M.A. Thesis Proposal: Comparing Theoretical and Data-Driven Demographic Regions

Becky Davies

Planning Meeting: 10/27/16, 1:30-2:30 pm, Gugg 201E

Thesis Committee: Seth Spielman, Fernando Riosmena, Carson Farmer

Summary: My thesis seeks to explore the intersection of theoretical and data-driven approaches to defining demographic regions. The boundaries of demographic data are firm, while regions informed by theory blur across space as the scale, thematic interests, and perspectives of the viewers change. By applying theoretical distinctions of places, potentially including landscape, cultural, social, or other attributes, I will evaluate the resemblance between statistical population units and regions such as neighborhoods and districts.

Given the somewhat arbitrary lines used to form statistical units for demographic data, I will use Spielman & Folch's regionalization algorithm to derive data-informed statistical units (*Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization*, 2015). The algorithm aggregates census tracts into larger regions based on a set of input variables and their associated uncertainty, provided by the U.S. Census Bureau's American Community Survey as margins of error. The algorithm optimizes uncertainty in the data (measured as the coefficient of variation) according to a user-defined threshold. The resulting regions constitute one solution that meets the constraints.

The algorithm yields different solutions based on a random starting point employed. By examining the results of repeat trials within defined study areas, my research seeks to evaluate the stability of regions formed. The stable regions that emerge provide a basis for comparing statistical, data-driven regions to theoretically-informed regions.

Background: Throughout the 2015-2016 school year and following summer, I used the regionalization algorithm to produce results for a variety of scenarios and metropolitan statistical areas. The scenarios include sets of variables each on a particular theme such as housing, transportation, or poverty. Changing the input variables typically changes the output solution.

Thematic Areas

- Neighborhood definition
- Regionalization
- Uncertainty
- American Community Survey

Potential broader questions to address

- What advantages might exist in using data-informed units rather than semi-arbitrary census units (aside from limiting uncertainty)?
- What is the correspondence between census demographics, correlated uncertainty, and community definitions?
- How might features of the urban/non-urban landscape correspond with data-derived units of analysis?

Intial Data Exploration

To test methods to identify stable groups, I first ran 100 trials of the regionalization algorithm using a housing scenario containing four computed variables for the Denver MSA. The result was a list of all of the regions that form. Using that list, I looked how often each pair of tracts were in the same group over the course of the 100 trials. This enabled visualizing the pairings as a network graph.

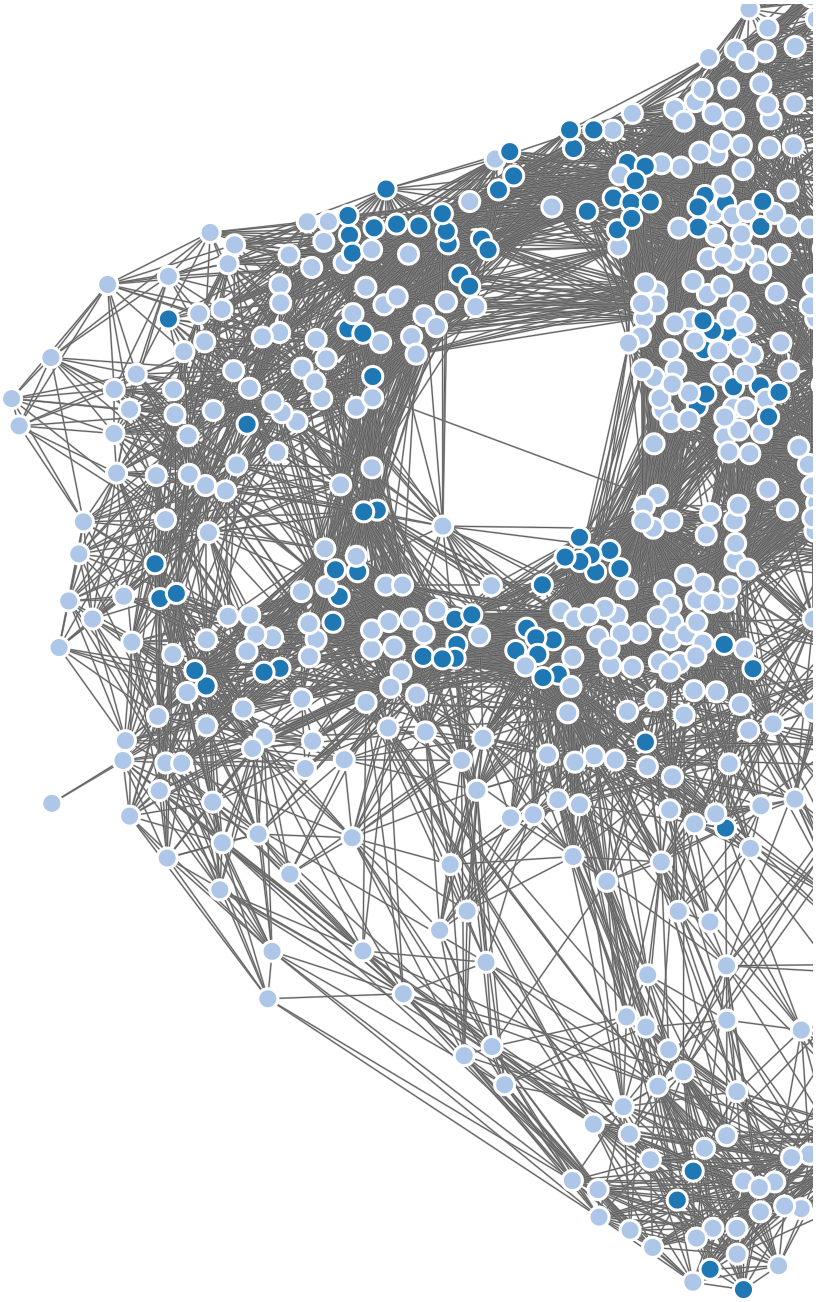
Network Graph View

The network graph below visualizes all of the tracts in the Denver MSA as nodes. The edges connecting the nodes indicate pairings between tracts at least once in the 100 trials. Tracts in dark blue are part of a pair of tracts that occur in the same group at least 90 out of the 100 trials, so they are high frequency/high stability pairs. The graph is very interconnected, so it is difficult to pull any meaning from it initially.

In [1]:
```
from IPython.display import IFrame, SVG, display
%matplotlib inline
import geopandas as gp
from pylab import rcParams
rcParams['figure.figsize'] = 20,20  #set the default map size
```

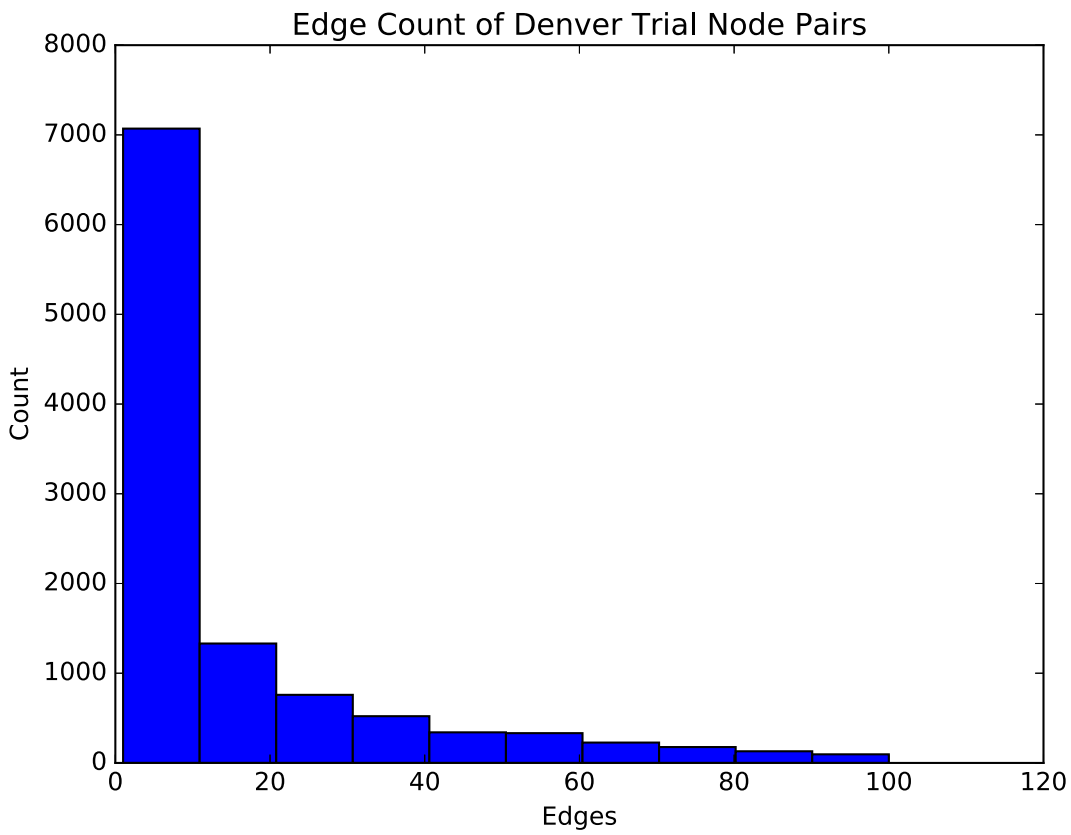
In [2]:
```
IFrame('force/force.html', width=950, height=700)
```

Out[2]:



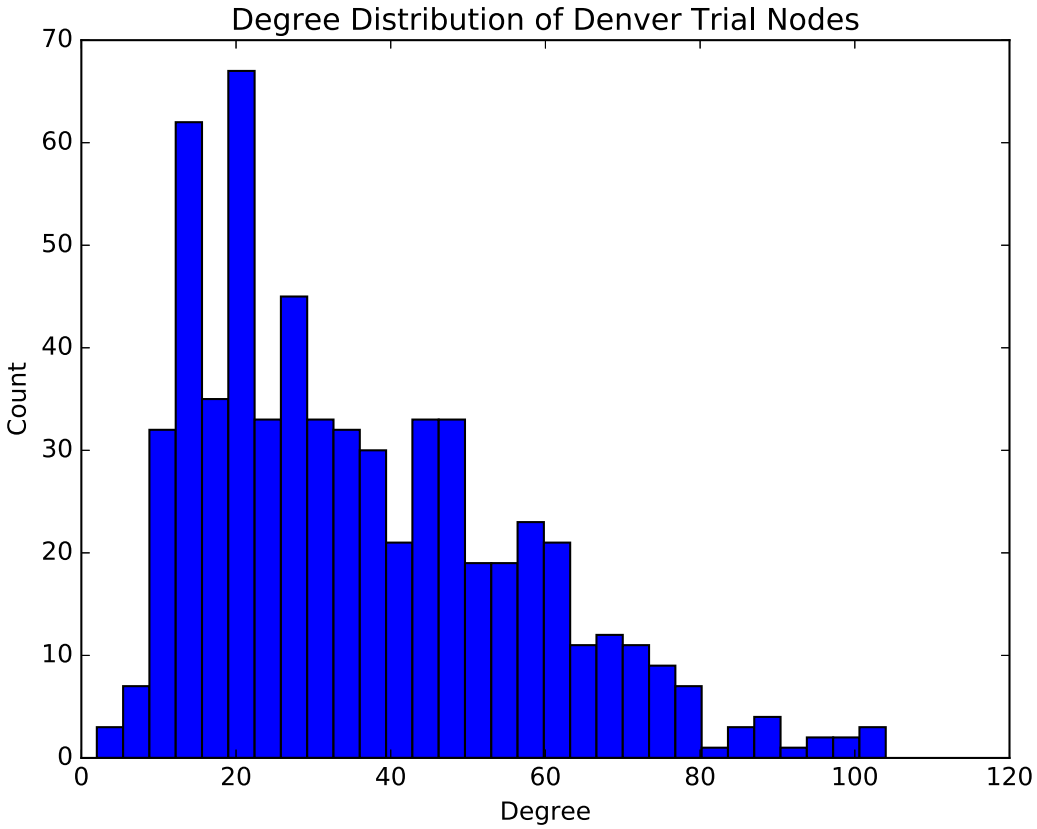
To understand the frequency of the pairings, the histogram below shows nodes by their membership in a pair. The majority of nodes are in pairs that occur 10 or fewer times in 100 trials, meaning they are not stable pairings. A small fraction occur greater than 60% of the time.

In [12]:
```
display(SVG('hist_19740.svg'))
```



Another way to look at the nodes is through their degree, or how many connections they have to other pairs. The nodes cluster around 20 degrees, with a small group of tracts that are in groups with over 100 other tracts and a small number that pair with less than 5 or 10 tracts, indicating greater isolation and thus stability.

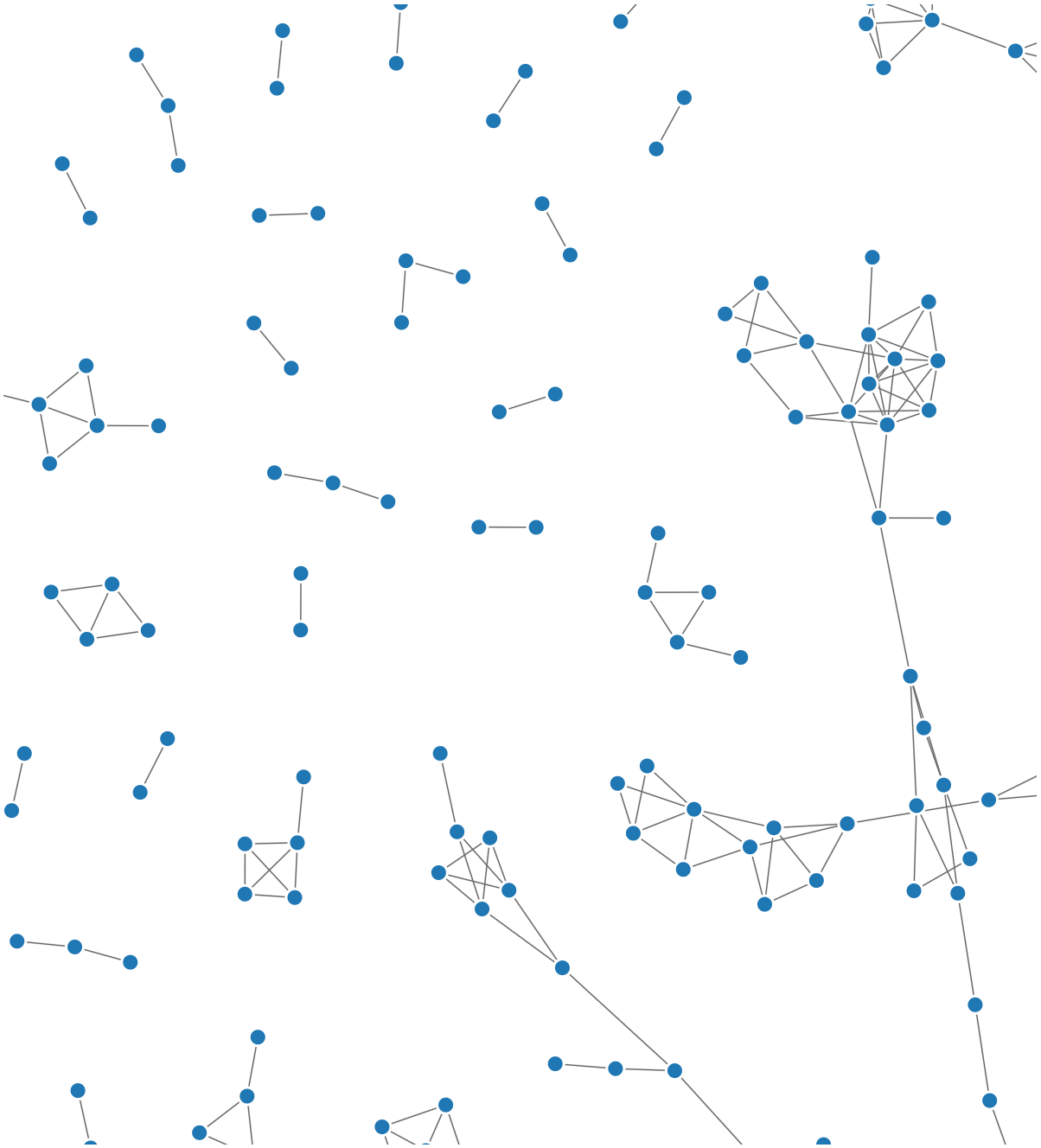
In [11]:
```
display(SVG('degHist_19740.svg'))
```



The three graphs above show low evidence for pair stability, indicating tract groupings aren't consistent and undermining my effort to identify stable groups from which I may be able to extract meaning. However, by removing the tracts that are not in stable pairs, a stronger picture of stability emerges. The network graph below shows all nodes that are in pairs that occur in at least 60 out of 100 trials. More groupings are visible, as well as many single pairs.

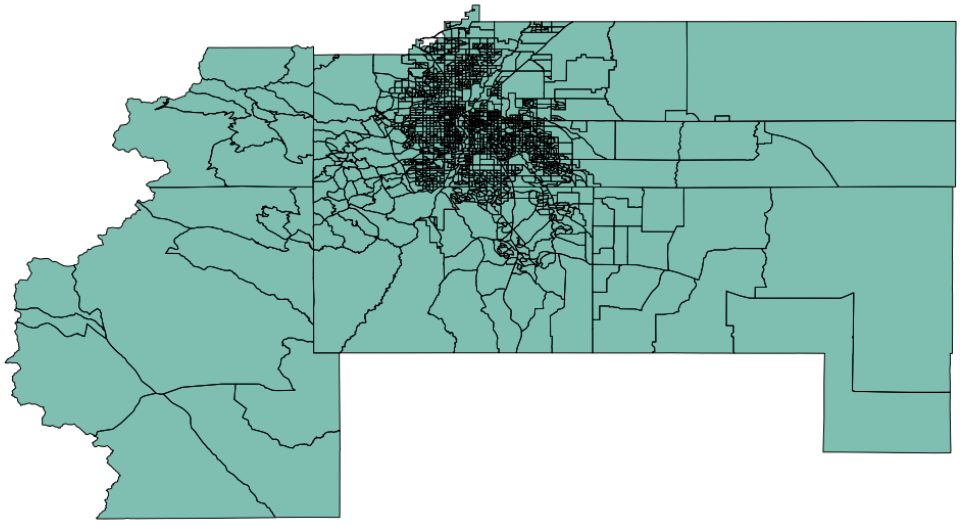
In [6]:
```
# Hover over nodes to see tract ID
IFrame('force/force60.html', width=950, height=700)
```

Out[6]:



In [13]:
```
# Map of Denver MSA
denver = gp.read_file('19740/19740_hous_tracts.shp')
denver['constant'] = 1
p = denver.plot('constant', alpha=0.5, cmap='summer')
p.axis('off')
p.plot()
```

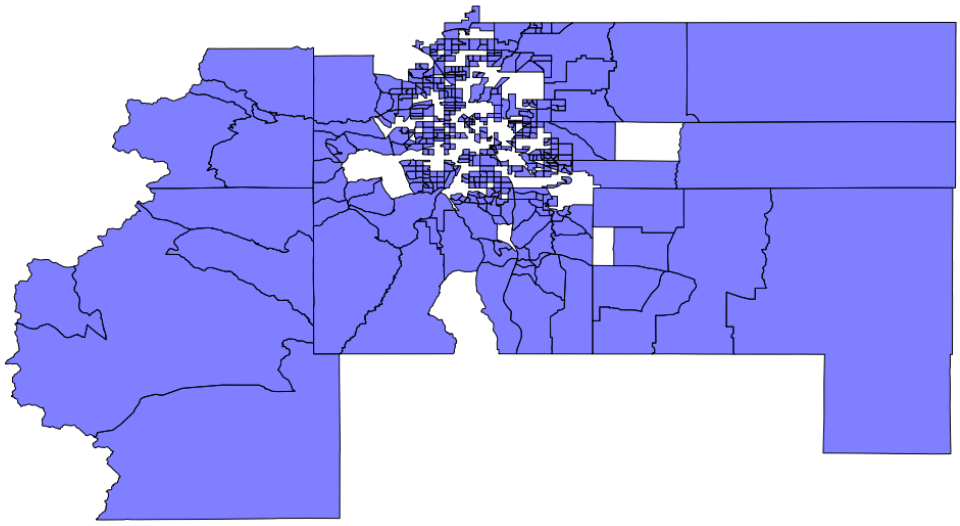
Out[13]: []



The second map, below, shows the tracts that appear in the network graph of >60% pair frequency. Over two-thirds of the tracts are part of one of these high frequency pairs, a more promising outcome for identifying stable groups. In other words, although most pairs occur infrequently, most tracts are also in pairs that occur frequently.

In [8]:
```
freq60 = gp.read_file('freq60.shp')
freq60['constant'] = 1
p = freq60.plot('constant', alpha=0.5, cmap='winter')
p.axis('off')
p.plot()
```

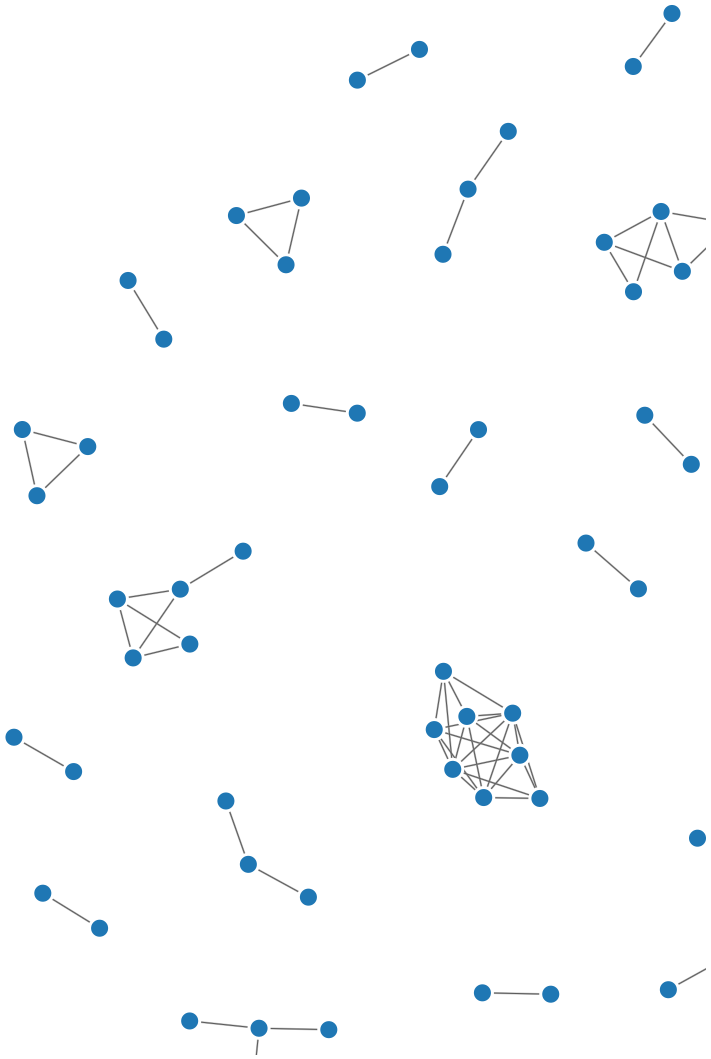
Out[8]: []



By isolating the pairings that occur often, a new story emerges that suggests despite a large number of tract pairs that are noisy or somewhat random, many tracts have at least a modest level of stability. Examining the very high frequency tracts, >90% occurrence, further reveals areas of high stability, though most of these highly stable groups contain only two tracts.

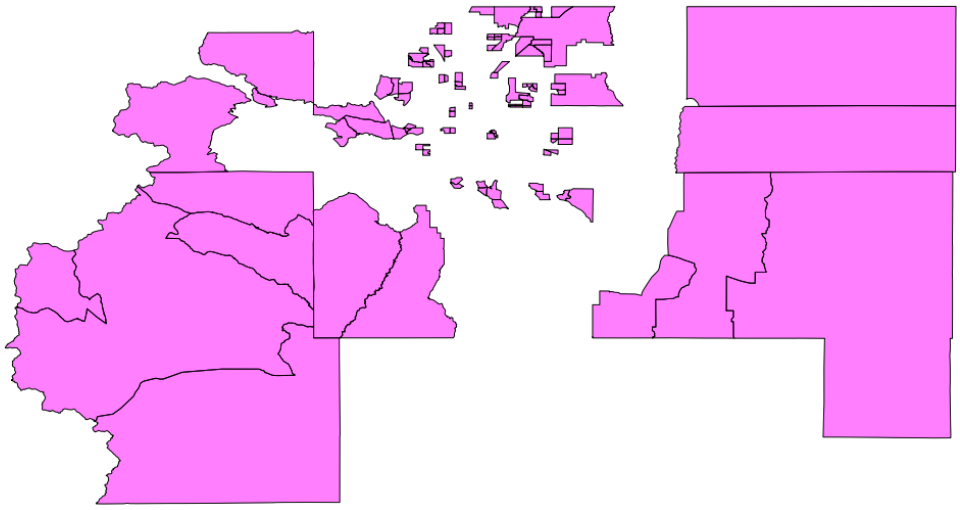
In [9]:
```
IFrame('force/force90.html', width=950, height=700)
```

Out[9]:



In [10]:
```
# Roughly 1/6 of tracts are involved in a very high frequency pair
freq90 = gp.read_file('freq90.shp')
freq90['constant'] = 1
p = freq90.plot('constant', alpha=0.5, cmap='spring')
p.axis('off')
p.plot()
```

Out[10]: []



Running regionalization trials for the same housing variables for the Portland, OR metropolitan area results in a similar distribution of pair frequencies. I intend to run additional trials for other MSAs and other variables to confirm whether the pattern of region formation described in the above example is typical.

Challenges

- Methods to examine group stability, testing variations in the output
- Determining which theoretical attributes to use for comparison
- Potential lack of data-driven region stability for comparison

Future plans

I don't plan on continuing on to a Ph.D. at this time. After completing my master's degree, I'd love to find an analytical role that addresses issues relevant to public resource allocation, which would most likely be in a government or a non-profit/research setting. Realistically, those jobs are scarce so I'll also be looking at roles in data analysis and/or product management in the technology sector.