US University & College Ancestry Visualizations

Paul Miller: pdm1@iu.edu, **Akshada Bhandari:** aksbhand@iu.edu, **Pranay Chopra:** choprap@iu.edu, **Renu Jaiswal:** rjaiswal@iu.edu,

Mansi Sarda: msarda@iu.edu

Abstract—The project aims to create visualizations to understand the evolutions of different universities and colleges in the United States from 1973 till 2021 based on the 10 updates of the Carnegie Classification of Institutions of Higher Education. The project is important for scholars and policymakers to better understand how the higher education environment in the US is changing by creating interactive visualizations of institutional evolution throughout the course of 10 updates to CCIHE. The visualizations include the Table Visualization, Sankey Diagram [1], Bar Chart and Decomposition Tree. The project will evaluate various publicly accessible ancestry visualization tools and resources to suggest how to use ancestral visualizations for advocacy and study.

Index Terms—Sankey Diagrams, Evolution, Bar Charts, Table Visualization

Introduction

Due to the expanding diversity of American college and university campuses and the development of many disciplines, ancestry visualizations are becoming increasingly important to understand the direction and focus of US educational institutions. By developing interactive visualizations of institutional evolution throughout the span of the 10 updates to the Carnegie Classification of Institutions of Higher Education, we can better inform researchers and policymakers about how the higher education landscape in the US is changing. Our visualizations could be an effective tool for stratification and sampling in higher education research, as well as for assessing institutional performance and holding institutions responsible. to ascertain which new subjects are developing, their appeal, and the transition from the university to the college.

Our visualizations include: Table Visualization, Sankey Diagram, Bar Chart and Decomposition Tree.

1 Motivation

The growing diversity of American college and university campuses, the emergence of new fields, the transformation of colleges to universities, and the evolution of different domains call for a better understanding of the changing higher education environment in the US. Ancestry visualizations can be a useful tool to comprehend the direction and focus of US education institutions, evaluate institutional performance, and hold institutions accountable. Therefore, in this research paper, we aim to evaluate and contrast various publicly accessible ancestry visualization tools and resources, pinpoint their advantages and disadvantages, and offer suggestions on how to use ancestral visualizations for advocacy and study. Through this research, we hope to help scholars and policymakers better understand how the higher education environment in the US is changing, identify emerging fields and their popularity, and maintain the ancestry of US educational institutions.

1.1 Related work

Investigating genealogical data is the main goal of the visualization. Individuals' ancestry can be discovered via genealogical information, which is compiled from either official or unofficial historical documents. The project's main goal is to depict the development of US colleges and universities. Since there is no existing direct study specifically relevant to this project, we will instead discuss visualizations of genealogical data.

In the study [2], a method for exploring and evaluating genealogical data known as Timenets is described. The complexity and variety of temporal links between family members are present in genealogical data. Users can explore and study the data more easily using Timenets, which give these correlations throughout time a graphical representation.

The authors [3] begin by reviewing related research in the fields of graph layout algorithms and genealogy visualization. They then go over their visualization method, which represents the genealogical graph using a node-link diagram. The authors explain their strategy, which combines force-directed and hierarchical layout techniques, for organizing the diagram's nodes.

The publication [4] offers VisAC, a brand-new interactive visualization tool that enables users to explore and examine consanguinity in an individual's history. The degree of kinship between people is known as consanguinity, and it plays a significant role in many fields, including genetics, family history, and medicine.

The paper [5] presents a new design for visualizing family trees in a contextual way. Family trees are commonly used to represent genealogical relationships between family members, but they can become complex and difficult to navigate when the family tree is large or includes multiple generations. This case coincides with our data and hence some of the design could prove useful as a basis for our visualizations.

From the article by Yan and Zhang [6], the use of visualizations in higher education research, particularly in understanding student mobility patterns and institutional diversity. The article also emphasizes the importance of institutional diversity in higher education, particularly in relation to student outcomes and the labor market. This literature helped in understanding how institutional diversity has evolved over time. The authors have highlighted the role of data analytics in higher education decision-making, particularly in relation to enrollment management and student success initiatives and thus this literature helps to understand the ancestry of US universities and colleges and inform strategic decision-making.

2 Data Analysis

The usefulness of the visualization can be considerably influenced by the caliber of the data used in it. The US University & College Ancestry Visualizations dataset contains duplicated and incomplete data, which could result in erroneous or deceptive findings.

Based on the analysis of the dataset, we can see that there are a total of 38,811 rows and 16 columns in the dataset. Some of the major attributes of the dataset include INSTNAME, UNITID, STATE, YEAR, CATEGORY, GrpCat, and GrpLbl. We can also see that there are some missing values in the dataset, particularly in the CITY, GrpLbl, Status, LinkUnit, and Notes columns. It is important to address these missing values appropriately to ensure that the analysis is accurate. We have organized and cleaned the data to address this problem. For instance, universities that formerly fell under the category of "Unclassified" are now classified as "Not yet born" and "Inactive". In order for our client to provide the necessary corrections at their level, we have informed them of the accuracy concerns.

One key visualization that can be created using this dataset is a Sankey diagram that shows the evolution of institutional classification from 1973 to 2021. This visualization can provide insights into how institutional classifications have changed over time, and which categories have seen the most growth or decline.

Through the data cleaning and analysis process, we can gain insights into trends in institutional naming changes over time, and how these changes fit into broader institutional trends. For example, the bar graph visualization of the number of institutional name changes over a specific year may reveal that there are certain periods when higher name transitions occurred, which could indicate changes in institutional priorities or focus.

Analyzing the CCIHE dataset can provide valuable insights into the evolution of institutional diversity in U.S. higher education over the past 50 years, and can help researchers better understand the factors that have driven these changes.

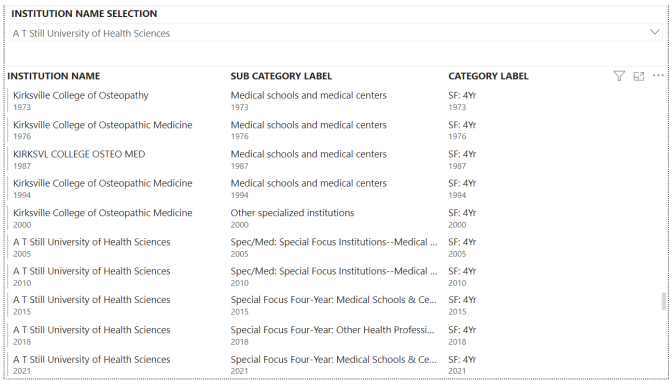
Attributes	INSTNAME	UNITID	STATE	YEAR	CATEGORY	GrpCat	GrpLbl
Count	38811	38811	38807	38811	38811	38811	38811
Unique Values	14739	6546	66	10	14739	8	8

- Table 1. Major Dataset Statistics

2.1 Data Visualizations

Our major visualizations focus on showing the evolution of universities' names, the coursework adapted and the university spread. We have four main visualizations which are explained below along with the insight needs.

2.1.1 Evolution in Name changes



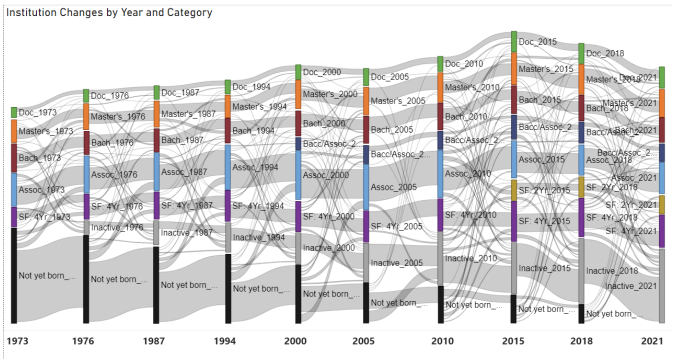
- Fig. 1 - Table representing each Institution Evolution over time

Tracking the evolution of institutions over time within CCIHE is very challenging. This problem was solved using table visualizations which helps in providing a clear and concise way of presenting this information. The visualization is structured as a table which includes name of institution, sub-category label and category label as shown in Figure 1. This allows viewers to easily track the evolution of each institution and see how changes in institution name correspond with changes in classification.

Examining the table can reveal patterns in institution name changes and expansions of academic curriculum over time. For example, some institutions may change their names more frequently than others, or there may be certain periods of time when name changes are more common. Analyzing the table can also provide insights into the reasons behind institutional name changes. Institutions may change their names to better reflect their academic programs or to distance themselves from controversial histories, among other reasons. There also might be cases where the names have not evolved for a specific year but the institutes upgraded to include higher education domains and levels such Master's or Doctoral programs. Changes in institutional names may correspond with changes in classification within the CCIHE framework. For example, an institution may change its name as part of an effort to improve its

classification. Name changes can have a significant impact on an institution's identity and brand. Analyzing the table can provide insights into how institutions have navigated these changes over time. Finally, analyzing the table can reveal geographic and demographic trends in institutional changes. For example, certain regions or types of institutions may be more likely to change their names than others.

2.1.2 University Evolution Category Wise



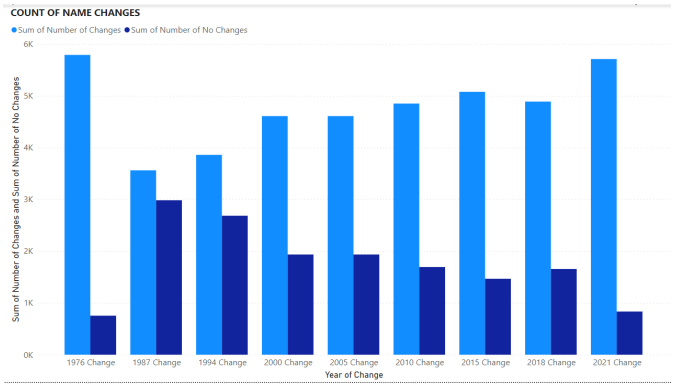
- Fig. 2 - Sankey diagram for Institution Category changes over time

Sankey diagrams are a type of data visualization that show the flow of data or information through a system. In the context of institutional evolution, a Sankey diagram could be used to show how institutions have changed over time in terms of their classification within the Carnegie framework as shown in Figure 2. The diagram shows the flow of institutions from one classification to another over time, highlighting which categories saw the most movement.

The diagram covered the 10 updates from 1973 with the most recent update as in 2021 according to the Carnegie framework. This provided a comprehensive overview of the evolution of universities in this timeframe. The diagram included all the categories from the data including two new categories labeled “Inactive” and “Not Yet Born”. The diagram helps analyze how institutions moved over a time period as institutions may remain in the same category or become part of a different classification. The width of the links are determined on the basis of weights which are summed up for every update of the university. The nodes for specific years are colored to identify the change over time more easily.

For example, the diagram could show that many institutions moved from a "Master's Colleges and Universities" classification to a "Doctoral/Professional Universities" classification over the past decade, while relatively few moved out of the "Doctoral" category. This provides insight into how the landscape of higher education is changing, and which category types of institutions are growing or declining. The visualization also helps to understand when some institutions were established and if any institutions were shut down over the years and labeled as Inactive.

2.1.3 Name Changes Over The Years



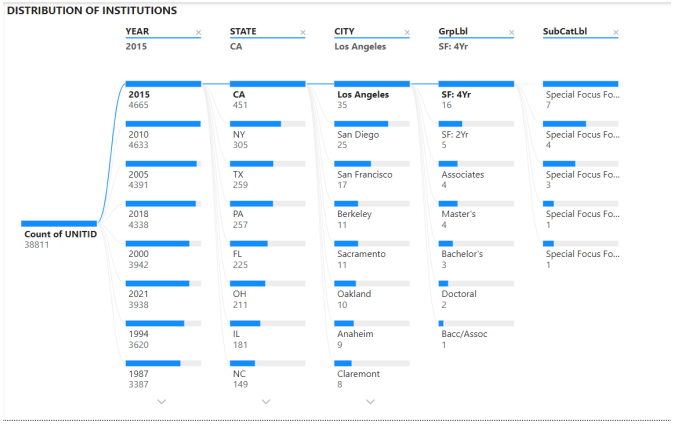
- Fig. 3 - Bar Chart for Distribution of Name Changes each year

The Bar graph visualization is used to track the count of institutional name changes over time within the Carnegie Classification of Institutions of Higher Education (CCHIE) as shown in Figure 3. The x-axis represents the year in which the name was updated as compared to previous year, while the y-axis represents the number of institutions which underwent the name changes as well as the count of universities which didn’t have any name change.

By analyzing the bar graph, stakeholders can identify trends and patterns in institutional name changes over time. For example, the graph could reveal whether there are certain periods when name changes were more common, or if there are specific years when many institutions underwent name changes. Comparing the number of name changes to other institutional changes, such as changes in classification or status, can also provide insights into how name changes fit into broader institutional trends. For example we can interpret that the maximum number of changes occurred from 1973 to 1976. And the count of changes decreased in the next year and after that displayed a gradual increase trend until 2021.

The bar graph can also be used to identify outliers or unique patterns. For instance, if there is a year with an unusually high number of name changes, it may be worth investigating the reasons behind this trend. Overall, a bar graph visualization can provide a clear and concise way of analyzing trends in institutional naming over time within the CCIHE framework, and can help researchers gain insights into the factors driving these changes.

2.1.4 Distribution of Institutions



- Fig. 4 - Decomposition Tree for Distribution of Institutions

The Decomposition Tree visualization is used to display the distribution of institutions across different classifications of the Carnegie Framework and locations for each year . This allows the comparison between the number and type of each institution within each state and city for all the years.

For example, the tree shows that California has the maximum number of institutions in the year 2015 in Los Angeles and has a max number of institutions in the "SF: 4yr Colleges and Universities" classification, while New York has a greater number of institutions in the "SF: 4yr Colleges and Universities" classification. Over time, the tree could reveal shifts in the distribution of institutions within each state, indicating changes in the higher education landscape.

To design this visualization, the tree is organized by this specific ordering to show major groups higher in the hierarchy. Hence the Year is represented first followed by State and City columns for geographic analysis. Then for each the count is shown for each category and sub category of Carnegie Framework.

The tree includes all the categories from the Carnegie framework. This gives a detailed analysis of how each state has moved from

different types of classification. This also helps in determining which institutes have been growing or diminishing over time. Also gives insight as to which state shows significant allocation of resources to education and academic variety.

3 Challenges

Data accuracy and completeness: The quality of the data used in the visualization can significantly impact its effectiveness. We have incomplete and duplicate data that can lead to incorrect or misleading conclusions.

To resolve this issue, we have restructured and cleaned the data. For instance, the universities which were under the "Unclassified" label are now categorized in "Not yet born" and "Inactive".

We have also reported some of the inaccuracy issues to our client so that they can rectify at their level.

Visual clarity and user-friendliness: The tool's interface and visual representation of data must be clear and easy to understand to ensure that users can effectively interpret and analyze the information presented. Our previous Sankey diagrams showed the course's evolution over the time frame of total universities. However, we were not able to show the information of a particular university.

4 Conclusion

The use of the above visualizations can provide valuable insights into the evolution of institutional names and classifications within the Carnegie Classification of Institutions of Higher Education (CCIHE) framework. These visualizations allow researchers to track patterns and trends in institutional name changes and understand how institutions have navigated changes in classification and location over time.

The Table visualization which is the primary deliverable for the stakeholders provides a clear and concise way of presenting the evolution of institutional names and classifications over time. It allows viewers to easily track the changes in the name of institutions, sub-category label and category label. By analyzing this visualization, stakeholders can gain insights into the reasons behind institutional changes, how they correspond with changes in classification, and geographic and demographic trends in institutional name changes.

A Sankey diagram is particularly useful for showing the flow of institutions from one classification to another over time. This visualization allows researchers to identify which categories saw the most movement and gain insights into how the landscape of higher education is changing.

Bar Chart visualizations can help researchers identify trends and patterns in institutional name changes over time. By comparing the number of name changes to other institutional changes, such as changes in classification or status, researchers can gain a deeper understanding of the factors driving these changes.

Finally, the Decomposition Tree helps in understanding the university distribution patterns based on location, year and classification. This will help policy makers to design the education curriculum accordingly which will vary state-to-state and assign resources to states which have lesser number of institutions and lack of academic domains coverage.

5 Future Scope

The following ideas can be implemented in future work to address the challenges:

1. Addressing complexity and scaling issues: As the project aims to track the evolution of 1000s of universities over the years, the data is enormous which makes visualization of data complicated. Future work can include processing and scaling data through code in more efficient ways.
2. Enhancing Data Quality: The data which was entered manually for CCHIE classification includes some discrepancies. This requires manual examination of the data collection process. Future work can include validating and processing data in more depth.
3. Designing user friendly visualizations: Using color for Sankey waves can help in identifying the insights more easily for users. To understand the University name, label and category change over time, the users can be provided with the option to filter the universities according to their interests.
4. Incorporating client feedback: As these visualizations will be used by various people, feedback can be taken to incorporate necessary changes and find out the way visualization can be deployed according to the client feedback.

Acknowledgments

The authors wish to thank Prof. Victor Borden, Prof. Andreas Bueckle and Daria Ivleva for actively engaging with us and guiding throughout the project.

References

[1] The Sankey Diagram

[2] Kim, Nam Wook, Stuart K. Card, and Jeffrey Heer. "Tracing genealogical data with timenets." In *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 241-248. 2010.

[3] McGuffin, Michael J., and Ravin Balakrishnan. "Interactive visualization of genealogical graphs." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 16-23. IEEE, 2005.

[4] Borges, José. "VisAC: An interactive tool for visual analysis of consanguinity in the ancestry of individuals." *Information Visualization* 21, no. 4 (2022): 354-370.

[5] Borges, José. "A contextual family tree visualization design." *Information Visualization* 18, no. 4 (2019): 439-454.

[6] Yan, W., & Zhang, Y. (2019). Exploring the Relationship Between Social Capital and Students' Sense of Belonging in Higher Education.