Who's My Neighbor: Data-Driven Community Detection

Ben Greene, Linden James, Aidan Shen
Project Manager: Dr. Ranthony Clark
Faculty Leads: Drs. Ranthony Clark, Gregory Herschlag, Jonathan Mattingly

Motivation

Communities of Interest (COI):
- Groups of people united by similar policy concerns.
- A supermajority of U.S. states list COI preservation as a redistricting criterion.

Existing COI Detection Methods:
- Disagreement on the definition of COI.
- Detection relies on narrative testimony, which rarely has geospatial components.

Goal:
- Develop a data-driven detection method to streamline COI preservation.

Objectives

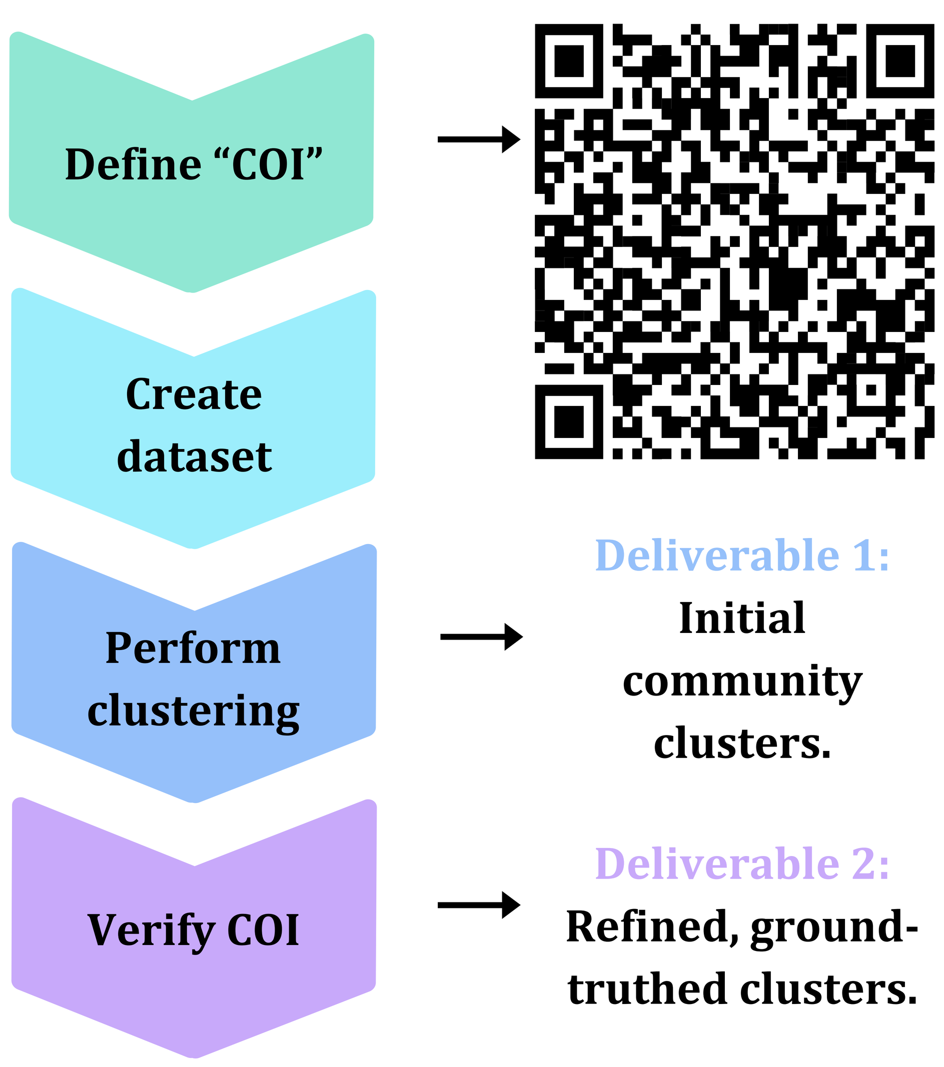

***Figure 1.* Objectives flowchart.**
For more information on the chosen definition of COI, please follow the QR code.

Data

Demographic:
- U.S. 2020 Census.
- 2023 American Community Survey.

Spatial:
- U.S. 2020 Census - Census Block Groups (CBG).
- OpenStreetMap - roadways, bike paths, places.

Tools Used:
GeoPandas • Pandas • QGIS • NetworkX
CDLib • Pickle • Numpy • Shapely • OSMnx
SciPy • Scikit Learn

Methodology

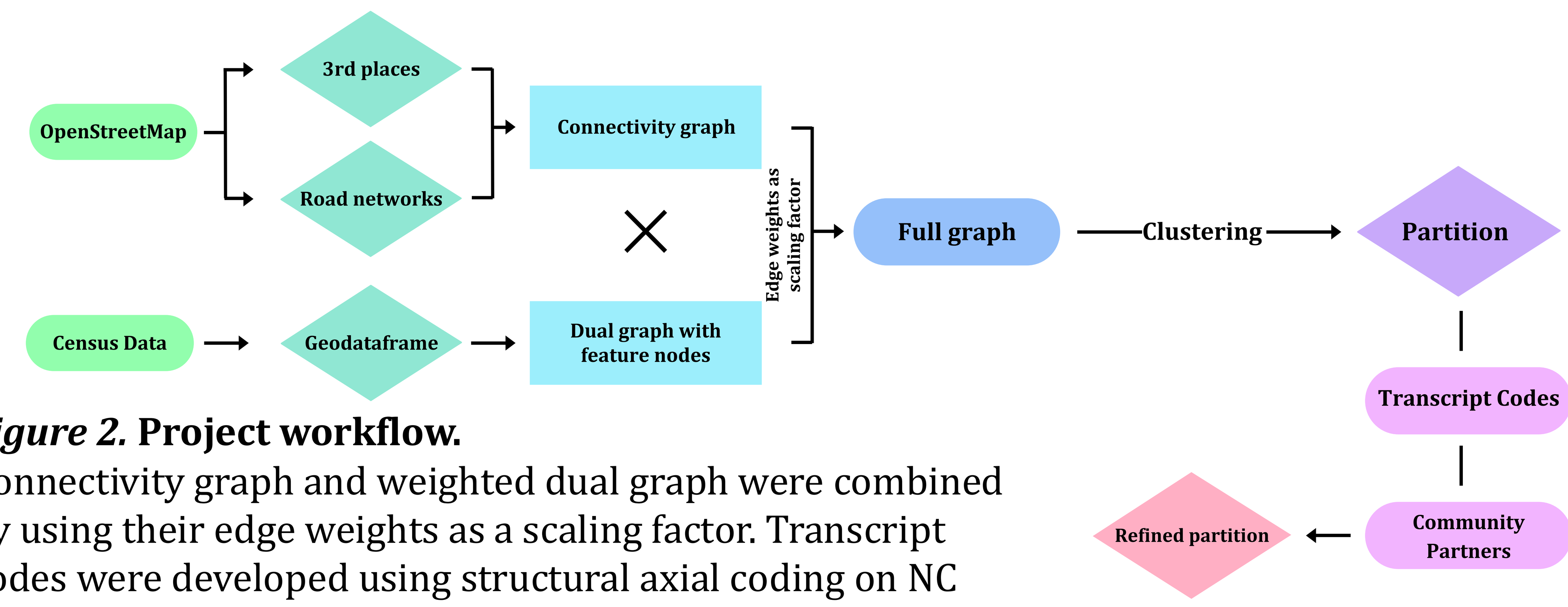

***Figure 2.* Project workflow.**
Connectivity graph and weighted dual graph were combined by using their edge weights as a scaling factor. Transcript codes were developed using structural axial coding on NC redistricting hearing transcripts.

Results

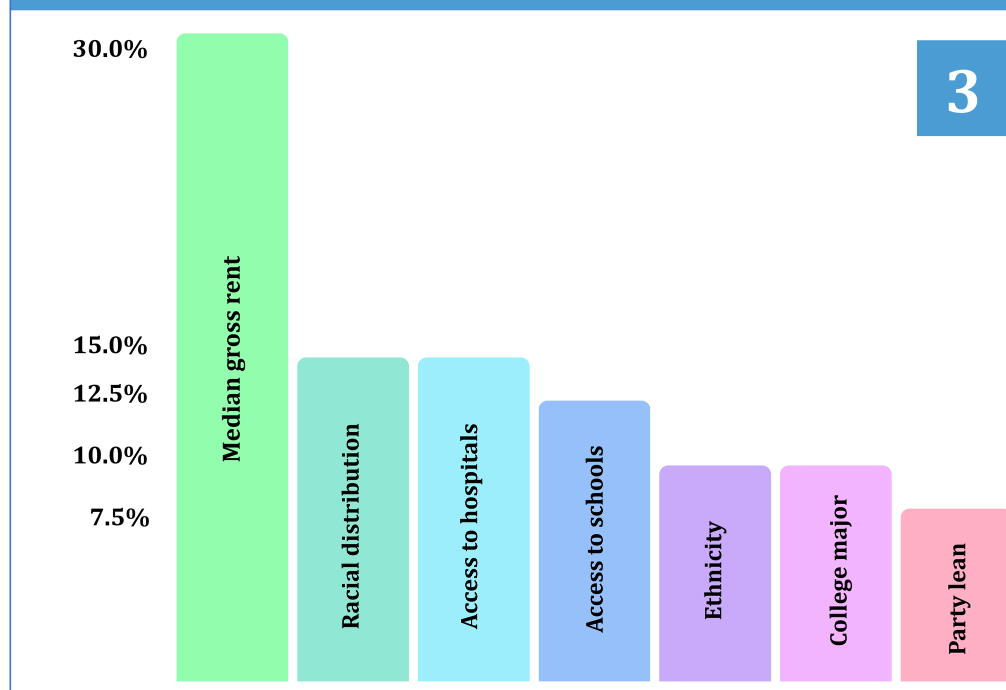

***Figure 3.* Chosen weights.**
Edge weights were determined using transcript codes. More frequently emphasized aspects of community received higher weights.

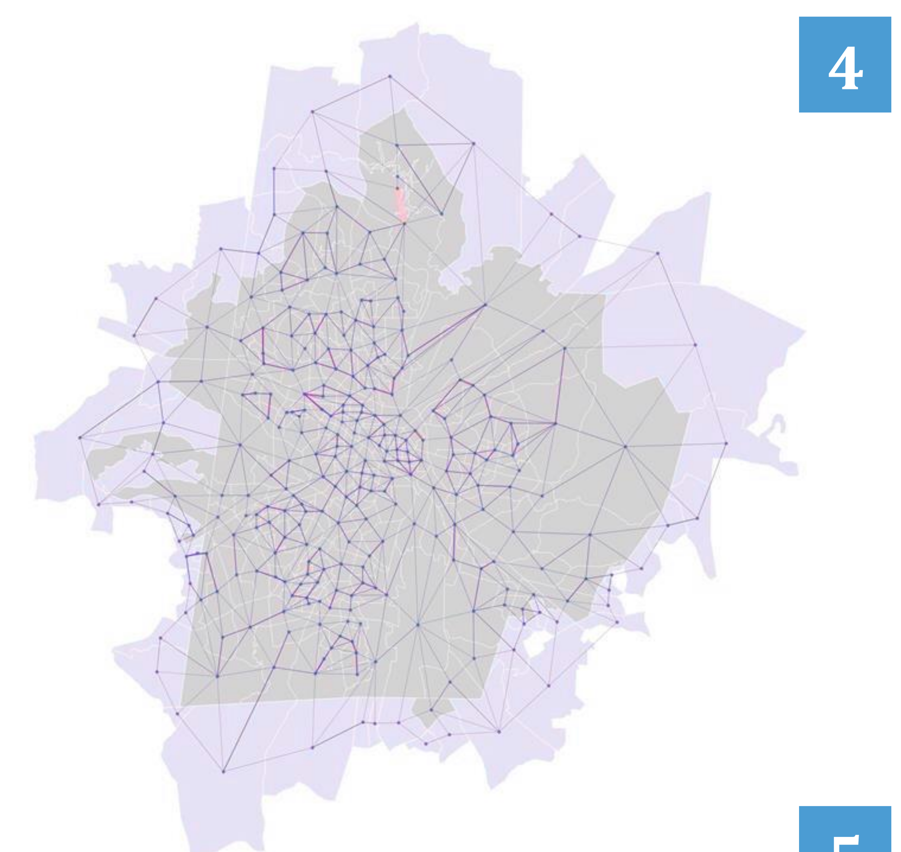

***Figure 4.* Connectivity network.**
The connectivity between the centroids of CBG, found via random walks using OSM road, bike path, and sidewalk coverage data.

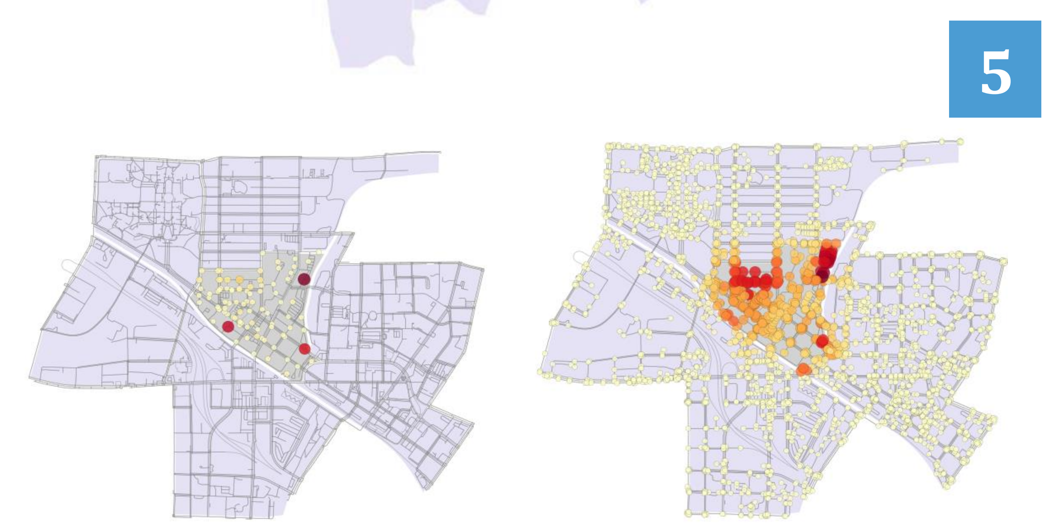

***Figure 5.* Diffusion Process.**
A closeup of the random walk to find connectivities between CBG. A buffer was added to counter artificially high connectivities of outer CBG.

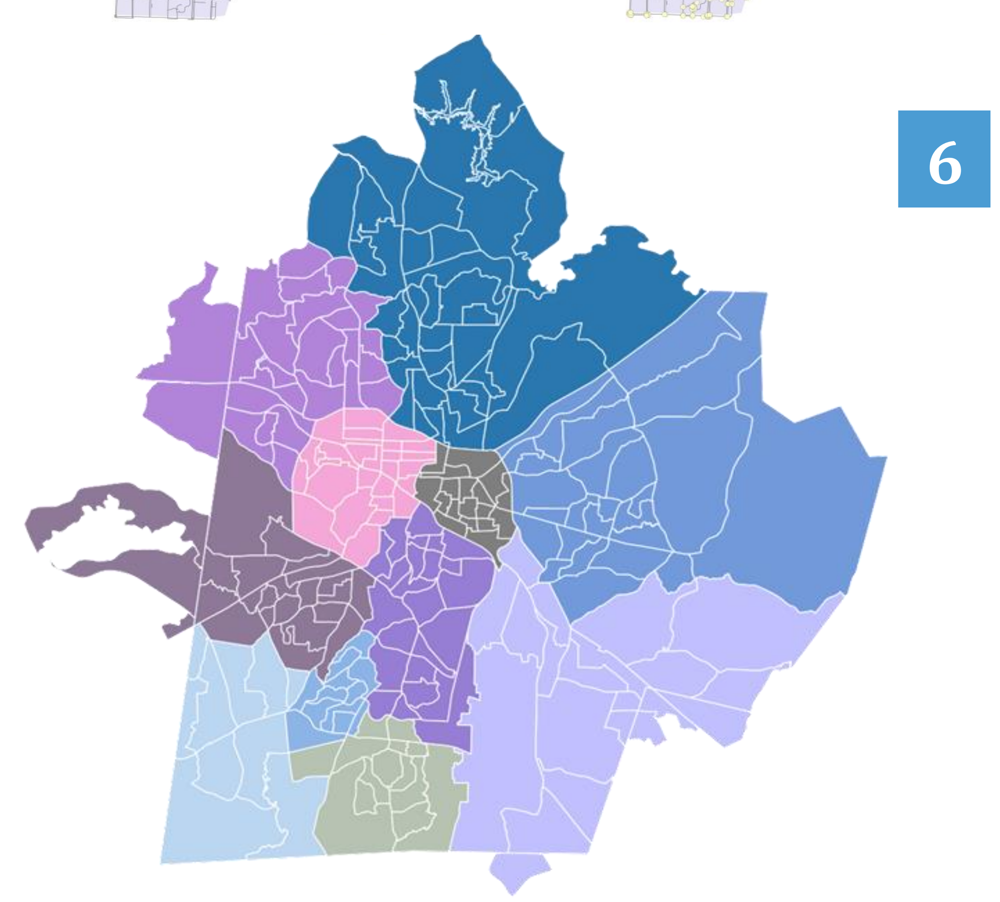

***Figure 6.* Community clusters in Durham, NC.**
Found via Leiden clustering.

Novel Approaches

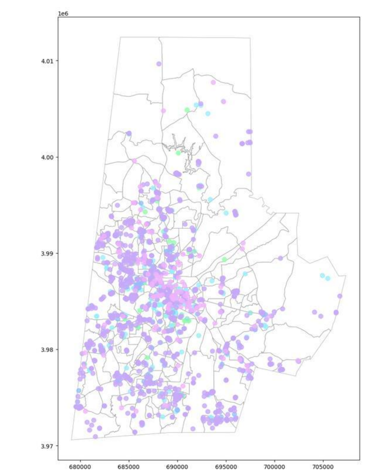

***Figure 7.* Third places in Durham, NC.**
Places of worship, schools, recreational areas, parks, libraries, and hospitals were considered demographic features.

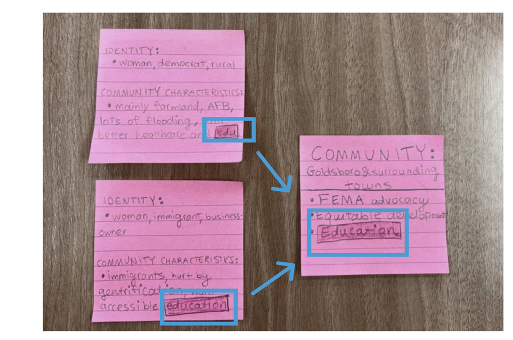

***Figure 8.* Transcript coding.**
Ground truths were built into the project workflow to ensure compatibility with lived experiences.

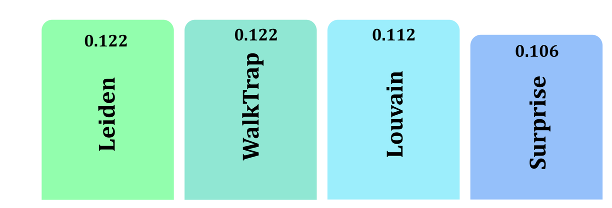

***Figure 9.* Link modularity score for Durham clusters.**
Consistent scores across algorithms.

Future Work

Connectivity Parameters
- Expand novel stop criteria for the random walk.
- Compare random walk vs. shortest path for improved travel decision modeling.

Analysis Over Time
- Measure the influence of housing/highway development on demographic statistics.
- Test which features are most sensitive to time.

Fuzzy Clustering
- Contrast dedicated fuzzy algorithms with noise-augmented crisp algorithms.
- Understand the practicality of fuzzy algorithms vs. overlapping algorithms.

Acknowledgements

The team members extend their gratitude to the Data+ Program, to Dr. Nicholas Eubank, to Dr. Thomas Weighill and his 2025 REU students, to community partners Kids Voting Durham, Common Cause NC, and NC Counts, and to the Data Science for DPS Operations and Planning team for their encouragement.