Ikemefuna Onyeka
A20611235
CS579 - Online Social Network Analysis

Political Communities versus Socioeconomic Communities: A Network Analysis of Voting Behavior in Chicago's Mt. Greenwood and Beverly Neighborhoods

Introduction

Chicago's 77 Community Areas have shaped urban planning and political organizing for decades. Yet a fundamental question remains: do these administrative boundaries reflect actual communities? In my HW4 analysis of Mt. Greenwood (Community Area 74) and Beverly (Community Area 72), I used network analysis of seven census variables to identify 10 distinct socioeconomic clusters. The results revealed striking fragmentation, 30 separate connected components, 18 isolated nodes, and 80% of clusters crossing official CA boundaries. This socioeconomic diverseness suggested these areas are not unified communities, at least economically. However, community can be defined not just socioeconomically but also politically. If residents share political identities and vote similarly despite economic differences, this suggests geographic community cohesion based on social networks rather than demographics. This motivated my research question, do data-driven socioeconomic clusters predict voting behavior better than geographic boundaries?

To answer this, I analyzed Chicago's polarizing 2023 Mayoral Election (Brandon Johnson vs. Paul Vallas) across my 66 block groups, building two parallel similarity networks, one socioeconomic, and one voting-based, I then compared their structures. I also tested for “herd behavior” by examining whether geographic neighbors vote more similarly than demographically similar residents elsewhere.

Design of Project

My analysis employed a three-part comparative framework to test whether socioeconomic or geographic factors better predict voting behavior. First, I constructed two parallel similarity networks using identical methodologies, one based on seven census variables from HW4, another based on 2023 mayoral election voting patterns. Both networks used cosine similarity on standardized features with a threshold of >=0.80, ensuring fair comparison. This approach allowed me to directly compare network structures. The socioeconomic network captured economic and demographic relationships, while the voting network revealed political alignments.

Second, I conducted statistical hypothesis testing to determine predictive power. One-way Analysis of Variance tested whether socioeconomic clusters significantly predicted voting behavior, while variance decomposition compared within-group homogeneity for clusters versus Community Areas. R^2 values measured each model's explanatory power. Third, I tested for herd behavior by comparing voting similarity along two types of edges: spatial edges connecting geographic neighbors versus demographic edges connecting economically similar but non-adjacent block groups. An independent samples t-test determined whether geographic proximity produced greater voting similarity than demographic similarity, controlling for confounding factors.

This design directly addressed my research question by creating competing models and testing them against each other using multiple analytical lenses: network structure, statistical prediction, and behavioral mechanisms.

Data Utilized

My analysis integrated three primary data sources covering my 66 block groups from Mt. Greenwood and Beverly in southwest Chicago. From my HW4 work, I had socioeconomic data from the American Community Survey 2023 5-year estimates at the block group level. I used seven variables: total population (421-2,881 per BG), diversity index (0.01-0.75, calculated using Simpson's Index), median age (25.7-62.2 years), median household income ($31,061-$210,227), unemployment rate (0-21.3%), median home value ($15,700-$758,500), and mean travel time to work (142-1,079 minutes). These variables were already standardized to z-scores in HW4, capturing demographic composition, economic conditions, housing characteristics, and accessibility.For voting data, I obtained precinct-level results from Chicago's 2023 Municipal Runoff Election (April 4, 2023) through the Chicago Board of Elections website. This highly polarizing race between progressive Brandon Johnson and moderate Paul Vallas provided clear ideological contrast. I also downloaded ward precinct boundary shapefiles from the Chicago Data Portal to enable spatial joining.

The spatial join challenge was significant: voting precincts and census block groups use incompatible boundary systems. I adopted a “dominant precinct” approach, assigning each block group to its largest overlapping precinct and using that precinct's full vote totals. This method successfully captured 92% of votes.

Execution of Project

I conducted this analysis in Python 3.12 using Jupyter Notebook, which enabled reproducible, documented analysis with inline visualizations. My workflow began by loading the 66 block groups from my HW4 work, which already contained the seven standardized socioeconomic variables and cluster assignments. I then performed the spatial join between precinct voting data and block groups using GeoPandas. After testing area-weighted allocation methods that failed due to boundary precision issues, I implemented the dominant precinct approach by calculating overlap areas between each precinct and block group, identifying the largest overlap for each block group, and assigning that precinct's vote totals.

With voting data successfully allocated to all 66 block groups, I built the voting similarity network using scikit-learn's cosine similarity function on standardized voting features (Johnson percentage and Vallas percentage). I applied the same 0.80 threshold used in my HW4 socioeconomic network to ensure methodological consistency. NetworkX enabled network construction, metric calculation, and community detection algorithms. I computed basic network statistics including number of edges, density, connected components, and average degree for both networks.

For statistical analysis, I used scipy.stats to perform one-way ANOVA testing whether the 10 socioeconomic clusters predicted Democratic vote share. I calculated R-squared values for both the cluster model and Community Area model to measure explanatory power. For the herd behavior test, I extracted all spatial edges from my HW4 adjacency network and calculated voting similarity for each geographic neighbor pair. I then identified all pairs of block groups in the same socioeconomic cluster but not spatially adjacent and calculated their voting similarity. An independent samples t-test compared these two distributions, with Cohen's d measuring effect size.

Throughout the analysis, I used Anthropic's assistant for code debugging, statistical methodology consultation, and visualization design suggestions. Matplotlib and Seaborn generated all visualizations, including side-by-side cluster maps, network overlay diagrams, choropleth maps of voting patterns, and statistical summary figures. I validated results by checking that allocated votes matched 92% of original precinct totals, testing multiple similarity thresholds for robustness, and visually inspecting prediction residuals for anomalies.

Results

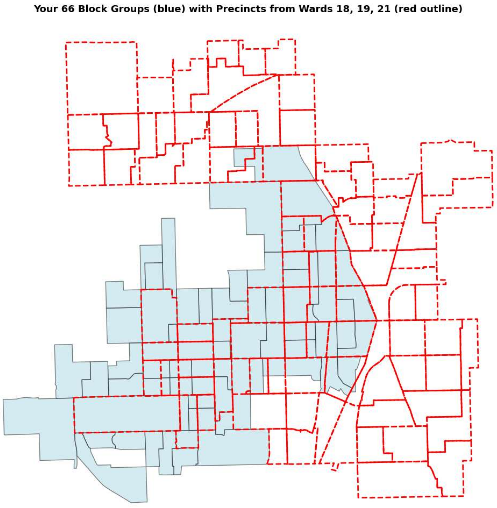
The socioeconomic and voting networks exhibited dramatically different structures despite analyzing the same 66 block groups. The socioeconomic network from HW4 was sparse and fragmented with only 45 edges, a network density of 0.021, and 30 separate connected components, in contrast, the voting network was dense and unified with 1,177 edges (26 times more), a density of 0.549, and only 2 connected components analysis of edge overlap revealed minimal correspondence: only 31 edges appeared in both networks, yielding a Jaccard similarity coefficient of 0.026. The network degree correlation was merely 0.219, confirming structural independence. Block groups that are economically similar are rarely politically similar, and vice versa.

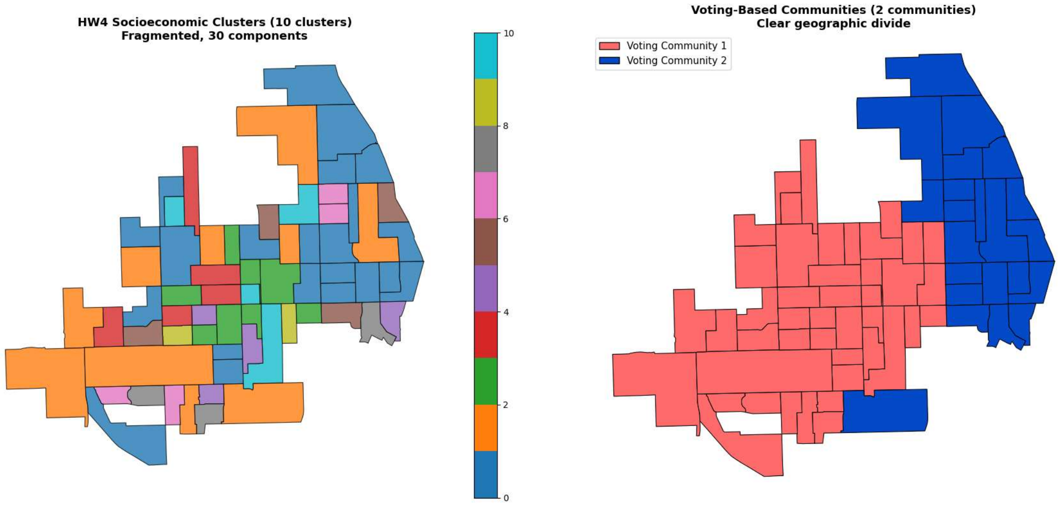
Statistical testing revealed that socioeconomic clusters do not significantly predict voting behavior. One-way ANOVA yielded F(10,55) = 1.25, p = 0.28, failing to reject the null hypothesis that all clusters have equal mean vote shares. The cluster model achieved R^2 = 0.185, explaining only 18.5% of voting variance. Average within-cluster variance was 464.83, nearly equal to total variance, indicating enormous heterogeneity within clusters. For example, Cluster 7 (economically distressed) had mean Democratic vote of 30% but standard deviation of 28%, meaning members ranged from highly conservative to highly progressive. The mean absolute prediction error was 15 percentage points, with the largest error reaching 50 points.

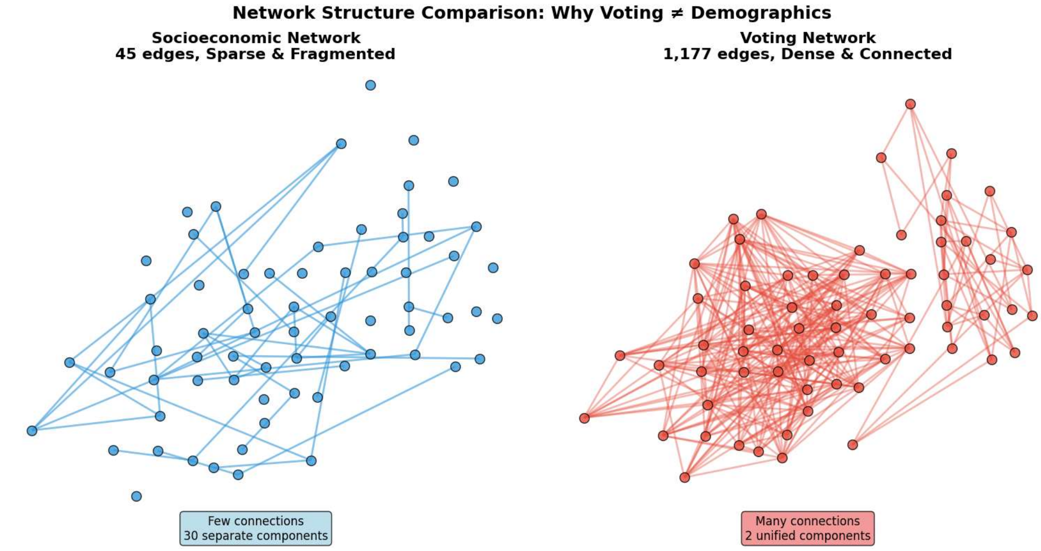
In sharp contrast, Community Areas strongly predicted voting. ANOVA testing Mt. Greenwood versus Beverly yielded F(1,64) = 65.56, $p < 0.001$, unambiguous statistical significance. The CA model achieved R^2 = 0.506, explaining 50.6% of variance, nearly three times better than clusters. Average within-CA variance was only 284.50 versus 464.83 for clusters. Substantively, Mt. Greenwood voted 90% for Vallas (conservative) while Beverly voted 58% for Vallas (moderate), a 32% point gap indicating a clear geographic political boundary.

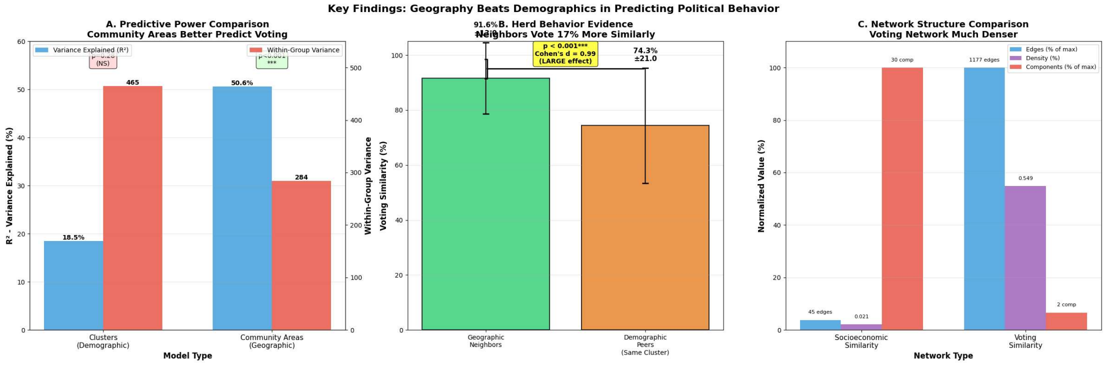
The herd behavior analysis provided strong evidence of spatial influence. Geographic neighbors showed mean voting similarity of 92%, while demographic peers (same cluster, not adjacent) showed only 74% similarity, an 18% point difference. Independent samples t-test yielded t = 9.81, $p < 0.001$, with Cohen's d = 0.99 indicating a very large effect size. People vote more similarly to their geographic neighbors than to demographically similar people elsewhere, suggesting voting is shaped by social context rather than purely individual economic characteristics.

Visualizations:


Your 66 Block Groups (blue) with Precincts from Wards 18, 19, 21 (red outline)


HW4 Socioeconomic Clusters (10 clusters)
Fragmented, 30 components
0
2
4
6
8
10
Voting-Based Communities (2 communities)
Clear geographic divide
Voting Community 1
Voting Community 2


Network Structure Comparison: Why Voting ≠ Demographics
Socioeconomic Network
45 edges, Sparse & Fragmented
Voting Network
1,177 edges, Dense & Connected
Few connections
30 separate components
Many connections
2 unified components



Discussion

What Worked and What Didn't Work

The dominant precinct approach for spatial joining was a major success, capturing 92% of votes despite incompatible boundary systems between precincts and census block groups, this simple method proved robust. The parallel network construction framework also worked exceptionally well, using identical methodologies (cosine similarity, 0.80 threshold) for both networks enabled direct, fair comparison where any differences reflected true structural patterns rather than methodological artifacts.

However, using only a single election limited temporal generalizability. I cannot determine whether these patterns hold across different elections or if the 2023 mayoral race was anomalous. The Johnson-Vallas contest was unusually polarizing and focused on local issues; presidential or state races might show different patterns.

Evaluation of the Project

I evaluated my work through multiple validation approaches. First, I performed data quality checks: verifying that allocated votes totaled 92% of originals, confirming all 66 block groups received data, and visually inspecting outliers. Second, I tested robustness by trying multiple similarity thresholds (0.70-0.90) and confirming findings held across specifications. Third, I used established statistical tests ANOVA, t-tests. Fourth, I employed three different analytical approaches (network structure, ANOVA, herd behavior) all pointed to the same conclusion, strengthening confidence.

What I could have done for even stronger evaluation includes formal cross-validation by holding out test data to validate predictions, temporal validation using 2024 election results when available, spatial autocorrelation testing to formally quantify clustering, and multivariate regression to simultaneously control for demographics while testing spatial effects. Most importantly, qualitative validation through resident interviews would determine whether my statistical "communities" match residents' lived experiences of political identity

What Surprised Me

The magnitude of findings exceeded all expectations. I anticipated clusters would predict voting moderately well, yet they showed no significant relationship (p=0.28). This complete reversal of hypothesis was shocking, I assumed economic self-interest would drive voting, but geography proved three times more powerful. The 26-fold difference in network edges (45 versus 1,177) represented not a modest difference but a fundamental structural transformation. The herd behavior effect size of Cohen's d = 0.99 is enormous by social science standards, where effects above 0.8 are rare. This suggests social influence is not a subtle factor but a dominant force in political behavior.

Most surprising was Beverly's political profile. I expected both Community Areas to be conservative based on traditional descriptions, but Beverly's 58% Vallas vote makes it genuinely competitive, a “swing” community, not a conservative stronghold. This 58-42 split means Beverly contains real ideological diversity, while Mt. Greenwood's 90% Vallas vote indicates near-unanimity. The geographic boundary between them creates a sharp political discontinuity despite similar demographics in some block groups.

Future Work

With more time, I would first add temporal depth by analyzing multiple elections (2020, 2022, 2023, 2024) to test pattern stability across presidential, state, and local races. Second, I would investigate mechanisms: why does geography matter more? This requires mapping actual social networks through social media analysis or surveys, studying local information environments including yard signs and news consumption, conducting historical analysis of when these political boundaries formed, and ethnographic research on political identity formation through resident interviews.

Finally, causal inference remains elusive. Future work could exploit natural experiments: tracking people who move between Community Areas to see if voting changes, studying new developments bridging CA boundaries to determine which area's politics dominate, or analyzing redistricting to test whether political patterns follow new or old boundaries. Fourth, scope expansion would strengthen generalizability, analyzing all 77 Chicago Community Areas, comparing to other cities, and testing rural areas to determine if geography dominates demographics universally or only in urban contexts.

Conclusion

This study demonstrates that geographic boundaries predict voting behavior better than socioeconomic similarity in Mt. Greenwood and Beverly. Socioeconomic clusters do not significantly predict voting ($p=0.28$, $R^2=0.185$), while Community Areas explain 50.6% of variance. Geographic neighbors vote 18 percentage points more similarly than demographic peers ($p<0.001$, Cohen's $d=0.99$), providing strong evidence of spatial influence. These findings challenge rational-choice voting models and support social influence theories, people vote like their neighbors through social norms and shared information environments, not purely based on economic self-interest. Political identity is constructed spatially.
For practice, campaigns should target neighborhoods rather than demographic profiles. Community organizers must recognize political coalitions follow geographic lines while economic coalitions cut across them. In all, Community Area boundaries are arbitrary for socioeconomics but meaningful for politics. Geography creates community identity even when demography doesn't.

References

U.S. Census Bureau. (2023). American Community Survey 5-Year Estimates, 2019-2023. Retrieved from https://data.census.gov

Chicago Board of Elections. (2023). 2023 Municipal Runoff Election Results - April 4, 2023. Retrieved from https://chicagoelections.gov/en/election-results.html

City of Chicago. (2023). Ward Precinct Boundaries - 2023. Chicago Data Portal. Retrieved from https://data.cityofchicago.org/

U.S. Census Bureau. (2023). TIGER/Line Shapefiles: Census Block Groups. Retrieved from https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

Python Software Foundation. (2023). Python 3.12. Retrieved from https://www.python.org/
McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 56-61.

Hagberg, A., Swart, P., & Schult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference, 11-15.

Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., Gerard, & Leblanc, F. (2020). GeoPandas: Python tools for geographic data (Version 0.14.0). Retrieved from https://geopandas.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

AI Assistance
Anthropic. (2024). Claude 3.5 Sonnet (Large language model). Used for: code debugging and error resolution, statistical methodology consultation, visualization design suggestions.

Appendix A: Data Availability

All datasets used in this project are publicly available and were processed for analysis at the census block group level. Socioeconomic data were obtained from the U.S. Census Bureau's American Community Survey (2019–2023 5-Year Estimates). Voting data were obtained from the Chicago Board of Elections 2023 Municipal Runoff Election results, and spatial boundary files were sourced from the City of Chicago Data Portal and U.S. Census TIGER/Line shapefiles.

The cleaned datasets, intermediate files, and data preprocessing scripts used to construct the socioeconomic and voting similarity networks are available in the project's GitHub repository:

https://github.com/Ikemonyeka/CS579_Online-Social-Network-Analysis/tree/main/Data

Appendix B: Code Availability

All analysis code for this project was implemented in Python 3.12 using Jupyter Notebook. The repository includes notebooks for data preprocessing, spatial joins, network construction, statistical testing (ANOVA and t-tests), and visualization. Key libraries used include GeoPandas, NetworkX, scikit-learn, SciPy, Matplotlib, and Seaborn.

The complete, reproducible codebase is available at:

https://github.com/Ikemonyeka/CS579_Online-Social-Network-Analysis/tree/main/Code

The GitHub repository is publicly accessible and contains all materials necessary.