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
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 $i = 1, 2, \ldots, 333$
  - Year $t = 1976, 1977, \ldots, 2005$
- Each firm $i$ (in each year $t$) patents across **430 USPC technological categories**
  - Class $c = 1, 2, \ldots, 430$
- Patenting activity of firm-year $(i, t)$ is a **430-vector**
  $$p_{i,t} = (p_{i,t,1}, p_{i,t,2}, \ldots, p_{i,t,430})$$

**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
Step 1: Apply filter function.

- Project 2-dimensional $X$ onto the horizontal axis (i.e., $\mathbb{R}^1$).
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.
Step 3: Perform clustering in each pre-image.

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.
Step 4: Represent clusters by nodes, and shared points by edges.

- Represent clusters by nodes (vertices) $V_{j,k}$.
- If clusters share the same points, connect them with an edge.
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.
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
## Table: Revenues and Flare Length \((n = 20)\)

<table>
<thead>
<tr>
<th>LHS variable:</th>
<th>Log(Revenue in 2005)</th>
<th>R&amp;D only</th>
<th>R&amp;D and M&amp;A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents acquired by:</td>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Flare length = 1</td>
<td></td>
<td>1.225</td>
<td>0.538</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.182)</td>
<td>(0.221)</td>
</tr>
<tr>
<td>Flare length = 2</td>
<td></td>
<td>2.145</td>
<td>1.210</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.382)</td>
<td>(0.409)</td>
</tr>
<tr>
<td>Flare length = 3</td>
<td></td>
<td>2.287</td>
<td>1.167</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.476)</td>
<td>(0.507)</td>
</tr>
<tr>
<td>Flare length = 4</td>
<td></td>
<td>2.082</td>
<td>0.897</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.487)</td>
<td>(0.522)</td>
</tr>
<tr>
<td>Flare length = 5</td>
<td></td>
<td>2.960</td>
<td>1.674</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.614)</td>
<td>(0.640)</td>
</tr>
<tr>
<td>Flare length = 6</td>
<td></td>
<td>3.632</td>
<td>2.258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.523)</td>
<td>(0.570)</td>
</tr>
<tr>
<td>Flare length = (\infty)</td>
<td></td>
<td>3.780</td>
<td>2.258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.595)</td>
<td>(0.646)</td>
</tr>
<tr>
<td>Log(Patents)</td>
<td></td>
<td>–</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(—)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>Adjusted (R^2)</td>
<td></td>
<td>0.440</td>
<td>0.487</td>
</tr>
<tr>
<td>Number of observations</td>
<td></td>
<td>286</td>
<td>286</td>
</tr>
</tbody>
</table>

*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.
We can summarize firms’ technological locations in a graph.

It preserves global data patterns in a high-dimensional space.

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 \((\text{Patstat by EPO})\)
- From top-333 to top-1000+
- From 430 USPC classes to 630 IPC subclasses
- Including product-market competition?
- Including non-practicing entities (NPEs)?