A Community-Based Approach for Hub Placements

by

Alissa Chavalithumrong

B.S. Mechanical Engineering, University of Nevada Reno, 2022

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE IN AERONAUTICS AND ASTRONAUTICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: Alissa Chavalithumrong
Department of Aeronautics and Astronautics
May 17, 2024

Certified by: Hamsa Balakrishnan
William E. Leonhard (1940) Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by: Jonathan P. How
R. C. Maclaurin Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

A Community-Based Approach for Hub Placements

by

Alissa Chavalithumrong

Submitted to the Department of Aeronautics and Astronautics
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE IN AERONAUTICS AND ASTRONAUTICS

ABSTRACT

Advanced Air Mobility (AAM) is a rapidly emerging sector in the aerospace industry that seeks to revolutionize transportation by integrating highly automated aircraft into the airspace. As AAM technology matures, establishing a network framework and strategic hub locations becomes crucial for transitioning from theoretical models to practical applications in transportation systems. This thesis investigates community-based strategies for hub placement within the AAM infrastructure. More specifically, it utilizes network segmentation to decompose a network into communities to simplify the hub selection process into more manageable sub-problems. Our first contribution is the development of a specialized community detection methodology called Directed Flow Communities (DFC), which is designed to accommodate the attributes of transportation networks. Next, we conduct a case study using the Freight Analysis Framework (FAF) dataset as a proxy for AAM demand. The empirical investigation focuses on three key sectors: pharmaceuticals, electronics, and comprehensive freight flows, each presenting distinct challenges and insights into the network's structure. The findings show the effectiveness of the community detection-based methods in unveiling cost-efficient hub locations.

Thesis supervisor: Hamsa Balakrishnan
Title: William E. Leonhard (1940) Professor of Aeronautics and Astronautics

Acknowledgments

First, I want to express my deepest gratitude to my advisor, Hamsa Balakrishnan, for her mentorship and guidance throughout my Master's program. Her support has been invaluable. I also extend my thanks to my fellow labmates in DiNaMo—Siddarth Nayak, Victor Qin, Geoffrey Ding, Chris Chin, Akila Saravanan, Shashank Deshpande, Allan Shtofenmakher, and Kevin Zimmer—for their camaraderie and collaboration during this thesis. Special thanks to Jasmine Jerry Aloor and Sydney Dolan for being the best colleagues, cubicle mates, advice-givers, and friends I could have asked for. Jasmine, I am so happy we were part of the same cohort and got to be neighbors. Sydney, thank you for tolerating my shenanigans and always lending a hand.

I am deeply thankful for my family: Mom, Dad, Deanna, Max, TJ, and Tina. I would not be where I am today without you all. I would also like to thank Michael and Jenny. Michael, the research articles and substack posts you sent me kept me engaged and inspired in my work. And Jenny, thank you for checking in on me throughout and continuously welcoming me into your home. Thank you to the friends I made in Cambridge—Yana, Amy, Ayaka, Fausat, Hamid, Jelle, Louis, and Blake—for filling my life with joy. I would also like to thank my friends from back home—Eunnise, Ruthie, Julia, Ginny-Mei, and Jojo—who make my life whole.

Finally, I thank my partner, Jasper. Your support, humor, unwavering confidence, and the effervescent way you exist in this world made my time writing this thesis easier.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Advanced Air Mobility (AAM) is a rapidly emerging aerospace industry sector that aims to integrate highly automated aircraft safely and efficiently into the national airspace. AAM is not a single technology but rather a collection of aircraft types, including electric vertical takeoff and landing systems, drones, and other Unmanned Aerial Vehicles (UAVs). AAM has the potential to alter transportation by introducing new strategies for rapid medical shipments, last-mile delivery services, the deployment of urban air taxis, and improved rural air connectivity. As we approach the next decade, the surge in AAM is expected to result in over 100,000 crewless flights for package delivery alone in the San Francisco Bay Area at maturity [1]. Given this projected growth, prioritizing safety, efficiency, and seamless integration with existing infrastructure becomes imperative.

As AAM technologies develop, their adoption in public sectors will mark a shift in how we can access transportation. Notably, the AAM industry's participation in global events, such as the demonstration of the Volocopter at the Olympic Games [2], highlights its growing significance and potential for widespread adoption. The proliferation of companies specializing in various segments of AAM, including intra-city travel (e.g., Joby Aviation, EHang, Volocopter), inter-city connections (e.g., Heart Aerospace), and cargo delivery services (e.g., UPS Flight Forward, Amazon), emphasizes the diverse applications and scalability of AAM. Importantly, the utility of AAM extends beyond passenger transport to critical public services

such as fire fighting, search-and-rescue operations, power line inspections, and the delivery of vaccines and medical equipment. A study conducted by Dulia et al. compares existing ground transportation with potential AAM operations in Ohio, considering factors like travel time savings, safety cost reductions, cargo delivery efficiencies, and environmental impacts, with findings suggesting that AAM could offer substantial benefits far outweighing its initial costs [3].

This emerging technology raises questions about the planning and development of large-scale AAM facilities. Central to this discourse is the consideration of how network science and existing knowledge of transportation systems can inform the strategic design and policy formulation for AAM infrastructure. AAM presents a rare opportunity to redesign transportation networks fundamentally. As a society, we have the choice to avoid replicating inequalities that marred historical transportation systems. The integration of network theory techniques, such as hub location problems and community detection, presents an intriguing approach to devising AAM networks that are efficient, equitable, and resilient.

1.2 Background

As AAM technology matures, the development of a robust network framework and hub placements, become crucial for transitioning from theoretical models to practical applications in transportation systems. Hubs serve as a central point in transportation networks where traffic (whether goods, information, or passengers) is consolidated, distributed, or switched. Hubs are foundational to the hub-and-spoke model, a system designed to optimize routes and connections, streamlining the flow within networks [4]. By centralizing traffic, hubs reduce operational costs associated with transportation and are crucial for achieving economies of scale in many-to-many distribution systems. This consolidation allows for the more efficient use of resources, as vehicles can operate at fuller capacities, and direct routes can be minimized.

The hub location problem (HLP) involves locating hub facilities and allocating demand nodes to hubs to route the traffic between origin–destination pairs [5]. HLPs generally classified as NP-hard due to the computational complexity involved in finding an optimal

solution, especially as the size of the network grows [6]. In this work, we consider the uncapacitated multiple allocation p-hub problem, which selects a predetermined number (p) of optimal hub locations from potential nodes in a network with no capacity limits. We investigate the use of community detection to divide the network into distinct clusters, to identify inherent groupings for hub positioning.

There are several factors to consider when dealing with hub location problems in transportation networks. HLPs experience combinatorial explosion, as even a network with only a moderate number of nodes can result in a large number of possible subsets, making exhaustive search methods impractical. Additionally, the interconnected nature of the network makes it challenging to establish hubs as the decision to establish a hub in one location can affect the optimal placement of other hubs. Therefore, decision-making must account for the network's connectivity and its impact on efficiency and costs. HLPs involve multiple objectives and constraints, such as minimizing transportation costs, maximizing service levels, and minimizing total travel time. These competing objectives and constraints further add to the complexity of solving the problem.

In order to address these challenges, we consider decomposing a large transportation network in to clusters. By segmenting the network into distinct communities with dense internal connections, we simplify the task of identifying optimal hub locations to smaller sub-problems. This tacticly acknowledges the interactions between nodes, offering a more context-sensitive hub selection process and aids in optimizing hub allocation by aligning with the network's inherent clustering.

1.3 Related Work

Hub location problems, and its ties to large industries, have been an active field of research for several decades. The first formal studies and formulations of HLPs can be traced back to the late 1970s and early 1980s, within the context of the airline industry's need to optimize their networks through hub-and-spoke systems [7]. Traditionally, hub location problems are formulated as integer linear programming (ILP) problems with binary decision variables. O'Kelly introduced the first formulation of the HLP in literature for single and two-hub sys-

tems as an optimization problem [8]. HLPs are a specialized extension of classical Facility Location Problems (FLPs). Analogous to well-known FLPs such as the p-median, uncapacitated facility location, p-center, and covering problems. HLPs include variants like the p-hub median, uncapacitated hub location, p-hub center, and hub covering problems. These HLP variants cater differ based on assumptions regarding network topology, the allocation of origin/destination nodes to hubs, and hub capacity constraints, among other factors. Due to the complex nature and inherent uncertainties in hub location problems, these issues are often addressed through stochastic analysis or heuristic techniques [9].

Transportation and shipping networks on both global and national levels are known to exhibit community structure [10], [11]. Initial clustering or community detection has become increasingly popular as a foundational step for optimizing hub placement in networks. Previous research by Zheng et al. [12] involved using the Girvan-Newman community detection method [13], [14] to identify communities within networks as a preliminary step towards optimizing hub placement. Wang et al. [15] uses K-means as a method for spatial analysis before running a heuristic based method for selecting hub placements within groups. O'Kelly considers the placement of hubs within a network by first clustering planar nodes into groupings that minimizes the sum of square distances between nodes, and then assigns the centroid of the groups as a hub [16]. Peker et al. identifies optimal hub locations by considering node centrality and demand through a spatial viewpoint [17].

Past works have independently explored the utility of specific community detection methods for preliminary community identification, but a systematic comparison across a spectrum of community detection methods remains largely unexplored in literature.

1.4 Thesis Contributions

Our contributions in this work are as follows:

1. **Community Detection in Directed, Weighted, Attributed Networks:** The first contribution is the development of a community detection methodology tailored to transportation networks.

2. **Comparative Analysis of Community Detection Based Hub Selection:** Building on the identified communities, we conduct a comparative analysis for hub placements on the Freight Analysis Framework. We explore two approaches within these communities: *centrality-based*, focusing on the strategic importance of nodes, and *integer programming-based*, emphasizing mathematical optimization.

1.5 Outline

The remainder of this document is structured as follows: Chapter 2 details the methodology behind Directed Flow Communities, a community detection approach tailored for transportation networks. Chapter 3 presents a case study for hub selection, utilizing the Freight Analysis Framework as a stand-in for Advanced Air Mobility (AAM) demand to compare various community decomposition strategies. Finally, Chapter 4 summarizes the key findings and discusses potential avenues for future research stemming from this thesis. Code, parameters, and additional supporting figures can be found in the Appendices.

Chapter 2

Community Detection in Network Analysis

2.1 Background

Community detection, also known as graph partitioning or clustering, is an essential component of network analysis. It entails identifying subgroups of nodes within a network that are more strongly interconnected than they are connected to the rest of the network [18]. This concept plays an important role in comprehending and interpreting the structural properties and dynamics of complex networks [19], [20].

Directed, weighted graphs are especially significant in representing systems where not only the connections matter but also the direction and magnitude of interactions. This includes a wide array of systems, from biological networks [21] to social media interactions [22] and transportation networks [23].

In the domain of transportation networks, the development and application of community detection methods tailored to the unique intricacies of these systems remain largely unexplored. This gap is particularly evident when considering the three pivotal issues that are characteristic of transportation networks: geographic location, connectivity, and dynamic traffic patterns.

- **Geographic Location:** Transportation networks are inherently tied to physical ge-

ography, where the spatial distribution of nodes and the distances between them play critical roles in network functionality and efficiency. Traditional community detection algorithms often overlook the geographical constraints and the physical realities of node placements, leading to suboptimal grouping that may not be feasible or efficient in real-world applications. A transportation-specific community detection method must incorporate geographic considerations, ensuring that identified communities reflect practical constraints such as physical distances, topographical barriers, and regional boundaries.

- **Connectivity:** Unlike many other types of networks, the connectivity within transportation networks is not merely defined by the existence of links between nodes but also by the capacity, frequency, and reliability of these connections. Different modes of transportation have varying levels of service, which influence the strength and significance of connections within the network.

- **Dynamic Traffic Patterns:** Traffic flow within transportation networks is highly dynamic, influenced by factors such as time of day, season, and changing economic activities. This dynamism introduces complexity in understanding the true structure of the network, as the significance of nodes and connections can fluctuate significantly.

2.1.1 Challenges in Directed, Weighted Networks

In directed graphs, the challenge of community detection extends beyond merely identifying densely connected groups. It also involves understanding the directional flow of information and discerning how this flow affects community formation and boundaries [20]. This understanding is critical for accurately interpreting the directional relationships and dependencies within the network.

Many community detection algorithms optimize for specific metrics that define communities, which may not be universally applicable across all network types. This limitation underscores the utility for more adaptable and generalized approaches that can cater to the diverse characteristics of complex networks.

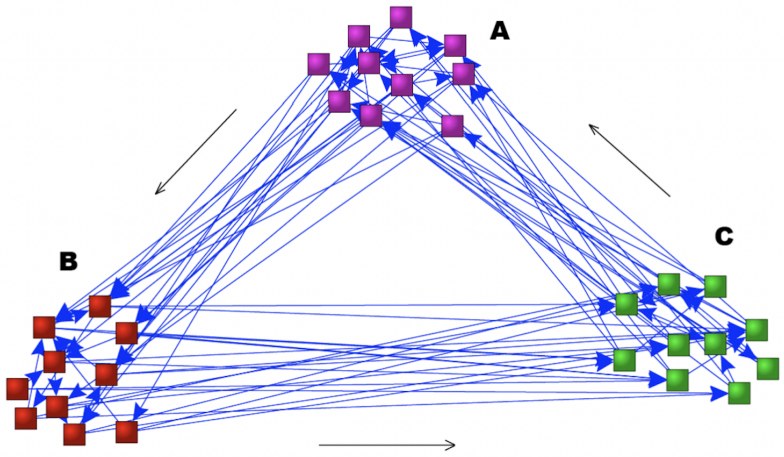


Figure 2.1: Example of a directed graph where there is a cyclical flow among three distinct node groups. Traditional community detection approaches fail to detect these group. For example, when dividing the graph into these three groups, the directed modularity measure is zero. In contrast, when the modularity value is maximized, these flow-based communities are lost. Figure from Lancichinetti and Fortunato [24].

2.1.2 Existing Community Detection Methods

Community detection in undirected, unweighted graphs has historically formed the foundation of network analysis, with quantifiers such as modularity [25] used to identify densely connected subgroups. Modularity juxtaposes the density of intracluster links with the expected density in an equivalent random graph, providing a measure of community strength. Given the NP-hard nature of optimizing modularity [26], which poses computational challenges in large-scale networks, approximation algorithms have become pivotal. Notable among these are the Louvain [27] and Leiden [28] algorithms, which offer solutions by approximating the optimal modularity score to identify meaningful community partitions efficiently. The concept of modularity has been expanded to include weighted and directed graphs [29], and has been incorporated into the Directed Louvain algorithm [30].

Spectral clustering operates on the principle of transforming the problem of community detection into a graph partitioning problem in a lower-dimensional space, achieved by the spectral decomposition of the graph's Laplacian matrix [31], [32]. This process presents issues when encountering directed graphs, which are characterized by asymmetrical matrices

[20]. Other traditional clustering methods include Girvan-Newman [13], [14] and hierarchical clustering [33].

Several methods use a dynamic approach to community detection, leveraging the concept of a random walk. In these methods, a community is a group of nodes within which a "random walker," an entity moving from node to node following edges, is more likely to remain, indicative of dense intra-community connections. Walktrap [34] does this by calculating distances between nodes and merging communities with the smallest distance. Approaches for directed graphs, such as Infomap [35] and Relaxmap [36] perform similarly, but instead optimize for an information theoretic function known as the Map Equation.

In recent years, higher-order clustering and deep learning techniques have been explored for community detection. Higher-order clustering goes beyond pairwise node connections, focusing instead on the overarching patterns of connectivity, known as "motifs," to delineate communities through motif significance [37]–[40]. Concurrently, integrating deep learning and graph neural networks (GNNs) for community detection has also shown promise [41], [42].

2.2 Graph Notation

A directed graph is represented as $G = (V, E, w)$, where: V is the set of nodes, with $|V|$ denoting the total number of nodes, E is the set of directed edges between nodes, with $|E|$ representing the total number of edges. $w : E \to R^+$ is a function assigning a weight to each edge, quantifying the strength or capacity of the interaction from the start node u to the end node v in each edge $(u, v) \in E$.

2.3 Proposed Community Detection Method: Flow Based Clustering

This model, Directed Flow Communities (DFC) treats the flow of mass (or information) through a network in a manner reminiscent of dye permeating fabric. In this analogy, the network nodes are likened to vats filled with fabric, each capable of absorbing dye, while the

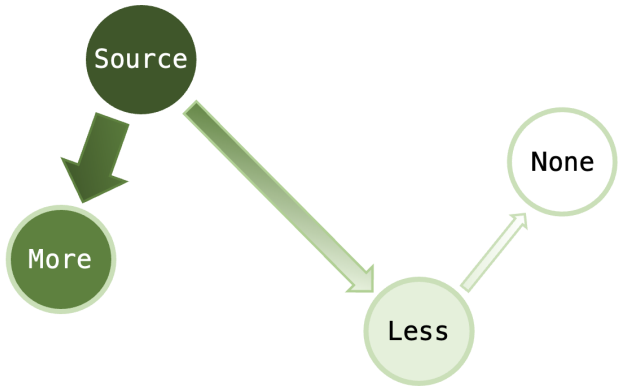


Figure 2.2: This schematic diagram represents the distribution of flow from a source node within a directed network. The "Source" node is the origin of the mass that is being distributed through the system. The "More" node, positioned along a shorter and wider edge, indicates substantial accumulation of flow mass, attributed to both its proximity to the source and the greater edge weight. In contrast, the "Less" node, connected by a longer and narrower edge and shown with a lighter color, receives less mass, due to the impact of increased distance and reduced edge magnitude on flow attenuation. The "None" node, isolated by its significant distance and absence of a substantial connecting edge, receives negligible mass, highlighting how edge length and network topology influence flow distribution.

edges represent the conduits through which the dye is transported between them. The edge weights symbolize the velocity at which the dye moves, influencing the rate and pattern of distribution across the network. Similar to how dye is dispersed through water, the distance and material of the fabric affect how swiftly and thoroughly the dye permeates. Figure 2.2 illustrates this concept with a schematic diagram that delineates the flow from a source node within a directed network.

DFC incorporates geographic considerations by utilizing the physical distances and spatial relationships between nodes as integral components of its algorithm. By considering the flow of traffic or goods through the network and how it traverses the geographical landscape, DFC ensures that the communities it identifies are structurally coherent and geographically plausible. This spatial awareness in community formation makes the method particularly suited to transportation networks, where the physical distance between nodes significantly impacts network dynamics and efficiency.

Through a flow-simulation-based approach, DFC accounts for both connectivity and dynamic traffic patterns by modeling directed and weighted edges that represent real-world

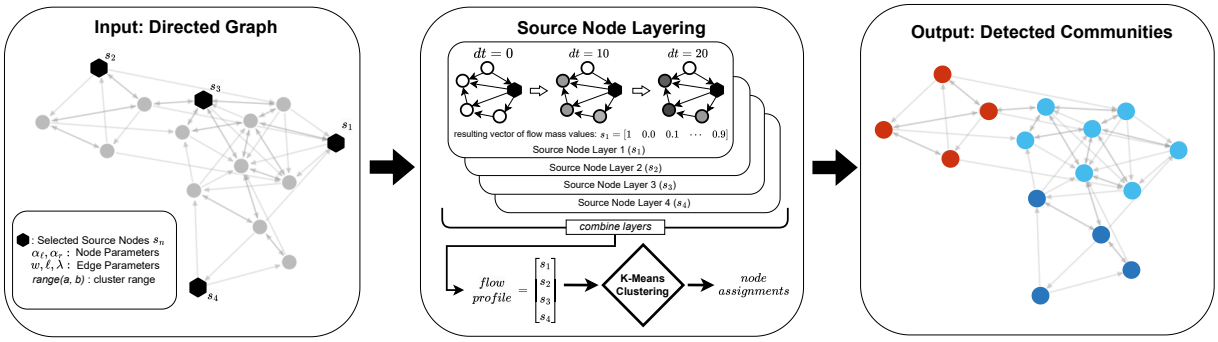


Figure 2.3: Overview of Directed Flow Communities. (Left) Input: A directed graph is initialized with designated source nodes and predefined parameters for nodes, edges, and clustering range. (Center) Source Node Layering: Each source node undergoes a discrete simulation over T timesteps, accumulating a mass vector s_n for every node within the graph. Flow profile represents the combined mass vectors across the network. K-Means clustering is applied to the flow profile, which partitions nodes into communities based on similarity, effectively detecting communities through their flow dynamics. (Right) Output: Final community assignments.

transportation routes and their capacities. The algorithm simulates the movement of traffic or goods across these routes, taking into account the directionality and volume of flow, which are critical in understanding the functional ties between different parts of the network. This process mirrors the utilization of transportation links, allowing DFC to adapt to the fluctuations in traffic volume and the varying importance of routes under different conditions through parameters.

2.3.1 Model Parameters

The model's behavior in network environments is governed by a set of parameters, classified into three primary groups that impact its performance and outcome.

- **Node Parameters:** *Absorption Rate and Absorption Limit*

  The absorption rate (α_r) specifies the quantity of mass a node retains at each timestep, while the absorption limit (α_ℓ) defines the maximum mass capacity a node can accommodate.

- **Edge Parameters:** *Velocity, Length, and Decay Rate*

Velocity (w), typically denoted by edge weight, refers to the rate at which mass travels along the edge, while length (d) represents the physical or conceptual distance between two nodes. Decay rate (λ) quantifies the rate at which mass diminishes as it moves through the edge and is generally constant throughout all edges. The specific units of edge weights (w) are inconsequential, as they function as abstract indicators of the mass transmission rate rather than concrete physical measurements. These parameters together influence the transmission of mass between nodes in the network.

- **Simulation Parameters:** *Timesteps*

  The number of timesteps (T) defines the discrete intervals at which the flow simulation state is updated.

The chosen parameters impact the model in two main aspects: spread dynamics and final community results. Specifically, node and edge parameters are crucial in setting the pace of the network's flow. However, the overall structure of the network ultimately dictates the formation and delineation of communities.

Parameters should be selected to align with the specific characteristics of the network under study. For instance, in the context of migration studies, the edge parameters for distance d can be adjusted to represent distances between geographic locations, and edge velocity w can represent the total amount of migration. In supply chain and logistics, decay rate λ can simulate the loss of goods during a shipment and absorption limit α_ℓ can represent the capacity of a hub.

The proposed model parameters offer a flexible yet robust framework for understanding and manipulating network dynamics. This model's effectiveness is contingent on carefully selecting its parameters, which should be tailored to the specifics of each network scenario. This approach ensures that the model can be effectively adapted and applied to a diverse range of network types and conditions.

2.3.2 Flow Simulation

The mass flow simulation begins with a selected subset of nodes, wherein each node is designated as a source for each distinct layer. In our simulation, "layering" refers to the process of sequentially conducting individual simulation runs for each source node and then aggregating these runs to discern community groupings. Specifically, each layer represents the outcome of a single run with one source node. During this run, we track the flow and absorption of mass throughout the network, originating from the source node. Once a simulation run is complete, its results are set aside as one layer of data. This process is repeated for each source node in the subset. After completing all individual runs, these layers are superimposed, or "stacked," creating a comprehensive flow profile of the network's mass flow dynamics). Analyzing this stacked data allows us to identify patterns and interactions between nodes, determining community groupings within the network (Figure 2.3).

Source nodes act as an infinite point of mass for the network and must have at least one outgoing edge.

As stated in the previous section, each edge has its own characteristics (d, w, λ). These characteristics influence how the mass travels along each edge. An upwind differential scheme is used to solve the advection-decay equation (2.1) for calculating the inflow $C_{\text{start}}(e)$ and outflow $C_{\text{end}}(e)$ mass concentrations along each edge. This computation is performed for each timestep dt over a total duration of T steps.

$$C_{\text{end}}(e) = C_{\text{start}}(e) - \left(\frac{w(e)}{d(e)} \cdot C_{\text{start}}(e) + \lambda(e) \cdot C_{\text{start}}(e) \right) \cdot dt \tag{2.1}$$

When mass is conveyed to a node via an edge, the node absorbs the mass I_{in} (2.2) at rate α_r until it reaches the absorption limit α_ℓ. Unabsorbed mass I_{out} is redistributed proportionally to outgoing edges (2.3). In cases where a node has reached α_ℓ, all arriving mass is passed along to outgoing edges.

$$I_{\text{in}}(n) = \sum_{e \in E_{\text{in}}(n)} C_{\text{end}}(e) \tag{2.2}$$

$$C_{\text{start}}(e) = \frac{w(e)}{\sum_{e \in E_{\text{out}}(n)} w(e)} \cdot I_{\text{out}}(n) \tag{2.3}$$

Absorption and redistribution occur at each timestep dt in the simulation. Upon initiating the run with a new source node, total absorbed values are reset to ensure a fresh start for each simulation cycle. Additionally, this method disregards self-loops to maintain the integrity of the simulation dynamics.

Once simulations for identified source nodes are completed, we apply K-means clustering [43] to the flow profile for community detection. When the number of communities is unknown, an approximate range can be used to determine the optimal number of clusters, which can be found using the elbow method or the highest silhouette score.

2.4 Validation and Testing

This section is divided into three parts. First, experiments are conducted on benchmark synthetic graphs, detailed in Section 2.4.1, to test the model in controlled environments and validate our method. Second, experiments on empirical graphs are presented in Section 2.4.2, offering insights into the model's performance in real-world scenarios. These experiments involve comparisons against three established community detection methods: Infomap [35], which shares flow-based properties with our model; Directed Louvain [44], a modularity-based clustering method; and MotifCluster [39], a technique that focuses on pattern-based clustering.

For networks with ground truth communities, normalized mutual information (NMI) and adjusted mutual information (AMI) are used as the basis for information-theoretic measures of clustering comparisons. NMI measures the similarity between two clustering results by comparing the mutual information normalized by the average of the entropy of each clustering. Mathematically, NMI between two clusterings U and V is defined as:

$$\text{NMI}(U, V) = \frac{2 \cdot I(U; V)}{H(U) + H(V)}$$

where $I(U; V)$ is the mutual information between U and V, and $H(U)$ and $H(V)$ are the entropies of U and V, respectively. NMI ranges from 0 (no mutual information) to 1 (perfect

correlation).

AMI is a variant of NMI that adjusts for chance, correcting for the fact that random clustering assignments could yield a non-zero mutual information [45]. The AMI is given by:

$$\text{AMI}(U, V) = \frac{I(U;V) - E[I(U;V)]}{\text{avg}(H(U), H(V)) - E[I(U;V)]}$$

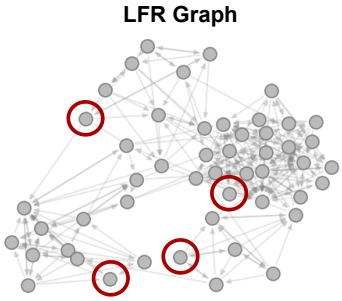
where $E[I(U;V)]$ is the expected mutual information between U and V under the null hypothesis of independence. This adjustment makes AMI more robust in comparing clustering results, as it accounts for the possibility of random agreements. There are various other measures for evaluating clustering performance, as discussed in comparative metric studies [40], [45], [46], but NMI and AMI are chosen due to their extensive application and proven reliability in community detection.

2.4.1 Synthetic Datasets

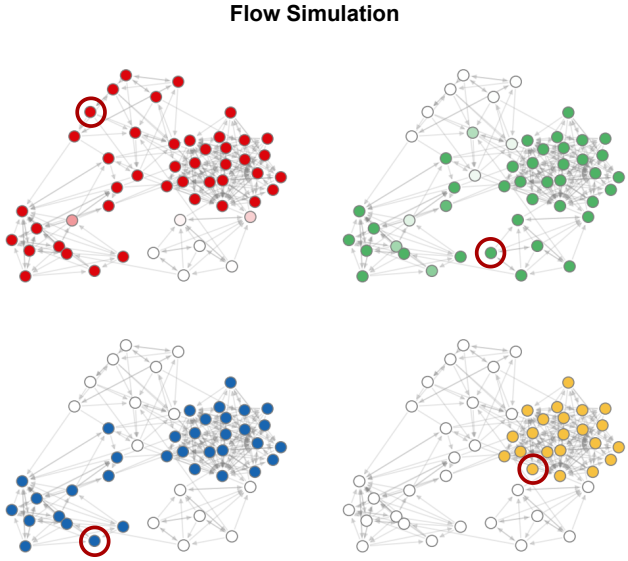
The Lancichinetti-Fortunato-Radicchi (LFR) benchmark graphs are synthetic weighted, directed networks with predefined community structures, designed to test community detection algorithms [24]. These graphs are notable for their ability to mimic real-world network characteristics, such as power-law distributions for node degrees and community sizes. The typical parameters for LFR benchmark graphs include a range of average node degrees, community sizes, mixing parameters, and other structural features. These parameters can be adjusted to create networks with varying degrees of complexity and community overlap, allowing researchers to tailor the graphs to specific experimental needs. The flexibility in parameter settings ensures that the LFR benchmark can closely replicate diverse real-world network scenarios.

A key feature of LFR graphs is the mixing parameter (μ), which controls the fraction of a node's links that connect to nodes in other communities. This parameter adjusts the network's community structure, with lower values indicating clearer community boundaries and higher values making community detection more challenging.

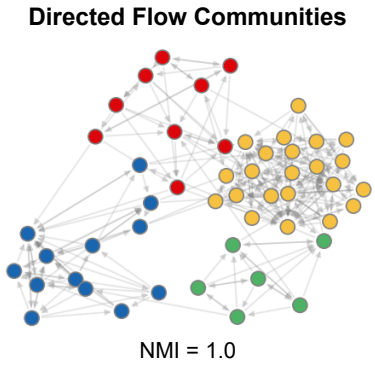
LFR benchmarks allow for customization of various aspects, including network size, community size distribution, and the level of community overlap. This flexibility makes LFR



(a) Input LFR graph with community ground truths. Selected source nodes are circled.



(b) Flow simulations representing the accumulation of flow from each source node, with node color intensity indicating mass accumulation.



(c) The output graph of detected communities with an NMI score of 1.0.

Figure 2.4: Community Detection with Directed Flow Communities on a 50-node LFR graph.

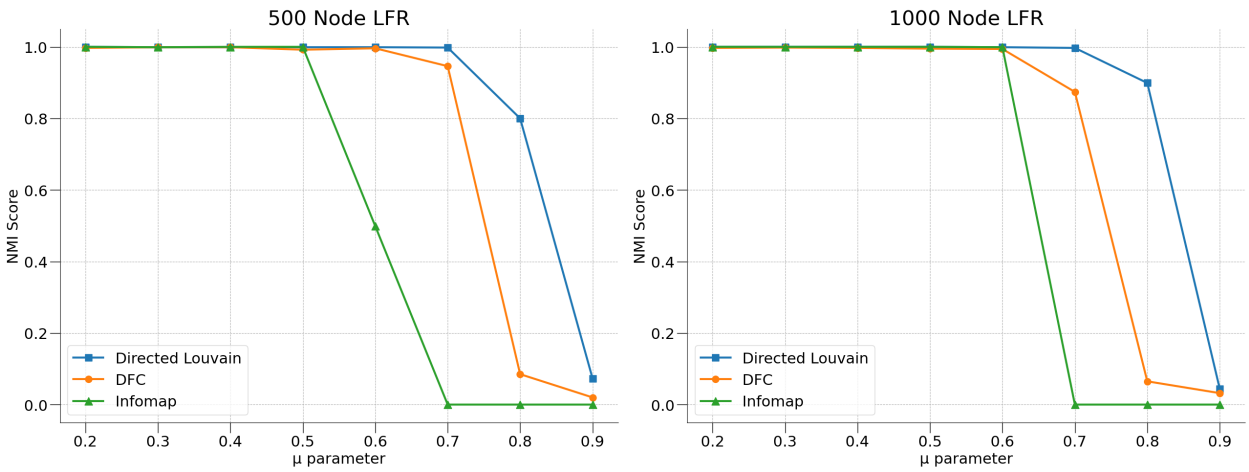


Figure 2.5: Graphs of community detection performance on 500-node and 1000-node LFR benchmark graphs with varying μ parameters on Directed Louvain, Infomap and Directed Flow Communities.

graphs a comprehensive tool for evaluating the performance of community detection algorithms across different scenarios, providing insights into their effectiveness in revealing the underlying community structure in complex networks.

Tests were systematically conducted on 500-node and 1000-node directed weighted LFR graphs, as shown in Figure 2.5. For each distinct μ value, 100 graphs were generated. In these tests, only the μ values were varied, while all other input parameters were kept constant. The μ value, often referred to as the mixing parameter, modifies the strength of the community structure within the graph. For the 500 node experiments, community size distribution was set to 20 and 50 nodes and the degree distribution was between 35 and 75. For the 1000 node experiments, community size distribution was set to 40 and 70 nodes and the degree distribution was between 35 and 75. Figure 2.4 showcases a demonstration of the Directed Flow Communities method on a smaller, 50-node LFR benchmark graph.

In the range of μ values from 0.2 to 0.6, the performance of the Directed Flow Communities method was comparable to that of other methods, but a decline in its effectiveness was observed for μ values exceeding 0.8. Notably, our method outperformed Infomap in scenarios with $\mu > 0.60$, but it's performance still lagged Directed Louvain.

Table 2.1: Comparative Results for Community Detection in the EU-Core Dataset. This dataset represents an email communication network from a European research institution, characterized by 42 communities (1005 nodes, 25571 edges).

	EU-Core			
Algorithm	No. of Comms.	AMI	NMI	% of Clustered Nodes
DFC	37	0.472	0.560	100%
Dir. Louvain	36	**0.565**	0.621	100%
Infomap	91	0.531	0.720	100%
Motif - Ms	42	0.330	0.331	98.2%
Motif - M3	42	0.516	0.517	77.0%
Motif - M8	42	0.345	0.338	96.9%

Table 2.2: Comparative Results for Community Detection in the Cora Dataset. The Cora dataset is a citation network of scientific publications categorized into seven communities (2708 nodes, 5429 edges).

	Cora			
Algorithm	No. of Comms.	AMI	NMI	% of Clustered Nodes
DFC	44	**0.277**	0.310	100%
Dir. Louvain	2707	0.000	1.0	100%
Infomap	176	0.237	0.329	100%
Motif - Ms	7	0.133	0.127	87.5%
Motif - M8	7	0.158	0.141	79.4%

2.4.2 Empirical Datasets

To extend this research to real world directed networks, we conducted tests on two graphs with ground truths: EU-Core [47] and Cora [48], as shown in Table 2.1 and 2.2. The EU-Core network is composed of email data from a European research institution, capturing the communication patterns among individuals. It features a dense connectivity structure with each individual belonging to one of 42 departments. The Cora dataset is a citation network consisting of scientific publications classified into one of seven classes. Each publication in the dataset is described by a binary word vector indicating the absence or presence of the corresponding word in a dictionary of 1,433 unique keywords. Cosine similarity was used to transform text similarity vectors into unique edge weights [49]–[51].

The unweighted nature of the EU-Core network, coupled with the absence of additional features, offers perspective into the intuitive spread of information within a network. However, this simplicity can lead to a reduction accuracy when detecting and evaluating community structures using Directed Flow Communities. In the Cora Dataset, Directed Flow Communities has a 16.87% improvement on AMI scores over Infomap.

An interesting takeaway from the comparative analysis concerns the potential inflation of NMI scores, especially in the context of algorithms like Directed Louvain and Infomap. In such algorithms, overestimating the number of communities can lead to high NMI scores, falsely suggesting a higher level of accuracy in community detection than actually exists. A more accurate representation of the scores can attained with AMI values, which account for the agreement between clusterings that occurs purely by chance.

Unlike community detection methods that discover how many communities exist in a given network, MotifCluster requires this value at the outset. In practice, the number of communities must be known beforehand or estimated using a methodological approach. Additionally, this method focuses on clustering the most strongly connected components of a network [39]. This bias is rooted in the assumption that the most meaningful and coherent community structures are likely to be found within these highly interconnected segments. MotifCluster with the M3 motif was not able to find communities for the Cora dataset.

2.5 Source Node Selection

Table 2.3: Experimental Comparison of Source Node Selection Strategies on an 802-Node LFR Network ($\mu = 0.3$). The table compares AMI and NMI with standard deviations derived from 100 iterations on an LFR graph. It evaluates the efficacy of various node selection methods for different quantities of source nodes.

(a) For 50 and 100 Source Nodes

	50 Nodes		100 Nodes	
Strategy	AMI $\pm\sigma$	NMI $\pm\sigma$	AMI $\pm\sigma$	NMI $\pm\sigma$
Low Btwns.	0.813 ± 0	0.820 ± 0	0.984 ± 0	0.985 ± 0
High Btwns.	0.781 ± 0	0.788 ± 0	0.882 ± 0	0.886 ± 0
Random	0.724 ± 0.169	0.773 ± 0.168	0.944 ± 0.047	0.947 ± 0.046
Hybrid	0.784 ± 0.116	0.792 ± 0.115	0.941 ± 0.042	0.944 ± 0.042

(b) For 250 and 500 Source Nodes

	250 Nodes		500 Nodes	
Strategy	AMI $\pm\sigma$	NMI $\pm\sigma$	AMI $\pm\sigma$	NMI $\pm\sigma$
Low Btwns.	1.0 ± 0	1.0 ± 0	1.0 ± 0	1.0 ± 0
High Btwns.	0.987 ± 0	1.0 ± 0	1.0 ± 0	1.0 ± 0
Random	0.994 ± 0.008	0.995 ± 0.008	0.997 ± 0.006	0.997 ± 0.006
Hybrid	0.993 ± 0.010	0.992 ± 0.011	0.998 ± 0.004	0.997 ± 0.005

In smaller networks, it is typically advantageous to apply the Directed Flow Communities algorithm to all nodes, as the computational load is manageable. But as network size increases, the processing time can become prohibitive. To maintain efficiency in larger graphs, judiciously selecting a representative subset of nodes for analysis is recommended.

Betweenness centrality [52], [53] quantifies a node's role as an intermediary across the shortest paths in a graph. It is denoted as $C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} represents the total shortest paths from node s to t, and $\sigma_{st}(v)$ is the count of those paths passing through node v. This metric finds nodes critical for information flow and is helpful for selecting source nodes in network flow simulations.

Selecting source nodes with high betweenness centrality for flow simulations often channels the flow along the network's most central paths, potentially overemphasizing well-connected communities and neglecting those on the periphery. This approach tends to

highlight the network's main arteries but can skew the detection towards larger, more visible communities, possibly overlooking smaller, less central communities.

On the other hand, using nodes with low betweenness centrality as sources can illuminate less conspicuous communities by tracing flow through less central pathways, offering insights into the network's more nuanced structures. An entirely random selection of source nodes introduces variability, affecting the consistency of detected communities and potentially introducing bias due to the stochastic nature of the selection process.

To counteract these limitations and ensure a balanced exploration of the network, one can adopt a hybrid strategy that combines a selection of nodes characterized by high and low betweenness centrality with nodes chosen at random. However, strategies that exclusively select nodes with low betweenness centralities have been the most effective. Table 2.3 shows the efficacy of different strategies within an 802-node LFR network. The table compares strategies through AMI and NMI metrics across 50, 100, 250, and 500 source nodes.

Chapter 3

Freight Analysis Framework Case Study

The Freight Analysis Framework (FAF) database provides comprehensive estimates of freight movement within the United States [54]. The Bureau of Transportation Statistics, in collaboration with the Federal Highway Administration, compiles the FAF for 42 different types of commodities, encompassing data across state and metropolitan regions. Data sources include the 2017 Commodity Flow Survey (CFS), foreign trade statistics, and sector-specific data from agriculture, extraction, utilities, construction, and other services. In this case study, we use version 5.1 of the FAF dataset. FAF version 5.1 contains data for the base year 2017 through to 2022, forecast year estimates (2023-2050), and state level historical trend estimates (1997-2012).

The FAF categorizes freight data into three principal dimensions: weight, value, and transportation activity. Freight weight is reported in thousands of tons, measuring the physical bulk of goods transported. The value of these goods is expressed in millions of dollars. Transportation activity is quantified in millions of ton-miles. Table 3.3 is a comprehensive representation of various commodities transported within the United States in 2017, per the FAF dataset. In the context of AAM, the FAF dataset can serve as a proxy for future AAM package demand. In another case study by Gunady et al., FAF is used for AAM demand modeling [55]. In this chapter, we conduct experiments on three specific areas of the FAF dataset.

3.0.1 Datasets

Table 3.1: Graph Characteristics of FAF Datasets

Dataset	Nodes	Edges	$\langle k \rangle$	σ_k	λ_h	τ	r	c
Pharmaceuticals	129	11192	173.51	49.27	248.33	908	-0.20	0.79
Electronics	129	16150	250.38	10.31	1161.25	325	-0.05	0.98
Total Freight	129	16496	255.75	1.00	138971.05	104	-0.02	0.99

Three sub-datasets from the FAF framework were selected for this case study. Pharmaceuticals were chosen because their critical importance in healthcare necessitates fast, secure, and regulated transport conditions, making them ideal for AAM adoption [56]. Electronics, a high-value commodity, are strategic investment points for AAM [3]. Lastly, an analysis of comprehensive freight flows across all commodities was conducted to identify key transportation corridors and regions with significant freight activity.

A graph characteristics is shown in Table 3.1. In this table, the average degree is denoted as $\langle k \rangle$ and the standard deviation of node degrees is denoted as σ_k. The largest eigenvalue of the adjacency matrix is denoted as λ_h. The mixing time, τ, estimates the time required for a random walk to reach its steady-state distribution. The assortativity coefficient, r, quantifies whether nodes tend to connect to others with similar degrees. If positive, this indicates that nodes are inclined to connect to other nodes with similar degrees. Lastly, the clustering coefficient, c, captures the likelihood that two neighbors of a node are also neighbors themselves, reflecting the density of local clusters in the graph.

3.0.2 Assumptions

Despite current technological constraints, this analysis assumes the feasibility of long-distance AAM flights for national-level transportation. This case study aims not only to understand the potential of AAM in the context of existing transportation networks but also to explore hub location problems within a more expansive framework of an entire country.

3.1 Approach

We begin by decomposing the transportation networks into clusters. By segmenting the network into distinct communities with dense internal connections, we simplify the task of identifying optimal hub locations to smaller sub-problems. This tactic acknowledges the interactions between nodes, offering a more context-sensitive hub selection process and aids in optimizing hub allocation by aligning with the network's inherent clustering. We then employ eigenvector centrality, degree centrality, and integer programming to designate hubs within each identified community for ten hubs. When community detection yields more or fewer than ten communities, we adjust by making multiple selections from the larger communities or excluding the smaller ones from the hub selection. Additionally, for comparative analysis, we apply an integer programming model to find hubs on a subset of 30 nodes with the highest degree.

To evaluate the cost-effectiveness and robustness of the hubs identified by the integer programming (IP) approach and centrality metrics (described in Sections 3.1.3 and 3.1.4, respectively), we undertake a comparative analysis. The nominal costs associated with all ten hubs and scenarios where the 1, 2, 3, and 4 largest inoperative hubs were tested. A lesser cost escalation from normal operations to scenarios with non-functional hubs suggests greater robustness of the hub configuration. The methodology of our approach is as follows:

1. Identify 10 Hubs: For each community configuration, 10 hubs were initially selected using methods such as integer programming and centrality metrics.

2. Select Largest Hubs: Among the selected hubs, those with the largest sum of incoming and outgoing edge weights were identified.

3. Simulate Hub Failures: The analysis simulated failures by sequentially removing the largest hubs from the network. Scenarios were created where 1, 2, 3, and 4 of the largest hubs were rendered inoperative.

4. Recalculate Costs: For each failure scenario, the expected transportation costs were recalculated. This involved determining new optimal routing paths and computing the

cost of freight movement across the network with the remaining operational hubs.

5. Evaluate Robustness: The robustness of each community configuration was assessed by examining the increase in transportation costs due to the hub failures. Lower increases in costs indicate higher robustness, as the network can maintain its efficiency and functionality despite the loss of key hubs.

6. Compare Configurations: The cost impacts for each community configuration were plotted to visualize and compare their robustness. The configurations with the smallest cost increases across different failure scenarios were deemed the most robust.

3.1.1 Community Detection

Infomap, Directed Louvain, and Directed Flow Communities (DFC) are used to examine and contrast the community structures across three sub-datasets of the Freight Analysis Framework. The assessment of community quality will hinge on two key metrics: conductance ($\phi(S)$) and the aggregate of freight flows within each community. Lower conductance values indicate more distinct separation between communities. Equation 3.1 defines the conductance of a subset S in a graph G.

$$\phi(S) = \frac{c(S, \bar{S})}{\min(a(S), a(\bar{S}))} \tag{3.1}$$

Where:

- $c(S, \bar{S})$ is the cut size, i.e., the number of edges with one endpoint in S and the other in $\bar{S} = V \setminus S$.
- $a(S)$ is the volume of S, defined as the sum of the degrees of vertices in S.
- $\bar{S}$ is the complement of S in G.

3.1.2 Hub Selection Costs

We consider the transportation costs within a traditional airline cargo hub-and-spoke network, employing a cost ratio of 1:1.7:10 for different flight types, corresponding to Boeing 747

(hub-to-hub), Boeing 767 (hub-to-spoke), and Cessna 208 Caravan (point-to-point) flights, respectively. Despite AAM flights potentially incurring lower costs, this analysis assumes the same cost ratio to maintain consistency with economies of scale. A script calculates and compares the direct route cost, the cost including a single hub stop, and the cost involving stops at two distinct hubs for each origin-destination (OD) pair, and opts for the most cost-effective route for each pair. Table 3.2 contains the costs per nautical mile per ton for each flight type.

Table 3.2: Cost per Nautical Mile per Ton for Different Flight Types

Flight Type	Cost ($ per nautical mile per ton)
Hub-to-Hub	0.496
Hub-to-Spoke	0.858
Point-to-Point	5.154

3.1.3 Optimal Hub Selection using Integer Programming

In the uncapacitated multiple allocation p-hub problem, each node can send and receive traffic through any hubs with no capacity limits. Yaman [57] created an IP for multiple allocation p-hub problem, in which given a network with a set of nodes N and a set of edges E, let t_{ij} be the amount of traffic to be routed from node i to node j. Let d_{ij} be its associated unit routing cost. We use a similar construction for our baseline IP but modify the cost structure to utilize a mileage-based cost factor described in the previous section. We define the following variables:

- t_{ij}: Represents the volume of traffic that needs to be routed from node i to node j.
- d_{ij}: Denotes the unit cost associated with routing traffic from node i to node j.
- z_{kk}: Equals 1 if node k is designated as a hub; otherwise, it is 0.
- z_{ik}: Set to 1 if a non-hub node i is allocated to a hub node k; it is 0 if there is no allocation.

- f_{ijkl}: The fraction of traffic t_{ij} from node i to node j that is routed through the path from $i \to k \to l \to j$, where both k and l are hubs.

- χ, α, δ: Represents the cost factors associated with routing traffic from an origin node to a hub (χ) between hubs (α), and from a hub to a destination node (δ), respectively.

The minimization problem is as follows:

$$\min \sum_{i \in N} \sum_{j \in N} \sum_{k \in N} \sum_{l \in N} t_{ij} \left(\chi d_{ik} + \alpha d_{kl} + \delta d_{lj} \right) f_{ijkl} \tag{3.2}$$

Subject to:

$$\sum_{k \in N} z_{ik} \leq r \quad \forall i \in N \tag{3.3}$$

$$z_{ik} \leq z_{kk} \quad \forall i, k \in N \tag{3.4}$$

$$\sum_{k \in N} z_{kk} = p \tag{3.5}$$

$$\sum_{k \in N} \sum_{l \in N} f_{ijkl} = 1 \quad \forall i, j \in N \tag{3.6}$$

$$\sum_{l \in N} f_{ijkl} \leq z_{ik} \quad \forall i, j, k \in N \tag{3.7}$$

$$\sum_{k \in N} f_{ijkl} \leq z_{jl} \quad \forall i, j, l \in N \tag{3.8}$$

$$f_{ijkl} \geq 0 \quad \forall i, j, k, l \in N \tag{3.9}$$

$$z_{ik} \in \{0, 1\} \quad \forall i, k \in N \tag{3.10}$$

The model has a complexity of $O(|N|^4)$. It has $O(|N|^4)$ variables and $O(|N|^3)$ constraints [57]. This level of complexity arises because each pair of nodes (i, j) must consider all possible hub pairs (k, l) for routing, resulting in the quartic growth of variables with the number of nodes. Each constraint type, such as ensuring that traffic flows through designated hubs and satisfying hub allocation conditions, contributes to the cubic growth in the number of constraints. Such complexity highlights the computational challenges in solving the multiple allocation p-hub problem optimally, especially for large networks.

3.1.4 Centrality Based Hub Selection

Centrality metrics offer a quantitative measure to identify influential nodes within a network, each providing a unique perspective on a node's role and importance.

Degree centrality (3.11) evaluates the prominence of a node by counting its direct connections, emphasizing nodes with a high potential for interaction within the network[58]. The degree centrality for a node v is defined as:

$$C_D(v) = \frac{|\{e \in E : v \in e\}|}{N - 1} \tag{3.11}$$

Eigenvector centrality (3.12) extends the concept of degree centrality by considering not just the quantity but the quality of connections, prioritizing nodes connected to other influential nodes [59].

$$\lambda x_i = \sum_{j=1}^{n} A_{ij} x_j \tag{3.12}$$

In which λ is the largest eigenvalue, x_i represents the eigenvector centrality of node i, and A_{ij} is an element of the adjacency matrix A, indicating the presence (1) or absence (0) of a link between nodes i and j.

The selection of centrality metrics for hub identification largely depends on the network's specific characteristics and the objectives of the hub location problem. Degree centrality is most effective in dense networks where hubs are expected to handle high volumes of traffic directly [60]. Eigenvector centrality is suited for networks where the influence of a node is derived from its connections to other influential nodes, often applicable in hierarchical networks [61]. In the context of HLPs, centrality-based hub selection considers the underlying network structure and the strategic importance of nodes.

A centrality based approach can identify potential hubs that are well-connected and hold significant influence over the flow of traffic through the network. However, one significant limitation is the static nature of most centrality measures, which may not adequately capture the dynamic changes in traffic patterns and network topology [62]. Additionally, reliance on centrality metrics alone may overlook other critical factors such as geographical constraints, necessitating a more comprehensive approach that considers a wider array of factors to

ensure the practicality and effectiveness of the selected hubs [63]. Addressing these challenges requires a balanced approach that combines centrality metrics with other decision-making criteria and leverages advancements in computational methods to manage the complexity and dynamics of transportation networks.

3.2 Results

This section presents the outcomes of comparing various hub selection methods. Details on the community structures and configurations of hubs for each method are provided in C. For our analysis, we chose to use 10 hubs for each configuration. Selecting 10 hubs was well-suited for the size and distribution of the network, offering a good balance between efficiency and computational feasibility. Additionally, this number is manageable within the computational limits of our integer programming model, ensuring that we can obtain optimal or near-optimal solutions without excessive computational overhead.

In the context of this case study, robustness describes the ability of the freight transportation network to maintain operational efficiency when there are failures in one or more hubs. This involves the network's capacity to adapt to changes, reconfigure routes, and continue functioning effectively under adverse conditions. The analysis evaluates robustness by examining the cost impact when the largest hubs are rendered inoperative, with lower cost increases indicating higher robustness. The ability to minimize cost escalations during these scenarios reflects the network's capacity to withstand and recover from disruptions.

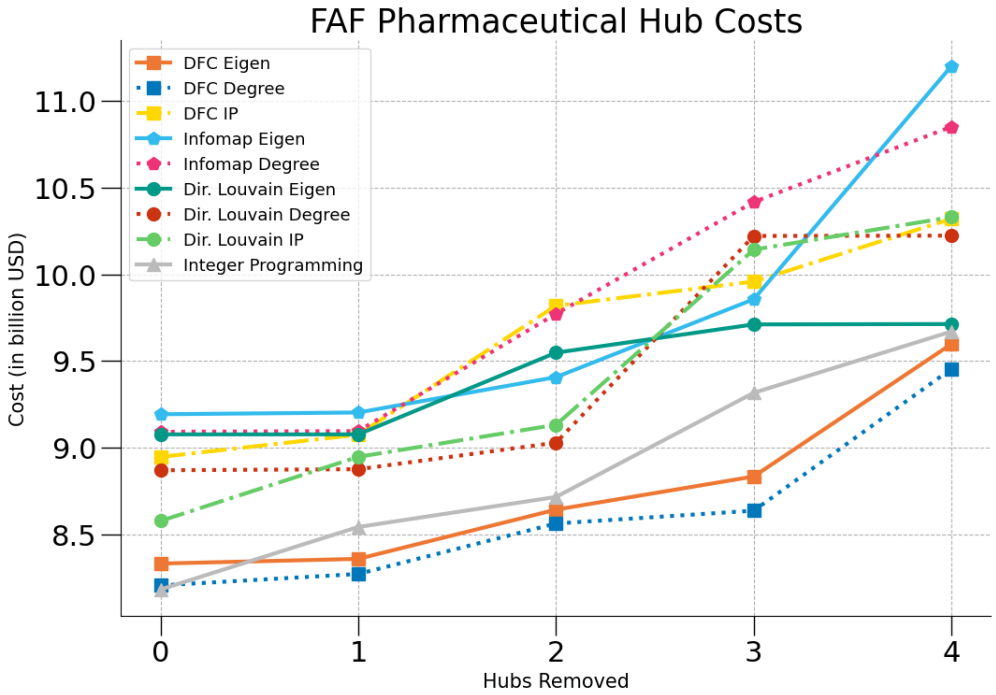


Figure 3.1: Impact of hub removal on pharmaceutical distribution costs (in USD) for the FAF pharmaceutical dataset, represented through various hub configurations. The graph compares the cost changes across different configurations derived from three community detection methods.

3.2.1 Pharmaceuticals

Figures C.2, C.3, and C.4 present the variations in community clustering derived from community detection methods applied to the Pharmaceutical dataset. Figure C.1 shows the 30 largest nodes used in the IP approach. Figure 3.1 compares the quantitative assessment of the network's robustness by illustrating the impact of removing the largest within each community configuration, as per the approach described in Section 3.1. Hub configurations are shown in Figures C.12, C.13, C.14, C.21, C.24, C.26, and C.29,.

Using Integer Programming alone resulted in the lowest initial costs. However, when compared with the hubs chosen through the Directed Flow Communities method, combined with degree-based hub selection, the latter's hubs exhibited greater robustness, with only a marginal 0.83% increase in initial costs.

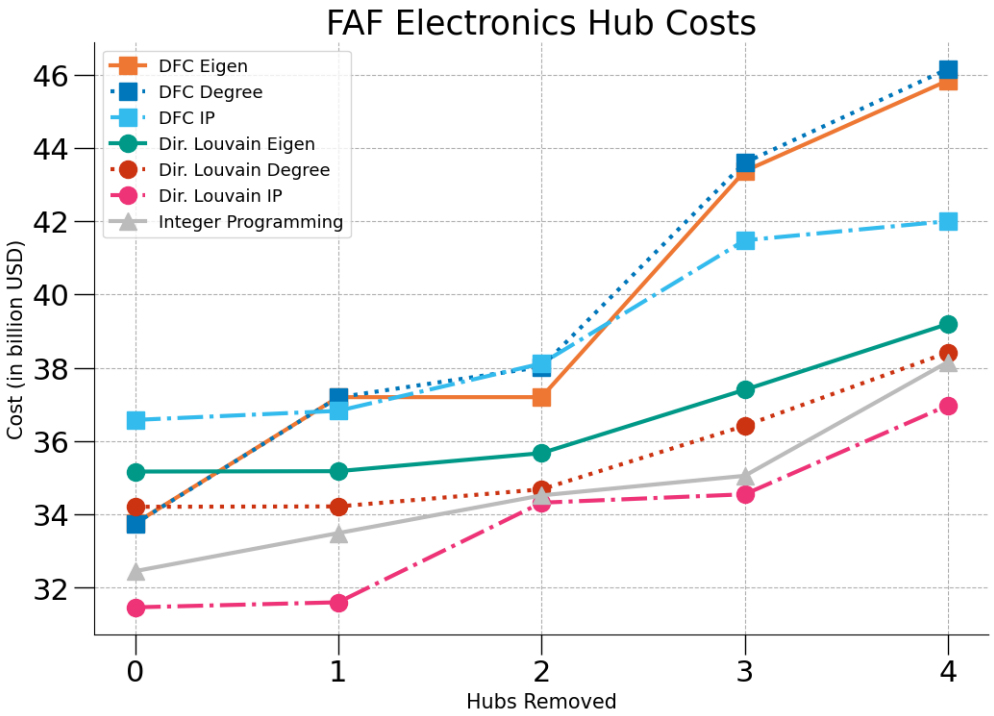


Figure 3.2: Impact of hub removal on electronics distribution costs (in USD) for the FAF electronic dataset, represented through various hub configurations.

3.2.2 Electronics

When applying the hub selection approach to the electronics dataset, the community structures revealed through network analysis are depicted in Figures C.6 and C.7. Figure C.5 shows the 30 largest nodes used in the IP approach. The Infomap algorithm did not yield meaningful communities for this dataset, and was omitted. Hub configurations are shown in Figures C.15, C.16, C.25, and C.28.

Figure 3.2 presents a comparative analysis of hub configuration costs, with the directed Louvain communities, when paired with IP hub selections, yielding the most optimal cost implications. Leveraging IP on the 30 largest nodes emerged as the second-best approach, highlighting the potential of scale and connectivity in influencing hub efficiency.

3.2.3 Total Freight Flows

Finally, hubs were selected using communities from the total freight flows dataset (Figures C.9, C.11, C.10). Figure C.8 shows the 30 largest nodes used in the IP approach. Hub

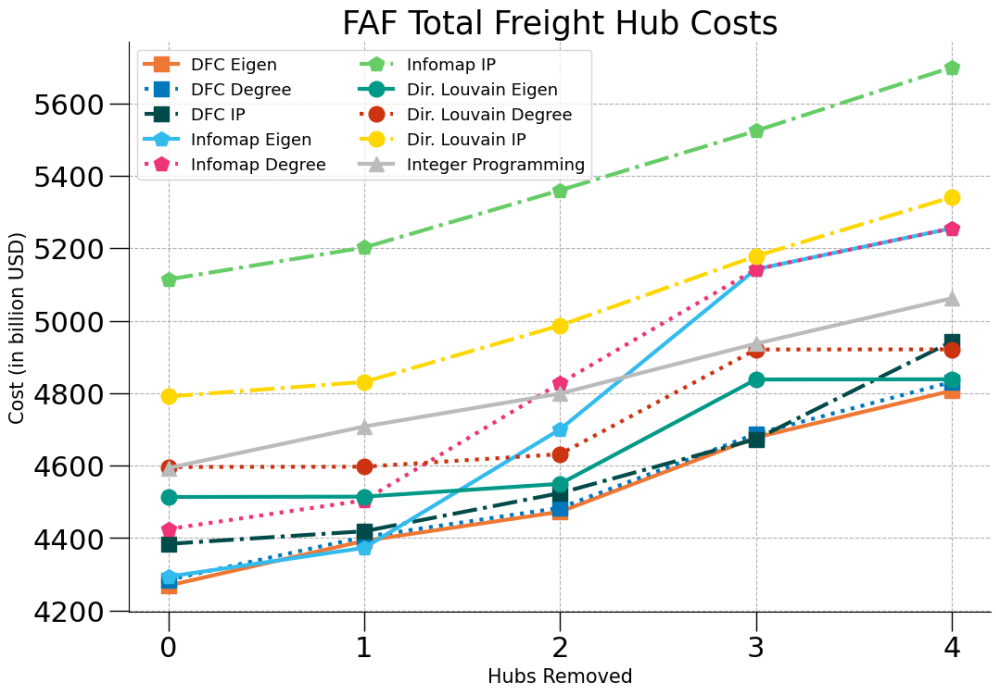


Figure 3.3: Impact of hub removal on total freight distribution costs (in USD) on the FAF total freight flows dataset, represented through various hub configurations.

configurations are shown in Figures C.17, C.18, C.19, C.22, C.23, C.27, and C.30.

An examination of the hub costs, depicted in Figure 3.3, reveals that the hubs identified through the Directed Flow Communities (DFC) method yielded the lowest overall costs.

3.3 Discussion

This chapter presented a case study using the Freight Analysis Framework dataset, emphasizing the potential of Advanced Air Mobility in transforming freight transportation in the United States. We delved into the feasibility of employing graph analysis techniques to deconstruct the hub selection problem within freight networks, utilizing community detection methods and centrality metrics. Through this exploration, we demonstrated how these analytical approaches can identify highly interconnected regions and central nodes, providing a strategic framework for optimizing hub placements. This methodological blend not only enhances our understanding of network structures but also offers actionable insights for developing more efficient and robust transportation systems.

In analyzing the FAF datasets, the Infomap algorithm tended to identify communities that were more geographically dispersed. This led to the formation of clusters that resembled outcomes typically associated with spatial analyses of hub location problems, as discussed in studies like that by Peker et al. [17].

Notably, the Integer Programming approach, when confined to the 30 largest nodes, did not yield promising hub configurations on the largest dataset, Total Freight Flows. This strategy aimed to reduce computational demands by concentrating on the nodes presumed to be most crucial. However, this approach may have overlooked essential aspects of network behavior, especially those not evident in the most prominent nodes. A case in point is the omission of Chicago, IL, as a hub (Figure 3.4). Despite being included in the 30 nodes selected for the dataset, it was not classified as a hub in the subgraph it serves. However, in the complete graph, the Chicago hub is a critical junction for numerous smaller nodes, dramatically reducing operating costs. In the smaller electronics and pharmaceutical datasets, the IP only approach did result relatively low operating nominal operating costs, but was less robust overall.

When applied to the pharmaceutical and total freight flow datasets, the Directed Flow Communities method identified hub configurations that either was, or was close to, the lowest operational costs. This indicates that DFC's approach to community detection, which emphasizes the direction and weight of connections, is particularly effective in pinpointing cost-efficient hub locations within these networks. However, it's noteworthy that the communities detected by DFC exhibited higher conductivity scores compared to those identified by other methods. Higher conductivity scores suggest that these communities might be less tightly knit, with more connections leading outside the community. This could imply a trade-off between cost efficiency and the internal cohesion of the communities, where DFC optimizes for the former at the expense of the latter.

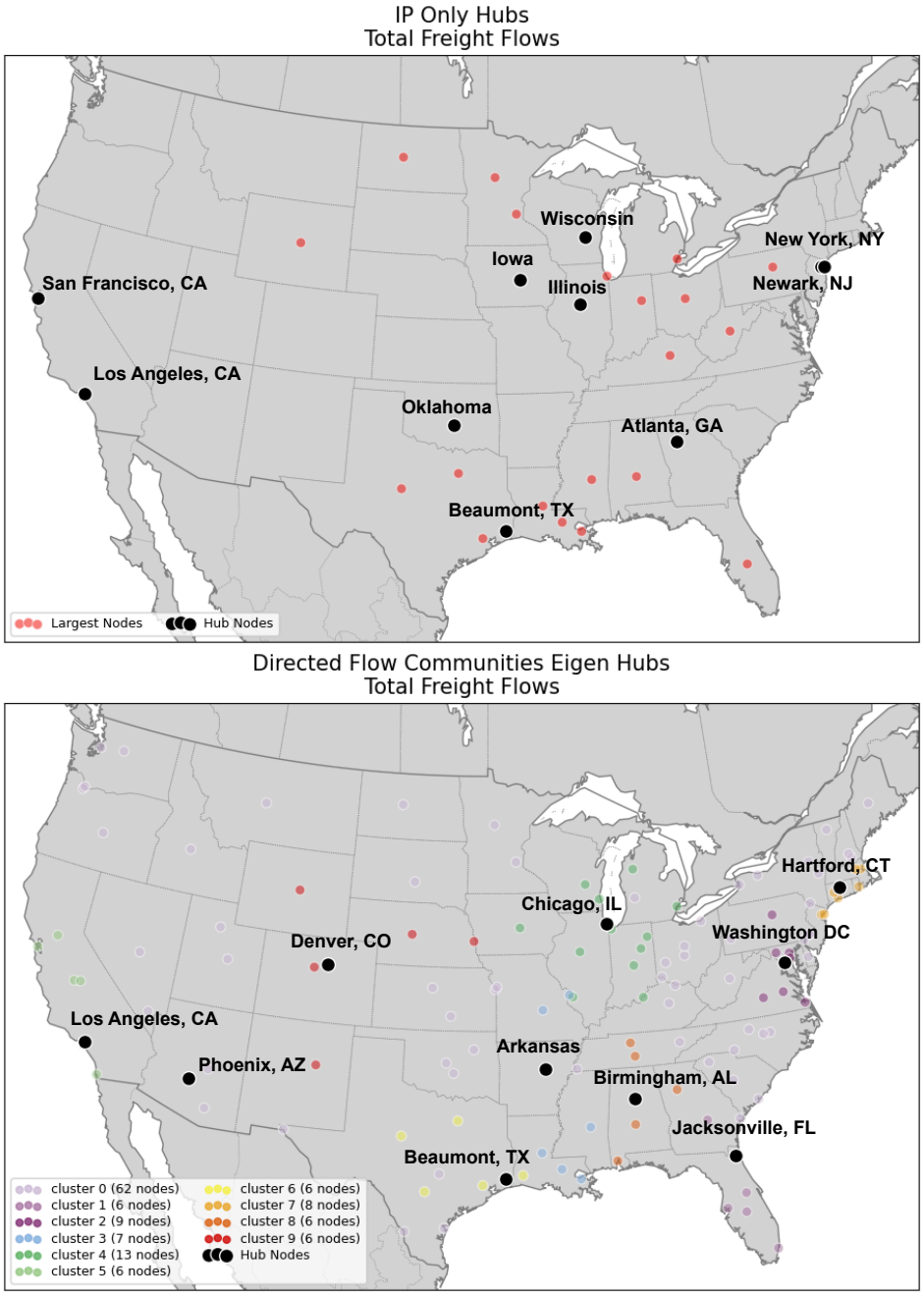


Figure 3.4: Comparison of Two Hub Configurations for the Total Freight Flows Dataset. (Left) Hub configuration based on integer programming from a subset of the 30 largest hubs, resulting in a total cost of 4593.55 billion USD. (Right) Hub configuration based on eigenvector centrality chosen from Directed Flow Communities, with a total cost of 4269.06 billion USD.

Table 3.3: Transportation details of various commodities within the United States for the year 2017, based on data from the Freight Analysis Framework. This table categorizes commodities into several types and provides three key metrics for each: the total weight transported in thousands of tons, the total value in million dollars, and the total transportation activity in million-ton/miles. Commodities listed in bold are used for experiments.

Commodity Type	**Thousand Tons**	**Million Dollars**	**Million-Ton/Miles**
01 - Live animals/fish	89915.84	178546.08	16814.77
02 - Cereal grains	1322356.77	177703.64	420353.98
03 - Other ag prods.	751996.13	420512.64	266485.48
04 - Animal feed	451597.80	158859.67	130306.53
05 - Meat/seafood	106863.43	403969.25	54937.03
06 - Milled grain prods.	142043.87	214826.67	63046.35
07 - Other foodstuffs	684143.23	706454.15	247106.11
08 - Alcoholic beverages	127735.60	253394.73	39590.45
09 - Tobacco prods.	4806.68	82644.65	983.16
10 - Building stone	16518.32	7220.96	2857.84
11 - Natural sands	624132.45	12263.05	130970.67
12 - Gravel	1906165.02	20024.53	186590.19
13 - Nonmetallic minerals	294926.52	28410.62	83785.26
14 - Metallic ores	92835.58	27521.89	43247.41
15 - Coal	894422.38	35910.46	526032.47
16 - Crude petroleum	1009183.33	323139.05	534915.05
17 - Gasoline	1476339.84	785848.51	206829.03
18 - Fuel oils	1052775.27	518485.29	155154.49
19 - Natural gas	2768256.06	668100.20	581697.22
20 - Basic chemicals	483088.12	360706.42	183997.17
21 - Pharmaceuticals	**23973.32**	**1188667.12**	**12162.53**
22 - Fertilizers	209079.22	66628.96	64989.35
23 - Chemical prods.	146624.84	483703.02	71813.66
24 - Plastics/rubber	282014.35	825242.10	148670.32
25 - Logs	481662.84	15499.93	67309.22
26 - Wood prods.	407390.71	263818.012	115905.68
27 - Newsprint/paper	170220.68	150043.38	97505.24
28 - Paper articles	86630.80	163057.84	30013.03
29 - Printed prods.	31463.39	146338.55	13094.23
30 - Textiles/leather	63710.02	692500.73	38525.89
31 - Nonmetal min. prods.	1275292.34	275028.36	175763.07
32 - Base metals	379031.56	524097.96	137608.14
33 - Articles-base metal	161531.98	496850.38	68066.53
34 - Machinery	133006.46	1225192.17	68693.22
35 - Electronics	**83059.28**	**1750327.30**	**50064.02**
Total Freight Flows	**19786384.60**	**18906784.40**	**5436308.30**

Chapter 4

Conclusion

4.1 Thesis Summary

In this thesis, we considered the application of community-based hub placement strategies within the emerging Advanced Air Mobility field. Central to this was the introduction of Directed Flow Communities, a methodology designed to incorporate the geographic positioning and flow of directed, weighted transportation networks. Through an empirical investigation utilizing the Freight Analysis Framework dataset, this research showed the potential of Directed Flow Communities and other community detection methods in unveiling strategic hub locations.

The case study provided was a preliminary examination of the DFC method's applicability to real-world data, presenting a promising avenue for optimizing AAM infrastructure. While the results from the case study are encouraging, suggesting the method's potential in identifying cost-efficient hub configurations, this represents an initial step toward validation. The process demonstrated the method's conceptual viability and its alignment with real-world network structures, but a more rigorous validation is essential to confirm its effectiveness comprehensively.

4.2 Future Work

Based on the results of this thesis, three directions for future research are identified:

- **Capacitated p-Hubs Problem:** The current research primarily addresses the uncapacitated p-hubs problem for simplicity. Future investigations could delve into the capacitated variant, incorporating limitations on hub capacities to mirror real-world constraints more accurately.

- **Holistic Evaluation of Hub Placement Impact:** Beyond assessing the cost implications on the network, it's crucial to understand the broader socio-economic effects of hub placements. Future studies could integrate demographic and socio-economic factors within the Directed Flow Communities framework. This approach would allow for strategic hub and community placements in underserved areas, facilitating an in-depth analysis of potential improvements in local living conditions.

- **Expansion of Directed Flow Communities Validation Testing:** The initial testing of the Directed Flow Communities methodology has shown promise. To solidify its validity and applicability, further testing on more extensive datasets, ranging from 10,000 to over 100,000 nodes, is recommended. This would provide a more comprehensive understanding of the methodology's scalability and performance in varied network sizes and complexities. Additionally, more empirical testing can also aid in validating this method.

Appendix A

Directed Flow Communities Code

Code for the Directed Flow Communities can be found on my github repository.

A.1 Directed Flow Communities Psuedo Code

```
// Initialize network with nodes and edges
INITALIZE network_nodes and network_edges

// Create a variable to save layer properties to
CREATE layers

// Create layer for each source node
FOR source_node in source_nodes
    // Set up simulation parameters and initial conditions
    INITIALIZE simulation_parameters
    SET source_node

    // Define simulation duration
    FOR timestep IN range(1, T+1)

        // Update edge concentrations
```

```
        FOR edge IN network_edges
            edge.update_step(edge.incoming_mass)  // Updates edge
               values
            SET edge.outgoing_mass // Set outgoing mass for edge

        // Update node values based on the incoming mass from edges
        FOR node IN network_nodes
            total_incoming_mass = 0
            FOR edge IN node.incoming_edges  // Sum up the outgoing
               mass from all incoming edges to this node
                total_incoming_mass += edge.outgoing_mass

            // Update the node concentration
            node.concentration = node.update_mass(total_incoming_mass
               )

            // Propagate the remaining node mass to outgoing edges
            FOR edge IN node.outgoing_edges
                // Divide the node output mass evenly among all
                   outgoing edges
                // NUM(node.outgoing_edges) calculates the number of
                   outgoing edges from the node
                edge.incoming_value = node.output_mass / NUM(
                   node.outgoing_edges)

        APPEND [node.concentration for node in network_nodes] TO
           layers

RUN kmeans ON layers_saves
RETURN community_assigments
```

Appendix B

Parameters for Experiments

Tables B.1, B.2 and B.4 show parameters used by Directed Flow Communities in benchmark graphs in Section 2.4.

Table B.1: 500 Node LFR Benchmark Parameters

μ	T	w	ℓ	α_r	α_ℓ	λ	no. source nodes
0.2	50	edge weight	1	1	0.4	0	200
0.3	50	edge weight	1	1	0.4	0	200
0.4	50	edge weight	1	1	0.4	0	200
0.5	25	edge weight	1	1	0.4	0	250
0.6	50	edge weight	1	1	0.4	0	250
0.7	50	edge weight	1	1	0.4	0	350
0.8	50	edge weight	1	1	0.4	0	350
0.9	50	edge weight	1	1	0.4	0	350

Table B.2: 1000 Node LFR Benchmark Parameters

μ	T	w	ℓ	α_r	α_ℓ	λ	no. source nodes
0.2	50	edge weight	1	1	0.4	0	250
0.3	50	edge weight	1	1	0.4	0	250
0.4	50	edge weight	1	1	0.4	0	250
0.5	50	edge weight	1	1	0.4	0	250
0.6	50	edge weight	1	1	0.25	0	250
0.7	50	edge weight	1	0.3	0.1	0	400
0.8	50	edge weight	1	0.5	0.1	0	700
0.9	50	edge weight	1	0.5	0.1	0	700

Table B.3: Empirical Datasets Parameters

network	T	w	ℓ	α_r	α_ℓ	λ	no. source nodes
Cora	50	cos. sim.	1	1	0.001	0	892
Eu-Core	1000	1	1	0.5	0.001	0	750

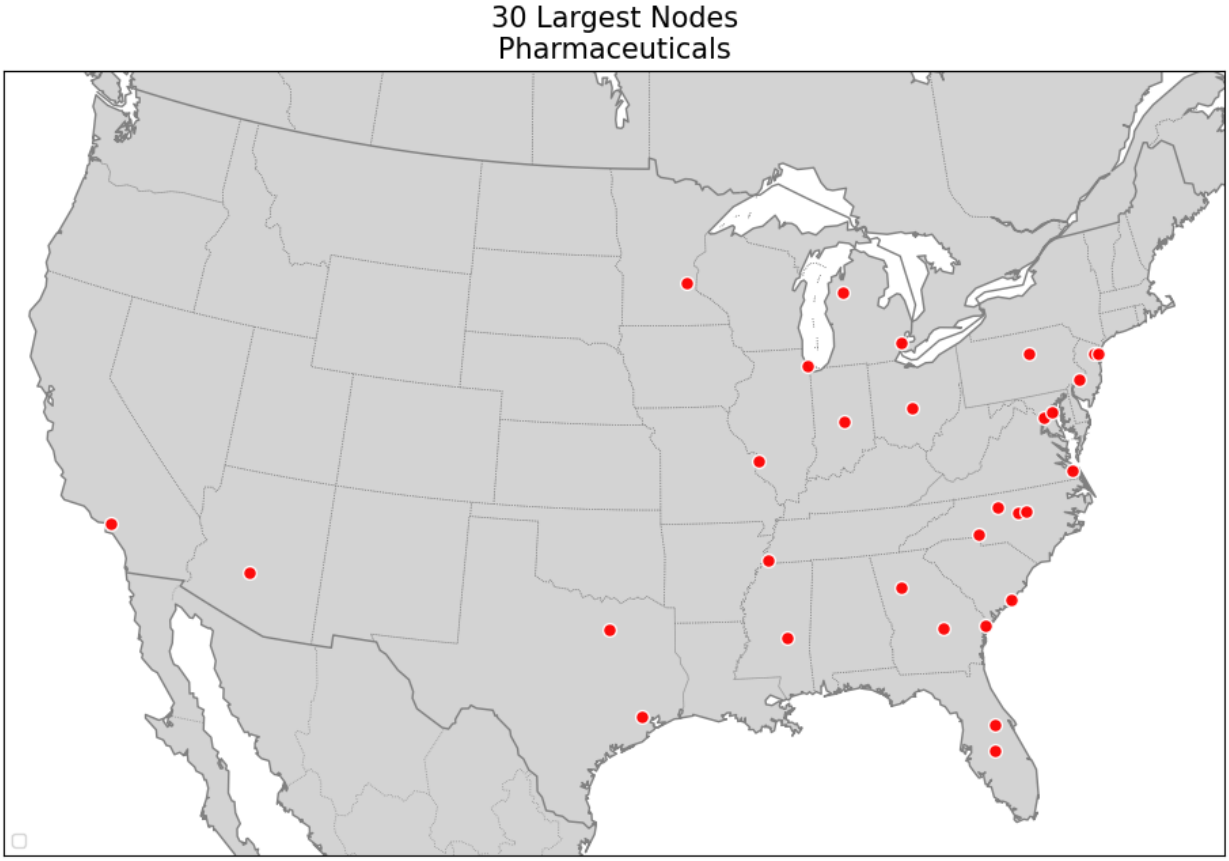
Table B.4: FAF Datasets Parameters

network	T	w	ℓ	α_r	α_ℓ	λ	no. source nodes
Pharmaceuticals	100	mil-tons	haversine distance	1	1e-10	2	10
Electronics	100	mil-tons	haversine distance	1	1e-10	2	10
Total Freight Flows	50	mil-tons	haversine distance	1	1e-5	2	50

Appendix C

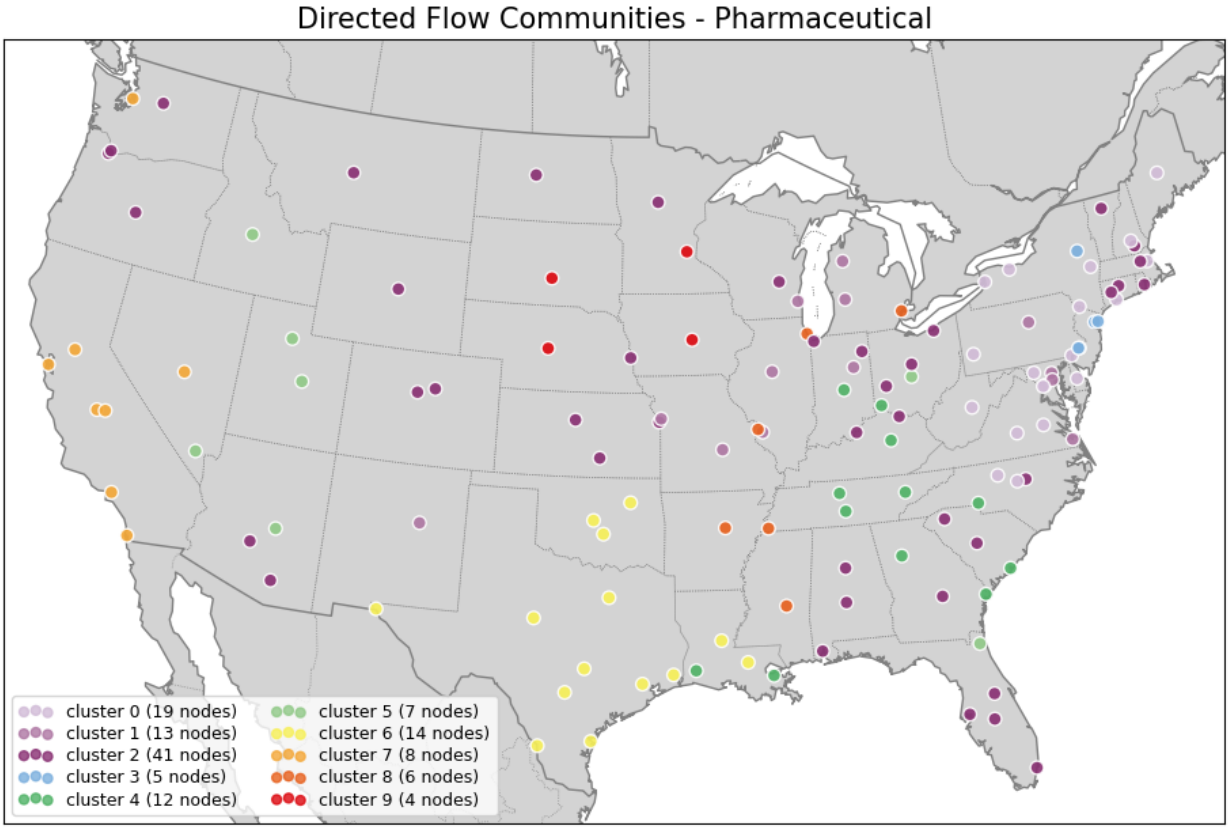
FAF Case Study - Community and Hub Figures

This appendix complements Chapter 3 by providing figures of the community structures and hub configurations that emerged from the Freight Analysis Framework (FAF) case study.



(a) 30 Largest Nodes on the FAF Pharmaceutical dataset.

Figure C.1: 30 Largest Nodes identified for the FAF Pharmaceutical dataset using the Integer programming method.

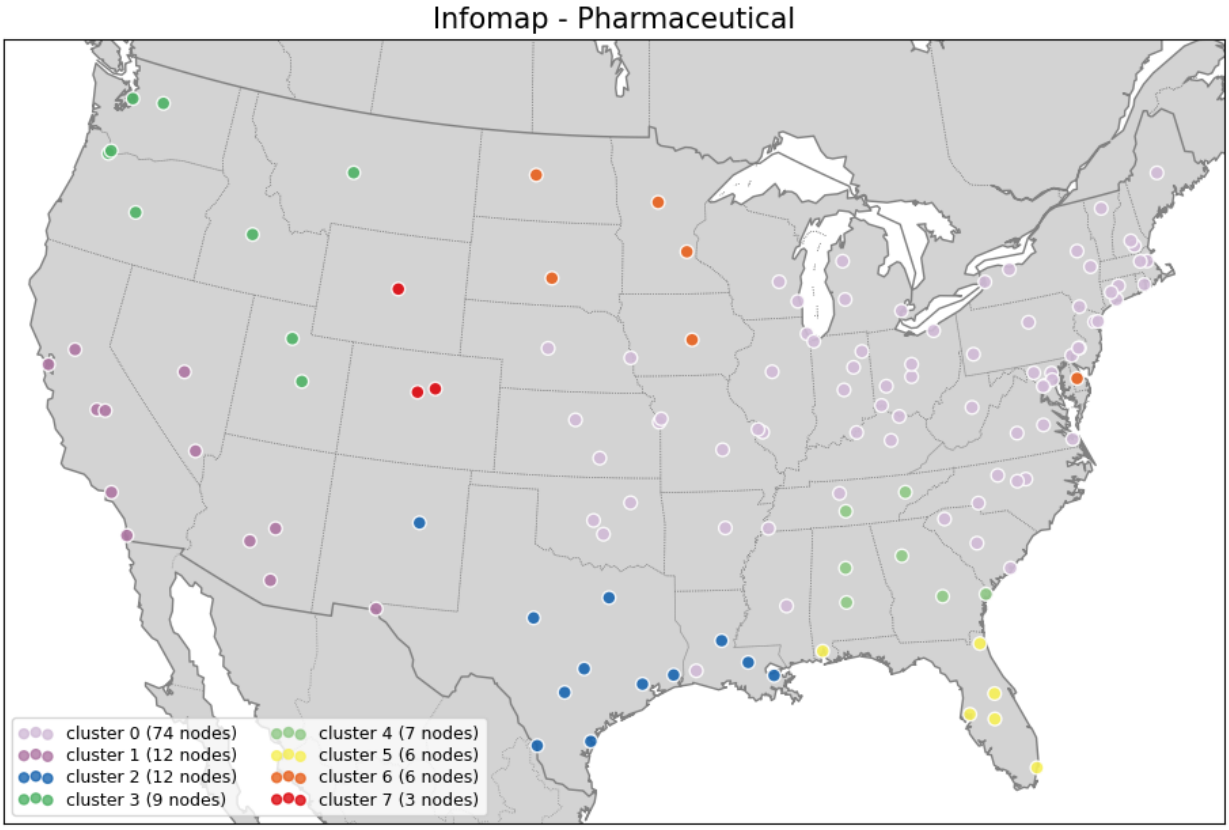


(a) Directed Flow Communities on the FAF Pharmaceutical dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.341	395.870
1	0.409	276.879
2	0.333	974.326
3	0.369	396.826
4	0.332	1,349.861
5	0.324	15.376
6	0.307	762.337
7	0.302	316.444
8	0.312	318.321
9	0.250	61.974

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of pharmaceutical traffic for each community.

Figure C.2: Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community

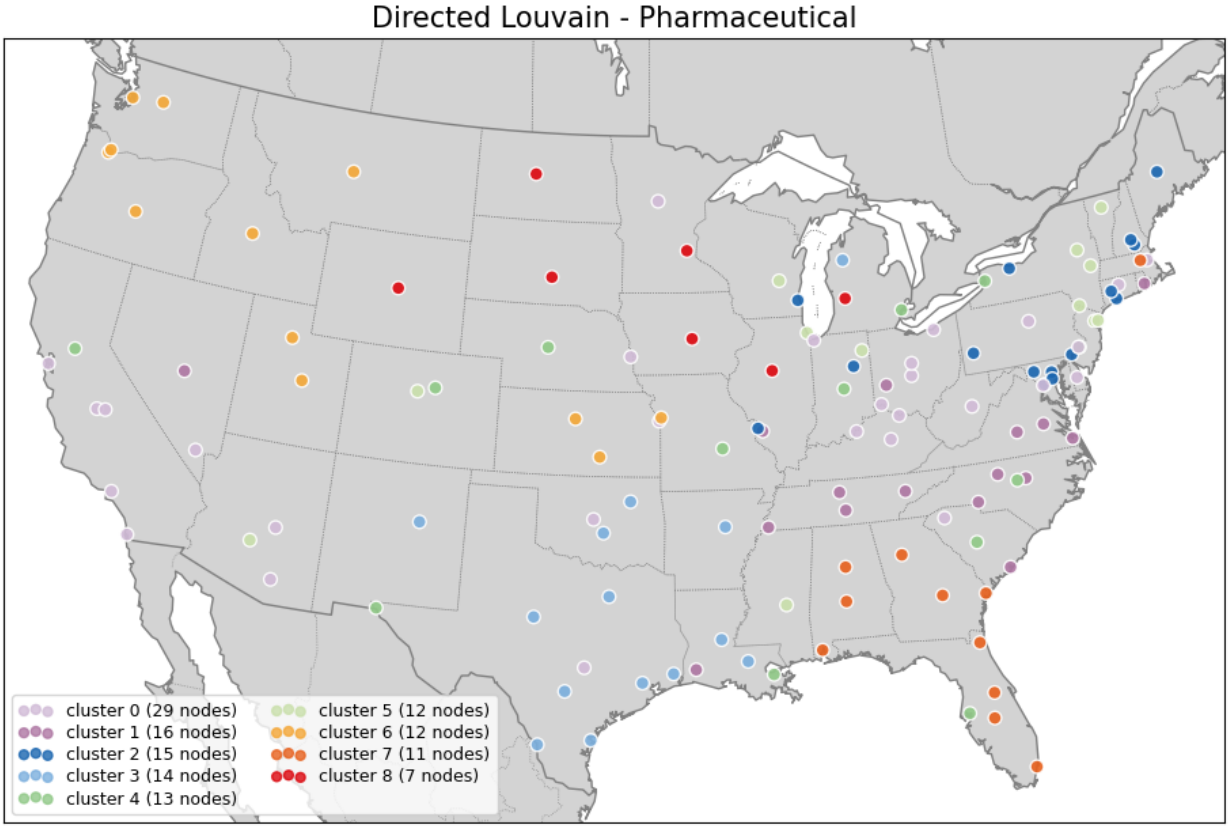


(a) Infomap Communities on the FAF Pharmaceutical dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.207	8,202.85
1	0.323	449.161
2	0.269	746.362
3	0.258	193.983
4	0.271	1,098.183
5	0.212	318.007
6	0.347	188.425
7	0.132	24.938

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of pharmaceutical traffic for each community.

Figure C.3: A visual and quantitative analysis of FAF pharmaceutical flow within the United States using Infomap.

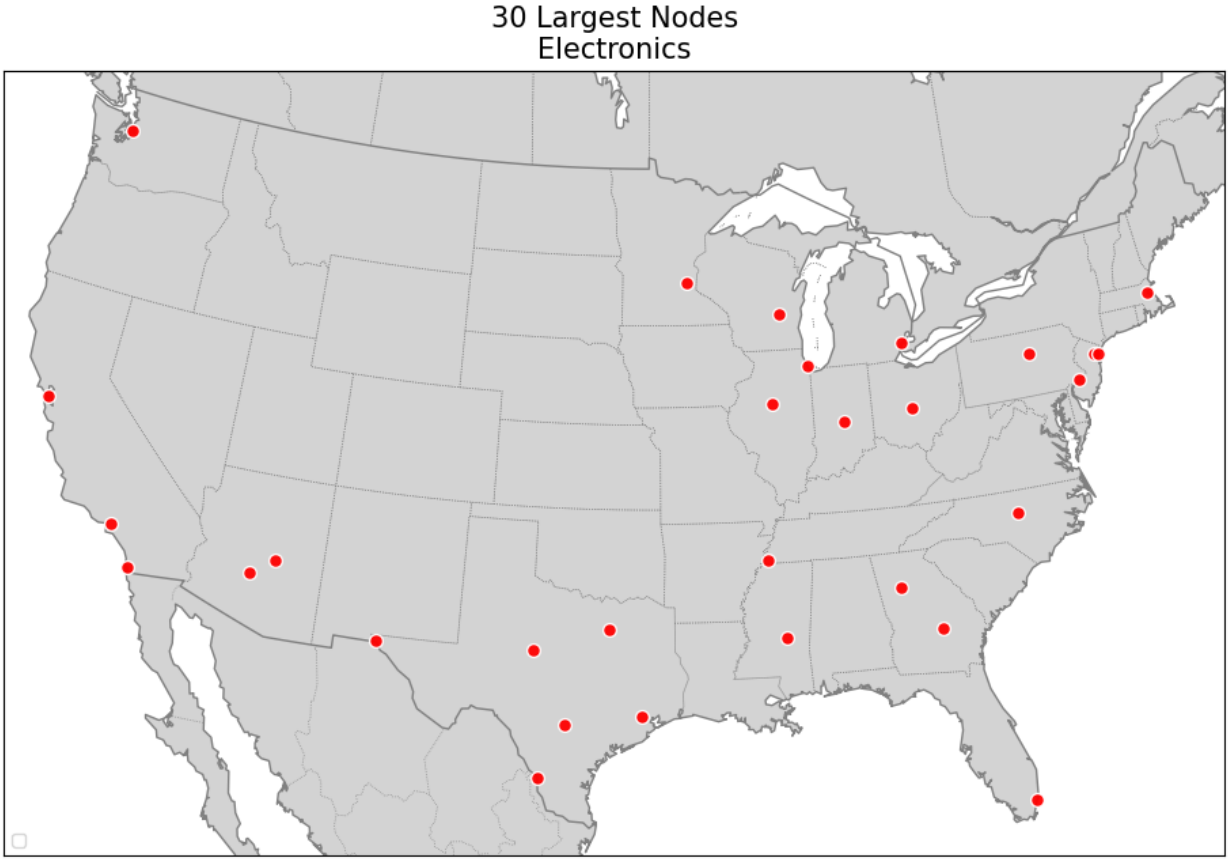


(a) Directed Louvain Communities on the FAF Pharmaceutical dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.357	1,282.182
1	0.335	1,035.36
2	0.345	383.36
3	0.233	713.16
4	0.364	580.731
5	0.374	937.205
6	0.341	230.723
7	0.247	1,411.989
8	0.416	98.551

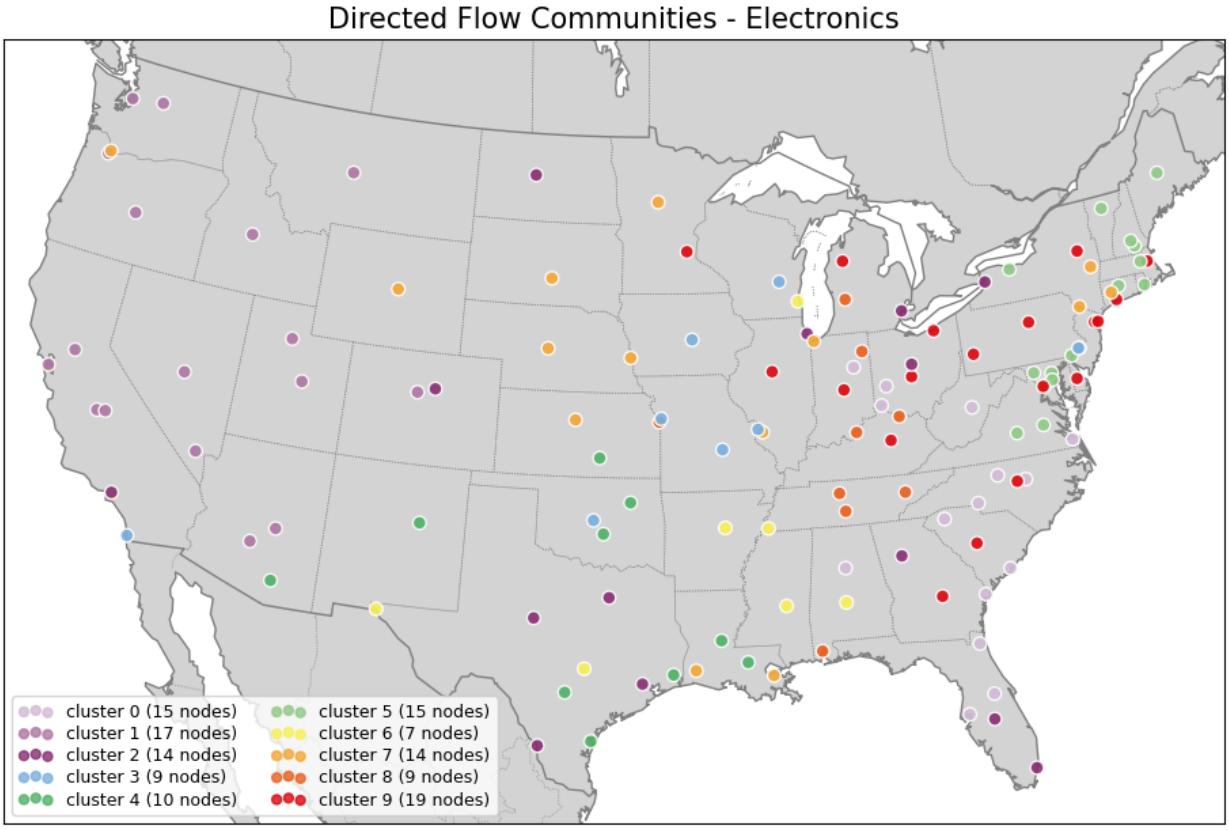
(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of pharmaceutical traffic for each community.

Figure C.4: A visual and quantitative analysis of FAF pharmaceutical flow within the United States using Directed Louvain.



(a) 30 Largest Nodes on the FAF Electronics dataset.

Figure C.5: 30 Largest Nodes identified for the FAF Electronics dataset using the Integer programming method.

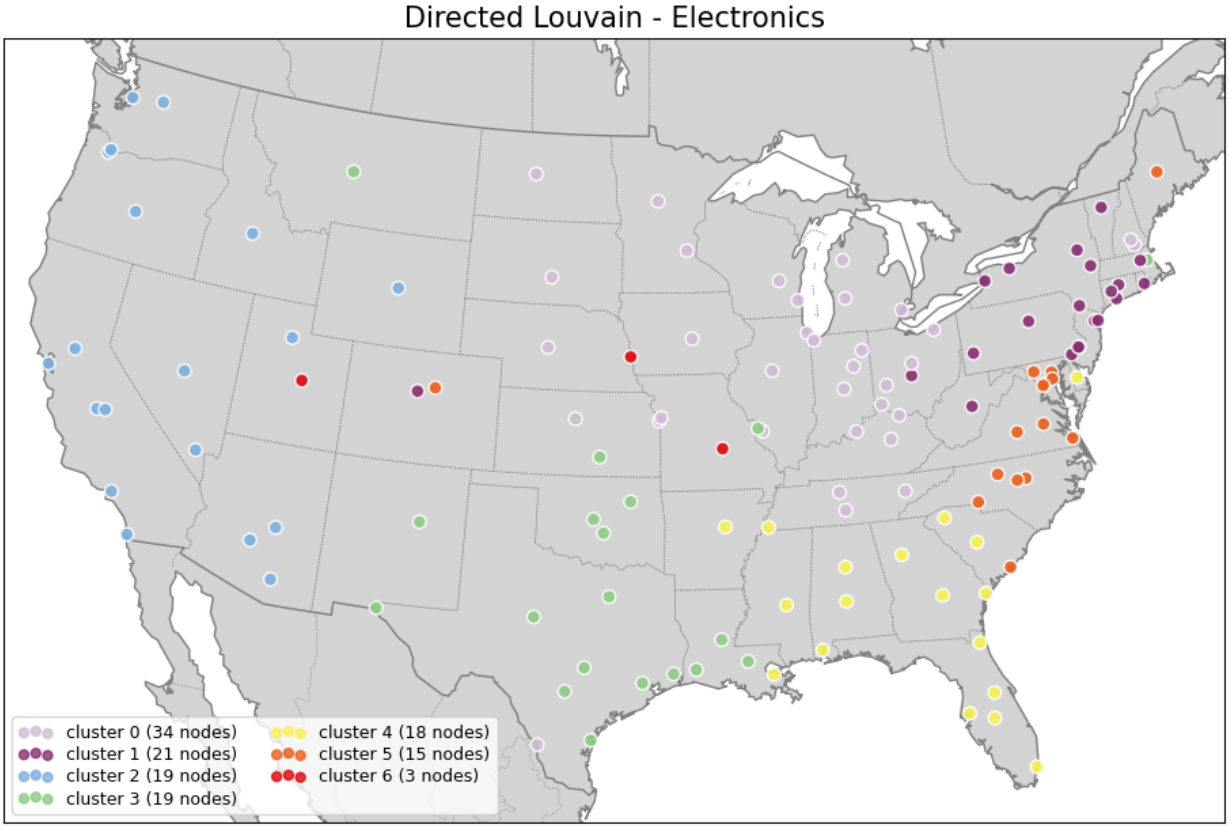


(a) Directed Flow Communities on the FAF Electronics dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.445	579.175
1	0.325	1,702.025
2	0.339	5,542.85
3	0.35	322.716
4	0.426	193.65
5	0.422	508.258
6	0.472	333.488
7	0.408	57.926
8	0.379	139.2922
9	0.309	3,161.221

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community.

Figure C.6: A visual and quantitative analysis of FAF electronics flow within the United States using Directed Flow Communities.

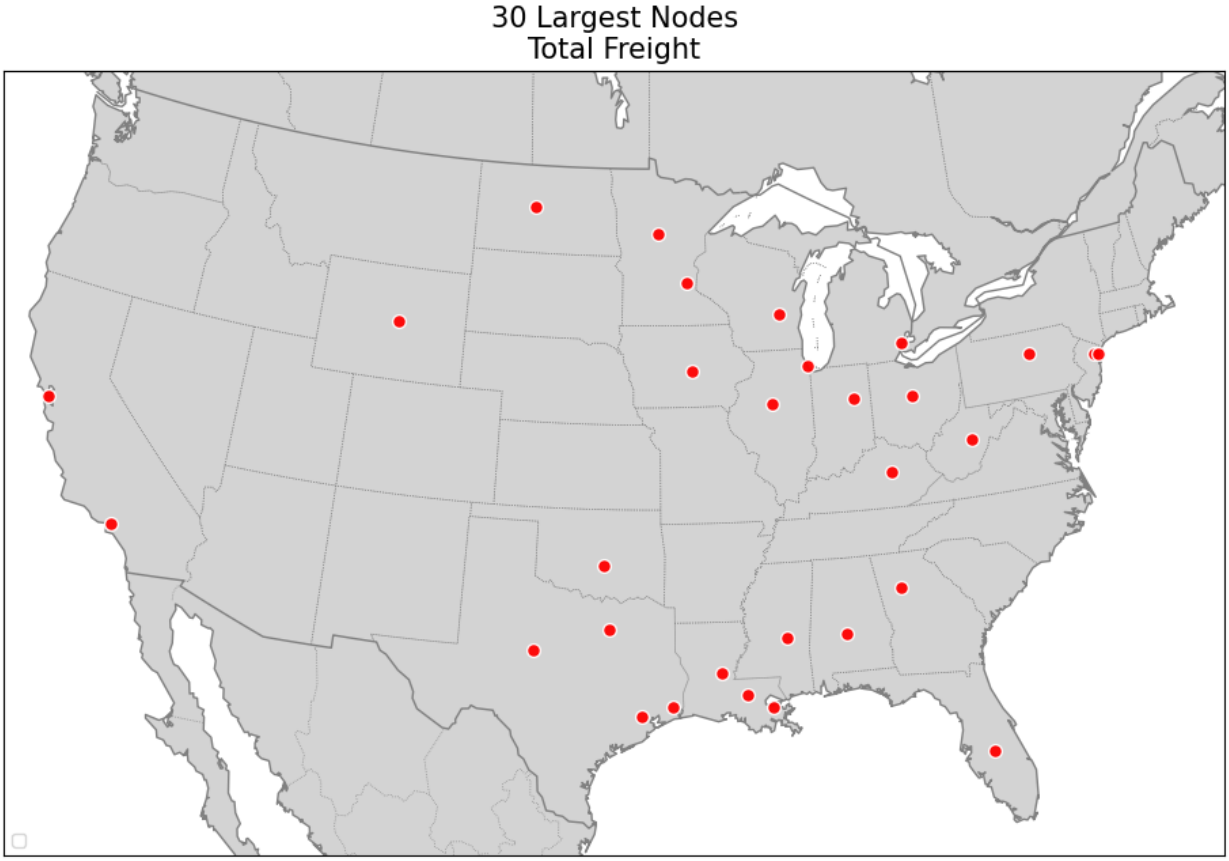


(a) Directed Louvain Communities on the FAF Electronics dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.306	6,143.848
1	0.303	3,836.163
2	0.191	7,367.732
3	0.301	5,254.402
4	0.335	3,167.129
5	0.354	1,302.54
6	0.327	92.628

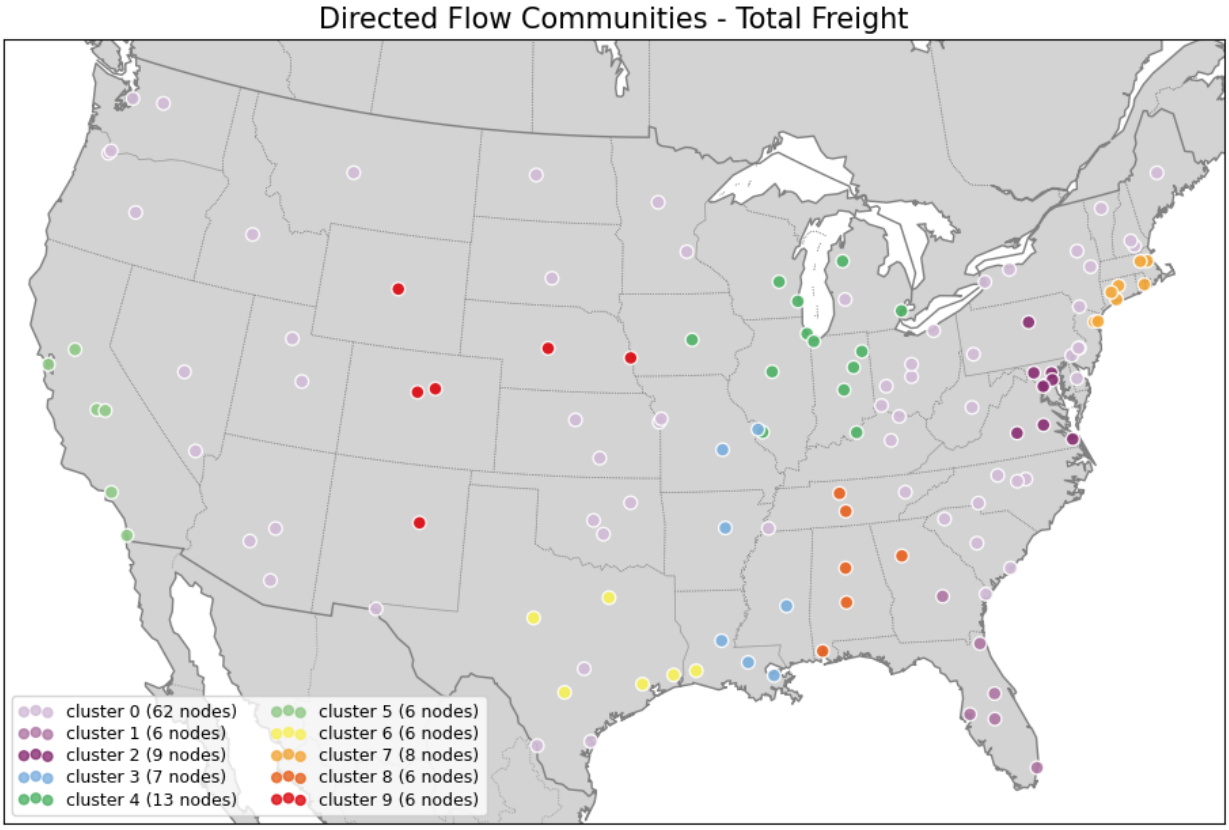
(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community.

Figure C.7: A visual and quantitative analysis of FAF electronics flow within the United States using Directed Louvain.



(a) 30 Largest Nodes on the FAF Total Freight dataset.

Figure C.8: 30 Largest Nodes identified for the FAF Total Freight dataset using the Integer programming method.

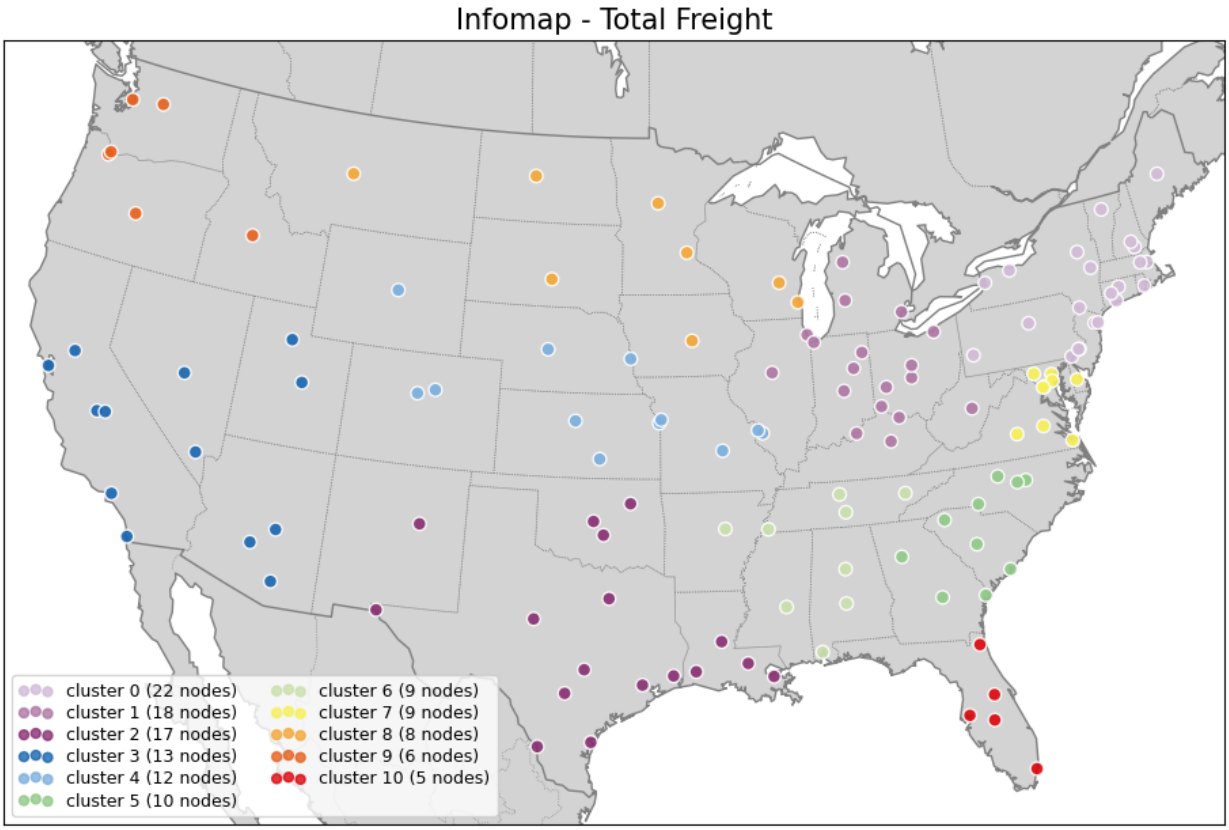


(a) Directed Flow Communities on the FAF total freight flows dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.303	1,682,746.760
1	0.227	206,689.788
2	0.358	188,753.358
3	0.312	427,118.404
4	0.337	644,370.413
5	0.210	299,952.080
6	0.326	559,023.207
7	0.274	139,620.501
8	0.357	164,043.276
9	0.186	179,505.781

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community.

Figure C.9: A visual and quantitative analysis of FAF total freight flow within the United States using Directed Flow Communities.

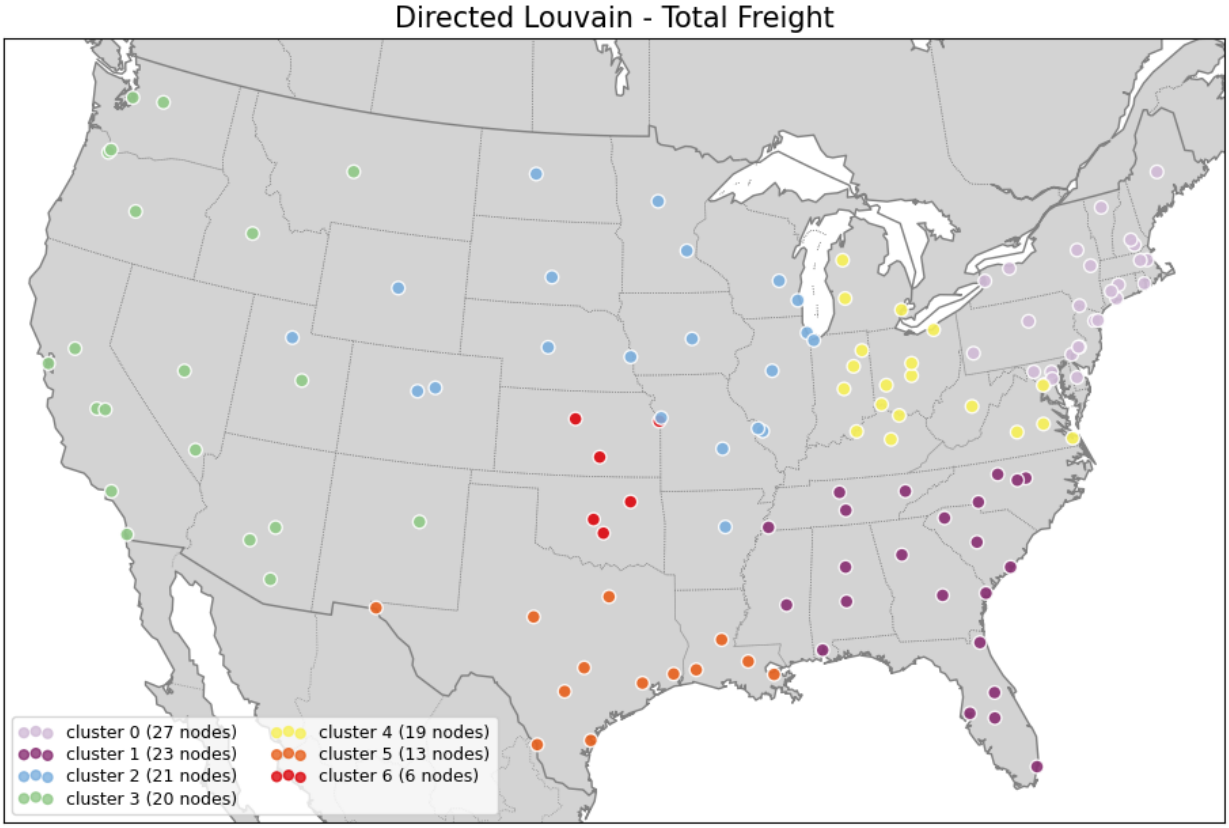


(a) Infomap Communities on the FAF total freight flows dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.217	707,041.959
1	0.298	970,552.974
2	0.206	1,516,247.111
3	0.173	474,861.215
4	0.235	400,181.107
5	0.252	257,485.792
6	0.333	326,198.530
7	0.234	47,733.031
8	0.211	427,802.0574
9	0.200	230,297.327
10	0.149	182,103.502

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community.

Figure C.10: A visual and quantitative analysis of FAF total freight flow within the United States using Infomap.



(a) Directed Louvain Communities on the FAF total freight flows dataset.

Cluster No.	Conductance ↓	Freight Flows
0	0.195	819,612.518
1	0.171	967,431.183
2	0.193	1,293,284.410
3	0.191	816,146.275
4	0.291	805,963.578
5	0.208	1,295,554.829
6	0.366	126,643.644

(b) Connectivity and freight flow (in thousand-tons) figures representing the volume of electronics traffic for each community.

Figure C.11: A visual and quantitative analysis of FAF total freight flow within the United States using Directed Louvain.

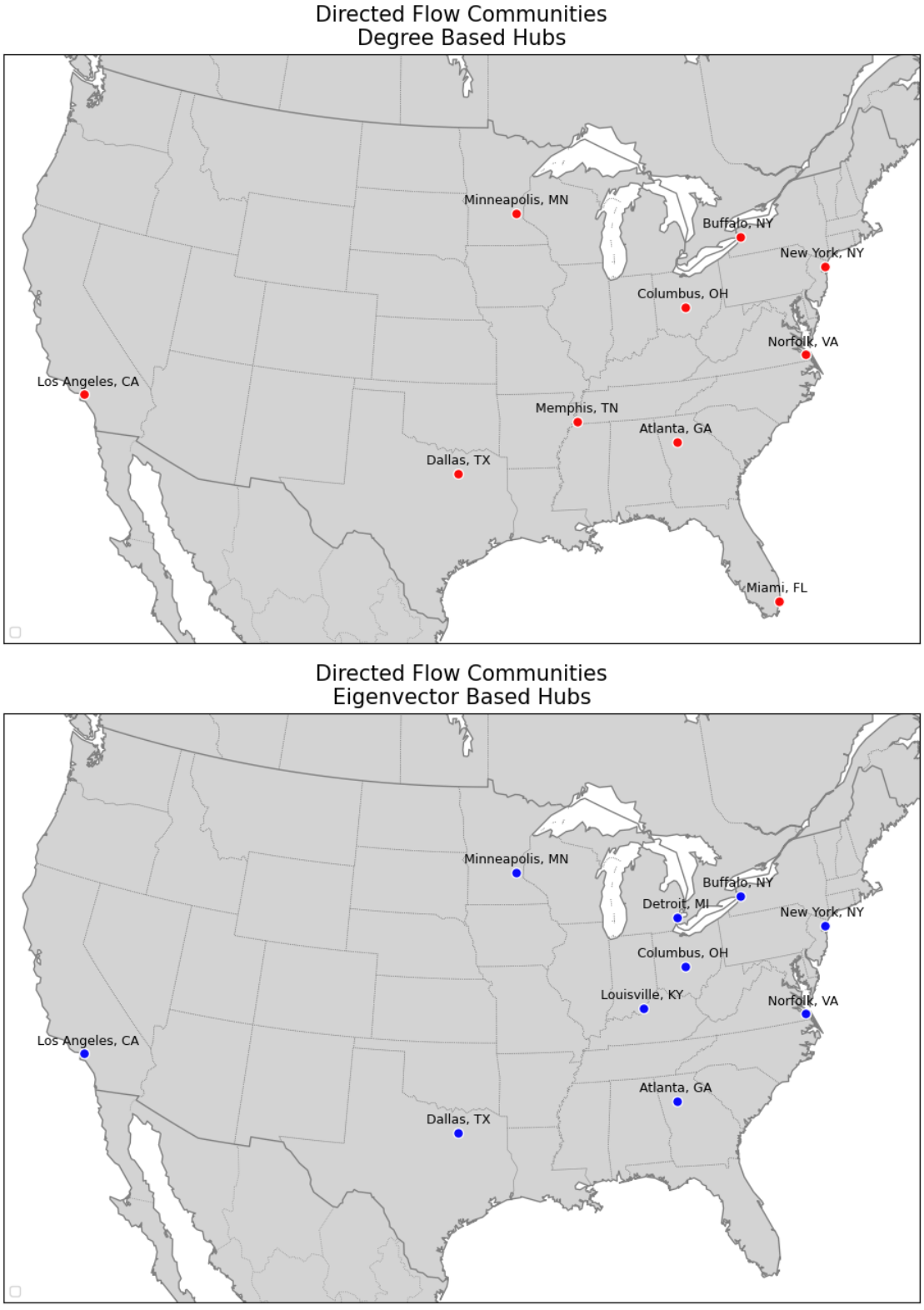


Figure C.12: Directed Flow Communities pharmaceutical derived maps comparing key transportation hubs across the United States.

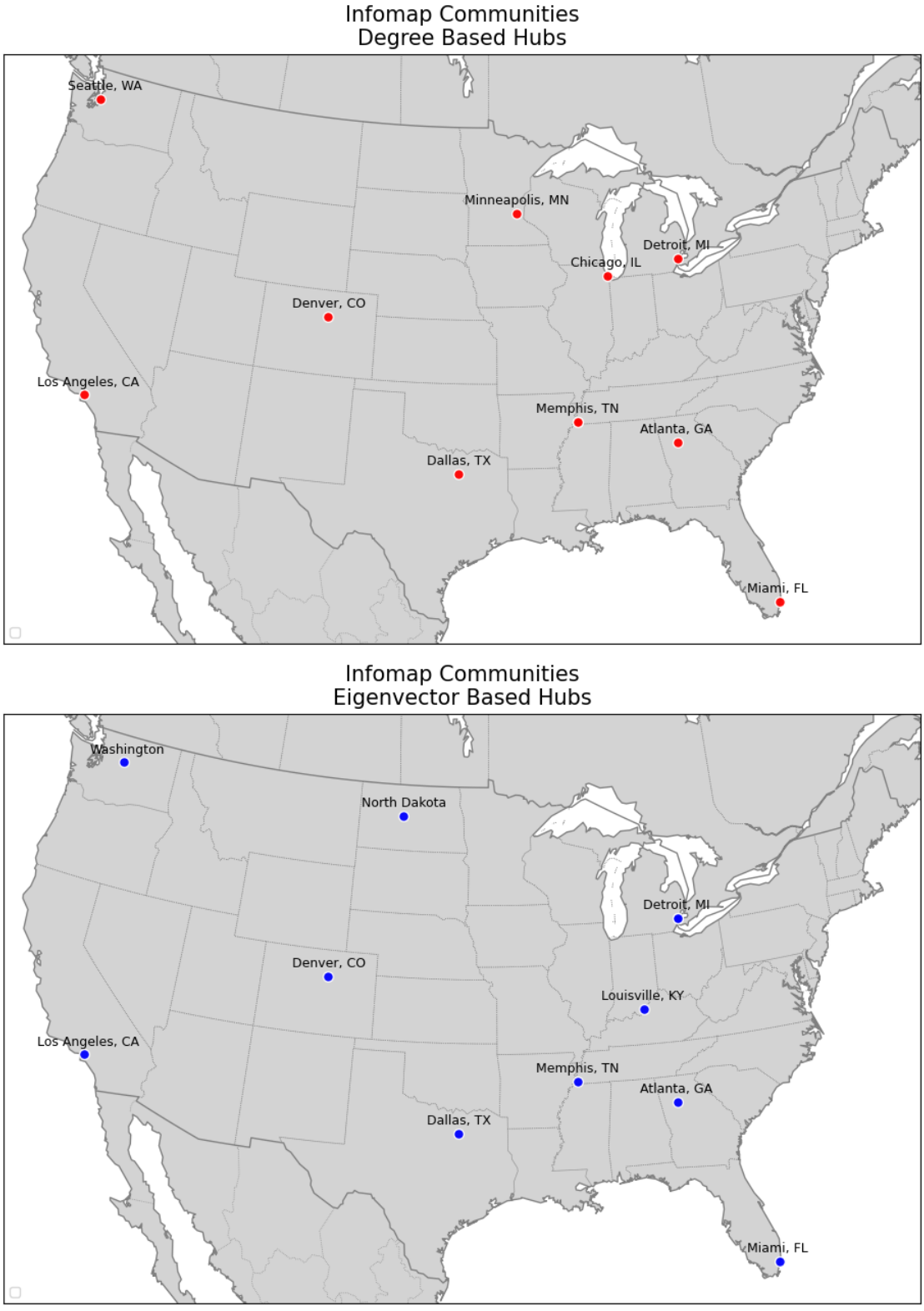


Figure C.13: Infomap pharmaceutical derived maps comparing key transportation hubs across the United States.

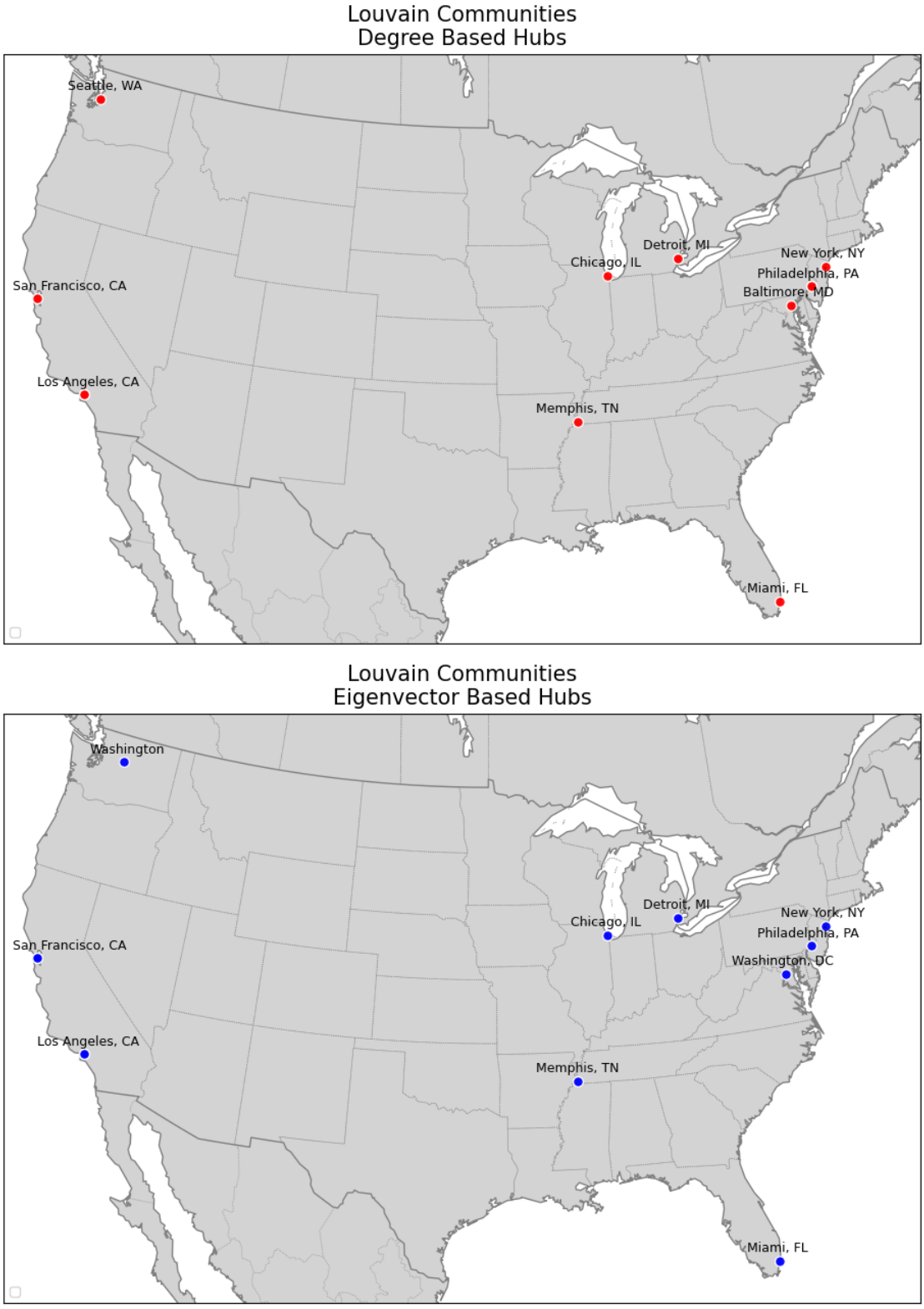


Figure C.14: Directed Louvain pharmaceutical derived maps comparing key transportation hubs across the United States.

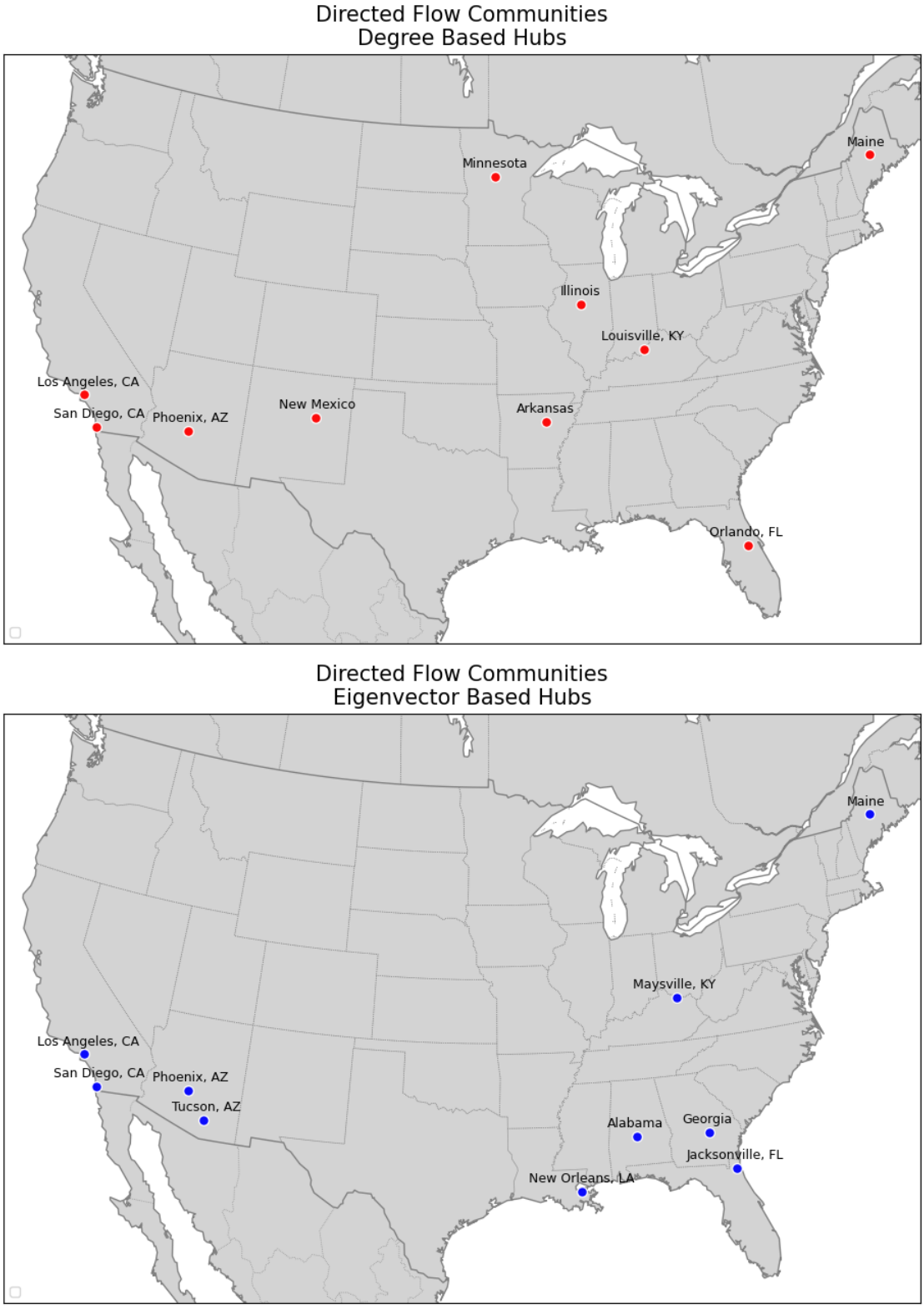


Figure C.15: Directed Flow Communities electronics derived maps comparing key transportation hubs across the United States.

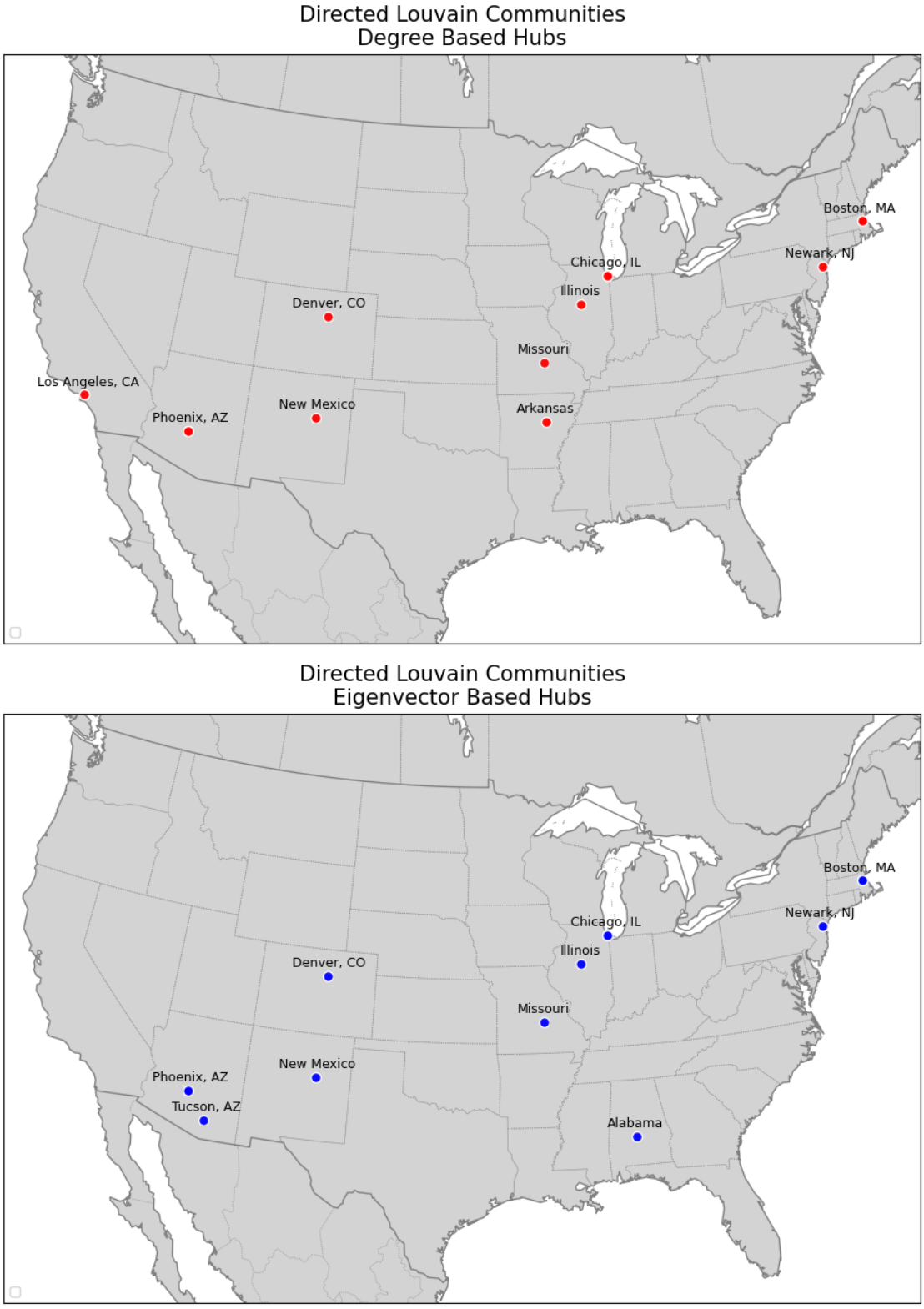


Figure C.16: Directed Louvain electronics derived maps comparing key transportation hubs across the United States.

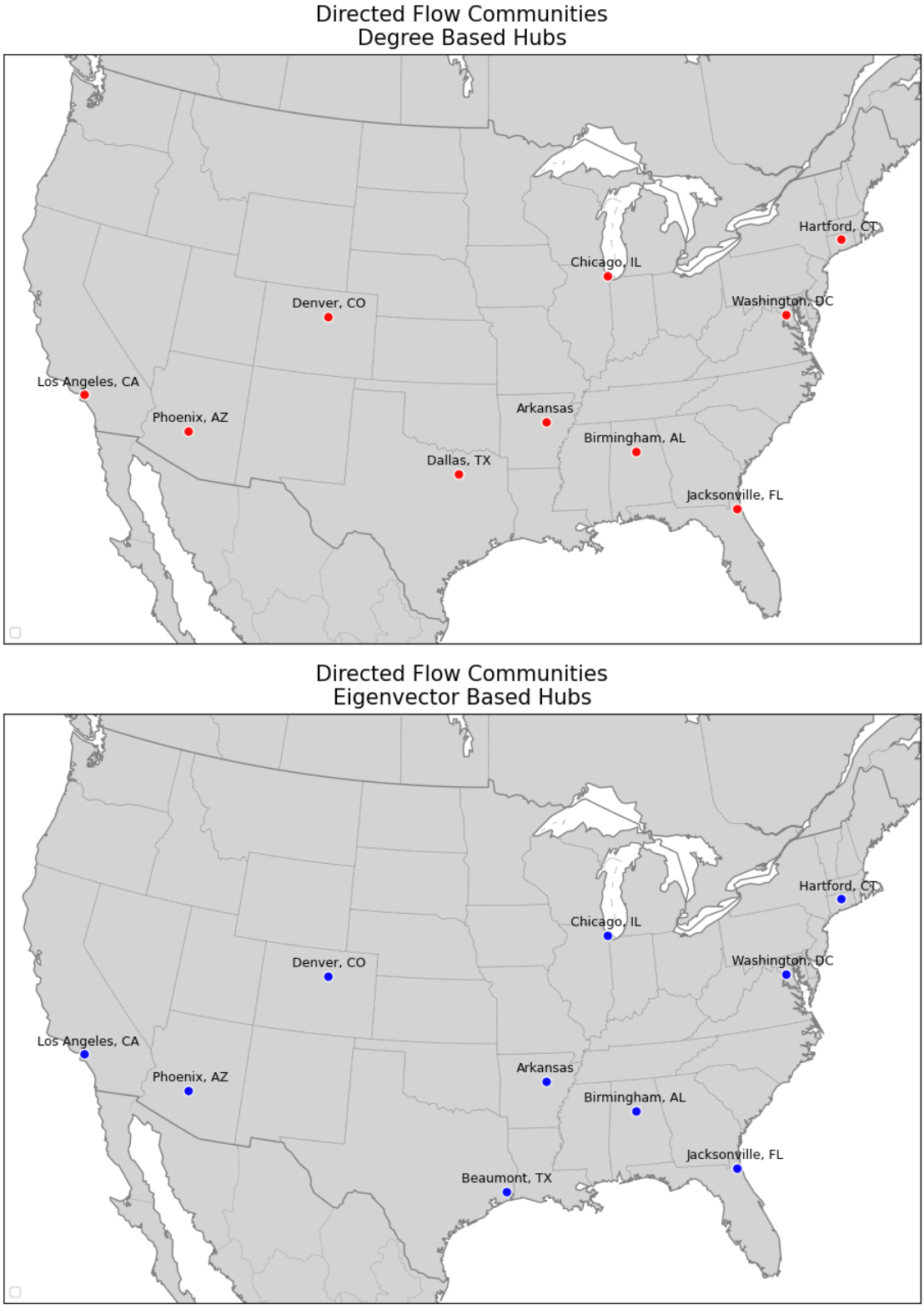


Figure C.17: Directed Flow Communities total freight flow derived maps comparing key transportation hubs across the United States.

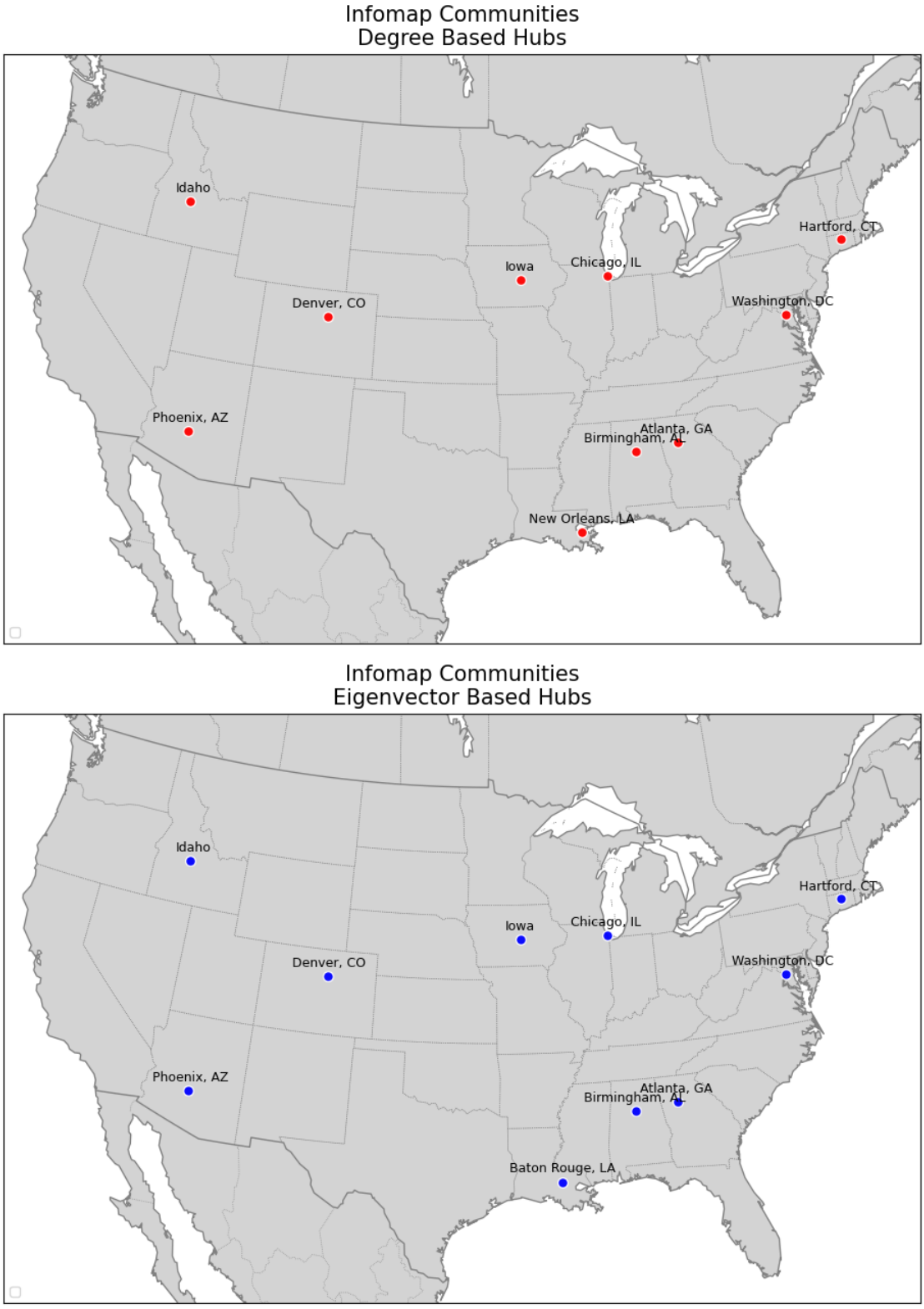


Figure C.18: Infomap total freight flow derived maps comparing key transportation hubs across the United States.

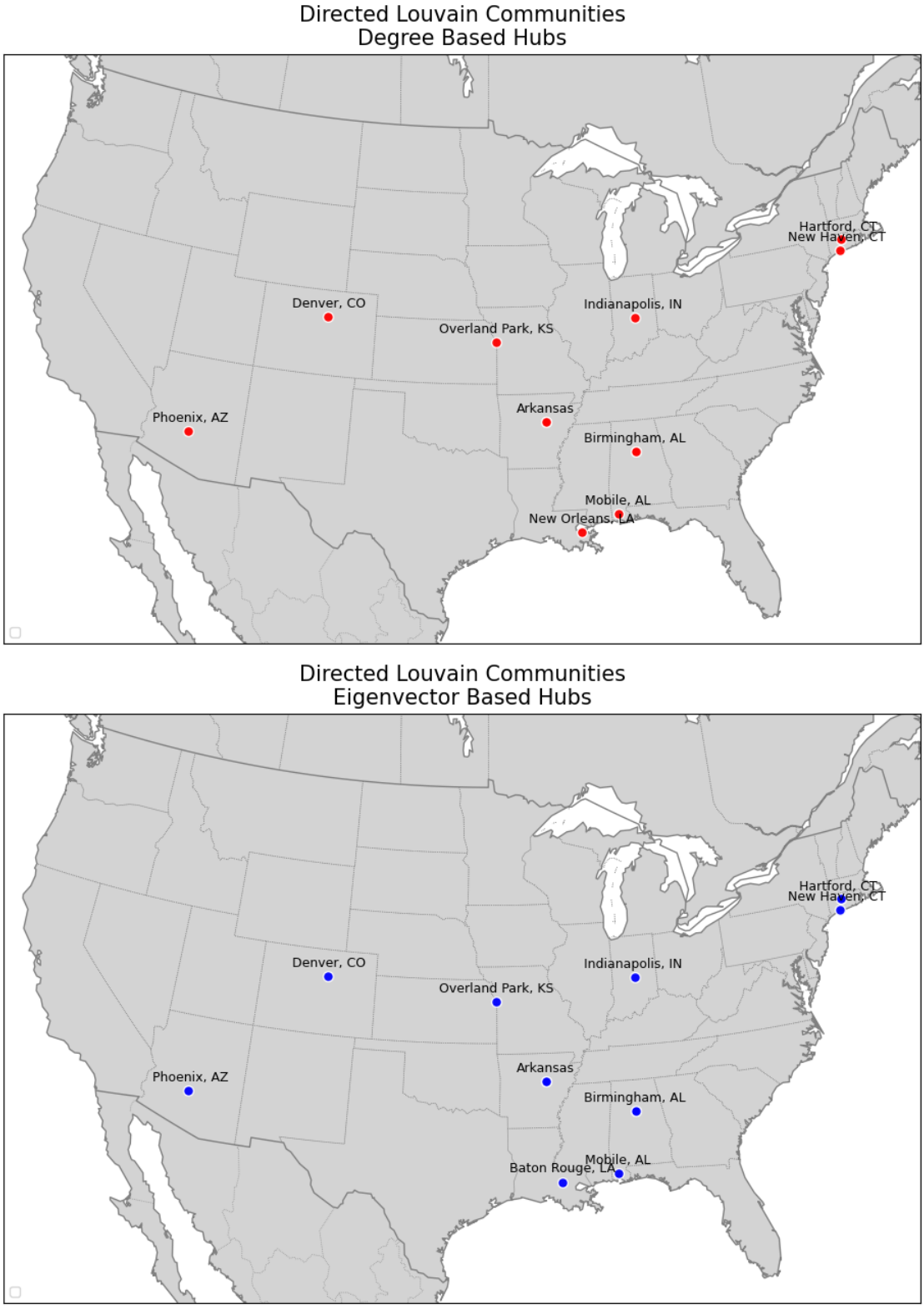


Figure C.19: Directed Louvain total freight flow derived maps comparing key transportation hubs across the United States.

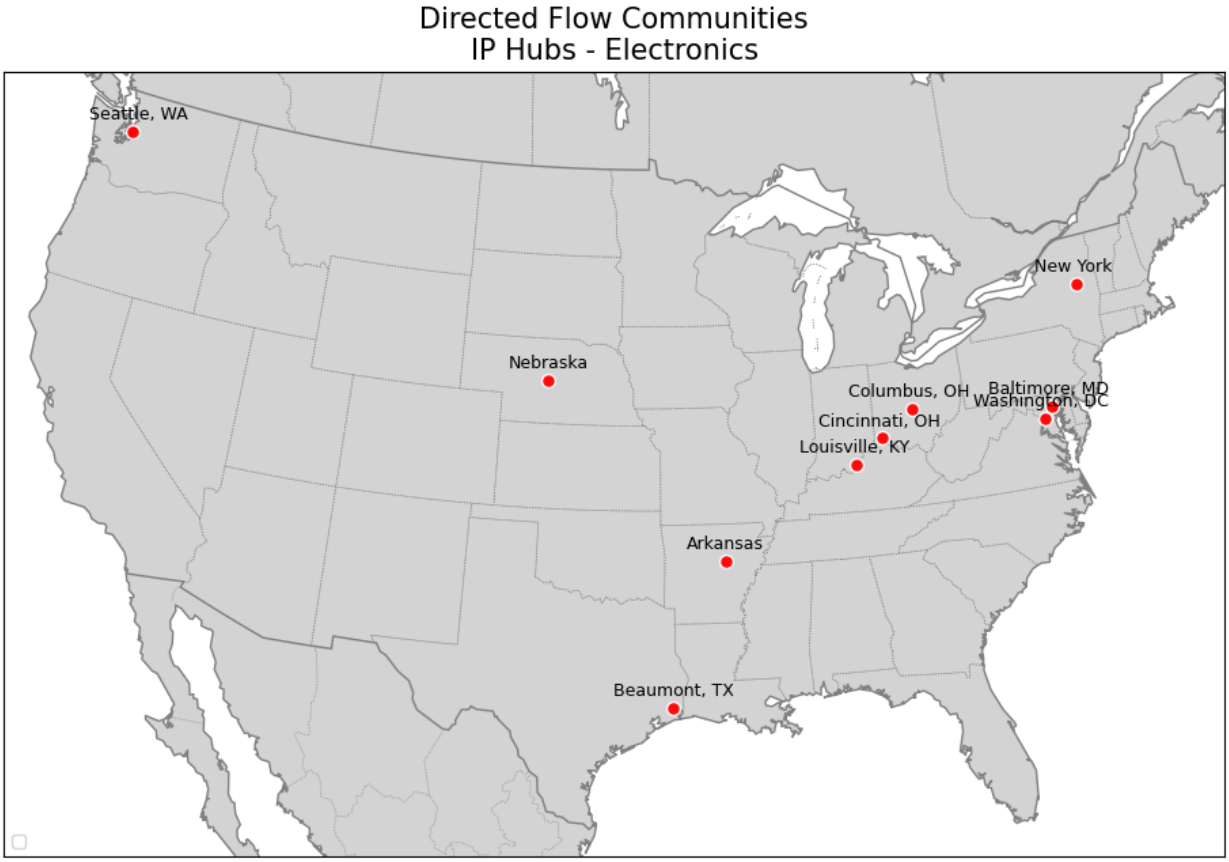


Figure C.20: Hubs identified from Directed Flow Communities on the FAF Electronics dataset using Integer programming.

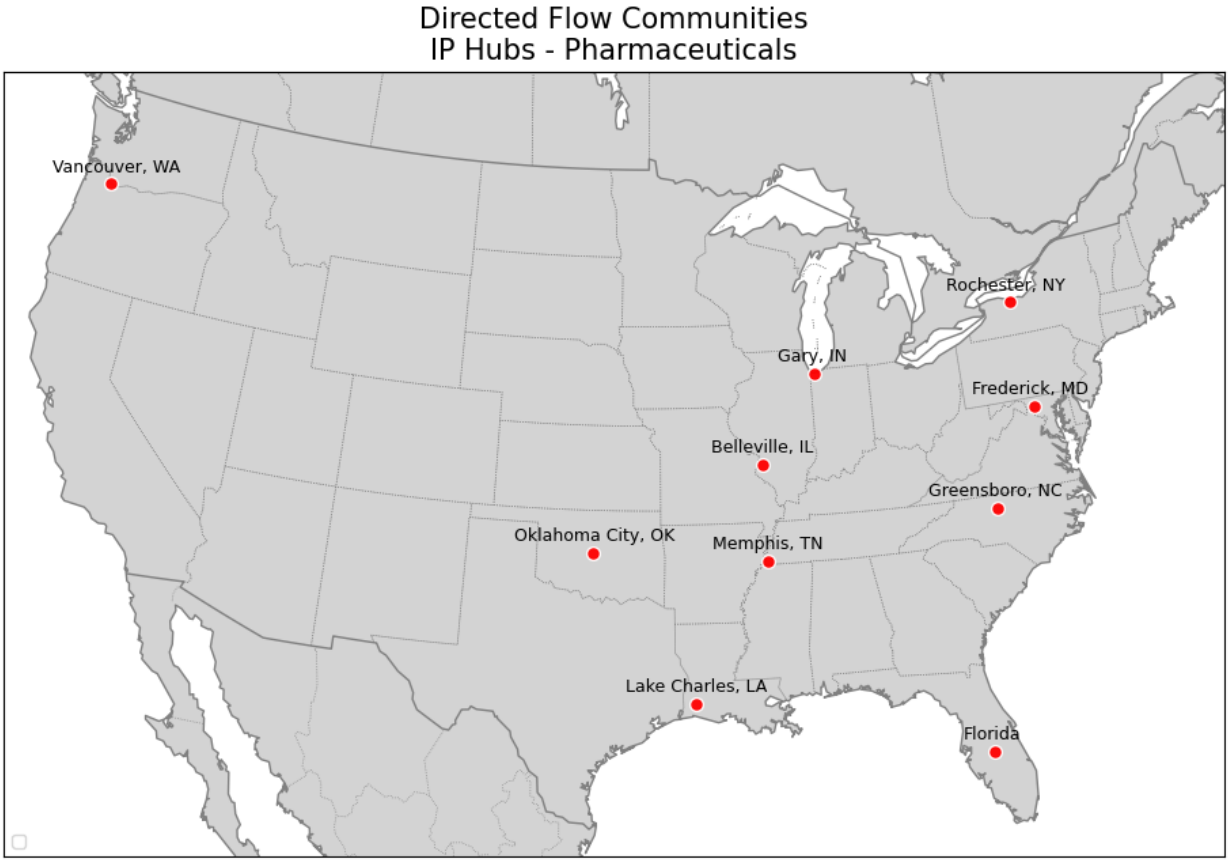
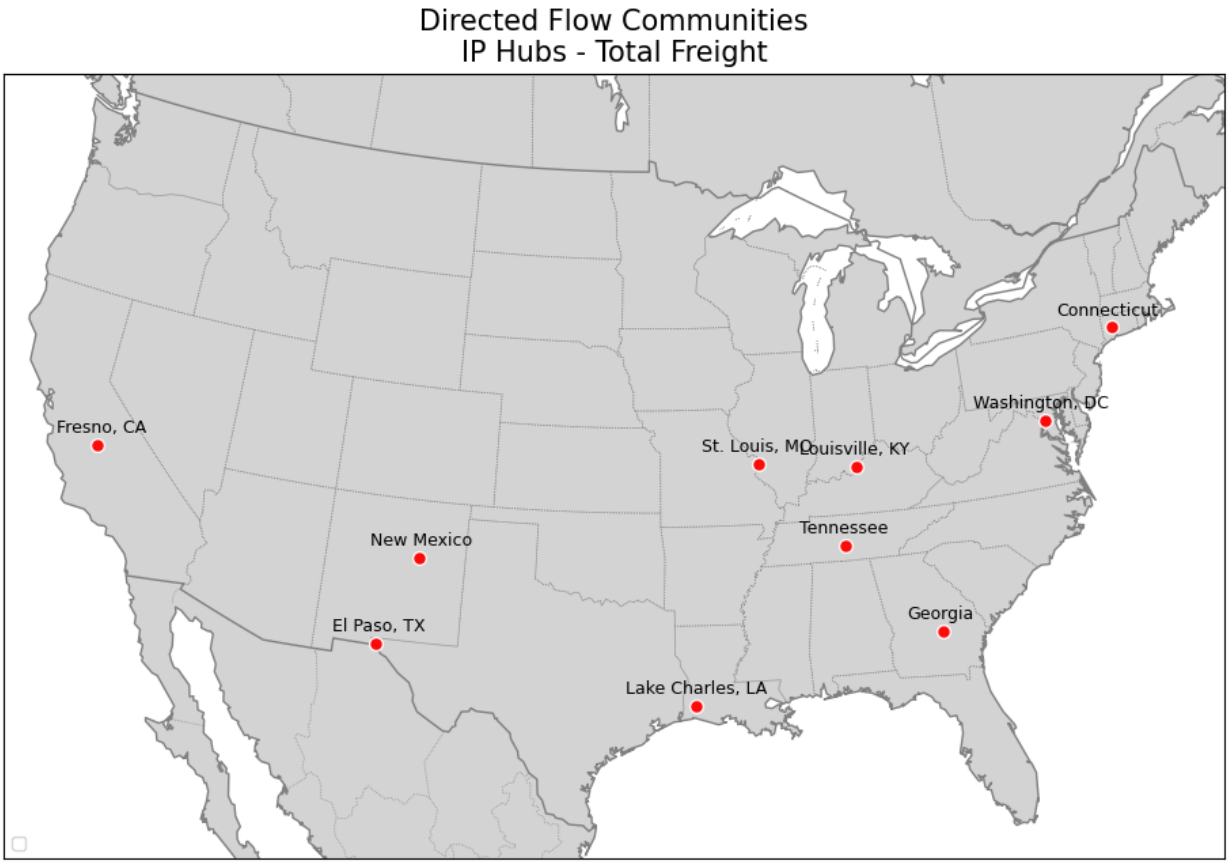


Figure C.21: Hubs identified from Directed Flow Communities on the FAF Pharmaceutical dataset using Integer programming.



(a) Directed Flow Communities hubs on the FAF Total Freight dataset using Integer programming.

Figure C.22: Hubs identified from Directed Flow Communities on the FAF Total Freight dataset using Integer programming.

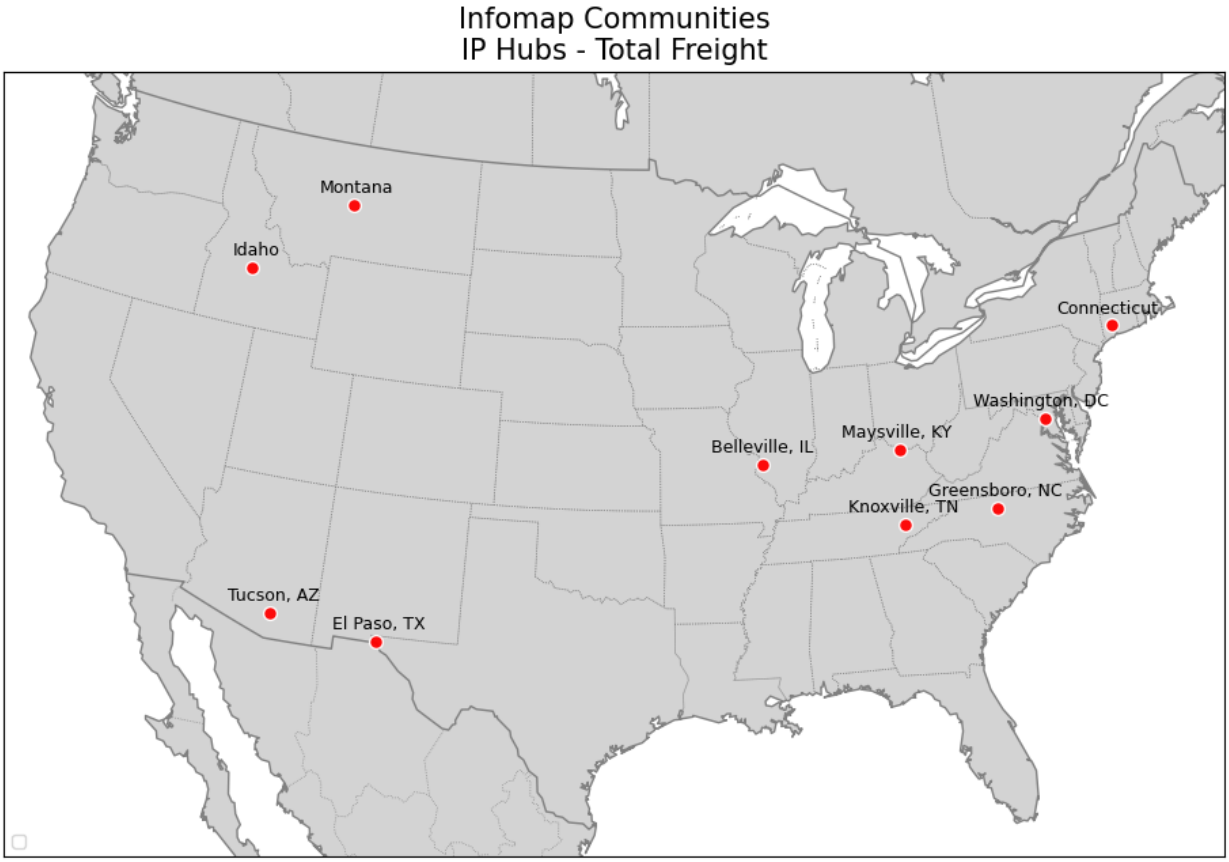


Figure C.23: Hubs identified from Infomap communities on the FAF Total Freight dataset using Integer programming.

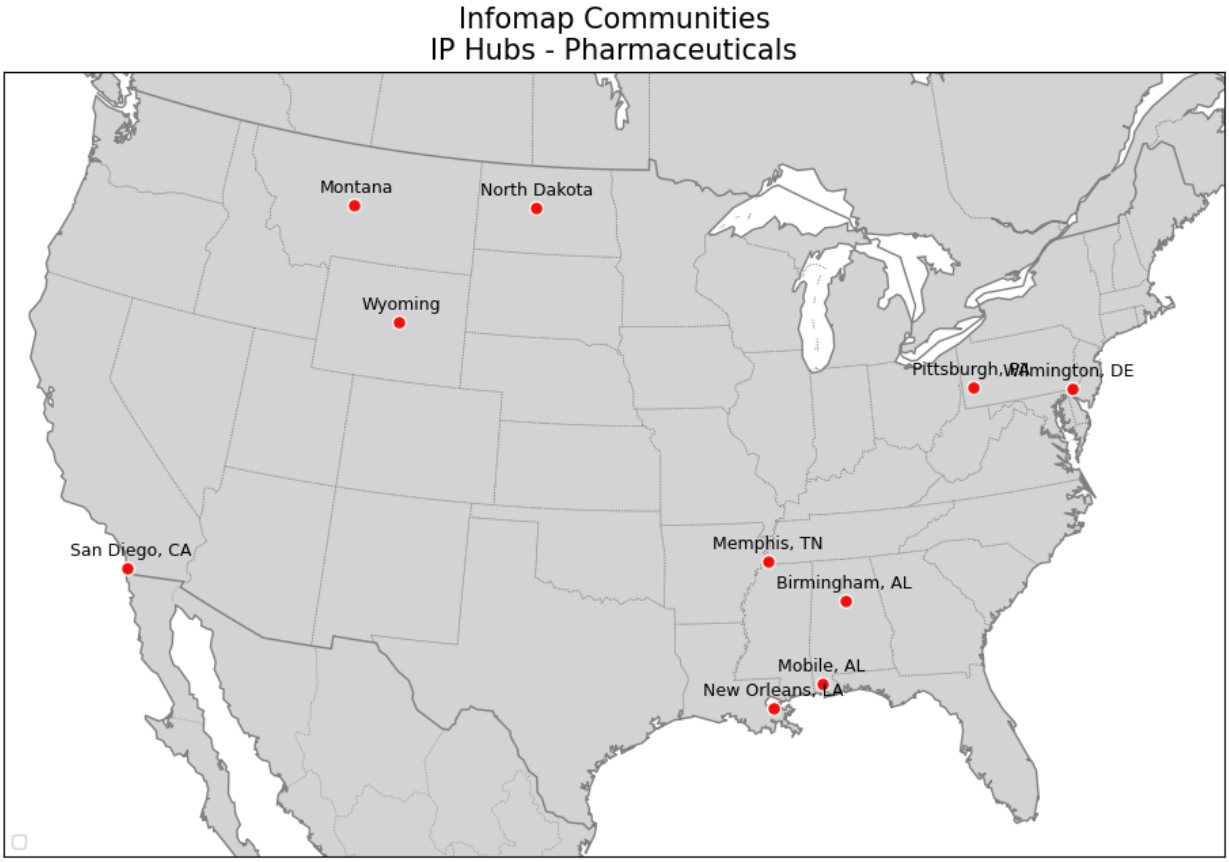


Figure C.24: Hubs identified from Infomap communities on the FAF Pharmaceutical dataset using Integer programming.

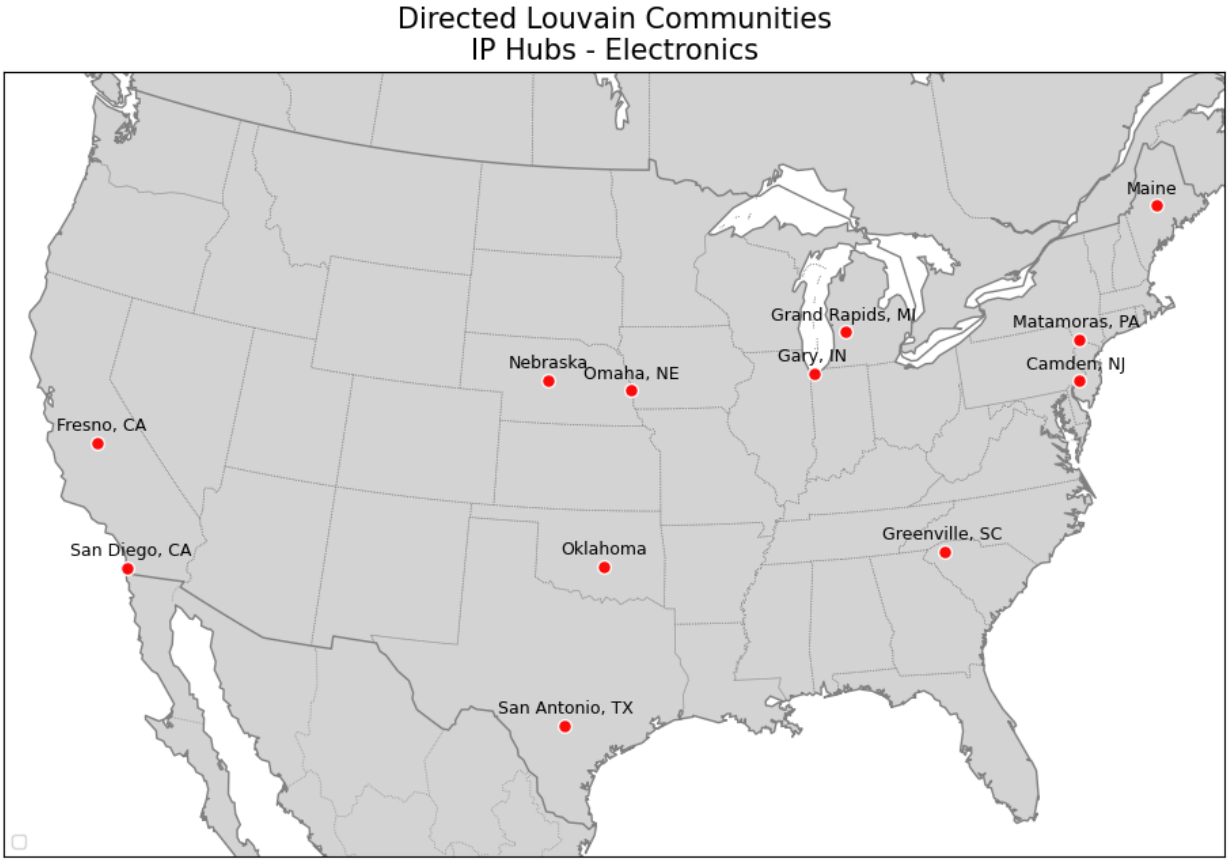


Figure C.25: Hubs identified from Directed Louvain Communities on the FAF Electronics dataset using Integer programming.

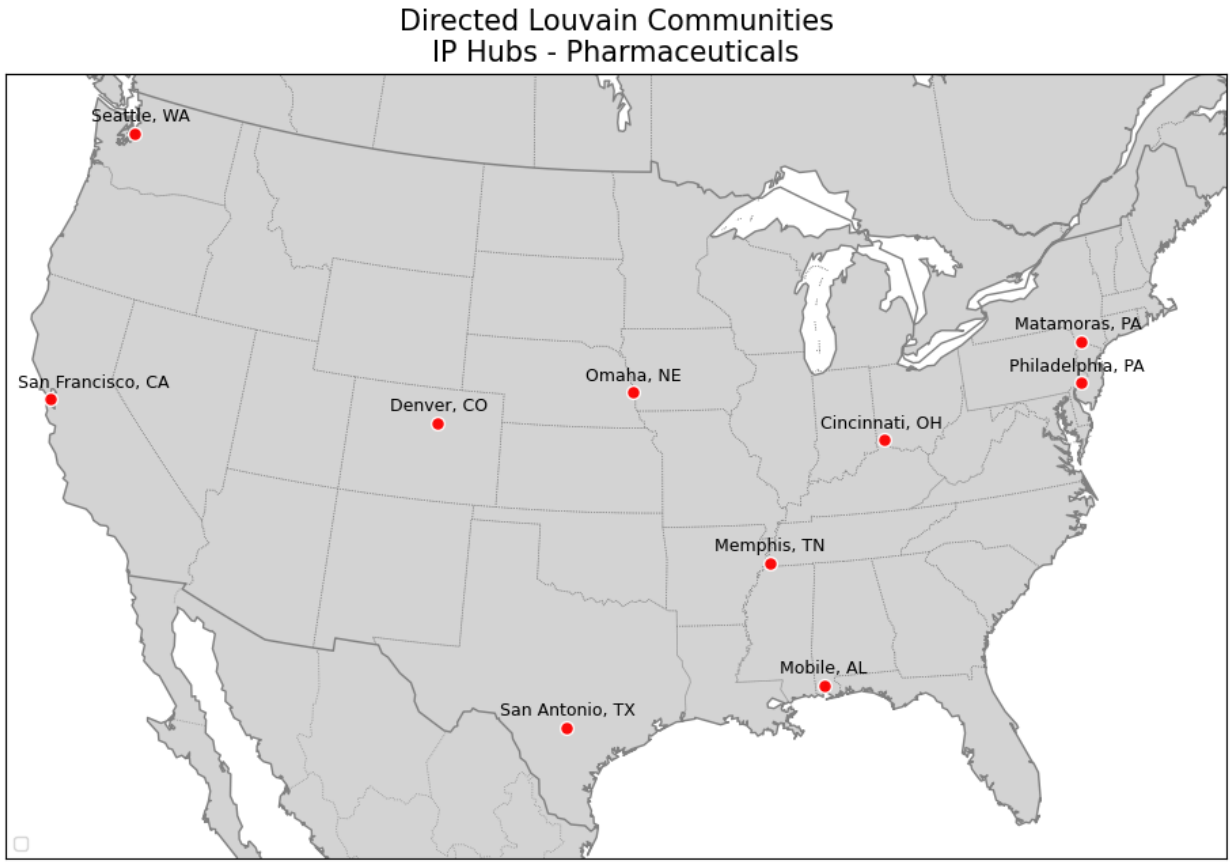


Figure C.26: Hubs identified from Directed Louvain Communities on the FAF Pharmaceutical dataset using Integer programming.

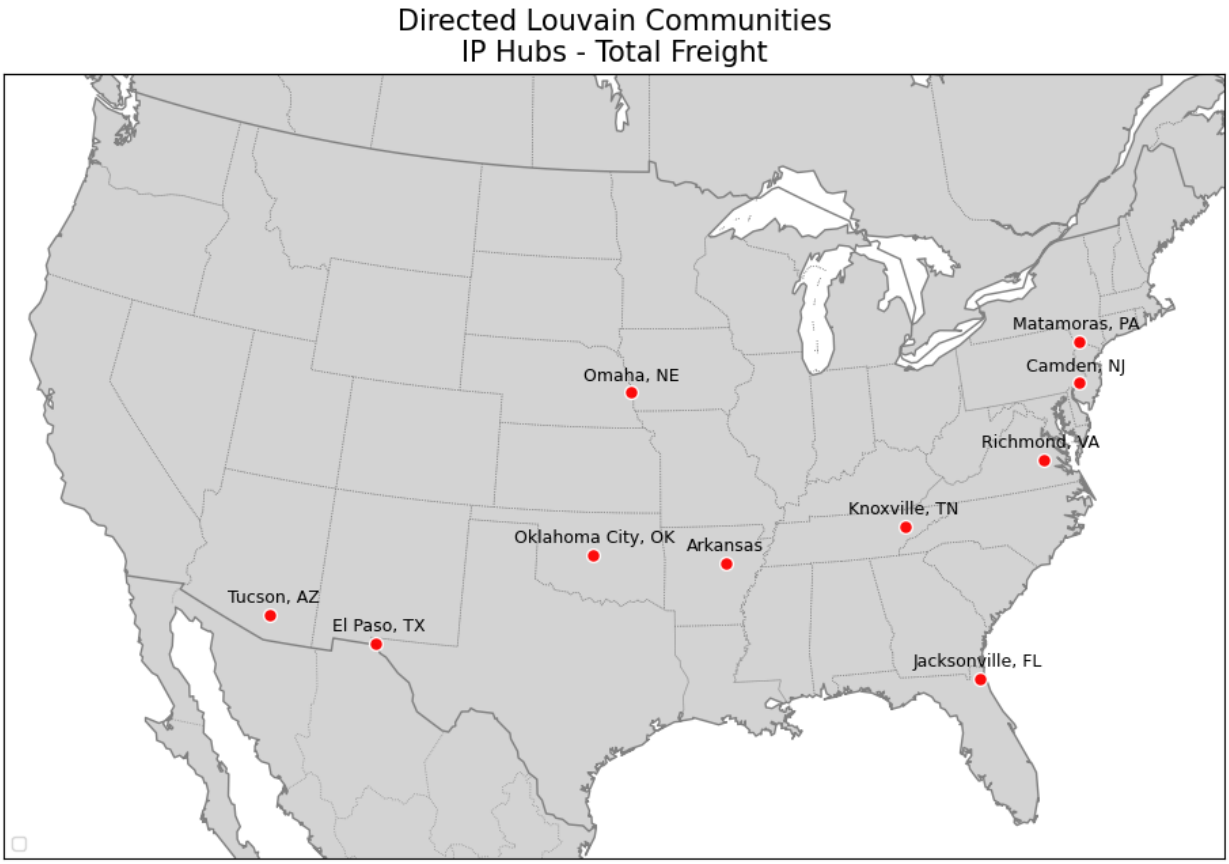


Figure C.27: Hubs identified from Directed Louvain Communities on the FAF Total Freight dataset using Integer programming.

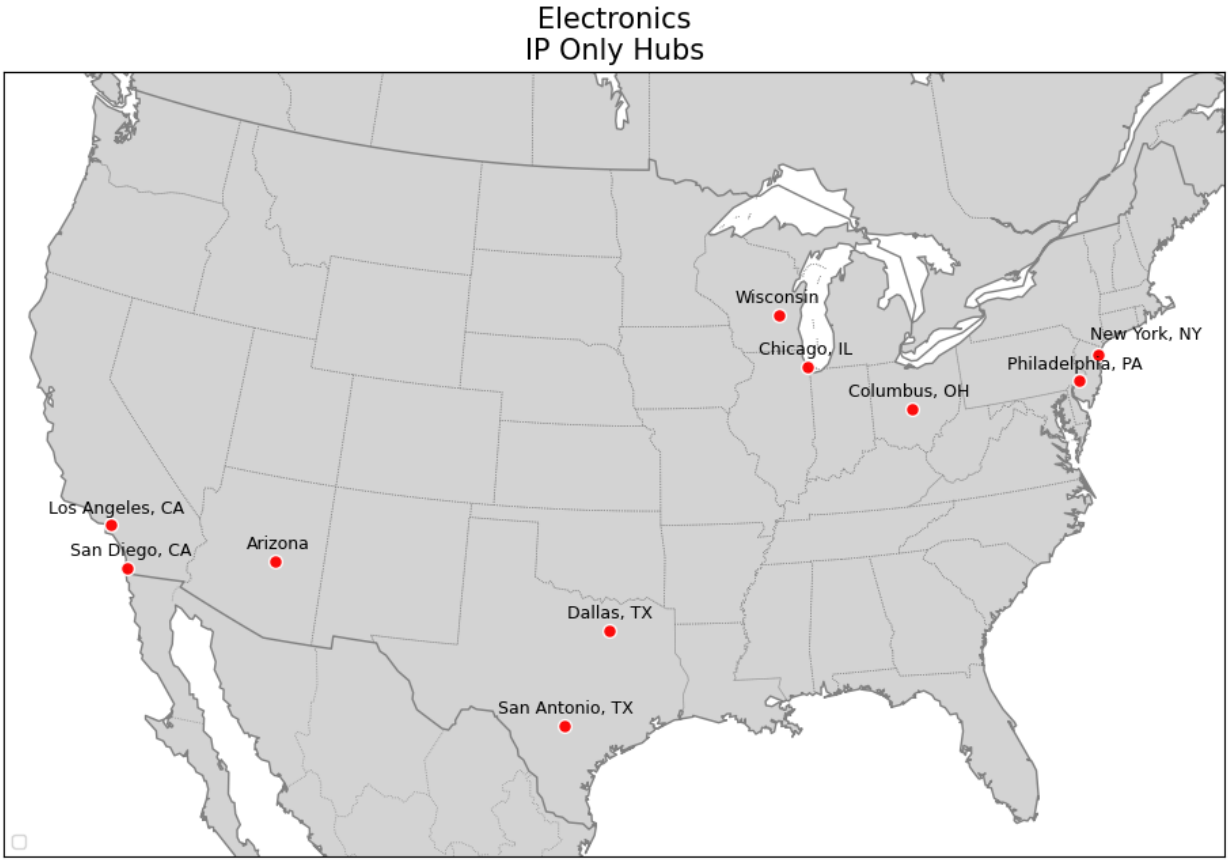


Figure C.28: Hubs identified from the 30 Largest Nodes on the FAF Electronics dataset using Integer programming.

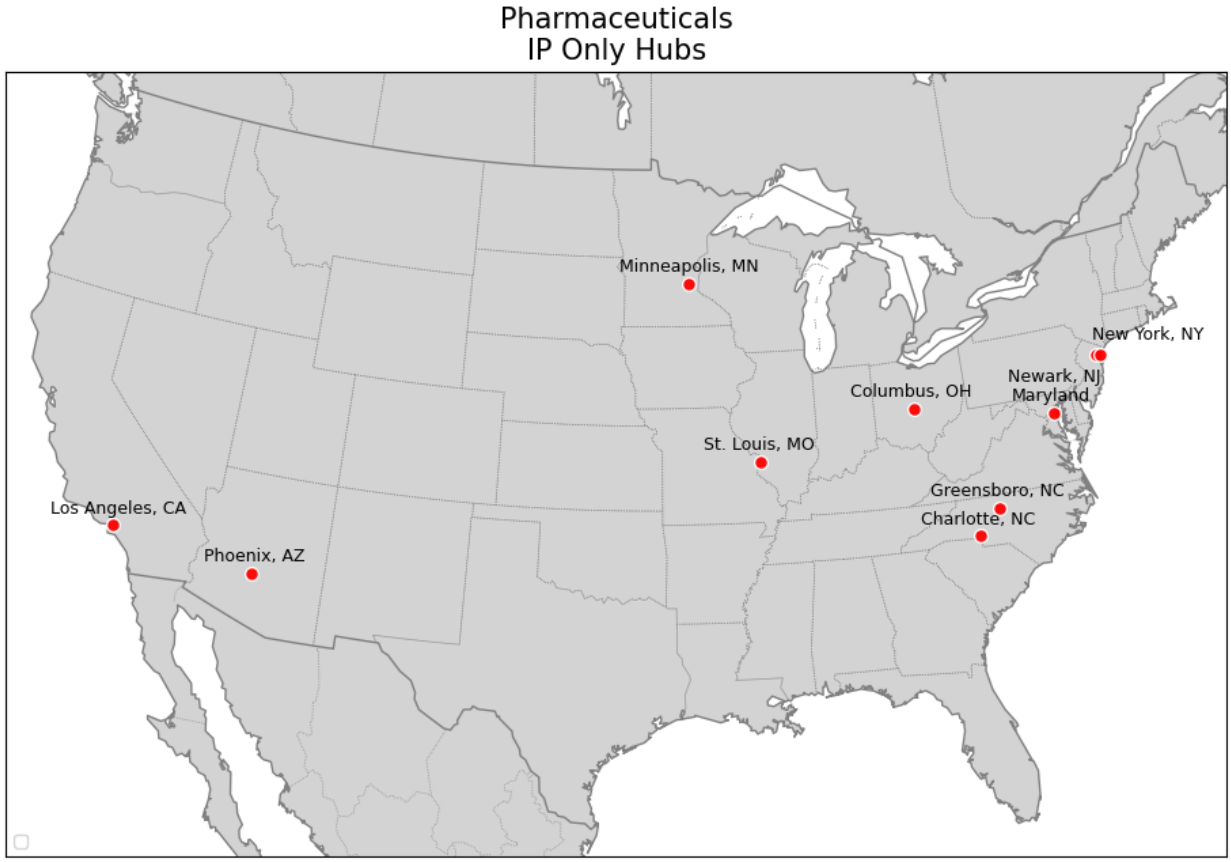


Figure C.29: Hubs identified from the 30 Largest Nodes on the FAF Pharmaceutical dataset using Integer programming.

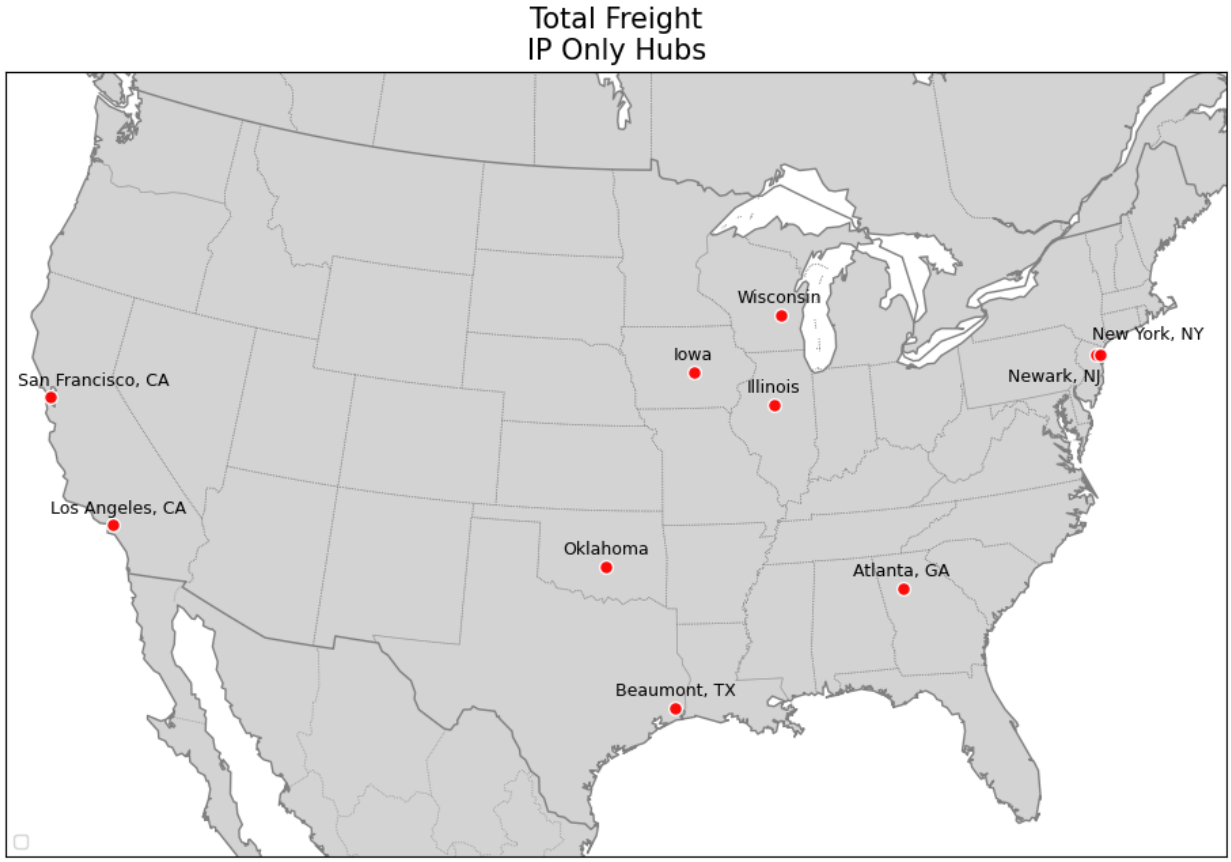


Figure C.30: Hubs identified from the 30 Largest Nodes on the FAF Total Freight dataset using Integer programming.

References

[1] V. Bulusu, R. Sengupta, and Z. Liu, "Unmanned aviation: To be free or not to be free?" In *7th International Conference on Research in Air Transportation*, 2016.

[2] A. Torsoli, "Air taxi plan for paris at risk of missing olympics deadline," *Bloomberg*, Feb. 2024. URL: https://www.bloomberg.com/news/articles/2024-02-26/air-taxi-plan-for-paris-at-risk-of-missing-olympics-deadline.

[3] E. F. Dulia, M. S. Sabuj, and S. A. Shihab, "Benefits of advanced air mobility for society and environment: A case study of ohio," *Applied Sciences*, vol. 12, no. 1, p. 207, 2021.

[4] D. L. Bryan and M. E. O'kelly, "Hub-and-spoke networks in air transportation: An analytical review," *Journal of regional science*, vol. 39, no. 2, pp. 275–295, 1999.

[5] J. F. Campbell and M. E. O'Kelly, "Twenty-five years of hub location research," *Transportation Science*, vol. 46, no. 2, pp. 153–169, 2012.

[6] S. Wolf, "On the complexity of the uncapacitated single allocation p-hub median problem with equal weights," 2007.

[7] R. S. Toh and R. G. Higgins, "The impact of hub and spoke network centralization and route monopoly on domestic airline profitability," *Transportation journal*, pp. 16–27, 1985.

[8] M. E. O'kelly, "The location of interacting hub facilities," *Transportation science*, vol. 20, no. 2, pp. 92–106, 1986.

[9] S. A. Alumur, J. F. Campbell, I. Contreras, B. Y. Kara, V. Marianov, and M. E. O'Kelly, "Perspectives on modeling hub location problems," *European Journal of Operational Research*, vol. 291, no. 1, pp. 1–17, 2021.

[10] R. Guimera, S. Mossa, A. Turtschi, and L. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 22, pp. 7794–7799, 2005.

[11] X. Feng, H. Jiang, and L.-p. Jiang, "Study on community detection of shipping network based on modularity," in *Green, Smart and Connected Transportation Systems: Proceedings of the 9th International Conference on Green Intelligent Transportation Systems and Safety*, Springer, 2020, pp. 365–374.

[12] J. Zheng, J. Qi, Z. Sun, and F. Li, "Community structure based global hub location problem in liner shipping," *Transportation Research Part E: Logistics and Transportation Review*, vol. 118, pp. 1–19, 2018.

[13] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[14] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026 113, 2004.

[15] Q. Wang, J. Zheng, and X. Liu, "Reliable liner shipping hub location problem considering hub failure," *Journal of Marine Science and Engineering*, vol. 11, no. 4, p. 818, 2023.

[16] M. E. O'Kelly, "A clustering approach to the planar hub location problem," *Annals of Operations Research*, vol. 40, no. 1, pp. 339–353, 1992.

[17] M. Peker, B. Y. Kara, J. F. Campbell, and S. A. Alumur, "Spatial analysis of single allocation hub location problems," *Networks and Spatial Economics*, vol. 16, pp. 1075–1101, 2016.

[18] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[19] S. E. Schaeffer, "Graph clustering," *Computer science review*, vol. 1, no. 1, pp. 27–64, 2007.

[20] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[21] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, "Using graph theory to analyze biological networks," *BioData mining*, vol. 4, pp. 1–27, 2011.

[22] W. Liu, A. Sidhu, A. M. Beacom, and T. W. Valente, "Social network theory," *The international encyclopedia of media effects*, pp. 1–12, 2017.

[23] S. Derrible and C. Kennedy, "Applications of graph theory and network science to transit network design," *Transport reviews*, vol. 31, no. 4, pp. 495–519, 2011.

[24] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, no. 4, Oct. 2008, ISSN: 1550-2376. DOI: 10.1103/physreve.78.046110. URL: http://dx.doi.org/10.1103/PhysRevE.78.046110.

[25] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[26] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2007.

[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008.

[28] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, p. 5233, 2019.

[29] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 03, P03024, 2009.

[30] N. Dugué and A. Perez, "Directed louvain: Maximizing modularity in directed networks," Ph.D. dissertation, Université d'Orléans, 2015.

[31] M. E. Newman, "Spectral methods for community detection and graph partitioning," *Physical Review E*, vol. 88, no. 4, p. 042 822, 2013.

[32] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[33] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[34] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, Springer, 2005, pp. 284–293.

[35] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[36] S.-H. Bae, D. Halperin, J. D. West, M. Rosvall, and B. Howe, "Scalable and efficient flow-based community detection for large-scale graph analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 3, pp. 1–30, 2017.

[37] H. Yin, A. R. Benson, and J. Leskovec, "Higher-order clustering in networks," *Physical Review E*, vol. 97, no. 5, p. 052 306, 2018.

[38] P.-Z. Li, L. Huang, C.-D. Wang, and J.-H. Lai, "Edmot: An edge enhancement approach for motif-aware community detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 479–487.

[39] W. G. Underwood, A. Elliott, and M. Cucuringu, "Motif-based spectral clustering of weighted directed networks," *CoRR*, vol. abs/2004.01293, 2020. arXiv: 2004.01293. URL: https://arxiv.org/abs/2004.01293.

[40] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 555–564.

[41] F. Liu, S. Xue, J. Wu, C. Zhou, W. Hu, C. Paris, S. Nepal, J. Yang, and P. S. Yu, "Deep learning for community detection: Progress, challenges and opportunities," *arXiv preprint arXiv:2005.08225*, 2020.

[42] L. Yang, X. Cao, D. He, C. Wang, X. Wang, and W. Zhang, "Modularity based community detection with deep learning.," in *IJCAI*, vol. 16, 2016, pp. 2252–2258.

[43] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[44] N. Dugué and A. Perez, "Direction matters in complex networks: A theoretical and applied study for greedy modularity optimization," *Physica A: Statistical Mechanics and its Applications*, vol. 603, p. 127 798, 2022, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2022.127798.

[45] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010, ISSN: 1532-4435.

[46] H. van der Hoef and M. J. Warrens, "Understanding information theoretic measures for comparing clusterings," *Behaviormetrika*, vol. 46, no. 2, pp. 353–370, 2019.

[47] J. Leskovec and A. Krevl, *SNAP Datasets: Stanford large network dataset collection*, http://snap.stanford.edu/data, 2014.

[48] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval*, vol. 3, pp. 127–163, 2000.

[49] W. W. Cohen, "Integration of heterogeneous databases without common domains using queries based on textual similarity," in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 1998, pp. 201–212.

[50] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Intelligent Data Engineering and Automated Learning–IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings 14*, Springer, 2013, pp. 611–618.

[51] M. Wijewickrema, V. Petras, and N. Dias, "Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora," *The Electronic Library*, vol. 37, no. 3, pp. 506–527, 2019.

[52] D. R. White and S. P. Borgatti, "Betweenness centrality measures for directed graphs," *Social networks*, vol. 16, no. 4, pp. 335–346, 1994.

[53] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of mathematical sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[54] U.S. Department of Transportation, Bureau of Transportation Statistics (BTS), Federal Highway Administration (FHWA), *Freight analysis framework, faf5*, https://doi.org/10.21949/1529116, Accessed: February 28, 2024, 2017.

[55] N. I. Gunady, S. R. Patel, and D. DeLaurentis, "A system-of-systems approach to analyzing future advanced air mobility cargo operations," in *2022 17th Annual System of Systems Engineering Conference (SOSE)*, IEEE, 2022, pp. 368–373.

[56] S. De Silvestri, M. Pagliarani, F. Tomasello, D. Trojaniello, and A. Sanna, "Design of a service for hospital internal transport of urgent pharmaceuticals via drones," *Drones*, vol. 6, no. 3, p. 70, 2022.

[57] H. Yaman, "Allocation strategies in hub networks," *European Journal of Operational Research*, vol. 211, no. 3, pp. 442–451, 2011.

[58] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[59] P. Bonacich, "Power and centrality: A family of measures," *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[60] M. Newman, *Networks*. Oxford university press, 2018.

[61] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of mathematical sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[62] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical review E*, vol. 72, no. 2, p. 027 104, 2005.

[63] J.-P. Rodrigue, *The geography of transport systems*. Routledge, 2020.