A Geometry-Driven Longitudinal Topic Model

Yu Wang[4] Conrad Hougen[3] Brandon Oselio[2] Walter Dempsey[25] Alfred Hero[134]

[1]Department of Biomedical Engineering

[2]Department of Biostatistics

[3]Department of EECS

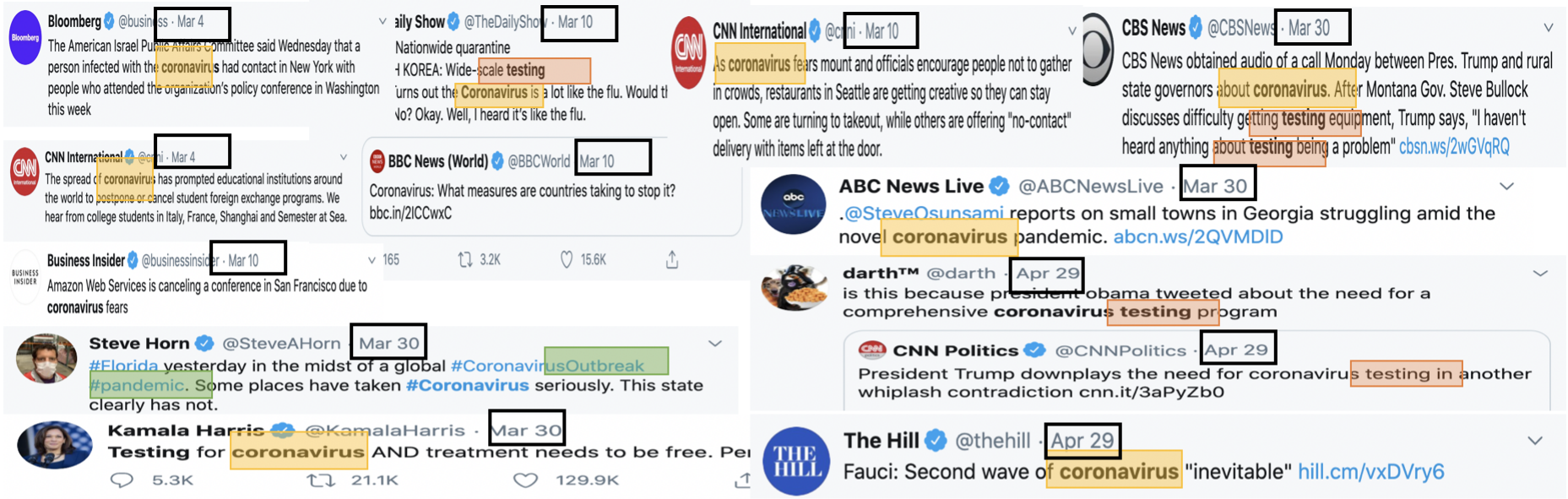
[4]Department of Statistics

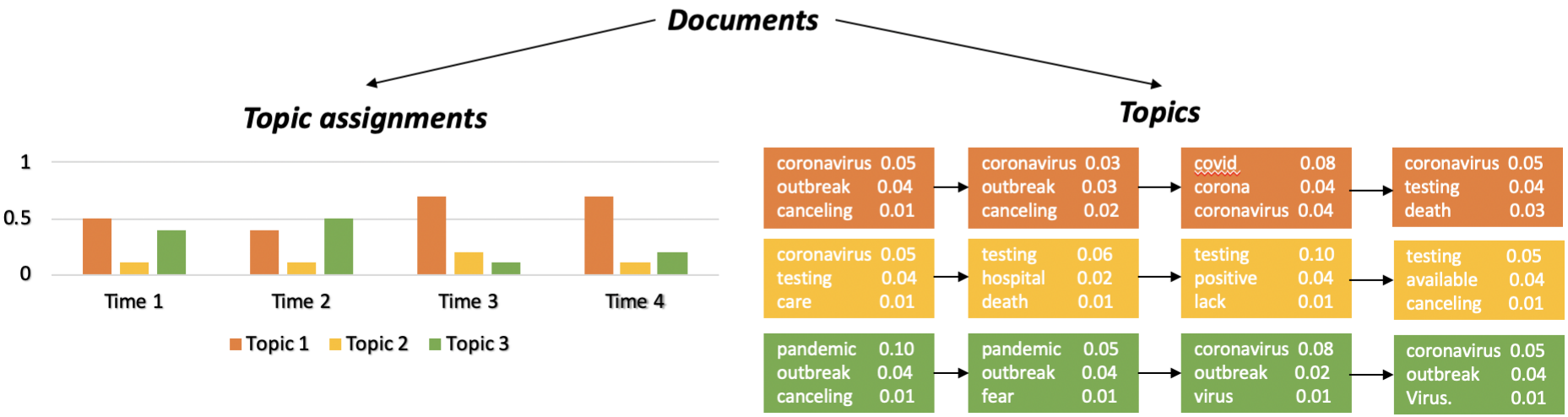
[5]Institute of Social Research

Reference

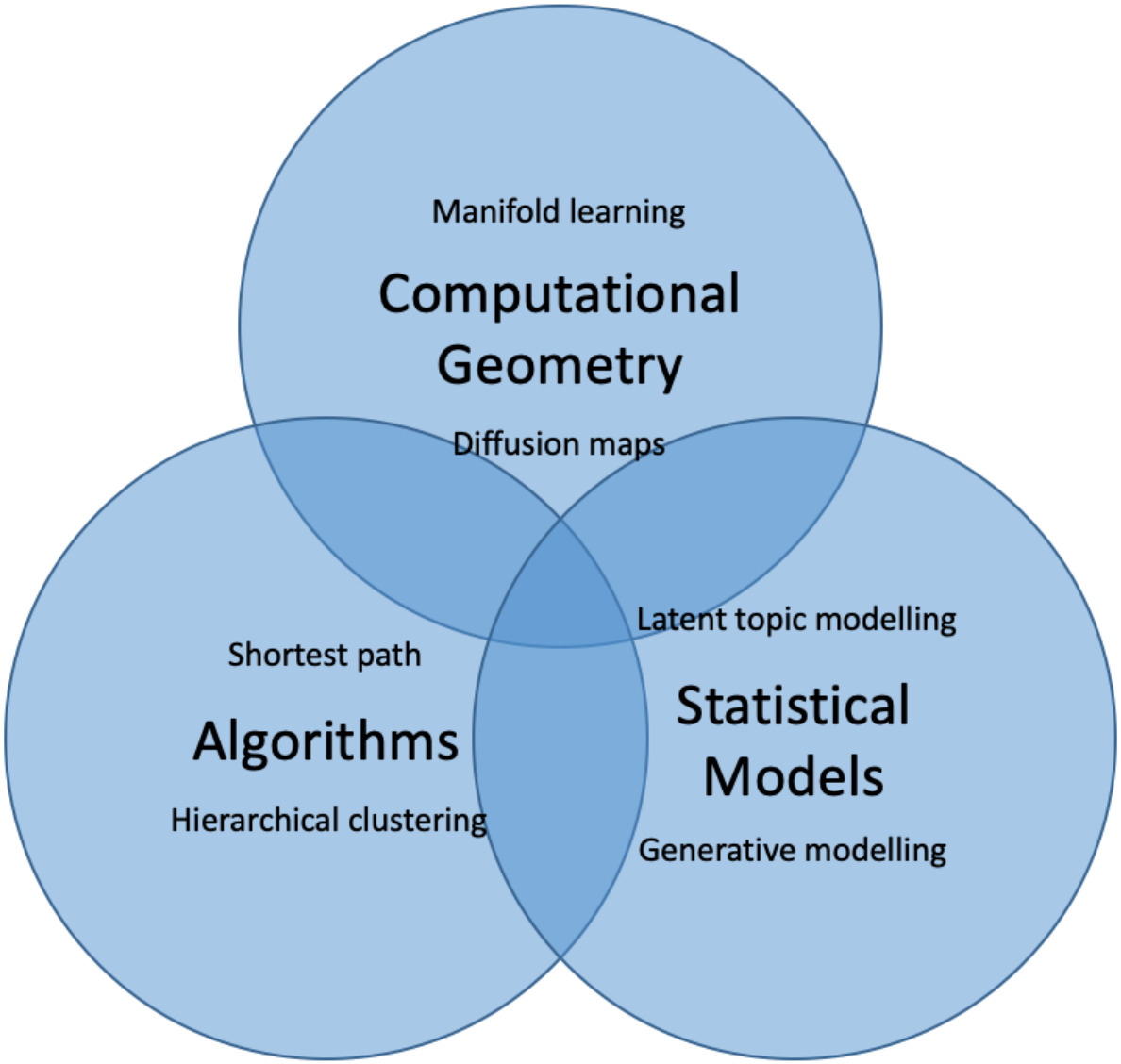
Wang, Y., Hougen, C., Oselio, B., Dempsey, W., Hero, A. O. (2021). A geometry-driven longitudinal topic model. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.b447c07e

Summarizing time-evolving texts





Data science tools



Outline

Outline

Two approaches[1][2] for time-varying topics

Dynamic topic model:

1. Draw $\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2\mathbf{I}), \forall k$
2. Draw $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2\mathbf{I})$
3. For each document $d = 1, \ldots, D$:
   1. Draw $\eta_{t,d} \sim \mathcal{N}(\alpha_t, a^2\mathbf{I})$
   2. For each word $n = 1, \ldots, N_d$:
      1. Draw topic $Z_{t,d,n} \sim \text{Multi}(\pi(\eta_{t,d}))$
      2. Draw word $W_{t,d,n} \sim \text{Multi}(\pi(\beta_{t,Z_{t,d,n}}))$.

$\Rightarrow$ Topics and temporal dynamics are learned *jointly*.
$\Rightarrow$ Computationally expensive.
$\Rightarrow$ Interpretability comes at a cost.

Time-unaware LDA:

1. Draw $\beta_k \sim \text{Dir}(\eta), \forall k$
2. Draw $\theta_d \sim \text{Dir}(\alpha), \forall d$
3. For each document $1, \ldots, D$:
   1. For each word $n = 1, \ldots, N_d$:
      1. Draw topic $Z_{d,n} \sim \text{Multi}(\theta_d)$
      2. Draw word $W_{d,n} \sim \text{Multi}(\beta_{Z_{d,n}})$.

$\Rightarrow$ Pre-divides the data into discrete time slices and use separate LDA to learn topics are learned *marginally*.
$\Rightarrow$ Post-align the topics from each time slice.
$\Rightarrow$ Accurate alignment could be difficult.

[1] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[2] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

Aligning topics from each time stamp

Goal: Apply time-unaware LDA – stitch together marginal topics – connecting a topic at time t to another topic at $t+1$ or a future time stamp such that the **distance** between them is small and **natural transition** is preserved.

Previous work:

- Euclidean or Cosine distance [3] is commonly applied[4][5][6].
- "Non-natural" mechanism for alignment via e.g., distance thresholding[7], clustering[8].

[3] $1 - \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}$.

[4] Jason Chuang et al. "Topic model diagnostics: Assessing domain relevance via topical alignment". In: *International Conference on Machine Learning*. 2013, pp. 612–620.

[5] Jason Chuang et al. "TopicCheck: Interactive alignment for assessing topic model stability". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 175–184.

[6] Michelle Yuan, Benjamin Van Durme, and Jordan L Ying. "Multilingual anchoring: Interactive topic modeling and alignment across languages". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8653–8663.

[7] Sana Malik et al. "TopicFlow: visualizing topic alignment of Twitter data over time". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2013, pp. 720–726.

[8] Weiwei Cui et al. "TextFlow: Towards better understanding of evolving topics in text". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2412–2421.

Aligning topics from each time stamp: Metric

Hellinger distance between discrete probability distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$:

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

Because:

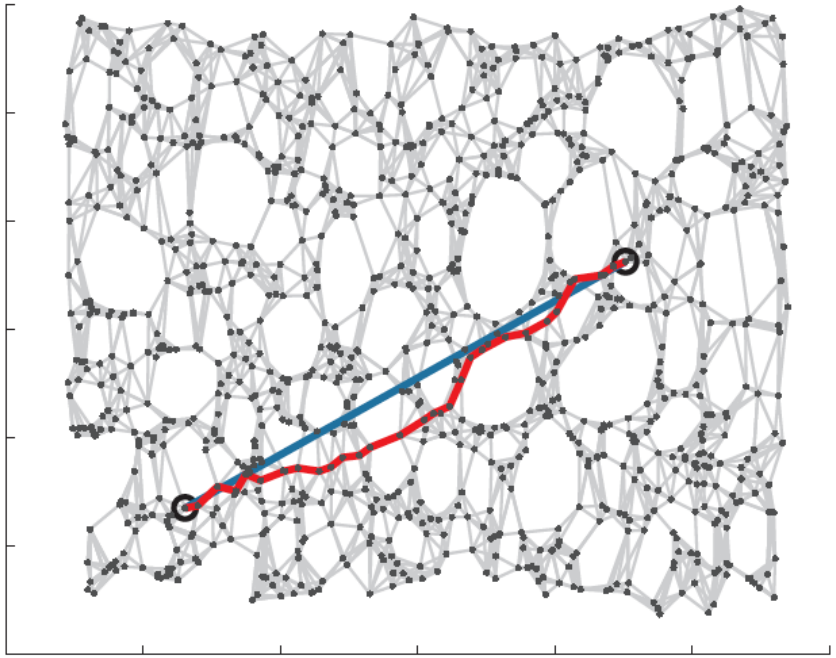
- Closely related to Bhattacharyya distance that is used to, e.g., measure the separability of classes in classification.
- A metric well adapted to probability distributions. (comparing to KL-divergence and Euclidean distance[9])
- Computationally simple. (comparing to Wasserstein distance)
- An "inferential" distance between PDFs in the absence of the geometry of the statistical manifold on which the PDFs lie[10].

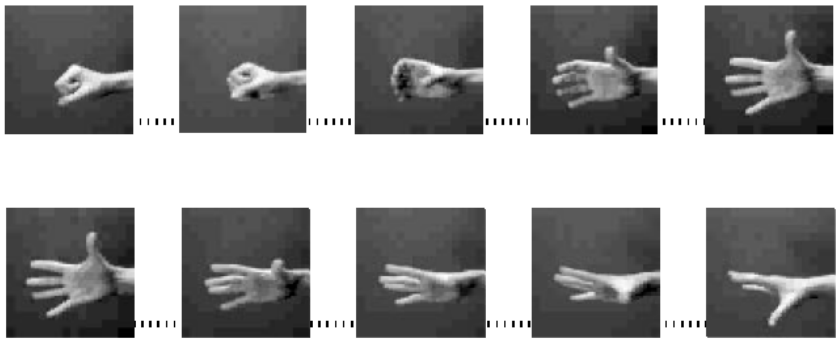
[9] Shun-ichi Amari. *Differential-geometrical methods in statistics*. Vol. 28. Springer Science & Business Media, 2012.

[10] Kevin M Carter et al. "Fine: Fisher information nonparametric embedding". In: *IEEE transactions on pattern analysis and machine intelligence* 31.11 (2009), pp. 2093–2098.

Aligning topics from each time stamp: Algorithm

Nearest neighbor graphs and shortest paths has been widely used to capture natural transitions[11]. For example:





Shortest path distance on neighborhood graph captures perceptually natural but highly nonlinear morphs of the corresponding high-dimensional observations by transforming them approximately along geodesic path.

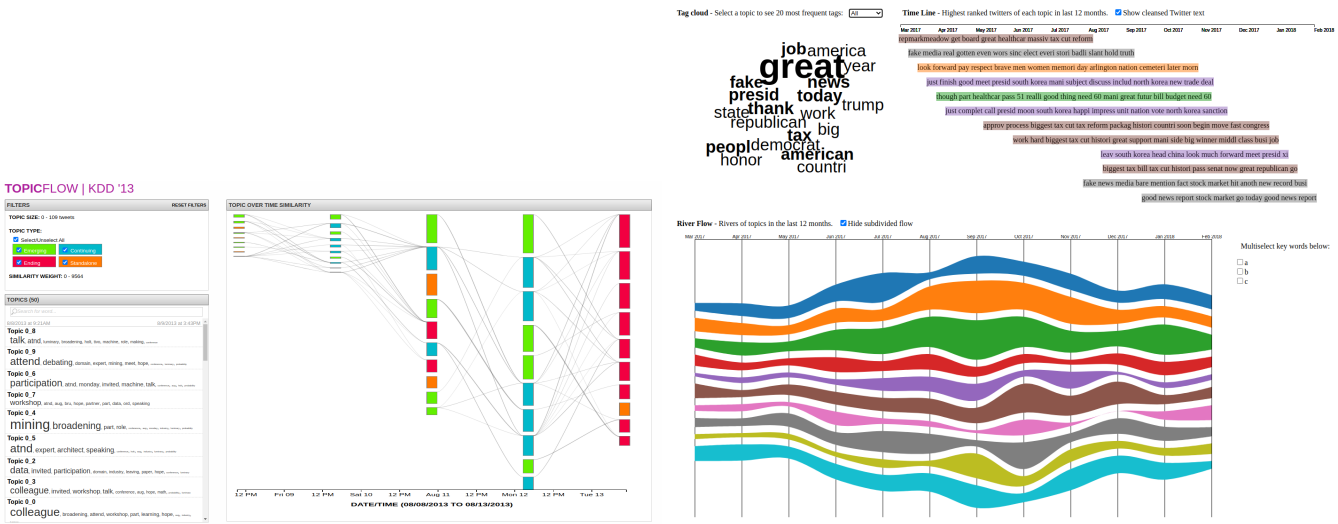
[11] Joshua B Tenenbaum, Vin De Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

Interpretation and visualization of topic trends

Goal: Visually effective ways to interpret the high-dimensional topic trends and possibly discover new trends.

Previous work:

- Based on graphical heuristics instead of principled math/stats models[12][13].
- Allows only visualizations of the summary statistics, e.g., topic volumes, keywords frequencies, etc.



[12] Sana Malik et al. "TopicFlow: visualizing topic alignment of Twitter data over time". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2013, pp. 720–726.

[13] Weiwei Cui et al. "TextFlow: Towards better understanding of evolving topics in text". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2412–2421.

Interpretation and visualization via dimensionality reduction

Our pipeline does facilitate effective interpretation and visualization of topic trends, but has the advantage of being based on principled methods from computational geometry.

Some considerations:

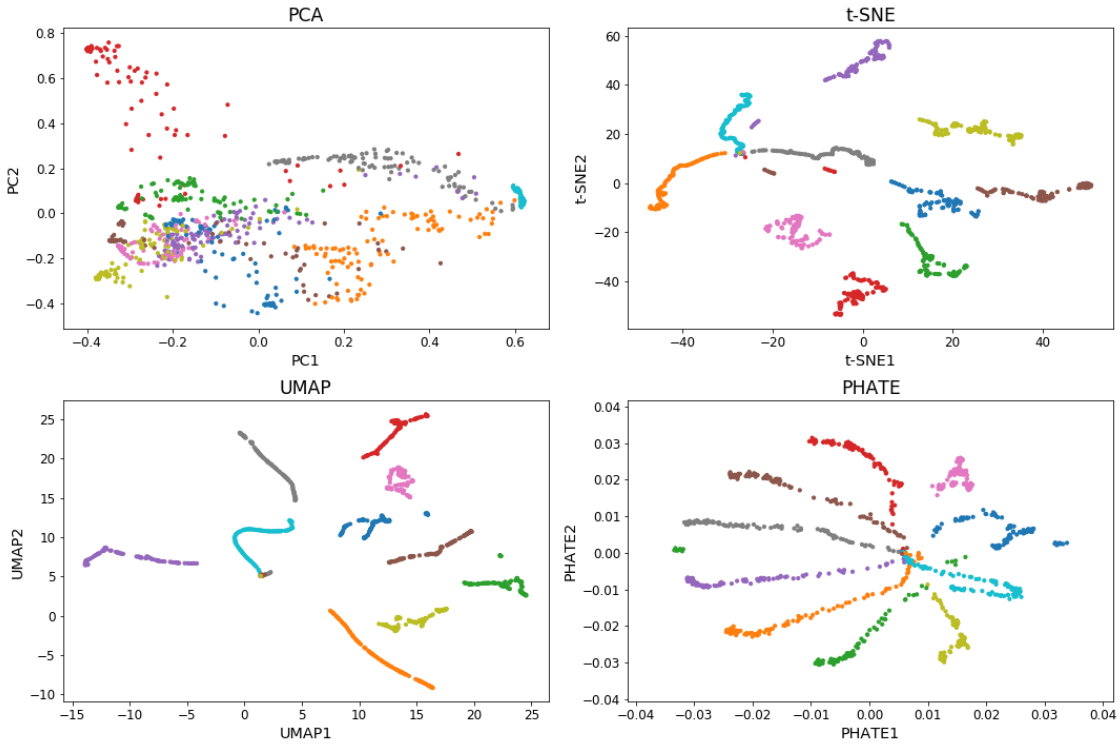
- To visualize and interpret high dimensional word distributions, need *lower-dimensional embedding* that capture the intrinsic high-dimensional *trajectory structure* of the data.
- Traditional methods like PCA assumes linearity.
- Nonlinear methods like t-SNE[14], UMAP[15], etc. do not naturally exhibits trajectory or progression.
- PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding)[16] is designed explicitly to preserve progression structure in data.

[14] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE.". In: *Journal of machine learning research* 9.11 (2008).

[15] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[16] Kevin R Moon et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature Biotechnology* 37.12 (2019), pp. 1482–1492.

PHATE illustration



PCA, t-SNE, UMAP, and PHATE dimensionality reduction methods applied to 2D embedding of simulated 10 trajectories (identified by color) of 100-dimensional probability vectors, all originating from a common initial point:

$$X_t^j | X_{t-1}^j \sim \mathcal{N}_{100}(X_{t-1}^j, \sigma_j^2 I) \quad j = 1, \ldots, 10 \quad t = 0, \ldots, 99.$$

Outline

Data preparation

Twitter Decahose API [17] was used. We focused on a time period from **February 15, 2020 to May 15, 2020** and applied following filtering:

- **US geographic area**: Tweets that are geotagged and originated in the US 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 pre-processing 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.

[17] `https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/decahose`. We thank the Michigan Institute of Data Science (MIDAS) for providing access to this dataset.

Procedure for Twitter trend analysis

Algorithm 1 Procedure for longitudinal analysis of Twitter data.

Input: Raw Twitter data

1: Pre-process Twitter data and organize Tweets into a temporally smoothed corpus [a].

2: Apply T-LDA [b] independently to each corpus with $K = 50$ topics [c]. This results in 4500 word distributions.

3: Compute all pairwise Hellinger distances for 4500 word distributions.

4: Compute:

a: k-nearest neighbor graph with $k = 10$ [d] from the 4500-by-4500 Hellinger matrix and find the shortest path of interest on the neighborhood graph using the Djikstra algorithm.

b: PHATE embedding of 4500 high-dimensional points in 2D or 3D.

Output: Shortest paths and PHATE coordinates.

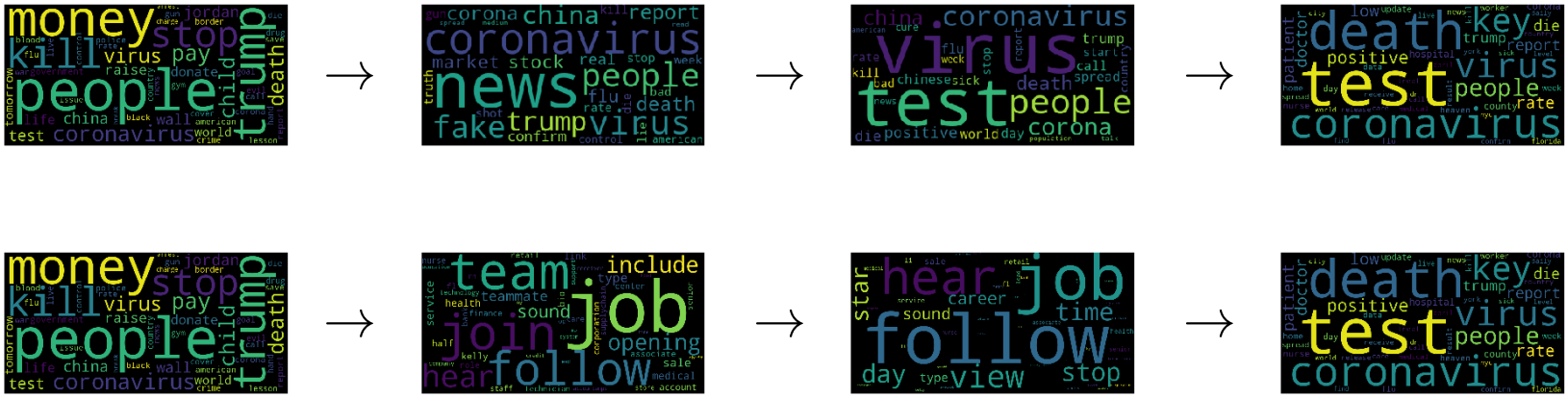
[a] Similar to the smoothing idea used in LOESS (locally estimated scatterplot smoothing) regression, for example.

[b] We restricted the model to allow for only one topic per Tweet.

[c] A BIC-type criteria was used to select the approximately optimal number of topics.

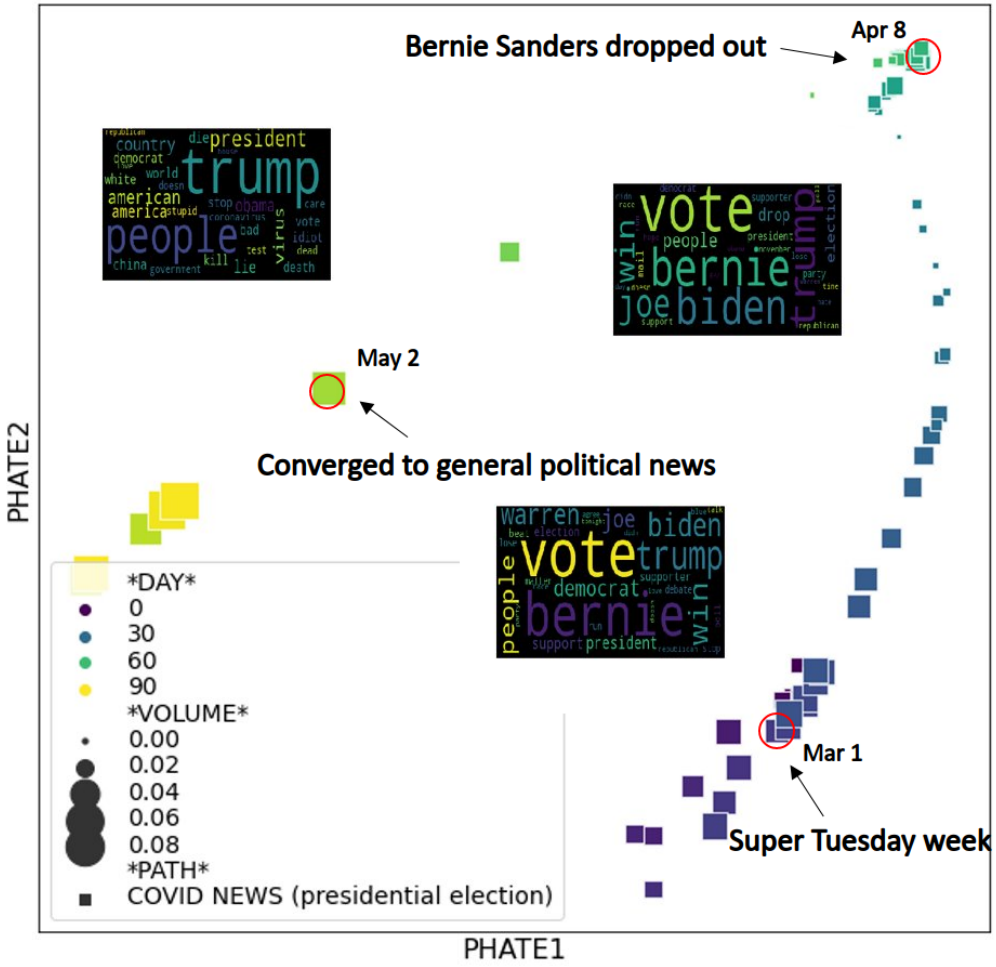
[d] A sensitivity analysis was performed for this choice of the number of neighbors.

Shortest paths of topics

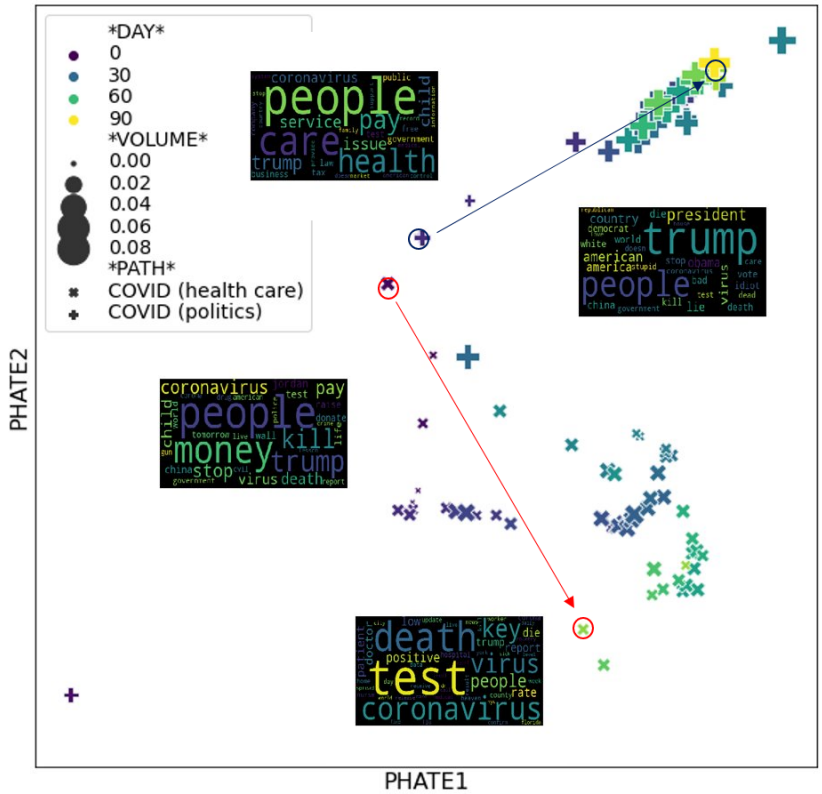


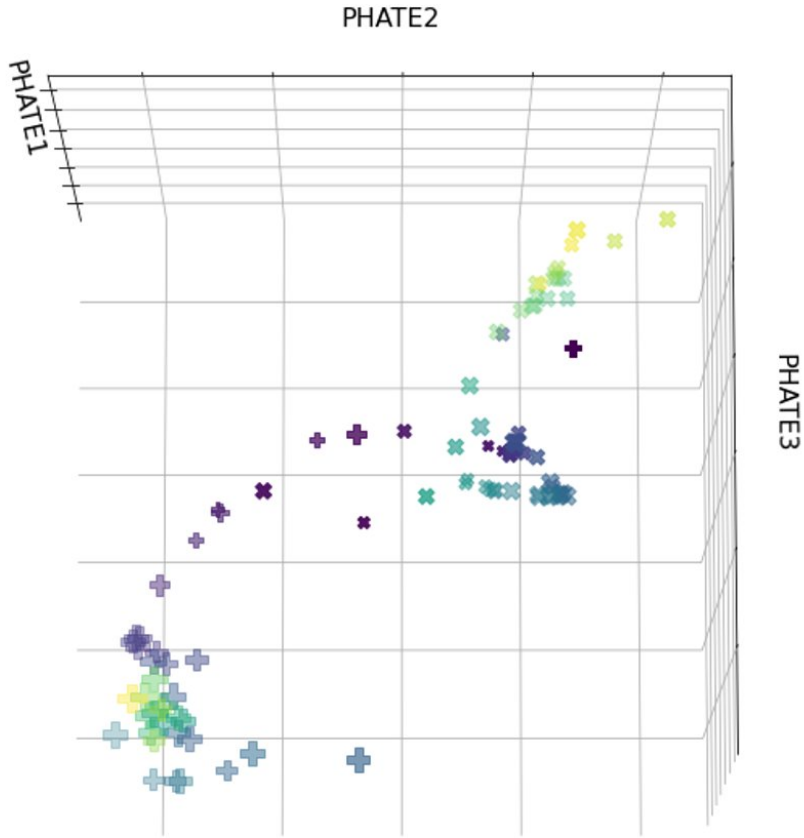
Evolution along the shortest paths of a COVID-19 topic on the first day to a COVID-19 health care focused topic on the last day illustrated as top 30 words, computed on a 10-nearest **neighbor graph (top)** and a **fully connected graph (bottom)**. The middle two wordclouds are illustrations of two of the topics on the paths. Note that the middle two topics on the top row represent natural transformations from the beginning to the end topics, whereas on the bottom row they are barely correlated with the those two topics.

Case study I: Presidential election path



Case study II: Two COVID-19 paths





Outline

What's covered

In this talk we have seen

- A modular framework that provides a wrapper for a suite of tools for learning, interpreting, and visualizing of temporal topic models. See `https://github.com/ywa136/twitter-covid-topics` for all the learned topics, code, and an accompanying ShinyApp for more exploration.
- 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.

What's next

Akin to Rohe et al.[18] where the consistency of spectral clustering algorithms is proved under the Stochastic Blockmodel for network data, we are analyzing the theoretical performance of the geometry-driven "nonparametric" topic model under the parametric DTM[19]:

- Random-walk "path" in DTM: $\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2\mathbf{I})$.
- How close are the shortest paths to the random-walk path in appropriate sense?

$\Rightarrow$ Distances between sequences of probability distributions vanish as D, $N_d, \forall d$, and K grow (high-dimensional setting).

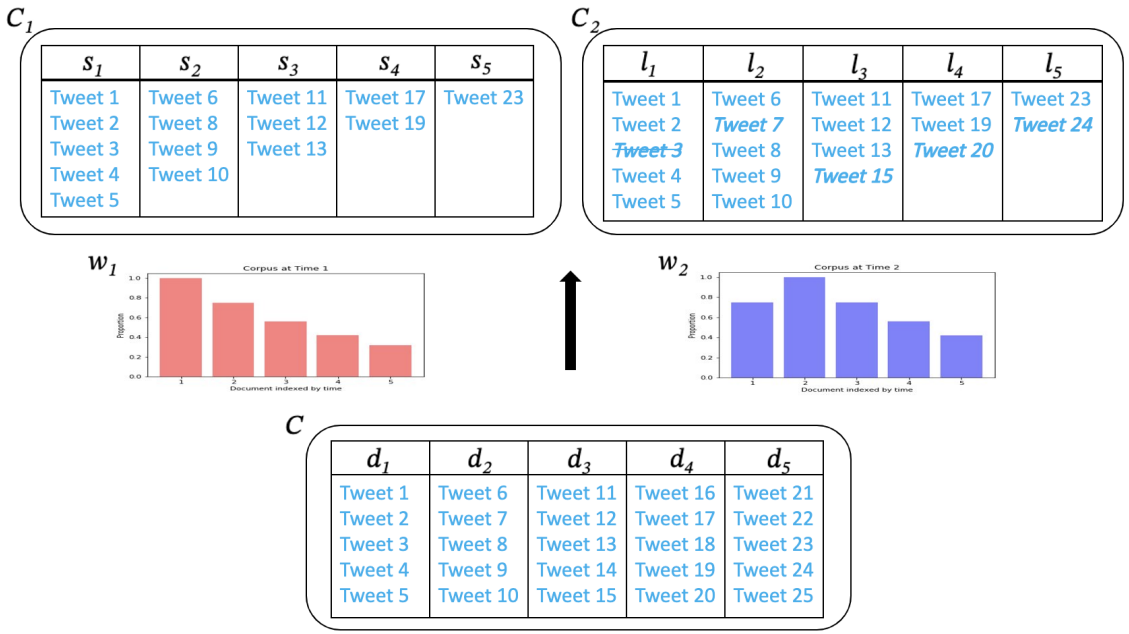
[18] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. "Spectral clustering and the high-dimensional stochastic blockmodel". In: *The Annals of Statistics* 39.4 (2011), pp. 1878–1915.

[19] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

The end

Thank you! Questions?

Temporally smoothed corpora



Conditional subsampling procedure using a hypothetical corpus composed of 5 documents each containing 5 Tweets ($C = \{d_1, \ldots, d_5\}$), e.g., d_1 aggregates Tweets from day 1, d_2 aggregates Tweets from day 2, etc. 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.

T-LDA details

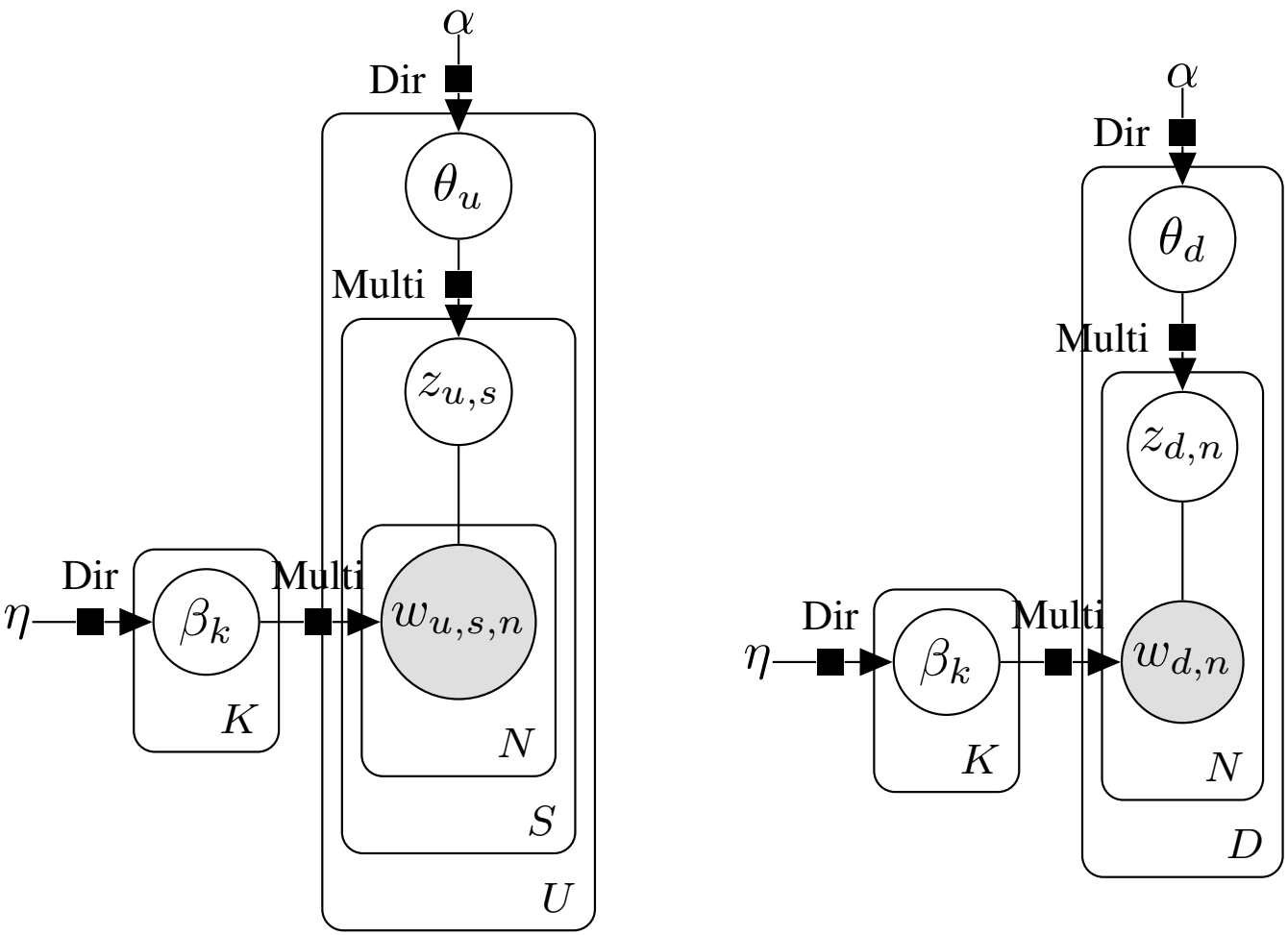


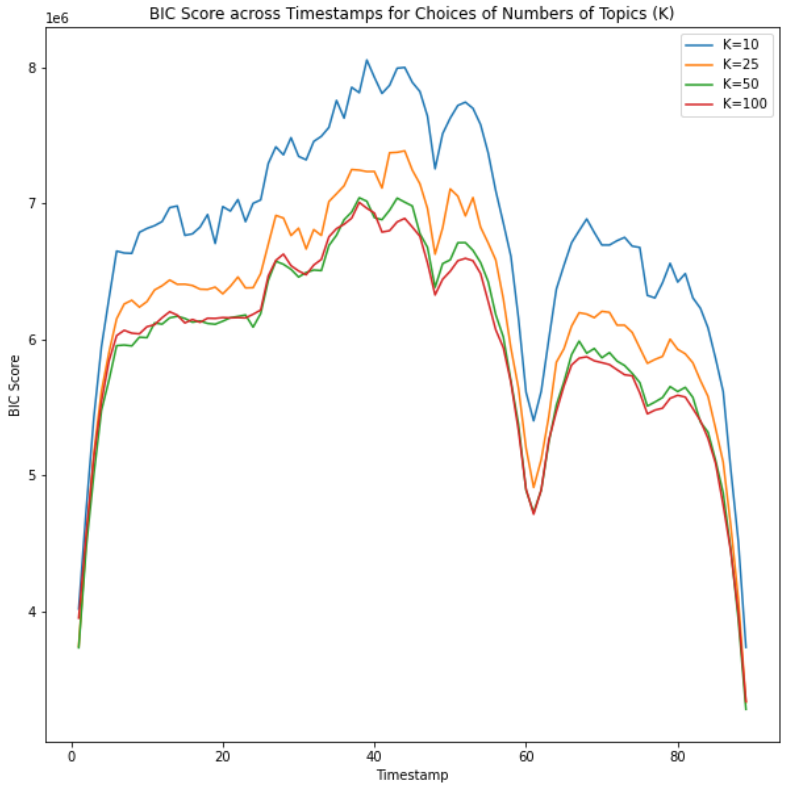
Plate notation comparison for the T-LDA (left) and the standard LDA (right) models.

T-LDA inference details

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.

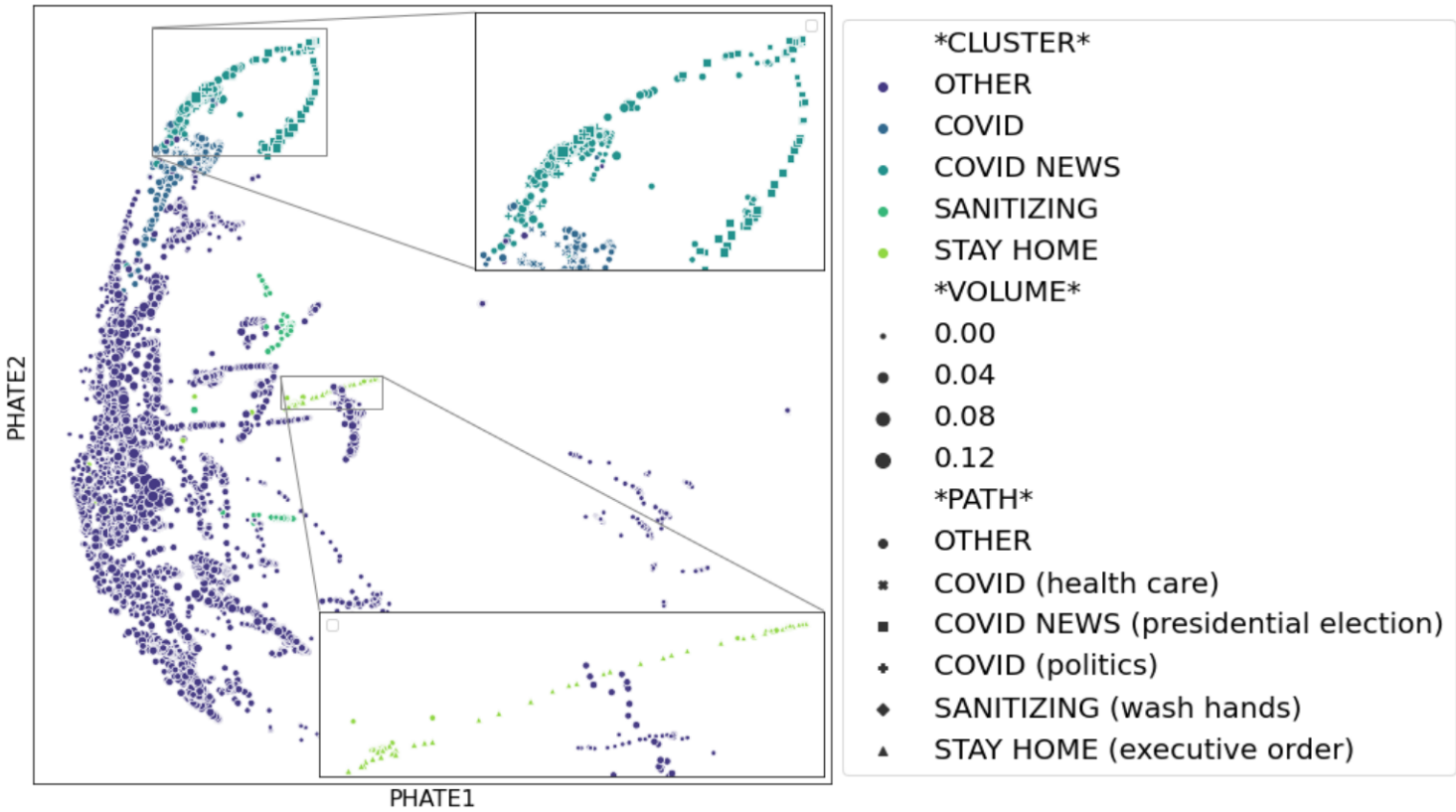
BIC for selecting K

We computed a BIC-type score[20] for topic model selection:



[20] Matt Taddy. "On estimation and selection for topic models". In: *Artificial Intelligence and Statistics*. PMLR. 2012, pp. 1184–1193.

The whole picture



PHATE embedding 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. Note that the embedding captures some important clustering/trajectory structures, e.g., branching, splitting, merging, etc.