# Go with the Flow? A Large-Scale Analysis of Health Care Delivery Networks in the United States Using Hodge Theory


Thomas Gebhart[1] Xiaojun Fu[2] Russell J. Funk[3]

[1]Computer Science and Engineering, University of Minnesota

[2]School of Physics and Astronomy, University of Minnesota

[3]Carlson School of Management, University of Minnesota

# Motivation

# Background

## Health Care Delivery in the United States

- Relative to comparable countries, the United States spends far more on health care, nearly 18% of its GDP in 2016.
- Yet it has little to show for that spending, ranking near the bottom of Western, industrialized nations on many critical health outcomes.
- While the problems are complex, many suggest that the fragmented nature of care delivery contributes significantly to the health care system's poor performance.
- Care fragmentation occurs when the delivery of services to patients is spread across multiple, disconnected providers.
- In settings with greater care fragmentation, communication and coordination among care team members is more difficult.
- Consequently, care fragmentation leads to higher spending and lower quality.

# Our approach

- In this study, we leverage recent advances in topological data analysis and the growing availability of "big data" on health care delivery to study care fragmentation at scale.
- Specifically, using claims data from Medicare, we map care delivery networks across regions (2014-2017), wherein edges track patient flows among local physicians.
- Subsequently, we use Hodge theory to decompose the observed patient flows into their local cyclic (curl), global cyclic (harmonic), and acyclic (gradient) components.
- We then examine associations between these three different flow patterns and measures of local care quality and spending.

# Data

# Data

- Our primary data are derived from Medicare claims.
- Bills (or claims) submitted to Medicare for reimbursement include detailed information about the billing providers and dates and locations of service.
- These data are exceptionally rich, allowing us to map hundreds of millions of provider-provider relationships across all 50 states, from 2014 to 2017.
- We also collected information on local care quality and spending from the Dartmouth Institute for Health Policy and Clinical Practice.
- In addition, basic data on providers (e.g., practice locations) were obtained from the National Plan and Provider Enumeration System (NPPES).

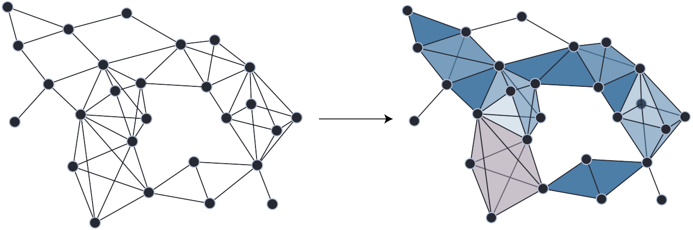
# Methods

# Mapping care delivery networks

- The referral data are formatted as edge lists, one for each year of observation.
- Nodes correspond to providers (indicated by NPIs).
- Edges are recorded between pairs of providers when they bill for the same patients within a defined time window, and are weighted by the number of shared patients.
  - For example, if NPI A saw 30 patients in one week, and 12 of those subsequently saw NPI B in the next week, we would record an edge between A and B with a weight of 12.
- There is a directionality to the edges, implied by the timing of patient visits, which motivates our view of these networks as tracking patient flows.
- Because health care delivery tends to be highly localized, we map care delivery networks within regions (Hospital Service Areas).
- For each observation year $\times$ HSA, we identify all local providers, based on practice addresses, and then map their relationships using the referral data.

# Combinatorial Hodge theory

- Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with $n_0 = |\mathcal{V}|$ nodes and $n_1 = |\mathcal{E}|$ edges.
- We define the *clique complex* $\mathcal{K}(G)$ of $G$ by "filling in" all $k$-cliques, treated as $(k-1)$-dimensional simplices.
- For each dimension $k$, define the space of $k$-chains $\mathcal{C}_k$ as a finite-dimensional Hilbert space with coefficients in $\mathbb{R}$.



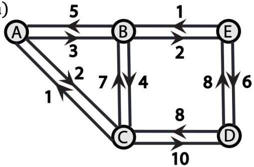
# Combinatorial Hodge theory

- $\mathcal{C}_k$ has a dual space of $k$-dimensional co-chains $\mathcal{C}^k$ composed of alternating functions $f : \mathcal{C}_k \to \mathbb{R}$.
- $\mathcal{C}^1$ may be interpreted as the space of edge flows on $G$.
- A *flow* $\boldsymbol{f} \in \mathbb{R}^{n_1}$ is an assignment of a real number each edge, negative values indicating flow in direction opposite to orientation.

# Combinatorial Hodge theory

- $\mathcal{C}_k$ has a dual space of $k$-dimensional co-chains $\mathcal{C}^k$ composed of alternating functions $f : \mathcal{C}_k \to \mathbb{R}$.
- $\mathcal{C}^1$ may be interpreted as the space of edge flows on $G$.
- A *flow* $\boldsymbol{f} \in \mathbb{R}^{n_1}$ is an assignment of a real number each edge, negative values indicating flow in direction opposite to orientation.
- The *boundary operator* takes $k$-chains to $(k-1)$-chains $\boldsymbol{B}_k : \mathcal{C}_k \to \mathcal{C}_{k-1}$.
- Dually, the coboundary map follows as $\boldsymbol{B}_k^\top : \mathcal{C}_k \to \mathcal{C}_{k+1}$.

(a)




bidirected edge flow

(b)

$B_1 =$

| | [A, B] | [A, C] | [B, C] | [B, E] | [C, D] | [D, E] |
|---|---|---|---|---|---|---|
| A | -1 | -1 | 0 | 0 | 0 | 0 |
| B | 1 | 0 | -1 | -1 | 0 | 0 |
| C | 0 | 1 | 1 | 0 | -1 | 0 |
| D | 0 | 0 | 0 | 0 | 1 | -1 |
| E | 0 | 0 | 0 | 1 | 0 | 1 |

$B_2 =$

| | [A, B, C] |
|---|---|
| [A, B] | 1 |
| [A, C] | -1 |
| [B, C] | 1 |
| [B, E] | 0 |
| [C, D] | 0 |
| [D, E] | 0 |

# The Hodge Laplacian

- The *Hodge Laplacian* is given by

$$\boldsymbol{\mathcal{L}}_k = \boldsymbol{B}_k^\top \boldsymbol{B}_k + \boldsymbol{B}_{k+1} \boldsymbol{B}_{k+1}^\top$$

- Of particular interest for our application is the *Hodge 1-Laplacian*:

$$\boldsymbol{\mathcal{L}}_1 = \boldsymbol{B}_1^\top \boldsymbol{B}_1 + \boldsymbol{B}_2 \boldsymbol{B}_2^\top$$

- The Hodge Laplacian generalizes the standard graph Laplacian: $\boldsymbol{\mathcal{L}}_0 = \boldsymbol{B}_1 \boldsymbol{B}_1^\top$.

(b)

$B_1 =$

| | [A, B] | [A, C] | [B, C] | [B, E] | [C, D] | [D, E] |
|---|---|---|---|---|---|---|
| A | -1 | -1 | 0 | 0 | 0 | 0 |
| B | 1 | 0 | -1 | -1 | 0 | 0 |
| C | 0 | 1 | 1 | 0 | -1 | 0 |
| D | 0 | 0 | 0 | 0 | 1 | -1 |
| E | 0 | 0 | 0 | 1 | 0 | 1 |

$B_2 =$

| | [A, B, C] |
|---|---|
| [A, B] | 1 |
| [A, C] | -1 |
| [B, C] | 1 |
| [B, E] | 0 |
| [C, D] | 0 |
| [D, E] | 0 |

$$\mathcal{L} = B_1^\top B_1 + B_2^\top B_2 = \begin{array}{|cccccc} 3 & 0 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & 0 & -1 & 0 \\ 0 & 0 & 3 & 1 & -1 & 0 \\ -1 & 0 & 1 & 2 & 0 & 1 \\ 0 & -1 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 1 & -1 & 2 \end{array}$$

# Hodge decomposition

- $\mathrm{im}(\boldsymbol{B}_k)$ defines the space of $(k-1)$ boundaries and $\ker(\boldsymbol{B}_k)$ the space of $k$-cycles.
- The vector space $\mathcal{H}_k = \ker(\boldsymbol{B}_k)/\,\mathrm{im}(\boldsymbol{B}_{k+1})$ has rank equal to the number of $k$-dimensional holes in $\mathcal{K}(G)$.
- Functions $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_k)$ are called *harmonic*, in reference to their status as solutions to the (discrete) Laplace equation $\boldsymbol{\mathcal{L}}_k \boldsymbol{h} = \mathbf{0}$.
- The harmonic functions are representatives of elements in $\mathcal{H}_k$.

# Hodge decomposition

- $\text{im}(\boldsymbol{B}_k)$ defines the space of $(k-1)$ boundaries and $\ker(\boldsymbol{B}_k)$ the space of $k$-cycles.
- The vector space $\mathcal{H}_k = \ker(\boldsymbol{B}_k)/\,\text{im}(\boldsymbol{B}_{k+1})$ has rank equal to the number of $k$-dimensional holes in $\mathcal{K}(G)$.
- Functions $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_k)$ are called *harmonic*, in reference to their status as solutions to the (discrete) Laplace equation $\boldsymbol{\mathcal{L}}_k \boldsymbol{h} = \boldsymbol{0}$.
- The harmonic functions are representatives of elements in $\mathcal{H}_k$.
- $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_k)$ requires that $\boldsymbol{h} \in \ker(\boldsymbol{B}_k)$ and $\boldsymbol{h} \in \ker(\boldsymbol{B}_{k+1})$, therefore we may decompose $\mathcal{C}_k$ as:
$$\mathcal{C}_k = \text{im}(\boldsymbol{B}_{k+1}) \oplus \text{im}(\boldsymbol{B}_k^\top) \oplus \ker(\boldsymbol{\mathcal{L}}_k)$$

# Hodge decomposition

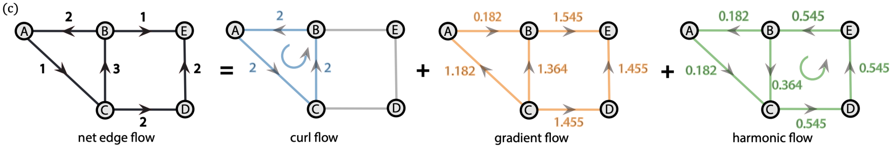
- $\mathrm{im}(\boldsymbol{B}_k)$ defines the space of $(k-1)$ boundaries and $\ker(\boldsymbol{B}_k)$ the space of $k$-cycles.
- The vector space $\mathcal{H}_k = \ker(\boldsymbol{B}_k)/\,\mathrm{im}(\boldsymbol{B}_{k+1})$ has rank equal to the number of $k$-dimensional holes in $\mathcal{K}(G)$.
- Functions $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_k)$ are called *harmonic*, in reference to their status as solutions to the (discrete) Laplace equation $\boldsymbol{\mathcal{L}}_k \boldsymbol{h} = \boldsymbol{0}$.
- The harmonic functions are representatives of elements in $\mathcal{H}_k$.
- $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_k)$ requires that $\boldsymbol{h} \in \ker(\boldsymbol{B}_k)$ and $\boldsymbol{h} \in \ker(\boldsymbol{B}_{k+1})$, therefore we may decompose $\mathcal{C}_k$ as:
$$\mathcal{C}_k = \mathrm{im}(\boldsymbol{B}_{k+1}) \oplus \mathrm{im}(\boldsymbol{B}_k^\top) \oplus \ker(\boldsymbol{\mathcal{L}}_k)$$
- On the space of edge flows $\mathcal{C}^1$ this becomes
$$\mathcal{C}^1 \cong \mathcal{C}_1 = \mathrm{im}(\boldsymbol{B}_2) \oplus \mathrm{im}(\boldsymbol{B}_1^\top) \oplus \ker(\boldsymbol{\mathcal{L}}_1)$$

# Hodge decomposition

- On the space of edge flows $\mathcal{C}^1$ this becomes

$$\mathcal{C}^1 \cong \mathcal{C}_1 = \mathrm{im}(\boldsymbol{B}_2) \oplus \mathrm{im}(\boldsymbol{B}_1^\top) \oplus \ker(\boldsymbol{\mathcal{L}}_1)$$

- $\mathrm{im}(\boldsymbol{B}_2)$ is the *curl* subspace consisting of weighted flows $\boldsymbol{r} \in \mathrm{im}(\boldsymbol{B}_2)$ which may be composed of local circulations along any 2-simplex (3-clique).
- $\mathrm{im}(\boldsymbol{B}_1^\top)$ is a weighted cut space of edges which disconnect the network or, equivalently, *gradient* flows $\boldsymbol{g} \in \mathrm{im}(\boldsymbol{B}_1^\top)$ which contain no cyclic component.
- *Harmonic* elements $\boldsymbol{h} \in \ker(\boldsymbol{\mathcal{L}}_1)$ are weighted global circulations that do not sum to zero around cycles but are inexpressible as linear combinations of curl flow around 2-simplices.



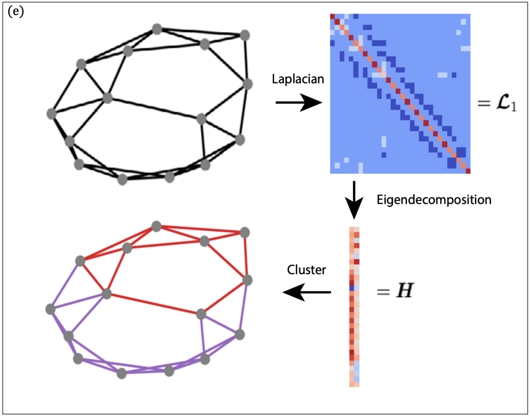
# Random walk normalization

- Note: for our analyses, we compute a normalized form of $\boldsymbol{\mathcal{L}}_1$ and the resulting decomposition known as the Random-walk normalization.
- This normalization mimics the random walk normalization of the graph Laplacian in higher dimensions by approximating the steady-state transition matrix of a random walker on $\mathcal{K}(G)$.
- We will not go into specifics here, but see the paper for more details.

For more information see Schaub et al. (2020).

# Harmonic Clustering

- The harmonic functions of $\mathcal{L}_1$ encode topological features of $\mathcal{K}(G)$, and by extension, $G$.
- Let $\mathcal{L}_1 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and collect the eigenvectors (harmonic functions) corresponding to the fist $d$ 0-eigenvalues $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2 \ldots \boldsymbol{h}_d)$.
- We can then cluster $\boldsymbol{H}$ using any standard clustering method, though subspace clustering.




For more information see Ebli and Spreemann (2019).

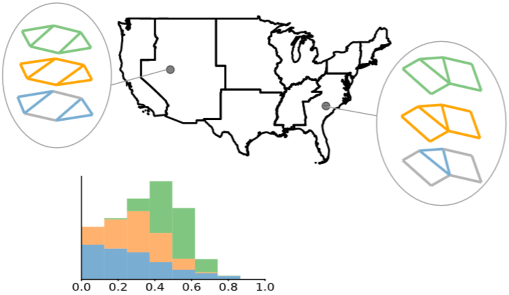
# Network-level measures

- We define network-level measures of flow, computed for each region $i$ and year $t$:
- gradient flow per edge $\bar{g}_{it} = \frac{1}{n_1} | \sum_e^{n_1} g_e |$
- harmonic flow per edge $\bar{h}_{it} = \frac{1}{n_1} | \sum_e^{n_1} h_e |$
- curl flow per edge $\bar{r}_{it} = \frac{1}{n_1} | \sum_e^{n_1} r_e |$

(d)

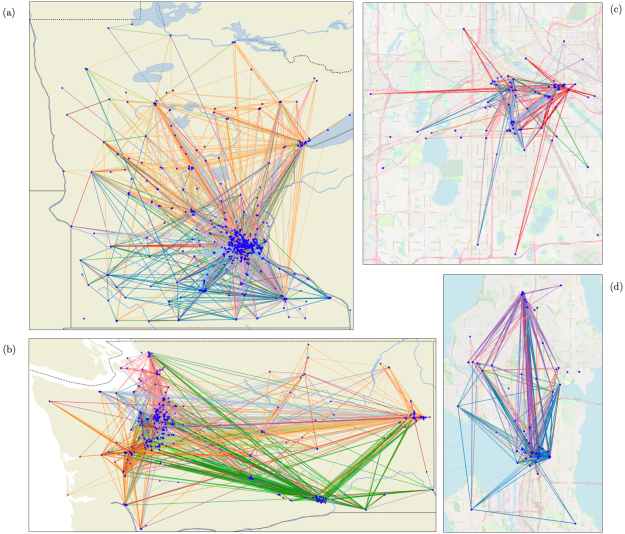
Network level

| | | | |
|---|---|---|---|
| Curl flow $r$ | 6 | Curl flow per edge $\bar{r}$ | 1 |
| Gradient flow $g$ | -1.365 | Gradient flow per edge $\bar{g}$ | 0.228 |
| Harmonic flow $h$ | -1.635 | Harmonic flow per edge $\bar{h}$ | 0.273 |



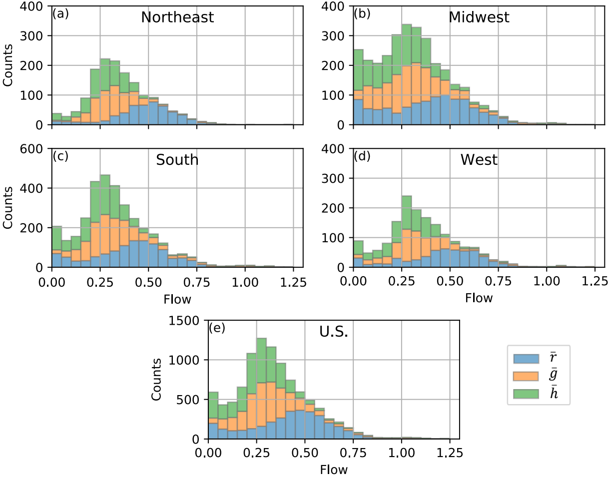
# Results

# Harmonic clustering of care delivery networks




Care delivery networks depicted for Minnesota (a), Washington (b), Minneapolis (c), and Seattle (d) as of 2017.

# Distribution of patient flows by subspace and region




- Harmonic flow per edge is the lowest in all regions, which seems reasonable, as global cyclic flow is likely harder to form.
- Curl flow per edge assumes larger values, which also seems plausible, as the formation of local cycles is probably natural in a care delivery network, where team coordination is important.

# Regression models

- Up to this point, our results have shown that there is substantial variability in the composition of patient flows across regions.
- We also considered whether this variation is correlated with spending and quality.
- To do so, we estimate a series of linear regression models.
- The unit of observation is the region $\times$ year.
- Each model includes three independent variables, which correspond to sums across flow values assigned to each edge, separately for each subspace (adjusted by network size).
- To adjust for temporal trends, each model includes year fixed effects.
- We also estimate models that control for socioeconomic conditions, which have been shown to be predictive of regional health care cost and quality.

| | (1) DV: Total spending per beneficiary | (2) DV: Inpatient spending per beneficiary | (3) DV: Outpatient spending per beneficiary | (4) DV: Readmission rate post-surgical treatment | (5) DV: ER visit rate post-surgical treatment |
|---|---|---|---|---|---|
| Gradient flow $\bar{g}$ per edge | 60.68 | 34.89 | -288.66*** | -0.13 | -0.37 |
| | (163.44) | (98.35) | (79.10) | (0.87) | (0.71) |
| Harmonic flow $\bar{h}$ per edge | 1147.09*** | 1137.64*** | 2828.18*** | 2.59** | 4.42*** |
| | (321.92) | (201.14) | (160.58) | (1.31) | (1.11) |
| Curl flow $\bar{r}$ per edge | -1702.83*** | -1416.81*** | -2712.92*** | -2.80*** | -3.64*** |
| | (216.07) | (135.18) | (105.29) | (0.52) | (0.51) |
| Constant | 10361.97*** | 4726.87*** | 2557.93*** | 11.54*** | 16.90*** |
| | (57.38) | (37.14) | (29.81) | (0.16) | (0.16) |
| Year fixed effects | Yes | Yes | Yes | Yes | Yes |
| N | 12952 | 12952 | 12952 | 6034 | 7776 |
| r2 | 0.08 | 0.05 | 0.22 | 0.01 | 0.03 |

Robust standard errors (clustered on region) are shown in parentheses; *p<0.1; **p<0.05; ***p<0.01

- Harmonic $\bar{h}$ and curl $\bar{r}$ flow are associated with spending, but in opposite directions.
- When harmonic flow is greater, spending is higher; when curl flow is greater, it's lower.
- For perspective, a 1 SD increase in curl flow is associated with a decrease of $354.75 in annual spending per beneficiary; for an average region, the savings works out to almost $3 million/year.
- Turning to quality, we find that greater curl $\bar{r}$ flow is associated with better outcomes, but again, the opposite holds for harmonic flow.
- Our models are robust to controls for socioeconomic factors, and are comparable in effect sizes.
- A 1 SD decrease in the population without a high school degree is associated with a $465.76 drop in spending/beneficiary, on par with the savings associated with a similar decrease in harmonic flow.

| | (1) DV: Total spending per beneficiary | (2) DV: Inpatient spending per beneficiary | (3) DV: Outpatient spending per beneficiary | (4) DV: Readmission rate post-surgical treatment | (5) DV: ER visit rate post-surgical treatment |
|---|---|---|---|---|---|
| Gradient flow $\bar{g}$ per edge | -196.34 | -127.45 | -184.56** | 0.43 | -0.09 |
| | (153.25) | (93.77) | (74.60) | (0.80) | (0.66) |
| Harmonic flow $\bar{h}$ per edge | 557.08* | 871.46*** | 2748.45*** | 0.67 | 2.42** |
| | (329.08) | (204.19) | (158.85) | (1.27) | (1.10) |
| Curl flow $\bar{r}$ per edge | -862.23*** | -1004.39*** | -2665.44*** | -1.51** | -1.82*** |
| | (238.69) | (146.99) | (111.30) | (0.60) | (0.59) |
| Median household income ($) | -0.00 | 0.00 | -0.00 | 0.00 | -0.00*** |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| Unemployment rate (%) | 7.53 | 2.98 | -50.70*** | 0.08*** | 0.12*** |
| | (9.86) | (5.89) | (4.22) | (0.02) | (0.02) |
| No high school degree (%) | 71.01*** | 44.43*** | -8.68*** | 0.05*** | 0.03*** |
| | (5.19) | (3.00) | (1.53) | (0.01) | (0.01) |
| Hispanic population (%) | -2.54*** | -1.25*** | -0.98*** | -0.00*** | -0.00 |
| | (0.68) | (0.37) | (0.21) | (0.00) | (0.00) |
| Black population (%) | 4.61*** | 2.37*** | 0.66*** | 0.01*** | -0.00 |
| | (0.97) | (0.50) | (0.22) | (0.00) | (0.00) |
| Constant | 9227.76*** | 4033.15*** | 3007.61*** | 9.94*** | 15.35*** |
| | (94.76) | (60.42) | (43.81) | (0.21) | (0.23) |
| Year fixed effects | Yes | Yes | Yes | Yes | Yes |
| N | 12950 | 12950 | 12950 | 6034 | 7776 |
| R2 | 0.18 | 0.16 | 0.29 | 0.07 | 0.06 |
| Wald tests for flow predictors | | | | | |
| F | 11.08 | 19.79 | 208.88 | 3.82 | 4.14 |
| d.f. | 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| p-value | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 |

# Wrapping up

# Wrapping up

- Care fragmentation is a critical problem facing health care delivery in the United States.
- Recently, the growing availability of "big data" has enabled unprecedented insight into care delivery, creating opportunities to better understand and address fragmentation.
- We utilized a novel framework from topological data analysis—the discrete Hodge decomposition—to study flows of patients among physicians in care delivery networks.
- We found substantial variation across broad regions of the country, perhaps corresponding to institutional differences in care delivery.
- Moreover, we observed that greater curl flow is associated with better performance (i.e., lower cost, higher quality), but the opposite holds for harmonic flow.
- Given our context, these patterns seem plausible.
  - The movement of patients around global cycles seems problematic from a care coordination perspective, potentially leading to higher cost, lower quality care.
  - By contrast, the movement of patients around local cycles (as indicated by greater curl flow) seems more conductive to close coordination among providers.
- While preliminary, our findings highlight the significant potential of emerging methods in topological data analysis for the study of health care delivery.