Investigating text data using Topological Data Analysis

Max Turgeon

11 March 2021

University of Manitoba
Departments of Statistics and Computer Science

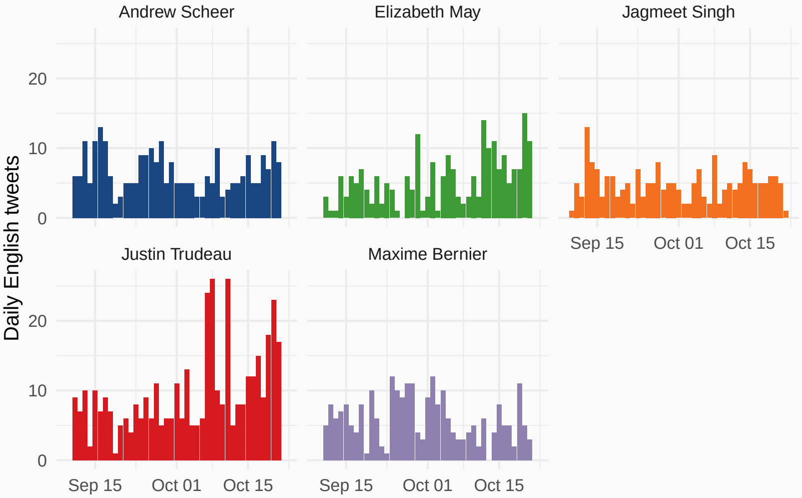
Motivation

- In my applied multivariate analysis course, students have to analyze a dataset of their choice using tools we learned about.
- One student (J. Hamilton) wanted to analyze tweets from Canadian party leaders during the last election campaign.
- **Main objective**: Can we uncover the main debate themes solely from the text data?
- This project was further expanded during his honours thesis.
  - The main conclusion: yes for some themes, but the signal is weak.
- But the question remains: *What is the best way to analyze such data¿?*

Dataset

- 1356 English tweets from 5 party leaders:
  - Andrew Scheer (CPC), Elizabeth May (GPC), Jagmeet Singh (NDP), Justin Trudeau (LPC) and Maxime Bernier (PPC)
  - Yves-François Blanchet (BQ) was ignored, since he only tweeted once in English
- Tweets were posted between September 10 and October 22 2019.
- JT tweeted the most (401), Jagmeet Singh tweeted the least (210)

Number of daily tweets



Data cleaning and preparation

- Each tweet was split into a collection of words (“bag-of-words” model)
- Hashtags, mentions, stop-words, emojis, and numerical digits were removed.
- Tweets with less then 4 tokens were removed.
- *Output*: a 1256 by 3932 document-term matrix.

Most common words

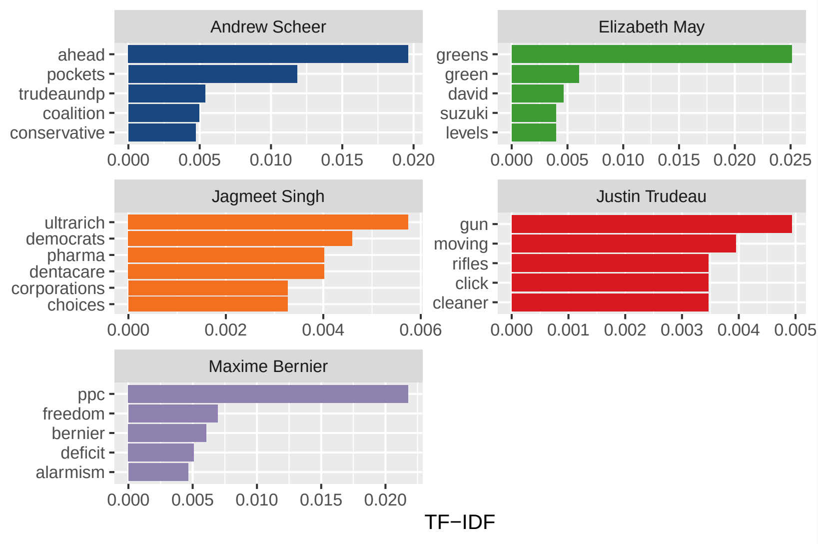
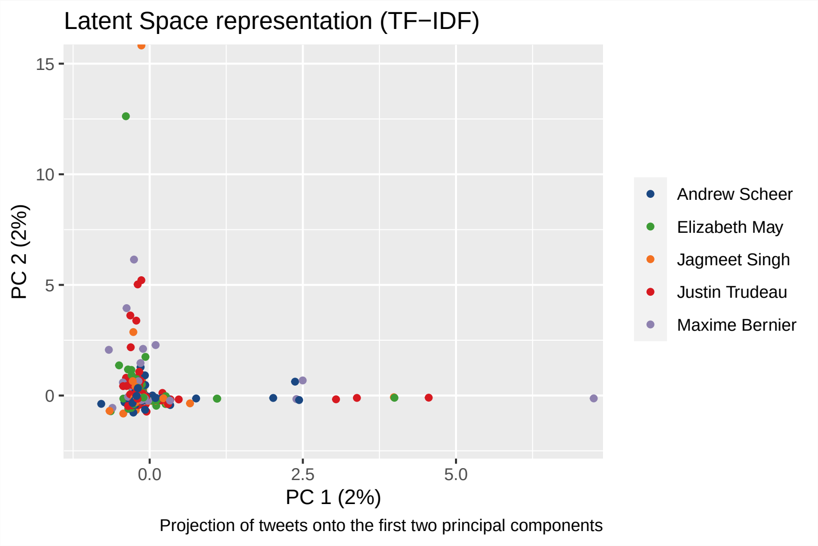


Figure 1: TF-IDF: Term frequency-Inverse Document Frequency

PCA of document-term matrix



Observations

- PCA doesn't work very well...
    - These rays are evidence of non-normality (e.g. Rohe & Zeng, 2020).
- The matrix is over 99% sparse!!!
- The data is high-dimensional.

Curse of Dimensionality

High-dimensional data suffers from the **curse of dimensionality**:

- Lots of empty space, neighbouring observations are far apart.
- Most points are far away from the mean.
- Computational challenges

Is the curse as bad as it seems?

Big data matrices are approximately low rank

- Empirical observations suggest that it's not.
- **Why?** These cursed results are derived by assuming variables are independent.
- This is clearly almost always false.
- And actually, big data matrices are approximately low-rank (Udell and Townsend, 2019)

Mitigating the curse

What does it mean? We can mitigate the effects of this curse by exploiting the *structure* in our data.

- Use sparse methods (e.g. penalized regression, graphical lasso)
- Use structured covariance estimators (e.g. Turgeon *et al*, 2018)

Topological Data Analysis

- **Topological Data Analysis** (TDA) allows to study the geometry of the sample space.
- It has its roots in computational geometry, computational linear algebra, etc.
- It has been successfully used to study *constraints* in the sample space.
- We will look at two different techniques from TDA:
    1. Persistent homology
    2. Mapper algorithm

Crash course in Algebraic Topology

- **Topology** is a field of (pure) mathematics studying shapes independently of coordinate systems and distance functions.
  - "Primitive" geometry
- **Algebraic topology** uses tools from abstract and linear algebra to study and classify shapes.
- **Homology** attaches a series of vector spaces to shapes.
  - The dimension of these vector spaces are called the **Betti numbers**.
- **Important**: these vector spaces are invariant under continuous deformations of the shapes.
- The Betti numbers count important topological features:
  - Connected components, holes, cavities, etc.

Where is the data?

- We will assume our high-dimensional data lives on, or near, a lower-dimensional *manifold*.
    - One way to model data constraints.
- A finite dataset is not a very interesting topological space...
- **Main idea**: From our data, construct a *simplicial complex*, which is a very interesting topological space!

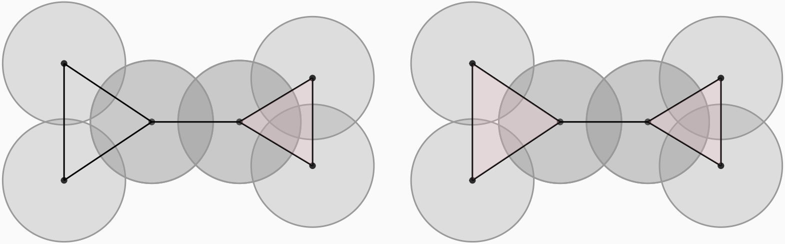
Simplicial complex

- A **simplicial complex** is a topological space constructed as the convex hull of subsets of points (called a simplex).
    - Plus some internal consistency rules
- We will consider two different constructions (let $r > 0$):
    - **Cech complex**: the data points are simplices; a subset of k points is a simplex if the intersection of open balls of radius r around them is non-empty.
    - **Vietori-Rips complex**: the data points are simplices; a subset of k points is a simplex if they are all at most distance $2r$ apart of each other.

Cech and Vietori-Rips complex

Cech

Rips



Nerve Theorem

- An important theorem in algebraic topology (called the *Nerve theorem*) can be leveraged to give consistency results about the topology of the Cech complex and that of the support of our data.
- **What about Vietori-Rips?** Because we have

$$\mathrm{Rips}_r(\mathcal{X}) \subseteq \mathrm{Cech}_r(\mathcal{X}) \subseteq \mathrm{Rips}_{2r}(\mathcal{X}),$$

  we can translate those consistency results to the Vietori-Rips complex.

In other words, given enough points, a well-behaved sample space, and an appropriate radius r, we can compute the Betti numbers of our sample space using the Cech (or Vietori-Rips) simplicial complex.

Persistent homology

- **How can we choose the right r?** We don't have to choose!
- **Persistent homology** computes the homology of a sequence of simplicial complexes:

$$\mathrm{Cech}_{r_1}(\mathcal{X}) \subseteq \mathrm{Cech}_{r_2}(\mathcal{X}) \subseteq \mathrm{Cech}_{r_3}(\mathcal{X}) \subseteq \cdots$$

- For very small r: $\beta_0 = n; \beta_k = 0$ for $k \leq 1$
- For very large r: $\beta_0 = 1; \beta_k = 0$ for $k \leq 1$
- We are looking for topological features that *persist* over long ranges of r.

Barcodes and diagrams

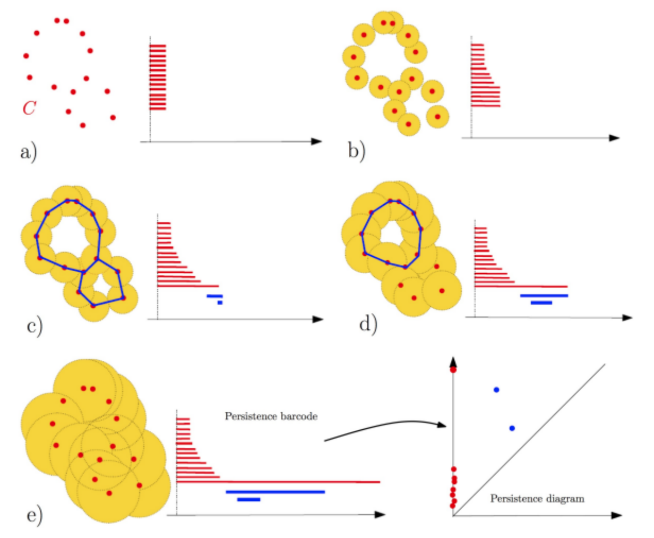
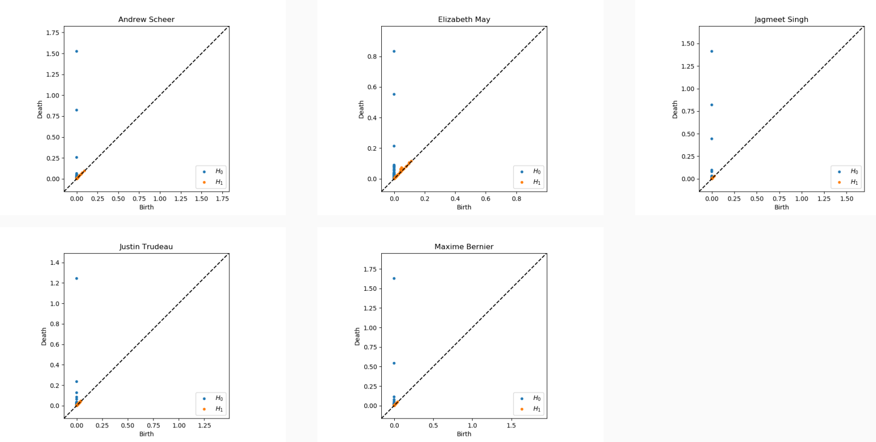


Figure 2: Chazal & Michel, Arxiv 2017

Data analysis

- Split document-term matrix into 5 smaller matrices
    - One for each leader
- Reduce dimension using PCA
- Compute persistent homology on reduced data
- Analysis conducted in Python using `scikit-tda`

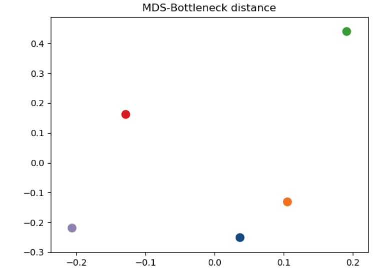
Persistence diagrams



Bottleneck distance

Bottleneck distance to compare 0-th homology; visualize using Multidimensional Scaling.

	AS	EM	JS	JT	MB
AS		0.67	0.20	0.42	0.30
EM	0.67		0.56	0.39	0.78
JS	0.20	0.56		0.41	0.28
JT	0.42	0.39	0.41		0.39
MB	0.30	0.78	0.28	0.39	



Discussion

- The two most distant leaders are from the two smallest parties (EM and MB)
- AS and JS have a similar goal: replacing JT. Could explain why they are close.
- JS is closer to EM than MB, and the opposite is true for AS.
- In fact, by rotating the points, the x-axis could order the leaders from left to right in the "expected" order.

Discussion

- More generally, persistent homology can be used for feature extraction in text data (Gholizadeh *et al*, 2020), document clustering and classification (Guan *et al*, 2016).
- Persistence diagrams can be embedded in Hilbert spaces, and therefore used with kernel methods for prediction
- It is also gaining popularity in other fields, e.g. social network analysis (Almgren *et al*, 2017), change-point analysis (Islambekov *et al*, 2019), understanding deep neural networks (Gebhart *et al*, 2019)

Mapper algorithm

- Data visualization for high-dimensional datasets
- Alternative to manifold learning and dimension reduction
- **Main idea**: Study the data by looking at its image under a function $f : \mathbb{R}^p \to \mathbb{R}$.

Mapper algorithm

Algorithm
Input: A dataset $\mathcal{X}$, a function $f : \mathbb{R}^p \to \mathbb{R}$, a set $\mathcal{U}$ of intervals covering the image $f(\mathcal{X})$.

- For each interval $U \in \mathcal{U}$, cluster the pre-image $f^{-1}(U)$.
- For each cluster, draw a node.
- Connect a pair of nodes if their corresponding clusters have a non-trivial intersection.

Output: A graph (or network).

Mapper algorithm

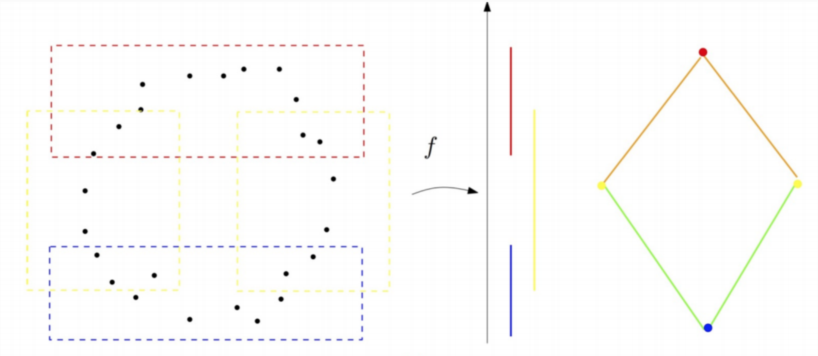


Figure 3: Chazal & Michel, Arxiv 2017

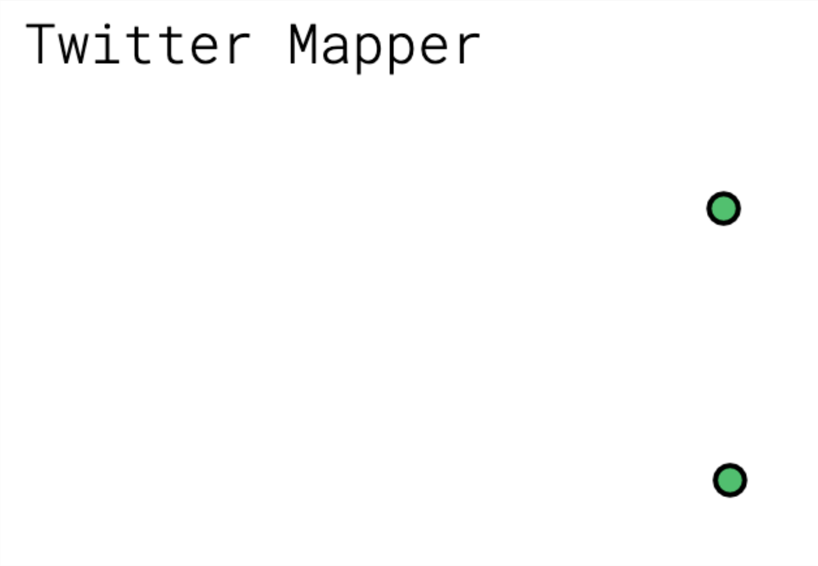
Choice of f

- The Mapper algorithm requires the user to make a few choices:
  - Cover $\mathcal{U}$
  - Clustering algorithm
- Choosing the right function f can have a significant impact on the resulting graph.
- Common choices include:
  - Density function
  - PCA (or manifold learning) coordinates
  - Distance to a fixed point

Data analysis

- Use the full document-term matrix
- *Can we recover leaders/topics from Mapper graph?*
- Choice of f: t-SNE coordinates
- For cover and clustering: use default from Kepler Mapper library.

Data analysis



Discussion

- Unfortunately, not much to see here...
- However, Mapper has been very successful in the literature.
- Used for topic detection (Torres-Tramòn *et al*, 2015), Bitcoin ransomware prediction (Akcora *et al*, 2019)

Room for improvement

- Sparsity is a feature of text data, but it may be possible to mitigate it via lemmatization.
- Use bootstrap on persistence diagrams.
- Tune dimension reduction steps and function in Mapper

Future inquiries

- Impact of dimension reduction on TDA calculations
- How far can we push algorithms?
- Can we reconstruct the manifold? Where are the holes?

Questions or comments?

For more information and updates, visit `maxturgeon.ca`**.**