Ivankovic, V. Dadic, L. Seric, and A. Ivanda. Web Based Spatio-Temporal Data Bidirectional Relationship Visualization-A Case Study of Oceanographic Data. *Applied Sciences-Basel*, 12(13), 2022.
- [37] X. Ji and Z. Wang. Research on Teaming Strategies Based on Network Science. In *Proc. 2022 International Conference on Electronics and Devices*, pages 209–213.
- [38] A. Komarek, J. Pavlik, and V. Sobeslav. Time-Based Change Visualization Techniques Survey. *ARPJ Journal of Engineering and Applied Sciences*, 14(Special Issue 9):10538–10543, 2019.
- [39] C. D. G. Linhares, J. R. Ponciano, D. S. Pedro, L. E. C. Rocha, A. J. M. Traina, and J. Poco. LargeNetVis: Visual Exploration of Large Temporal Networks Based on Community Taxonomies. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):203–213, 2023.
- [40] Mailrecipe LLC. Formfacade - Customize UI, 2024. Available at <https://formfacade.com/website/>, last accessed in April 2024.
- [41] Mailrecipe LLC. Formsign - Formprefill, 2024. Available at <https://formesign.com/prefill/>, last accessed in April 2024.
- [42] Maze. Maze - Enrich product decisions with intuitive user research, 2024. Available at <https://maze.co/>, last accessed in May 2024.
- [43] Mike Bostock and Observable, Inc. D3 - The JavaScript library for bespoke data visualization, 2014. Available at <https://d3js.org/>, last accessed in April 2024.
- [44] M. Mokaddem, I. Idris Khodja, H. Amar Setti, B. Atmani, and C. E. Mokaddem. Communities Detection in Epidemiology: Evolutionary Algorithms Based Approaches Visualization. In *Lecture Notes in Networks and Systems*, volume 593, pages 319–332, 2023.

- [45] N. Nachar. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, 4, 2008.
- [46] H. Nurrahmi, R. Wijayanti, A. F. Rozie, and A. Arisal. Twitter data transformation for network visualization based context analysis. In *2018 International Conf. on Information and Communications Technology*, pages 525–530.
- [47] NYT. The Clubs that connect the World Cup, June 2014. Available at <https://www.nytimes.com/interactive/2014/06/20/sports/worldcup/how-world-cup-p-layers-are-connected.html>, last accessed in January 2024.
- [48] I. Portugal, P. Alencar, and D. Cowan. A Graph-based Analysis Approach to Cluster Lifetime Dynamics. In *Proc. 2022 IEEE International Conf. on Big Data*, pages 3831–3837.
- [49] PRISMA. PRISMA - transparent reporting of systematic reviews and meta-analyses, 2023. Available at <http://www.prisma-statement.org/>, last accessed in October 2023.
- [50] R. C. Sprinthall. *Basic Statistical Analysis*. Pearson, 2011.
- [51] K. V. Ryabinin, K. I. Belousov, and S. I. Chuprina. Novel circular graph capabilities for comprehensive visual analytics of interconnected data in digital humanities. *Scientific Visualization*, 12(4):56–70, 2020.
- [52] M. Ryoo, N. Kim, and K. Park. Visual analysis of soccer players and a team. *Multimedia Tools and Applications*, 77(12):15603–15623, 2018.
- [53] S. Tselova. 7 creative data visualization ideas for your World Cup coverage - Visualizing FIFA World Cup stats, from match history to goals and player info, November 2022. Available at <https://flourish.studio/blog/visualizing-world-cup-2022/>, last accessed in January 2024.
- [54] S. M. Shaheen, S. Alhalawani, N. Alnabet, and D. Alhenaki. Analytical Experiments on the Utilization of Data Visualizations. In *Communications in Computer and Information Science*, volume 1097 CCIS, pages 148–161, 2019.
- [55] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. 1996 IEEE Symposium on Visual Languages*, pages 336–343.
- [56] A. J. Slopek, C. Winkelholz, and M. Varga. Trade-Off Between Mental Map and Aesthetic Criteria in Simulated Annealing Based Graph Layout Algorithms. volume 10904, pages 110–125, 2018.
- [57] Tableau. Sports + Data Viz Gallery. Available at <https://www.tableau.com/sports-data>, last accessed in January 2024.
- [58] S. Tandon, A. Abdul-Rahman, and R. Borgo. Measuring Effects of Spatial Visualization and Domain on Visualization Task Performance: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):668–678, 2023.
- [59] UIUX Trend. Measuring and Interpreting System Usability Scale (SUS), 2024. Available at <https://uiuxtrend.com/measuring-system-usability-scale-sus/>, last accessed in June 2024.

- [60] M. S. Ulizko, E. V. Antonov, A. A. Artamonov, and R. R. Tukumbetova. Visualization of graph-based representations for analyzing related multidimensional objects. *Scientific Visualization*, 12(4):133–142, 2020.
- [61] Vasco Asturiano. GitHub - vasturiano/force-graph: Force-directed graph rendered on HTML5 canvas, 2014. Available at <https://github.com/vasturiano/force-graph>, last accessed in February 2024.
- [62] Y. Z. Wang, G. Baciú, and C. H. Li. Visualizing the Temporal Similarity Between Clusters of Dynamic Graphs. In *Proc. 2019 IEEE 18th International Conf. on Cognitive Informatics*, pages 205–210.
- [63] K. Watanabe, N. Sakamoto, J. Nonaka, Y. Maejima, and I. C. Soc. Angular-based Edge Bundled Parallel Coordinates Plot for the Visual Analysis of Large Ensemble Simulation Data. In *Proc. 2022 IEEE 12th Symposium on Large Data Analysis and Visualization*, pages 5–14.
- [64] Y. Wei, X. Du, D. Xu, and X. Wang. A speedup spatial rearrangement algorithm for dynamic network visualization. In *Proc. 2018 IEEE 3rd International Conference on Data Science in Cyberspace*, pages 749–753.
- [65] Z. Yang, Y. Q. Xie, M. Li, and G. Q. Huang. GazeGraphVis: Visual analytics of gaze behaviors at multiple graph levels for path tracing tasks. *Advanced Engineering Informatics*, 57, 2023.
- [66] Z. H. Zhu, Y. Shen, S. J. Zhu, G. F. Zhang, R. H. Liang, and G. D. Sun. Towards better pattern enhancement in temporal evolving set visualization. *Journal of Visualization*, 2022.
- [67] J. Zong, J. Pollock, D. Wootton, and A. Satyanarayan. Animated Vega-Lite: Unifying Animation with a Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):149–159, 2023.
- [68] ZOS. ZOS, Lda, 2023. Available at <https://www.zos.pt/>, last accessed in October 2023.

Appendix A

Prototype Code and Resources

This appendix provides access to the prototype resources, including the source code, the Docker image of the prototype's version used in the experiments phase, and the online version.

- **Prototype Source Code:** <https://github.com/Pdoliveira21/thesis-hub/tree/d4a76b6317c30e6d8c0778a0daed5789c31ddfc1/prototype>.
- **Prototype Docker Image:** <https://hub.docker.com/repository/docker/pdoliveira21/clubs-in-nations/general>.
- **Online Prototype Version:** <https://clubs-in-nations.zerozero.pt/>.

Appendix B

Formative Questionnaire

This appendix presents the figures of the formative questionnaire used in the experimental procedure from the participants' perspective. Originally in Portuguese, the questionnaire was automatically translated into English using browser capabilities to facilitate its interpretation in this document.

Clubs in Nations - Evaluation Survey

Welcome to the "Clubs in Nations" Evaluation Survey.

The **purpose** of this test is to evaluate the accuracy, clarity, effectiveness, and ease of use of a multi-view weighted graph-based visualization. This visualization aims to convey complex and highly interconnected data with temporal evolution. As a test case, it focuses specifically on the dynamics of players within clubs and national teams over time.

By proceeding, you voluntarily agree to participate in this questionnaire and consent to providing the necessary information. You understand that your responses will be used for research and improvement purposes. You are aware that your participation is completely voluntary and that you may withdraw at any time without providing a reason.

up201905427@g.uporto.pt [Switch account](#)

Not shared

* Indicates required questions

Do you agree to provide demographic information and details about your experience interpreting representations for the purpose of this experiment? *

☐ I agree to the terms.

Are you willing to participate in the experience, understanding that your responses and interactions with the visualization will be recorded and analyzed for research purposes? *

☐ I agree to the terms.

Do you agree to share your subjective evaluation of the usability of the visualization using the System Usability Scale (SUS), acknowledging that your feedback will be used for research and improvement purposes? *

☐ I agree to the terms.

[Next](#) [Clear form](#) Page 1 of 4

Figure B.1: Evaluation survey - Informed Consent Form.

[1] Pre-Test Questionnaire

In this questionnaire, we will only collect basic information about you for statistical purposes. All information provided will remain anonymous and will not be sold or shared with third parties.

Please DO NOT change fields that are already auto-populated!

Date you started the questionnaire

Do not change! Field automatically filled in by the system for statistical purposes.

11/04/2024, 15:52:53

Gender *

☐ Man

☐ Woman

☐ Prefer not to say

☐ Others: _____

Age *

Your answer _____

How would you rate your experience and/or knowledge in the field of **data visualization** ? *

1 2 3 4 5 6 7

I have no knowledge or experience on the subject. ☐ ☐ ☐ ☐ ☐ ☐ ☐ I have extensive knowledge and experience on the subject.

How would you rate your experience and/or knowledge in the field of **sports** , particularly **football** ? *

1 2 3 4 5 6 7

I have no knowledge or experience on the subject. ☐ ☐ ☐ ☐ ☐ ☐ ☐ I have extensive knowledge and experience on the subject.

Session identifier

Do not change! Field automatically filled in by the system for statistical purposes.

ati8tppqh9u3som0c62ecan76h

Type of interaction with the view

Do not change! Field automatically filled in by the system for statistical purposes.

movimento do rato

Back

Next

Page 2 of 4

Clear form

Figure B.2: Evaluation survey - Pre-test Questionnaire.

[2] Tasks and task-specific quiz

If the preview is not already open, please open it in a new tab using the following [link](#). **The purpose of this quiz is to interact with the visualization to discover the correct answers to the following tasks**. You can take some time to explore the visualization before answering the questions.

Keep in mind that:

- A "guide" or "tutorial" on how to interpret the visualization will not be provided to ensure an unbiased assessment of the visualization's ease of use, intuitiveness, and effectiveness. You must interpret the visualization yourself.
- The scoring criteria for the tasks in the quiz are simple: 1 for correct answers, 0 for incorrect answers.
- Tasks and questions are intentionally constructed with different levels of complexity, progressing from lower to higher levels, to suit the target audience. Don't feel discouraged if you can't answer something. This is also valid feedback!

T1 - In the 2022 Qatar World Cup, which club contributed the most players to the national teams?

Choose ▼

T2 - In the 2022 Qatar World Cup, how many national teams did Real Madrid supply with players?

Choose ▼

T3 - In the 2022 Qatar World Cup, from which club did the German national team receive the most players?

Choose ▼

T4 - In the 2022 Qatar World Cup, for which national team did Manuel Akanji play, knowing he was playing for Manchester City?

Choose ▼

Figure B.3: Evaluation survey - Task-specific Questionnaire (Part 1).

T5 - How many different clubs did Cafu play for, considering his participation in the World Cups?

Choose ▼

T6 - In the 1994 USA World Cup, how many forwards did the Spanish national team receive from Barcelona?

Choose ▼

T7 - Which was the first World Cup to feature a national team from Oceania?

Choose ▼

T8 - Considering only the 13 World Cups in which a Portuguese club contributed players, which club was always present, always providing at least one player to a national team?

Choose ▼

T9 - Analyzing the journey of players in national teams and clubs over time, from the first World Cup (Uruguay 1930) to the last (Qatar 2022), what phenomenon can be observed?

Choose ▼

Did these tasks help you discover visualization features that you didn't know about? If so, which ones?

Your answer

Back

Next

Page 3 of 4

Clear form

Figure B.4: Evaluation survey - Task-specific Questionnaire (Part 2).

[3] Post-Test Questionnaire

Your answers to this questionnaire will help us understand how well the visualization meets your expectations and how user-friendly it is. Please approach this test with an open mind, providing honest and constructive feedback on the usability of the visualization.

Please rate your agreement with the following statements about the usability of the system on a scale from 'Strongly Disagree' to 'Strongly Agree'. *

	Totally Disagree	I disagree	Neither Agree Nor Disagree	I agree	I totally agree
I think I would like to use this system.	☐	☐	☐	☐	☐
I found the system unnecessarily complex.	☐	☐	☐	☐	☐
I thought the system was easy to use.	☐	☐	☐	☐	☐
I think I would need the support of a technical person to be able to use this system.	☐	☐	☐	☐	☐
I found the various functions in the system to be well integrated.	☐	☐	☐	☐	☐
I thought there was a lot of inconsistency in the system.	☐	☐	☐	☐	☐
I would imagine that most people would learn to use this system very quickly.	☐	☐	☐	☐	☐
I found the system very cumbersome to use.	☐	☐	☐	☐	☐
I felt very confident using the system.	☐	☐	☐	☐	☐
I needed to learn many things before I could use this system.	☐	☐	☐	☐	☐

Do you have any additional comments or feedback about the preview? Feel free to share with us, any feedback is greatly appreciated!

Your answer

Back
Submit

Page 4 of 4
Clear form

Figure B.5: Evaluation survey - Post-test Questionnaire.

Appendix C

Supplementary Results

This appendix provides additional charts and tables that support and elaborate on the findings presented in the results section.

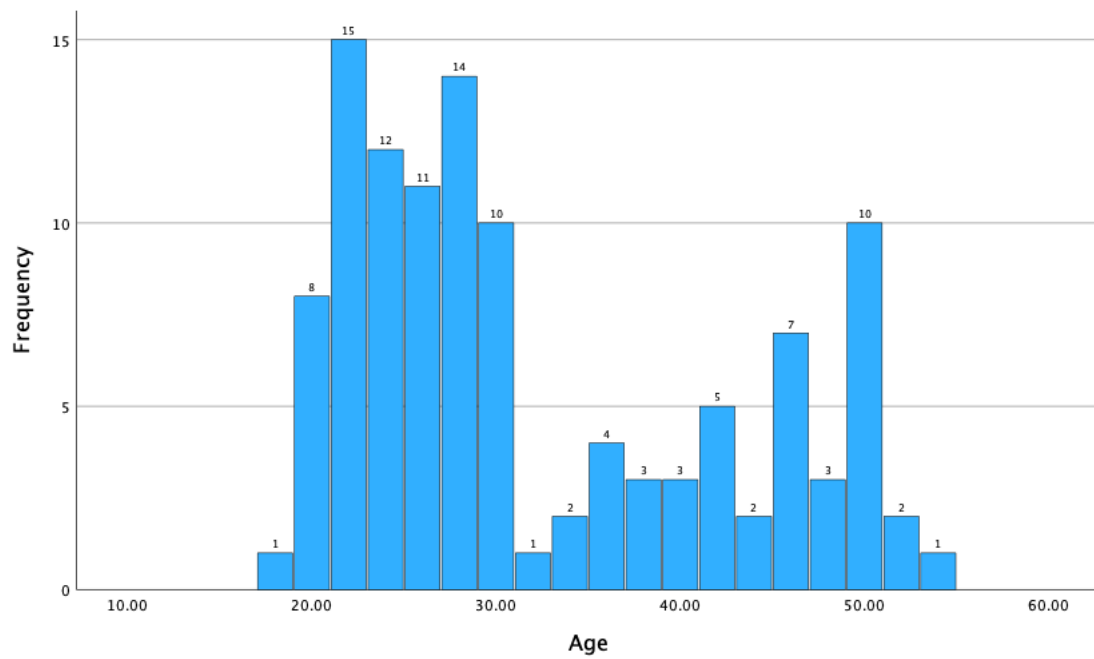


Figure C.1: Histogram of the distribution of age among participants.

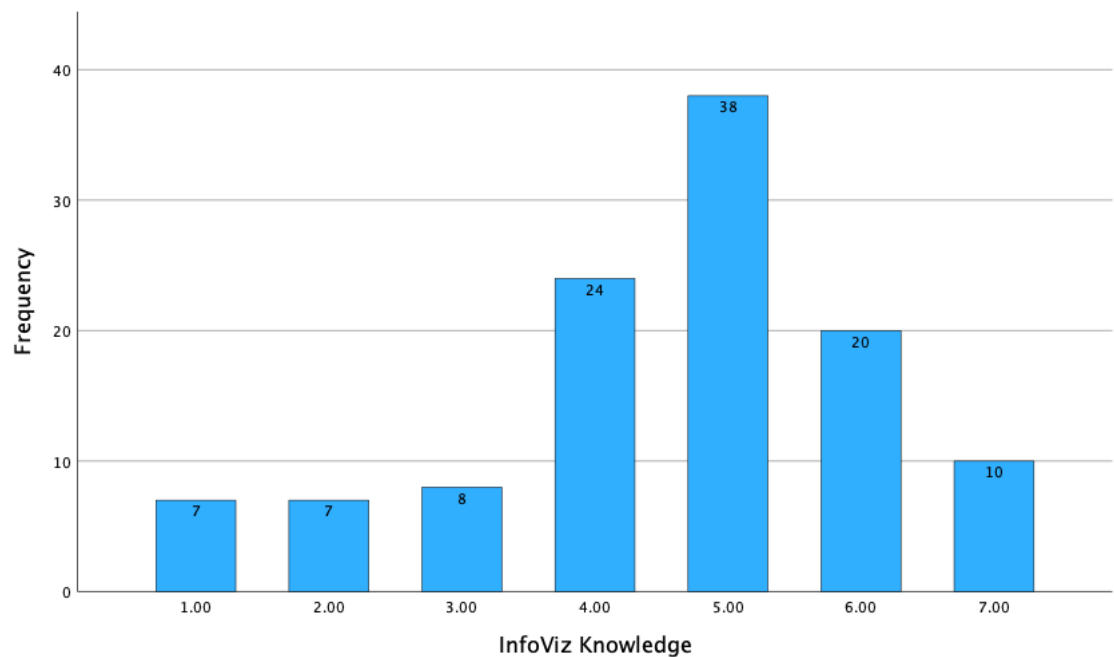


Figure C.2: Bar chart of the distribution of information visualization knowledge among participants.

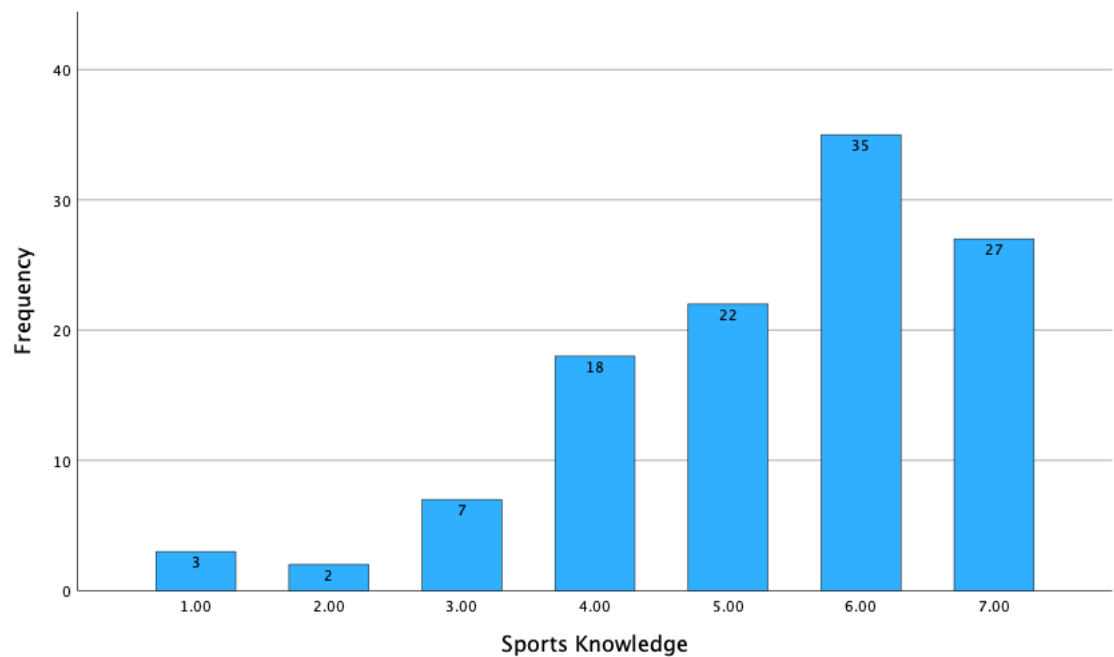


Figure C.3: Bar chart of the distribution of sports knowledge among participants.

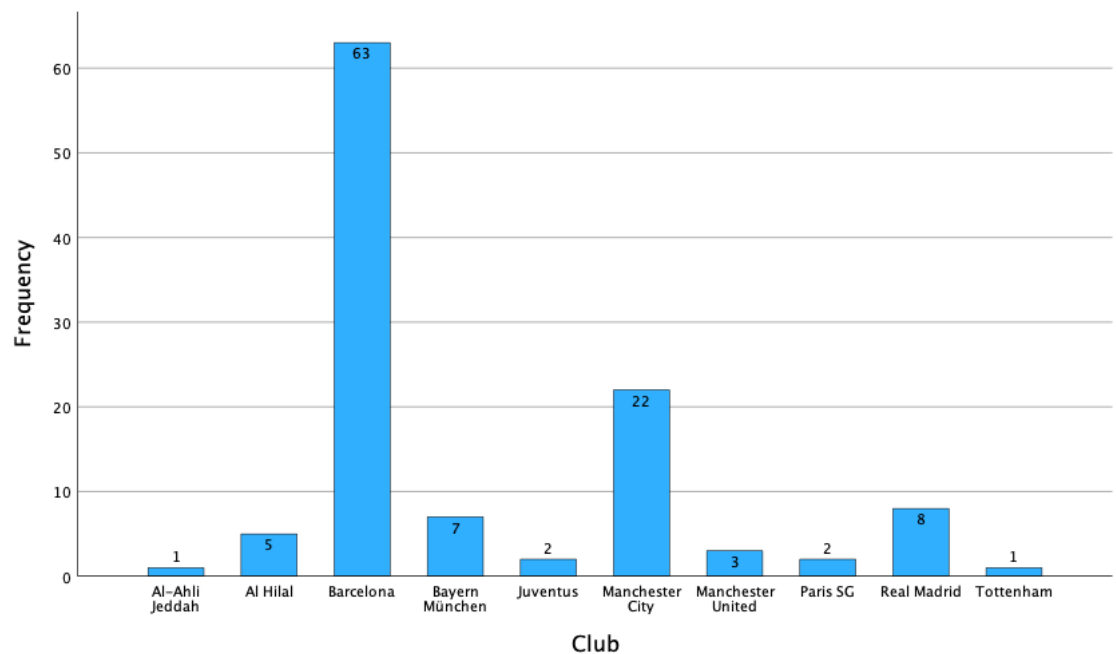


Figure C.4: Bar chart of the distribution of responses to question T1 (Task 1) among participants.

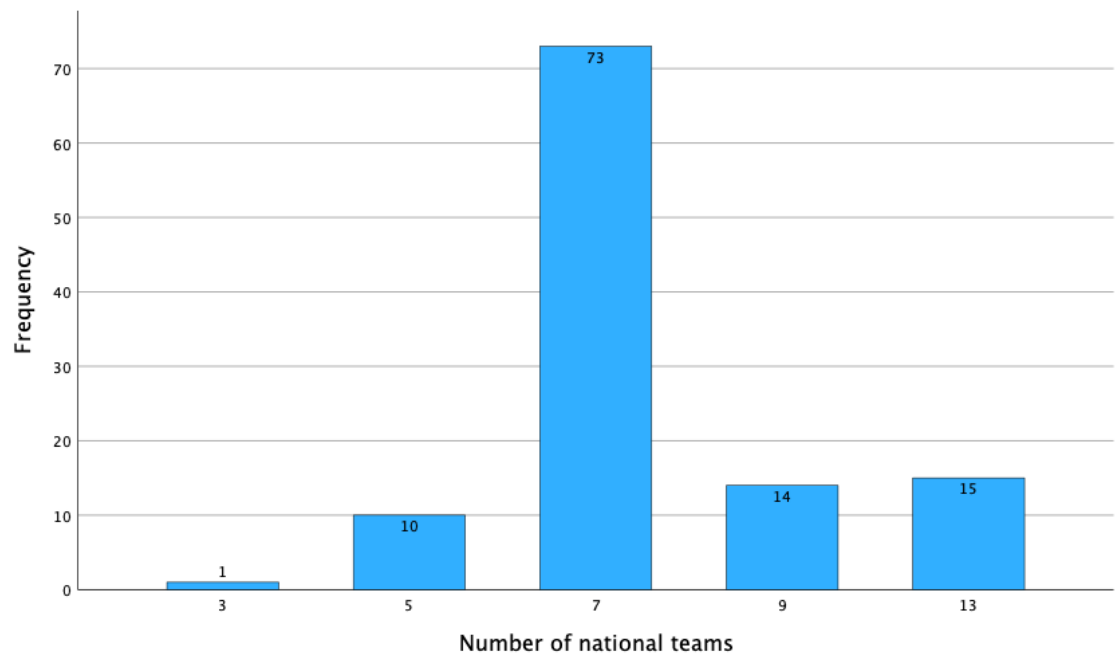


Figure C.5: Bar chart of the distribution of responses to question T2 (Task 2) among participants.

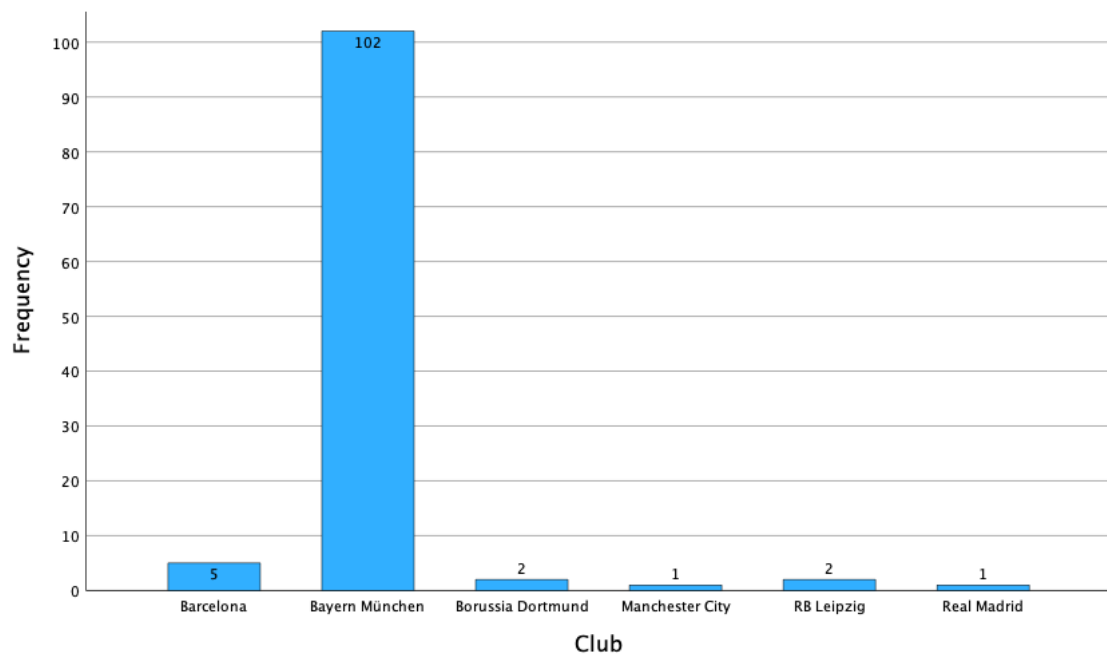


Figure C.6: Bar chart of the distribution of responses to question T3 (Task 3) among participants.

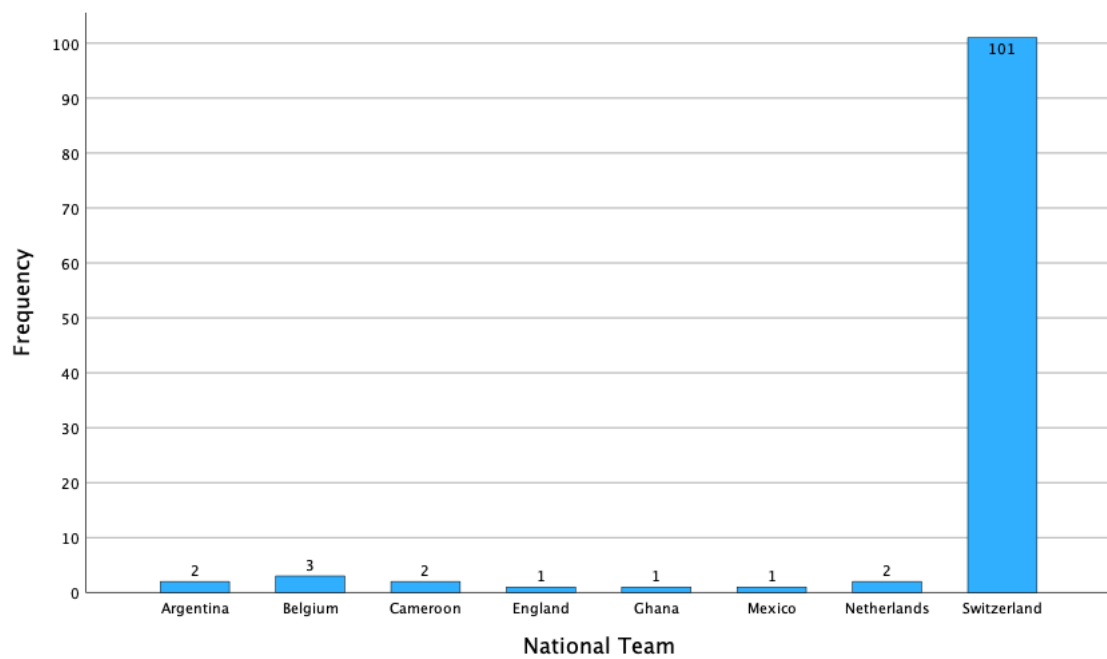


Figure C.7: Bar chart of the distribution of responses to question T4 (Task 4) among participants.

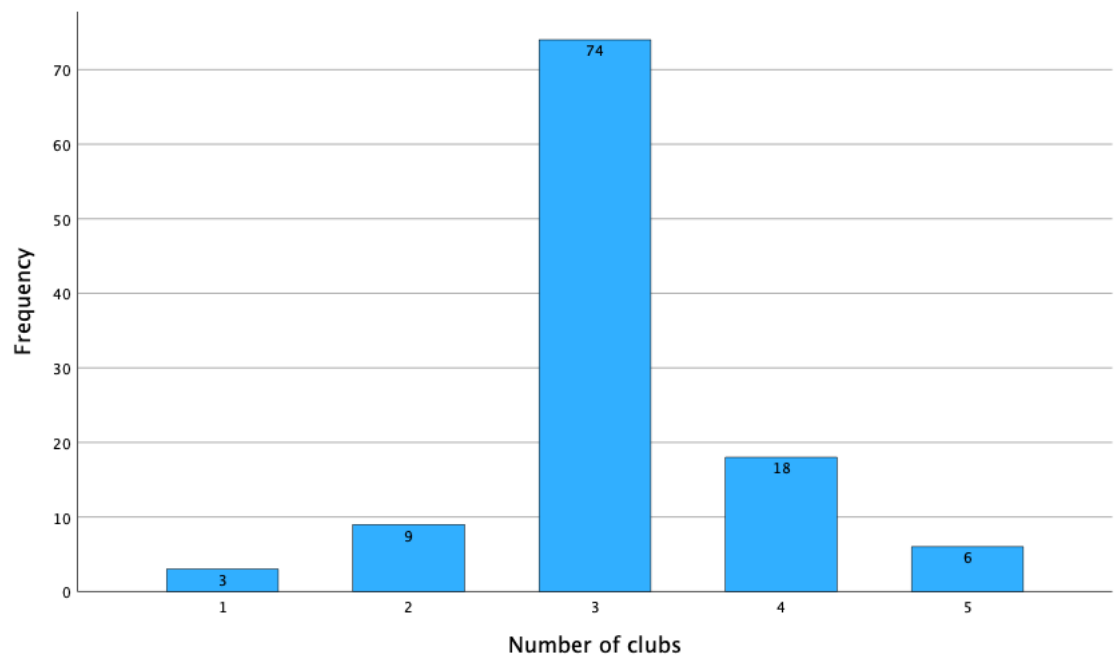


Figure C.8: Bar chart of the distribution of responses to question T5 (Task 5) among participants.

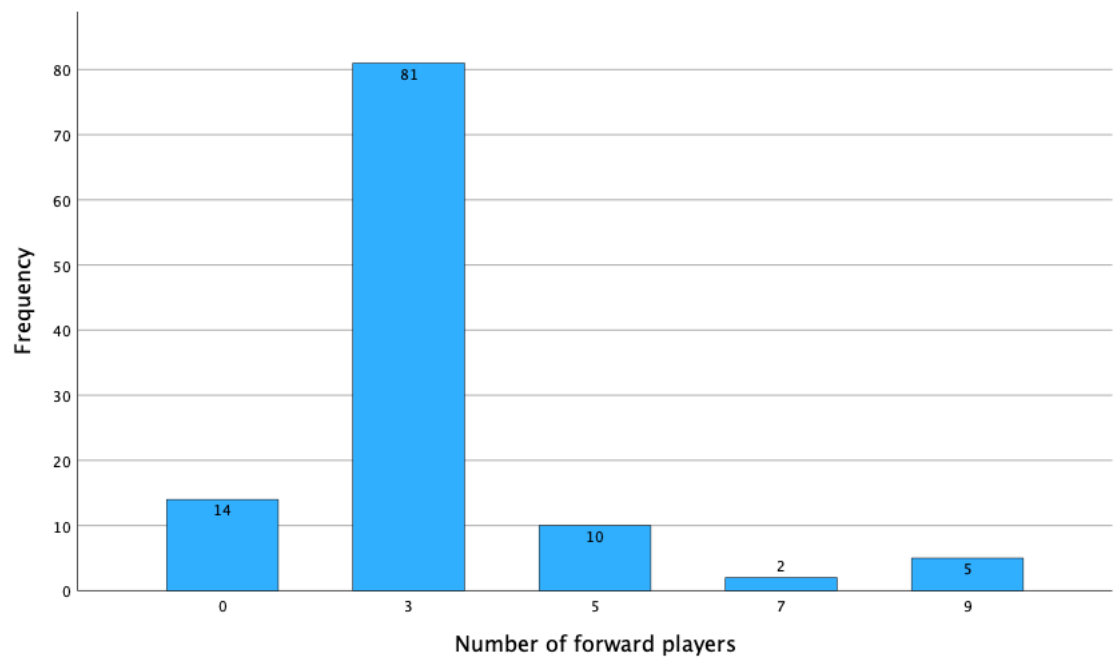


Figure C.9: Bar chart of the distribution of responses to question T6 (Task 6) among participants.

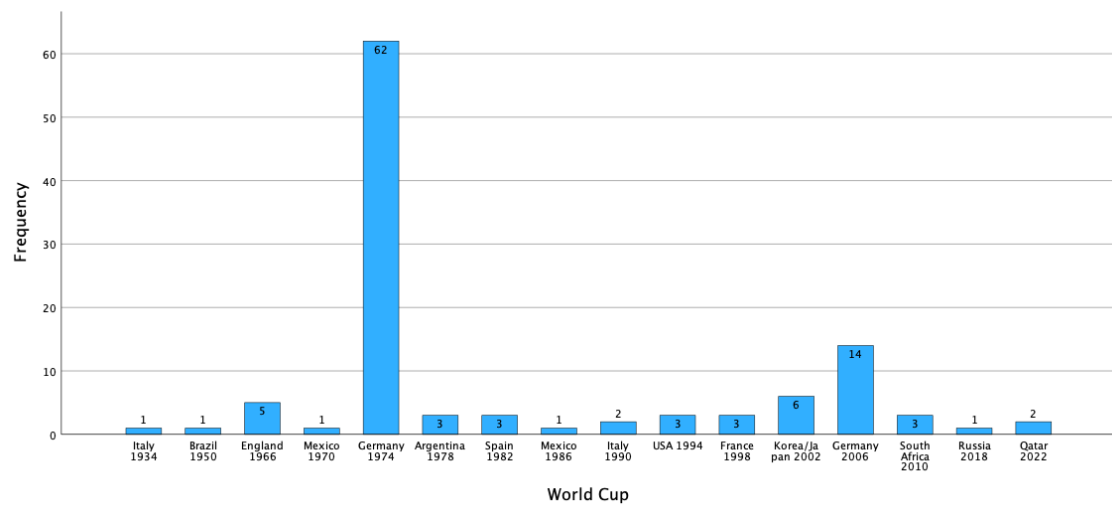


Figure C.10: Bar chart of the distribution of responses to question T7 (Task 7) among participants.

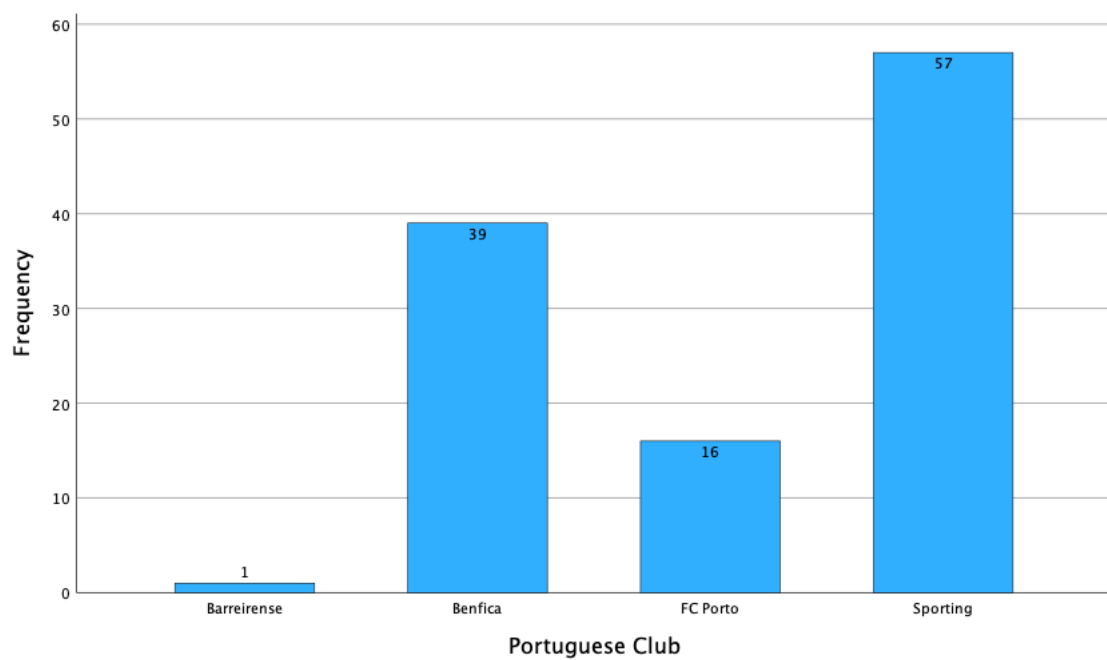


Figure C.11: Bar chart of the distribution of responses to question T8 (Task 8) among participants.

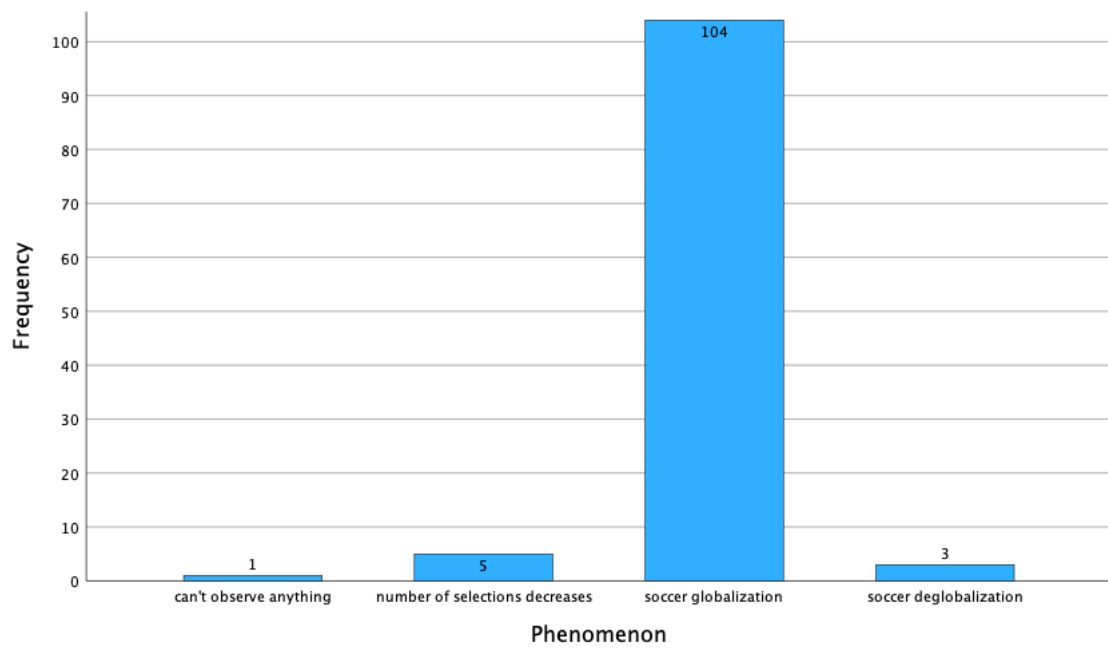
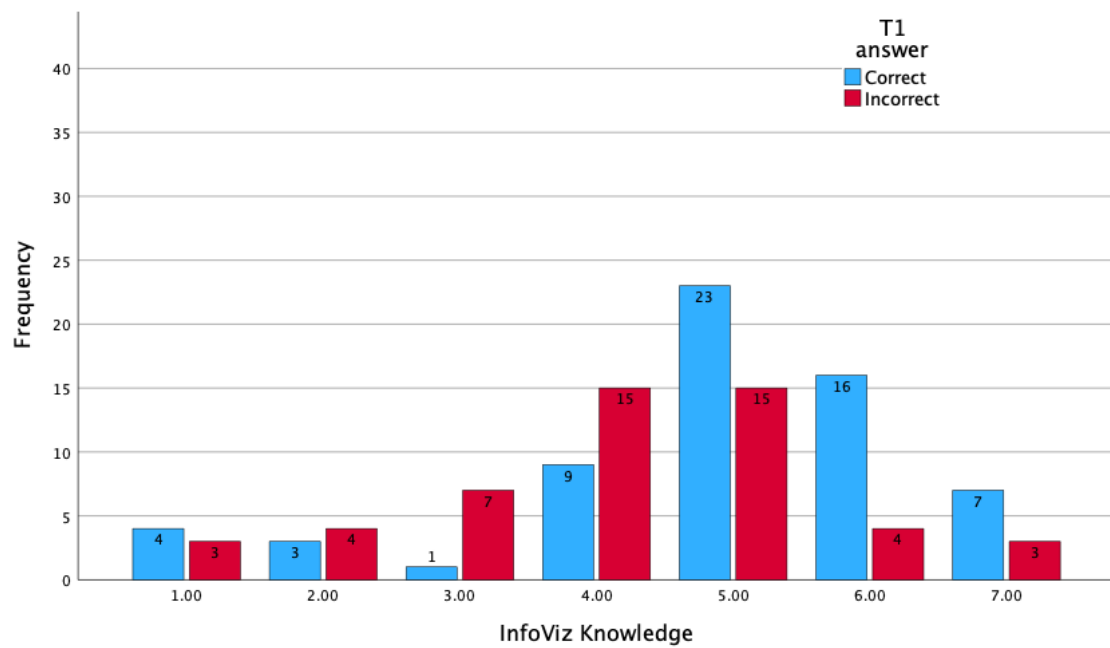
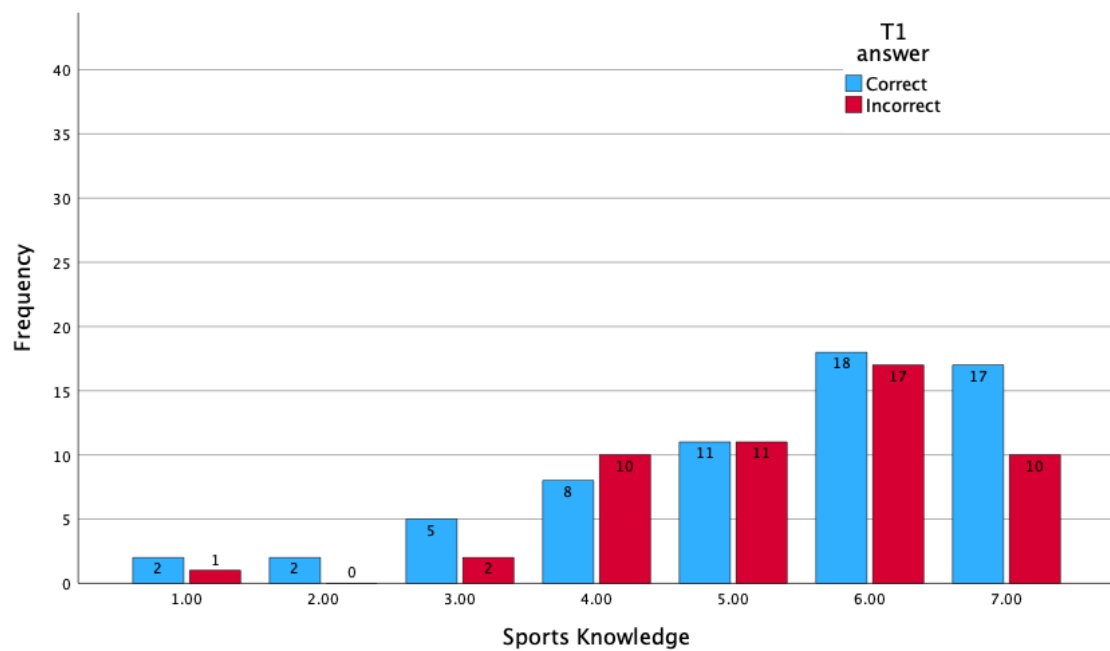


Figure C.12: Bar chart of the distribution of responses to question T9 (Task 9) among participants.

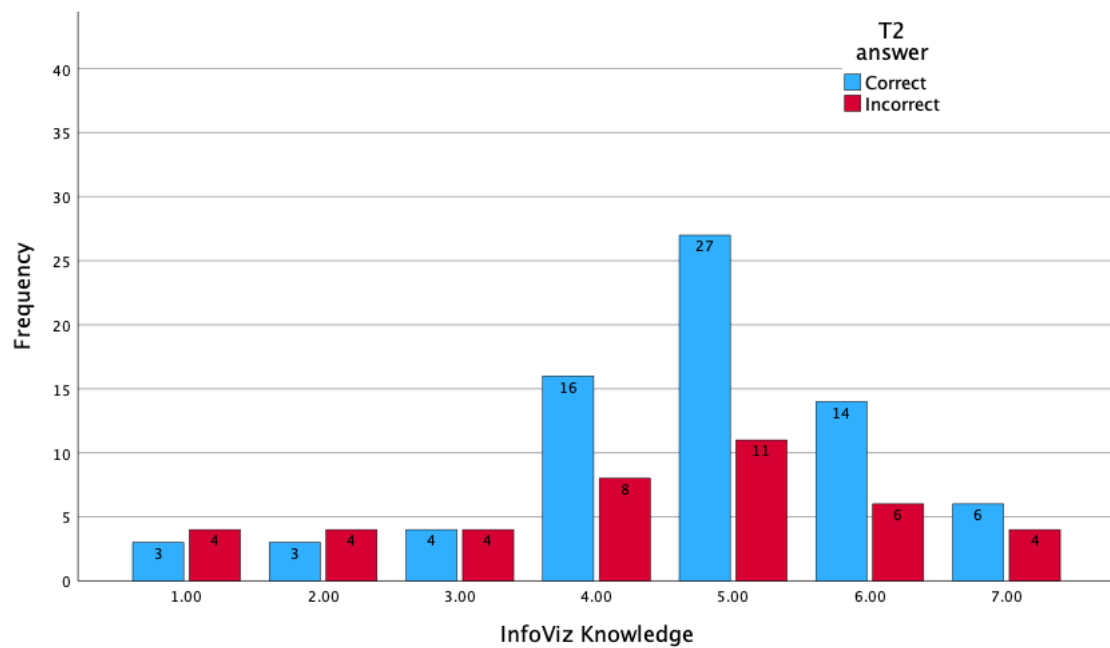


(a) Correct and incorrect answers per information visualization knowledge to question T1 (Task 1).

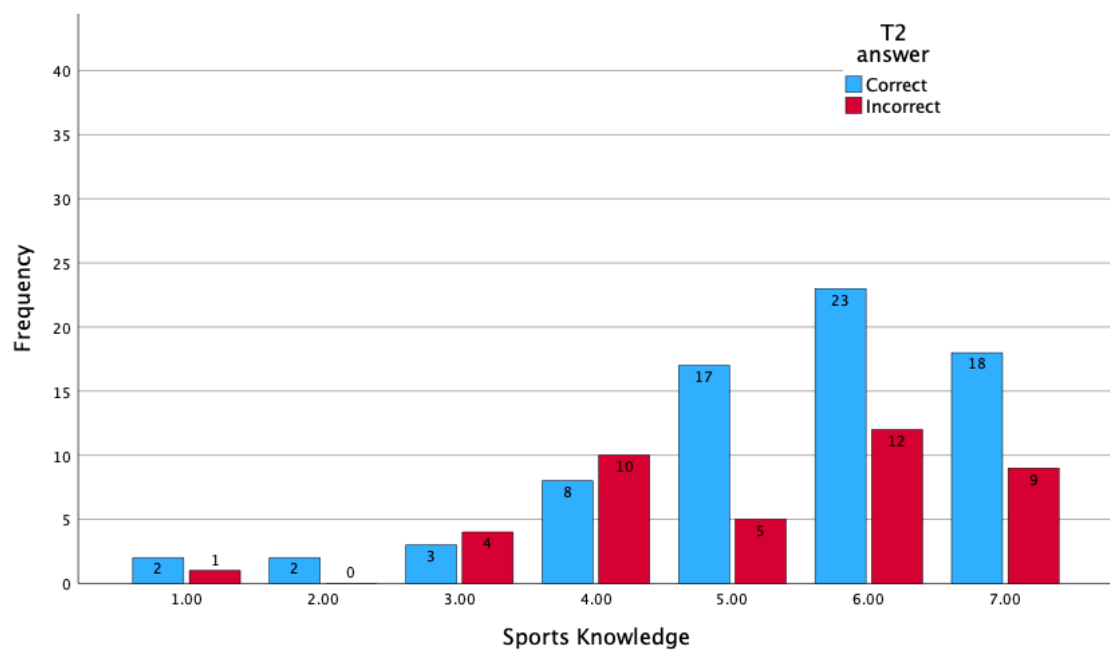


(b) Correct and incorrect answers per sports knowledge to question T1 (Task 1).

Figure C.13: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T1 (Task 1) among participants.

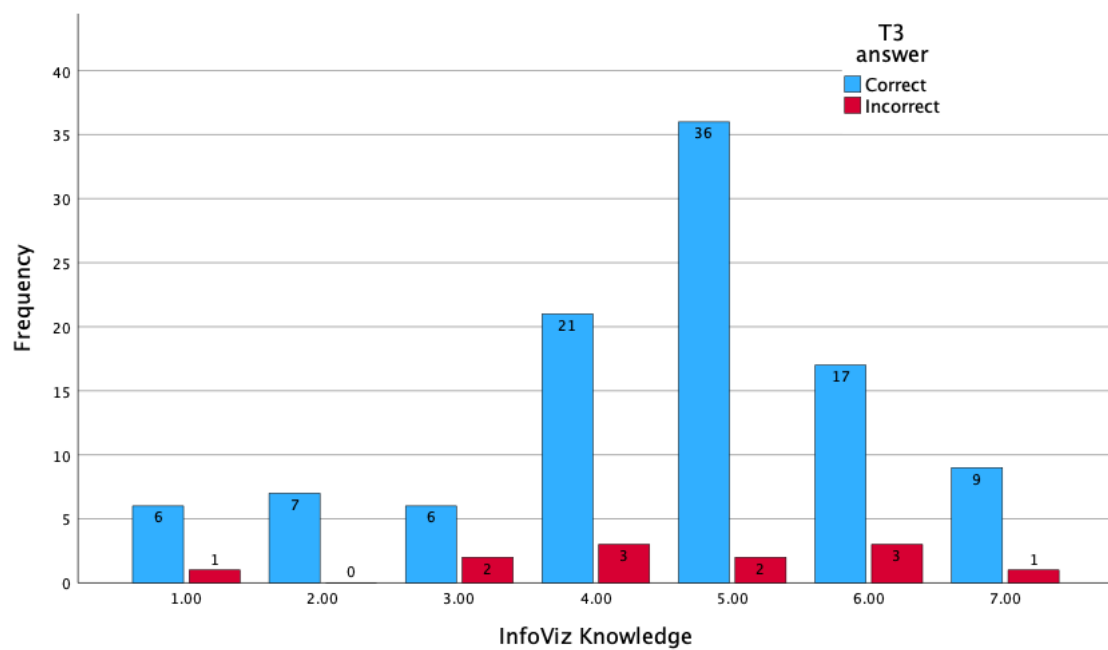


(a) Correct and incorrect answers per information visualization knowledge to question T2 (Task 2).

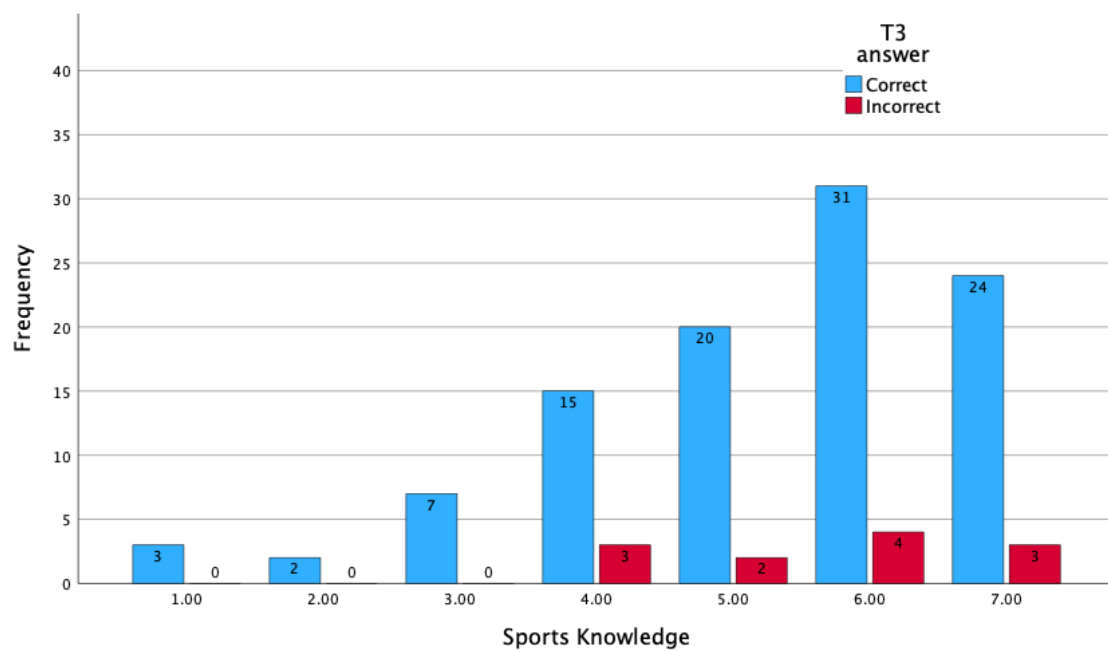


(b) Correct and incorrect answers per sports knowledge to question T2 (Task 2).

Figure C.14: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T2 (Task 2) among participants.

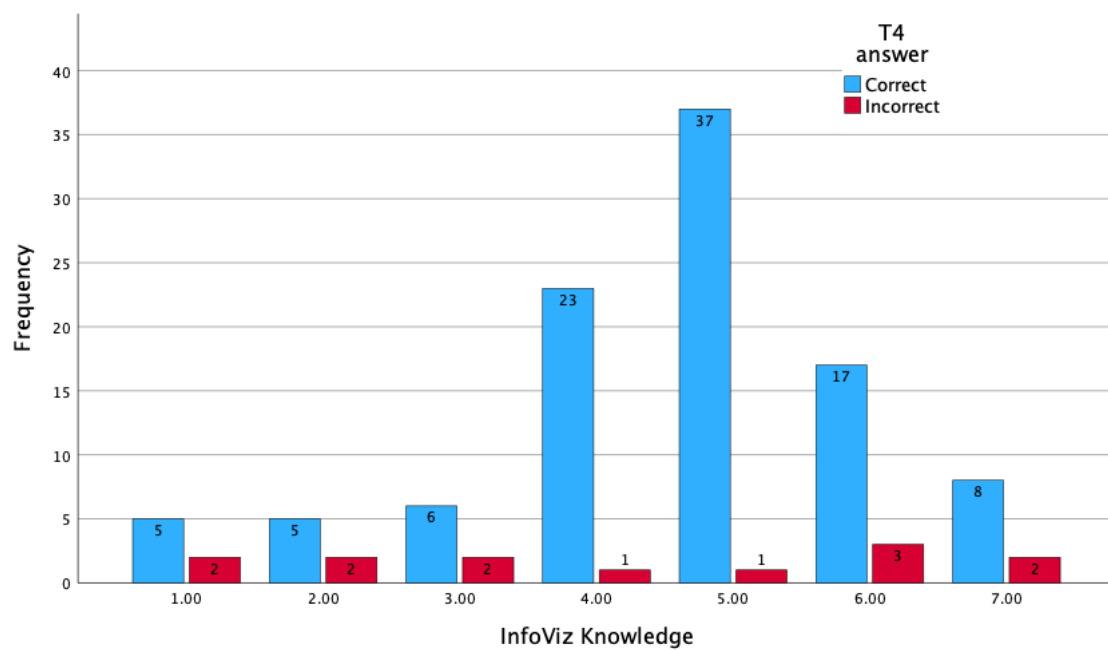


(a) Correct and incorrect answers per information visualization knowledge to question T3 (Task 3).

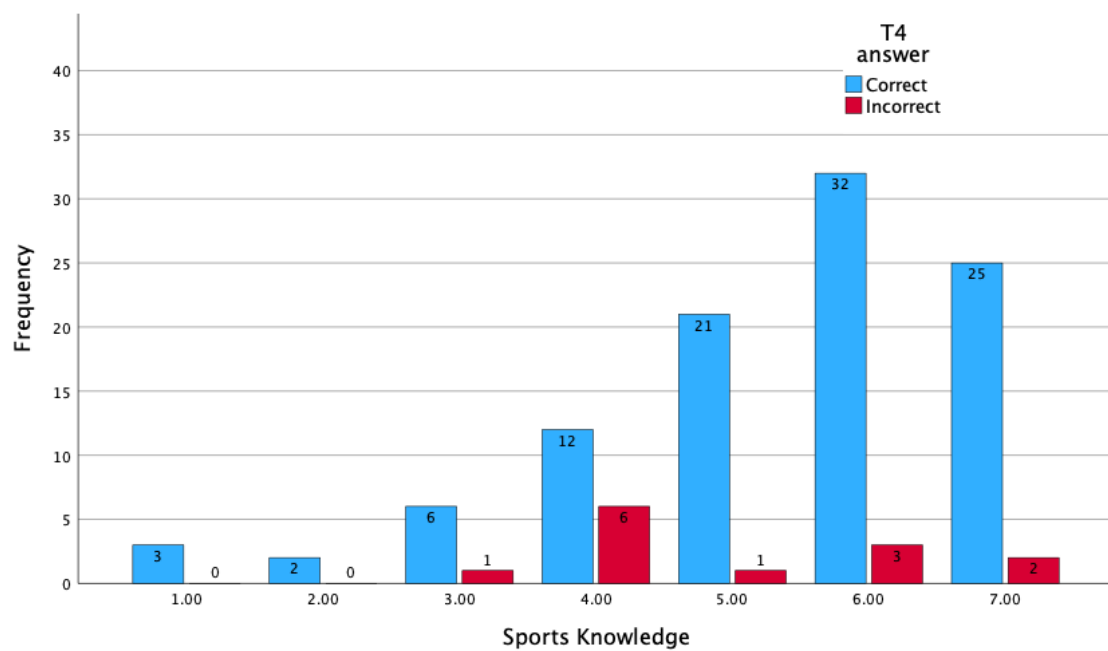


(b) Correct and incorrect answers per sports knowledge to question T3 (Task 3).

Figure C.15: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T3 (Task 3) among participants.

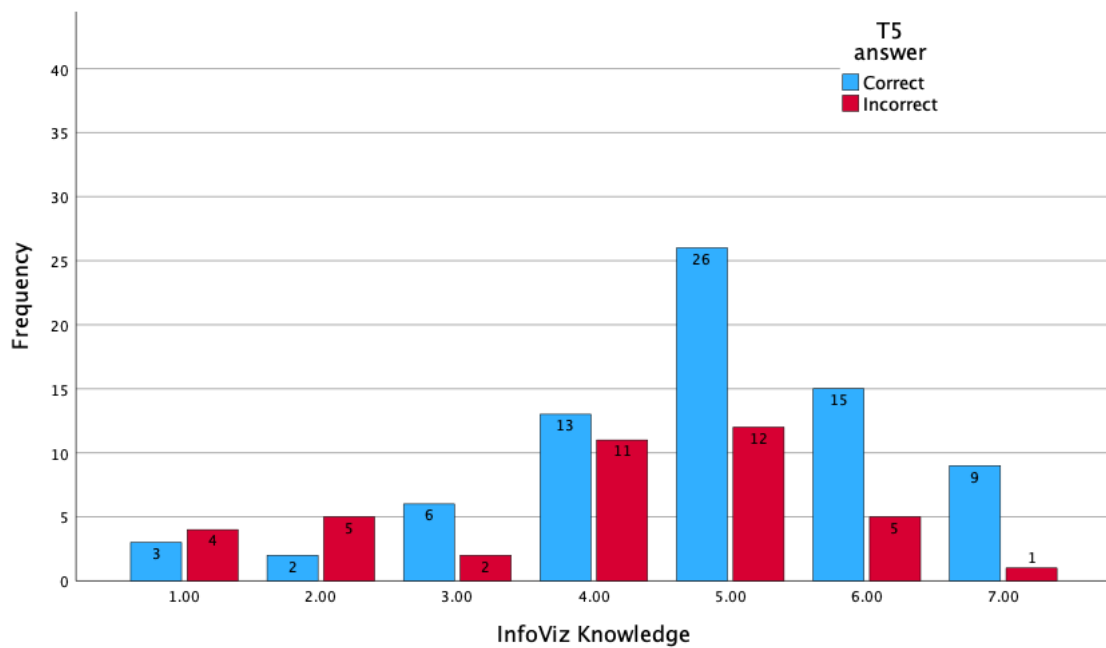


(a) Correct and incorrect answers per information visualization knowledge to question T4 (Task 4).

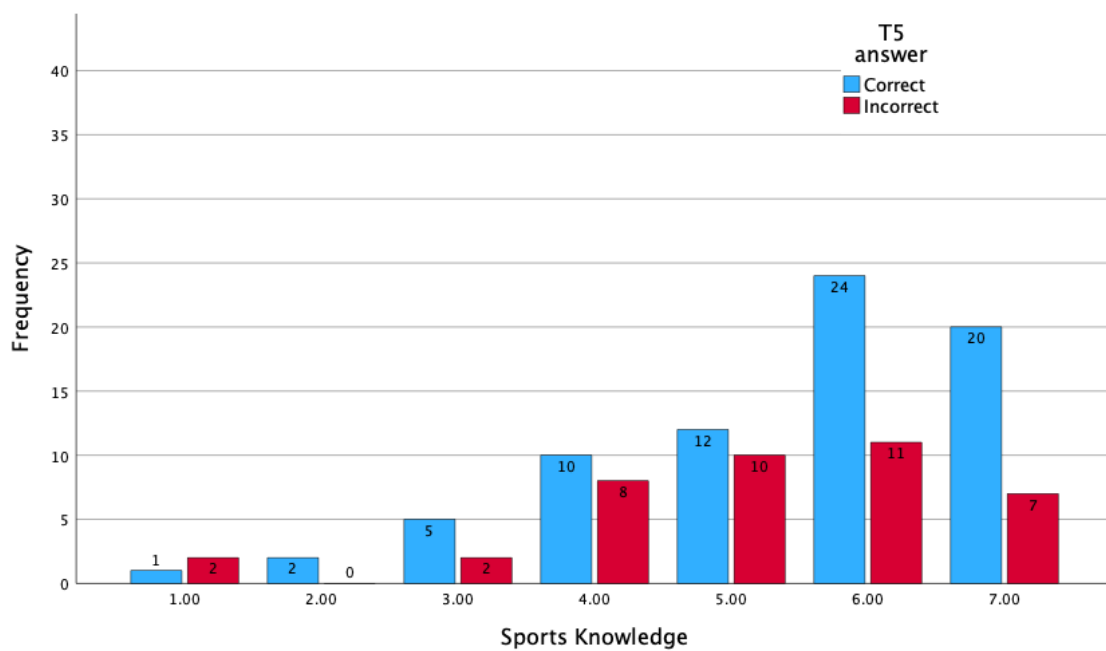


(b) Correct and incorrect answers per sports knowledge to question T4 (Task 4).

Figure C.16: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T4 (Task 4) among participants.

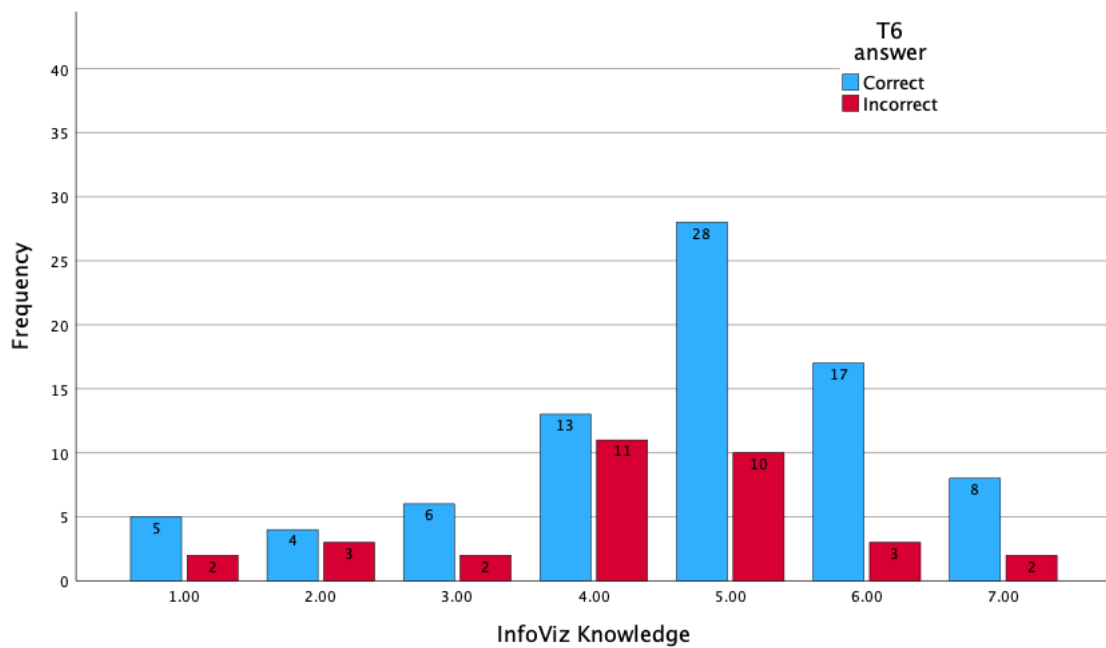


(a) Correct and incorrect answers per information visualization knowledge to question T5 (Task 5).

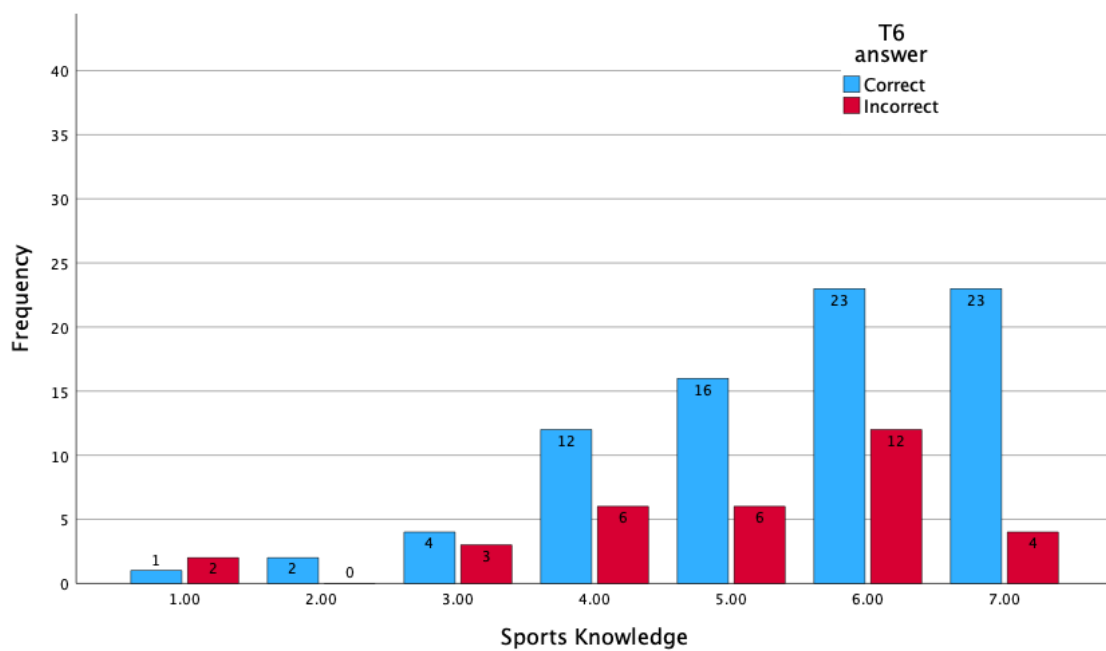


(b) Correct and incorrect answers per sports knowledge to question T5 (Task 5).

Figure C.17: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T5 (Task 5) among participants.

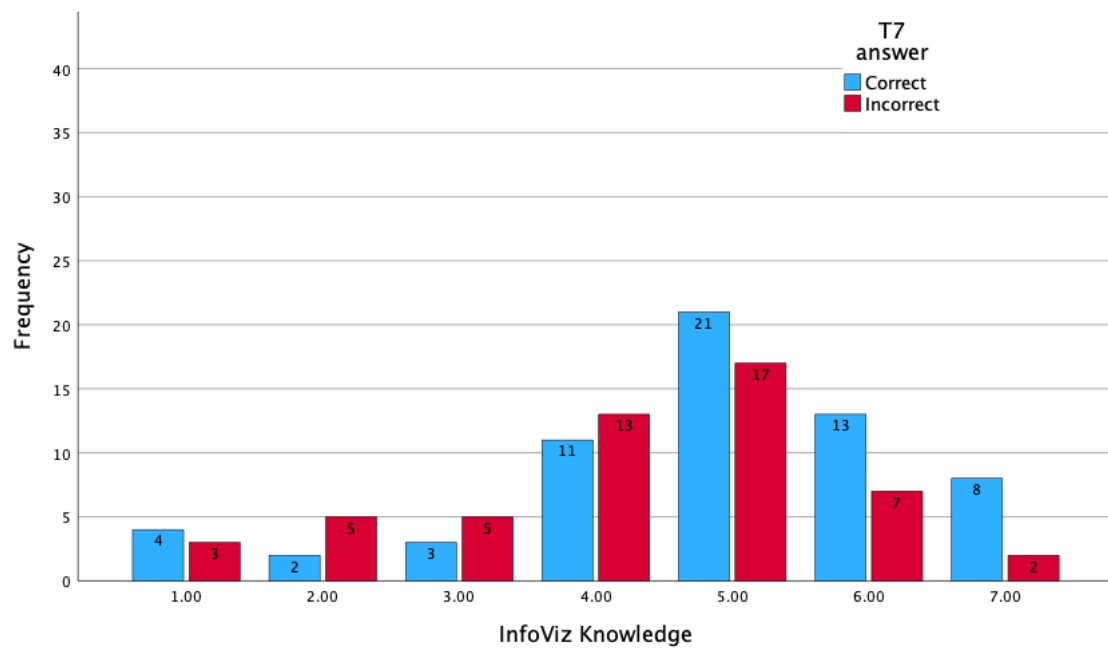


(a) Correct and incorrect answers per information visualization knowledge to question T6 (Task 6).

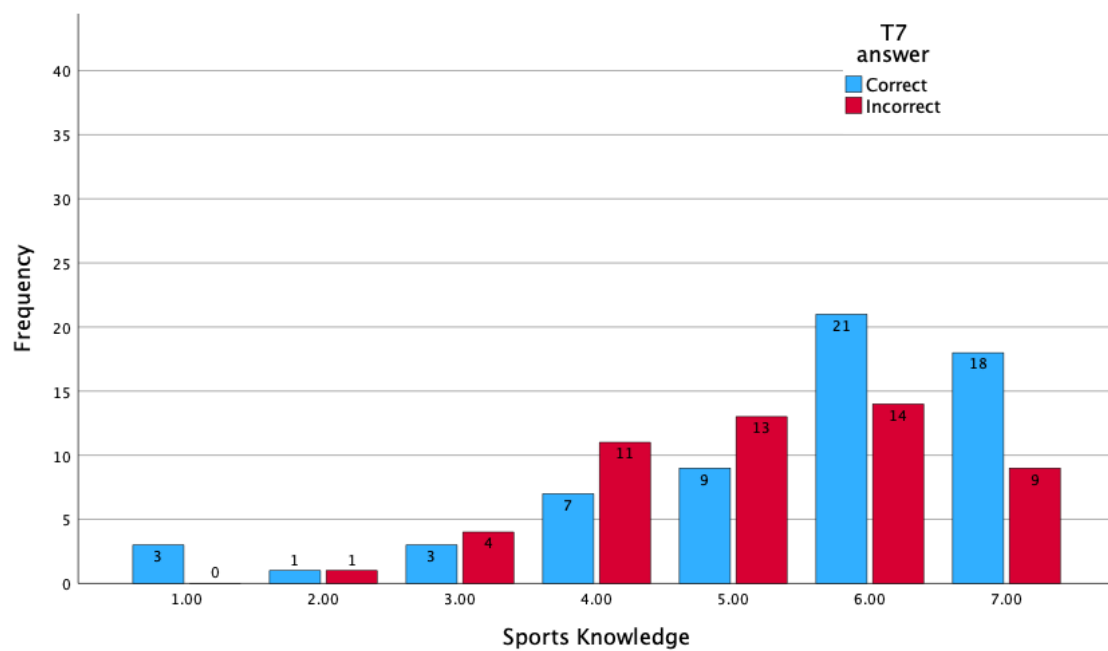


(b) Correct and incorrect answers per sports knowledge to question T6 (Task 6).

Figure C.18: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T6 (Task 6) among participants.

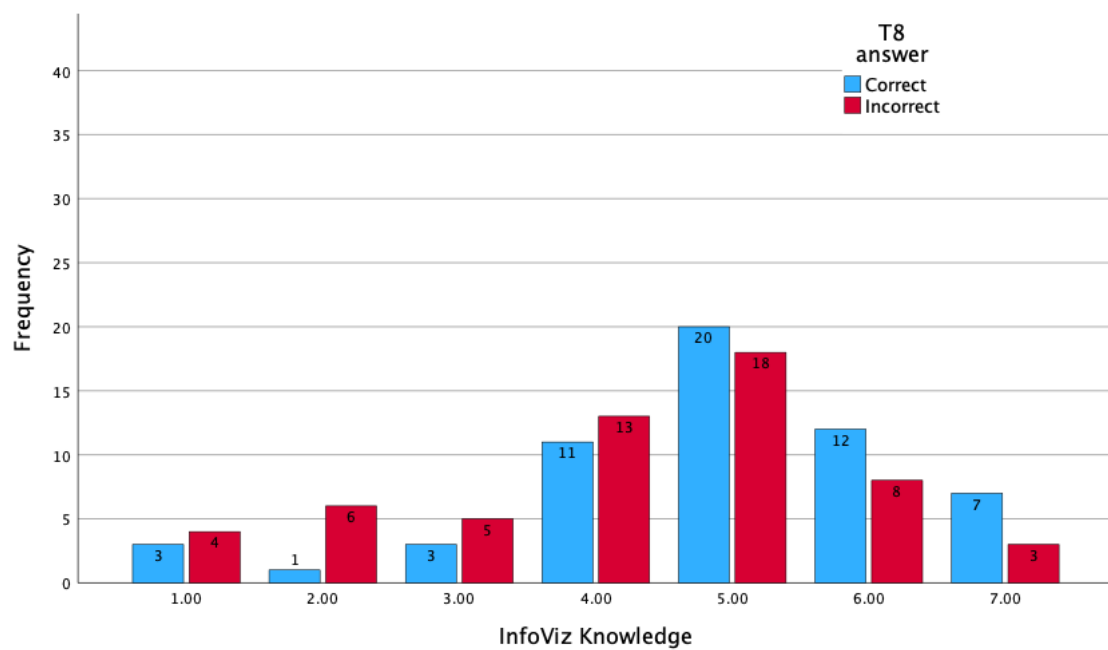


(a) Correct and incorrect answers per information visualization knowledge to question T7 (Task 7).

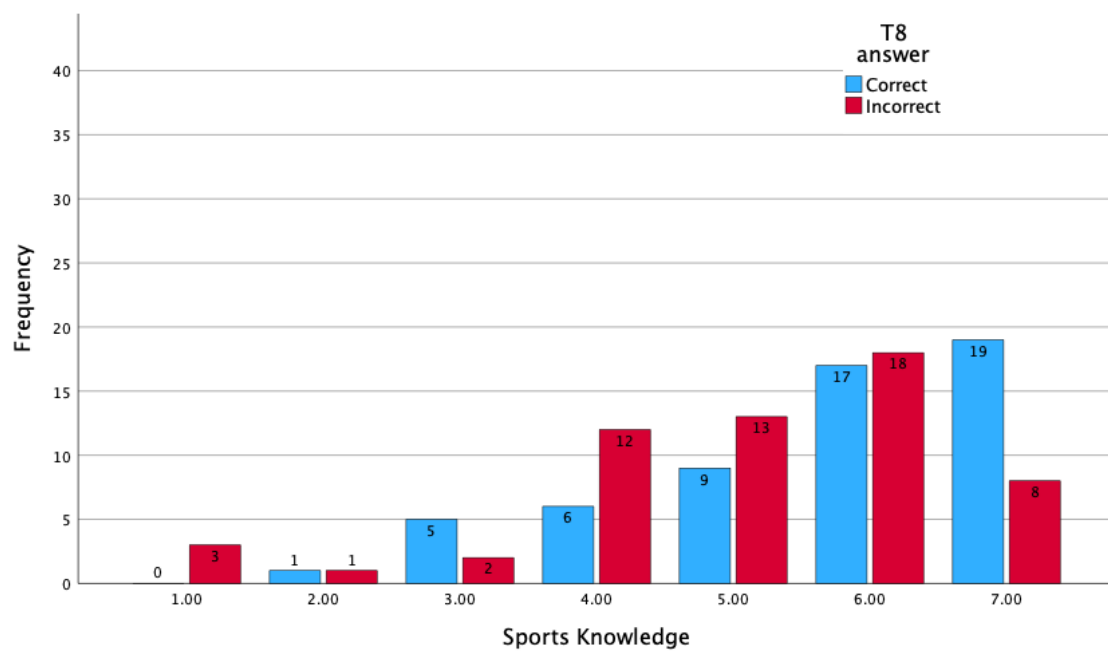


(b) Correct and incorrect answers per sports knowledge to question T7 (Task 7).

Figure C.19: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T7 (Task 7) among participants.

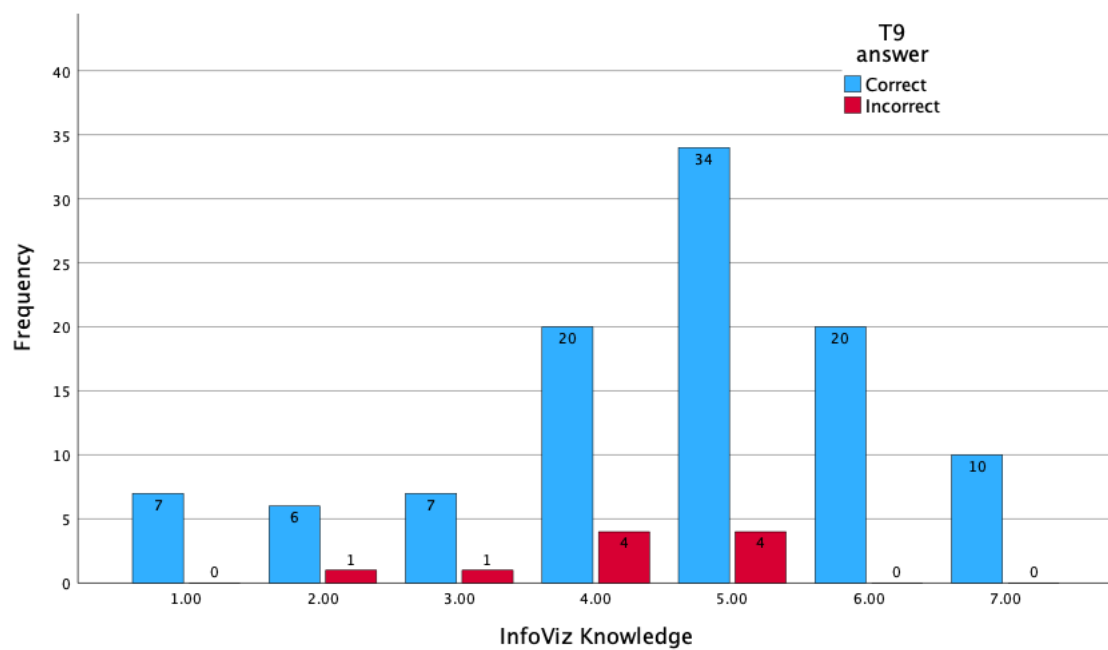


(a) Correct and incorrect answers per information visualization knowledge to question T8 (Task 8).

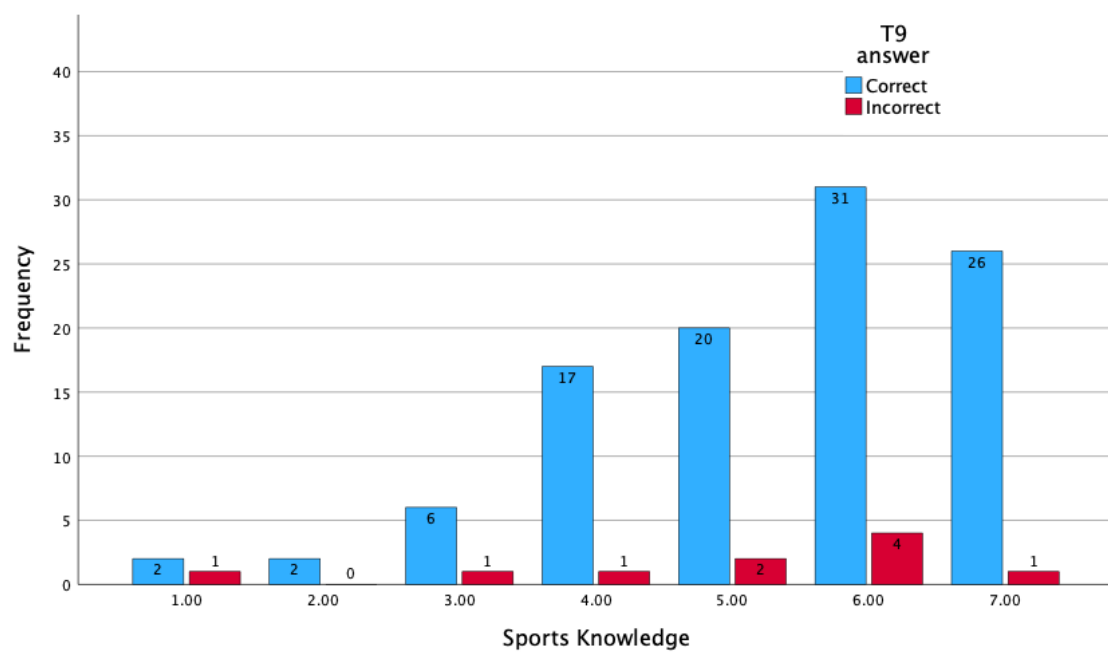


(b) Correct and incorrect answers per sports knowledge to question T8 (Task 8).

Figure C.20: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T8 (Task 8) among participants.



(a) Correct and incorrect answers per information visualization knowledge to question T9 (Task 9).



(b) Correct and incorrect answers per sports knowledge to question T9 (Task 9).

Figure C.21: Bar chart of the distribution of the correct and incorrect answers per domain knowledge to question T9 (Task 9) among participants.

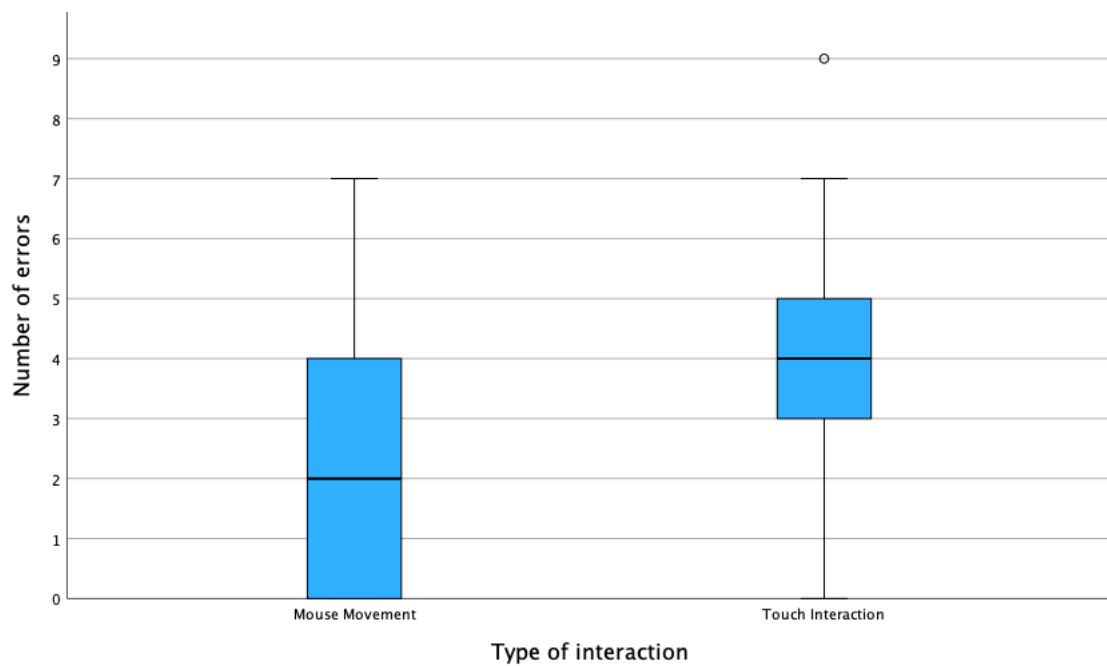


Figure C.22: Box-plot of the distribution of the number of incorrect answers (errors) per interaction type among participants.

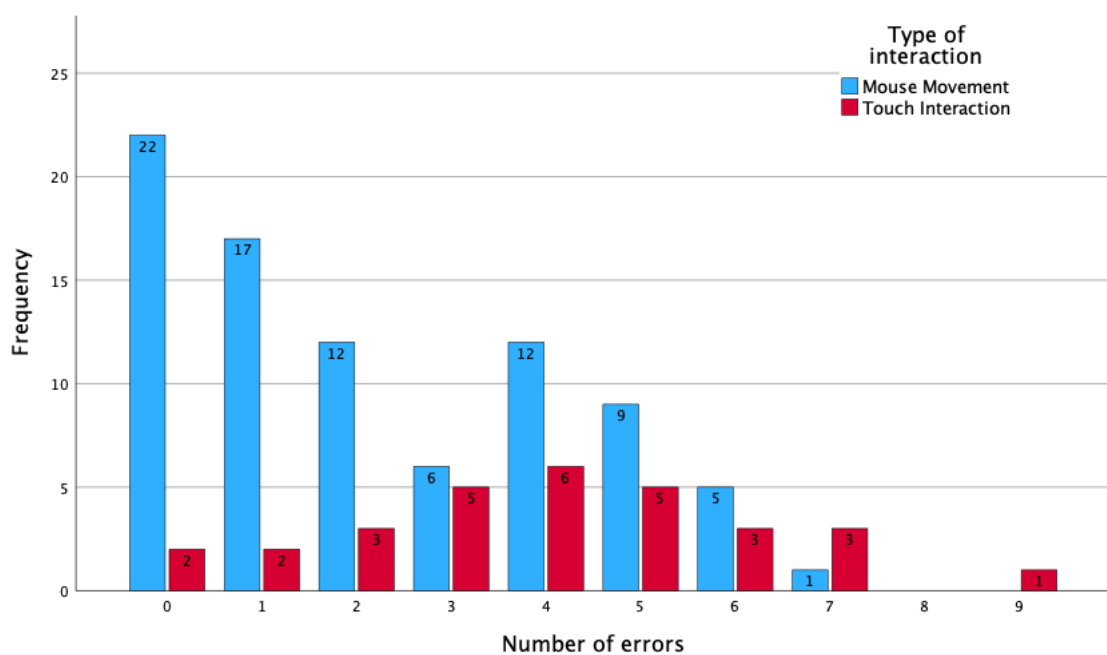


Figure C.23: Bar chart of the distribution of the number of incorrect answers (errors) per interaction type among participants.

	Event name	Event Value	+ Event count	Active users	Event count per user
	Totals		5,029	342	14.7
1	control_timeline	Catar 2022	259	122	2.12
2	control_timeline	México 1970	255	155	1.65
3	control_timeline	Inglaterra 1966	251	151	1.66
4	control_timeline	Chile 1962	243	148	1.64
5	control_timeline	Suiça 1954	235	144	1.63
6	control_timeline	Espanha 1982	234	163	1.44
7	control_timeline	Suécia 1958	234	145	1.61
8	control_timeline	Rússia 2018	231	177	1.31
9	control_timeline	Alemanha 1974	230	153	1.5
10	control_timeline	EUA 1994	230	161	1.43
11	control_timeline	França 1998	230	169	1.36
12	control_timeline	Brasil 1950	228	143	1.59
13	control_timeline	Argentina 1978	227	156	1.46
14	control_timeline	Coreia/Japão 2002	225	163	1.38
15	control_timeline	França 1938	221	134	1.65
16	control_timeline	Itália 1934	218	129	1.69
17	control_timeline	Itália 1990	217	157	1.38
18	control_timeline	México 1986	216	158	1.37
19	control_timeline	África do Sul 2010	213	168	1.27
20	control_timeline	Alemanha 2006	212	163	1.3
21	control_timeline	Uruguai 1930	211	127	1.66
22	control_timeline	Brasil 2014	209	171	1.22

Figure C.24: Values of *control_timeline* events in *next* category.

	Event name	Event Value	+ Event count	Active users	Event count per user
	Totals		3,526	326	10.82
1	control_timeline	Catar 2022	330	248	1.33
2	control_timeline	Rússia 2018	259	203	1.28
3	control_timeline	Brasil 2014	238	189	1.26
4	control_timeline	Alemanha 2006	216	158	1.37
5	control_timeline	África do Sul 2010	216	171	1.26
6	control_timeline	Coreia/Japão 2002	198	144	1.38
7	control_timeline	França 1998	187	132	1.42
8	control_timeline	EUA 1994	166	124	1.34
9	control_timeline	Itália 1990	154	118	1.31
10	control_timeline	México 1986	147	105	1.4
11	control_timeline	Alemanha 1974	145	98	1.48
12	control_timeline	Espanha 1982	140	106	1.32
13	control_timeline	México 1970	140	97	1.44
14	control_timeline	Argentina 1978	139	102	1.36
15	control_timeline	Inglaterra 1966	120	78	1.54
16	control_timeline	Chile 1962	113	76	1.49
17	control_timeline	Suécia 1958	109	75	1.45
18	control_timeline	França 1938	108	69	1.57
19	control_timeline	Itália 1934	107	71	1.51
20	control_timeline	Brasil 1950	106	71	1.49
21	control_timeline	Suiça 1954	102	69	1.48
22	control_timeline	Uruguai 1930	86	47	1.83

Figure C.25: Values of *control_timeline* events in *prev* category.

	Event name	Event Value	+ Event count	Active users	Event count per user
	Totals		2,000	466	4.29
1	control_timeline	Uruguai 1930	281	187	1.5
2	control_timeline	Catar 2022	264	205	1.29
3	control_timeline	África do Sul 2010	100	83	1.2
4	control_timeline	Alemanha 2006	94	78	1.21
5	control_timeline	Inglaterra 1966	94	82	1.15
6	control_timeline	Brasil 2014	93	79	1.18
7	control_timeline	Coreia/Japão 2002	89	79	1.13
8	control_timeline	EUA 1994	89	74	1.2
9	control_timeline	Itália 1990	89	76	1.17
10	control_timeline	México 1986	85	76	1.12
11	control_timeline	França 1998	81	70	1.16
12	control_timeline	Rússia 2018	75	68	1.1
13	control_timeline	Espanha 1982	74	62	1.19
14	control_timeline	México 1970	68	59	1.15
15	control_timeline	Chile 1962	67	59	1.14
16	control_timeline	Alemanha 1974	66	59	1.12
17	control_timeline	Brasil 1950	56	48	1.17
18	control_timeline	Argentina 1978	52	48	1.08
19	control_timeline	Itália 1934	50	42	1.19
20	control_timeline	França 1938	47	42	1.12
21	control_timeline	Suécia 1958	44	38	1.16
22	control_timeline	Suiça 1954	42	40	1.05

Figure C.26: Values of *control_timeline* events in *change* category.

	Event name	Event Value	+ Event count	Active users	Event count per user
	Totals		736	355	2.07
1	control_timeline	Catar 2022	308	255	1.21
2	control_timeline	Uruguai 1930	107	80	1.34
3	control_timeline	Inglaterra 1966	28	27	1.04
4	control_timeline	México 1986	27	23	1.17
5	control_timeline	Alemanha 2006	26	26	1
6	control_timeline	Coreia/Japão 2002	25	25	1
7	control_timeline	Brasil 2014	21	19	1.11
8	control_timeline	África do Sul 2010	21	20	1.05
9	control_timeline	EUA 1994	20	19	1.05
10	control_timeline	Rússia 2018	19	17	1.12
11	control_timeline	Espanha 1982	18	17	1.06
12	control_timeline	Alemanha 1974	17	15	1.13
13	control_timeline	Itália 1990	17	15	1.13
14	control_timeline	França 1998	13	13	1
15	control_timeline	França 1938	12	12	1
16	control_timeline	Itália 1934	11	11	1
17	control_timeline	Argentina 1978	9	9	1
18	control_timeline	Chile 1962	9	7	1.29
19	control_timeline	México 1970	9	7	1.29
20	control_timeline	Brasil 1950	8	7	1.14
21	control_timeline	Suécia 1958	8	8	1
22	control_timeline	Suiça 1954	3	3	1

Figure C.27: Values of *control_timeline* events in *play* category.

Event name	Event Value	± Event count	Active users	Event count per user
Totals		324	144	2.25
1	control_timeline Uruguai 1930	28	25	1.12
2	control_timeline México 1986	26	19	1.37
3	control_timeline França 1938	22	21	1.05
4	control_timeline Inglaterra 1966	21	20	1.05
5	control_timeline Alemanha 2006	20	20	1
6	control_timeline Itália 1934	20	20	1
7	control_timeline Coreia/Japão 2002	19	19	1
8	control_timeline África do Sul 2010	19	18	1.06
9	control_timeline Brasil 2014	16	15	1.07
10	control_timeline EUA 1994	14	13	1.08
11	control_timeline Espanha 1982	14	13	1.08
12	control_timeline Rússia 2018	14	14	1
13	control_timeline Brasil 1950	12	12	1
14	control_timeline Itália 1990	12	11	1.09
15	control_timeline Catar 2022	10	10	1
16	control_timeline Alemanha 1974	9	8	1.13
17	control_timeline Chile 1962	9	7	1.29
18	control_timeline França 1998	9	9	1
19	control_timeline Suécia 1958	9	9	1
20	control_timeline México 1970	8	6	1.33
21	control_timeline Argentina 1978	7	5	1.4
22	control_timeline Suíça 1954	6	6	1

Figure C.28: Values of *control_timeline* events in *pause* category.

Event name	Event Value	± Event count	Active users	Event count per user
Totals		380	88	4.32
1	control_timeline EUA 1994	102	46	2.22
2	control_timeline França 1998	71	35	2.03
3	control_timeline Coreia/Japão 2002	68	45	1.51
4	control_timeline Alemanha 2006	60	40	1.5
5	control_timeline Catar 2022	38	20	1.9
6	control_timeline Espanha 1982	8	2	4
7	control_timeline África do Sul 2010	8	6	1.33
8	control_timeline Rússia 2018	7	5	1.4
9	control_timeline México 1986	5	2	2.5
10	control_timeline Inglaterra 1966	4	4	1
11	control_timeline Itália 1990	4	3	1.33
12	control_timeline Brasil 2014	3	3	1
13	control_timeline Alemanha 1974	1	1	1
14	control_timeline Chile 1962	1	1	1

Figure C.29: Values of *control_timeline* events in *goto* category.