Geospatial Network Analysis of US Megaregions in 40 Years

Pawornwan Thongmak
Walker Department of Mechanical Engineering
The University of Texas at Austin
pt8527@utexas.edu

Yinshuang Xiao
Walker Department of Mechanical Engineering
The University of Texas at Austin
yinshuangxiao@utexas.edu

Phillip Gavino
Walker Department of Mechanical Engineering
The University of Texas at Austin
phillip.gavino@utexas.edu

Ming Zhang
School of Architecture
The University of Texas at Austin
zhangm@austin.utexas.edu

Zhenghui Sha
Walker Department of Mechanical Engineering
The University of Texas at Austin
zsha@austin.utexas.edu

Abstract

This paper proposes a network analysis framework based on geographic information systems (GIS) to study the development of megaregions in support of urban planning and policy-making. The framework includes a new approach to model geo-shaped polygon data of census places as the Place Geo-Adjacency Network (PGAN). In particular, the integration of descriptive network analysis and degree distribution analysis supports the study of spatial connections, geospatial growth, hub effects, and expansion patterns in megaregions. To demonstrate this framework, a case study was conducted on four US megaregions to study their growth and expansion in the last 40 years since 1980. The degree distribution analysis captures the small-world property and quantifies the level of geospatial connectivity influenced by the hub effects. Policymakers can use the model as a decision support for urban planning and policy design to reduce disparities and improve connectivity in megaregion areas.

Keywords: Geographic information systems (GIS), megaregion, complex networks, urban planning, geospatial network analytics, degree distribution.

1. Introduction

Geographic information systems (GIS) – engineered for storing, retrieving, manipulating, analyzing, and mapping geographic data – have experienced significant advancements and widespread adoption in the past five decades [Lü et al., 2019]. The central component of GIS, which involves using a location referencing system to analyze data in relation to other locations, makes it a powerful tool for facilitating urbanization research [Church, 2002]. For example, GIS was integrated with artificial neural networks (ANN) to model urban expansion, taking into account transportation factors and the density of important landscape features [Pijanowski et al., 2009]. In another study, Jat et al. explored the application of remote sensing and GIS in assessing the spatial and temporal phenomena of urbanization and its impact on groundwater [Jat et al., 2009]. Through a case study, the authors validated the utility of these technologies to reveal the correlation between the decline in the water table and quality and urbanization. These examples demonstrate the power of GIS to uncover valuable information to facilitate urban research.

In the latter half of the 20th century, megaregions [1] have emerged as a new scale of geography resulting from the continuing expansion of metropolitan regions. They are often identified based on population density, population growth, employment growth, etc. [Hagler, 2009]. Given the significance of megaregions as a catalyst for economic growth, innovation, and collaboration in driving regional development, megaregion analysis has become a focal point for urban systems design and planning, which have attracted substantial attention from both researchers and policymakers. For example, Guo and Zhang utilized spatial cluster analysis and mixed-effect regression analysis to investigate the factors influencing the expansion of the Texas Triangle megaregion, employing various data sources including land coverage and imperviousness data, transportation data, and socio-demographic data [Guo and Zhang, 2021]. The

[1] A megaregion consists of two or more metropolitan areas and their integrated hinderland. The U.S. Census Bureau has not yet formally adopted megaregion as a census geographic unit. Delineating megaregion boundaries remains an academic exercise. In this study, we follow the working definition of megaregions provided in [Yaro et al., 2022].

URI: https://hdl.handle.net/10125/107064
978-0-9981331-7-1
(CC BY-NC-ND 4.0)

results found that economic growth, population, and highway density are three key factors to motivate the expansion of the Texas Triangle. In another paper, Woodall et al. conducted a comprehensive review study on megaregions, highlighting three prominent approaches. These approaches include an interconnected analysis of megaregions that integrates demographical, economic, environmental, and infrastructural factors. They also discussed a cohesive urban nodes analysis of megaregions based on transportation infrastructure data, as well as a boundary definition approach using satellite data. This review emphasizes the increasing popularity of interconnectedness analysis in megaregion studies [Woodall et al., 2023].

Although existing studies on interdependence in megaregions have provided deep insight into economic, environmental, infrastructural, cultural, and historical relationships [Woodall et al., 2023], geographic interactions between cities have received little attention. However, it is a crucial factor influencing the development and expansion of megaregions. For example, these geographic interplays unveil varying constraints and opportunities for hub cities versus rural cities. Hub cities benefit from shorter distances to their surrounding cities, facilitating socio-economic exchanges, but encounter challenges when it comes to expanding their physical boundaries. In contrast, rural cities face fewer limitations in land expansion, but their isolation from other cities hampers their economic development. Planners and policymakers must consider these geographic distinctions and relationships when framing development policies and strategies.

To gain a deeper understanding of the geographic interdependencies within megaregions, we propose a GIS-based network analysis framework to model shared boundary relationships among census places[2] that are stored as geo-shape polygon data in GIS. We name this network model as the place geo-adjacency network (PGAN) model. This analytical framework offers several advantages. First, it provides a computational representation of the spatial connections between places, enabling us to quantify the role of each place within the megaregion based on network metrics, such as the node degrees (the number of connections of a place). Second, by studying the network evolution, we can trace and track the geospatial growth and expansion of megaregions. This is particularly valuable in identifying spatiotemporal patterns during urbanization. Third, the network model is computationally and visually effective in handling large-scale datasets, providing a solution to the use of GIS data to improve transportation and policy decision-making in urban planning.

This study provides new knowledge on the expansion of megaregions in the last 40 years and contributes to the literature in two aspects. First, we performed degree analyses on the PGAN topology and hub effects. The analysis results show that the proposed PGAN carries a small-world network property where the hub cities play a key role in determining the level of geospatial connectivity of a megaregion. The quantification of hub effects and the characteristics of PGAN topology offers valuable insight for urban planning purposes. For example, the findings can guide the strategic development of transportation infrastructure and the optimization of transportation networks in urban systems, taking into account the effects of the centers. Additionally, the degree analysis identifies areas with lower connectivity and spatial integration, allowing planners to develop intervention strategies to improve connectivity and reduce disparities in less well-connected regions.

Second, a comparative analysis of the evolution of four megaregions in the US over a span of 40 years was conducted. They are the Texas Triangle (Texas), Northeastern Region (Northeast), Northern California (NorCal), and Southern California (SoCal) megaregions. The analysis finds that the PGAN of each megaregion exhibits a distinct topology that remains stable over time. For example, unlike the other three regions with a right-skewed degree distribution, the Northeast shows a more even degree distribution, indicating weaker hub effects in that region. The evaluation of hub effects in the other megaregions shows that Texas has a number of highly connected cities, resulting in shared responsibilities between hubs, while Los Angeles is the only major hub for SoCal. Because SoCal relies heavily on Los Angeles, it experiences a reduced influence of the hub effect and becomes more susceptible to the consequences of losing hub functions. The presence of a solitary hub poses challenges to network connectivity in SoCal, evident in its larger average path length compared to that of Texas. Furthermore, in the time dimension, by tracking the evolution of PGAN over time, the rise of new hub cities can be identified. For instance, the number of places sharing boundaries with Austin, TX has risen from 3 to 19, while Bakersfield, CA has seen an increase from 1 to 23 in four decades.

The remainder of the paper consists of three sections. Section 2 provides an introduction to the research methodology. In Section 3, a case study is demonstrated,

[2]In this study, we follow the U.S. Census Bureau to define a place as a concentration of population that is assigned with a name, locally recognized, and is not a part of any other place. By census definition, cities, incorporated or unincorporated, are specific types of places in certain context [Ratcliffe et al., 2016].

and results and discussions are presented. Finally, Section 4 summarizes the findings and concludes the paper with closing remarks and future work.

2. Research Methodology

Figure 1 displays the proposed GIS-based PGAN analysis framework for megaregions. The detailed explanations are presented in the following sections.

2.1. PGAN modeling and visualization

To create and visualize the PGAN model, we first obtain the geo-shape polygon data of census places, which are the units of analysis for this study, from NHGIS (National Historical GIS, http://www.nhgis.org/). We then select and extract the topological data of the places (including cities) inside megaregions. Finally, we create the network measures of the places based on their spatial adjacency attributes. The network measures of the places are then used as inputs for PGAN generation in Gephi, an open-source network visualization and analysis software. In PGAN, each node represents a unique place. Given that geographic adjacency is reciprocal, an undirected link is established between two nodes when the places share boundaries. The nodes are mapped geographically with their latitude and longitude using a Gephi plugin called *Geo Layout*. The PGAN generation process quantitatively captures geographic connectivity and interdependence between different places within a specific region. That is, a standalone place without shared boundaries is considered a zero-degree node. On the contrary, a high-degree node represents a place with high connectivity to other surrounding places and a higher tendency to behave as a hub city, connecting many smaller cities, towns, and settlements.

2.2. Analysis method

Descriptive network analysis After obtaining the PGAN model, we first perform a descriptive network analysis to gain a better understanding of the network characteristics.

- Network size: The number of nodes and edges enables us to mathematically determine the size of the network and the scale of urban expansion over time.
- Location of hubs: The location of nodes with a high degree demonstrates a significant area with high activities and flow volumes. A hub is an intuitively interesting phenomenon because of its role as a traffic facilitator of activities and information. The loss of well-connected hubs can cause significant disruption to network function [O'Kelly, 2015]. Capturing such a hub effect allows us to better quantify the interdependence of cities in megaregions and observe its trend over time.
- Network metrics: In this study, the average path length and the average clustering coefficient are utilized to examine the structural characteristics of the PGAN. For example, by calculating the average path length in PGAN, we can gain insight into the geospatial interconnections and proximity of places within the megaregion. A reduction in an average path length signifies a higher level of connection between places within the region. In contrast, a longer average path length suggests that places are more distant and separated from each other within the megaregion.

Network degree distribution analysis Degree distribution, along with descriptive network analysis, is of significant importance in understanding network properties and plays a central role in network theory [Barabási and Pósfai, 2016]. It offers valuable statistical insights by determining the probability that a randomly selected node in the network has a specific degree. In urban studies, degree distribution analysis has been effectively employed to examine the linkage properties of urban street networks [Porta et al., 2006] and assess the connectivity of urban bus transport networks [Chen et al., 2007]. These applications have demonstrated the viability of using degree distribution analysis in urban research. This paper employs degree distribution analysis to quantitatively characterize and evaluate PGAN connectivity, investigate the trend of PGAN expansion, and examine the impact of hub nodes.

The degree distribution is plotted through a complementary cumulative distribution (CCDF) on logarithmic-scaled axes [Fornito et al., 2016]. The CCDF conveys the probability that a random variable node X will take a degree that is larger than a random value, x, defined in Equation 1.

$$F_X(x) = P(X \geq x) \tag{1}$$

Degree distributions tend to be highly skewed or asymmetric in most real networks due to the existence of hubs, a characteristic of scale-free types. [Albert and Barabási, 2002, Bettencourt, 2013, Mori et al., 2020]. Unlike a random network where the resulting degree distribution follows the Poisson

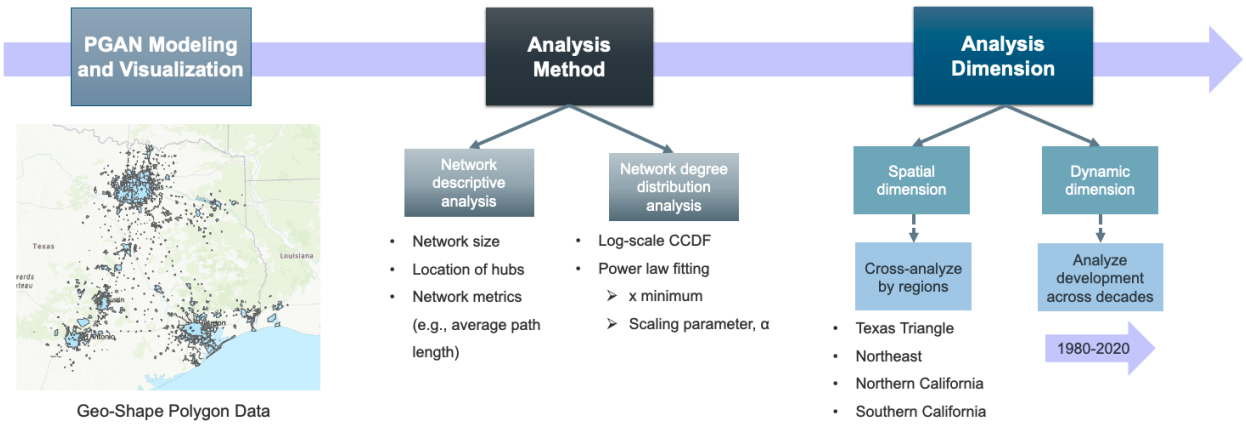


Figure 1: Megaregion Analysis Framework Empowered by GIS and PGAN

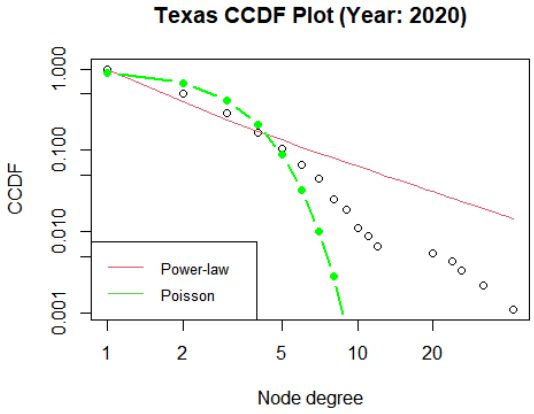


Figure 2: Example of Poisson and power-law distribution fits to node degree CCDF

distribution [Barabási and Pósfai, 2016], indicating that most nodes have comparable degrees and nodes with a large number of links are absent, a scale-free network follows the power-law distribution where it has the majority of the nodes with only a few links while a few highly connected hubs hold most of the links. Figure 2 shows a comparison of fits between Poisson distribution and power-law distribution of Texas node degree distribution in year 2020. In our case, if PGAN performs as a scale-free network, it obeys a discrete power-law with a probability distribution defined in Equation 2, where C is a normalization constant, and α is known as the scaling parameter of the power-law.

$$p(x) = \Pr(X = x) = Cx^{-\alpha} \tag{2}$$

It is rare for an empirical phenomenon to follow the power-law distribution for all values of x, evidently shown in Figure 2 where points along higher node degrees are not properly fitted to the power-law distribution. Hence, a power-law is commonly fitted to the tail of the distribution starting at some value greater than the minimum value of x, called x_{min} [Clauset et al., 2009]. To further investigate the fitting of the heavy-tailed power-law for the PGAN of each region, we first estimate their acceptable x_{min}, the minimum value where the scaling relationship of the power-law begins. Then, we conduct a goodness-of-fit test for the power-law distribution fitting and compare their hub effect with the corresponding α.

The *poweRlaw* package in R [Gillespie, 2015] is implemented to estimate the optimal lower bound that minimizes the distance between CDFs of empirical data. It also reports the fitted model along with the Kolmogorov-Smirnov (KS) statistic for the fit. Then we apply a bootstrapping procedure outlined by [Clauset et al., 2009] to perform a goodness-of-fit test for power-law fitting with parameters including x_{min}, power-law exponent α, and its empirical distance. After running through 1000 power-law distributed synthetic data sets using predetermined parameters, the test calculates the *P-value* that indicates the fraction of time when synthetic distances are larger than empirical distances. The power-law is considered not a good fit to the data and is ruled out when *P-value* ≤ 0.1 [Conklin and Bressler, 2021], and we accept the first x_{min} where the *P-value* for power-law fitting is greater than or equal to 0.1 for future analysis.

2.3. Analysis dimension

After obtaining the outlined analysis, we propose to conduct the analysis in two dimensions.

- Spatial dimension: Investigate similarities and

differences across different megaregions driven by their unique network topology.

- Time dimension: Examine the evolution of PGAN by tracking the trend of change in CCDF curves and assessing how the effect of hub nodes has evolved, whether it has increased, decreased, or remained unchanged.

By cross-analyzing various megaregions spatially and dynamically, we can gain a deeper understanding of megaregions' characteristics, their interdependence, and urban development trend over time.

3. Case Study

The proposed framework is adopted and presented in a case study concerning sample megaregions in the United States.

3.1. Data source

The case study focuses on the degree distribution of 4 out of 13 megaregions in reference to [Yaro et al., 2022], including Texas, Northeast, NorCal, and SoCal, as shown in Figure 3. Texas is selected first due to its unique triangular geometry formed by three major metros including Dallas, Houston, and San Antonio-Austin, which incur considerably high commuting volumes along their edges [Zhang and Lan, 2022]. Apart from Texas, which is located in the south central region of the United States, Northeast and California are selected to represent megaregions on the East and West Coasts, respectively. In accordance with the definition presented in Figure 3, California is further divided into NorCal and SoCal. These four choices of megaregions represent county clusters with high flow volumes in different parts of the country. For PGAN modeling, we focus primarily on GIS datasets including:

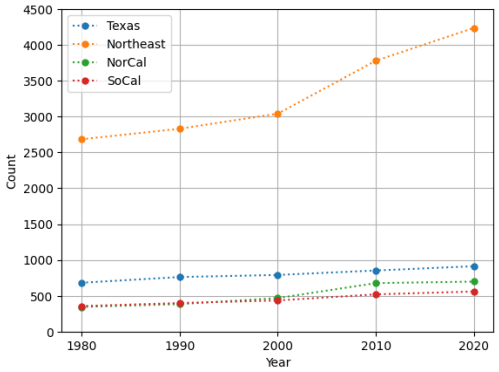
- Polygon shape files of census places in Texas, Northeast, NorCal, and SoCal that lie within each megaregion, spanning four decades across five timestamps from 1980 to 2020. The data source is the National Historical Geographic Information System (IPUMS NHGIS) and the polygon shape files are census-designated places [Manson et al., 2017].
- Topologial data for constructing network measures and latitude and longitude data for network geographic mapping.

Figure 3: Location of the four megaregions studied

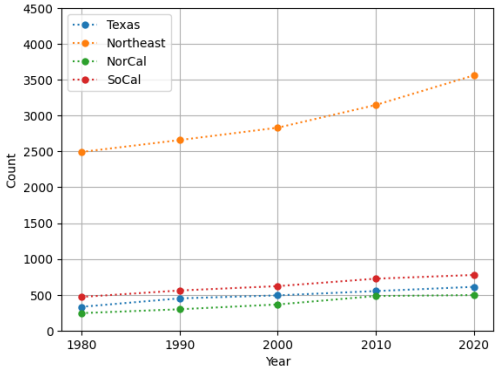
3.2. PGAN visualization and descriptive analysis

PGAN size and visualization The number of nodes and edges for each megaregion at five timestamps is summarized in Figure 4. Figure 5 represents the established PGAN visualizations for each region in 2020. Because the Northeast is a large region spanning multiple states, the network size is reflected through its significantly higher number of nodes and edges. Texas, NorCal, and SoCal all have comparable network sizes throughout the years, with Texas having the highest number of nodes between the three, but the edge size in the medium between California megaregions. Overall, we are able to capture the expansion in network size across the regions within four decades.

Location of hubs Geographic location is an essential underlying factor that affects the characteristics of each megaregion and the development of the hubs in it. The locations of the highly connected places in 2020 are marked in Figure 5. Texas has a number of major places with comparable degrees, namely Houston (41), Fort Worth (31), Dallas (25), and San Antonio (23). These four metros have remained the top most connected places since 1980. Meanwhile, by tracking the evolution of PGAN over time, we observed the rise of new places. For example, the number of places that share boundaries with Austin increases from 3 to 19 in four decades. In terms of SoCal, it comparatively has one major hub which has always been Los Angeles (42), followed by Bakersfield (23) and San Diego (17) in 2020. The other places apart from these well-connected metros are distributed with lower degrees for both megaregions. More specifically for SoCal, Los Angeles' degree has more than double that of the second-ranking place. Because of such dispersion in degree distribution, we notice distinct tail shapes for Texas and SoCal. SoCal, however, is more susceptible to network disruption as Los Angeles is solely bearing



(a) Number of nodes



(b) Number of edges

Figure 4: Network size across years

the major role of traffic facilitator compared to the distribution of activities between different hubs in Texas.

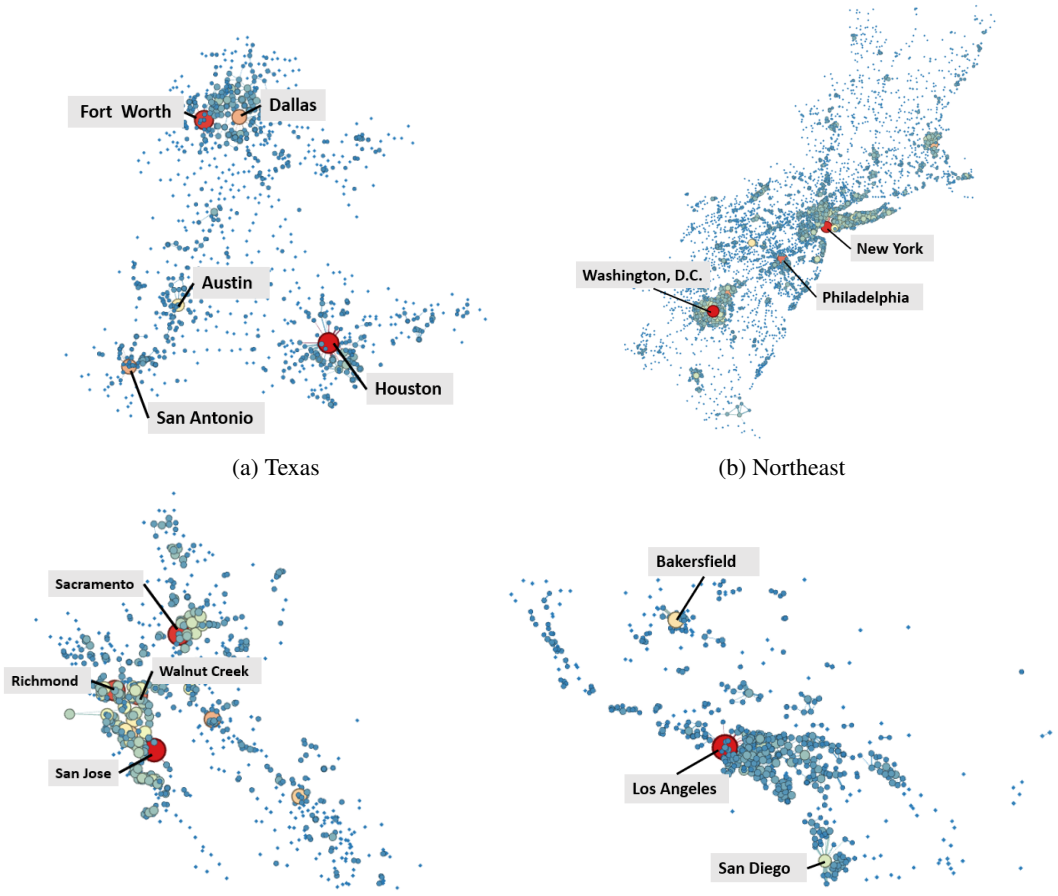
The Northeast and NorCal share similar megaregion shapes which allow them to expand mostly vertically in the north and south directions. The main hubs of the two regions, New York (19) and San Jose (14), are bordered by a body of water. Therefore, the city expansion is more restricted and naturally dispersed to other places, namely Washington, D.C. (20) and Philadelphia (16) for the Northeast, and Sacramento (13), Richmond (12), and Walnut Creek (12) for NorCal. On the contrary, Texas has an advantage compared to the other regions because of its substantial land area. Aside from Houston, the other places are less confined, allowing them to expand in any direction. Therefore, Texas constitutes more high-degree nodes than other megaregions.

Average path length In addition to network size and geographic locations, average path lengths and average clustering coefficients, shown in Figure 6(a) and 6(b), contribute to our deeper understanding of the expansion pattern and geospatial interconnections of megaregions. The average path length ($\langle d \rangle$) quantifies the average distance between all pairs of nodes in the network [Barabási and Pósfai, 2016]. There is an obvious trend over the years where the Northeast has the highest $\langle d \rangle$, followed by NorCal, SoCal, and Texas, showing that the overall geospatial connectivity of Texas is much higher than other megaregions. Specifically, Texas and the Northeast have an overall increasing trend, indicating that as network size increases, the path length to get from one place to another becomes greater, on average. Meanwhile, NorCal and SoCal have an almost identical trend that starts out with a rise in $\langle d \rangle$, then the values stabilize from 2000 to 2020 at approximately 8.4 and 6.5, respectively. This means that, regardless of the megaregion expansion, the average path length between each pair of places in NorCal and SoCal remains steady in the latter decades.

Global clustering coefficient The level of place connectivity is explored further with the clustering coefficient (CC). In this study, we are specifically interested in the global CC which represents the degree of connectivity for the whole network, instead of at any local node. The CC captures the probability that any two randomly chosen neighbors of a node with a degree of at least two are linked together [Latapy, 2008]. In support of an increase in average path lengths as time progresses, the CCs of Texas PGAN and Northeast PGAN have an overall decreasing trend from 0.553 to 0.507 for Texas and 0.484 to 0.431 for the Northeast. NorCal has the most obvious increase trend in average CCs, rising from 0.460 in 1980 to 0.503 in 2020. SoCal, on the other hand, has a more fluctuating trend but ends up with the highest value of 0.546 in 2020. The expansion pattern of the places could influence such a fluctuation. For example, an expansion surrounding a well-established area explains the higher probability that neighbors of a node are connected. In summary, the Northeast and Texas networks have been shown to experience an increase in average path length and a decrease in average CC over time. NorCal and SoCal networks can maintain greater resilience in geospatial interconnections. We continue our investigation with the network degree distribution analysis in the following section to gain more insight regarding network connectivity and expansion.

3.3. PGAN degree distribution analysis

Log-scale CCDF analysis We conduct the degree distribution analysis of log-log CCDFs of the PGAN and fit a power-law distribution to capture the scale-free property. The resulting CCDF plots are shown in Figure 7 and are accompanied by that of the year 2020 with



(a) Texas (b) Northeast

(c) NorCal (d) SoCal

Figure 5: 2020 PGAN for each megaregion

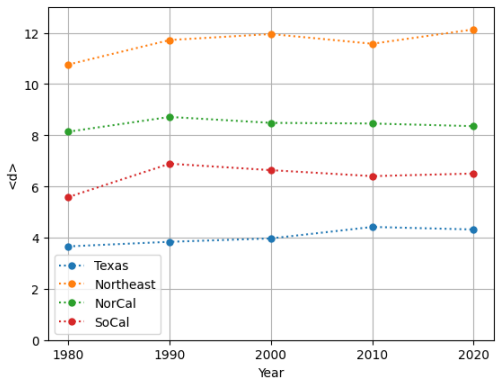
its fitted power-law line. It is interesting to observe that each megaregion has uniquely shaped degree distributions, while commonly sharing a relatively stable trend along the beginning of the distribution. The curves follow a similar downward slope across five timestamps for each megaregion when node degrees are less than or equal to 10. More noticeable deviations occur along the tail of the distributions, where the degree of the node ranges from 10 to 45. This supports the characteristic of the long-tailed distribution of PGANs. The degree distribution is evenly distributed for the Northeast and NorCal, that is, number of nodes with low and high degrees are comparable. Meanwhile, Texas and SoCal have a more distinct tail shape, displaying a more apparent pivot point that indicates an abrupt shift to a node or various nodes with larger degrees (i.e., the hubs), where the rest of the node distributions are concentrated on the lower node degrees.

Fitting the power-law distribution We identify the acceptable x_{min} values from the bootstrapping procedure introduced in Section 2.2 and use this method to fit the power-law distribution across four decades for all regions. The resulting values for each megaregion are as follows:

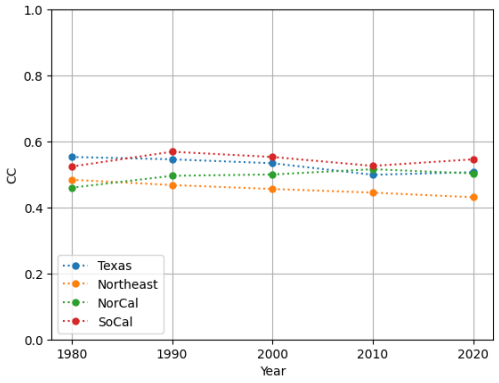
$$\mathrm{x}_{min,Texas} = 3,\ \mathrm{x}_{min,Northeast} = 10,$$
$$\mathrm{x}_{min,NorCal} = 5,\ \mathrm{x}_{min,SoCal} = 6$$

It is interesting that Texas, despite its larger network size, has a x_{min} at 3, lower than that of both NorCal and SoCal. From this finding coupled with its CCDF shape shown in Figure 7, we claim that Texas has a higher proportion of nodes with higher degrees, while NorCal and SoCal's nodes are concentrated more heavily towards the lower degree distribution. The power-law fitting occurs late at x_{min} of 10 for the Northeast, considering that the most connected node has the maximum degree of 20 from 2000 to 2020. Therefore, we observe that Northeast has a tendency to behave as a random network, where most of the nodes have comparable degrees.

The corresponding x_{min} values are used to fit the



(a) Average Path Length



(b) Average Clustering Coefficient

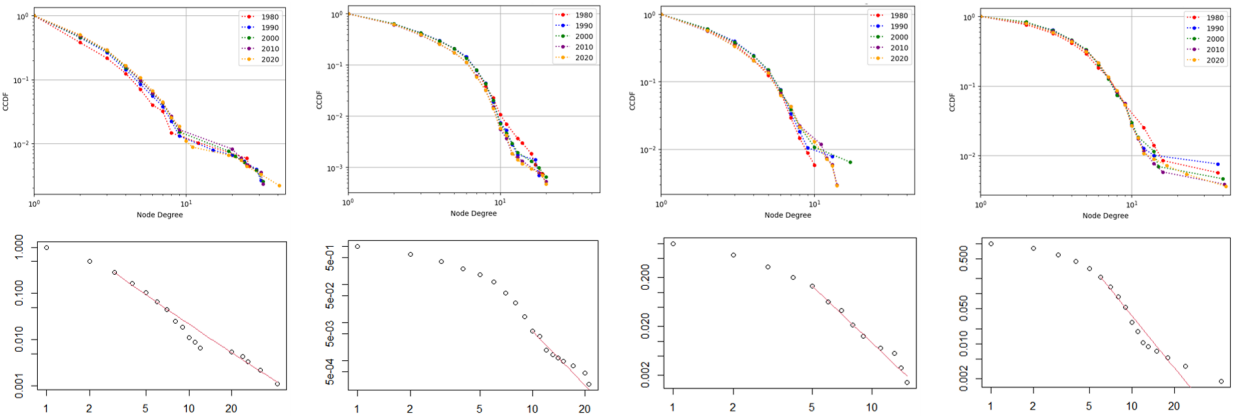
Figure 6: Network Metrics Investigated

power-law distribution to the degree distribution of each region and decade, and the power-law exponents, α, are extracted. The values are shown in Figure 8, ranging from 2.965 to 5.822. To categorize our analysis of α values in a proper regime, the average path length of four megaregions across five timestamps is plotted against their number of nodes in log-scale, presented in Figure 9. It is observed that the lnN curve, representing the small-world regime [Barabási and Pósfai, 2016], is the closest to our megaregion data points. A small-world network is characterized by a small average shortest path length between nodes and a high CC when compared to a random graph. For all megaregions, we notice the high average CC values ranging between 0.4 and 0.6. Therefore, our networks are expected to carry a small-world property, where the network hubs effectively shrink the distances between nodes as they become more pronounced. We utilize the trend of α to quantify the extent of the scale-free property under a small-world regime and the effect of hubs in the urban network context.

Discussion As α which is the slope of the fitted power-law line in Figure 7 decreases, there is a higher probability of observing a hub city in the region, hence a stronger scale-free property or hub effect. Figure 8 shows a stable trend over time for each megaregion, where the Northeast has the highest α, followed by NorCal, SoCal, and Texas, respectively. Dynamically, we observe that each megaregion has its scaling parameter at a considerably stable rate from 1980 to 2020. It can be interpreted that as the region expands in size, as reflected by the increase in the number of nodes and edges, the strength of the scale-free property is preserved. Therefore, the network topology and characteristics of the megaregions of interest are maintained despite their geographical differences over time.

The Northeast has the weakest effect of the scale-free property, where the loss of a hub is most destructive to the network cohesion. A decrease in α from 5.673 in 1980 to 5.583 in 2020 has a negligible effect on the shrinkage of the network's average distance. Following the Northeast, NorCal has the second highest α values over decades. As we reference back to Figure 7, we notice that both megaregions have evenly distributed curves contributed by nodes with comparable degrees, but NorCal's network is proven to be more resilient to the loss of hub than the Northeast. Despite their similar geographic characteristics, NorCal's hubs are able to improve interconnections and increase proximity between places as the network size expands, as outlined in Section 3.2. This supports the competitiveness of NorCal hubs in facilitating activities and information effectively across the megaregion. The lowest values of Texas's α support the claim of a higher proportion of high-degree nodes. The distribution tail at node degree between 30 and 45 of Texas indicates multiple highly connected places, whereas SoCal only has one sole candidate, Los Angeles. This explains the higher slope, weaker scale-free property, and higher average path length of SoCal in comparison to Texas.

For the four megaregions of interest, including the Northeast, Texas, NorCal, and SoCal, we discover that the size of the network plays an important role in urban expansion, influencing the level of network connectivity. We find that the linear, vertical stretching of both the Northeast and NorCal regions impacts their network connectivity. Specifically, the Northeast region is more vulnerable to network disconnection in the event of a hub loss. Compared to SoCal, Texas has a higher probability of encountering a hub, but the expected proximity between places does not materialize effectively as reflected by an increase in the average path length and a decrease in the average CC, conflicting with the small-world property. To optimize these



(a) Texas (b) Northeast (c) NorCal (d) SoCal

Figure 7: The 4-decade CCDFs (first row) and power-law fitted line (second row) for each region

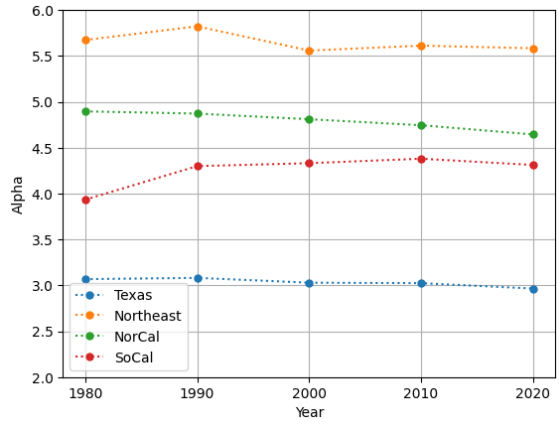


Figure 8: α trend across decades

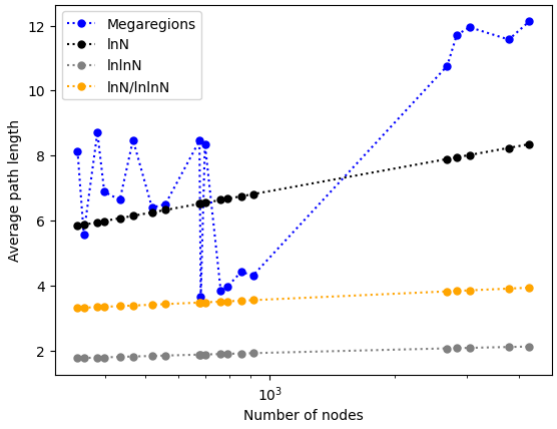


Figure 9: The power-law regimes based on the increase of average path length as network size increases

large megaregions, we recommend adopting a strategic expansion pattern that leverages well-established areas to improve average connectivity, reduce average path length, and improve the effectiveness of network hubs. In urban expansion planning efforts, it is crucial to consider geographic characteristics that align with the strengths of each region, ultimately boosting the effectiveness of the hub functions and strengthening network resilience.

4. Conclusion

This paper presents a GIS-based network analysis framework for megaregion research. This framework models geo-shaped polygon data of census places as the PGAN. By integrating descriptive network analysis and degree distribution analysis, we aim to understand spatial connections, geospatial expansion, and hub effects during the urbanization process. The proposed framework is demonstrated using a case study that includes representative megaregions in the United States. In the case study, geographic connectivity and interdependence of places are captured as networks, generated by PGAN modeling. The descriptive network analysis and degree distribution analysis quantitatively assess geospatial interconnections, geographic characteristics, and proximity between places. This study offers a new approach to megaregion analysis and generates new knowledge on megaregion expansion, which is beneficial for urban planning and policy making. Based on the distinct characteristics and dynamics of the Northeast, Texas, NorCal, and SoCal, several policy recommendations can be proposed. First, to strengthen connectivity, investments in transportation infrastructure such as road networks and railways

should be prioritized in the Northeast and Texas regions to improve connectivity and reduce average distance between places. Second, to enhance resilience, maintaining and strengthening existing connectivity in NorCal and SoCal is crucial. Third, in SoCal, where there is heavy reliance on Los Angeles as a major hub, policies should promote the development of additional hubs to distribute functions and reduce vulnerability. Incentives for businesses and industries to establish operations in other cities can contribute to this goal. Lastly, as new hub cities such as Austin, TX and Bakersfield, CA emerge, monitoring and support should be provided through investments in infrastructure, economic development, and connectivity to facilitate their integration into the network. These policy recommendations aim to foster connectivity, resilience, and balanced growth in the urban networks of the respective megaregions. In future work, we aim to apply the framework to all the megaregions within the United States. Our goal is not only to uncover trends and gain a deeper geospatial network understanding, but also to derive insightful policy recommendations that can inform and benefit practice.

References

[Albert and Barabási, 2002] Albert, A. and Barabási, A. (2002). *Statistical Mechanics of Complex Networks*. Reviews of Modern Physics.

[Barabási and Pósfai, 2016] Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge University Press, Cambridge.

[Bettencourt, 2013] Bettencourt, L. (2013). The origins of scaling in cities. *Science (American Association for the Advancement of Science)*, (340(6139)):1438–1441.

[Chen et al., 2007] Chen, Y.-Z., Li, N., and He, D.-R. (2007). A study on some urban bus transport networks. *Physica A: Statistical Mechanics and its Applications*, 376:747–754.

[Church, 2002] Church, R. L. (2002). Geographical information systems and location science. *Computers & Operations Research*, 29(6):541–562.

[Clauset et al., 2009] Clauset, A., Shalizi, C., and Newman, M. (2009). Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics*, (51(4)):661–703.

[Conklin and Bressler, 2021] Conklin, B. and Bressler, S. (2021). Organization of areal connectivity in the monkey frontoparietal network. *NeuroImage*, (118414):1–9.

[Fornito et al., 2016] Fornito, A., Zalesky, A., and Bullmore, E. (2016). *Fundamentals of Brain Network Analysis*. Elsevier.

[Gillespie, 2015] Gillespie, C. (2015). Fitting heavy tailed distributions: The powerlaw package. *Journal of Statistical Software*, (64(2)):1–16.

[Guo and Zhang, 2021] Guo, J. and Zhang, M. (2021). Exploring the patterns and drivers of urban expansion in the texas triangle megaregion. *Land*, 10(11):1244.

[Hagler, 2009] Hagler, Y. (2009). Defining us megaregions. *America*, 2050:1–8.

[Jat et al., 2009] Jat, M. K., Khare, D., and Garg, P. (2009). Urbanization and its impact on groundwater: a remote sensing and gis-based assessment approach. *The Environmentalist*, 29:17–32.

[Latapy, 2008] Latapy, M. (2008). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407:458–473.

[Lü et al., 2019] Lü, G., Batty, M., Strobl, J., Lin, H., Zhu, A., and Chen, M. (2019). Reflections and speculations on the progress in geographic information systems (gis): a geographic perspective. *International journal of geographical information science*, 33(2):346–367.

[Manson et al., 2017] Manson, S., Schroeder, J., Riper, D., and Ruggles, S. (2017). Ipums national historical geographic information system: Version 12.0 [database].

[Mori et al., 2020] Mori, T., Smith, T., and Hsu, W. (2020). Common power laws for cities and spatial fractal structures. *Economic Sciences*, (117(12)):6469–6475.

[O'Kelly, 2015] O'Kelly, M. E. (2015). Network hub structure and resilience. *Networks and Spatial Economics*, 15:235–251.

[Pijanowski et al., 2009] Pijanowski, B., TAYEBI, A., Delavar, M., and Yazdanpanah, M. (2009). Urban expansion simulation using geospatial information system and artificial neural networks.

[Porta et al., 2006] Porta, S., Crucitti, P., and Latora, V. (2006). The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866.

[Ratcliffe et al., 2016] Ratcliffe, M., Burd, C., Holder, K., and Fields, A. (2016). Defining rural at the us census bureau. *American community survey and geography brief*, 1(8):1–8.

[Woodall et al., 2023] Woodall, B., Borowitz, M., Watkins, K., Costa, M., Howard, A., Kemerait, P., Lee, M., Rolls, G., Takubo, Y., Titshaw, R., et al. (2023). The megaregion–forms, functions, and potential? a literature review and proposal for advancing research. *International Journal of Urban Sciences*, pages 1–23.

[Yaro et al., 2022] Yaro, R., Zhang, M., and Steiner, F. (2022). *Megaregions and America's Future*. Lincoln Institute of Land Policy.

[Zhang and Lan, 2022] Zhang, M. and Lan, B. (2022). Detect megaregional communities using network science analytics. *Urban Science*, (6(12)):1–14.