LiveRamp Checks

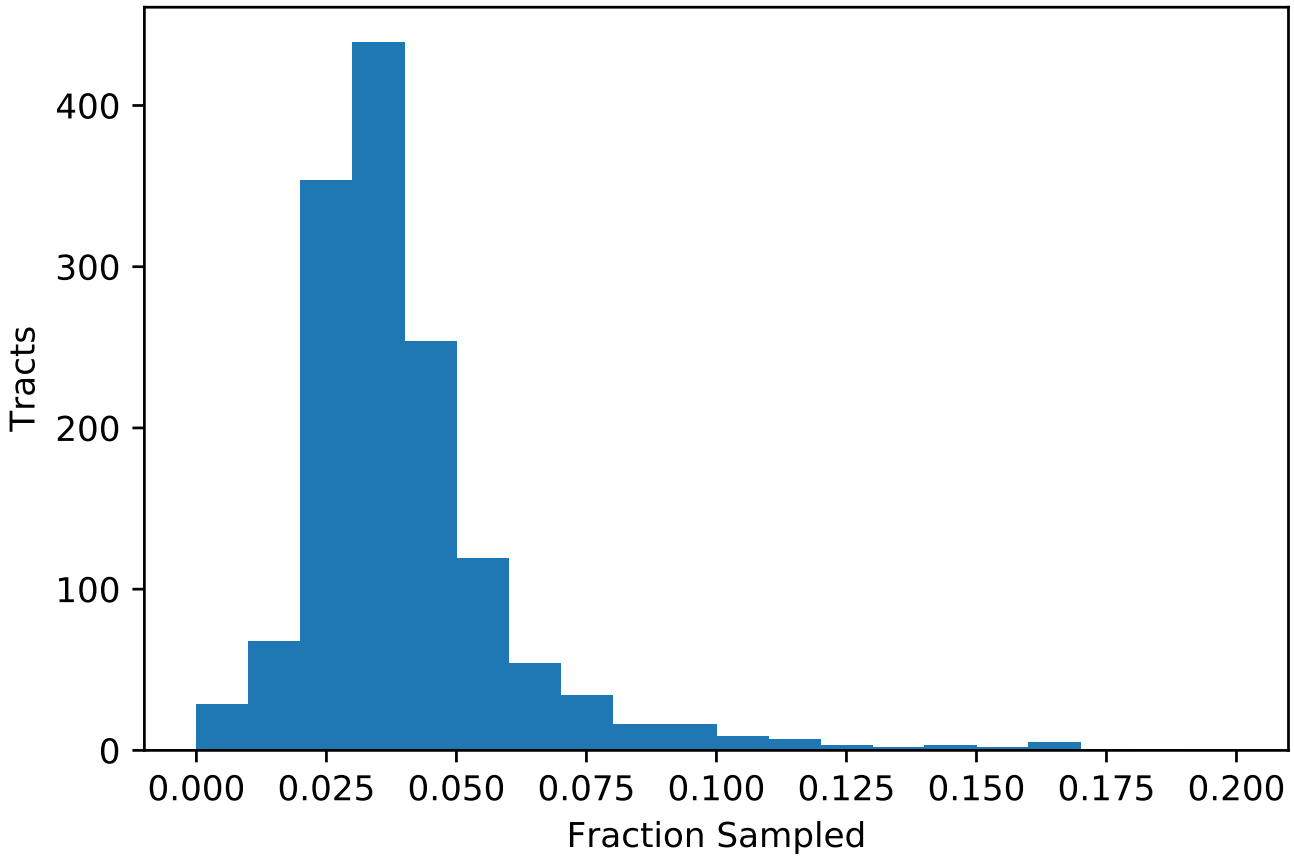
Jamie Saxon

University of Chicago

February 27, 2018

LiveRamp: Sample Size

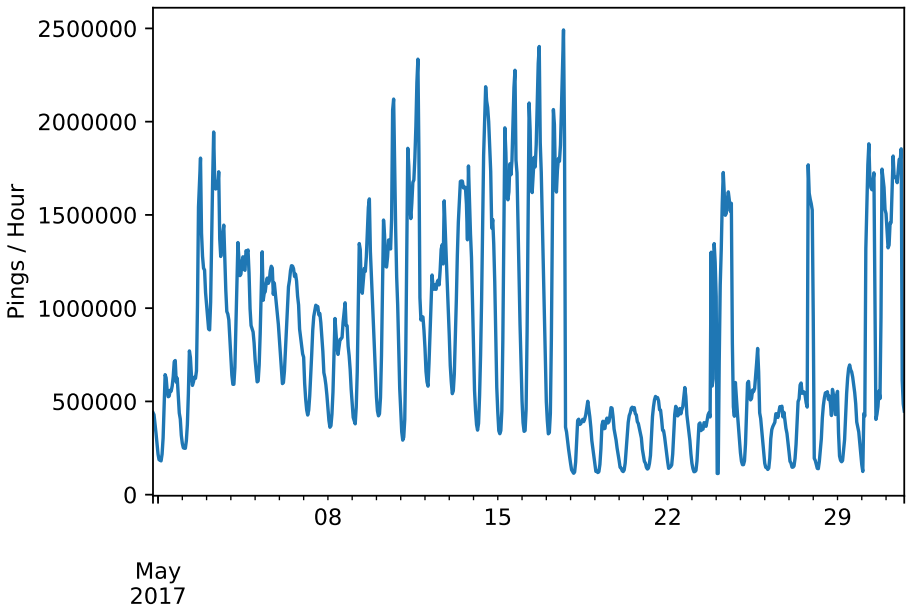
- 1 month of data in the coverage area is $\sim$300k individual users (after some cleaning).
- A $\sim$3% sample across Chicagoland, though quite uneven.
- A one-month sample with size competitive with the ACS!!



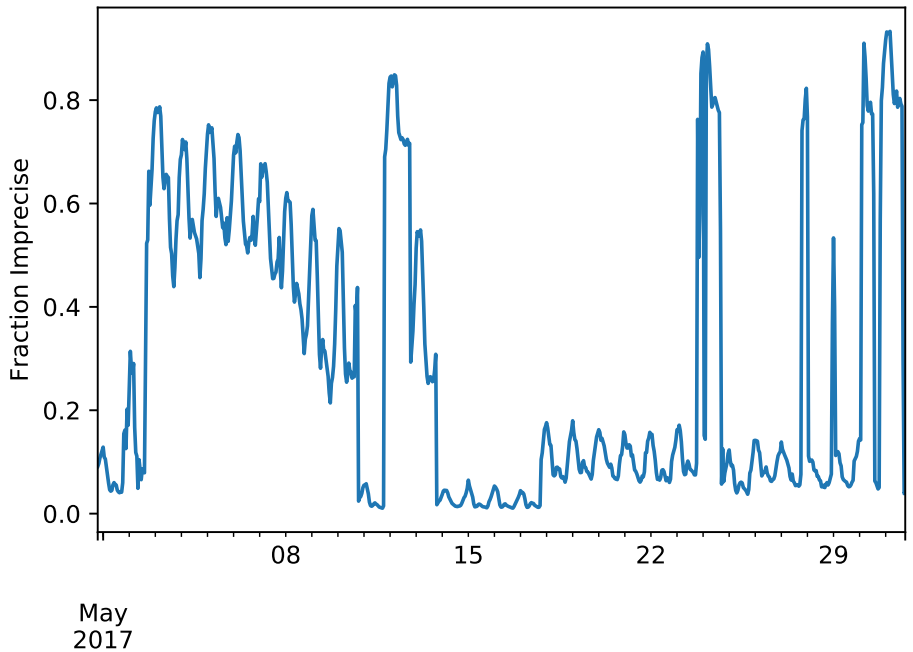
Time Trends

Expected daily structure along with enormous (baffling) variation in the number of pings, and the fraction of precise pings.

- **Am I missing data? Are some apps selling data to Liveramp only occasionally?**



Volume



Precision

Precision

- Roughly 1/3 of all pings in Chicago come from "imprecise" locations (precision = 0).
- In the first 10M lines of Chicago.csv0000_part_00, the top three (floating-point identical) locations constitute 15% of the sample.
  - Locations are not "evocative" – 2 and 3 are sheds in west Chicago.
- These come from across users, OSs, and times, but almost all are flagged as precision = 0. (What does that signify?)

[%]	Lon.	Lat.
5.92	41.8815	-87.6244
4.92	41.7775	-87.7093
4.02	41.8086	-87.7118
1.14	41.9184	-87.7560
1.02	41.7421	-87.6555

Top five locations, all "imprecise."

Low-Precision Apps

- Precision 0 by app is completely bi-modal: close to 0 or 100%.
- It appears to me that a collection of apps are basically not location-aware.
- Are these coming from wi-fi? First IP hop from a cell tower/data center?

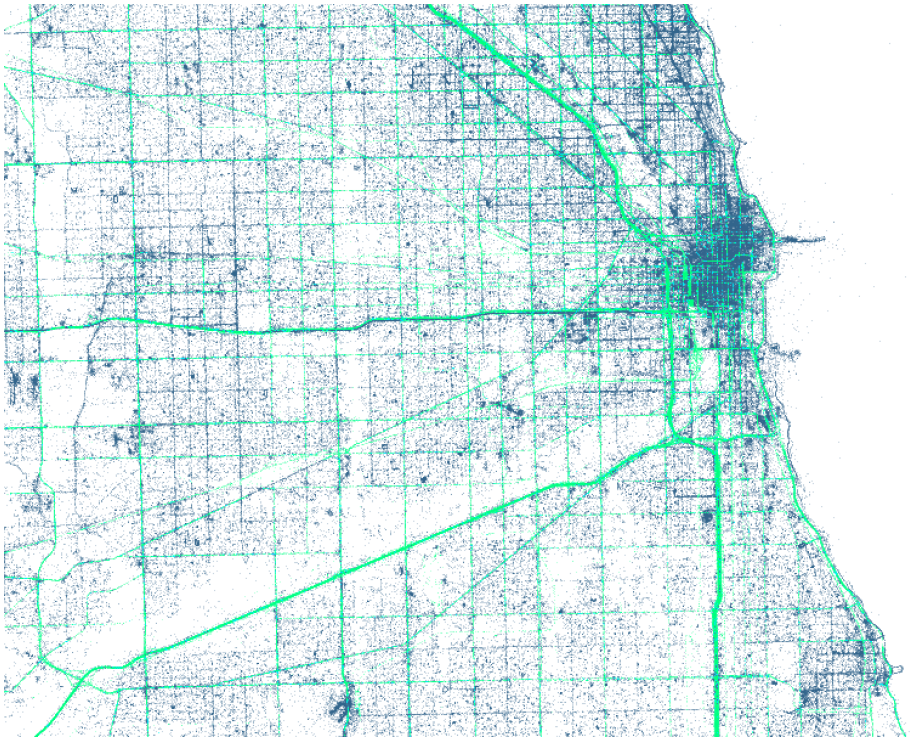
App	Fraction Precise	N
1138	1.00	416857
2728	1.00	77732600
1159	1.00	888986
2003	1.00	4563069
1065	1.00	9696314
1099	1.00	426233
1145	1.00	970297
1018	1.00	134538746
2080	1.00	4307762
1079	1.00	604798
3418	1.00	1192838
1172	1.00	620020
3251	1.00	32732984
1153	1.00	447012
2748	0.99	921448
1156	0.99	86468679
1003	0.91	36123869
1154	0.06	215710982
2076	0.00	2343430
1022	0.00	2251385
1169	0.00	1361618

Precision for Top 20 Apps by Volume

Excluding Highways

Highway traffic represents a huge fraction of out-of-home pings. I am not primarily interested in driving through places on the expressway, so I exclude these from subsequent calculations.

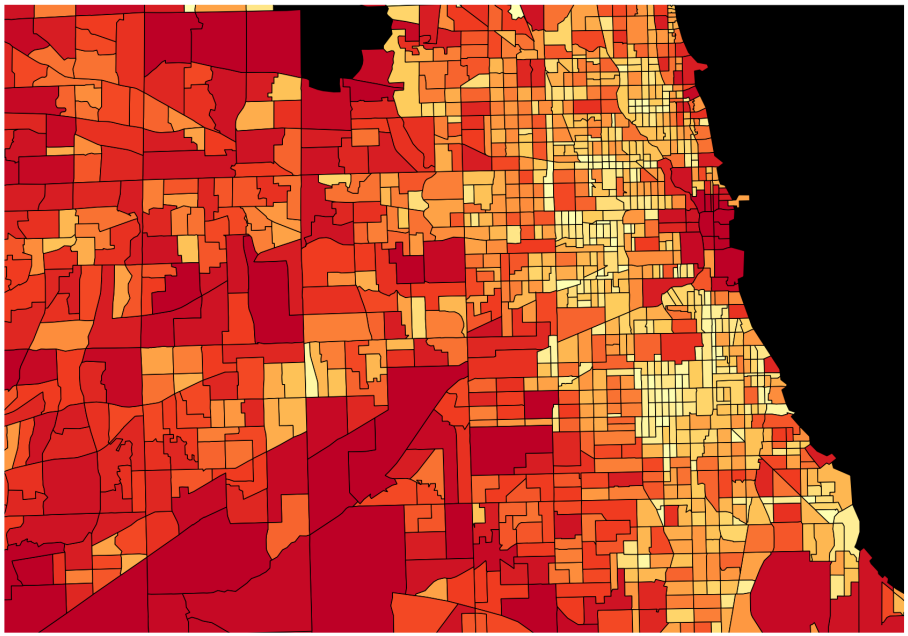
- Run a wider buffer on major highways (using OSM highway type).



Highways merged to a buffer around the OSM highway grid, and other **pings**.

Defining "Homes"

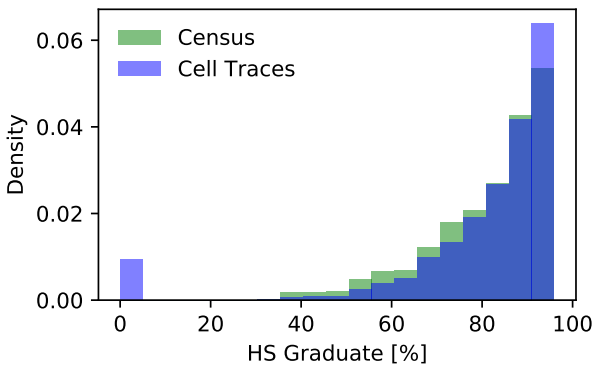
- Point-in-polygon merge to census tracts.
- Call users' modal precise 1am-5am location their "home."
- This is not perfect. Night-staff, especially in transportation hubs – airports, rail yards, and intermodal depots – assigned to work tract.
- Over-represent the loop (college students at Roosevelt, RMU, Columbia College?) and the suburbs. Hypothesis: young user base.

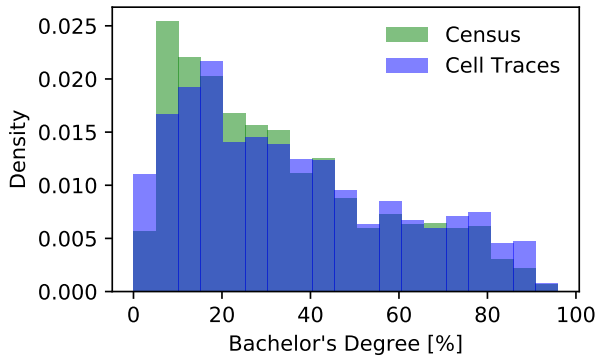


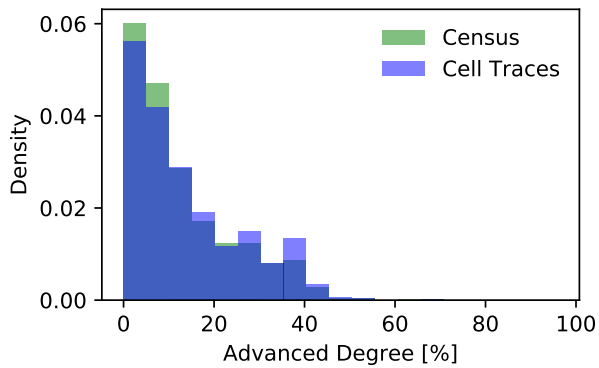
Population: Education

Weight census characteristics by "home" fraction or Census population.

- App users (mobile phone carriers) are, on the whole, more educated.

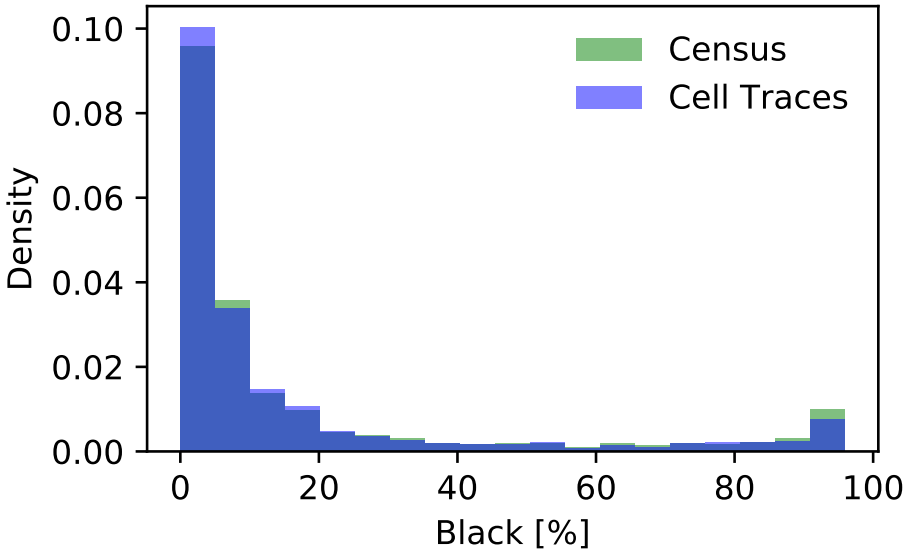


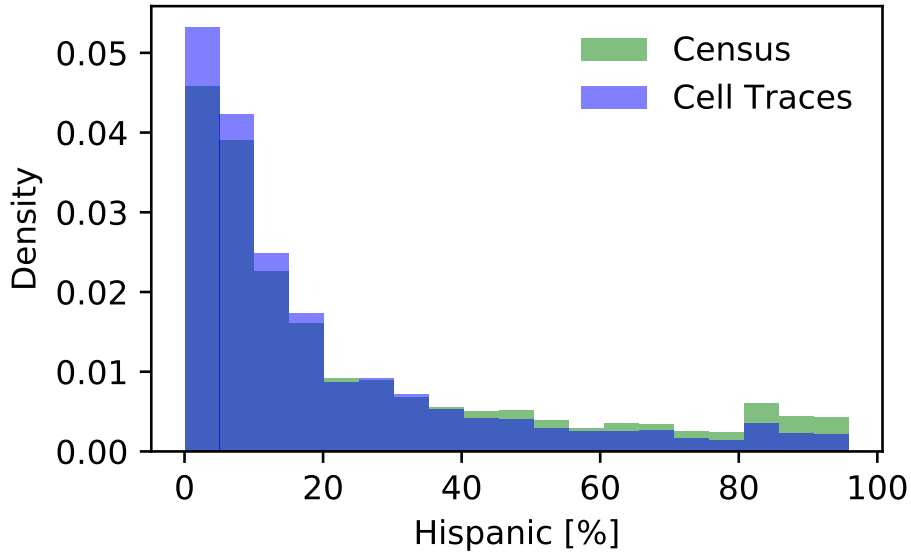




Population: Race/Ethnicity

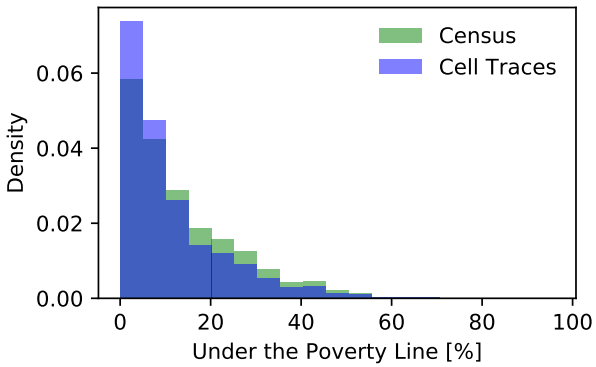
Blacks and Hispanics are both under-represented, but it is more severe for Hispanics.

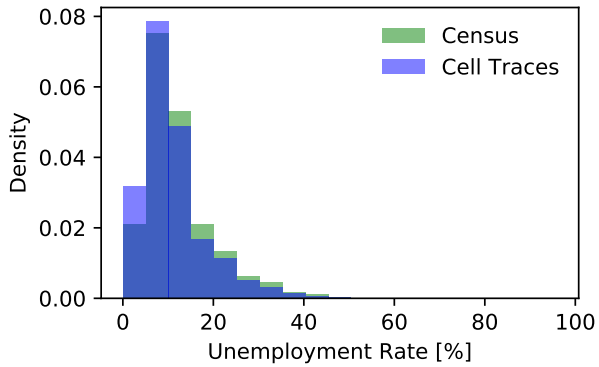


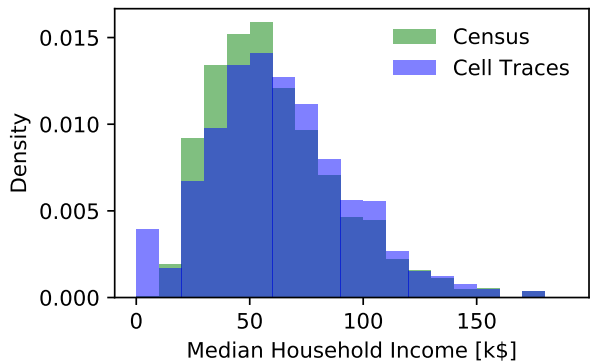


Population: Economic Status

The general population has higher poverty, higher unemployment, and lower median household income than the mobile carriers.

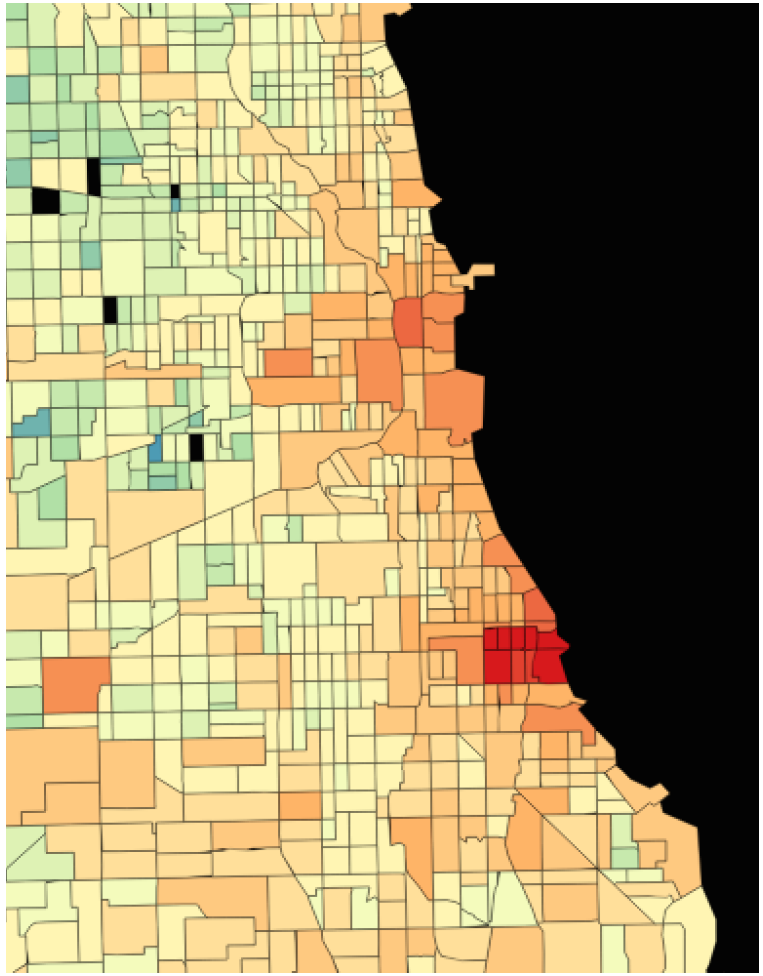






Adjacency Matrix

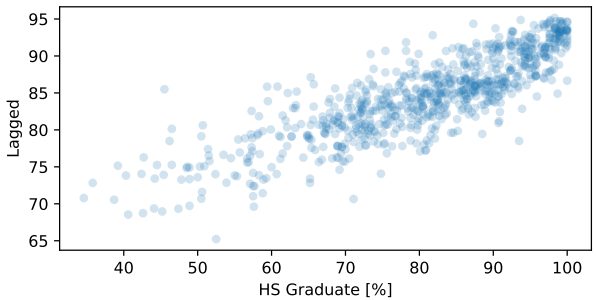
- Construct an adjacency matrix matching home tracts to other neighborhoods in the city, from residents' fractional pings in other places.
- Each user counts for 1 (heavy users are not more important).

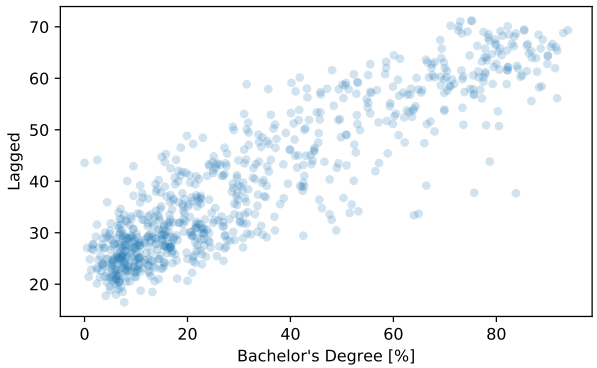


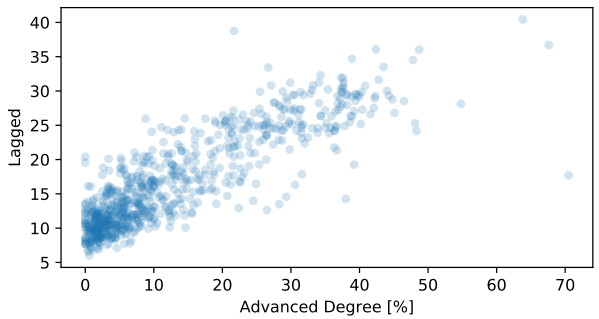
Hyde Park Adjacency

Lagged Rates: Education

- Use the adjacency matrix to construct lagged rates: average over neighbors, not including self.
- Clear trend but also reversion to the mean.
- R^2s are 0.15 for HS, 0.19 for BA, and 0.15 for advanced.



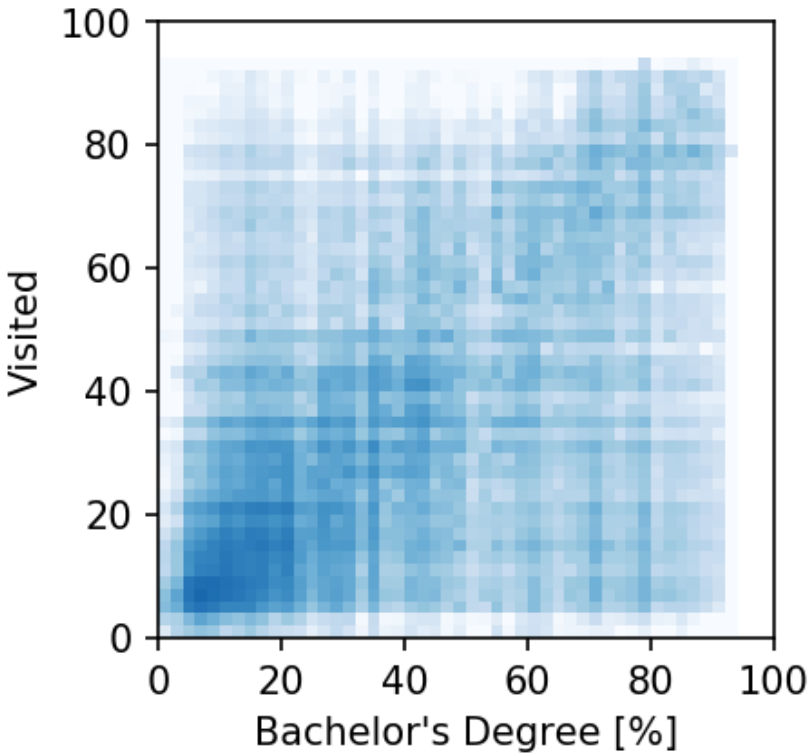


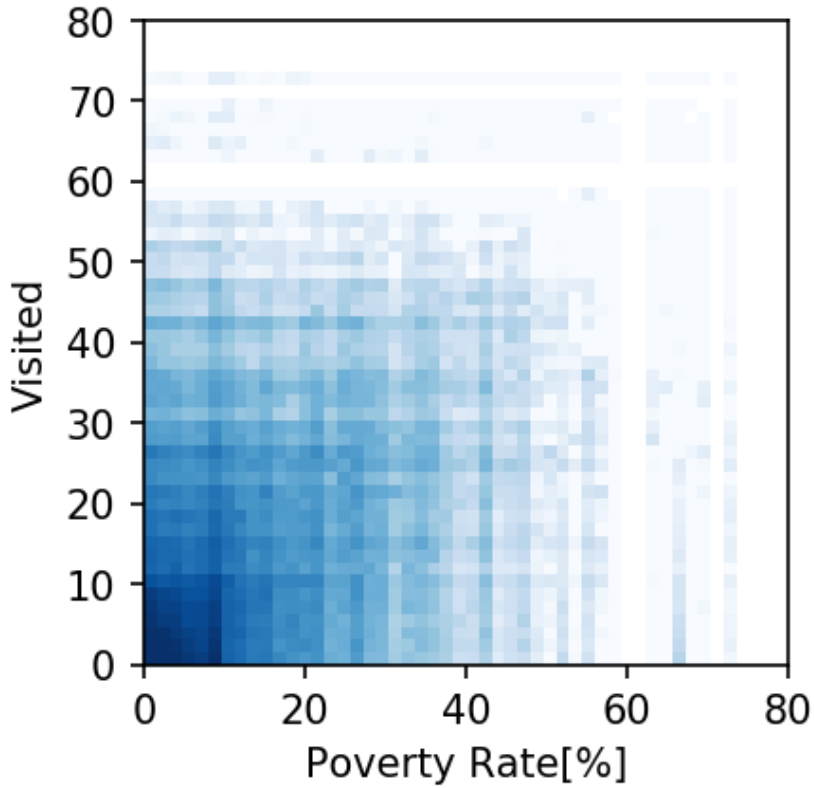


Crossing Divides

We can also just weight crossed places (homes/destinations) by population $\times$ visit weight. This shows the full distribution of individuals who actually visit different tracts.

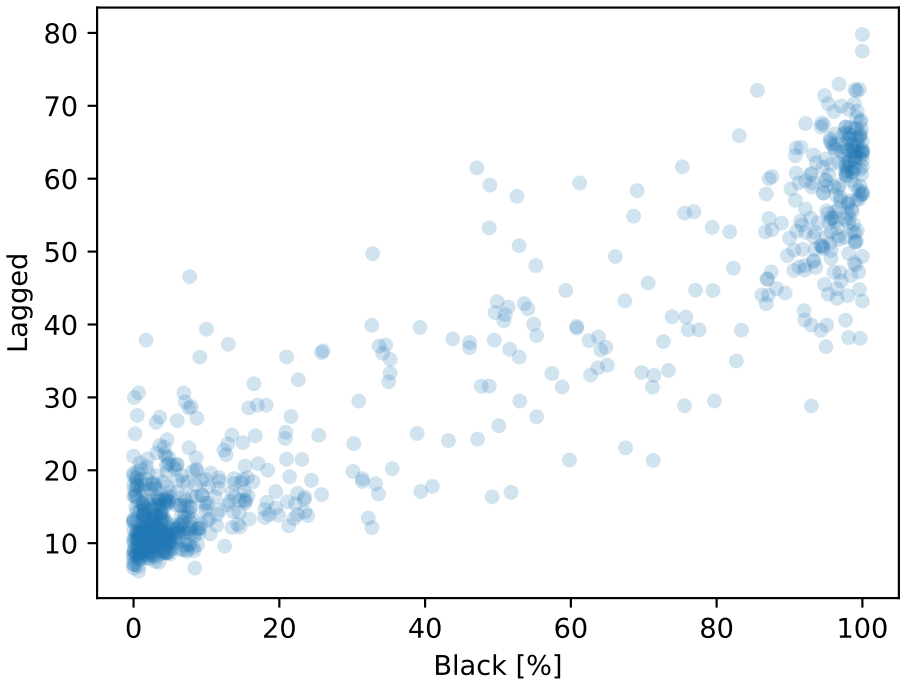
- If you live in a $< 5\%$ poverty tract, do you ever visit a neighborhood with higher poverty?

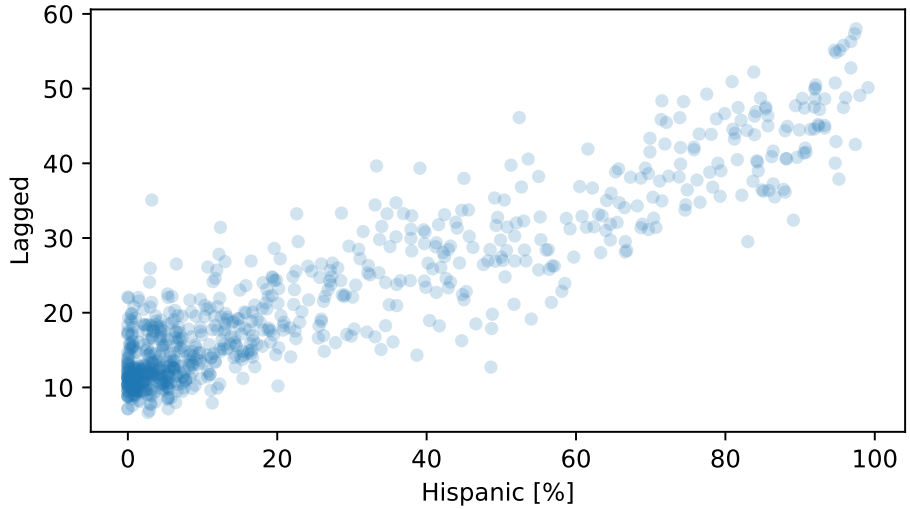




Lagged Rates: Race/Ethnicity

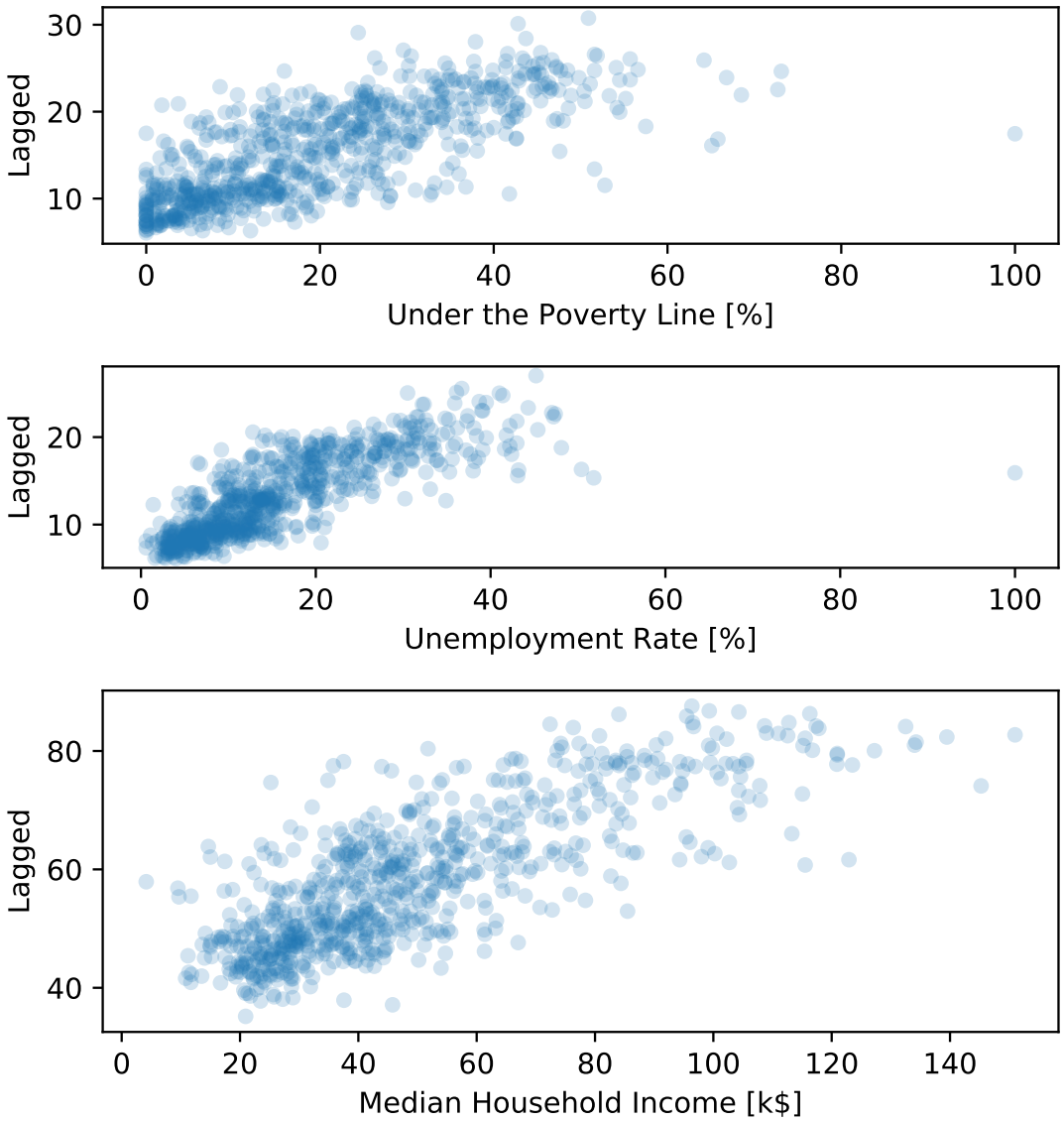
- R^2's here are higher: 0.27 for black and 0.16 for hispanic.
- Note the overwhelming segregation of blacks, and how little whites see black neighborhoods.





Lagged Rates: Economic Status

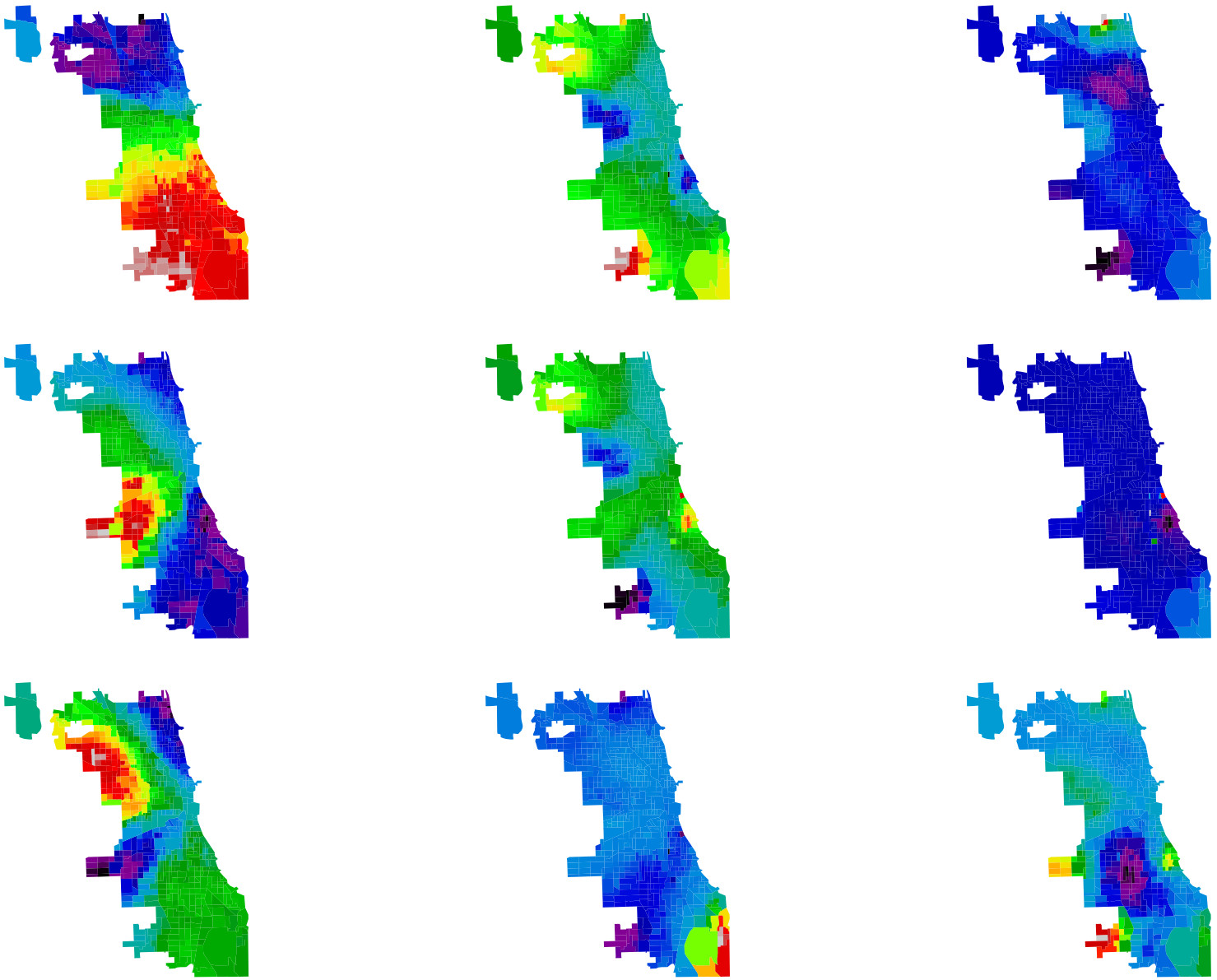
R^2's here are 0.13 for poverty, 0.16 for unempoyment, and 0.14 for MHI.



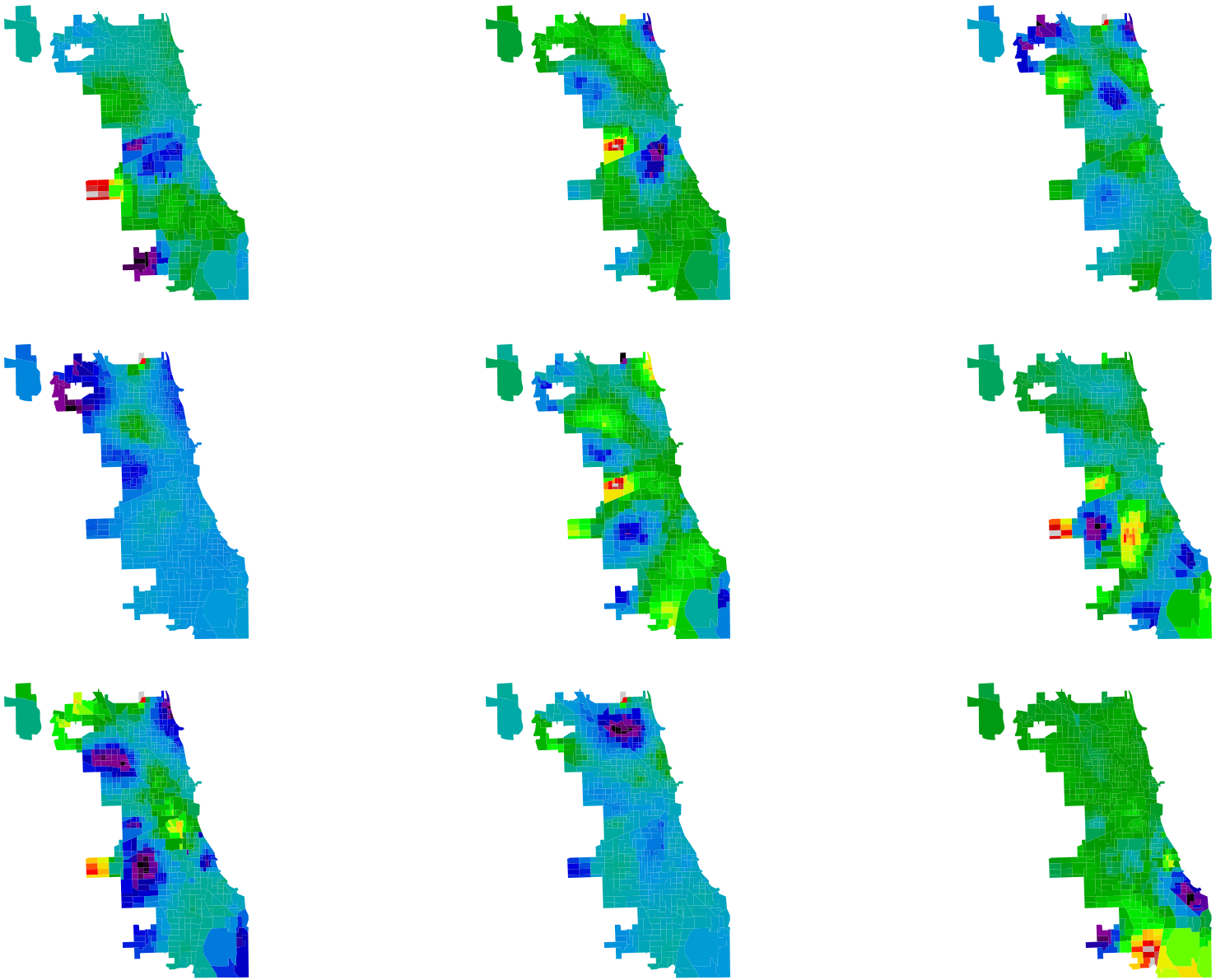
Spectral Clustering

- Form the random walk Laplacian from the adjacency matrix.

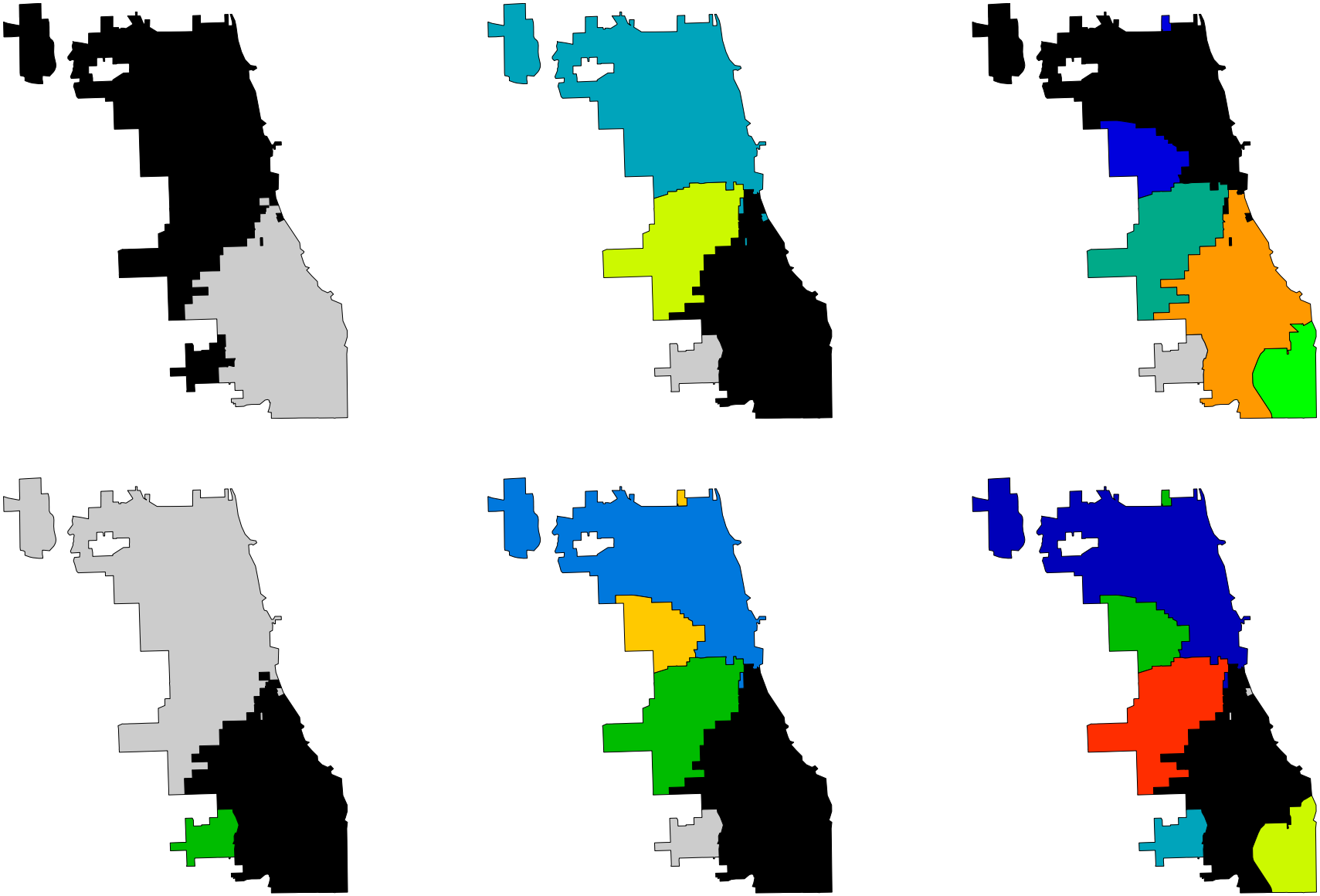
Eigenmodes 1-9



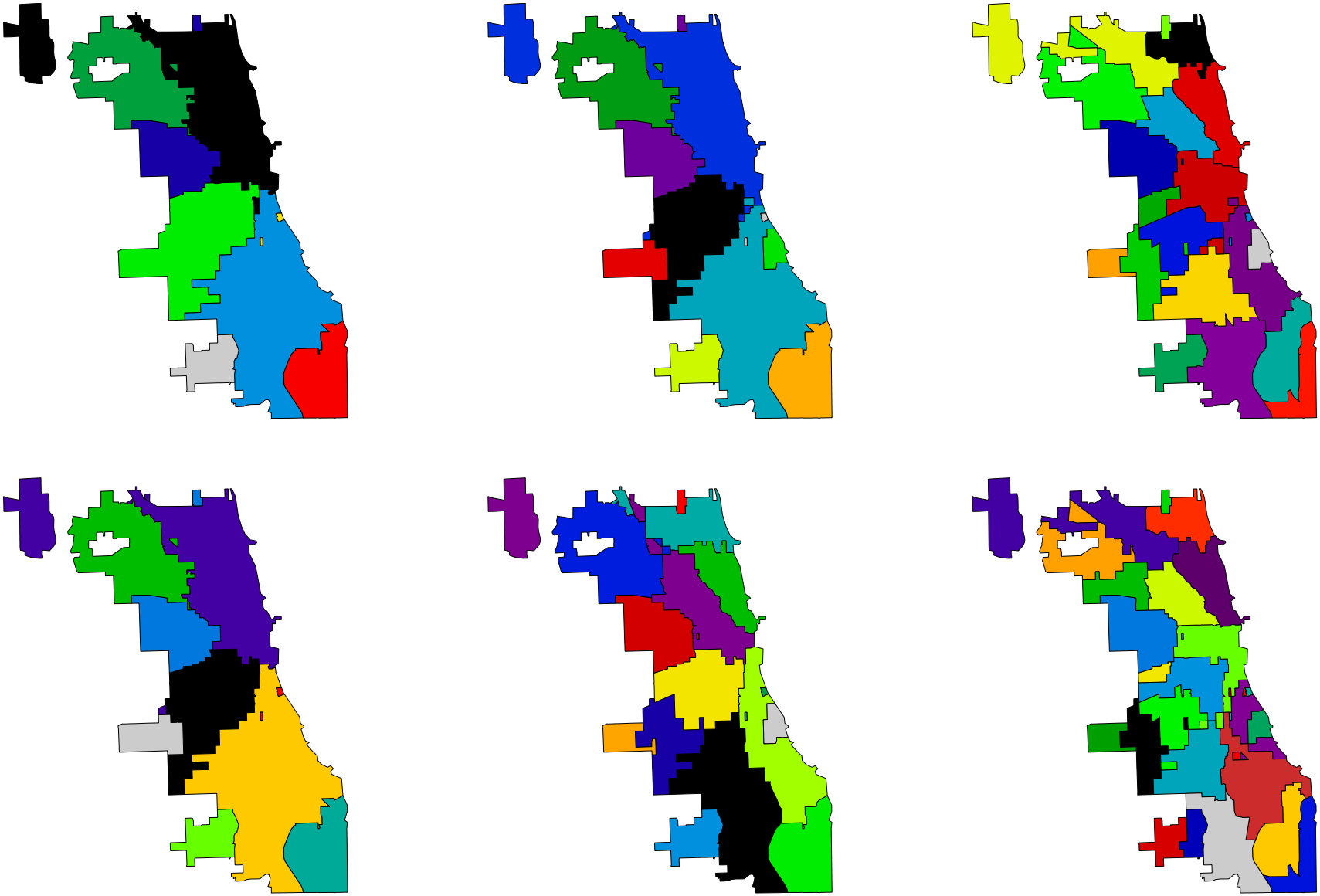
Eigenmodes 10-18



Random Walk Laplacian (2-7 k-means Clusters)

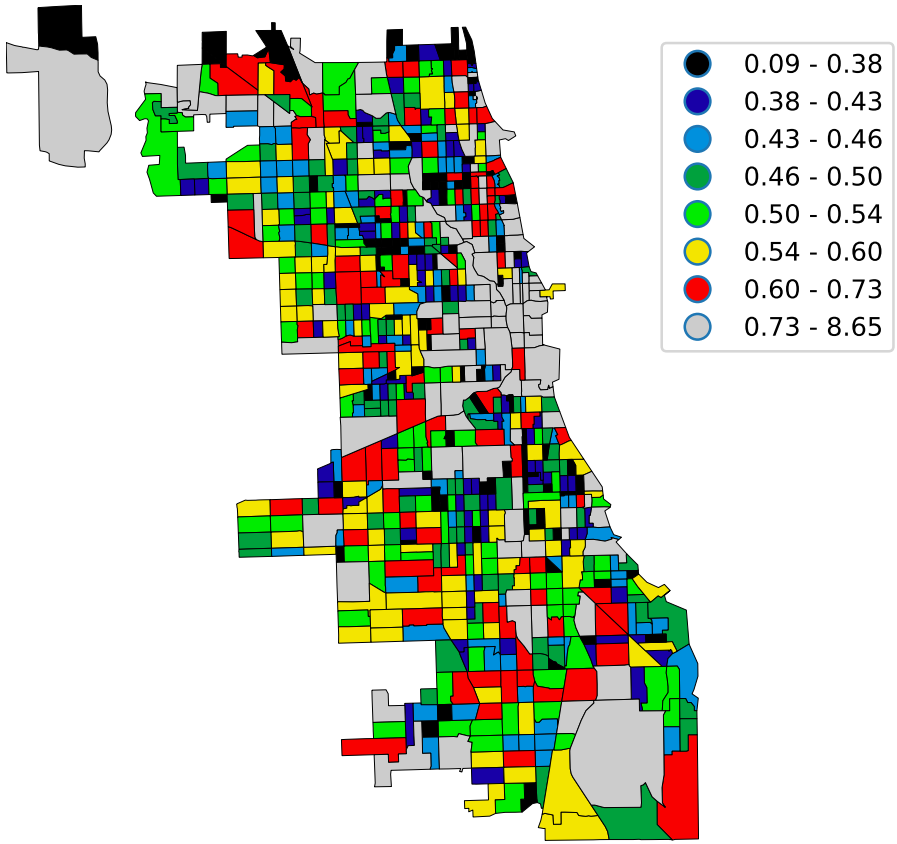


Random Walk Laplacian (8, 9, 10, 15, 20, 25 Clusters)



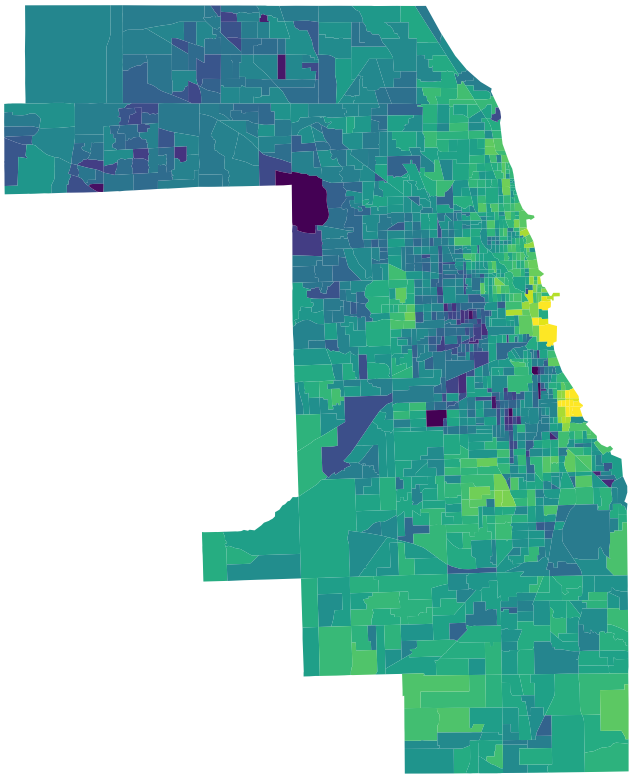
Network Influence / Katz Centrality

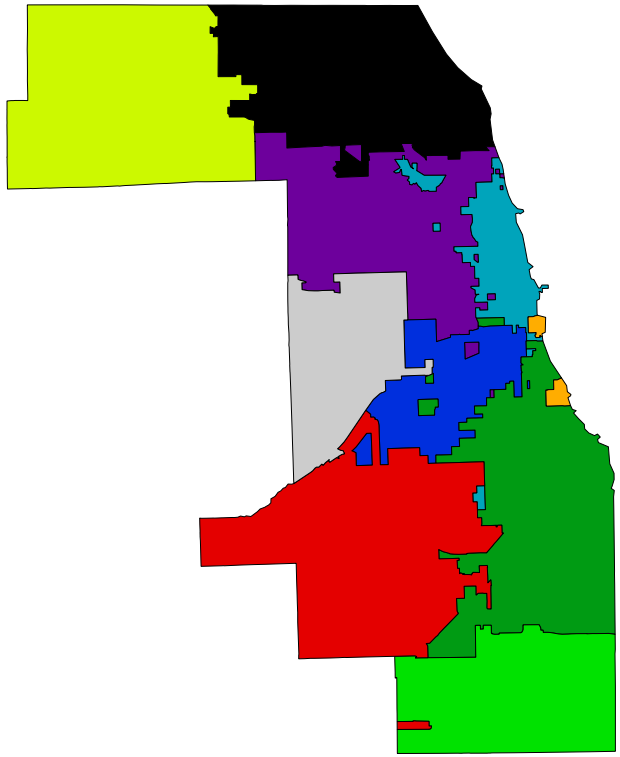
- Highways show clearly (high influence, grey), despite the mask. This underscores the need for the better highway buffer.
- Will drop O'Hare.



LODES Commuting Data

- LEHD Origin-Destination Employment Statistics measure employment and residence from unemployment insurance and Social Security administrative data.
- Where do co-workers live? Consistent picture with traces.





Check out the simple web map for the trace and LODES adjacencies.

Next Up

- Improving highway buffer based on lanes.
- After this is done, correlate centrality to economic status and race.
- Graph properties – triads etc.! Can we get a cut further southwest into Illinois, to compare rural and urban areas?
- Check if users are young: flag schools/students by buffer.
- Does Carto have any more months? More data would help on some sparse tracts.
  - Would be really useful for an IV approach we're considering immigrants...