

Introduction

Background

- Population-scale crowdsourced family tree datasets extract information from publicly available genealogy websites and can offer advanced insights into human evolution, population dynamics, and migration (Wohns et al., 2022)
- Current family tree datasets are skewed to overrepresent certain groups and are can contain a range of errors such as duplicates and missing data (Koylu et al., 2025)

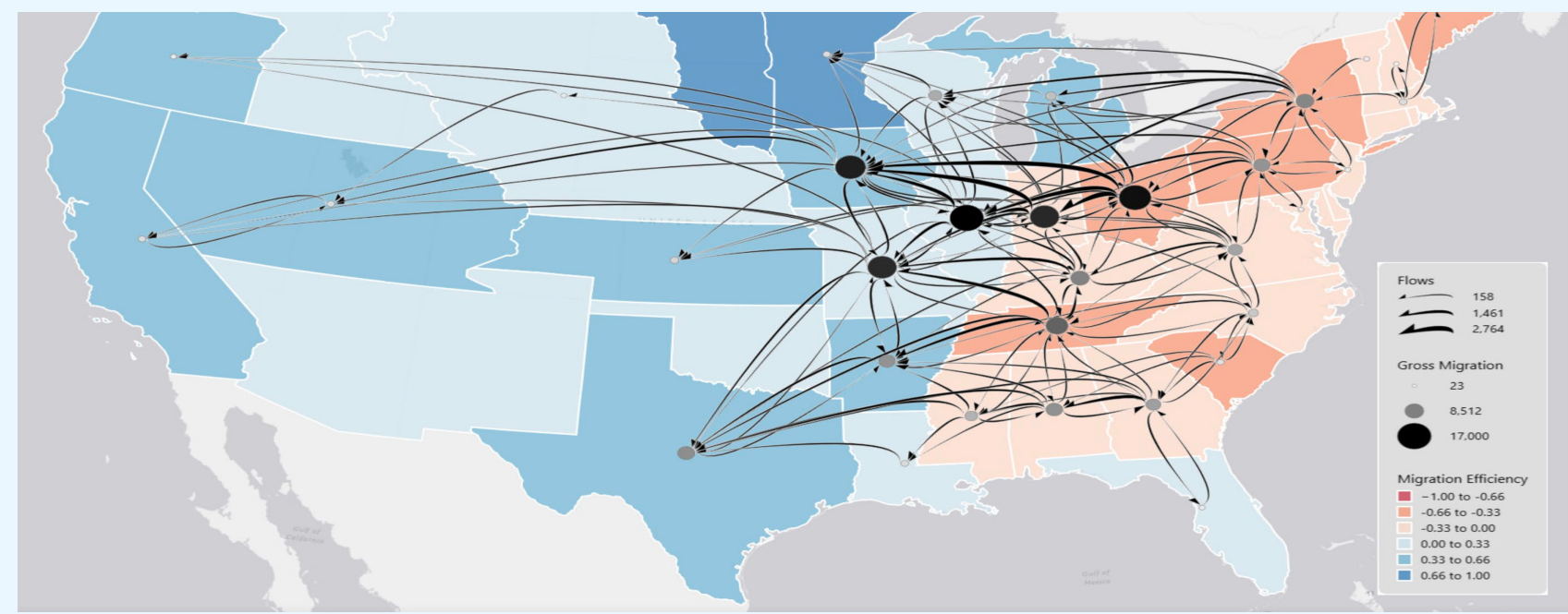


Figure 1. Migration Flow Map Developed with Family Tree Data. Koylu et al. (2024) created a plot of migration flows from 1830-1857 using only family tree data

Existing Literature

- Currently, the largest visualization for a single pedigree of a family tree population dataset displays approximately 6,000 individuals.

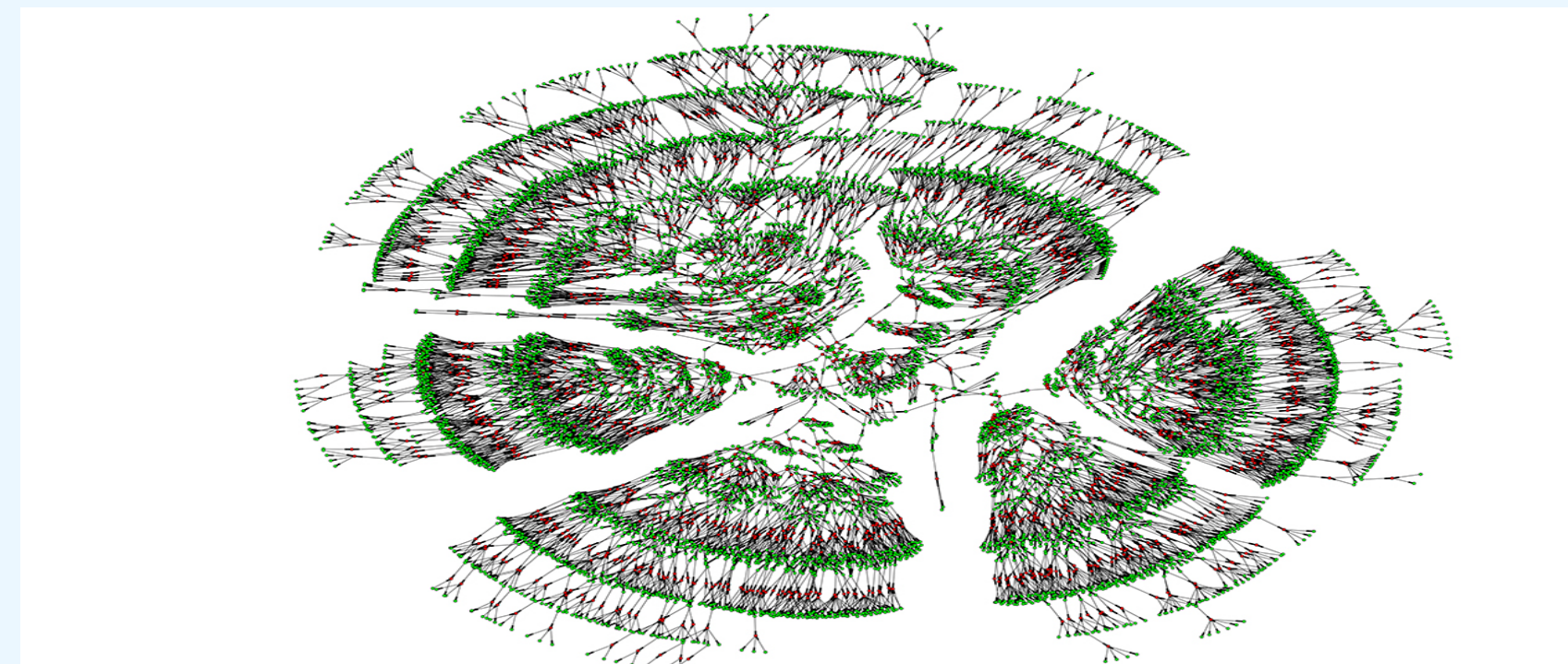


Figure 2. Seven Generation Visualization. Visualization by Kaplanis et al. (2018) that represents the connections between seven generations of over 6,000 individuals

Objectives

Purpose Statement

- In this study, we develop a framework for visualizing population-scale genealogical networks using a compilation of crowdsourced family tree datasets, highlighting direct lineages and spousal connections within the population

Research Questions

- How can we create an effective visualization of population-scale family tree data that can display spatiotemporal changes in kinship and population dynamics?
- How can visualizing population-scale family tree data reveal inconsistencies and biases in the genealogical network data that are otherwise difficult to notice?

Methodology

Data Summary

- The data used was compiled by Koylu et al. (2021), who created a family tree dataset using crowdsourced genealogical profiles from rootsweb.com
- The dataset contains multigenerational pedigrees representing over 40 million individuals, spanning many centuries

Visualization Workflow

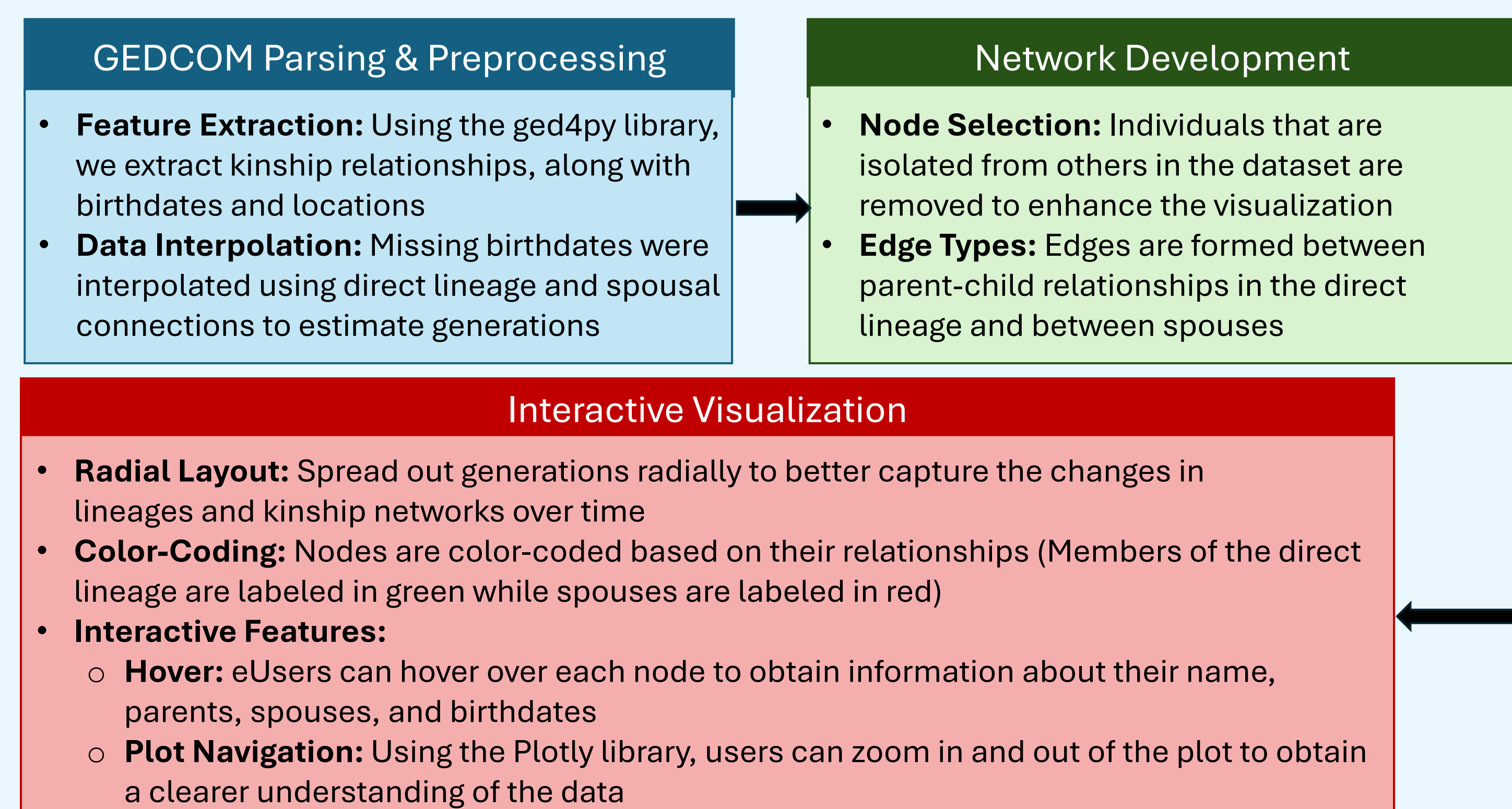


Figure 3. Visualization Creation Workflow. The key steps in the process of interconnecting individuals in the dataset and developing the visualization are processing the GEDCOM files, constructing the network, and interactively visualizing the connections between each node

Results

Framework

- Developed a framework for visualizing large-scale genealogical networks, to better understand their structure and identify issues in data quality

Structures of Multigenerational Family Trees

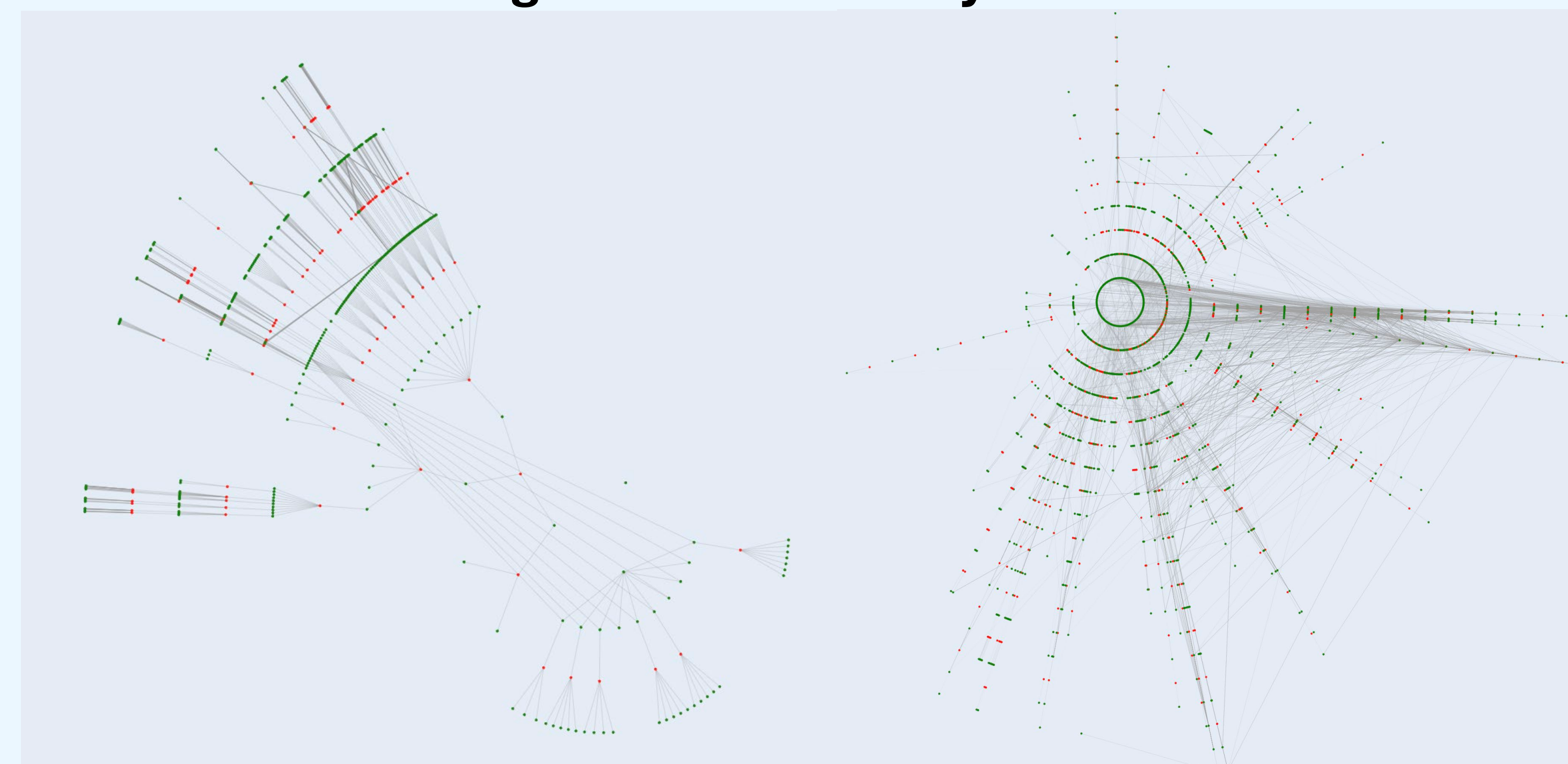


Figure 4. Plotly and NetworkX Visualizations of Two Sets of Individuals. Two sample visualizations of a set of GEDCOM files representing over 700 and 26,000 individuals, respectively. Nodes are color-coded based on type of connection to the lineage and are spread out radially by generation.

Conclusions

Discussion

- Created the largest visualization to-date of a large-scale genealogical network, consisting of over 26 thousand individuals
- Enables in-depth analysis of the anatomy of family tree datasets, along with the impact of historical events on population dynamics across space and time

Limitations

- The data that comprised the visualizations contains bias towards the underrepresentation of specific minorities (Koylu et al. 2022)

Future Work

- Utilizing this visualization framework onto more diverse and representative datasets, to get a better understanding of how populations as a whole respond to historical events
- Implement geocoding into the visualization to not only see changes in kinship ties, but also in spatiotemporal migration patterns in response to various evolutionary events

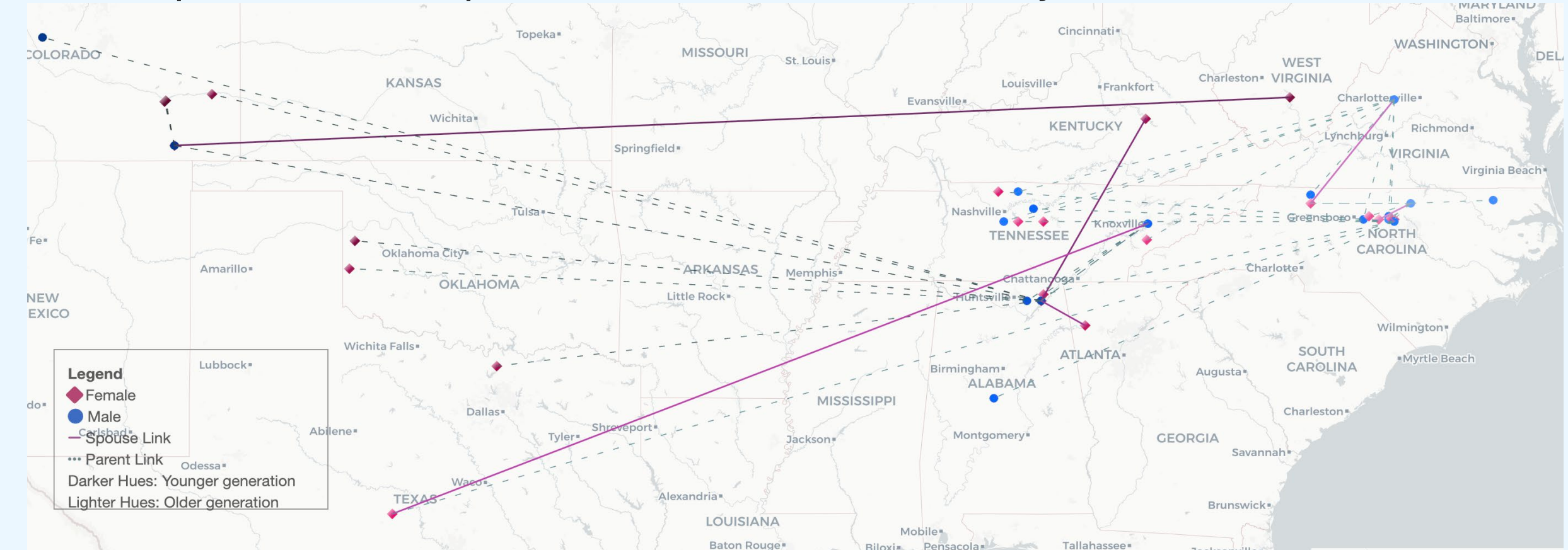


Figure 5. Visualization of Geocoded Family Tree Data. In the study conducted by Clio et al. (under review), family tree data is plotted geographically to observe spatiotemporal variations

Acknowledgements

I would like to acknowledge the support from my mentors and fellow researchers at the Geo-Social Lab. Namely, I want to thank Dr. Koylu, Loretta, Maryam, and Jinyi for their guidance throughout this study.

References

- Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, A., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Erlich, Y., & Carmel, L. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385), 171–175. <https://doi.org/10.1126/science.aam9309>
- Koylu, C., Guo, D., Huang, Y., Kasakoff, A. B., & Grieve, J. (2021). Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science*, 35(12), 2380–2423. <https://doi.org/10.1080/13658816.2020.1821885>
- Koylu, C., & Kasakoff, A. B. (2022). Measuring and mapping long-term changes in migration flows using population-scale family trees. *Cartography and Geographic Information Science*, 46(1), 57–71. <https://doi.org/10.1080/15230406.2021.2011419>
- Koylu, C., & Kasakoff, A. B. (2025). Ethical challenges in analyzing and mapping historical demographic changes and migration using population-scale family trees. *Cartographic Perspectives*, 105, 55–62. <https://doi.org/10.14714/CP105.1945>
- Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375(6583), eabi8264. <https://doi.org/10.1126/science.abi8264>