Mapping Firms' Locations in Technological Space: A Topological Analysis of Patent Statistics

Emerson G. Escolar   Yasuaki Hiraoka   Mitsuru Igami   Yasin Ozcan

Riken AIP, Kyoto University, Yale University, MIT Sloan

December 7, 2019

Question & Approach

- Basic descriptive question
  - Where do firms innovate?
  - Where are they “located” in technological space?
- To answer this question, we use:
  - Patent statistics
  - Mapper algorithm: a new method from topological data analysis (TDA)
- Can handle any distance metrics
  - Looking forward to working with *text-based* metrics, too

The Problem

- Data from USPTO on top 333 firms (by count) in 1976–2005
  - Firm $i = 1, 2, ..., 333$
  - Year $t = 1976, 1977, ..., 2005$
- Each firm i (in each year t) patents across 430 USPC technological categories
  - Class $c = 1, 2, ..., 430$
- Patenting activity of firm-year (i, t) is a 430-vector

$$p_{i,t} = (p_{i,t,1}, p_{i,t,2}, ..., p_{i,t,430})$$

Challenge: How do we map all $p_{i,t}$'s in technological space?

Existing Methods

1. Normalization

$$x_{i,t} = f(p_{i,t})$$

   - E.g., taking log, converting to % shares, moving window, ...

2. Distance metric

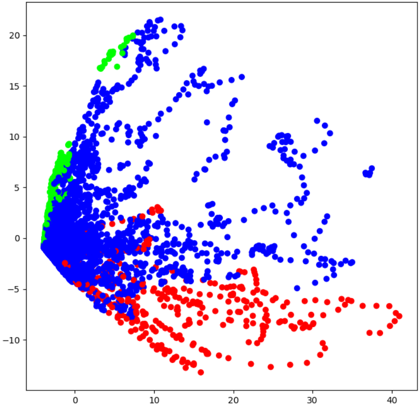
$$\delta(x_{i,t}, x_{i',t'})$$

   - E.g., Euclidean, Mahalanobis, correlation, cosine, min-complement, ...

3. Dimensionality reduction
   - E.g., PCA (principal-component analysis), MDS (multi-dimensional scaling), k-means clustering, ...

Note: The final step is needed because of high dimensionality (430) of data.

Example: Log(.) + Corr(.) + PCA



- Interesting 2-dimensional plot, but what about the other 428 dimensions?

Our Proposal: Computational Topology

1. Normalization [same as before]

$$x_{i,t} = f(p_{i,t})$$

   - E.g., taking log, converting to % shares, moving window, ...

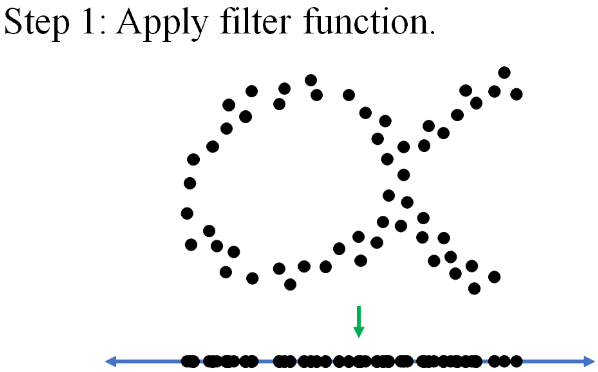
2. Distance metric [same as before]

$$\delta(x_{i,t}, x_{i',t'})$$

   - E.g., Euclidean, Mahalanobis, correlation, cosine, min-complement, ...

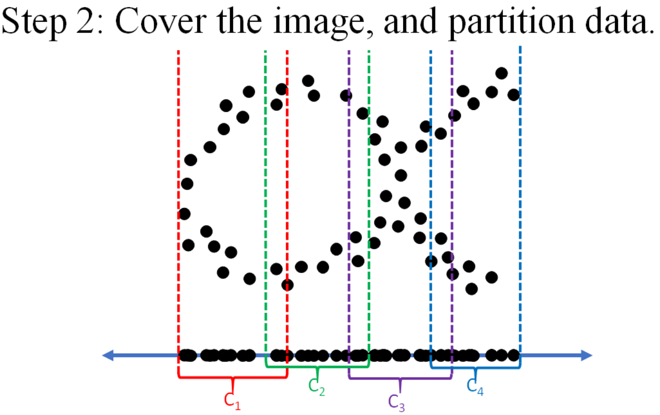
3. Dimensionality reduction [NEW!]
   - Not "just PCA/MDS"
   - Not "just clustering"
   - But combine them in a clever way

Mapper Algorithm: A Two-dimensional Example



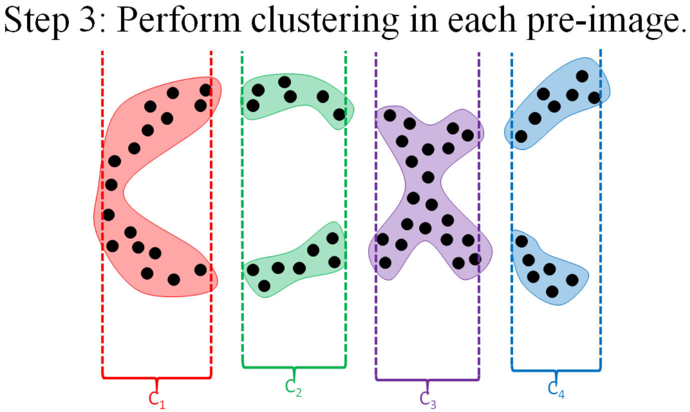
- Project 2-dimensional X onto the horizontal axis (i.e., $\mathbb{R}^1$).

Mapper Algorithm: A Two-dimensional Example

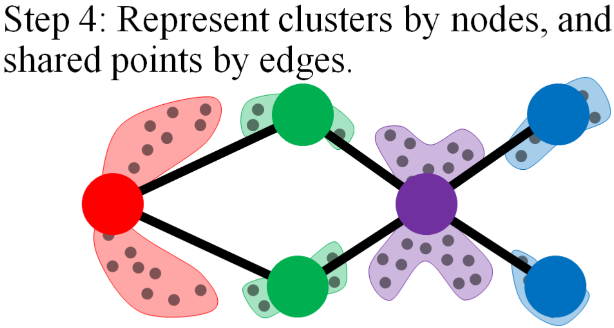


- Cover the image $f(X)$ (the points on the horizontal axis) by equal-sized intervals C_1, C_2, C_3, & C_4 with overlaps.

Mapper Algorithm: A Two-dimensional Example



- For each interval C_j, apply clustering algorithm to its pre-image $f^{-1}(C_j)$. That is, adjacent points in the original 2-dimensional space are bundled.

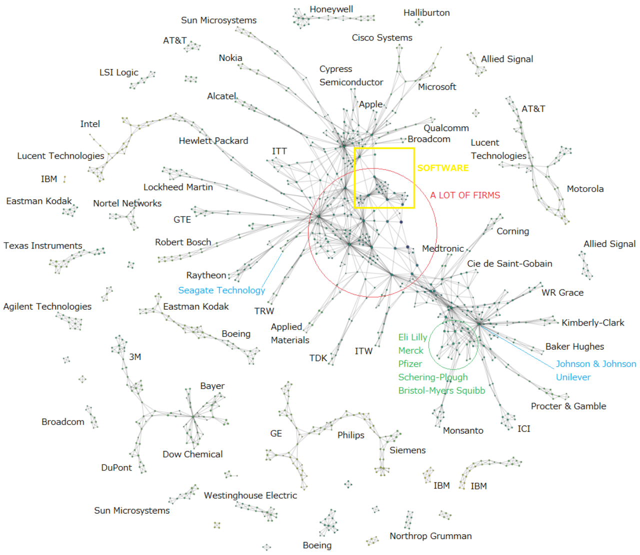


- Represent clusters by nodes (vertices) $V_{j,k}$s.
- If clusters share the same points, connect them with an edge.

Mapper Algorithm: Background

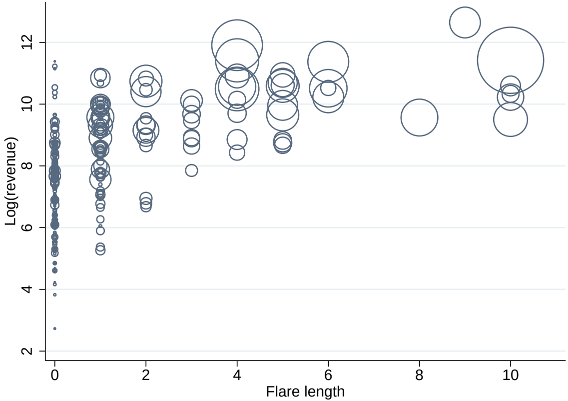
- Singh, Mémoli, & Carlsson ('07) proposed it.
  - Yao et al. ('09): an RNA folding pathway
  - Nicolau, Levine, & Carlsson ('11): the DNA microarray data of breast cancer
  - Rizvi et al. ('17): cellular differentiation & development
- Lum et al. ('13):
  - (i) gene expression of breast tumors
  - (ii) voting in the US Congress
  - (iii) NBA players' performances
  - They also propose a (generic, graph-theoretic) algorithm to detect "flares".
- By contrast, our "flare" definition & detection algorithm exploit the Mapper graph's particularities.

The Mapper Graph of 333 Firms in 1976–2005



[Interactive version]

Do Flares Matter?



- Each circle represents a firm.
  - Horizontal axis (X): Flare length of its patenting history in 1976–2005
  - Vertical axis (Y): Its financial performance (revenue) as of 2005
  - Circle size (Z): Its total patent count 1976–2005

Do Flares Matter?

Table: Revenues and Flare Length ($n = 20$)

LHS variable:	Log(Revenue in 2005)			
Patents acquired by:	R&D only		R&D and M&A	
	(1)	(2)	(3)	(4)
Flare length = 1	1.225	0.538	1.440	0.403
	(0.182)	(0.221)	(0.190)	(0.231)
Flare length = 2	2.145	1.210	2.141	0.703
	(0.382)	(0.409)	(0.354)	(0.387)
Flare length = 3	2.287	1.167	2.242	0.790
	(0.476)	(0.507)	(0.462)	(0.476)
Flare length = 4	2.082	0.897	3.225	1.246
	(0.487)	(0.522)	(0.438)	(0.495)
Flare length = 5	2.960	1.674	2.586	0.716
	(0.614)	(0.640)	(0.441)	(0.489)
Flare length = 6	3.632	2.258	3.757	1.746
	(0.523)	(0.570)	(0.637)	(0.656)
Flare length = ∞	3.780	2.258	3.511	1.523
	(0.595)	(0.646)	(0.571)	(0.600)
Log(Patents)	–	0.252	–	0.446
	(–)	(0.050)	(–)	(0.065)
Adjusted R^2	0.440	0.487	0.478	0.555
Number of observations	286	286	288	288

Note: Standard errors are in parentheses. S&P economic sector dummies are included. Estimates for flare lengths 8, 9, & 10 are suppressed due to space constraint.

Conclusion

1. We can summarize firms' technological locations in a graph.
2. It preserves global data patterns in a high-dimensional space.
3. The method works with:
   - ANY distance metrics
   - ANY ways to codify patents
   - ANY clustering & PCA/MDS methods

   It complements all existing (& new) measures/methods!

- Extensions
  - From US data to world data (*Patstat* by EPO)
  - From top-333 to top-1000+
  - From 1976–2005 to 1966–2015
  - From 430 USPC classes to 630 IPC subclasses
  - Including product-market competition?
  - Including non-practicing entities (NPEs)?