HDSR
HARVARD DATA SCIENCE REVIEW

Volume 3 Issue 2
DOI: 10.1162/99608f92.b447c07e
ISSN: 2644-2353

A GEOMETRY-DRIVEN LONGITUDINAL TOPIC MODEL

Yu Wang[†], Conrad Hougen[‡], Brandon Oselio[◇], Walter Dempsey[◇], Alfred Hero[†,‡,♯,*]

[†] Department of Statistics, University of Michigan, Ann Arbor, MI

[‡] Department of EECS, University of Michigan, Ann Arbor, MI

[◇] Department of Biostatistics, University of Michigan, Ann Arbor, MI

[♯] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI

Abstract. A simple and scalable framework for longitudinal analysis of Twitter data is developed that combines latent topic models with computational geometric methods. Dimensionality reduction tools from computational geometry are applied to learn the intrinsic manifold on which the latent, temporal topics reside. Then shortest path distances on the manifold are used to link together these topics. The proposed framework permits visualization of the low-dimensional embedding, which provides clear interpretation of the complex, high-dimensional trajectories that may exist among latent topics. Practical application of the proposed framework is demonstrated through its ability to capture and effectively visualize natural progression of latent COVID-19–related topics learned from Twitter data. Interpretability of the trajectories is achieved by comparing to real-world events. In addition, the framework permits study of spatial variation in Twitter behavior for learned topics. The analysis demonstrates that the proposed framework is able to capture granular-level impact of COVID-19 on public discussions. We end by arguing that Twitter data, when analyzed within the proposed framework, can serve as a valuable supplementary data stream for COVID-related studies.

Keywords: topic models, latent Dirichlet allocation, computational geometry, social media analysis, COVID-19

Media Summary

As technology advances, there is an ever-increasing demand to acquire, analyze, and generate complex unstructured data. These necessarily large data sets must be amenable to efficient processing, analysis, and implementation in a variety of settings such as multidimensional modeling and high-resolution visualization. Traditionally, machine learning approaches for these problems relied on user-defined heuristics to extract features encoding structural information about the data. More recently, there has been an increasing interest in developing geometry-based methods that automatically learn signals from these complex and large data sets. This article introduces a flexible and scalable framework to extract patterns from time-evolving, complex, high-dimensional data

[*]hero@umich.edu

associated with documents and social media archives. Our framework combines latent Dirichlet allocation (LDA) and computational geometric representations to time-align thematic information extracted from each time slice of the data. Under this framework, we achieve two goals. First, we are able to apply machine learning and data analytic tools that preserve geometric information intrinsic to the structure of the data. Second, we introduce a modular approach to a complex statistical modeling problem, separating topic modeling and temporal dynamics. The utility of this framework is demonstrated on Twitter data collected during the COVID-19 outbreak in the United States. Using geotagged tweets, the geometry-driven longitudinal model reveals temporal patterns in the evolution of topics of conversation as they are affected by the outbreak.

1. Introduction

The continued digitization of public discourse in news feeds, books, scientific reports, social media, blogs, microblogs, and web pages creates opportunities to discover meaningful patterns and trends of public opinion. Methods of probabilistic topic modeling have been used to extract such patterns using a suite of algorithms that aim to automatically discover and annotate large collections of documents with thematic labels (Blei, 2012). Topic modeling algorithms are computational methods that manipulate word frequencies in document corpora to discover the themes that run through them, quantify how those themes are connected to each other, and how they change over time.

1.1. Probabilistic topic models and computational geometry.

A probabilistic topic model that has seen success in many applications is the latent Dirichlet allocation (LDA) model (Blei et al., 2003), which uses a latent topic model to extract thematic information from document corpora to infer an underlying generative process that explains hidden relationships among documents. Many real-world document corpora, however, have complex structure and include temporal information that is ignored by traditional LDA models. For example, discussions of COVID-19 on Twitter between February and May 2020 involve the emergence, evolution, and extinction of multiple topics over time. Moreover, tweets are short bursts composed in micro-text (Ellen, 2011), which traditional LDA models struggle to model effectively.

Extensions of the standard LDA have been proposed to learn latent topics in the context of complex structure and temporal information. An early modeling strategy is to assume a temporally Markovian relationship where the state of the process at time $t+1$ is independent of past history given the state at time t. Blei and Lafferty (2006) proposed the dynamic topic model (DTM) for modeling time-varying topics, where the topical-alignment over time is captured by a Kalman filter procedure. Further improvements have been in various directions, including: (1) relaxation of the Markov assumption, as discussed by X. Wang and McCallum (2006), who introduced a non-Markov continuous-time model called the topics-over-time (TOT) model, capturing temporal changes in the occurrence of the topics themselves, and (2) circumvent of time discretization, as proposed by C. Wang et al. (2008) that improved the DTM using a continuous time variant, called cDTM, formulated on Brownian motion to model the latent topics in a longitudinal collection of documents. These approaches rely on spatiotemporally coupled stochastic processes for modeling the evolution of topics over time. Such integrated models employ a global joint parameterization of time evolution and word co-occurrence, producing a unified generative probabilistic model for both temporal and topical dimensions.

However, global parameterized DTMs have several deficiencies that motivate the model proposed in this article. The main issue is that global parameterization can increase the computational

complexity of parametric inference. C. Wang et al. (2008) and Blei and Lafferty (2006) argued that applying Gibbs sampling to perform inference on DTMs is more difficult than on static models, principally due to the nonconjugacy of the Gaussian and multinomial distributions. As an alternative, they proposed the use of inexact variational methods, in particular, variational Kalman filtering and variational wavelet regression, for inference. These approximate inference procedures face two issues: 1) they usually involve assumptions on the correlation structures among latent variables, for example, mean-field, which undermines uncertainty quantification; 2) the resulting optimization problems are usually nonconvex, which means that the approximate posterior distribution found might only be locally optimal–trapping the topic parameters in a neighborhood of a local optima. An additional issue is that posterior inference via variational approximation usually relies on batch algorithms that need to scan the full data set before each update of the model. This increases the computational burden, especially for long time sequences, and parallel computing cannot be easily exploited (Bhadury et al., 2016). Such issues can lead to numerical instability and lack of interpretability of the model predictions.

Rather than jointly modeling word co-occurrence and the temporal dynamics, there exist alternatives that adopt simpler analysis strategies that motivate our proposed approach. Most of these approaches to nonglobal modeling involve fitting a local time-unaware topic model to predivided discrete time slices of data, and then examining the topic distributions in each time-slice in order to assemble topic trends that connect related topics (Cui et al., 2011; Griffiths & Steyvers, 2004; Malik et al., 2013; X. Wang et al., 2005). A difficulty with these approaches is that aligning the topics from each time slice can be challenging, even though several strategies have been proposed. Malik et al. (2013) proposed a framework to connect every pair of topics from adjacent time slices whose similarity, measured by the cosine metric, exceeds a certain threshold. Cui et al. (2011) used a semiparametric clustering algorithm to identify similar topics at adjacent time slices. However, these approaches suffer from an inherent inflexibility in modeling diverse dynamical structures that exist in a potentially large collection of temporal topic sequences. Such methods are developed to model and visualize specific, and relatively rare, types of temporal dynamics and are often not able to capture all types of variations, for example, anomalies, bifurcations, emergence, convergence, and divergence.

We propose a flexible and scalable computational geometry framework that remedies the above mentioned issues and complements the existing methods in the dynamic topic modeling toolbox. Specifically, in this article a time-evolving topic model is introduced that uses a local LDA-type model for discrete time slices of collections of documents, and a geometric proximity model to align the topics from time to time. In contrast to global parametric dynamic latent variable approaches to summarizing time-evolving unstructured texts, our framework offers a wrapper for a suite of tools. The proposed wrapper framework has the flexibility to allow any particular topic model to be applied locally to each time slice of documents. It then implements a fast and scalable shortest path algorithm to stitch together the locally learned LDA topics into an integrated collection of temporal topic trends.

To facilitate visualization and interpretation of the learned topic trends, an emphasis of this article, the proposed framework also implements a recent geometric embedding method called PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) that projects the high-dimensional word distributions representing latent topics to lower dimensional coordinates. The PHATE embedding has been shown to preserve the intrinsic geometry of high-dimensional time-varying data (Moon et al., 2019), which provides a clear and intuitive visualization of any

progressive structure that exists among the topics. We note that similar computational geometric representations of data have been used in unsupervised, semisupervised, and supervised learning, both as principal learning models and as supplementary regularizers of other models. In manifold learning, geometric affinity (or distance) between data points drives dimensionality reduction (Donoho & Grimes, 2003; Tenenbaum et al., 2000) and dimensionality estimation methods (Costa & Hero, 2006). Several deep learning architectures, like the deep k-nearest neighbors (Papernot & McDaniel, 2018, DkNN), use interpoint distances and the kNN classifier to induce interlayer representational continuity and robustness against adversarial attacks. Semisupervised classification approaches adopt geometric measures over reproducing kernel Hilbert space (RKHS) to associate unlabeled data with labeled data in geometry-regularized empirical loss frameworks (Belkin & Niyogi, 2004). Geometry is the driver for many missing data models, for example, synthetic minority oversampling technique (Chawla et al., 2002, SMOTE) and more generally, nearest neighbor interpolations.

We point out that dimensionality reduction is the basis for latent semantic analysis (LSA) in computational linguistics. In particular, Doxas et al. (2010) had a similar objective to ours, to explore temporal evolution of discourse, but in long text with labeled corpora. The authors constructed semantic spaces for various corpora, and then calculated the intrinsic dimensionality of the paragraph trajectories through these corpora. The work focuses on investigating the intrinsic dimension of the trajectories and they used LSA to construct representations of the texts. However, they did not address the topic alignment or trajectory clustering problems for which our PHATE and Hellinger shortest path framework is designed.

1.2. **Application to COVID-19 using Twitter data.** Enabled by our proposed longitudinal dynamic topic model, we leverage recent activity on social media to understand the impact of the COVID-19 pandemic and, in particular, its impact on social discourse. The utilization of novel data sources is vital, as the current data landscape for understanding the pandemic remains imperfect. For example, public databases maintained by Johns Hopkins University (https://bit.ly/2UqFSuA) and *The New York Times* (https://bit.ly/2vUHfrK) provide incoming county-level information of confirmed cases and deaths. Unfortunately, these data streams are of limited utility due to limited testing capacity and selection bias (Dempsey, 2020). The public health community requires auxiliary sources of information to improve national and local health policy decisions. A critical question is whether there are complementary data streams that may be leveraged to better understand the COVID-19 pandemic in the United States. Social media platforms, such as Twitter, Reddit, Facebook, and so on, are examples of such data streams. These platforms generate high resolution spatiotemporal data sets that concern public opinions on various societal issues, including health care, government decisions, and politics, all of which could be highly relevant to understanding the impact of COVID-19.

Although use of these novel data streams create new challenges due to limitations such as high noise level, high volume, and selection bias, many recent efforts have explored social media data as a complementary source to traditional health care data and applied topic models to understand public concerns toward COVID-19 (Boon-Itt & Skunkan, 2020; Doogan et al., 2020; Jang et al., 2020; Stokes et al., 2020; Xue et al., 2020), as well as related socioeconomic issues (Liu et al., 2020; Sha et al., 2020; Su et al., 2021). Here, we extract information from Twitter, a particularly popular social media platform, and focus on studying its spatiotemporal behaviors that are believed to be affected by COVID-19. We use subsamples of tweets generated from February 15, 2020, to May 15, 2020, a period over which a large volume of COVID-19- related tweets occurred. An extended analysis of data

collected from May to August 2020 can be found on https://github.com/ywa136/twitter-covid-topics. We apply our temporal topic modeling framework to discover sets of COVID-19-related latent topics that impact public discourse.

We highlight key contributions of this article:

- A modular framework that provides a wrapper for a suite of tools for interpretation and visualization of temporal topic models.
- A new approach for aligning independently learned topic models over time based on computational geometry.
- A scheme for visualizing and understanding temporal structures of the aligned topics via manifold learning.

The remainder of the article is organized as follows: Section 2 introduces the methods and tools that have been applied in our analysis framework. Section 3 presents numerical results and visualizations with several case studies. Section 4 gives some concluding remarks.

2. Methods

In this section, we discuss the building blocks for the proposed framework: Section 2.1 briefly describes the LDA model and its variants for dealing with micro-text; Section 2.2 introduces two key components of the framework for propagating and associating topics over time; and Section 2.3 reviews and applies a dimension reduction technique to visualize the temporal trajectories of the evolving topics.

2.1. LDA for micro-text documents.

Since the literature in probabilistic topic models and their dynamic variants is enormous (see Blei (2012) for a survey), we focus our discussion on the LDA (Blei et al., 2003), which is the building block for all other algorithms targeting similar applications. A graphical model representing its generating process is presented in Appendix A. The idea of LDA is: from a collection of documents (each composed of set of words $w_{d,n}$), one is able to infer the per-word topic assignment $z_{d,n}$, the per-document topic proportions θ_d, and the per-corpus topic distributions β_k, through a joint posterior distribution $p(\theta, z, \beta|w)$. Numerous inference algorithms are developed to handle data at scale, for example, variational methods (Blei et al., 2003; Hoffman et al., 2013; Mimno et al., 2012; Srivastava & Sutton, 2017; Teh et al., 2008), expectation propagation (Minka & Lafferty, 2002), collapsed Gibbs sampling (Griffiths & Steyvers, 2002), distributed sampling (Ahmed et al., 2013; Newman et al., 2008), and spectral methods (Anandkumar et al., 2014; Arora et al., 2012). The posterior expectations can then be used to perform the task at hand: information retrieval, document similarity, exploration, and so on.

The standard LDA, however, may not work well with micro-text like tweets. In particular, each tweet usually concentrates on a single topic, and it is not reasonable to consider one tweet as a document in the traditional sense as there is limited data (e.g., word co-occurrences) from which the latent topics can be learned. To overcome this "data sparsity" issue, efforts have been made along on three major directions (Qiang et al., 2020): 1) methods predicated on the assumption that each text (e.g., tweet) is sampled from only one latent topic; 2) methods utilizing global (i.e., the whole corpus) word co-occurrences structures; 3) methods based on aggregation/pooling of texts into 'pseudo-documents' prior to topic inference.

In this article, we apply the Twitter LDA model (Zhao et al., 2011, T-LDA), for modeling topics at each time slice. T-LDA can be categorized along the directions 1) and 3) mentioned above. But we note that the proposed framework works with any topic model that outputs word distributions

representing learned latent topics. We selected T-LDA since it has been widely used in many related applications, including aspect mining (Yang et al., 2016), user modeling (Qiu et al., 2013), and bursty topic detection (Diao et al., 2012). The generative model underlying T-LDA assumes that there are K topics in the Tweets, each represented by a word distribution, denoted as β_k for topic k and β_B for background words. Let θ_u denote the topic assignment distribution for user u. Let π denote a Bernoulli distribution that governs the choice between background words and topic words. The generating process for a tweet is as follows: a user first chooses a topic based on its user-specific topic assignment distribution. Then the user chooses a bag of words one-by-one based on the chosen topic or the background model. The generation process is summarized in Algorithm 1, and a plate notation comparison between the T-LDA and standard LDA is included in Appendix A. Similarly to a standard LDA algorithm, parameters in each multinomial distribution are governed by symmetric Dirichlet priors. The model inference can be performed using collapsed Gibbs sampling (code available at https://github.com/minghui/Twitter-LDA). Due to space limitations we leave out derivation details and sampling formulas. More details on the implementation can be found in Appendix A.

Algorithm 1 Generating process for T-LDA

Input: Constants η, γ
  Draw $\beta_B \sim \text{Dir}(\eta)$, $\pi \sim \text{Dir}(\gamma)$
  for topic $k = 1, \ldots, K$ **do**
    Draw $\beta_k \sim \text{Dir}(\eta)$
  end for
  for user $u = 1, \ldots, U$ **do**
    Draw $\theta_u \sim \text{Dir}(\alpha)$
    for Tweet $s = 1, \ldots, S_u$ **do**
      Draw $z_{u,s} \sim \text{Multi}(\theta_u)$
      for word $n = 1, \ldots, N_{u,s}$ **do**
        Draw $y_{u,s,n} \sim \text{Multi}(\pi)$
        if $y_{u,s,n} = 0$ **then**
          Draw $w_{u,s,n} \sim \text{Multi}(\beta_B)$
        else
          Draw $w_{u,s,n} \sim \text{Multi}(\beta_{z_{u,s}})$
        end if
      end for
    end for
  end for

2.2. **Time evolution of topics and shortest paths.** Instead of explicitly building the temporal structures into the model as in the globally parameterized DMT and its variants, we propose a two-stage approach: 1) construct a new corpus at each time point via subsampling the documents and independently fitting a topic model to each new corpus; 2) link each of these time points together via shortest distance paths through topics.

Temporal smoothing by subsampling. A subsample of tweets is constructed at each time point by conditional sampling of all the tweets with a sampling distribution that is inversely proportional to

the temporal proximity of the tweet. This produces subsamples that are local mixtures of tweets at nearby time points, accomplishing a degree of temporal smoothing prior to topic analysis.

To clarify the subsampling procedure, we give a simple example. Assume that the corpus composed of five tweets per day over a 5-day period. We write d_t as the set of five tweets on day t. On the first day ($t = 1$), exponential weights are computed $w_1 = \{1.000, 0.7500, 0.5625, 0.4219, 0.3164\}$ and normalized by their sum, defining the sampling distribution used to construct the subsample. That is, at time 1, the subsample consists of 100% of all the tweets from day 1, 75% of the tweets from day 2, $75\%^2 = 56.25\%$ of the tweets from day 3 and so on. This subsample is denoted $C_1 = \{s_1, \ldots, s_5\}$. To construct the subsample on day 2, we condition on the subsample C_1 on day 1 and we construct exponential sampling weights for day 2 of the form $w_2 = \{0.7500, 1.000, 0.7500, 0.5625, 0.4219\}$. We combine the subsample C_1 and weights w_2 to construct the subsample on day 2, denoted C_2, by either randomly removing extra tweets if the sampling weights on a particular day i decreased or randomly adding more tweets from $d_i - s_i$ if the weights on a particular day i increased. This algorithm ensures a (tuneable) degree of smoothness over the subsamples generated at each time in the sense that subsamples which are close in time are likely to contain similar tweets. The procedure is illustrated in Figure 1. Specifically, using notations developed above, $C_1 = \{s_1, \ldots, s_5\}$ and $C_2 = \{s_1 - \text{Tweet 3}, s_2 + \text{Tweet 7}, s_3 + \text{Tweet 15}, s_4 + \text{Tweet 20}, s_5 + \text{Tweet 24}\}$, where Tweet 3, Tweet 7, Tweet 15, Tweet 20, and Tweet 24 were randomly chosen.

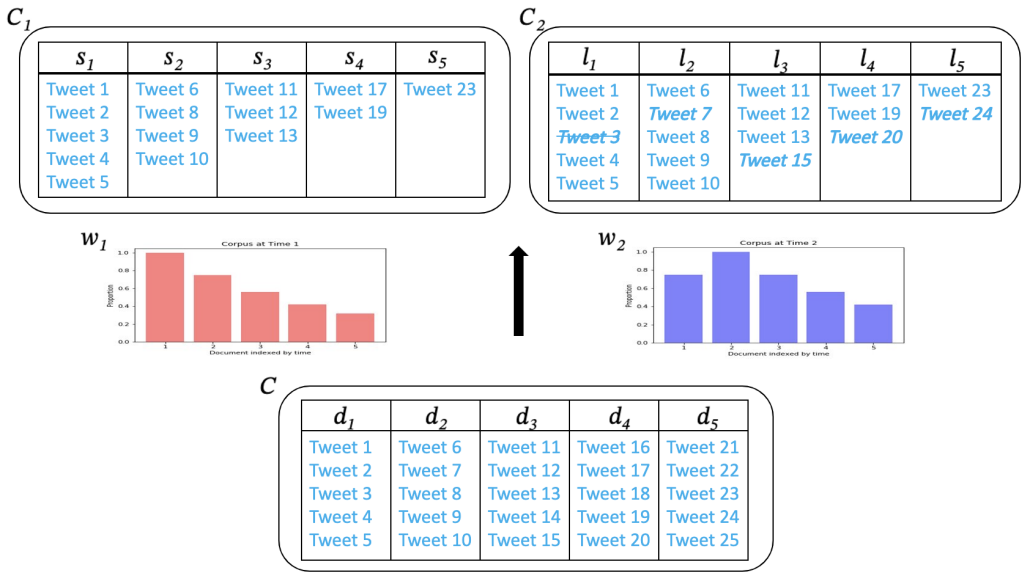


Figure 1. Conditional subsampling procedure using a hypothetical corpus composed of five documents each containing five tweets. For example, $C = \{d_1, \ldots, d_5\}$, d_1 aggregates tweets from day 1, d_2 aggregates tweets from day 2, and so on. The subsampling weights for each document are shown in the bar plots and are exponentially decaying with a factor of 0.75, centered at day 1 (left, w_1) and day 2 (right, w_2), respectively. Each newly generated corpus is a proportionally weighted random sample and a realization of these samples are shown in the tables (C_1 and C_2). Note that the two corpora differ only by those highlighted and italicized tweets.

After constructing the temporally smoothed corpus, any topic modeling algorithm, such as LDA, can be independently applied to each of the extracted corpora, using either the same set of

parameters across all LDA runs, or seeding the next LDA with estimated parameter values (e.g., the posterior mean of the Markov chain Monte Carlo samples) from the current LDA, as described in Song et al. (2005).

Our proposed temporal smoothing technique is similar to smoothing approaches introduced in spatiotemporal statistics. For example, in time-series analysis the idea of exponential smoothing was proposed in the late 1950s (Brown, 1959; Holt, 2004; Winters, 1960), and has motivated some of the most successful forecasting methods. Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations become older. As another example, kriging (Krige, 1951) or Gaussian process regression is a widely used method of interpolation in geostatistics. The basic idea of kriging is to predict the value of a function at a given point by computing a weighted average of the known values of the function in the neighboring points (Cressie, 2015). Lastly, in nonparametric regression analysis, the locally estimated scatterplot smoothing (LOESS) is a widely used method that combines multiple regression models in a k-nearest neighbor based framework. Particularly, at each point in a data set a low-degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is being estimated. The subsets used for each of these polynomial fits are determined by a nearest neighbor algorithm. A user-specified 'bandwidth' or smoothing parameter determines how much of the data is used to fit each local polynomial. This smoothing parameter is the fraction of the total number of data points that are used in each local fit.

Dissimilarity between topic word distributions. After applying local LDA to each of the time localized subsamples of the smoothed corpus, we stitch together the local LDA results. The alignment of topics with different time stamps is accomplished by creating a weighted graph connecting all pairs of topics where the edge weights are a measure of topic similarity, to be described below. Assume that each local model generates K topics, resulting in a total of $K \times T$ topics across T time points. A weighted adjacency matrix is constructed from the similarities between $\binom{K\times T}{2}$ topic pairs. The similarities between topics will allow the alignment algorithm to relate topics together across time and enable us to track topic evolution.

As each topic is characterized by the LDA word distribution, any metric that measures dissimilarity between discrete distributions could be used to construct a similarity measure. It is well known that the Euclidean distance is not well adapted to measuring dissimilarity between probability distributions (Amari, 2012). As an alternative, we propose using the Hellinger metric on the space of distributions, which we justify as follows. The LDA word distribution is conditionally multinomial and it lies on a statistical manifold called an *information geometry*, that is endowed with a natural distance metric, called the Fisher-Rao Riemannian metric. Unlike the Euclidean metric, this Riemannian metric characterizes intrinsic minimal (geodesic) distances between multinomial distributions and it depends on the Fisher information matrix $[\mathcal{I}(\theta)]$, θ is the multinomial probability vector. Carter et al. (2009) showed that this metric can be well approximated by the Hellinger distance between multinomial distributions. The Hellinger distance between discrete probability distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$ is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{n=1}^{N} (\sqrt{p_n} - \sqrt{q_n})^2}, \quad 0 \leq H(\cdot, \cdot) \leq 1.$$

The major advantages of the Hellinger distance are threefold: 1) it defines a true metric for probability distributions, as compared to, for example, the Kullback-Leibler divergence; 2) it is

computationally simple, as compared to the Wasserstein distance; 3) and it is a special case of the f-divergence, which enjoys many geometric properties and has been used in many statistical applications. For example, Liese (2012) showed that f-divergence can be viewed as the integrated Bayes risk in hypothesis testing where the integral is with respect to a distribution on the prior; Nguyen et al. (2009) linked f-divergence to the achievable accuracy in binary classification problems; Jager and Wellner (2007) used a subclass of f-divergences for goodness of fit testing; Rao (1995) demonstrated the advantages of the Hellinger metric for graphical representations of contingency table data; Srivastava and Klassen (2016) adopted the Hellinger distance to measure distances between functional and shape data; Shemyakin (2014) showed the connection of the Hellinger distance to Hellinger information, which is useful in nonregular statistical models when Fisher information is not available; and finally, Servidea and Meng (2006) derived an identity between the Hellinger derivative and the Fisher information that is useful for studying the interplay between statistical physics and statistical computation.

We note that in previous work on aligning topics the $L2$ or cosine distance is commonly applied (Chuang et al., 2013; Chuang et al., 2015; Yuan et al., 2018). As discussed above, these distances are practically and theoretically deficient for aligning distributions. A simulation study is presented in Appendix E that compares use of these distances to the Hellinger distance, showing that the latter better preserves topic trend coherence.

Nearest neighbor graphs and shortest paths. We use the topic graph with Hellinger weights to identify natural progressions from one topic to another over time. We use Dijkstra shortest paths through a nearest neighbor subgraph to identify these progressions. These paths can be interpreted as trajectories of public discourse on the topics identified. This is of interest because we want to understand how the conversation around COVID-19 evolves over time. Shortest path analysis allows us to do this with minimal assumptions on the data. In particular, we do not assume or further encourage temporal smoothness in the data beyond the temporally smoothed corpora described in Section 2.2.

Due to the noisy nature of Twitter data and the wide range of topics, we pay special attention to local neighborhoods of data points. Hence, instead of working with a fully connected graph induced by the full $N \times N$ Hellinger distance matrix of pairwise distances between topics, we build a k-nearest neighbor graph from it. Natural evolution of a topic over time can then be inferred by finding a shortest path of topics on the weighted k-nearest graph, where Hellinger distances represent edge weights. Here, a shortest path is a path between two vertices (i.e., two topics) in a weighted graph such that the total sum of edges weights is minimum, and can be computed efficiently using, for example, Dijkstra's algorithm. The approach of using neighborhood graphs for estimating the intrinsic geometry of a data manifold is justifiable both empirically and theoretically. In manifold learning similar ideas are used to reconstruct lower dimensional geometry from data. For example, the isometric feature mapping (Tenenbaum et al., 2000, ISOMAP) extends metric multidimensional scaling (MDS) by replacing the matrix of Euclidean distances in MDS with the matrix of shortest path distances between pairs of vertices in the Euclidean k nearest neighbor graph. Using such embedding, ISOMAP is able determine lower dimensional structure in high-dimensional data and capture perceptually natural but highly nonlinear "morphs" of the corresponding high-dimensional observations (see figure 4 in Tenenbaum et al. (2000)). Such shortest path analysis is supported by substantial theory (Bernstein et al., 2000; Costa & Hero, 2006; Hwang et al., 2016). Under the assumption that the data points are random realizations on a compact and smooth Riemannian

manifold, as the number of data points grows, the shortest paths over the k nearest neighborhood graph converge to the true geodesic distance along the manifold.

In the context of our topic alignment application, this theory suggests that the analogous Hellinger shortest paths should be able to achieve alignment if the empirical LDA word distributions can themselves be interpreted as random draws from an underlying distribution that varies continuously and smoothly over time along a statistical manifold. To illustrate, Figure 2 demonstrates how a COVID-19-related topic learned from the corpus on February 15, 2020 (far left), evolves to a COVID-19-related health care–focused topic learned from the corpus on May 15, 2020 (far right). The top row in the figure was constructed by computing the shortest Hellinger distance path on a 10-nearest neighbor graph, whereas the bottom row was constructed using the full graph. As expected, the shortest path on the neighborhood graph captures perceptually natural but highly nonlinear 'morphs' of the corresponding high-dimensional word distributions by transforming them approximately along geodesic paths. On the other hand, the shortest path on the full graph connects the two observations through a sequence of apparently unrelated and nonintuitive topics. In Appendix F, we compare the proposed Hellinger shortest path topic alignment method with TopicFlow (Malik et al., 2013), a common method for topic alignment that uses local matching and Euclidean distances.

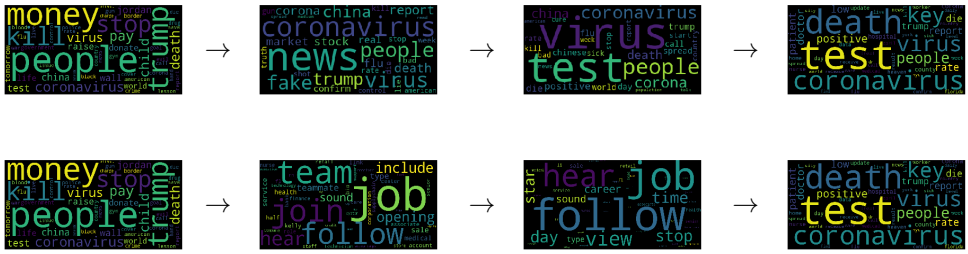


Figure 2. Evolution along the Hellinger shortest paths of a COVID-19 topic on February 15, 2020, to a COVID-19 topic on May 15, 2020. The paths are computed on a 10-nearest neighbor graph (top) and a fully connected graph (bottom). Each word cloud image represents a topic at a particular time, showing the word distribution encoded by font size (only the top 30 words in each topic are shown). The middle two word clouds represent two intermediate topics on the respective paths and illustrate the benefit of using the k nearest neighbor graph. The middle two topics on the top row seem naturally connected to the beginning and the end topics, in contrast to the bottom row.

The choice of k for the neighborhood graph affects the approximation to the Hellinger geodesic path: choosing a k that is too large creates short circuits in the graph, resulting in a noisy path like the bottom row of Figure 2; choosing a k that is too small results in a graph that is disconnected for which there might not exist a path between two points of interest. The problem of selecting an optimal value of k remains open, although several computational data-driven approaches have been proposed for ISOMAP (Gao & Liang, 2011; Samko et al., 2006; Tenenbaum et al., 2000). Here we use $k = 10$, which exceeds the connectivity threshold, to induce the most natural approximation to the true geodesic path between topics of interest. In Appendix H we establish that our results are robust to perturbations around this value of k.

We also note that the Hellinger shortest paths may differ in length, which is the number of topics that they connect over time. This variation is due to the occasional time skips in the path that occur when the shortest path algorithm does not find an adequate match between topics at successive time points. Such skipping can occur when a topic thread wanes temporarily, merges with another thread, or dies. In Appendix B we provide statistics on the occurrences of skips for a subset of paths.

2.3. **Interpretation and visualization of topic trends via low-dimensional embedding.** In LDA each latent topic is represented by a vector that lies on a simplex that constitutes a discrete probability distribution over words. This vector could be very high dimensional depending on the size of the vocabulary. Dimensionality reduction methods are useful for visualization, exploration, and interpretation of such high-dimensional data, as they enable extraction of critical information in the data while discarding noise. Many popular methods are available for visualizing high dimensional data, such as principle component analysis (PCA), MDS, uniform manifold approximation and projection (McInnes et al., 2018, UMAP), and t-distributed stochastic neighbor embedding (Van der Maaten & Hinton, 2008, t-SNE). These methods use spectral decompositions of the pairwise distance matrix to embed the data into lower dimension. PHATE (Moon et al., 2019), on the other hand, is designed to visualize high-dimensional time-varying data. As demonstrated by the authors, it is capable of uncovering hidden low-dimensional embedded temporal progression and branching structure.

Here we embed the estimated LDA word distributions into lower dimensions using a novel application of PHATE to the Hellinger distance matrix. For details on our implementation, see Appendix D. Here, using simulated data, we demonstrate the power of the proposed PHATE-Hellinger embedding for visualization of temporal evolution patterns as compared to other embedding methods. Specifically, we simulate 10 trajectories of 100-dimensional probability vectors using the model

$$X_t^j | X_{t-1}^j \sim \mathcal{N}_{100}(X_{t-1}^j, \sigma_j^2 I)$$

and

$$P_{t,i}^j = \frac{\exp(X_{t,i}^j)}{\sum_{i=1}^p \exp(X_{t,i}^j)}, \qquad i = 1, \ldots, 100$$

for $j = 1, \ldots, 10$ and $t = 0, \ldots, 99$. Each trajectory starts at the same point $X_0 \in \mathbb{R}^{100}$ and differs from realization to realization depending on σ_j. We project all 1000 vectors onto a hypersphere by computing the element-wise square root of each probability vector and using the mapping $P_{t,1} + \cdots + P_{t,100} = 1 \Leftrightarrow (\sqrt{P_{t,1}})^2 + \cdots + (\sqrt{P_{t,100}})^2 = 1^2$. Figure 3 presents the 2D embeddings of this synthetic dataset using PCA on the Euclidean distance matrix, and t-SNE, UMAP, and PHATE on the Hellinger distance matrix. Observe that, among all methods, only PHATE correctly captures the temporal progressions as distinct trajectories originating from a common initial point X_0. Additional simulation studies comparing PCA, t-SNE, UMAP, and PHATE with and without Hellinger distance are included in the Appendix E. In particular, the benefit of using the Hellinger distance instead of the Euclidean distance is demonstrated.

3. Results and Discussion

The entire pipeline for our analysis is described in Algorithm 2. The implementation requires setting several hyperparameters. The temporal smoothing parameter was selected as $\gamma = 0.75$, which corresponds to smoothing approximately one month of tweets into the current time point, in inverse proportion to temporal proximity. In Appendix H we show relative insensitivity to the

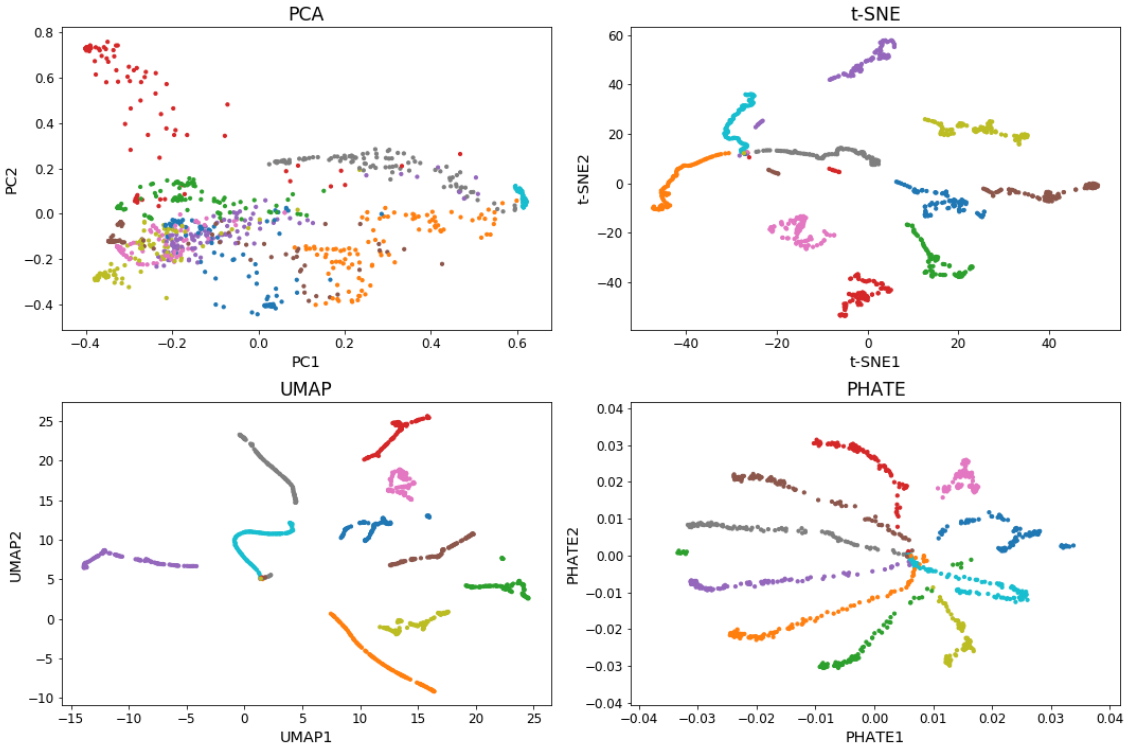


Figure 3. Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) for dimensionality reduction. The methods are applied to 2D embedding of simulated 10 trajectories (identified by color) of 100-dimensional probability vectors, all originating from a common initial point. Except for PCA, all these methods are applied to the matrix of Hellinger distances. Only PHATE correctly captures the temporal progressions as distinct trajectories originating from a common initial point.

choice of this smoothing parameter. The parameter for the number of topics was set to $K = 50$ at every time point. Although not explored here, one could also vary K over time, for example, selected by minimizing perplexities or Bayesian information criteria (BIC) scores at each time (see Appendix H for a further discussion).

3.1. **Data preparation.** We use data from the Twitter Decahose Stream API (https://developer.twitter.com/en/docs/Tweets/sample-realtime/overview/decahose). The Decahose includes a random sample of $\sim 10\%$ of the tweets from each day, resulting in a sample of $300 - 500$ millions of tweets per day. Among all tweets that are sampled, between $\sim 0.1\%$ and 0.5% (see Appendix G for details) of them contain geographic location information, called geotags, that localize the tweet to within a neighborhood of the user's location when the tweet was generated. Note that Twitter's precise location service that uses GPS information has been turned off by default (https://twitter.com/TwitterSupport/status/1141039841993355264). We consider here the more common Twitter "Place" object that consists of 4 longitude-latitude coordinates that define the general area from which the user is posting the tweet (https://developer.twitter.com/en/docs/tutorials/filtering-Tweets-by-location).

Algorithm 2 Procedure for longitudinal analysis of Twitter data.

Input: Raw Twitter data

1: Preprocess Twitter data and organize tweets into a temporally smoothed corpus as described in Section 2.2, with smoothing parameter $\gamma = 0.75$.

2: Apply T-LDA described in Algorithm 1 independently to each corpus with $K = 50$ topics. This results in $4,500$ word distributions.

3: Compute all pairwise Hellinger distances for $4,500$ word distributions.

4: Compute:

a: k-nearest neighbor graph with $k = 10$ from the $4,500$-by-$4,500$ Hellinger matrix and find the shortest path of interest on the neighborhood graph using the Djikstra algorithm.

b: PHATE embedding of $4,500$ high-dimensional points in 2D and 3D.

Output: Shortest paths and PHATE coordinates.

In this study, we focus on a time period from February 15, 2020, to May 15, 2020, where we expect there to be a large volume of tweets that are COVID-19 related. Figure G.1a in Appendix G shows the number of tweets for each day in the study period. The following filtering was used:

- **U.S. geographic area**: Tweets that are geotagged and originated in the United States as indicated by the Twitter location service.
- **English language Tweets**: Tweets from users who selected English as their default language.
- **Non-retweets**: Tweets that contain original content from the users and are not a retweet of other tweets.

The following text preprocessing steps were undertaken: 1) we remove stop words (e.g., *in, on, and*, etc., which do not carry semantic meaning); 2) we keep only common forms of words (lemmatization); 3) we remove words that occurred less than 5 times in a document. As a result, the average vocabulary length per timestamp was been reduced from around 300000 to 3000. Further, the union of the unique words from each timestamp has been used as the common vocabulary with word frequencies zeroed out on days where those words do not occur.

3.2. **Hellinger-PHATE embedding for all topics.** Figure 4 shows the 2D Hellinger-PHATE embedding of 4500 word distributions. We labeled the points on the plots with different colors, sizes, and styles for visualization and interpretation of various time points, tweet volumes, and shortest paths. The full labeling scheme is included in Appendix I. Figure 4 also shows (as insets) two zoom-ins onto selected COVID-19 topics. We observe several interesting trajectory patterns in the PHATE embeddings. For example, the "STAY HOME (executive order)" cluster (bottom inset) is organized along a straight line, where the points are more dense at the beginning as well as at the end of the line while sparser in between. The COVID and COVID NEWS clusters (top inset) behave like a splitting between two branches of a tree, and the COVID NEWS (presidential election) path in those clusters exhibits a 'hook' or a 'U' shape. Within the COVID NEWS cluster, the presidential election path also splits and diverges from other points in the same cluster. The following two subsections will focus on these two clusters and paths therein to illustrate the advantages of the proposed framework. Additional visualizations for the SANITIZING (wash hands) and STAY HOME (executive order) paths are included in Appendix J.

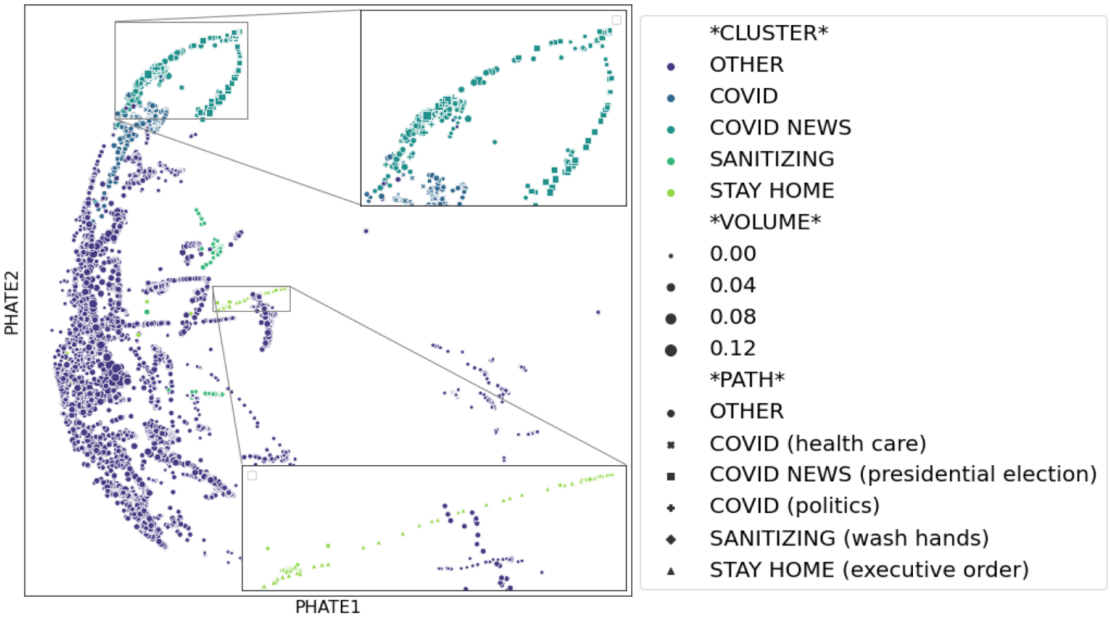


Figure 4. Potential of heat-diffusion for affinity-based transition embedding (PHATE) for all word distributions. Here the two bounding boxes and insets highlight two of the COVID-19-related topic clusters/paths (COVID/COVID NEWS and STAY HOME). The colors, sizes, and styles signify various clusters, tweet volumes, and shortest paths, as given in the dictionary in Appendix I. Note that the embedding captures some important clustering/trajectory structures, for example, branching, splitting, merging, and so on.

3.3. **Case study I: presidential election topic path.** Here, we focus on a cluster of topics that is implicitly COVID-related but can be well understood from associated real-world events. We call this the presidential election topical path. The subset of topics lying on this shortest path is illustrated in Figure 5. Here continuous color scales are used to illustrate temporal evolution, which exhibits a smooth transition from the beginning to the end points on the path. The PHATE embedding exhibits three subclusters on the path: 1) an early March cluster that groups topics related to Super Tuesday; 2) an April cluster that groups topics related to or triggered by the "Bernie Sanders dropped out of the presidential race" event; 3) an early to mid-May cluster that groups topics converging to more general COVID-related political topics.

Additionally, in terms of tweet volume generated by COVID NEWS topic, there exists again a U-shaped trend: starting at a high level in mid-February the tweet volumes dropped down after the Super-Tuesday week and started to rise near the time when Bernie Sanders dropped out and eventually peaked in mid-May. We believe that this modulation of the presidential election path can be explained by the COVID-19 pandemic in the United States, which accelerated through March when many states issued stay-at-home orders. This then triggered public discourse around COVID-19, increasing the volume of COVID-19-related topics. However, starting in May, as many stay-at-home orders were lifted, more mainstream political news topics reentered the discourse.

Following we present results of spatial analysis, showing county-level tweet volume in California, illustrating that the Hellinger-distance shortest path combined with PHATE is able to capture more granular-level variations in both space and time. In Figure 6 we plot smoothed choropleth maps

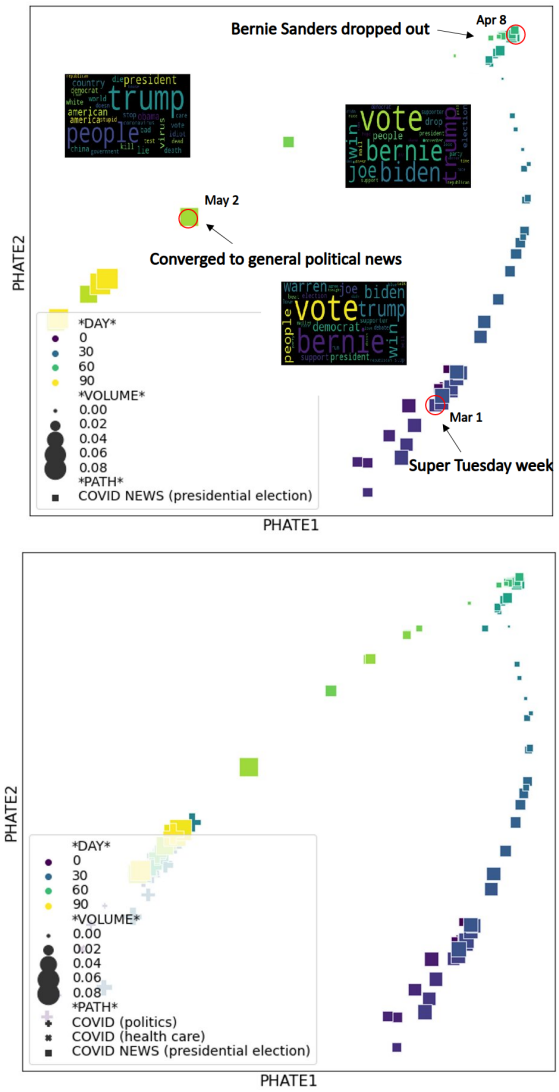


Figure 5. Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics in the COVID NEWS cluster (bottom) and the presidential election path (top) within the cluster. Colors and sizes highlight time and tweet volumes, respectively. Here three word clouds containing top 30 words in corresponding topics are shown for the time points highlighted by red circles, showing important real-word events that are annotated. Note the plot at the bottom shows (near lower left) the merge and split of different paths (labeled by filled squares, crosses, and pluses) within the same cluster.

for the same three topics that were highlighted in Figure 5, where the color changes with respect to tweet proportions (the estimated tweet volumes generated from the given topics normalized by the total tweet volumes for the given days for each county). Here raw tweet proportions have been smoothed using a simple Markov random field (MRF) smoother (Wood, 2017), which regularizes neighboring counties (i.e., regions with contiguous boundaries, that is, sharing one or

more boundary point) to have similar tweet proportions. This smoothing procedure is used to identify hot spots, or areas whose tweet volumes have a high likelihood of differing over neighboring locations. The regularization removes some of the variance one would normally see in a choropleth, and gives a bird's eye view of the entire state. For visualization of similar choropleth maps for other states, as well as a comparison of the maps between states, we include interactive maps at https://wayneyw.shinyapps.io/mrf_smooth_map_app/.

From the top row of Figure 6 we observe two 'presidential election' hot spots in counties near the Bay Area and in counties near Los Angeles. The local trend in tweet volume for California is similar to the global trend overall in the United States, as indicated by Figure 5 above as well as Figure K.1 in Appendix K.

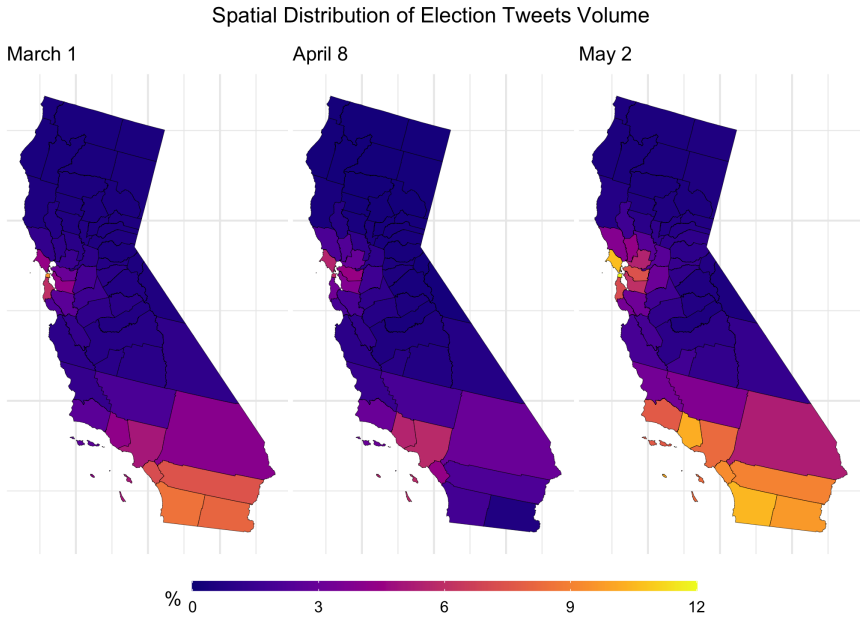


Figure 6. County-level maps for California. It shows the spatial distribution of proportional tweet volumes for the three time points on the COVID NEWS (presidential election) path.

3.4. **Case study II: general COVID-19 topic path.** In this case study we focus on an explicit COVID-19 topic cluster and shortest paths therein. Figure 7 shows the PHATE embedding for subsets of topics in the COVID cluster. The embedding identifies two paths that together exhibit splitting behavior, which can be considered as types of structures built into PHATE a priori. In this case, two similar discussions around COVID-19 split into a path that focused on health care, for example, testing, deaths, hospital, and so on, and a path that focused on politics, for example, government, Trump, president, and so on, respectively. The split of the two paths into two different sets of topics is revealed by naive clustering algorithms, such as hierarchical clustering. We emphasize here that such bifurcation behavior would be difficult to model explicitly, for example, using a time-varying global LDA-type model, but appears naturally in the PHATE embedding of the shortest paths using Hellinger distance.

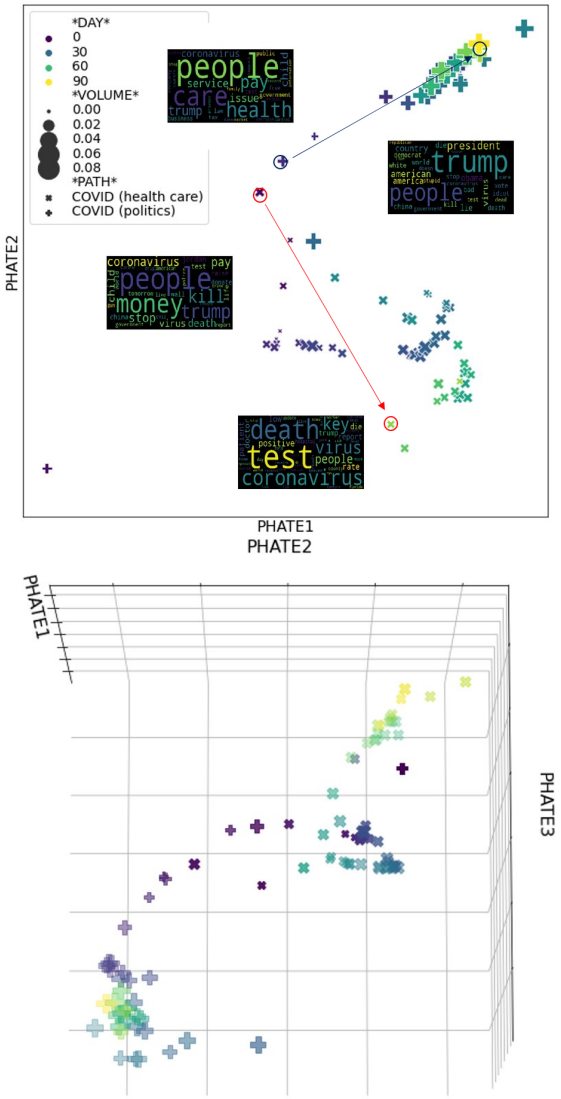


Figure 7. Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics in the COVID cluster. The plots demonstrate a 2D (top) and a 3D (bottom) embedding of two different paths (i.e., health care and politics). Colors and sizes highlight time and tweet volumes, respectively. Here four word clouds containing top 30 words in corresponding topics are shown for the time points (with arrows connecting the beginning and the end topics on the same path) highlighted by red (health care) and black (politics) circles. Note the plots show divergent behavior of public discourse around COVID-19, where two similar discussions diverge to different discussions (indicated by the word clouds). The 3D embedding illustrates nonlinear paths, that is, spirals and loops, for this topic.

The two separated paths can be more clearly observed in the 3D view, where a 'spiral' structure in the path labeled by filled circles is revealed. This spiral as well as the 'loop' presented in Figure 5

capture sharp transitions of discussions within a topic path, in contrast to more linear structures such as those exhibited in the SANITIZING (wash hands) and the STAY HOME (executive order) clusters, where the discussion is stable over time. In particular, the health care trajectory transitioned from a discussion on general concerns about the coronavirus to testing-focused discussions on a similar topic; the discussions along the presidential election trajectory transitioned from politicians in the presidential race to more general politics. On the other hand, as illustrated in Appendix J, for more linear 'wash hands' and 'executive order' trajectories, discussions along the paths are quite stable in terms of the most relevant words. We conjecture, more formally, that linear paths geometrically constitute a one-dimensional subspace over which a single multinomial word distribution propagates over time, unaffected by nearby clusters. This represents stability in the discussions of the topic. Nonlinear paths like spirals, on the other hand, likely constitute a nonlinear subspace where the multinomial word distribution changes smoothly over time, affected by proximity to other clusters.

For county-level spatial analysis, three examples of events can be visualized in Figure 8 (following the list of relevant events found at https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California):

- Spatial distribution of COVID tweet proportions on March 1, where the Bay Area is identified as a relative hot spot in the state. Around late February and early March, counties near the Bay Area were first hit by the coronavirus pandemic. For example, cases were reported in Alameda and Solano Counties on that day; a case was reported in Marin County, who was a passenger on the Grand Princess cruise.
- On March 11, the first death due to coronavirus was reported in LA County, and Ventura County reported their first case on the day before. These 'light up' the two counties on the map as a hot spot.
- On March 20 to March 21, Los Angeles County, which is nationally the second-largest municipal health system, announced that it could no longer contain the virus and changed their guidelines for COVID-19 testing to not test symptomatic patients if a positive result would not change their treatment. Note that the Bay Area hot spot before started to 'fade away' in terms of Tweets volume proportions.

3.5. **Limitations.** There are several limitations of our analysis that deserve additional attention. First, as with most statistical algorithms, there are user tuning parameters that must be selected. There are three tuning parameters that the user must provide: 1) the numbers of nearest neighbors k in the k nearest neighbor graph; 2) the data smoothing parameter γ; and 3) the number of topics K for the T-LDA algorithm. We have shown that our results are robust to perturbations about the parameters we chose, but there may be better choices. These include comprehensive cross-validation methods which, with sufficient computational resources, can be used to reliably select parameters that minimize a loss function. Such methods have been proposed for selecting k. For selection of K, a promising option is the hierarchical Dirichlet process (HDP), a nonparametric Bayesian model for the number of topics that could vary over time and model birth and death of topics (Teh et al., 2006). Our wrapper framework could easily incorporate an HDP in place of the T-LDA model, but at the expense of increased computation. Two more challenging limitations are those of selection bias and model bias.

Selection biases. The use of Twitter data for studying public discourse may be subject to selection bias as users of Twitter may not be representative of the U.S. population. Additionally, users of Twitter may be engaged in different types of public discourses around COVID-19 than users of other

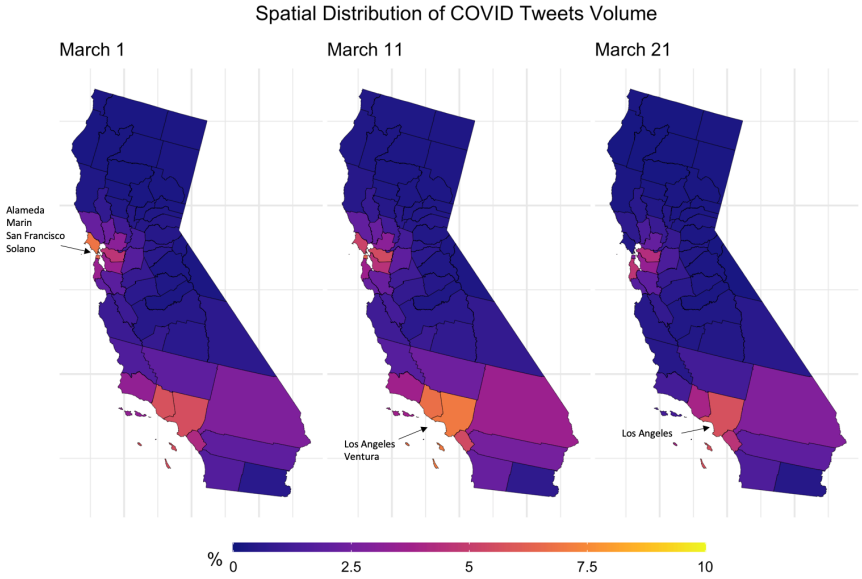


Figure 8. County-level maps for California. It shows the spatial distribution of proportional tweet volumes for three time points on the COVID (health care) path. Note that counties' names are given for spatial hot spots (in terms of tweet volume).

social media platforms, for example, Facebook and Reddit, which have different user demographics and privacy policies. Different types of subsampling of Tweets may create their own biases. For example, subsampling based on retweet status, geotag information, country, and time range (e.g., Feb 15 to May 15) are all subject to selection biases. Our subsampling procedure may leave out some important information. For example, we did not consider any retweets, which may contain information on how popular a particular topic might be. Retweets could possibly shed light on a particular topic, which can be measured, for example, by the longitudinal distribution of retweet frequencies for the topic. However, we could not perform a retweet analysis on our geotagged tweets since Twitter does not allow retweets to be geotagged. We also leave out tweets that are generated from U.S. users who are outside of the United States.

Model biases. The LDA algorithm we have applied to topic modeling summarizes unstructured texts by themes or topics using a *bag-of-words* approach. This particular approach is computationally scalable but it ignores the relative order of words. For example, a topic about 'vaccines are not available' can be very close to a topic on 'vaccines are available, but not to me.' The issue may be alleviated by using more sophisticated representations, for example, bigrams or latent semantic analysis. This would result in higher computational burden–the length of unique phrases would increase exponentially as the word order dimension. Other approaches that attempt to model the semantic meaning of topics, such as deep neural networks, could also be used. Additionally, our construction of the smoothed corpora assumes temporal similarities between tweets generated at adjacent time points. Similar types of smoothing assumptions are common in other areas of spatiotemporal statistics as described in Section 2.2. The manifold hypothesis is also essential in our model for the shortest path algorithms to recover the intrinsic similarities between topics over time.

4. Conclusion

We proposed a framework for longitudinal analysis of Twitter data, combining tools from graph algorithms, statistics, and computational geometry. The proposed procedure works by linking together marginal topic-word distributions discovered by a 'regularized' LDA model designed for microtext, via Hellinger distances and shortest paths on neighborhood graphs. The resulting chain of topics can then be visualized by PHATE dimensionality reduction, which preserves the progressive nature of the input data. With this framework, we discovered and interpreted how certain conversations split and merged under the impact of the COVID-19 pandemic, which can be validated by associating with real-world events. Granular-level spatial analyses showed that our framework is able to capture both global (in the United States) and local variations of COVID-19-related discussions. We think that social media data could be used to supplement traditional health care or census data to provide fresh insights into the impact of events, like the pandemic, on society.

Future directions include conducting theoretical analysis on how well and to what extent local LDA alignment using Hellinger shortest paths can approximate a global spatiotemporal LDA model. This will put the framework on a strong theoretical footing and provide mathematical insights that support the empirical results presented here. It will also be worthwhile to study other aspects of the Twitter data, for example, the retweeting network. This could elucidate how public conversation is affected by certain key users/sources and how this explains the spread of COVID-19-related discussions. Lastly, it will be interesting to study whether Twitter data can be useful as supplementary surveillance data. For example, spatiotemporally specific information on the number of complaints on Twitter about not being able to be tested or vaccinated for COVID-19 may be of value to public health authorities.

Disclosure Statement. The authors have no conflicts of interest to declare. The research in this paper was partially supported by grants from ARO W911NF-15-1-0479 and DOE DE-NA0003921.

Acknowledgments. The authors thank the Michigan Institute for Data Science (MIDAS) at the University of Michigan for providing access to the Twitter Decahose data. The authors thank James Chu for assistance in building the web application for spatiotemporal topic visualization.

Contributions. A.H., B.O., and W.D. devised the project, the main conceptual ideas and analysis outline. Y.W. and C.H. carried out the analyses and experiments. Y.W. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

References

Ahmed, N. K., Neville, J., & Kompella, R. (2013). Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(2), 1–56.

Amari, S.-I. (2012). *Differential-geometrical methods in statistics* (Vol. 28). Springer Science & Business Media.

Anand, K., Bianconi, G., & Severini, S. (2011). Shannon and von Neumann entropy of random networks with heterogeneous expected degree. *Physical Review E*, *83*(3), Article 036109.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, *15*, 2773–2832.

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models – Going beyond SVD. *2012 IEEE 53th Annual Symposium on Foundations of Computer Science* (pp. 1–10). IEEE Computer Society.

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, *56*(1-3), 209–239.

Bernstein, M., Silva, V. D., Langford, J. C., & Tenenbaum, J. B. (2000). Graph approximations to geodesics on embedded manifolds. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.6460

Bhadury, A., Chen, J., Zhu, J., & Liu, S. (2016). Scaling up dynamic topic models. *Proceedings of the 25th International Conference on World Wide Web* (pp. 381–390). International World Wide Web Conferences Steering Committee.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). Omni Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, *6*(4), Article e21978.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw/Hill.

Carter, K. M., Raich, R., Finn, W. G., & Hero III, A. O. (2009). Fine: Fisher information non-parametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 2093–2098.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chuang, J., Gupta, S., Manning, C., & Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In S. Dasgupta & D. McAllester (Eds.), *International Conference on Machine Learning* (pp. 612–620). PMLR.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In R. Mihalcea, J. Chai, & A. Sarkar (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 175–184). Association for Computational Linguistics.

Costa, J. A., & Hero, A. O. (2006). Determining intrinsic dimension and entropy of high-dimensional shape spaces. In H. Krim & A. Yezzi (Eds.), *Statistics and analysis of shapes* (pp. 231–252). Birkhäuser Boston.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., & Tong, X. (2011). TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2412–2421.

Dempsey, W. (2020). *The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts*. arXiv: https://arxiv.org/abs/2005.10425.

Diao, Q., Jiang, J., Zhu, F., & Lim, E.-P. (2012). Finding bursty topics from microblogs. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 536–544). Association for Computational Linguistics.

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, *100*(10), 5591–5596.

Doogan, C., Buntine, W., Linger, H., & Brunt, S. (2020). Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: A topic modeling analysis of Twitter data. *Journal of Medical Internet Research*, *22*(9), Article e21419.

Doxas, I., Dennis, S., & Oliver, W. L. (2010). The dimensionality of discourse. *Proceedings of the National Academy of Sciences*, *107*(11), 4866–4871.

Ellen, J. (2011). All about microtext - a working definition and a survey of current microtext research within artificial intelligence and natural language processing. In J. Filipe & A. Fred (Eds.), *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (pp. 329–336). SciTePress.

Gao, X., & Liang, J. (2011). The dynamical neighborhood selection based on the sampling density and manifold curvature for isometric data embedding. *Pattern Recognition Letters*, *32*(2), 202–209.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *24*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303–1347.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10.

Hwang, S. J., Damelin, S. B., & Hero III, A. O. (2016). Shortest path through random points. *The Annals of Applied Probability*, *26*(5), 2791–2823.

Jager, L., & Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, *35*(5), 2018–2053.

Jang, H., Rempel, E., Carenini, G., & Janjua, N. (2020). Exploratory analysis of COVID-19 related tweets in North America to inform public health institutes. In K. Verspoor, K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, & B. Wallace (Eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 (part 2) at EMNLP 2020*. Association for Computational Linguistics.

Krige, D. G. (1951). *A statistical approach to some mine valuation and allied problems on the Witwatersrand* [Doctoral dissertation, University of the Witwatersrand].

Liese, F. (2012). Phi-divergences, sufficiency, Bayes sufficiency, and deficiency. *Kybernetika*, *48*(4), 690–713.

Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C. J. P., & Ming, W.-K. (2020). Health communication through news media during the early stage of the COVID-19 outbreak in China: Digital topic modeling approach. *Journal of Medical Internet Research*, *22*(4), Article e19118.

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). TopicFlow: Visualizing topic alignment of Twitter data over time. In J. Rokne & C. Faloutsos (Eds.), *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 720–726). IEEE.

McInnes, L., Healy, J., & Melville, J. (2018). *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv: https://arxiv.org/abs/1802.03426.

Mimno, D., Hoffman, M. D., & Blei, D. M. (2012). Sparse stochastic inference for latent Dirichlet allocation. In J. Langford & J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning* (pp. 1515–1522). Omni press.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In A. Darwiche & N. Friedman (Eds.), *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (pp. 352–359). Morgan Kaufmann Publishers Inc.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, *37*(12), 1482–1492.

Newman, D., Smyth, P., Welling, M., & Asuncion, A. U. (2008). Distributed inference for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 1081–1088.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2009). On surrogate loss functions and f-divergences. *The Annals of Statistics*, *37*(2), 876–904.

Papernot, N., & McDaniel, P. (2018). *Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning*. arXiv: https://arxiv.org/abs/1803.04765.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). *Short text topic modeling techniques, applications, and performance: A survey*. arXiv: https://arxiv.org/abs/1904.07695.

Qiu, M., Zhu, F., & Jiang, J. (2013). It is not just what we say, but how we say them: LDA-based behavior-topic model. In J. Ghosh, Z. Obradovic, J. Dy, Z.-H. Zhou, C. Kamath, & S. Parthasarathy (Eds.), *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 794–802). SIAM.

Rao, C. R. (1995). The use of Hellinger distance in graphical displays of contingency table data. In E. M. Tiit, T. Kollo, & H. Niemi (Eds.), *New Trends in Probability and Statistics* (pp. 143–161). VSP, Utrecht.

Samko, O., Marshall, A. D., & Rosin, P. L. (2006). Selection of the optimal parameter value for the isomap algorithm. *Pattern Recognition Letters*, *27*(9), 968–979.

Servidea, J. D., & Meng, X.-L. (2006). Statistical physics and statistical computing: A critical link. In J. Fan & H. L. Koul (Eds.), *Frontiers in statistics* (pp. 327–344). World Scientific.

Sha, H., Hasan, M. A., Mohler, G., & Brantingham, P. J. (2020). *Dynamic topic modeling of the COVID-19 Twitter narrative among us governors and cabinet executives*. arXiv: https://arxiv.org/abs/2004.11692.

Shemyakin, A. (2014). Hellinger distance and non-informative priors. *Bayesian Analysis*, *9*(4), 923–938.

Song, X., Lin, C.-Y., Tseng, B. L., & Sun, M.-T. (2005). Modeling and predicting personal information dissemination behavior. In R. Grossman, R. Bayardo, & K. Bennett (Eds.), *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 479–488). Association for Computing Machinery.

Srivastava, A., & Sutton, C. (2017). *Autoencoding variational inference for topic models*. arXiv: https://arxiv.org/abs/1703.01488.

Srivastava, A., & Klassen, E. P. (2016). *Functional and shape data analysis* (Vol. 1). Springer.

Stokes, D. C., Andy, A., Guntuku, S. C., Ungar, L. H., & Merchant, R. M. (2020). Public priorities and concerns regarding COVID-19 in an online discussion forum: Longitudinal topic modeling. *Journal of General Internal Medicine*, *35*(7), 2244–2247.

Su, Y., Venkat, A., Yadav, Y., Puglisi, L. B., & Fodeh, S. J. (2021). Twitter-based analysis reveals differential covid-19 concerns across areas with socioeconomic disparities. *Computers in Biology and Medicine*, *132*, Article 104336.

Taddy, M. (2012). On estimation and selection for topic models. In N. D. Lawrence & M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (pp. 1184–1193). PMLR.

Teh, Y. W., Kurihara, K., & Welling, M. (2008). Collapsed variational inference for HDP. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (pp. 1481–1488). Curran Associates, Inc.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

von Neumann, J. (2013). *Mathematische grundlagen der quantenmechanik* (Vol. 38). Springer-Verlag.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. In D. McAllester & P. Myllymaki (Eds.), *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (pp. 579–586). AUAI Press.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In T. Eliassi-Rad, L. Ungar, M. Craven, & D. Gunopulos (Eds.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 424–433). Association for Computing Machinery.

Wang, X., Mohanty, N., & McCallum, A. (2005). Group and topic discovery from relations and text. In J. Adibi, M. Grobelnik, D. Mladenic, & P. Pantel (Eds.), *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 28–35). Association for Computing Machinery.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, *6*(3), 324–342.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID-19 pandemic: Using latent Dirichlet allocation for topic modeling on Twitter. *PloS One*, *15*(9), Article e0239441.

Yang, Y., Chen, C., & Bao, F. S. (2016). Aspect-based helpfulness prediction for online product reviews. In N. Bourbakis, A. Esposito, A. Mali, & M. Alamaniotis (Eds.), *2016 IEEE 28th International Conference on Tools with Artificial Intelligence* (pp. 836–843). IEEE.

Yuan, M., Van Durme, B., & Ying, J. L. (2018). Multilingual anchoring: Interactive topic modeling and alignment across languages. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 8653–8663). Curran Associates, Inc.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *Advances in Information Retrieval* (pp. 338–349). Springer Berlin Heidelberg.

Appendix A. Additional details of the Twitter Latent Dirichlet Allocation (T-LDA) algorithm

The generation processes for a T-LDA and an LDA are illustrated side-by-side in Figure A.1. Here, the key differences exhibited in T-LDA are aggregation (pooling tweets from users) and regularization (restricting a tweet to be generated from only one topic). In our study, the aggregation is done by pooling tweets generated from the same day.

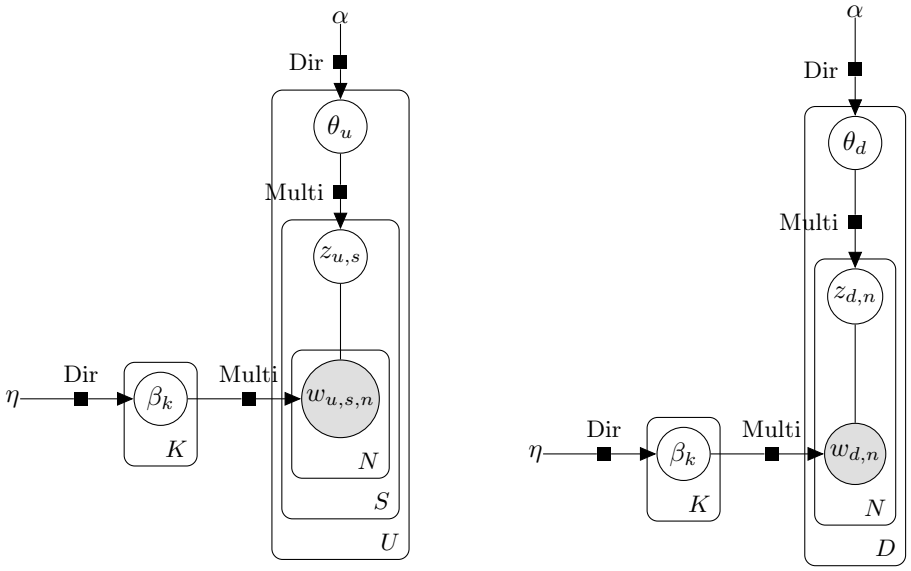


Figure A.1. Plate notation comparison for the Twitter Latent Dirichlet Allocation (T-LDA) (left) and the standard Latent Dirichlet Allocation (LDA) (right) models. Here nodes are random variables; edges indicate dependence through probability distributions (e.g., Dirichlet or multinomial). Shaded nodes are observed; unshaded nodes are latent. Plates indicate replicated variables. Note that the T-LDA model aggregates tweets from each user into a document and constrains each tweet to be drawn from only one topic.

All numerical results presented in the article were produced with the following implementation details of the T-LDA algorithm: the collapsed Gibbs sampler has been run for 2000 iterations with the first 1000 samples discarded as burn-in. The latent variable β is assumed to be symmetric Dirichlet with hyperparameter $\eta = 0.01$ for all topics; and θ is assumed to be symmetric Dirichlet with hyperparameter $\alpha = 0.5$ for all time stamps.

Appendix B. Summary statistics characterizing shortest paths

To characterize the smoothness and continuity of the learned shortest paths of topics, we present summaries of the 'skips' (days where there are no topics connected to either a topic immediately before or after the current timestamp) they made. Table B.1 depicts the number of skips and the length of the skips for four topic paths (see Appendix I for details on the path names). We note that the length of a whole path (number of topics connected) could be different because 1) the different numbers of skips, and 2) the different time span as some topics appeared only for a certain time range (e.g., the wash hands topic). The lengths of those paths shown in the table are: COVID NEWS (presidential election), 70; COVID (health care), 58; STAY HOME (executive order), 59; SANITIZING (wash hands) 19. Clearly, longer paths could make longer skips. However, the paths remain fairly continuous (small numbers of short skips) during their time span. This is partly due to the corpora smoothing being applied–the topics learned at time t should usually be very similar to those learned at nearby timestamps.

Table B.1. Summary of the number of skips along with the length of those skips for four different topic paths. The paths are discovered by the shortest path algorithm using 10-nearest neighbor weighted graph. Note that all paths exhibit small numbers of short-length skips.

Path Name	Days Skipped			
	1	2	3	4
COVID (health care)	10	0	1	1
COVID NEWS (presidential election)	3	1	2	2
SANITIZING (wash hands)	1	1	0	0
STAY HOME (executive order)	4	0	0	0

Appendix C. Shortest path on multidimensional scaling (MDS), isometric feature mapping (ISOMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE)

We desire a low-dimensional embedding that preserves the trajectory structures of shortest paths, so that we can visualize and interpret any results computed using methods described in Section 2.2. Here we compare PHATE with MDS and ISOMAP. MDS does not take any local structural information into account when building the embedding; ISOMAP applies MDS using shortest path distances computed on neighborhood graphs; finally, PHATE applies MDS on potential distances computed on neighborhood graphs while striking a balance between local and global trajectory structures. Figure C.1 shows that MDS failed to identify any path between two points. ISOMAP identifies a cleaner structure but there are interrupting background points on the path. PHATE identifies a clean path that is also well separated from background points. The comparison also highlights the importance of working with neighborhood graphs, instead of the fully connected graph, when trying to identify local structures in data.

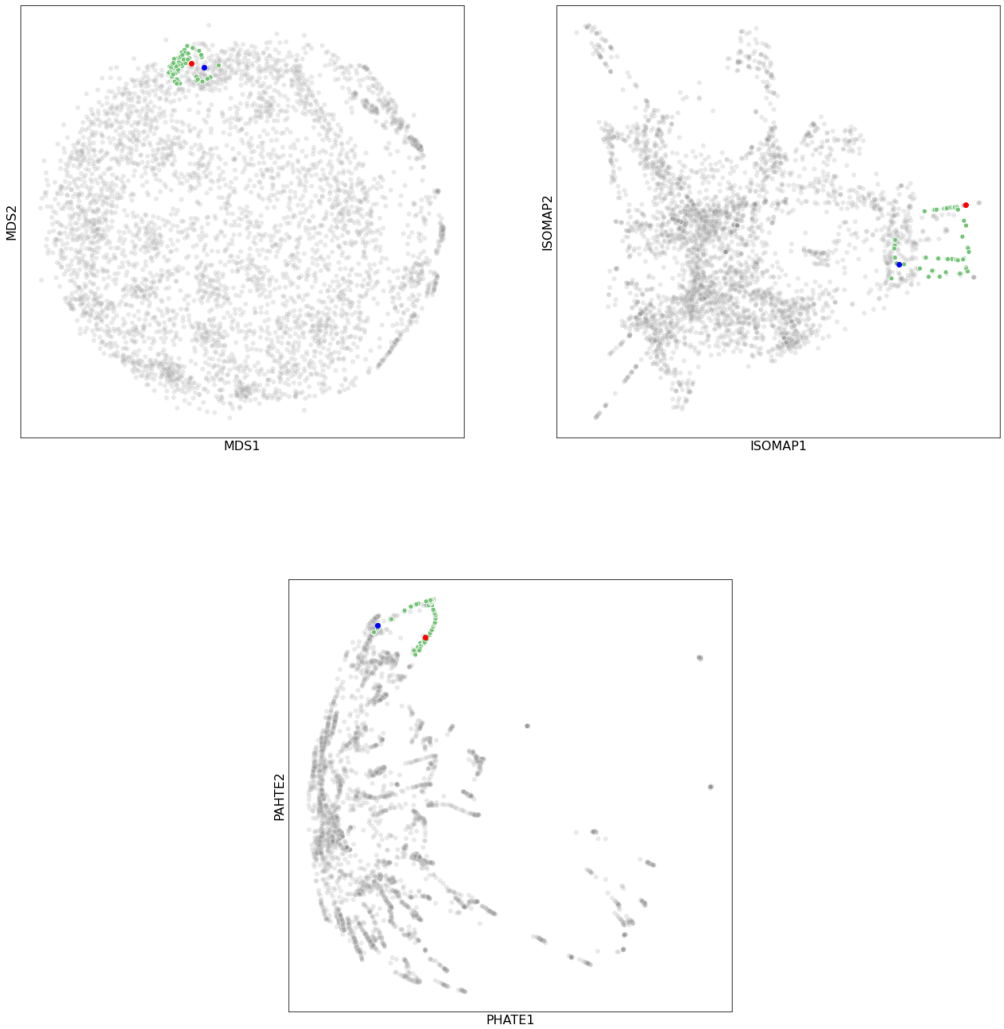


Figure C.1. Multidimensional scaling (MDS), isometric feature mapping (ISOMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) for the same set of word distributions. A shortest path computed on 10 nearest neighbors graph is highlighted on each embedding with red and blue points indicating the starting and ending points of the path. Note that PHATE identifies the cleanest path connecting the red and blue points, with minimal background noises (grey points) included in between.

Appendix D. Detailed descriptions of PHATE

Algorithm D.1 outlines the steps for obtaining a low-dimensional embedding using PHATE with the Hellinger distance metric.

Algorithm D.1 PHATE with Hellinger distance

Input: N observations of some objects

1: Compute pairwise Hellinger distance matrix (denoted as D) from all pairs of multinomial topic distributions (stored as columns in a matrix X).
2: Compute k-nearest neighbor distance (denoted as $\epsilon_k(x)$) from each column of X.
3: Compute local affinity matrix $K_{k,\alpha}$ from D and ϵ_k.
4: Form a diffusion operator P, which is a Markov transition matrix computed by normalizing $K_{k,\alpha}$.
5: Compute time scale via Von Neumann Entropy. The time scale is then used to diffuse P to obtain P^t.
6: Compute potential representation of the diffusion matrix as $U_t = -\log(P^t)$ and compute potential distance matrix $D_{U,t}$ from U_t.
7: Apply MDS on $D_{U,t}$ to embed the data in lower dimension.

Output: An $N \times L$ matrix that contains L-dimensional coordinates for each observation.

In Algorithm D.1 we use the Hellinger distance to compute D and ϵ_k. This ensures that PHATE is being used to perform dimension reduction on a statistical manifold (Amari, 2012).

The PHATE construction is based on computing local similarities between data points, and then diffusing through the data using a Markovian random-walk diffusion process to infer more global relations. The local similarities between points are computed by first computing pairwise distances and then transforming the distances into similarities, via a kernel named the α-decaying kernel with locally adaptive bandwidth. It is defined as

$$K_{k,\alpha}(x,y) = \frac{1}{2}\exp\Big(-\Big(\frac{\|x-y\|}{\epsilon_k(x)}\Big)^{\alpha}\Big) + \frac{1}{2}\exp\Big(-\Big(\frac{\|x-y\|}{\epsilon_k(y)}\Big)^{\alpha}\Big). \tag{D.1}$$

Here the k-nearest neighbor distance ϵ_k is used to ensure that the bandwidth is locally adaptive and varies based on the local density of the data. The exponent α controls the rate of decay of the tails in the kernel $K_{k,\alpha}$. Setting $\alpha = 2$ is equivalent to the use of a Gaussian kernel and choosing $\alpha > 2$ results in lighter tails in the kernel. The kernel is then normalized by row-sums that results in a row-stochastic matrix $P = P_{k,\alpha}$ (the diffusion operator), which is used for following steps.

In Step 5, the diffusion operator is powered by a time scale t. In particular, for a data point x and diffusion operator P, and let δ_x be the Dirac delta that is defined to be a row vector of length N (length of the data) with a one at entry corresponding to x and zero elsewhere. The t-step distribution of x is the row in P^t corresponding to x:

$$p_x^t := \delta_x P^t = [P^t]_{(x,\cdot)}. \tag{D.2}$$

This distribution captures multiscale (where t serves as the scale) local neighborhoods of data points, where the local neighborhoods are explored by randomly walking or diffusing over the intrinsic manifold geometry of the data. The scale parameter t affects the embedding. It can be selected based on any prior knowledge of the data or, as proposed in Moon et al. (2019), by quantifying the information in the powered diffusion operator with different values of t, via computing the Von

Neumann Entropy (Anand et al., 2011; von Neumann, 2013) of the diffusion affinity, and choosing the one that explains the maximum amount of variability in the data.

Finally, a new type of distance, called the potential distance in Moon et al. (2019), is recovered in the end from the powered diffusion operator, which is obtained by taking the negative log of the transition probabilities. This transforms these transition probabilities into the heat-potential context.

Appendix E. Additional simulation studies for PHATE

To illustrate the idea of probability vectors on a sphere, in Figure E.1 we present a simple example of a sphere in 3D and probability vectors (simulated as in Section 2.3) lying on the sphere. The trajectories in this simulated example exhibit different progressive structures. In particular, the trajectory in dark blue evolves smoothly and remains roughly on the same path; the trajectory in brown exhibits a sharp turn in the direction at a certain position; finally, the trajectory in light blue behaves more chaotically and exhibits clustering structures. The PHATE embedding presented in Figure 3 of Section 2.3 was able to uncover all these types of structures in low dimension.

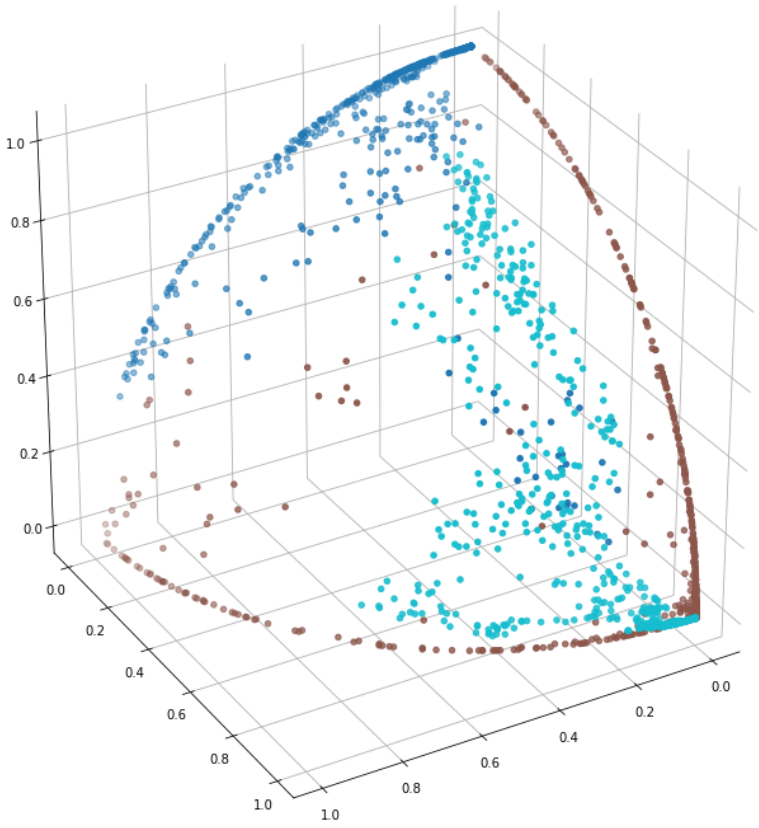


Figure E.1. Three simulated trajectories of probability vectors on a sphere. Each color signifies a trajectory simulated using a specific σ in the random-walk structure described in Section 2.3. Here, three trajectories started at the same point exhibit different progressive structures: stable (dark blue), chaotic and clustering (light blue), and sharp transition (brown).

To further demonstrate the advantage of PHATE over traditional methods for uncovering progressive structures, we present a similar example to that in Moon et al. (2019), which uses artificial tree-structured data and compare principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and PHATE in constructing low-dimensional embedding.

In particular, we generate tree-structured data with 10 branches and 200 dimensions, and each branch has length 300. Thus, we have 3000 observations of 200-dimensional data, and the goal is to find a 2-dimensional embedding for visualization. Figure E.2 shows the results of embedding for three different methods. PCA is good for finding an optimal linear transformation that gives the major axes of variation in the data. However, the underlying data structure in this case is nonlinear in which case PCA is not ideal. t-SNE is able to embed nonlinear data; however, it is optimized for cluster structure and as a result will destroy any continuous progression structure in the data. PHATE for this example separates the clusters and is able to clearly represent the trajectory structure of the data. Additionally, PHATE neatly captures the branching/splitting points of different trajectories. This feature is vital for our study of tweeting behaviors as we are interested in learning how different conversations converge to a similar one or diverge to different topics.

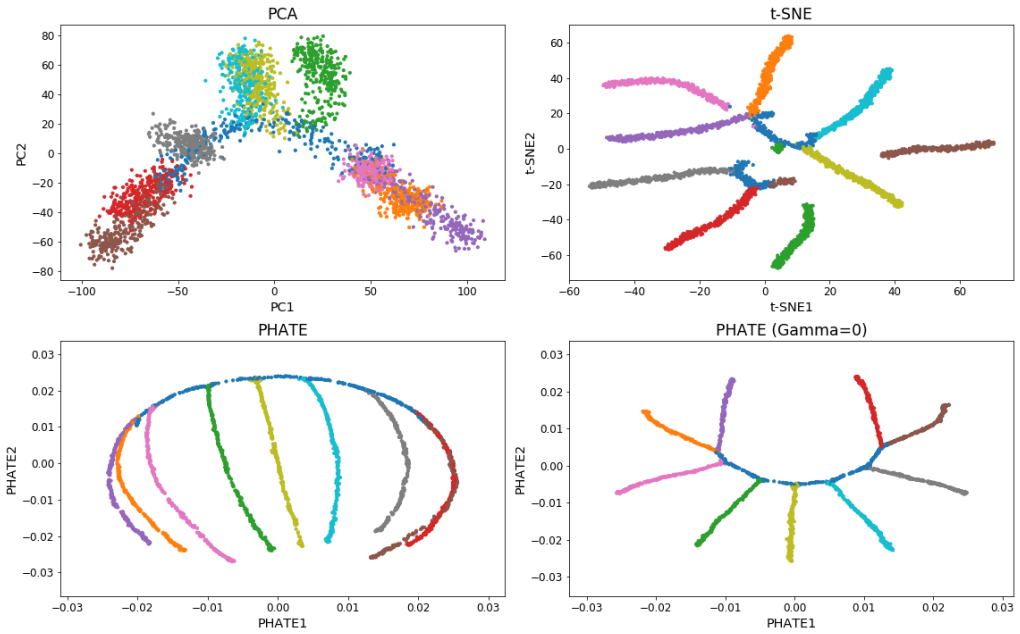


Figure E.2. Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and potential of heat-diffusion for affinity-based transition embedding (PHATE). Two versions of PHATE with different tuning parameters are illustrated. The data are 3000 tree-structured observations with 10 branches. Various branches are colored differently. Note that for this truly trajectory-based data, PHATE gives the clearest low-dimensional representation of the data.

Additionally, we also demonstrate that PHATE does not 'create' spurious trajectories, although it does not preclude the existence of such structures. Here, 3000 independent data points were simulated from a 3-component (with weights $0.6, 0.3, 0.1$) 10-dimensional Gaussian mixture model and transformed through softmax (i.e., $z_j \to \frac{\exp(z_j)}{\sum_{i=1}^{10}\exp(z_i)}, j = 1, \ldots, 10$). Figure E.3 depicts 2-dimensional embedding computed by PCA, t-SNE, uniform manifold approximation and projection (UMAP), and PHATE using Hellinger distance. Clearly, PHATE did not artificially 'trajectorize' the data; t-SNE seems to perform the best in terms of clustering as it often tries to separate data as much

as possible; UMAP separated the clusters well but generated artificial segments and trajectories in the embedding.

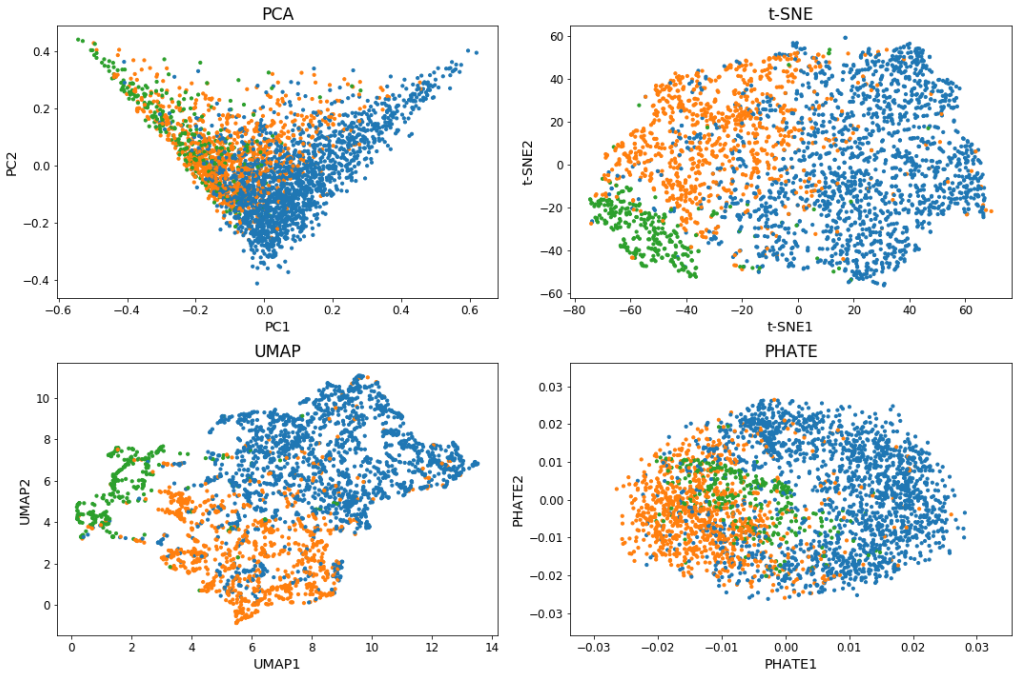
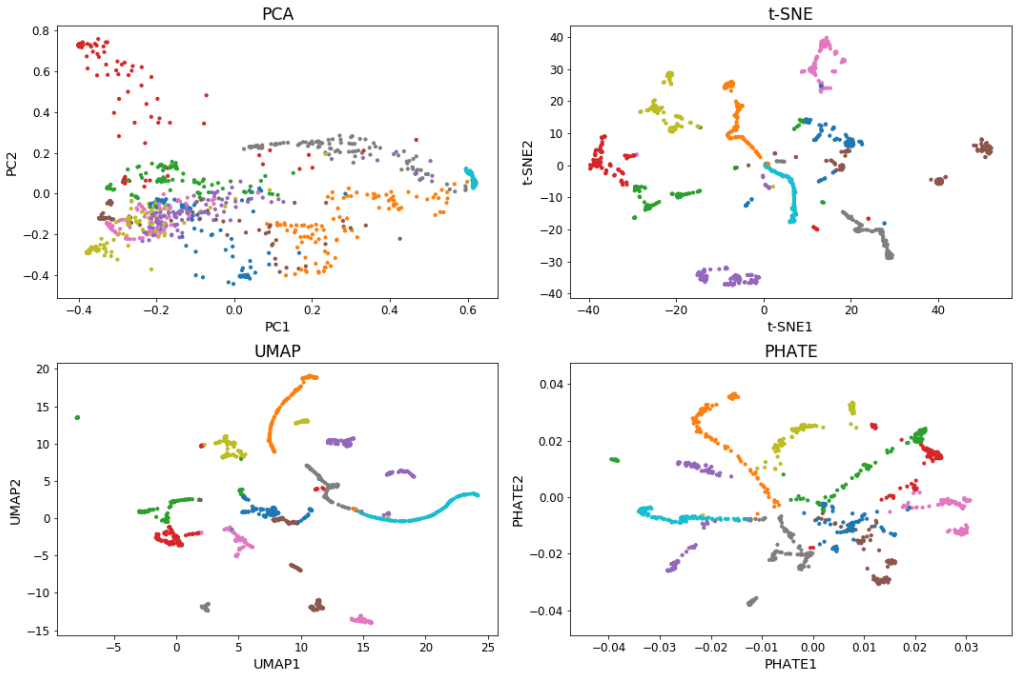
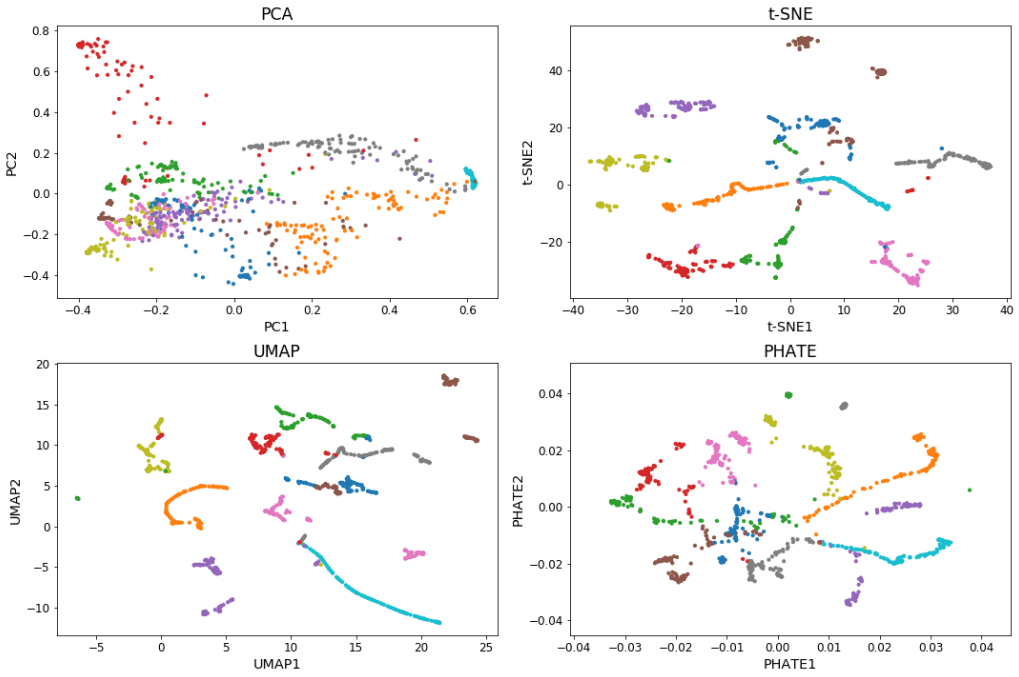


Figure E.3. Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE). Here $3,000$ independent data points were generated from a 3-component (with weights $0.6, 0.3, 0.1$) 10-dimensional Gaussian mixture model. Here, data were transformed via softmax to resemble a probability vector. Note that for this random nonstructured data, PHATE did not 'create' spurious trajectories in the low-dimensional embedding.

Lastly, we compare PHATE (and other) embeddings using different distance metrics. In particular, we compute 2-dimensional embeddings for the data generated in Section 2.3 using Euclidean and cosine distances/similarities. Figure E.3 depicts the results comparing PCA, t-SNE, UMAP, and PHATE. It shows that the Hellinger metric (for t-SNE, UMAP, and PHATE) outperforms the other two in terms of generating the clearest low-dimensional embedding that preserves the true data geometry.



(A) Embedding using Euclidean metric.



(B) Embedding using Cosine metric.

Figure E.4. Principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) using Euclidean and cosine metrics. Here 10 trajectories of 100-dimensional probability vectors are generated, where the trajectories are colored differently. PHATE gives the clearest 2D representation of the inputs that preserves their high-dimensional progressive structures, regardless of the distance metric used. Comparing with Figure 3, the Hellinger metric outperforms the other two metrics in recovering the data geometry.

Appendix F. Comparison with TopicFlow for topic trend mining

TopicFlow (Malik et al., 2013) is an analysis framework for Twitter data over adjacent time slices, binned topic models, and alignment, which is an application of LDA to timestamped documents at independent time intervals and alignment of the resulting topics. The key differences between TopicFlow and the proposed framework are: 1) a different similarity measure between topics, that is, cosine similarity metric for TopicFlow; 2) a different mechanism for topic alignment and connection–TopicFlow connects every pair of adjacent topics that has similarity above a certain threshold. The advantages of Hellinger metric over other metrics for comparing/embedding word distributions have been made clear in the previous section. Here, we demonstrate the advantages of the proposed shortest path mechanism over TopicFlow for obtaining natural temporal evolution of topics.

We analyze a particular topic cluster–the presidential election cluster discussed in Section 3–and compare the connections computed by the proposed shortest path algorithm and the TopicFlow algorithm. Here, for a fair and direct comparison, we fix the bins and the topic detection algorithms to be the same for both frameworks–using the smoothed temporal corpus and the T-LDA; the shortest path is performed on a 10-nearest neighbor weighted graph and the TopicFlow is performed with a connection threshold of 0.2. For the latter, we obtain a path by localizing the connection that has the largest cosine similarity at each pair of adjacent timestamp. For illustration, in Table F.1, we highlight a time segment that exhibits differences between two paths. In particular, the shortest path skipped 3 days, March 23 to March 26, while the TopicFlow remain continuously connected. The top row of Figure F.1 depicts the top word clouds of topics at timestamps March 23, 24, 27, and May 15 on the TopicFlow path. It shows a sharp transition from a voting/election topic to general political topics and finally to a relatively nonpolitical topic. On the other hand, the shortest path automatically skipped the timestamps where these new topics emerged and maintained the major theme of the path, which is voting/election and later on general politics. This offers a more natural and much smoother transition.

This comparison demonstrates particularly that the mechanism for topic trend discovery used by TopicFlow is restrictive as it potentially results in nonsmooth and nonintuitive transitions. Although one could tune the connection threshold, it increases the computational burden and there is no obvious objective (e.g., prediction score, loss, etc.) that could help with the tuning process.

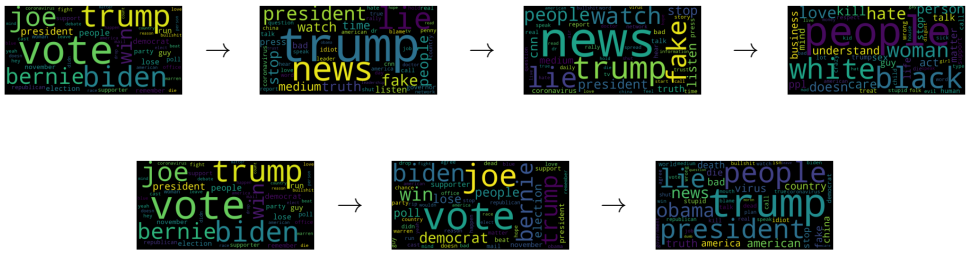


Figure F.1. Top word clouds showing evolution of topics on the presidential election topic paths computed via the shortest path algorithm (bottom) and the TopicFlow (top) algorithm. The sample timestamps at which the topics are learned are March 23, 24, 27, and May 15 (top); March 23, 27, and May 15 (bottom). Note that the shortest path algorithm produces much smoother and more intuitive transitions among topics within a general theme.

Table F.1. A portion of connected presidential election topics via the shortest path mechanism (left column) and the TopicFlow mechanism (right column). Here topics are indicated by their indices, e.g., $0 - 49$, at each timestamp (row index). *NA* indicates that no connection has been made by the algorithm.

	SP topic index	**TF topic index**
Feb 15	37	37
⋮	⋮	⋮
Mar 23	1	1
Mar 24	NA	44
Mar 25	NA	27
Mar 26	NA	26
Mar 27	22	26
⋮	⋮	⋮
May 15	2	15

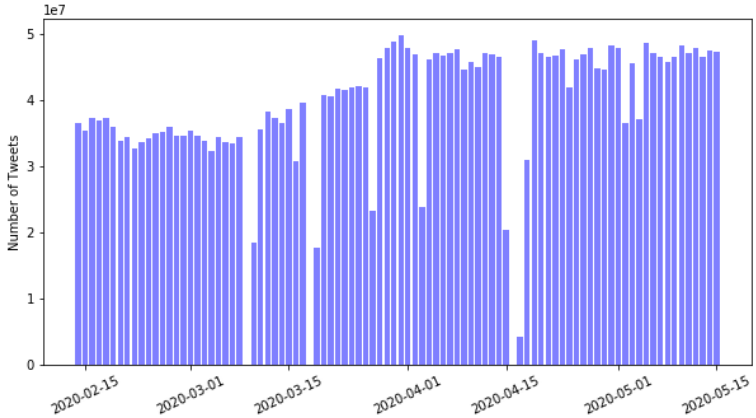
To further investigate the two different topic aligning methods, we fix the distance metric to be Hellinger, and compare the shortest path mechanism and the TopicFlow mechanism for the same set of topics. Table F.2 depicts a similar pattern for the time range March 23 to 27, where the restrictive TopicFlow mechanism for topic connection exhibits a sharp transition as shown in Figure F.1. Similar to Table F.1, from February to March 23, the two paths are mostly the same. However, we observe that the two paths also exhibit similar topics near the end of the time period. This again demonstrates the superiority of Hellinger distance for measuring topic similarity.

Table F.2. A portion of connected presidential election topics via the shortest path mechanism (left column) and the TopicFlow mechanism (right column) using the same distance metric (Hellinger). Here topics are indicated by their indices, e.g., $0 - 49$, at each timestamp timestamps (row index). *NA* indicates that no connection has been made by the algorithm. Note that the restriction imposed by TopicFlow impacts the topic path similar (from March 23 to 27) to that in Table F.1

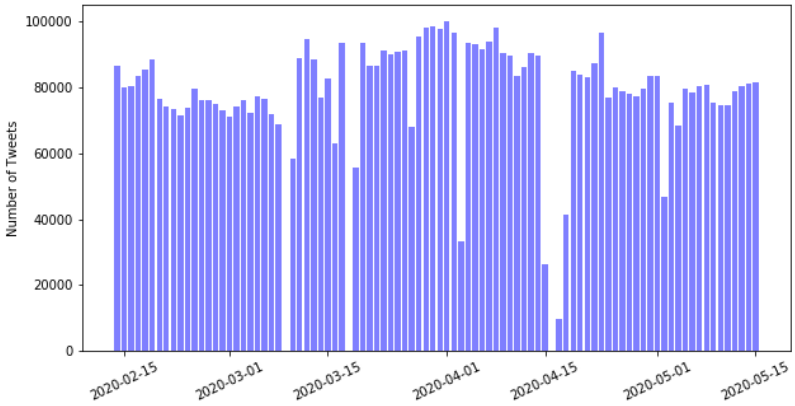
	SP topic index	**TF topic index**
Feb 15	37	37
⋮	⋮	⋮
Mar 23	1	1
Mar 24	NA	44
Mar 25	NA	27
Mar 26	NA	26
Mar 27	22	26
⋮	⋮	⋮
May 12	8	8
May 13	42	42
May 14	0	0
May 15	2	2

Appendix G. Volume plots of raw Twitter Decahose data

Figure G.1b shows the Decahose Twitter volume plots before (top) and after (bottom) processing. Although Twitter officially claims the percentage of geotagged tweets to be around 1-2% of the total tweets (https://developer.twitter.com/en/docs/tutorials/Tweet-geo-metadata), we found the percentage to much smaller. Note that there are several time points where the data is either incomplete (i.e., low volumes) or missing (i.e., 0 volumes).



(A) Raw Decahose tweets volume (on a scale of 10^7 tweets) from Feb 15 to May 15.



(B) Geotagged U.S., non-retweet, English Decahose tweets volume from Feb 15 to May 15.

Figure G.1. Volume of all and geotagged Decahose tweets for each day during the study period. The Decahose stream generates around $30 - 50$ million raw tweets and $50 - 100$ thousand geotagged English language tweets per day, except for several missing/incomplete cases with 0 or abnormally small volumes.

Appendix H. Sensitivity analysis for hyperparameters

In this section, we perform sensitivity analyses for the hyperparameters k and γ in Algorithm 2, namely the number of neighbors in the nearest neighbor graph and the smoothing parameter for constructing new corpora. Further, we perform model selection for varying choices of K, the number of topics in T-LDA.

In Figure H.1, the two shortest paths computed using neighborhood graphs of $k = 8$ and 12 are illustrated. For comparison, the same starting and ending topics as well as the two intermediate topics at the same time points as those in Figure 2 are used. It is clear from the word clouds that the shortest paths are not sensitive to the choice of k in the neighborhood of 10.

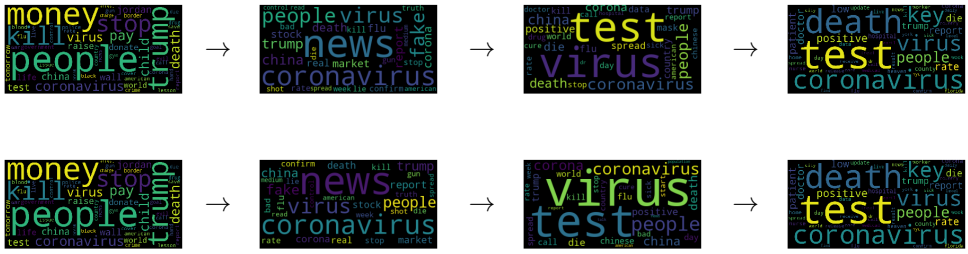


Figure H.1. Evolution along the shortest paths of a COVID-19 topic on the first day to a COVID-19 health care focused topic on the last day illustrated as top word clouds. The paths are computed on a 8- (top) and a 12- (bottom) nearest neighbor graph. The middle two word clouds are illustrations of two of the topics on the paths at the same time points as those in Figure 2. Note that the intermediate topics in both cases represent natural transformations from the beginning to the end topics, confirming that the shortest path is not sensitive to small perturbations of k around 10.

Additionally, we quantify the similarities between any two shortest paths computed on different neighborhood graphs by computing the average Hellinger distance between topics (at the same time point) on the paths. Particularly, in Table H.1 we show the average Hellinger distances. For this particular cluster of topics, the average Hellinger distances are negligible and are stable across all pairs of different paths, which suggests that the shortest path is not sensitive to different k in the neighborhood of $k = 10$.

Table H.1. Average Hellinger distances between any two topics paths generated using various neighborhood parameters k as the column/row indices. Examples are shown for the COVID (health care) topics. Note that the average Hellinger distances are identically 0 across all pairs of paths, indicating that the shortest paths are stable under different choices of k.

	8	**10**	**12**
8	0	0	0
10	0	0	0
12	0	0	0

Figure H.2 shows the contributions (in terms of the number of tweets) from each document to the temporally smoothed corpus constructed for March 31, using smoothing parameters of $0.65, 0.75, 0.85$.

With 0.75, the contents span the whole study period (Feb 15 to May 15) but concentrate on tweets within a month, centered at March 31.

Moreover, Figure H.3 shows the PHATE embedding of all topics learned by T-LDA, using corpus constructed with smoothing parameters 0.65 and 0.85. Here we highlight two clusters (COVID and COVID NEWS) and one shortest path (presidential election) similar to Figure 4. Comparing the three PHATE plots, the overall structures are similar and the highlighted trajectories remain relatively stable (i.e., presidential election paths exhibit similar 'U' shapes in all cases). Note that the 'split-and-merge' behaviors within the COVID NEWS cluster are being captured in all cases as well. The only notable difference in the PHATE produced with different temporal smoothing is the length of the trajectories, with those in the embedding produced using smoothing parameter 0.85 being the longest. This is reasonable because a larger smoothing parameter assumes a longer range temporal dependence structure of the data.

Additionally, we quantify the similarities between any two shortest paths from different smoothed corpora by computing the average Hellinger distance of the topics on the paths. Particularly, in Table H.2 we show the average Hellinger distances between any two paths computed under different smoothing conditions, for the COVID NEWS (presidential election) and COVID (health care) topics. In these two cases, the average Hellinger distances are around 0.35 and are stable across all pairs, which suggests that the shortest paths of key topics of interest are not sensitive to different smoothing parameters.

Table H.2. Average Hellinger distances between any two topics paths generated from corpora with various smoothing parameters as the column/row indices. Examples are shown for the COVID NEWS (presidential election) and the COVID (health care) topics in the top and bottom tables, respectively. Note that the average Hellinger distances are both relatively small and stable in the sense that all pairwise distances are similar in magnitude, indicating that the shortest paths are stable under different choices of smoothing parameters.

	0.65	**0.75**	**0.85**
0.65	0	0.3520	0.3578
0.75	0.3520	0	0.3056
0.85	0.3578	0.3056	0

	0.65	**0.75**	**0.85**
0.65	0	0.3697	0.4112
0.75	0.3697	0	0.3652
0.85	0.4112	0.3652	0

Lastly, for the choice of the number of topics for T-LDA, we propose to compute a Bayesian Information Criteria (BIC) score at each timestamp defined as

$$-\text{log-likelihood} + \frac{C \log(D)}{2}$$

where the model complexity is computed by $C := Kp + (K-1)D$ with p denoting the length of the vocabulary. The log-likelihood of the T-LDA model is defined as

$$\prod_{k=1}^{K} \text{Dirichlet}(\beta_k; \eta) \prod_{d=1}^{D} \text{Dirichlet}(\theta_d; \alpha) \prod_{s=1}^{S_d} \text{Categorical}(z_{s,d}; \theta_d) \prod_{n=1}^{N_s} \text{Categorical}(w_{n,s,d}; \beta_{z_{s,d}}).$$

Here, the categorical distribution is a special case of the multinomial distribution, in that it gives the probabilities of potential outcomes of a single drawing rather than multiple drawings; S_d denotes the number of tweets in document d ; and N_s denotes the number of words in a tweet s. This criteria is similar to the topic model selection criteria proposed in Taddy (2012).

In Figure H.4, we show the computed scores across all timestamps for various choices of the numbers of topics. The model with $K = 50$ consistently produces the lowest scores for the first half of the time range and is comparable to the model with $K = 100$ for the second half.

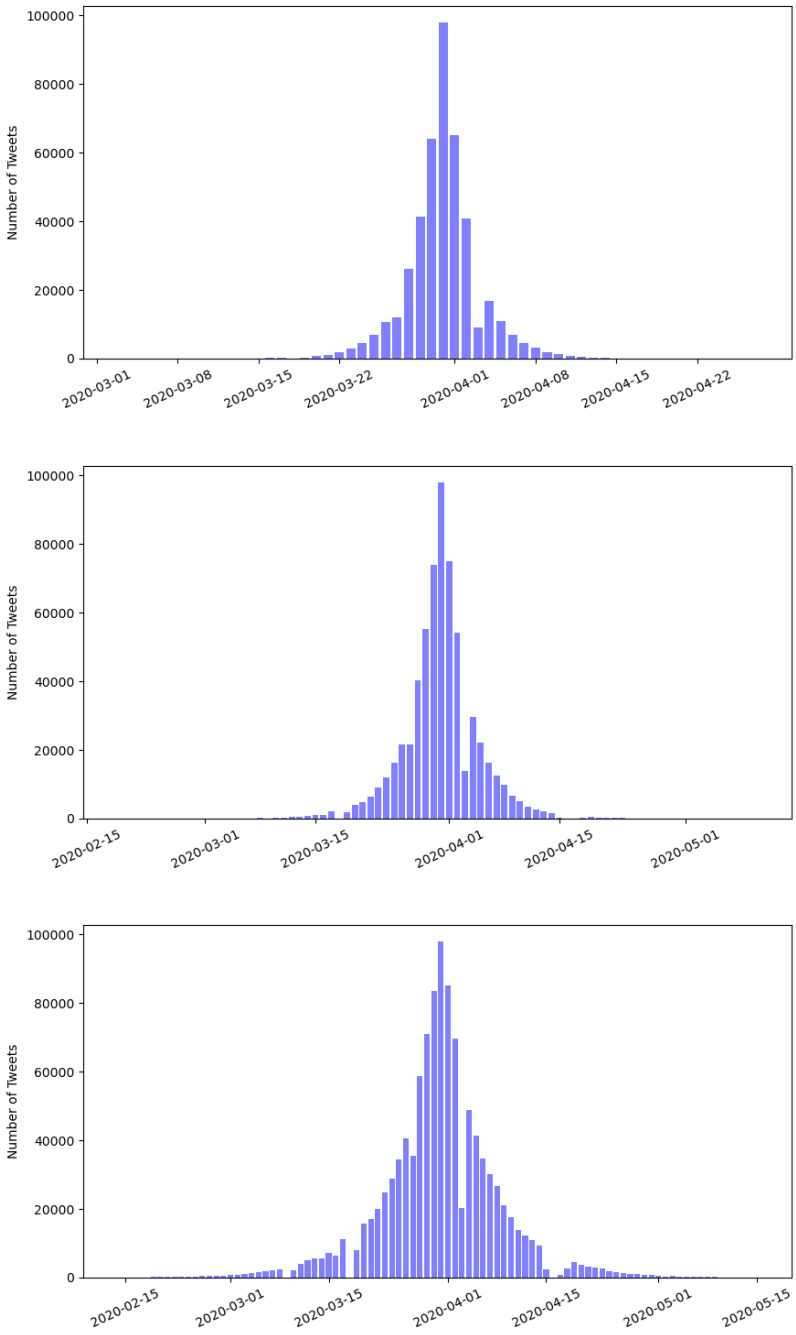


Figure H.2. Contributions of tweet volume from various time points for temporally smoothed corpora. The examples are constructed for March 31, using smoothing parameters $0.65, 0.75, 0.85$ (from top to bottom). Although the plots exhibit different resolutions and spans of the histograms, the shapes of the contribution distributions are similar in all cases. This illustrates robustness of the proposed method to the choice of smoothing parameters.

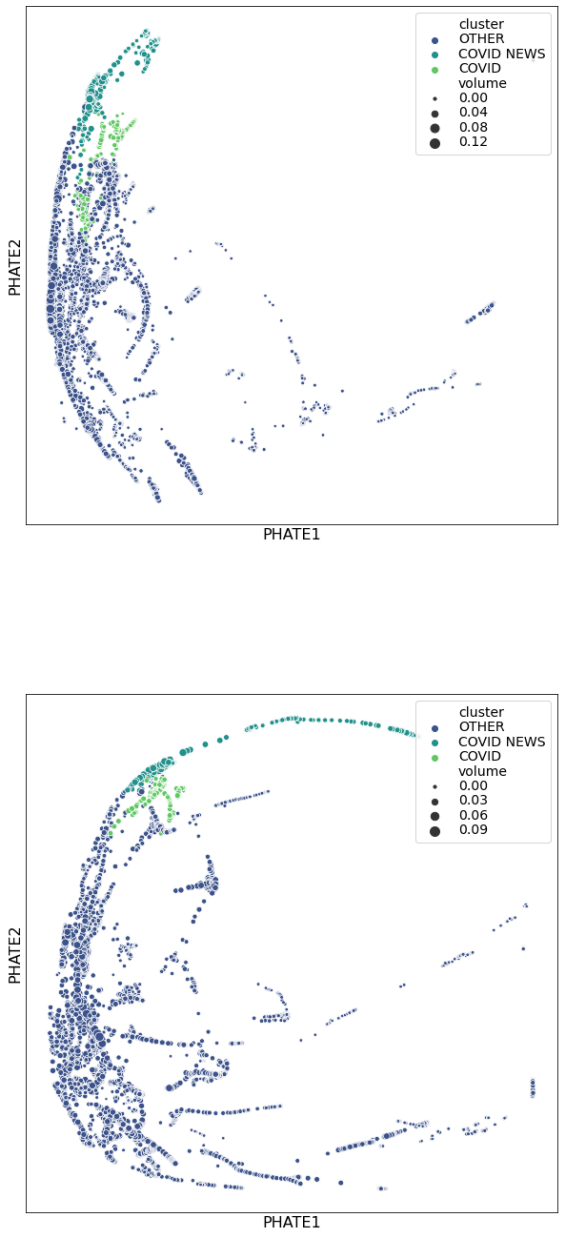


Figure H.3. Potential of heat-diffusion for affinity-based transition embedding (PHATE) for all word distributions. The topics here are learned by T-LDA on tweet collections constructed with smoothing parameters 0.65 (top) and 0.85 (bottom). Here two clusters and one shortest path are highlighted for comparison with Figure 4. Note that the overall structures as well as the trajectories for highlighted points are similar in all three cases, while the lengths of the trajectories are different, which are the result of different assumptions on the range of the temporal dependence (i.e., a smoothing using 0.85 assumes longer range dependence by including more old tweets).

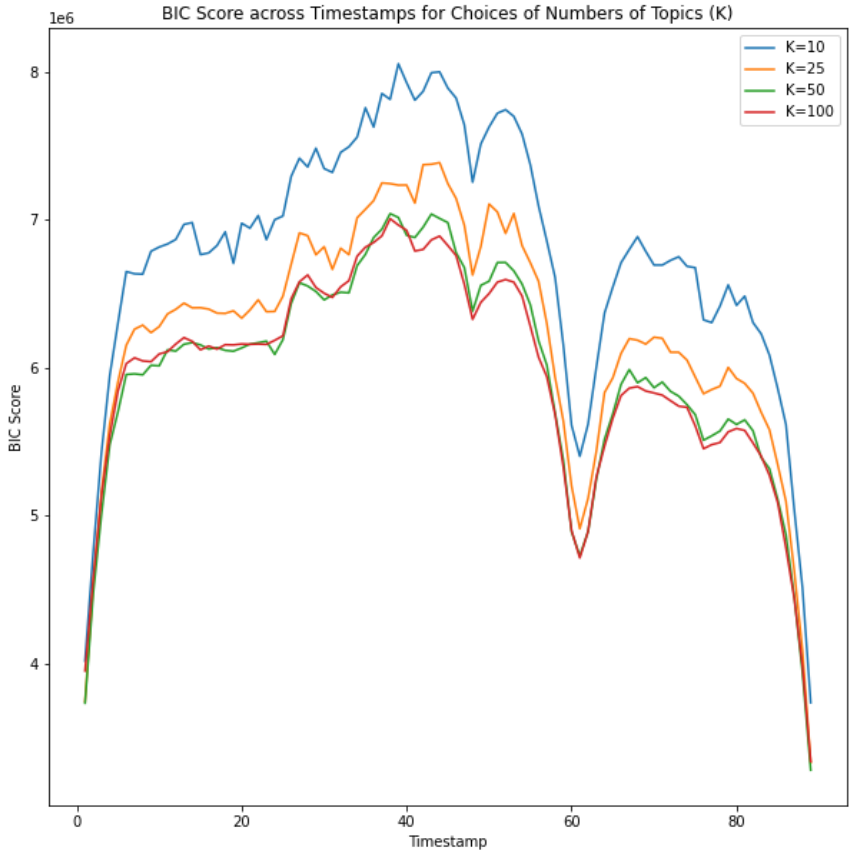


Figure H.4. Bayesian information criteria (BIC) scores across timestamps for different choices of the numbers of topics.

Appendix I. PHATE dictionary of clusters and trajectories

We explain the labeling of the PHATE plots for visualization and interpretation:

- Colors signify clusters of topics. Clusters are computed by a hierarchical clustering algorithm using Hellinger distance between topics. Only selected COVID-19 topics are colored differently, and all others are grouped into a single color. Selected COVID-19 topics are:
  - COVID: topics where the top words are mostly general COVID-19 terms such as coronavirus, virus, covid, etc.
  - COVID NEWS: topics where the top words are related to government officials or politicians discussing COVID-19 related issues. Typical top words include: Trump, government, news, covid, etc.
  - SANITIZING: topics where the top words are mostly wash hands, sanitizing, virus, etc.
  - STAY HOME: topics where the top words are mostly stay home, safe, covid, etc.
- Sizes represent normalized number of tweets that is generated from each topic.
- Shapes highlight selected COVID-19 related shortest paths computed on the neighborhood graph. Different shapes represent
  - COVID (health care): a subset of topics in the COVID topic cluster that are all on a shortest path starting from a general COVID topic at the first time point and finishing at a health care focused COVID topic (e.g., testing, death).
  - COVID (politics): a shortest path that starts from a topic that is second-closest in distance to the starting topic of the COVID (health care) and finishing at a politics focused COVID topic (e.g., president, news, etc.).
  - COVID NEWS (presidential election): a subset of topics in the COVID NEWS cluster that are all closely related to presidential election and are on a shortest path starting from a election-related topic at the first time point.
  - SANITIZING (wash hands): a subset of topics in the SANITIZING cluster that are on a shortest path starting from a topic related to washing hands due to COVID-19.
  - STAY HOME (executive order): a subset of topics in the STAY HOME cluster that are on a shortest path starting from a topic related to stay home executive order due to COVID-19.
  - General: topics that are not on selected shortest paths of interest.

Appendix J. Additional PHATE trajectories

Figure J.1 shows two linear trajectories, the SANITIZING (wash hands) and the STAY HOME (executive order), on the PHATE embedding. In contrast to nonlinear trajectories presented in Figure 5 and Figure 7, topics on linear trajectories exhibit no obvious deviation in terms of the top words.

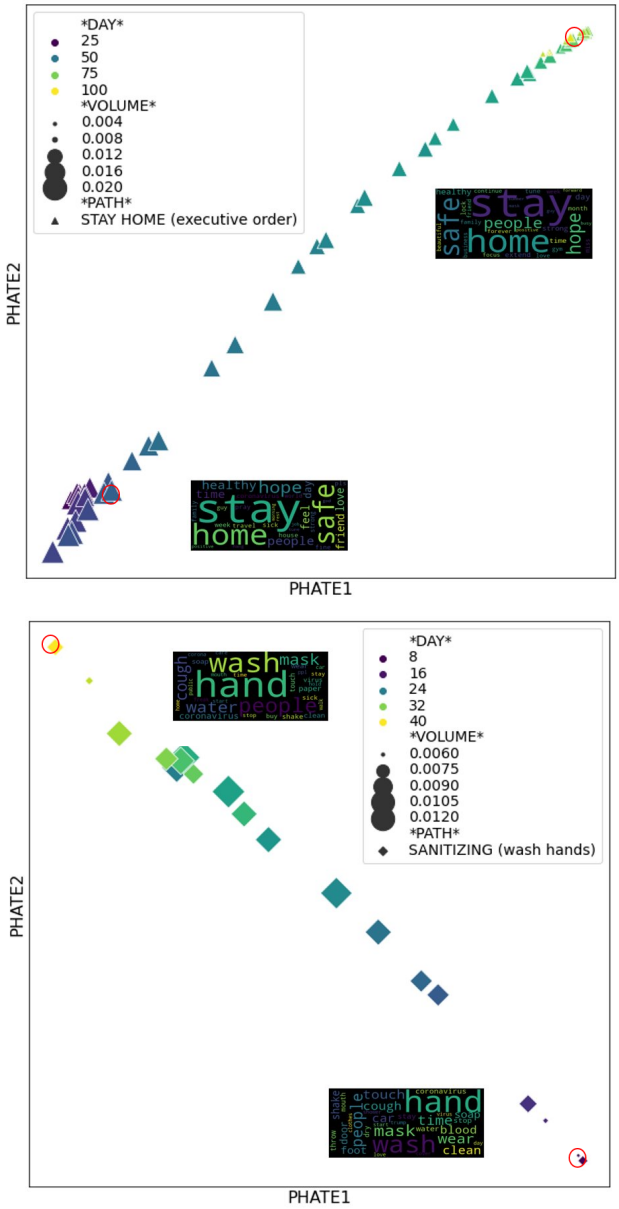


Figure J.1. Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics lie on the executive order path (top) and the wash hands path (bottom). Colors and sizes of points highlight time and tweet volume, respectively. Here two word clouds containing top 30 words in corresponding topics are shown for the time points highlighted by red circles in each path. Note that in both cases, the topic near the beginning of the study period is similar to that near the end of the study period. This shows the stability of topics on linear trajectories.

Appendix K. State-level trend in Tweet proportions

Here, we illustrate state-level variations in estimated tweet volumes generated by topics on the presidential election path, normalized by total tweet volume at each time point. From Figure K.1, we see that although tweet proportions vary state by state, the overall trend is clear with peaks roughly correspond to the time points of key event highlighted.

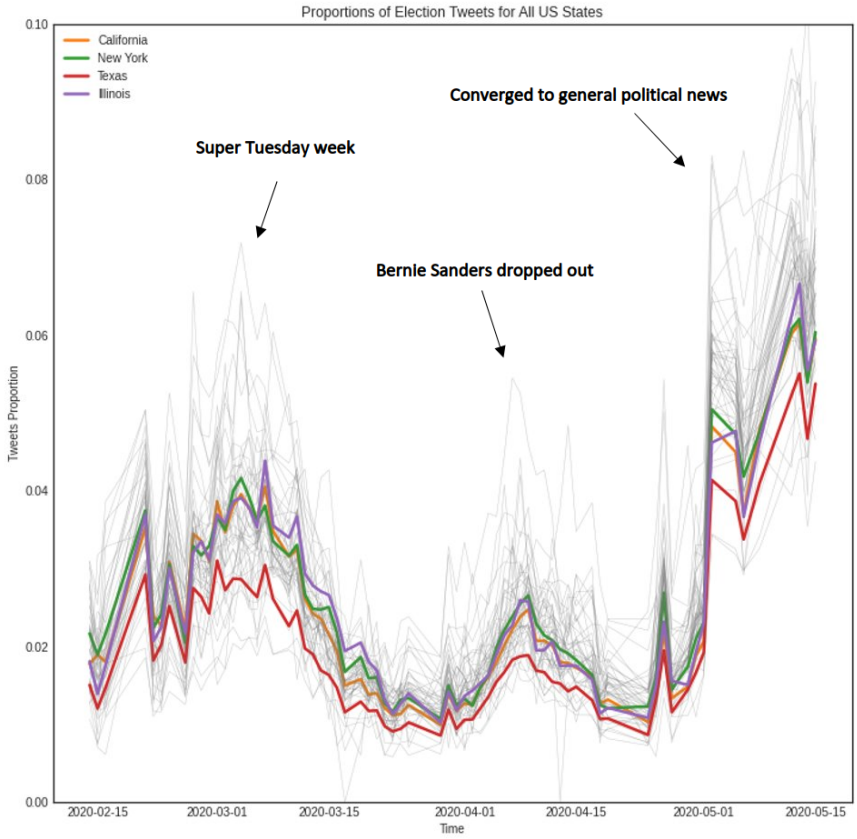


Figure K.1. State-level spatial distribution of tweet proportions generated from all topics on the COVID NEWS (presidential election) path. California, New York, Texas, and Illinois are highlighted for illustration, while all other states are ll plotted in grey. Note that similar three events (annotated using texts) as in Figure 5 correspond roughly to the three peaks in the time-course plot, indicating validations of the quality of the shortest path using real-world events.

For the COVID (health care) topic path, at the state level, tweet proportions follow global trends at the beginning of the study period in February and March but become chaotic starting in April. One possible explanation is that the COVID-19 pandemic in the United States started in several hot spots but quickly spread into other states, which then started to implement state-specific control measures. In addition, the overall new cases and death toll in the country reached a few record highs in April, starting with New York, which became an epicenter of the pandemic, with a record 12274 new cases reported on April 4 (https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state)).

This explains the difference in tweet proportions trend in New York, compared with the other three highlighted states.

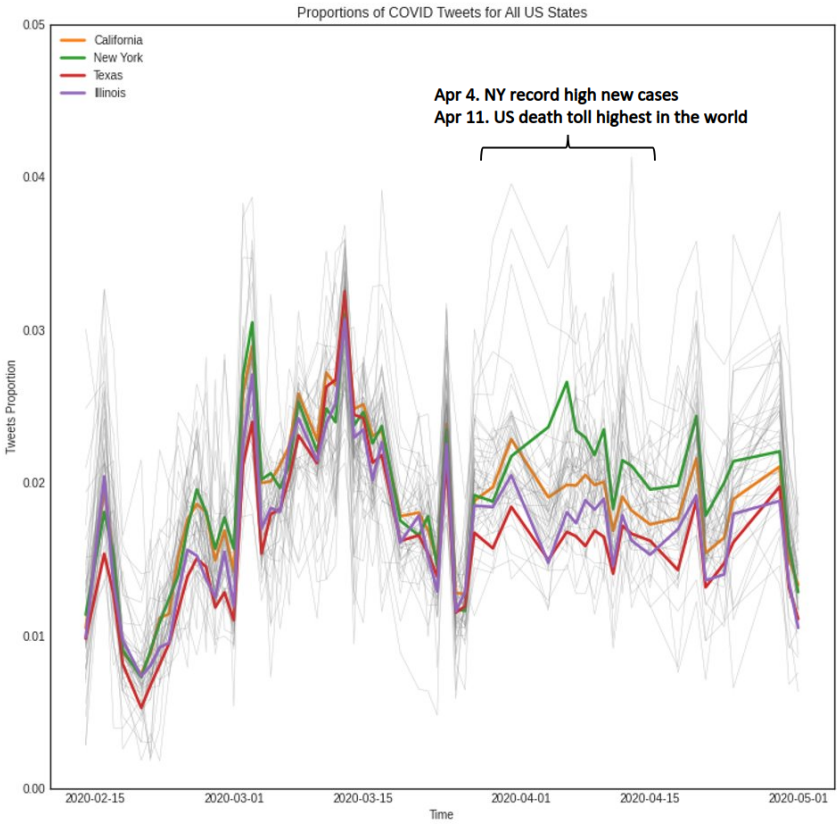


Figure K.2. State-level spatial distribution of Tweet proportions generated from all topics on the COVID (health care) path. California, New York, Texas, and Illinois are highlighted for illustration, while all other states are plotted in grey. Note that a time period in April is annotated with relevant events explaining the surge in tweet proportions in many states. This validates the quality of this shortest path using real-world events.