**GEOG498G Project Proposal**

**Background and Problem**

Human mobility is an important, highly dynamic factor to consider in urban planning problems. Activities that require travel through physical space -- like commuting, dining, and tourism -- are forces that drive a city. Spatial interaction (SI) models, commonly used by academics and urban planners alike, provide a framework for analyzing the varied dimensions of human movement. Among other applications, these models have been used to “optimize” cities, raising the natural question of what makes a city “optimal.” The COVID-19 pandemic, for example, has revealed that previous notions of optimality required citizens to be too close together and brought about rapid disease transmission.

With new sources of spatiotemporal data, we are now able to leverage insights from urban activity data with a higher temporal volume and spatial coverage than ever before. Yet such a fine resolution and detail often comes at the cost of computational complexity and fitness for use in urban models. One option that balances these concerns is to aggregate these individual level data in a way that ensures anonymity, maximizes data quality, and maintains tractability for real-time decision-making. This aggregation problem is known by many names (such as region-building or spatial clustering) and algorithms that solve it are called regionalization algorithms. Therefore, the overall goal of my project is to interface SI models and regionalization algorithms to better estimate human movement patterns and reimagine criteria for building optimal cities while balancing data quality, privacy, and representativeness.

**Methodology**

SI modeling provides a conceptual and technical foundation for explaining and predicting the flow of human movement (Fotheringham 1989; Oshan 2016). Importantly, SI models require only data at the aggregate level in order to make accurate predictions about movement patterns between places. This can be a mixed blessing: from a privacy standpoint, data aggregation means that an individual’s behavior can no longer be traced back to them; but it raises the new issue of defining appropriate spatial units or aggregation procedures. An ideal definition would successfully anonymize the individual level data and create a set of regions that accurately represents the communities at the heart of important policy goals. Algorithms to accomplish this task fall under the umbrella of regionalization.

Regionalization is the process of spatially “aggregating areas into homogeneous regions” (Duque 2012) subject to various constraints, such as contiguity, number and scale of aggregated regions, or determining which areas need to be analyzed. In the context of SI models, regionalization procedures could be used to define novel functional regions -- e.g., central business districts versus predominantly residential districts -- that fulfill multiple criteria, potentially diverging from traditional boundaries, such as census geographies. This yields a meso-scale community-driven approach to building meaningful regions as opposed to a purely top-down administrative approach or a bottom-up data-driven approach. Viewing SI problems through this new lens could empower analysts to extract more pertinent insights from movement data and create more diverse solutions to urban problems.

In particular, my project will begin this process of applying regionalization methods to SI problems. As of now, I will use mobility data from the SafeGraph/GeoDS partnership to power my analysis of the algorithms in question. While these datasets may not be highly meaningful from a social perspective, they provide a solid foundation for testing the ideas outlined in this proposal. In addition, these datasets are very large and require little cleaning or processing, which will allow me to focus on the algorithms more. I will use the Python libraries numpy, pandas, geopandas, matplotlib, geosnap, pysal.spint, pysal.spopt, and potentially more. There may be a need to write custom algorithms depending on what I find in the literature as I read, but I’m unsure as of yet if this is the case. For now, the algorithms I envision using happen to all be coded in these libraries.

One challenge I envision is gathering and setting up the data for this project. This—that is, dealing with data that is not my own—is a personal weakness and something I need to get better at. It will also be important to be thorough in my literature survey so as to pin down what algorithms are relevant and promising versus the algorithms that are more lackluster. I especially want to investigate the preexisting methods which already do the backwards direction of this proposal (using SI data as information for regionalization), as they will have important insights into the linkages between these two modelling philosophies.