Identify an interesting set of network data for analyzing and comparing centrality measures across nodes.

There should be at least one categorical variable available for each node

* Identify your data source,
* Create a high level plan that describes how you would load the data for analysis
* Describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

The file nycdoe\_school\_nodelist.csv is a list of NYC public middle schools along with testing and acceptance statistics of 7th graders to the coveted Specialized High Schools. There are comprehensive data sets available with demographic data by school which could be used as a predictive indicator of acceptance to the Specialized High Schools but in this study, geolocation be a proxy for such data.

|  |  |  |  |
| --- | --- | --- | --- |
| DBN | School Name | Testers | Offers |
| 01M034 | P.S. 034 FRANKLIN D. ROOSEVELT | 6 | 0 |
| 01M140 | P.S. 140 NATHAN STRAUS | 6 | 0 |
| 01M184 | P.S. 184M SHUANG WEN | 67 | 23 |
| 01M188 | P.S. 188 THE ISLAND SCHOOL | 0 | 0 |
| 01M839 | TOMPKINS SQUARE MIDDLE SCHOOL | 66 | 14 |
| 02M104 | J.H.S. 104 SIMON BARUCH | 260 | 53 |
| 02M114 | EAST SIDE MIDDLE SCHOOL | 124 | 75 |

nycdoe\_school\_nodelist.csv

The file nycdoe\_school\_edgelist.csv is a list of geolocation relationships of feeder middle schools including their xy coordinates, district number, address, long/lat and zipcode.

nycdoe\_school\_edgelist.csv

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| School Name | x-coord | y-coord | DISTRICT | Addr/long/lat |
| P.S. 034 Franklin D. Roosevelt | 991163 | 203782 | DISTRICT 01 | 730 EAST 12 STREET MANHATTAN, NY 10009 (40.726008, -73.975058) |
| P.S. 140 Nathan Straus | 988749 | 201282 | DISTRICT 01 | 123 RIDGE STREET MANHATTAN, NY 10002 (40.719148, -73.983769) |
| P.S. 184m Shuang Wen | 988287 | 198359 | DISTRICT 01 | 327 CHERRY STREET MANHATTAN, NY 10002 (40.711125, -73.985438) |
| Tompkins Square Middle School | 989351 | 202733 | DISTRICT 01 | 600 EAST 6 STREET MANHATTAN, NY 10009 (40.72313, -73.981597) |

By analyzing the graph structure and relationships, the idea is to locate and rank districts and zipcodes with deficits of students being sent to SHSAT high schools, show relationships with key predictor variables i.e. ethnicity, rank, poverty and 7th grade assessments.

This is the basis for a recommender system showing where the location and intensity of DOE resources could be applied in order to balance access to Specialized High Schools for students throughout the city.

The edges in this database are reciprocal relations two schools in similar location have equal proximity to one-another. The edges are not of an undirected graph.

Finally, the expectation is that schools will form dense clusters where population density is greatest but access to Specialized High Schools will concentrate in neighborhoods that have greater wealth and resources. This is intended to be applied in a recommender model that illustrates where high achieving schools that don’t send many children to these High Schools are and how to get resources to these schools so more such students are admitted.