Graph Frequency Analysis of COVID-19 Incidence in the United States

Yang Li and Gonzalo Mateos
Dept. of Electrical & Computer Engineering, University of Rochester
Rochester, NY, US
{yli131,gmateosb}@ur.rochester.edu

ABSTRACT

The COVID-19 pandemic markedly changed the way of life in the United States (US). From early isolated regional outbreaks to ongoing country-wise spread, the contagion exhibits different patterns at various timescales and locations. Thus, a close study of the COVID-19 spread patterns can offer valuable insights on how counties were affected by the virus. In the present work, a graph frequency analysis was conducted to investigate the spread pattern of COVID-19 in the US. A geographical graph was constructed by computing the geodesic distance between 3142 US counties. The numbers of daily confirmed COVID-19 cases per county were collected and represented as graph signals, then mapped into the frequency domain via the graph Fourier transform. The concept of graph frequency in Graph Signal Processing (GSP) enables the decomposition of graph signals (i.e. daily confirmed cases) into modes with smooth or rapid variations with respect to the underlying graph connectivity. Follow-up analysis revealed the relationship between graph frequency components and the COVID-19 spread pattern within and across counties. Specifically, our preliminary graph frequency analysis mined (and learned from) confirmed case counts to unveil spatio-temporal contagion patterns of COVID-19 incidence for each US county. Overall, results here support the promising prospect of using GSP tools for epidemiology knowledge discovery on graphs.

CCS CONCEPTS

• **Computing methodologies** → *Graph signal processing*; • **Information systems** → *Graph data mining*; • **Networks** → *Graph frequency analysis.*

KEYWORDS

Graph data mining, graph signal processing, frequency analysis, contagion pattern recognition

ACM Reference Format:
Yang Li and Gonzalo Mateos. 2020. Graph Frequency Analysis of COVID-19 Incidence in the United States. In *Proceedings of MLG '20: 16th International Workshop on Mining and Learning with Graphs, August, 2020, San Diego, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

MLG '20, August, 2020, San Diego, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Networks are ubiquitous and their graph representations offer an ideal tool to record and analyze massive amounts of data from almost every aspect of human life [15]: social networks [16, 27], traffic networks [2, 3] and biological networks [10, 26], just to name a few. Network data usually reside on irregular and complex structures, requiring graph algorithms for in-depth analysis [14].

Graphs enable modeling complex interactions within data by defining nodes as data entities and edges as relations between nodes. It is often beneficial to also consider nodal attributes that represent certain features of the elements of interest. Such attributes are often conceptualized as signals defined on graphs [15]. Unlike in classical signal processing (SP), the underlying graph topology provides a fair amount of prior information about the said graph signals, while the graph signals themselves can also determine and update pairwise node relationships embedded within graph edges [20]. Accordingly, the field of Graph Signal Processing (GSP) [19, 22] emerged to fruitfully leverage the relational structure encoded in the graph when carrying out information processing tasks. Fundamental concepts in classical SP were generalized to accommodate graph data, notably the graph Fourier transform (GFT) to enable characteristic operations such as filtering and sampling. Noteworthy GSP advances include inference and generation of graph signals from network structures [6, 9], network topology inference from graph signals [4, 12], and integration of both graph signals and topology for knowledge discovery in various timely applications [5, 18, 21, 25]. The required GSP background is briefly introduced in Section 2; see also [15, 19, 22] for further details.

As COVID-19 spreads on United States (US) soil and severely impacts a multitude of counties, there has been a great amount of interest in understanding the spread patterns of the virus. Most current work has focused on analyzing pathologies from a biological perspective [13, 28], or on studies of contagion within a specific location [11, 17]. In this work, we bring to bear recent GSP advances to investigate the spread pattern of COVID-19 across all counties in the US, providing a comprehensive spatio-temporal analysis of the contagion that is still ongoing. Specifically, we contribute via:

• **Spatio-temporal study**. Data from 3142 US counties was collected, offering a macroscopic view of the contagion within the nation. The number of daily confirmed cases studied ranges from January 22, 2020 to April 30, 2020. This 100 day window facilitates analysis of the evolution of contagion patterns across time.

• **Graph frequency analysis**. A graph frequency analysis was conducted to extract valuable information from frequency domain, beyond traditional vertex or time domain analyses. Specifically, we

established the correspondence between graph frequency components (via low/high-pass graph filtering) and spatial contagion patterns (within/across counties, respectively) in the network. GFT coefficients reveal fundamentally different contagion patterns across locations, and also help identify counties at risk of major outbreaks that were not readily apparent via simple temporal analysis.

2 GRAPH-THEORETIC PRELIMINARIES

As the Data Science revolution keeps gaining momentum, it is only natural that complex signals with irregular structure become increasingly of interest. While there are many possible sources and models of added complexity, a general proximity relationship between signal elements is not only a plausible but a ubiquitous model across science and engineering. In this section, we briefly review needed graph-theoretic fundamentals and introduce the concepts of GSP, GFT and filtering operations in the graph frequency domain.

2.1 Graph signal fundamentals

Consider signals whose values are associated with nodes of a weighted, undirected, and connected graph. Formally, we consider the signal $\mathbf{x} = [x_1, \ldots, x_N]^T \in \mathbb{R}^N$ and the weighted graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = \{1, \ldots, N\}$ is a set of N vertices or nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. Scalar x_i denotes the signal value at node $i \in \mathcal{V}$. The map $\mathbf{W} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ from the set of unordered pairs of vertices to the nonnegative reals associates a weight $W_{ij} \geq 0$ with the edge $(i, j) \in \mathcal{E}$, while $W_{ij} \equiv 0$ for $(i, j) \neq \mathcal{E}$. The symmetric coefficients $W_{ij} = W_{ji}$ represent the strength of the connection (i.e., the similarity or influence) between nodes i and j. Henceforth the graph nodes will be US counties and edge weights correspond to the geodesic distance between counties; see also Section 3.2 for details on the simple graph construction process. In terms of the signal $\mathbf{x}$ defined by the daily number of confirmed cases across all counties, this means that when the weight W_{ij} is small the signal values x_i and x_j tend to be similar, based on the assumption that infections happen via localized contacts. Conversely, when the weight W_{ij} is large, the signal values x_i and x_j are not directly related except for what is implied by their weak connections to other nodes. Such an interpretation of the edge weights establishes a link between the signal values and the graph topology.

2.2 Graph Fourier transform and smoothness

An instrumental GSP tool is the GFT, which decomposes a graph signal into orthonormal components describing different modes of variation with respect to the graph topology . The GFT allows to equivalently represent a graph signal in two different domains – the vertex domain consisting of the nodes in $\mathcal{V}$, and the graph frequency domain spanned by the spectral basis of $\mathcal{G}$. Therefore, signals can be manipulated in the frequency domain to induce different levels of interactions between neighbors in the network; see Section 2.3 for more on graph filters for frequency decomposition.

To elaborate on this concept, consider the eigenvector decomposition of the combinatorial graph Laplacian $\mathbf{L} := \text{diag}(\mathbf{W1}) - \mathbf{W}$ to define the GFT and the associated notion of graph frequencies. With $\mathbf{\Lambda} := \text{diag}(\lambda_1, \ldots, \lambda_N)$ denoting the diagonal matrix of non-negative Laplacian eigenvalues and $\mathbf{V} := [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ the orthonormal matrix of eigenvectors, one can always decompose the symmetric graph Laplacian as $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$.

Definition (Graph Fourier transform): The GFT of $\mathbf{x}$ with respect to the combinatorial graph Laplacian $\mathbf{L}$ is the signal $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_N]^T$ defined as $\tilde{\mathbf{x}} = \mathbf{V}^T\mathbf{x}$. The inverse iGFT of $\tilde{\mathbf{x}}$ is given by $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}$, which is a proper inverse by the orthogonality of $\mathbf{V}$.

The iGFT formula $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}} = \sum_{k=1}^N \tilde{x}_k \mathbf{v}_k$ allows one to synthesize $\mathbf{x}$ as a sum of orthogonal frequency components $\mathbf{v}_k$. The contribution of $\mathbf{v}_k$ to the signal $\mathbf{x}$ is the GFT coefficient $\tilde{x}_k$. The GFT encodes a notion of signal variability over the graph akin to the notion of frequency in Fourier analysis of temporal signals. To understand this analogy, define the total variation of the graph signal $\mathbf{x}$ with respect to the Laplacian $\mathbf{L}$ (also known as Dirichlet energy) as the following quadratic form

$$\text{TV}(\mathbf{x}) := \mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_{i \neq j} W_{ij}(x_i - x_j)^2. \tag{1}$$

The total variation $\text{TV}(\mathbf{x})$ is a smoothness measure, quantifying how much the signal $\mathbf{x}$ changes with respect to the presumption on variability that is encoded by the weights $\mathbf{W}$ [15, 22].

Back to the GFT, consider the total variation of the eigenvectors $\mathbf{v}_k$, which is given by $\text{TV}(\mathbf{v}_k) = \mathbf{v}_k^T\mathbf{L}\mathbf{v}_k = \lambda_k$. It follows that the eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N$ can be viewed as graph frequencies, indicating how the eigenvectors (i.e., frequency components) vary over the graph $\mathcal{G}$. Accordingly, the GFT and iGFT offer a decomposition of the graph signal $\mathbf{x}$ into spectral components that characterize different levels of variability.

2.3 Graph filtering

GFT encodes a notion of variability of the graph signals with respect to $\mathcal{G}$. For graph signal $\mathbf{x}$ with GFT coefficients $\tilde{\mathbf{x}}$, filtering can be done in the frequency domain akin to classical SP of time-varying signals. As discussed in Section 2.2, eigenvalues of the Laplacian correspond to graph frequencies and eigenvectors serve as frequency basis. For instance, a low-pass filter can be designed by isolating the lowest N_L eigenvalues and their corresponding eigenvectors [5]. Define a spectral operation $\tilde{\mathbf{x}}_L = \tilde{\mathbf{H}}_L\tilde{\mathbf{x}}$, where $\tilde{\mathbf{H}}_L = \text{diag}(\tilde{\mathbf{h}}_L)$ and $\tilde{h}_{L,n} = \mathbb{I}\{n < N_L\}$ ($\mathbb{I}\{\cdot\}$ is an indicator function). This is equivalent to the following convolution operation in the vertex domain

$$\mathbf{x}_L = \mathbf{V}\tilde{\mathbf{x}}_L = \mathbf{V}\tilde{\mathbf{H}}_L\tilde{\mathbf{x}} = \mathbf{V}\tilde{\mathbf{H}}_L\mathbf{V}^T\mathbf{x} = \mathbf{H}_L\mathbf{x}, \tag{2}$$

where $\mathbf{H}_L = \mathbf{V}\tilde{\mathbf{H}}_L\mathbf{V}^T$ is the low-pass graph filter. In addition to $\mathbf{H}_L$, a graph band-pass filter $\mathbf{H}_M$ and high-pass filter $\mathbf{H}_H$ can also be defined analogously. In this way, all graph frequencies are decomposed and assigned to each graph filter where $\mathbf{H}_L$ takes the lowest N_L frequencies, $\mathbf{H}_M$ takes the middle $\mathbf{N}_M$ frequencies and $\mathbf{H}_H$ takes the highest N_H frequencies, with $N_L + N_M + N_H = N$. As these filters are mutually exclusive and span all graph frequencies, we can map signals to the spectral domain via the GFT, filter them and use the iGFT to map each frequency component back to the vertex domain. This decomposes the original graph signal into

$$\mathbf{x} = \mathbf{x}_L + \mathbf{x}_M + \mathbf{x}_H, \tag{3}$$

which increases the resolution of the signal by partitioning it into components $\mathbf{x}_L, \mathbf{x}_M, \mathbf{x}_H$ that exhibit low, medium and high variability with respect to the underlying graph topology. In Section 3.3,

we perform this graph frequency decomposition of COVID-19 data to investigate various contagion patterns across US counties.

3 GSP ANALYSIS OF COVID-19 DATA

In this section, a graph frequency analysis is carried out on COVID-19 data. First, we define the graph signals in this context as well as the network graph constructed for the study. Then a thorough frequency analysis is conducted to identify contagion patterns.

3.1 COVID-19 data as graph signals

The raw data[1] is the cumulative number of confirmed COVID-19 cases per 100k residents for each of the $N = 3142$ counties in US from Jan 22 to April 30 (100 days in total); see Fig. 1. Due to the highly skewed distribution of case counts that severely hindered visualization, each one of the 3142 values was assigned to one of five severity levels via range partitioning. Darker color (corresponding to higher severity level) represents higher number of cases.

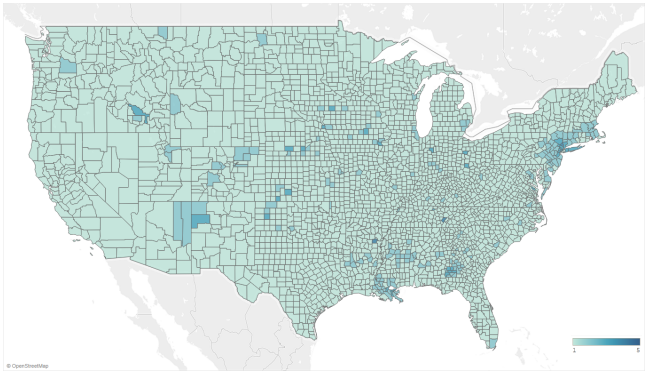


Figure 1: Cumulative number of confirmed COVID-19 cases per 100k residents for each county by Apr 30. For better visualization, Alaska and Hawaii are removed. This and the subsequent map visualizations share the same legend definition and are generated with Tableau.

From the cumulative data in Fig. 1, we compute the number of daily confirmed cases per 100k residents of each county. The daily graph signals can be stacked as columns of the matrix $\mathbf{X} \in \mathbb{R}^{3142\times 100}$, where row i is a time series of length 100 recording the daily confirmed cases in county i normalized by its population size. The time series are depicted in Fig. 2.

From Fig. 1 and Fig. 2 we may generate a rough picture of which counties suffer the most from COVID-19 infections. However, both the snapshot in Fig. 1 and the trends in Fig. 2 offer limited amount of information. There are hidden relationships between the signals of each county that can contribute to the analysis if we carry out a network-analytic study. As the spread of epidemic diseases is often related to the proximity of contacts, it is natural to take into account

[1]COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, https://systems.jhu.edu/research/public-health/ncov/

[2]Note that some signals have negative values because in the raw data the number of total confirmed cases at certain regions may be corrected later, which causes negative daily increases at certain dates. Due to the limited amount of negative values and the lack of ground truth to verify the correction, we kept these values in our analysis.

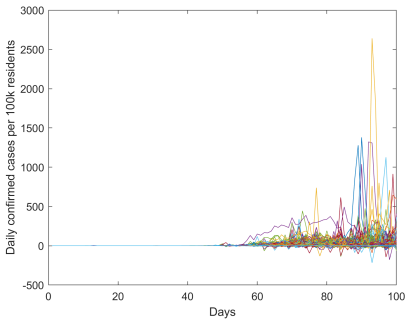


Figure 2: COVID-19 signals of all counties. Each line represents the daily confirmed number of cases per county. Counties exhibit different trends of signals in the forms of smooth lines or sudden spikes[2].

the geodesic distance between counties in building a graph. In this way, we may find different features of graph signals corresponding to the geographic graph and evidences of different patterns of the virus contagion in different locations within the country.

3.2 Graph construction

By extracting the latitude and longitude coordinates of each county, we compute the geodesic distance between counties. A weighted undirected graph $\mathcal{G}$ was constructed with $N = 3142$ counties as nodes and $\binom{3142}{2}$ edges, where edge weights represent the geodesic distance $dist(i,j)$ as an angular arc length on Earth between each pair of counties (i,j). A thresholded Gaussian kernel weghting function was adopted to yield the edge weights in $\mathbf{W}$ [22], namely

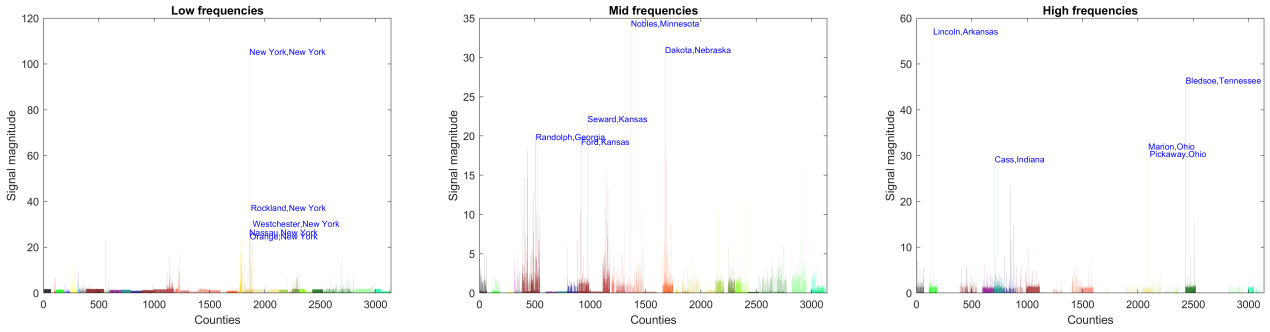
$$W_{ij} = \begin{cases} \exp\left(-\frac{[dist(i,j)]^2}{2\theta^2}\right) & \text{if } dist(i,j) \le k, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where k is a threshold and θ is a bandwidth (scale) parameter. Future work shall consider and include population mobility patterns to help determine the graph structure beyond geodesic connectivity.

3.3 Frequency decomposition of graph signals

With the graph signals $\mathbf{X}$ and graph topology $\mathcal{G}$, we can follow the procedure in Section 2.2 and Section 2.3 and carry out frequency decomposition of the graph signals. First, the Laplacian matrix is formed using $\mathbf{W}$ and its eigenvalues and eigenvectors were computed. A low/band/high-pass filter was constructed by taking the lowest/middle/highest one third of the eigenvalues, respectively. After graph filtering described in Section 2.3, the original graph signal $\mathbf{X}$ is now decomposed into $\{\mathbf{X}_L, \mathbf{X}_M, \mathbf{X}_H\} \in \mathbb{R}^{3142\times 100}$, which represents signal components that change slowly/mildly/rapidly with respect to the underlying geographical graph $\mathcal{G}$ [7]. Following the same procedure in [5, 7, 8], we take the row-wise average of the absolute values in $\mathbf{X}_L, \mathbf{X}_M, \mathbf{X}_H$ and thus obtain three vectors of length 3142 that quantify the signal magnitude per county with respect to their energy occupancy in each spectral band; see Fig. 3.

With GFT, we can partition the signal of each county into three frequency components and see if its signal concentrates more in



(a) (b) (c)

Figure 3: Magnitude of (a) low-pass (b) band-pass (c) high-pass signals of each county. Each color represents one US state. Some counties with high signal magnitudes were labeled out.

a single frequency component. For example, for NYC and nearby counties (yellow spikes before node No. 2000 in Fig. 3a) we can see very high magnitude in low-pass signals while relatively very low magnitude in the rest frequency components. As low-pass signals exhibit low variability with respect to the underlying graph topological connections, the signal patterns at NYC and nearby counties indicate that they share similar signals, which in turn reflects the assumption that regions with high magnitude of low-pass signals have similar number of daily confirmed cases normalized by population size with neighboring counties. On the other hand, high-pass signals are related with high variability regardless of the graph structure. Thus counties with high magnitude of high-pass signals shall present very distinct and abnormal signals compared with counties around them. Analysis in the next section combined multi-resources information to establish the connection between graph frequency (low/high-pass) and contagion spread patterns (i.e. due to across county spread or within county outbreak).

For simplicity, in the following discussion, we use the term **LP regions** to represent counties with high magnitude of low-pass signals, and **HP regions** for counties with high magnitude of high-pass signals. As low-pass signals and high-pass signals correlate with distinct contagion patterns and band-pass signals result from a mixture of them, we spent less effort on the band-pass signals.

3.4 Frequency analysis w.r.t contagion patterns

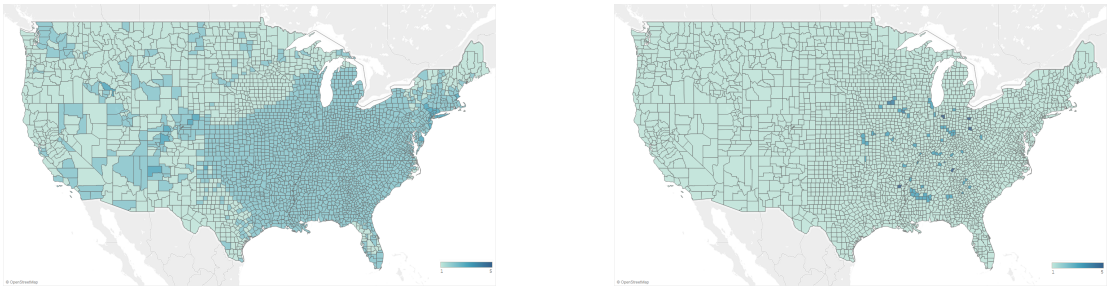
In this section, we further analyzed the correspondence between graph frequency components and the contagion patterns. Signals surviving low-pass filtering shall be smooth above the graph, indicating that low-pass signals shall not vary much between neighboring nodes. On the other hand, high-pass signals shall exhibit significant difference between neighboring nodes. In this analysis, we expect to build a correspondence between low-pass signals and across-county contagion, as one possible result of across-county contagion is that nearby counties will have similar number of confirmed cases, leading to smooth signals on graph. Meanwhile, high-pass signals shall relate to within-county outbreak, which makes the signal of the current county very dissimilar to nearby counties, leading to high-variability signals. In a more specific way, LP regions suffer more from across-county spread and HP regions suffer more from within-county outbreak.

3.4.1 ***Regions rank top in each frequency components.*** Fig. 1 shows a snapshot of the cumulative confirmed cases for each county. From the original data we can only see which county has the most severe situation. Using GFT, more information can be revealed from frequency domain which motivates us to speculate and explain why these counties suffer the most with respect to the connectivity encoded in the geographical graph. Table 1 offers a peek insight of the counties ranked top with the most cumulative confirmed cases, and identifies which frequency component that the signals of the counties concentrate in. Fig. 4 presents the magnitudes of low-pass signals and high-pass signals of all counties on US map.

Table 1: Frequency component assignment to regions ranked top by cummulative confirmed cases

County	State	Frequency component
New York	New York	LP region
Lincoln	Arkansas	HP region
Bledsoe	Tennessee	HP region
Rockland	New York	LP region
Marion	Ohio	HP region
Pickaway	Ohio	HP region
Westchester	New York	LP region
Cass	Indiana	HP region
Nassau	New York	LP region
Passaic	New Jersey	LP region
Louisa	Iowa	HP region
Hudson	New Jersey	LP region
Union	New Jersey	LP region

It is clear from Fig. 4a that counties labeled as LP regions locate mostly from mid and east US, and low-pass signals are smoothly spread on the map. This indicates that these regions share similar signals with nearby counties. On the other hand, counties identified as HP regions are distributively located in the central-east area in Fig. 4b. Further investigations of local news reveal a consistent finding that all these HP regions have concentrated outbreaks at prison, nursing home and food plants etc., to list a few [1, 23, 24]. This makes their signals very different from neighboring counties, thus standing out after high-pass graph filtering.



(a) (b)

Figure 4: Magnitude of (a) low-pass (b) high-pass signals of each county. Higher frequency components tend to be more localized in the vertex domain, whereas the signal energy distribution in the low-pass signal is more spatially diffused.

The analysis above shows what we can learn from the existing data using GFT. Besides, GFT can also tell us what is not shown in the raw data. Among the counties whose signal magnitudes ranking in the top 50 in each frequency components, a further mining was conducted to see if any of these counties did not appear in the top 150 counties ranked by the total cumulative confirmed cases per 100k residents. In the high frequency component, 20 such regions show up as included in Table 2.

Table 2: High-risk counties revealed by frequency analysis

County	State
Jasper	Illinois
Moniteau	Missouri
Adair	Kentucky
Jackson	Kentucky
Muhlenberg	Kentucky
Wabaunsee	Kansas
Kemper	Mississippi
Dallas	Iowa
Dearborn	Indiana
Washington	Oklahoma
Leavenworth	Kansas
Lake	Illinois
Attala	Mississippi
Jackson	West Virginia
Randolph	Illinois
Davidson	Tennessee
Carroll	Mississippi
Franklin	Indiana
Oglethorpe	Georgia
Hopkins	Kentucky

These counties may be ignored within raw data, but via GFT, they stand out to show irregular signals with respect to local geographical connections, indicating that these regions have very different contagion patterns from nearby counties. News search also returns evidences that these regions as well suffer from outbreaks in nursing home, or weirdly very high confirmed cases compared with other counties in the same state, etc. These regions may be regarded as in âĂŸhigh-riskâĂŹ since they are easily ignored with total confirmed cases not high enough to catch attention but relatively high confirmed cases per 100k residents compared with neighboring regions. Thus, these regions, if no cautions and interventions are made, could become new hot spots causing serious outcomes.

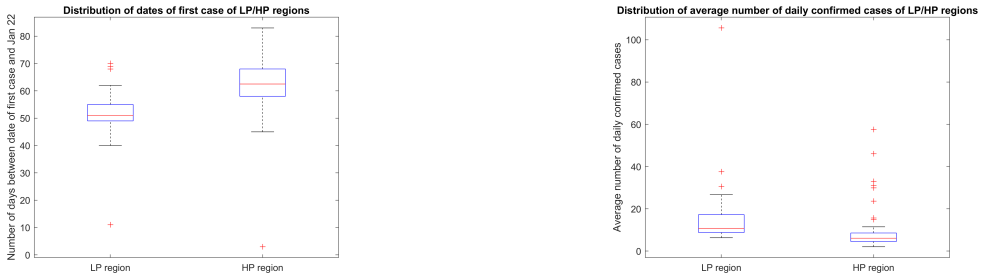
3.4.2 ***Start date of contagion per frequency band.*** One key feature of the investigation of pandemic is the date when the first confirmed case was recorded, as it marks the beginning of the contagion and also reflects the total duration. For the top 50 LP and HP regions ranked by signal magnitude, their start dates were extracted and the corresponding distribution was shown in Fig. 5a. T-test confirms that LP regions have significantly ($p = 0.05$) earlier dates of the first confirmed case, suggesting that in LP regions, the spread of COVID-10 began relatively earlier than HP regions.

3.4.3 ***Daily number of confirmed cases per frequency band.*** Another key feature is the daily confirmed cases that reflect the increase rate of COVID-19 cases. For the same LP and HP regions, the average number of daily confirmed cases was computed as shown in Fig. 5b. Again, LP regions have much higher daily confirmed cases, suggesting relatively faster spread of the virus.

The analysis in Section 3.4 connected the graph frequency components with descriptive findings from raw data such as dates of first cases and daily increase rate. The fact that LP regions have typically earlier start of the contagion and much faster spread, also considering that LP regions are more densely gathered as shown in Fig. 4a, makes it a valid conclusion that LP regions suffer from across-county spread. Since the contagion happened early in these regions, not sufficient prevention procedures were applied, which gradually leads to a relatively fast and wide spread. Meanwhile, HP regions located distributively in the central inland area. The reasons that these regions suffer from COVID-19 are all due to local outbreaks in prison, nursing home and food plants. This also explains that why these counties have high magnitude of high-pass signals as these outbreaks make their signals very dissimilar to nearby regions.

3.5 Temporal analysis of low-pass signals

In the above static analysis, a scalar value was assigned to each county in each frequency component by taking the row average of $\mathbf{X}_L, \mathbf{X}_M, \mathbf{X}_H$ introduced in Section 3.3. Such implementation helps find the link between low/high graph frequency and contagion pattern of across/within-county spread. However, we lose rich temporal information by taking the average of the complete signal



(a) (b)

Figure 5: (a) Distribution of dates of first case and (b) distribution of daily increase rate for LP and HP regions. T-tests with p = 0.05 were carried out in both comparison to prove the statistically difference between LP and HP regions.

timeseries. For example, in the result from the static analysis above, we see that LP regions include few counties on the west coast. Since west coast regions have more cases in the beginning (e.g State of California, Washington), we expect to see the trend that in the beginning, west coast suffers from local across-county virus spread and then the situation on the east coast got much worse.

For temporal analysis, previously in each frequency component, we have a $\mathbf{X}_L, \mathbf{X}_M, \mathbf{X}_H \in \mathbb{R}^{3142\times 100}$, which is the result of the process (GFT, filtering, iGFT). Instead of taking the average of each row, all 100 days were first partitioned into 5 windows with a size of 20 days to represent different temporal stages. The row-wise average of each window was taken. In this case, instead of a single 3142-by-1 vector recording signal magnitude per county in a frequency component, now we have five such vectors. Here, the focus is placed on the low-pass signals since previous section, we have related it with across-county spread. The 5 stages capturing the evolution of magnitude of low-pass signals on US map were visualized in Fig. 6.

We can see the pandemic center migrates from west coast to each coast and spread from a few counties to a much larger area including most counties. As the low-pass signals are related with across-county spread, we can conclude that as time goes by, most US counties are having more cases and the spread of the virus among counties is happening and getting worse.

4 CONCLUSION

In this work, we investigated a powerful tool of Graph Signal Processing and applied it in a timely research on analysis of COVID-19 contagion patterns. Novel information extracted from graph frequency domain leads to new findings regarding the contagion patterns of COVID-19. By establishing the link between graph low/high frequency and the across/within-county contagion spread, we were able to determine the spread patterns of 3142 US counties, and identify 'high-risk' regions which may be ignored in classical temporal signal analysis. Specifically,

(1) GFT helps partition raw graph signals into different graph frequency components with respect to the underlying graph structure. Signals in low/high frequency components show few/rapid variation regarding the connectivity encoded in the underlying graph, thus can be further related with across-county spread and within-county outbreaks.

(2) GFT can help identify regions that are not among the top regions with highest total confirmed cases. This offers new insights beyond raw graph signals to locate vital regions that may have relatively low cases compared to large cities like NYC but may very likely become regional hots spot due to their highly irregular signal (i.e. COVID-19 confirmed cases).

(3) LP regions suffer from early confirmed cases and faster increase rate. Based on the concept of low graph frequency (i.e. signals not vary much w.r.t graph connectivity), this indicates more geographically spread patterns of contagion.

(4) HP regions have relatively late confirmed cases and slower rate. By searching news, we find that these regions suffer from outbreaks in prison, nursing home, food plants etc. This fact perfectly matches the concept of high graph frequency signals w.r.t graph topology, indicating that the virus situation in these regions are due to concentrated within county outbreaks.

Generally speaking, the use of GSP and GFT exploits the frequency domain information embedded in the raw data, and in this specific cases of COVID-19 study, helps identify spread patterns of the virus among US counties. More importantly, the framework used in this work can also mine 'high-risk' regions which require timely intervention operations to cutoff the contagion before it reaches the stage of across-county spread. Future work shall be devoted to construct graphs incorporating mobility patterns such as airline connections to reveal in-depth patterns of the contagion.

ACKNOWLEDGMENTS

Work in this paper was supported by the NSF awards CCF-1750428 and ECCS-1809356.

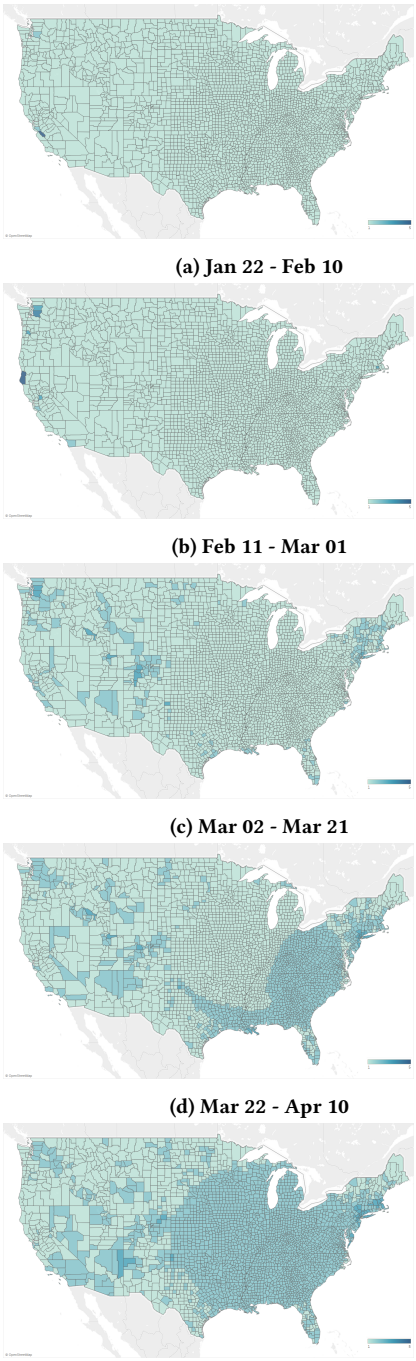
REFERENCES

[1] Max Brantley. 2020. *Coronavirus cases explode at Cummins prison.* https://arktimes.com/arkansas-blog/2020/04/13/coronavirus-cases-explode-at-cummins-prisonl

[2] Mark Crovella and Eric Kolaczyk. 2003. Graph wavelets for spatial traffic analysis. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, Vol. 3. IEEE, 1848–1857.

[3] Xiaowen Dong, Antonio Ortega, Pascal Frossard, and Pierre Vandergheynst. 2013. Inference of mobility patterns via spectral graph wavelets. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3118–3122.

[4] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. 2016. Learning Laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing* 64, 23 (2016), 6160–6173.



(a) Jan 22 - Feb 10

(b) Feb 11 - Mar 01

(c) Mar 02 - Mar 21

(d) Mar 22 - Apr 10

(e) Apr 11 - Apr 30

Figure 6: Temporal evolution of low-pass signal magnitudes. Note that the pandemic center migrates from west coast to each coast and spread from a few counties to a much larger area including most counties.

[5] Leah Goldsberry, Weiyu Huang, Nicholas F Wymbs, Scott T Grafton, Danielle S Bassett, and Alejandro Ribeiro. 2017. Brain signal analytics from graph signal processing perspective. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 851–855.

[6] Christopher J Honey, Olaf Sporns, Leila Cammoun, Xavier Gigandet, Jean-Philippe Thiran, Reto Meuli, and Patric Hagmann. 2009. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences* 106, 6 (2009), 2035–2040.

[7] Weiyu Huang, Thomas AW Bolton, John D Medaglia, Danielle S Bassett, Alejandro Ribeiro, and Dimitri Van De Ville. 2018. Graph signal processing of human brain imaging data. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 980–984.

[8] Weiyu Huang, Thomas AW Bolton, John D Medaglia, Danielle S Bassett, Alejandro Ribeiro, and Dimitri Van De Ville. 2018. A graph signal processing perspective on functional brain imaging. *Proc. IEEE* 106, 5 (2018), 868–885.

[9] Yang Li, Rasoul Shafipour, Gonzalo Mateos, and Zhengwu Zhang. 2019. Mapping brain structural connectivities to functional networks via graph encoder-decoder with interpretable latent embeddings. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.

[10] Yang Li, Rasoul Shafipour, Gonzalo Mateos, and Zhengwu Zhang. 2020. Supervised Graph Representation Learning for Modeling the Relationship between Structural and Functional Brain Connectivity. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9065–9069.

[11] Weiyong Liu, Qi Zhang, Junbo Chen, Rong Xiang, Huijuan Song, Sainan Shu, Ling Chen, Lu Liang, Jiaxin Zhou, Lei You, et al. 2020. Detection of Covid-19 in children in early January 2020 in Wuhan, China. *New England Journal of Medicine* 382, 14 (2020), 1370–1371.

[12] Gonzalo Mateos, Santiago Segarra, Antonio G Marques, and Alejandro Ribeiro. 2019. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine* 36, 3 (2019), 16–43.

[13] Mandeep R Mehra, Sapan S Desai, SreyRam Kuy, Timothy D Henry, and Amit N Patel. 2020. Cardiovascular disease, drug therapy, and mortality in COVID-19. *New England Journal of Medicine* (2020).

[14] Mark Newman. 2018. *Networks*. Oxford university press.

[15] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. 2018. Graph signal processing: Overview, challenges, and applications. *Proc. IEEE* 106, 5 (2018), 808–828.

[16] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[17] Andrea Remuzzi and Giuseppe Remuzzi. 2020. COVID-19 and Italy: what next? *The Lancet* (2020).

[18] Jonas Richiardi, Sophie Achard, Horst Bunke, and Dimitri Van De Ville. 2013. Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal Processing Magazine* 30, 3 (2013), 58–70.

[19] Aliaksei Sandryhaila and Jose MF Moura. 2014. Discrete signal processing on graphs: Frequency analysis. *IEEE Transactions on Signal Processing* 62, 12 (2014), 3042–3054.

[20] Santiago Segarra, Antonio G Marques, Gonzalo Mateos, and Alejandro Ribeiro. 2017. Network topology inference from spectral templates. *IEEE Transactions on Signal and Information Processing over Networks* 3, 3 (2017), 467–483.

[21] Rasoul Shafipour, Raiyan A. Baten, M. Kamrul Hasan, Gourab Ghoshal, Gonzalo Mateos, and M. Ehsan Hoque. 2018. Buildup of speaking skills in an online learning community: a network-analytic exploration. *Palgrave Commun* 4, 63 (2018), 1–10.

[22] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* 30, 3 (2013), 83–98.

[23] Mariah Timms. 2020. *Number of coronavirus cases at Bledsoe County Correctional Complex more than triple.* https://www.tennessean.com/story/news/local/2020/04/23/coronavirus-bledsoe-county-prison-inmates/3003595001/l

[24] Sarah Volpenhein. 2020. *Coronavirus spreads through private Marion prison as testing questions persist.* https://www.marionstar.com/story/news/local/2020/05/24/coronavirus-ohio-prison-spreading-testing/5239050002/

[25] Junqi Wang, Vince D Calhoun, Julia M Stephen, Tony W Wilson, and Yu-ping Wang. 2018. Integration of network topological features and graph Fourier transform for fMRI data analysis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 92–96.

[26] Lu Wang, Feng Vankee Lin, Martin Cole, and Zhengwu Zhang. 2020. Learning Clique Subgraphs in Structural Brain Network Classification with Application to Crystallized Cognition. *BioRxiv* (2020).

[27] June Zhang and José MF Moura. 2014. Diffusion in social networks as SIS epidemics: Beyond full mixing and complete graphs. *IEEE Journal of Selected Topics in Signal Processing* 8, 4 (2014), 537–551.

[28] Yan Zhang, Meng Xiao, Shulan Zhang, Peng Xia, Wei Cao, Wei Jiang, et al. 2020. Coagulopathy and antiphospholipid antibodies in patients with Covid-19. *New England Journal of Medicine* 382, 17 (2020), e38.