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

Emerson G. Escolar† Yasuaki Hiraoka‡ Mitsuru Igami§ Yasin Ozcan¶

August 31, 2019

**Abstract**

Where do firms innovate? Mapping their locations in technological space is difficult, because it is high dimensional and unstructured. We address this issue by using a method in computational topology called the Mapper algorithm, which combines local clustering with global reconstruction. We apply this method to a panel of 333 major firms' patent portfolios in 1976–2005 across 430 technological areas. Results suggest the Mapper graph captures salient patterns in firms' patenting histories, and our measures of their uniqueness (the type and length of "flares") are correlated with firms' financial performances in a statistically and economically significant manner.

*Keywords*: Innovation, Merger, Patent statistics, R&D, Topological data analysis.

*Journal of Economic Literature (JEL) classifications*: C65, C88, L10, O30.

*This draft is preliminary; comments are welcome. We thank Alan Chiang for research assistance, and the National Institute of Science and Technology Policy (NISTEP) of Japan for providing an occasion for Hiraoka and Igami to meet in the summer of 2017.

†RIKEN Center for Advanced Intelligence Project and Kyoto University Institute for Advanced Study. E-mail: emerson.escolar@riken.jp.

‡WPI-ASHBi, Kyoto University Institute for Advanced Study, Kyoto University and Center for Advanced Intelligence Project, RIKEN. E-mail: hiraoka.yasuaki.6z@kyoto-u.ac.jp.

§Yale Department of Economics. E-mail: mitsuru.igami@yale.edu.

¶MIT Sloan School of Management and FTI Consulting. Email: ozcan@alum.mit.edu.

# 1 Introduction

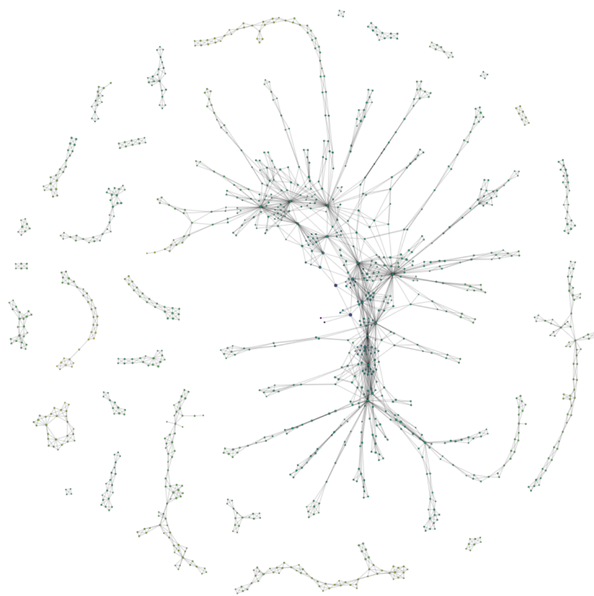
Where do firms innovate? Existing studies have answered this question by using geographical information in patent data, such as the locations of firms and inventors, and their agglomeration economies.[1] The same question can be asked in terms of technological space: in which technological areas do firms conduct research and development (R&D)? Mapping firms' technological positions is difficult, however, because technological space is high dimensional and unstructured. The US Patent and Trademark Office (USPTO) categorizes patented inventions into more than 400 classes and hundreds of thousands of subclasses. Existing methods, such as principal component analysis (PCA) and clustering algorithms, help reduce dimensionality, but they may not preserve the shape of the original data. This paper proposes a new approach to represent firms' technological locations in a graph by applying and extending a method in computational topology. It allows us to answer some of the most basic questions about innovation, including how their technological locations evolve over time, how they interact with each other, and whether these positions are related to their performances.

Specifically, we study a panel of 333 major firms' patenting histories across 430 USPTO patent classes between 1976 and 2005. Firms such as IBM and GE obtain thousands of patents in more than 100 classes every year; hence, manually inspecting the raw data is all but impossible, and analyzing firms' interactions in technological space would seem hopeless. We use a method in topological data analysis (TDA) called the Mapper algorithm, which reduces high-dimensional and complicated data into graphs (or simplicial complexes, more generally). This reduction is performed by a combination of local clustering and global reconstruction in such a way to capture and preserve the shape of data (i.e., topological and geometric information).

Figure 1 is an example of the Mapper graph, based on the 333 firms' patenting behaviors in 1976–2005. Its central part features a large connected component with many nodes that are densely linked in the shape of a spine or a "trunk." This trunk contains the majority of the firms in our sample, whose patenting patterns are similar to each other. However, many "flares" branch out of the trunk. Each of these flares typically consists of only one or two firms whose patent portfolio is similar to others in the 1970s and the 1980s, but becomes increasingly differentiated in the 1990s and the 2000s. Many "islands" surround the main trunk and flares. Some of these small connected components contain multiple firms, but most of them represent individual firms. Firms whose portfolios are undifferentiated from

[1] See Jaffe, Trajtenberg, and Henderson (1993) and the large literature on the geography of innovation.

Figure 1: Trunks, Flares, and Islands



*Note*: This graph is an output of the Mapper algorithm (based on correlation distance and resolution level $n = 20$) on 333 major firms' portfolios of US patents that are acquired by in-house R&D between 1976 and 2005. See sections 3 and 4 for further details.

others appear in the densely populated trunk part of the graph, whereas firms with unique patenting patterns are represented by flares and islands. We supplement these informal characterizations of the graph by proposing formal definitions of flares and their "lengths." We also propose a method to detect and measure these topological features in the graph, and use them as measures of "uniqueness" of each firm's patenting history.

Our examination of the Mapper graph suggests these topological characteristics reflect important aspects of firms' innovation strategies, including the size and diversity of patent portfolios, as well as their focus, consistency, and evolution over time. We then statistically assess whether our measures of uniqueness contain any relevant information. Following a common practice in the economic literature on patent statistics, such as Pakes and Griliches (1984), we look for correlations between our topological measures (flare types and lengths) and the firms' performance metrics, including revenues, profits, and stock market values. Results suggest flares are almost always associated with high revenues, profits, and values. In particular, longer flares are frequently associated with more than 100% higher performances even after controlling for the total count of patents (i.e., the size of such activities).

In summary, the Mapper provides a useful and intuitive map of firms' locations in technological space. Some firms' innovation histories are more unique and differentiated than others, and these firms tend to exhibit superior financial performances both statistically and economically significantly. The types and the length of flares convey additional information beyond what can be predicted based on patent count alone.

We organize the paper as follows. The rest of section 1 explains related studies. Section 2 explains our data, reports descriptive statistics, and shows examples of firms' patenting histories in raw data. Section 3 explains our method. Section 3.1 provides an overview of the Mapper procedure. Section 3.2 provides the details of our application. Section 3.3 presents our method to detect flares. Section 3.4 presents our method to measure the length of flares. Section 4 reports main results, including the Mapper graph (section 4.1) and the type and length of flares (section 4.2). Section 5 investigates the correlations between these topological measures and firms' financial performances, in the full sample (section 5.1), by economic sector (section 5.2), and at narrowly defined industry level (section 5.3). Section 6 concludes.

## Related Literature

Patent statistics have been used as an indicator of innovation in many economic studies. For an overview and surveys, see Griliches (1990), Cohen (2010), Nagaoka, Motohashi, and Goto (2010), and Lerner and Seru (2017). The most closely related work to ours is Jaffe (1989), who pioneered the use of patent-class data to characterize firms' technological positions and their distances (angular separation) from each other, as well as the use of k-means clustering.

More recently, Benner and Waldfogel (2008) scrutinize the USPTO's classification procedures, investigate statistical biases in the analysis of patent-class data, and offer practical suggestions. Bar and Leiponen (2012) propose a new measure of technological distance, called the min-complement distance, which satisfies a desirable property (independence of irrelevant patent classes) that no other conventional measures satisfy. Bloom, Schankerman, and Van Reenen (2013) propose the use of the Mahalanobis distance in their study of R&D spillovers. We contribute to this large literature by proposing a new method and measures to characterize firms' locations in technological space in an intuitive and tractable manner. Our method can be used with any of these distance metrics; hence, it directly complements all of these methodological proposals.

The Mapper algorithm was introduced by Singh, Mémoli, and Carlsson (2007), and has been applied to a range of scientific fields. Yao et al. (2009) used it to explore an RNA folding pathway. Nicolau, Levine, and Carlsson (2011) applied it to the DNA microarray data of breast cancer, and identified a new subclass of the disease. Lum et al. (2013) studied gene expression data of breast tumors, voting data from the United States House of Representatives, and performance data of players in the National Basketball Association. Rizvi et al. (2017) analyzed cellular differentiation and development.

Methodologically, Lum et al. (2013) is the most closely related work to ours, because they also propose a flare detection algorithm. Their method uses global graph-theoretic properties that are applicable to any graph, without using any additional information from the Mapper algorithm.[2] By contrast, our algorithm takes advantage of particularities of our Mapper graph, where each node is a set of firm-years. We enforce each flare that we identify to be associated with a specific firm. Hence, it can be interpreted as *a flare of that firm.*

# 2 Data

## 2.1 Patents

We use Ozcan's (2015) data on patents that are granted by the USPTO between 1976 and 2010.[3] We use their application years (instead of years in which they are granted) in our analysis, because the former is closer to the time of actual invention than the latter. We focus on patents that are applied through 2005, because a substantial fraction of later applications would still be under review as of 2010, which raises concerns about sample selection. We call these patents "R&D patents" to distinguish them from "M&A patents."

## 2.2 Mergers and Acquisitions (M&As)

Aside from conducting in-house R&D and applying for patent protection, firms often obtain patents by acquiring firms that have their own portfolios of patents. We incorporate such M&A-related acquisitions of patents into our analysis as well, by using Ozcan's (2015) dataset, which links the USPTO data to the Securities Data Company's M&A data module.

[2]Specifically, their flare detection algorithm uses the 0-dimensional persistent homology (Edelsbrunner, Letscher, and Zomorodian, 2000) of the graph filtered by an eccentricity measure on its nodes. An eccentricity measure tends to give a higher value to nodes that are "eccentric" (on tips of flares) compared to central nodes (on the trunks).

[3]Ozcan (2015) uses the USPTO's Patent Data Files, which contain raw assignee names at the individual patent level. By contrast, the NBER Patent Data File (another commonly used source of patent data) records standardized assignee names at the "pdpass" (unique firm identifier) level, which is less granular than the original assignee name.

This part of the dataset contains M&A deals between 1979 and 2010 in which both the acquiring firm and the target firm have at least one patent between 1976 and 2010.[4]

## 2.3 Firms' Performance Metrics

This paper's main focus is on proposing a topological method to map firms' locations in technological space and to characterize firms' patent portfolios. Nevertheless, we also study their eventual financial performances as a means to assess the immediate relevance of our proposed measures. We collect Compustat data on the firms' revenues, EBIT (earnings before interest and taxes), and stock-market capitalization between 1976 and 2005.

## 2.4 Descriptive Statistics of the Top 333 Firms

We focus on firms that acquired at least four firms with patents between 1976 and 2005. This criterion keeps 333 "major" firms, whose descriptive statistics are shown in Table 1.

Table 1: Summary Statistics

| Variables | Number of observations | Mean | Median | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Patents: In-house R&D | 333 | 1,872 | 172 | 5,264 | 0 | 62,382 |
| Patents: Acquired by M&A | 333 | 259 | 54 | 884 | 0 | 9,453 |
| Patents: Both R&D and M&A | 333 | 2,131 | 302 | 5,542 | 0 | 62,561 |
| Revenue (million US$) | 315 | 9,888 | 2,371 | 24,160 | 15 | 309,979 |
| EBIT (million US$) | 315 | 1,336 | 252 | 3,454 | −450 | 37,159 |
| Market capitalization (million US$) | 313 | 18,282 | 3,493 | 39,989 | 2 | 367,474 |

*Note*: Financial-performance metrics are as of 2005 or the firm's last available fiscal year.

The average patent count is 1,872 for in-house R&D and 259 for M&A, respectively. The medians (172 and 54) are much lower than the means, which suggests their distributions are highly skewed. Relatively few firms have a lot of patents even within our selective sample. The three financial-performance metrics exhibit similar skewness. The average revenue, EBIT, and market capitalization in 2005 are $9.9 billion, $1.3 billion, and $18.3 billion, respectively, whereas their medians are $2.4 billion, $252 million, and $3.5 billion. This aspect of the data leads us to use the natural logarithm of patent count and financial metrics in our subsequent analysis to mitigate potential econometric issues related to heteroskedasticity.

[4] The data include merger, acquisition, acquisition of majority interest, acquisition of assets, and acquisition of certain assets, but exclude incomplete deals, rumors, and repurchases. We use data on these transactions through 2005.

## 2.5 Where Do Firms Patent?

Let us illustrate with examples what these firms' patent portfolios look like. Figure 2 visualizes the evolution of patenting activities at six major firms. Each plot lists the 430 USPTO patent classes on the vertical axis, and the year of application (for R&D patents) or acquisition (for M&A patents) on the horizontal axis. The circle size represents the number of patents in each class-year.

The top panels show two IT firms. Cisco Systems makes network equipment (e.g., routers) and is famous for its active use of M&As to acquire new products and talents; it acquired the largest number of target firms with patents in our sample. Nevertheless, most of Cisco's patents are obtained by in-house R&D and are concentrated in classes 370 (multiplex communications) and 709 (electrical computers and digital processing systems: multicomputer data transferring). Seagate Technology makes hard disk drives (HDDs) and is another example of specialized IT firms. Its main patent class is 360 (dynamic magnetic information storage or retrieval), which is central to the HDD technology, but its portfolio gradually diversified as the firm intensified efforts to manufacture key components as well, including heads, media, and their interface.[5]

The middle panels show two health care firms. The pharmaceutical industry is R&D-intensive, but the patent portfolio of Pfizer looks simpler than the IT examples. Most of the drug patents are in classes 424 and 514 (drug, bio-affecting, and body treating compositions), and drug makers hardly patent elsewhere. By contrast, medical devices rely on a variety of technologies, even though their main classes are relatively few (600–607). The plot shows Medtronic, a leading medical-device maker, is active in many areas.

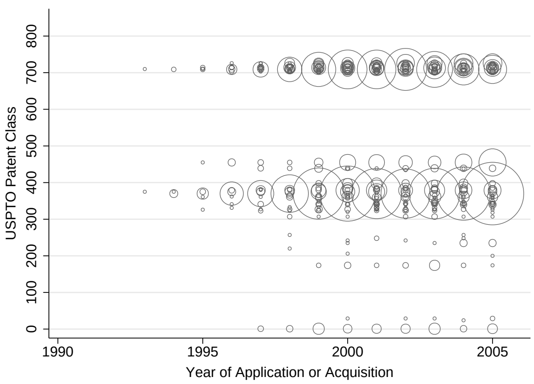
The bottom panels present extreme cases, for a reference. GE, a conglomerate, has the most diversified portfolio in our sample, with patents in more than 300 classes. The picture becomes too messy for human eyes to draw insights. Finally, IBM has by far the largest number of patents in our sample, but its portfolio looks more organized than GE's, because its activities are more focused. Most of the computers and electronics technologies are in the 300s and the early 700s, which are where IBM's portfolio is concentrated.

These examples suggest the portfolio aspect of patents and technologies is interesting and contains potentially important information. However, the high dimensionality of technological space makes data analysis difficult. "Where do firms patent?" is a basic question, but answering it turns out to be challenging. In the next section, we explain our topological method to address this problem.

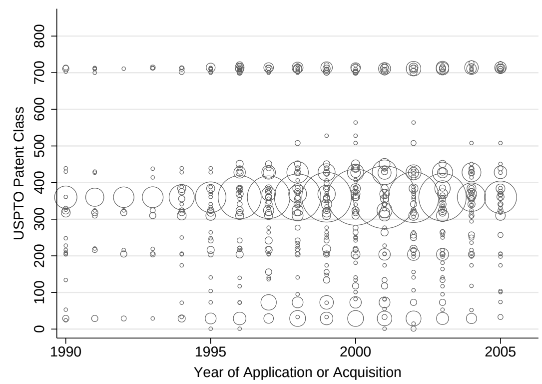
[5] See Igami and Subrahmanyam (2019) for the details of patents and innovation in the HDD industry.

Figure 2: Acquiring a String of Pearls

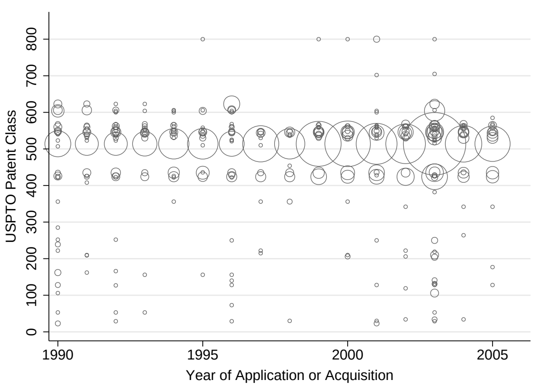
(a) Cisco Systems



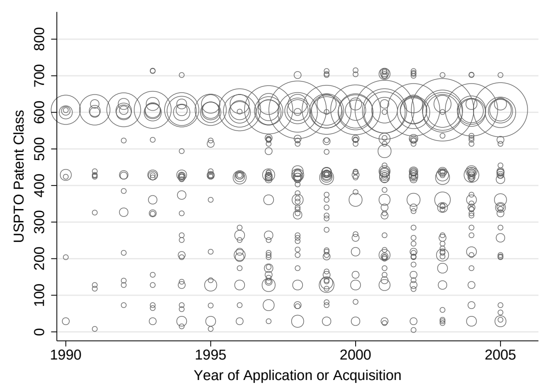

(b) Seagate Technology



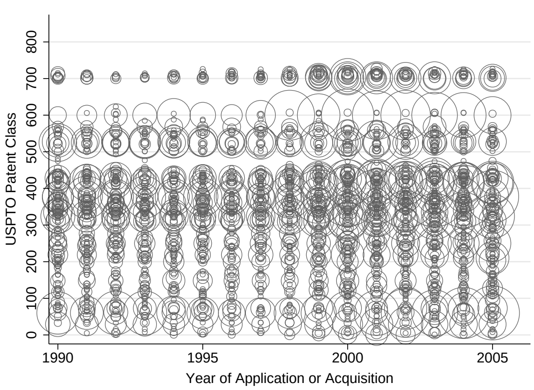

(c) Pfizer



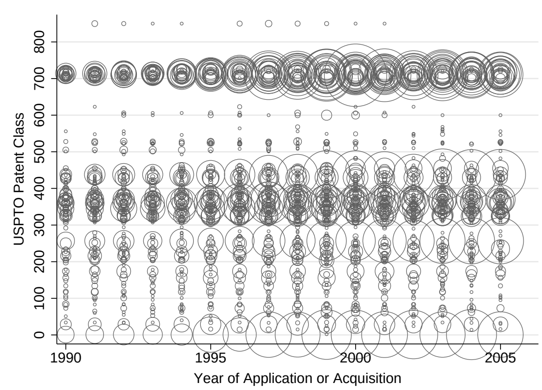

(d) Medtronic




(e) GE




(f) IBM




*Note*: The circle size represents the number of patents in each class-year. Based on our method and analysis in sections 3 and 4, the "flare lengths" (our proposed measure of "uniqueness") of these firms' portfolios are: 3 (Cisco), 2 (Seagate), 1 (Pfizer), 2 (Medtronic), 4 (GE), and $\infty$ (IBM).

# 3 Method

## 3.1 The Mapper Algorithm

We provide a quick review of the Mapper method introduced by Singh, Mémoli, and Carlsson (2007). Given some complicated and high-dimensional data, Mapper provides a simplified representation of the data via a graph that captures some of its important "topological features" such as branching, flares, and islands.

We assume the data are given as a set of points $X$ together with a dissimilarity function $\delta : X \times X \to \mathbb{R}_{\geq 0}$. The Mapper graph is constructed in four steps.

1. Project $X$ into $\mathbb{R}^d$ by some filter function $f : X \to \mathbb{R}^d$.
2. Cover the image $f(X)$ using an overlapping cover $\mathcal{C} = \{C_j\}_{j=1}^{J}$.
3. For each cover element $C_j$, apply some clustering algorithm to its pre-image $f^{-1}(C_j)$ based on the dissimilarity function $\delta$ to obtain a partition of $f^{-1}(C_j)$ into $K_j$ clusters, $V_{j,k}$ $(k = 1, \dots, K_j)$:

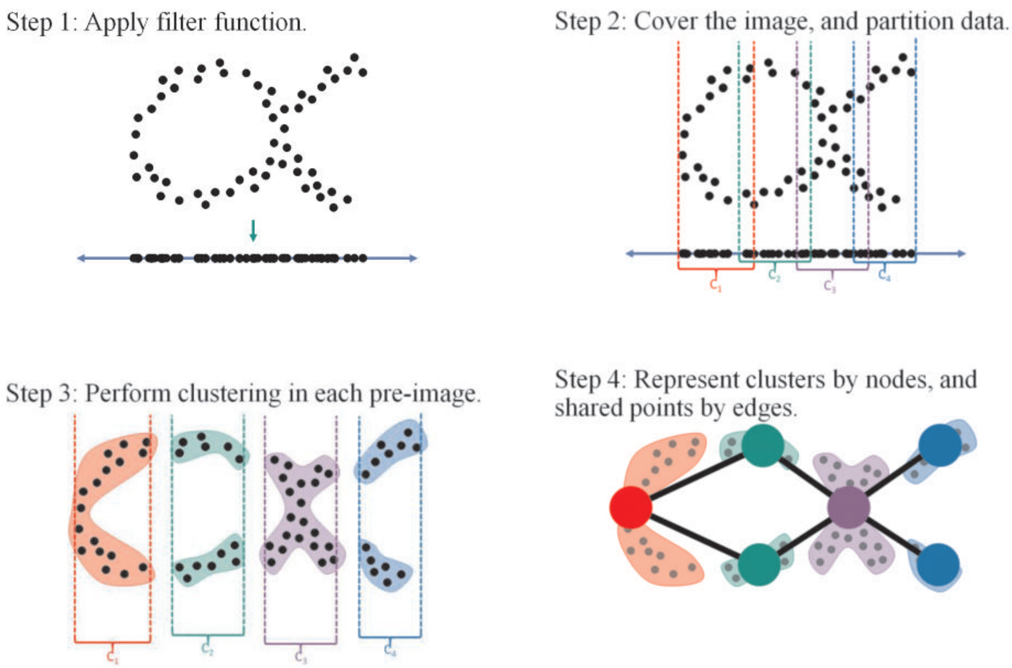
$$f^{-1}(C_j) = \bigsqcup_{k=1}^{K_j} V_{j,k}. \tag{1}$$

4. Construct the graph $G$ with nodes (vertices) consisting of all $V_{j,k}$s. Connect two nodes, $V_{j,k}$ and $V_{j',k'}$, by an edge if $V_{j,k} \cap V_{j',k'} \neq \emptyset$.

Figure 3 illustrates this procedure with an example. Let us start with data $X$ given by the points in two-dimensional space. Our goal is to obtain a simplified representation of $X$ while preserving its topological features, such as holes and branches. In step 1, we project $X$ onto the horizontal axis (i.e., $d = 1$). This operation reduces the dimensionality of the data by eliminating the second dimension (i.e., information on the vertical axis in this case). In step 2, we cover these points on the horizontal axis by four equal-sized intervals (i.e., cover elements) $C_1, C_2, C_3$, and $C_4$ (i.e., $J = 4$) with overlaps.[6] In step 3, we look at each interval $C_j$, and cluster adjacent points *in the original data space* with two dimensions (*not* on the horizontal axis). In step 4, we represent these clusters by nodes, and connect them with edges whenever adjacent clusters share the same points within their overlapping regions.

[6] The degree of overlap is approximately 20% in the pictured example, but the analyst can alter it.

Figure 3: Illustration of the Mapper Procedure




The resulting graph is much simpler than the original data and amenable to graph-theoretic analyses, but it still preserves the global structure of $X$. By contrast, using conventional techniques for dimensionality reduction alone would be similar to performing only step 1. Likewise, directly performing clustering in the original data would be the same as skipping steps 1 and 2, which would probably generate a single big cluster for the entire data. Neither approach would be able to recover the *shape* of the data. For this particular example, the usefulness of the Mapper graph is limited, as the original data itself is already two-dimensional and can be drawn directly. However, for more complicated high-dimensional data, a simplified graph representation helps us understand the data

One way to interpret the Mapper procedure is to view it as a kind of local clustering together with global reconstruction. The choice of the filter function and cover determines the local regions $f^{-1}(U_j) \subset X$ of the data. Then, the clustering algorithm is applied only locally, to each local region. The construction of the graph $G$ recovers some of the global information by connecting nodes (each of which is a cluster of points in $X$) whenever they share points in the original data.

## 3.2 Application of Mapper to Our Data

Recall that our data are a panel of 333 firms' yearly patent applications (and/or acquisitions) across 430 technology classes between the years 1976 and 2005. For each firm $i$, each year $t$, and each patent class $c$, we have patent count $p_{i,t,c}$. We regard each firm-year as a single observation represented by a vector $p_{i,t} \in \mathbb{R}^{430}$. Hence, each firm-year is a point $p_{i,t}$ in the 430-dimensional technology space.

Firms' patent applications in any single year tend to be volatile and less representative of their underlying R&D activities. This issue is particularly important in the use of patent-class data, as Benner and Waldfogel (2008) point out. We follow their recommendation to smooth out yearly fluctuations by using a five-year moving window:

$$\tilde{p}_{i,t} = p_{i,t} + p_{i,t+1} + \ldots + p_{i,t+4}. \tag{2}$$

Another practical consideration is the highly skewed distribution of patent count, as section 2.4 made clear. We address this issue by applying a logarithmic transform,

$$\tilde{p}_{i,t} \mapsto \ln(\tilde{p}_{i,t} + 1) =: x_{i,t}. \tag{3}$$

Let $X = \{x_{i,t}\}$ be the point cloud consisting of the transformed data. We use the following specifications in constructing a mapper graph for $X$. We use the Python implementation, KeplerMapper, by Van Veen and Saul (2019).

1. The filter function is $f : X \to \mathbb{R}^2$, which projects $X$ to its first two principal axes as obtained by PCA.[7]

2. For the cover of the image of $f$, we use the default cover implementation in KeplerMapper. We set the resolution level (called the "number of cubes," $n$) to 20 in our baseline specification, and the degree of overlap to 50%.[8]

3. For the clustering algorithm, we use single-linkage clustering together with the heuristic as proposed by Singh, Mémoli, and Carlsson (2007) for choosing the number of clusters.

For the dissimilarity measure between points in $X$, we use the correlation distance in our baseline specification, because it is not directly affected by the size of the firm/portfolio.[9]

[7]Note we use PCA only for the purpose of determining local regions. The subsequent clustering is performed in each pre-image in the original space and *not* in the PCA space.

[8]We set $n$ to 15 and 25 in sensitivity analysis.

[9]We try to purge the influence of firm size here, because the most conventional measure of innovation,

## 3.3 Detection of Flares

As we have seen in section 3.1, Mapper provides a simplified representation of complicated data via a graph $G$ that captures some of its more important topological features. In this section, we discuss the detection of one such feature: "flares."

Recall some basic concepts from graph theory. In general, a *graph* $G = (V, E)$ is a set $V$ of nodes (vertices) and a set $E$ of edges. We assume that each edge $e \in E$ of $G$ is assigned the weight $w(e) = 1$.[10] For $u, v \in G$, the *length* $\ell(p)$ of a path $p$ from $u$ to $v$ is the sum of the weights of the edges of $p$. The *distance* $d_G(u, v)$ between $u$ and $v$ is the minimum length of all paths $p$ in $G$ from $u$ to $v$. For simplicity, we write $d(u, v)$ for $d_G(u, v)$.

For a graph $G$ and a subset of the nodes of $G$, $V'$, the *full subgraph* of $G$ with nodes $V'$, denoted by $G[V']$, is the graph with the set of nodes $V'$ and edges consisting of all edges of $G$ whose endpoints are both in $V'$. It is the maximal subgraph of $G$ with set of nodes $V'$.

**Definition 1** (Ball)**.** Let $r \in \mathbb{R}$ and $u \in G$. The (closed) ball $B_r(u)$ in $G$ is

$$B_r(u) = G[\{v \in G \mid d(u, v) \leq r\}].$$

In words, it is the full subgraph of $G$ of all nodes at most distance $r$ from $u$.

Now, consider a graph $G = (V, E)$ obtained from the Mapper algorithm applied to our data. From the construction of the Mapper graph, each node $v \in V$ will consist of points (firm-years) of the form $x_{i,t}$. To simplify, we adopt the following notation, because we want to consider firms and not firm-years for the analysis.

**Notation 2.** In the setting above, firm $i$ is said to be in node $v$, or, equivalently, $v$ contains firm $i$ if node $v$ contains an observation of firm $i$ at some time $t$, that is, $x_{i,t} \in v$ for some $t$. In this situation, we write $i \in v$.

For each firm $i$, we want to determine whether $i$ appears as a "flare" in $G$. One way to extract "flares" is to use global graph-theoretic properties of $G$, as in the method proposed in Lum et al. (2013) using 0-persistence of eccentricity (or centrality). Instead, we start with the assumption that we can only consider a structure to be a "flare" of $i$ if each node in the "flare" contains $i$.

total patent count (without class information), already reflects this aspect of R&D activities. See Jaffe (1989) for further discussions. We use the Euclidean distance in sensitivity analysis.

[10] The theory can be easily modified to handle positive weights $w(e) > 0$ possibly different for each edge $e$.

Thus, we can restrict our attention to a smaller graph $G_i$ defined below, which contains only nodes that involve $i$, and look for "flares" therein. We see later that this perspective simplifies computations.

**Definition 3** (Induced subgraph $G_i$ of firm $i$)**.** Let $i$ be a firm. Define $G_i$ to be

$$G_i = G[\{v \in G \mid i \in v\}].$$

That is, $G_i$ is the full subgraph of $G$ formed by nodes that contain firm $i$. We decompose the nodes of $G_i$ into "interior" and "boundary."

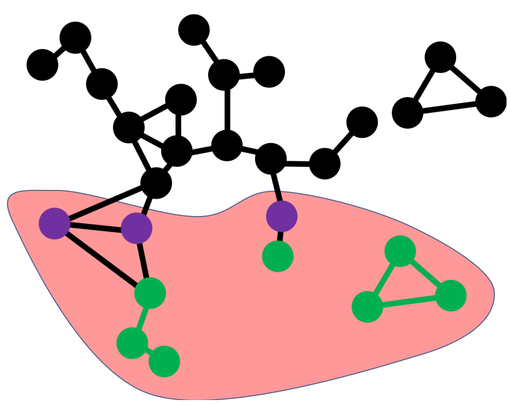
**Definition 4** (Interior and boundary of $G_i$)**.**

1. The *interior* $F_i$ of $i$ in $G$ is defined to be $F_i = G[\{v \in G_i \mid B_1(v) \subseteq G_i\}]$.
2. The *boundary* of $i$ in $G$ is $G_i \setminus F_i$.

In words, the interior $F_i$ contains all nodes $v$ of $G_i$ satisfying the property that $G_i$ contains the ball of radius 1 around $v$. Lemma 12 in the Appendix shows that the boundary $G_i \setminus F_i$ indeed serves as a "boundary" for $F_i$: to get outside of $G_i$, one always needs to go through the boundary.

Figure 4 illustrates the definitions of interior and boundary. The pink region represents firm $i$'s subgraph $G_i$, the green nodes are in the interior $F_i$, and the purple nodes are in the boundary $G_i \setminus F_i$.

Figure 4: Interior and Boundary



Let us define flares (and islands, by negation) in graph-theoretic terms.

**Definition 5** (Flares and Islands)**.** A connected component $L$ of the interior $F_i$ of firm $i$ is said to be an *island of firm* $i$ if $L$ is also a connected component of $G$, and said to be a *flare of firm* $i$, otherwise.

For example, two flares and one island (the triangle on the right) exist in Figure 4. In the next subsection, we refine these notions using numerical indices. As defined above, a flare may not always "look like" what one may imagine to be a "flare."

## 3.4 Measuring Flares

We introduce the following definition and theorem, which serve as the foundations for defining our concept of flare length.

**Definition 6** (Exit distance)**.** Let $u \in F_i$ be a node in the interior of firm $i$. The exit distance of $u$ in $F_i$ is

$$e_i(u) = \min\{d(u, w) \mid w \in G \setminus F_i\}.$$

In the case in which no path exists from $u$ to any $w \in G \setminus F_i$, we put $e_i(u) = \infty$.

**Theorem 7.** *Let $u \in F_i$. Then,*

$$e_i(u) = \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\},$$

*where $d_{G_i}(u, v)$ is the distance between $u$ and $v$ in $G_i$.*

See Appendix A for the proof. Using Theorem 7, we can compute $e_i(u)$ using only the information of $G_i$, because the distance $d_{G_i}(u, v)$ is the minimum weight of all <u>paths in $G_i$</u> from $u$ to $v$. By contrast, directly using Definition 6 would necessitate the computation of $d(u, w)$, the minimum weight of all <u>paths in $G$</u> from $u$ to $w$.

We use the exit distance $e_i(u)$ to refine our notion of flares.

**Definition 8** (Flare index)**.** For a connected component $L$ of $F_i$ (a flare or island of firm $i$), the *flare index* of $L$ is defined to be

$$k_i(L) = \max_{u \in L} e_i(u).$$

We immediately obtain the following characterization of islands using $k_i$.

**Lemma 9.** *Let $L$ be a connected component of $F_i$. Then, $k_i(L) = \infty$ if and only if $L$ is an island of firm $i$.*

*Proof.* Immediate from the definitions. □

Finally, to aggregate all the information, we define flare signature.

**Definition 10** (Flare signature)**.** Let $F_i = L_1 \sqcup L_2 \sqcup \ldots \sqcup L_M$ be a decomposition of $F_i$ into its connected components. The *flare signature* of $i$ is the multiset

$$\vec{k}_i = \{\{k_i(L_j) \mid j = 1, \ldots, M\}\}.$$

Note that if $F_i$ is empty, we simply put the empty multiset as the flare signature of $i$.

We link the flare signature to the following "types."

1. $\vec{k}_i$ **is empty.** This case occurs if and only if $F_i = \emptyset$, meaning every node containing firm $i$ neighbors at least one node not containing $i$. We call this case **Type 0: no flare or island**.

2. $\vec{k}_i$ **contains only finite elements**. In this case, each connected component $L$ of $F_i$ is connected to some point $w \in G \setminus G_i$, meaning each $L$ itself cannot be a connected component of $G$. Thus, each $L$ is not an island; it is a flare. We call this case **Type 1: flares only**.

3. $\vec{k}_i$ **contains finite elements, and some copies of** $\infty$. This case corresponds to **Type 2: flares and islands**.

4. $\vec{k}_i$ **contains only copies of** $\infty$. This case corresponds to **Type 3: islands only**.

The flare signature is defined as a multiset of flare indices. Sometimes, having one number describing how much firm $i$ looks like a flare in the Mapper graph may be convenient. Thus, we define the following.

**Definition 11** (Flare length)**.** The *flare length* (or just *length*, for short) of firm $i$ is

$$k_c = \begin{cases} 0 & \text{if } \vec{k}_c \text{ is empty,} \\ \operatorname{finmax}(\vec{k}_c) & \text{if } \vec{k}_c \text{ has at least one finite element,} \\ \infty & \text{otherwise,} \end{cases}$$

where $\operatorname{finmax}(\vec{k}_c)$ is the maximum among all finite elements of $\vec{k}_c$.

Type 0 gets flare length 0, type 3 is sent to index $\infty$, and types 1 and 2 occupy the range in between, where the flare length of a firm is determined by the "longest" flare of firm $i$.

**Computation of Flare Signatures** Let $G = (V, E)$ be the Mapper graph of our data $X$. For each firm $i$, the computation of the subgraph $G_i$ involving $i$ can be done by iterating through all nodes $v \in V$ and checking membership of firm $i$ in $v$. The interior-boundary decomposition of $G_i$ can be computed by considering the boundary first. For each $v \in G_i$, we simply check if $v$ has a neighbor that is not in $G_i$; if so, $v$ is part of the boundary $G_i \setminus F_i$. The nodes of $G_i$ not in the boundary are then automatically part of the interior.

Next, let us consider the computation of the flare signature $\vec{k}_i$ of firm $i$. First, we need a decomposition of $F_i$ into its connected components:

$$F_i = L_1 \sqcup L_2 \sqcup \ldots \sqcup L_M,$$

which can be done, for example, via a breadth-first search. For each connected component $L$ of $F_i$, its flare index is given by

$$k_i(L) = \max_{u \in L} e_i(u).$$

Because we need to do the same for each connected component $L$ of $F_i$, we compute $e_i(u)$ for all $u \in F_i$. By Theorem 7, the exit distance is

$$e_i(u) = \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\},$$

which can be computed using a multi-source version of Dijkstra's shortest-path algorithm, with sources $G_i \setminus F_i$.

# 4 Results

## 4.1 Trunks, Flares, and Islands

We use the Mapper algorithm (sections 3.1 and 3.2) to generate a graph representing the position of each firm $i$ in each year $t$, which is a vector of patent counts across 430 classes, $p_{i,t} = (p_{i,t,c})_{c=1}^{430}$.[11] Figure 1 in section 1 uses only in-house R&D patents, whereas Figure 5 incorporates patents acquired by both R&D and M&A. The two graphs are broadly similar, because relatively few patents are acquired through mergers.

In section 4.2, we detect and measure flares by using our method (sections 3.3 and 3.4).

[11] More precisely, we use its logarithmic transform with a five-year time interval starting from year $t$ (instead of single year $t$), $x_{i,t}$, in our computation. See section 3.2 for details.

Before proceeding to formal analysis, let us understand this graph better by investigating the identities of the firms and industries that show up saliently.

#### 4.1.1 IT

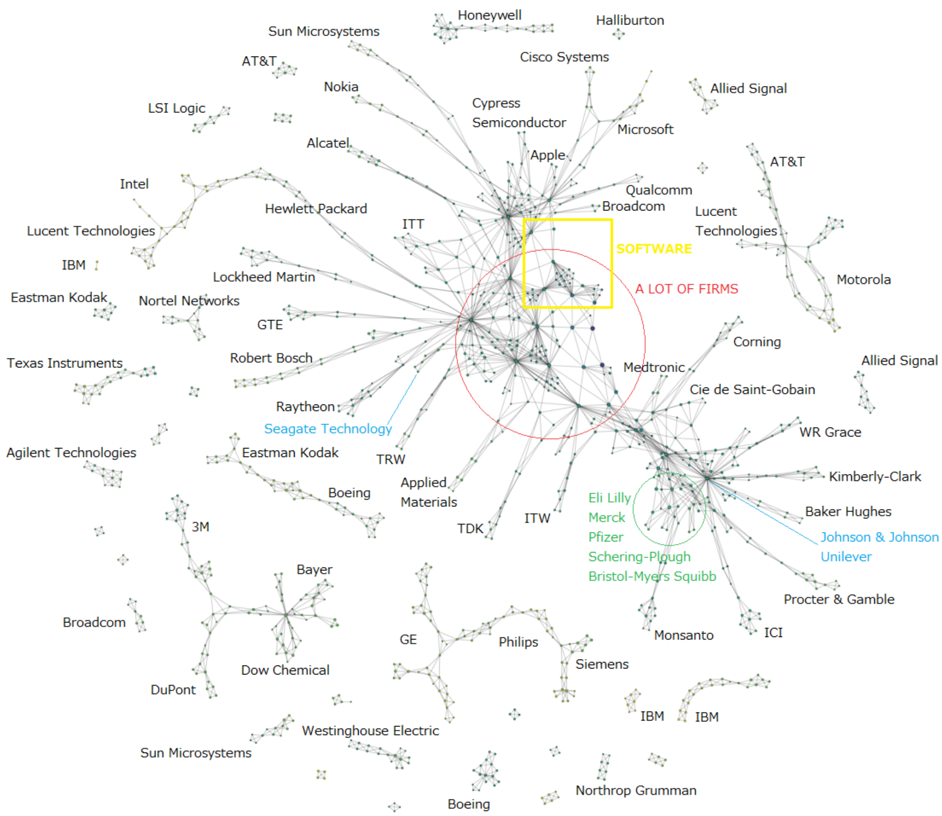
The main trunks are difficult to investigate visually, because each of the large nodes in the central part (circled in red) contains dozens of firm-years. Most of them operate in either IT or capital goods industries. By contrast, famous high-tech firms spike out in the form of flares (see long branches in the upper half of the graph). Some of the longest flares include Sun Microsystems, Nokia, Alcatel, Hewlett Packard, Intel, Lucent Technologies, Lockheed Martin, Robert Bosch, and Raytheon. Their patenting behaviors typically start from somewhere in the densely populated trunks in the 1970s, but start diverging from the majority in the 1980s, and evolve into something unique in the 1990s and the 2000s. This pattern agrees with the underlying technological trend in which computers and electronics emerged as an important sector with many opportunities to pursue innovations in different directions.[12]

Slightly less conspicuous flares include Cisco Systems, Microsoft, and Apple Computer (see the northeastern end of the main trunk), which might seem puzzling, because they are usually considered technological leaders in the internet age, especially at the turn of century. The reason Cisco and Microsoft appear in relatively short flares is that both of them heavily patented in classes 370 (multiplex communications) and 709 (electrical computers and digital processing systems: multicomputer data transferring) in the late 1990s, which connect their flares in the middle. Apple is another interesting example. Many of its products are perceived as unique and innovative, usually because of their aesthetics and functionality, but these product characteristics do not necessarily embody new patented inventions. Hence, Apple's patent portfolio may not look as unique as its products.

Other IT firms are relatively new during the sample period and do not have time to develop patent portfolios that are sufficiently unique to generate flares. The area between Apple and the red circle (the yellow square labeled "software") contains many software and internet firms around 2000, such as Oracle, SAP, America Online, Adobe, Symantec, McAfee, Google, and eBay. Intel's activities in the 1970s and the 1980s also appear in these nodes, which suggests Intel might have engaged in some of the path-breaking inventions in this area

[12] Many USPTO classes relate to computers and electronics. According to the NBER patent database, 35 classes (mostly in the 300s and the 700s) belong to "computers and communications" technologies, and 52 classes (mostly in the 300s as well) belong to "electrical and electronic." By contrast, only 14 classes are categorized as "drugs and medical."

Figure 5: The Mapper Graph of 333 Firms' Patent Portfolios (Both R&D and M&A)




*Note*: This graph uses correlation distance and $n = 20$. See section 3 for details.

20 years before the internet boom. We investigate IT firms more formally and closely in section 5.3.

### 4.1.2 Health Care, Basic Materials, and Consumer Goods

Health care is another R&D-intensive sector. Unlike IT firms, however, pharmaceutical firms do not appear in flares. Large drug-makers, such as Eli Lilly, Merck, and Pfizer, are clustered in the opposite side from IT firms (circled in green, at the southeast end of the main trunk). Patents are crucial for their business model, but most of the drug patents are in either class 424 or 514 (both are labeled "drug, bio-affecting, and body treating compositions"), which limits the extent to which their patent portfolios could differ from each other. Hence, further investigation into pharmaceutical innovations would need a higher resolution and/or subclass-level data.

Monsanto, an agrochemical firm, appears in a flare that connects with drug-makers, which suggests its patent portfolio is different but related. A closer look into Monsanto's time evolution reveals its flare is moving inward over time, rather than outward, as is the case for most other flares. More specifically, its patents in the 1970s and the 1980s are mostly unrelated to drugs, but those in the 1990s and the 2000s are filed in areas in which drug-makers patent. Monsanto is one of the few centripetal flares in the graph.[13]

Medtronic, a medical-device manufacturer, is in a short flare that is north of pharmaceuticals, in the neighborhood of Corning and Cie de Saint-Gobain, two large glass makers. Medical devices are made of a variety of materials, and hence Medtronic's R&D spans areas that are shared with the basic materials sector. Likewise, many consumer goods are based on materials and chemistry, and therefore are located close to the health care sector.[14]

#### 4.1.3 Conglomerates

Nevertheless, not all chemical firms are located near the health care sector. Bayer, Dow Chemical, DuPont, and 3M form their own connected component (see the large island in the southwest of Figure 5), which suggests their patent portfolios are similar to each other but substantially different from those of drug-makers and consumer-good makers.

GE, Philips, and Siemens form another large island (see the center-bottom part of the graph). GE, an archetypical conglomerate, holds the most diversified portfolio in our data. Its only peers are similarly diversified manufacturers of electronic and capital goods. Finally, IBM has no peers and appears as a set of islands by itself (see small islands in the southeast part).

### 4.2 Flare Characteristics

#### 4.2.1 Full-Sample Results

The Mapper graph helps us understand the firms' technological locations visually and intuitively, which is its main appeal. However, not all firms appear in flares or islands that are obvious to human eyes, and manually counting their lengths is difficult. The formal definitions and computational methods in sections 3.3 and 3.4 allow us to recognize the graph patterns of *all* firms, including those that are located within the densely populated trunks, and characterize them more precisely.

[13] We are currently investigating how to analyze the directions of flares (e.g., centripetal or centrifugal).

[14] For example, Johnson and Johnson, Unilever, Procter and Gamble, and Kimberly-Clark hold patents in not only classes such as 510 (cleaning compositions for solid surfaces, auxiliary compositions therefor, or processes of preparing the compositions), but also 424 (drug, bio-affecting, and body treating compositions) and 604 (surgery).

Table 2: Firm Count by Flare Type and Length

| Patents acquired by: | In-house R&D only | | | Both R&D and M&A | | |
|---|---|---|---|---|---|---|
| KeplerMapper resolution: | $n = 15$ | $n = 20$ | $n = 25$ | $n = 15$ | $n = 20$ | $n = 25$ |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Type | | | | | | |
| 0: No flares | 211 | 200 | 196 | 199 | 218 | 187 |
| 1: Flares only | 112 | 114 | 114 | 117 | 95 | 126 |
| 2: Flares and islands | 3 | 13 | 17 | 12 | 15 | 18 |
| 3: Islands only | 7 | 6 | 6 | 5 | 5 | 2 |
| Length | | | | | | |
| 0 (No flares) | 211 | 200 | 196 | 199 | 218 | 187 |
| 1 | 69 | 83 | 76 | 82 | 64 | 85 |
| 2 | 21 | 13 | 18 | 28 | 14 | 22 |
| 3 | 9 | 8 | 15 | 5 | 8 | 19 |
| 4 | 4 | 8 | 6 | 6 | 9 | 6 |
| 5 | 6 | 5 | 4 | 2 | 9 | 6 |
| 6 | 3 | 7 | 6 | 3 | 4 | 4 |
| 7 | 0 | 0 | 4 | 1 | 0 | 0 |
| 8 | 1 | 0 | 1 | 0 | 1 | 0 |
| 9 | 1 | 1 | 1 | 2 | 1 | 0 |
| 10 | 1 | 2 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 2 |
| $\infty$ (Islands only) | 7 | 6 | 6 | 5 | 5 | 2 |

*Note*: The underlying Mapper graphs are based on correlation distance. Column (2) corresponds to Figure 1 in the introduction. Column (5) corresponds to Figure 5.

Table 2 reports the results of our formal characterization. The first four rows classify each firm into the four flare types, whereas the rest of the table uses a finer classification system based on the lengths of the flares. Each of the six columns represents a different version of the graph. The first three columns use R&D patents only, and the last three use both R&D and M&A patents. In the spirit of sensitivity analysis, we use three different resolution levels ($n = 15$, 20, and 25 in KeplerMapper), where a higher number indicates higher resolution. Because higher $n$ leads to the construction of nodes and edges in smaller and more numerous local neighborhoods, higher $n$ tends to entail fewer cases of "no flares" (i.e., the method detects minor differences among broadly similar patent portfolios).[15] Higher $n$ also leads to fewer "islands only" firms, because the same firm's $p_{i,t}$ at different $t$s are more likely to be split into tiny fragments, some of which connect with other firms' $p_{j,t}$, thereby making the firm "flares and islands" instead of "islands only" (i.e., purely isolated subgraphs).

Our visual inspection in section 4.1 identified only a few dozen flares and islands, but the more systematic examination reveals the existence of many more. For the main version of

[15]The case of $n = 20$ with both R&D and M&A patents (column 5) is an exception to this tendency, which suggests these relationships are nonmonotonic.

the graph in Figure 5, Table 2 (column 5) reports 95 flares-only, 15 flares-and-islands, and 5 islands-only firms. Hence, more than a third of the 333 firms exhibit some uniqueness. The source of discrepancy between the manual and the automatic procedures is the presence of very short flares with only length 1 or 2.

Such negligible flares might appear unimportant and less interesting than long ones, but whether the presence and absence of these tiny spikes carry any information is an open empirical question. We investigate this issue in section 5 as part of our regression analysis. The benefit of our formal characterization is the ease and precision with which we can conduct a statistical analysis on flares at firm level.

### 4.2.2 Patents and Flares by Sector

The patenting behaviors and their topological patterns vary across industries. Table 3 lists 10 sectors used by Standard and Poor's (S&P), a credit-rating agency, in the descending order of firm count in our sample (see the second to last column). Technology and capital goods account for nearly two thirds of all firms, followed by health care, consumer goods, and basic materials, which collectively account for another third. Each of the four other sectors contains only a few firms, which makes their statistics highly idiosyncratic. Hence, we omit these sectors from our discussion below. Finally, 18 firms do not have a sector designation, because they are not publicly traded.

Table 3: Patents and Flares by Sector

| | Patent count (mean) | | | Flare type (# firms) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| S&P economic sector | R&D | M&A | Total | No flares | Flares only | Flares & islands | Islands only | Total | % flares |
| Technology | 2,665 | 339 | 3,004 | 60 | 37 | 7 | 2 | 106 | 43.4 |
| Capital goods | 1,836 | 241 | 2,077 | 65 | 21 | 2 | 3 | 91 | 28.6 |
| Health care | 890 | 248 | 1,138 | 33 | 11 | 0 | 0 | 44 | 25.0 |
| Consumer cyclicals | 453 | 41 | 494 | 22 | 6 | 0 | 0 | 28 | 21.4 |
| Basic materials | 2,389 | 239 | 2,628 | 13 | 9 | 1 | 0 | 23 | 43.5 |
| Consumer staples | 1,406 | 212 | 1,617 | 10 | 2 | 2 | 0 | 14 | 28.6 |
| Energy | 2,803 | 694 | 3,497 | 1 | 2 | 1 | 0 | 4 | 75.0 |
| Communication services | 5,456 | 43 | 5,500 | 0 | 2 | 1 | 0 | 3 | 100.0 |
| Transportation | 13,500 | 40 | 13,540 | 0 | 1 | 0 | 0 | 1 | 100.0 |
| Utilities | 711 | 243 | 954 | 0 | 1 | 0 | 0 | 1 | 100.0 |
| No S&P sector | 297 | 259 | 556 | 14 | 3 | 1 | 0 | 18 | 22.2 |

*Note*: "No S&P sector" means the firm does not have an S&P economic sector designation. Flare types and lengths are based on Figure 5 (i.e., the Mapper graph of both R&D and M&A patents with $n = 20$ and correlation distance).

The first three columns show the mean patent counts by sector. Technology and capital-goods firms lead the league table with their large patent portfolios. Basic materials and consumer staples are similarly patent-intensive. By contrast, health care and consumer cyclical firms have relatively few patents. The flare-type counts in the next four columns show analogous patterns. The tendency of more patent-intensive sectors to feature more flares and islands highlights an obvious relationship between patent count and flares: fewer patents mean a lower "degree of freedom" (i.e., fewer opportunities to construct a unique portfolio). Hence, we try to isolate the part of uniqueness that can be explained simply by the number of patents, by including patent count as a control variable in our subsequent analysis.

# 5 Correlations with Performance Measures

This section investigates whether flare types and flare length contain any relevant information. Following a common practice in the patent statistics literature (e.g., Pakes and Griliches 1984), we look for correlations between these topological characteristics and the firms' performance metrics, including revenues, profits, and stock market values.

## 5.1 Full-Sample Results

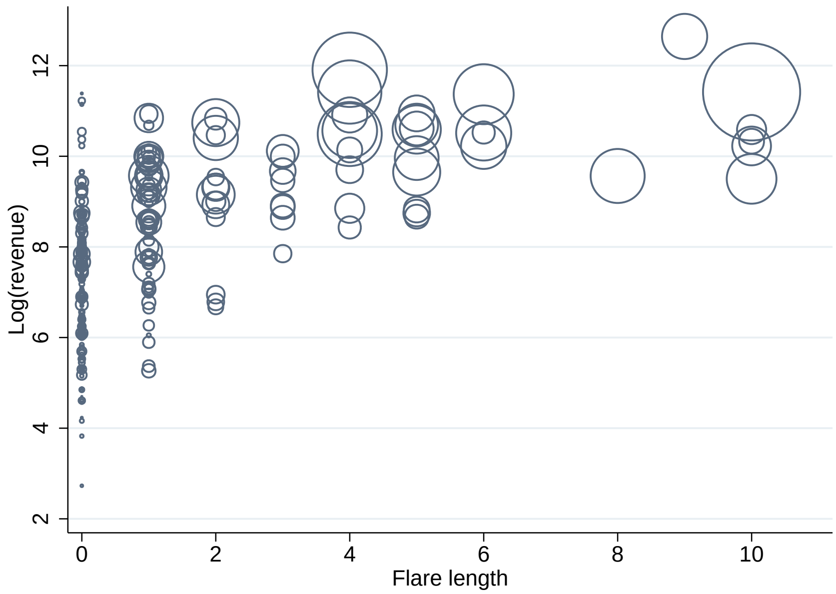
Figure 6 plots each firm's revenue in 2005 (on the vertical axis) against the flare length of its patenting patterns in 1976–2005 (on the horizontal axis). The circle size reflects the total count of patents. The maximum finite flare length of all firms in our baseline graph is 9; the figure shows infinitely long flares (i.e., islands-only type) at "length 10" for ease of visualization. Two patterns emerge. First, the upper-triangle-like shape of the scatter plot suggests long flares always entail high revenues, but the reverse is not true. Some high-revenue firms show short or no flares. Second, the prevalence of large circles in the upper-right region suggests long flares are frequently associated with many patents. However, some firms have relatively large portfolios but only short flares of length 1 or 2. Thus, long flares predict high revenues and many patents, but not all "large" firms exhibit long flares. We examine the details by sector in section 5.2.

Let us further investigate these correlations by using regressions of the following form:

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k d_{i,k} + \gamma \log(p_i) + \varepsilon_i, \tag{4}$$

where $y_i$ is the natural logarithm of firm $i$'s revenue (or other performance metrics) in 2005, $\alpha$ is a constant, $d_{i,k}$ is a dummy variable for firm $i$'s topological characteristic $k$ (i.e., $d_{i,k} = 1$

Figure 6: Revenues and Flares




*Note*: The center of each circle represents the firm's revenue in 2005 and the flare length of its patent portfolio in 1976–2005 (based on both R&D and M&A patents, correlation distance, and $n = 20$). The circle size reflects the total patent count across all classes and all years. Infinitely long flares (i.e., islands-only type) are shown at length 10 for illustration purposes.

if firm $i$'s patent portfolio in 1976–2005 belongs to type/length $k$, and $d_{i,k} = 0$ otherwise), $\beta_k$ is its coefficient, $p_i$ is the total count of firm $i$'s patents in 1976–2005 (i.e., $p_i = \sum_t \sum_c p_{i,t,c}$), $\gamma$ is its coefficient, and $\varepsilon_i$ is an error term.[16]

Table 4 shows the flare patterns are positively correlated with the firms' revenues in 2005. Column 1 uses each firm's flare type (no flare, flares only, flares and islands, or islands only), with "no flare" as the reference category ($k = 0$). The estimates suggest the revenues of the latter three types of firms are significantly higher than those of no-flare firms. Column 2 includes $\log(p_i)$ to control for the size of their patenting activities.[17] The coefficient estimates are 0.571, 0.996, and 1.828, which imply flares-only, flares-and-islands, and islands-only firms outperform no-flare firms by 77%, 171%, and 522%, respectively (e.g., $e^{0.571} = 1.77$). The flare types convey additional information beyond what can be predicted based on patent count alone. Columns 3 and 4 use a finer categorization scheme based on the

[16]Note we do not intend to prove causal relationships. Our purpose is to assess whether our uniqueness measures can predict any part of these performance metrics.

[17]The empirical literature on innovation has shown larger firms tend to patent more. Hence, to the extent that higher $p_i$ allows the firm to exhibit more uniqueness (i.e., a higher degree of freedom in shaping the distribution of patents across different classes), we should control for this connection between our uniqueness measures and the performance measures via $p_i$.

Table 4: Revenues and Flares ($n = 20$)

| LHS variable: | Log(Revenue in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.584 | 0.571 | – | – | 1.870 | 0.434 | – | – |
| | (0.174) | (0.222) | (–) | (–) | (0.172) | (0.227) | (–) | (–) |
| Flares and islands | 2.483 | 0.996 | – | – | 2.386 | 0.265 | – | – |
| | (0.418) | (0.448) | (–) | (–) | (0.375) | (0.416) | (–) | (–) |
| Islands only | 3.724 | 1.828 | – | – | 3.479 | 1.194 | – | – |
| | (0.624) | (0.645) | (–) | (–) | (0.594) | (0.593) | (–) | (–) |
| 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 = 8 | – | – | – | – | – | – | 3.143 | 0.939 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (1.258) | (1.205) |
| Flare length = 9 | – | – | 3.485 | 2.036 | – | – | 3.722 | 1.722 |
| | (–) | (–) | (1.343) | (1.317) | (–) | (–) | (1.299) | (1.234) |
| Flare length = 10 | – | – | 3.700 | 2.160 | – | – | – | – |
| | (–) | (–) | (0.927) | (0.938) | (–) | (–) | (–) | (–) |
| Flare length = $\infty$ | – | – | 3.780 | 2.258 | – | – | 3.511 | 1.523 |
| | (–) | (–) | (0.595) | (0.646) | (–) | (–) | (0.571) | (0.600) |
| Log(Patents) | – | 0.316 | – | 0.252 | – | 0.516 | – | 0.446 |
| | (–) | (0.047) | (–) | (0.050) | (–) | (0.060) | (–) | (0.065) |
| Adjusted $R^2$ | 0.384 | 0.469 | 0.440 | 0.487 | 0.434 | 0.551 | 0.478 | 0.555 |
| Number of observations | 286 | 286 | 286 | 286 | 288 | 288 | 288 | 288 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included. See Appendix C for additional results with (i) EBIT, (ii) market capitalization, (iii) $n = 15$, (iv) $n = 25$, and (v) Euclidean distance.

length of flares, where the reference category is "no flare" (i.e., zero-length flares), and the last category (infinitely long flares) represents "islands only." Their coefficients are positive and tend to increase with flare length.

These first four columns use patent statistics from the firms' in-house R&D activities only, whereas columns 5 through 8 also include patents that are obtained from target firms by M&As. The estimates are broadly similar to the first four columns, and the fit is better. For example, column 8 shows flares of lengths 4, 6, and $\infty$ are associated with $e^{1.246} - 1 = 248\%$, $e^{1.746} - 1 = 473\%$, and $e^{1.523} - 1 = 359\%$ higher revenues than zero-flare firms, respectively. Standard errors for some length categories are large, because few firms belong to them (e.g., Figure 6 shows only one firm each in length 8 and 9). Nevertheless, the overall pattern seems clear: long flares are frequently associated with more than 100% higher revenues even after controlling for the total number of patents.

Appendix C reports additional results using the other performance metrics: EBIT and market capitalization. We also report a sensitivity analysis with $n = 15$, $n = 25$, and Euclidean distance.

## 5.2 Sector-Level Analysis

Figure 7 decomposes the full-sample scatter plot in the previous subsection (Figure 6) by the S&P sector. The positive correlations between flares, revenues, and patent count are preserved within each sector, but important heterogeneity exists. First, the technology and capital-goods sectors (the top panels) contain many firms with long flares, and the positive correlations seem particularly clear.

Second, the health care and consumer-goods sectors (the middle panels) have only a few firms with flares longer than 1. The two health care firms with length-2 flares are Johnson and Johnson and Medtronic. All major pharmaceutical firms (e.g., Pfizer, Merck, Amgen, and Teva) have either length-1 or no flares, because only a few patent classes exist for drugs. Likewise, the consumer-goods sector has only a few firms with relatively long flares (e.g., Procter and Gamble). Several firms in this category earn high revenues but show no flares (e.g., Walt Disney, Nestle, Assa Abloy, and Cardinal Health), which present interesting counter-examples to the generally positive correlations.[18]

Third, most of the firms in basic materials and the rest of the sectors (the bottom panels) tend to be relatively large, have many patents, and feature some flares. Uniqueness is a relative concept; hence, giant firms in oligopolistic markets with capital-intensive technologies tend to have patent portfolios that are differentiated by construction.

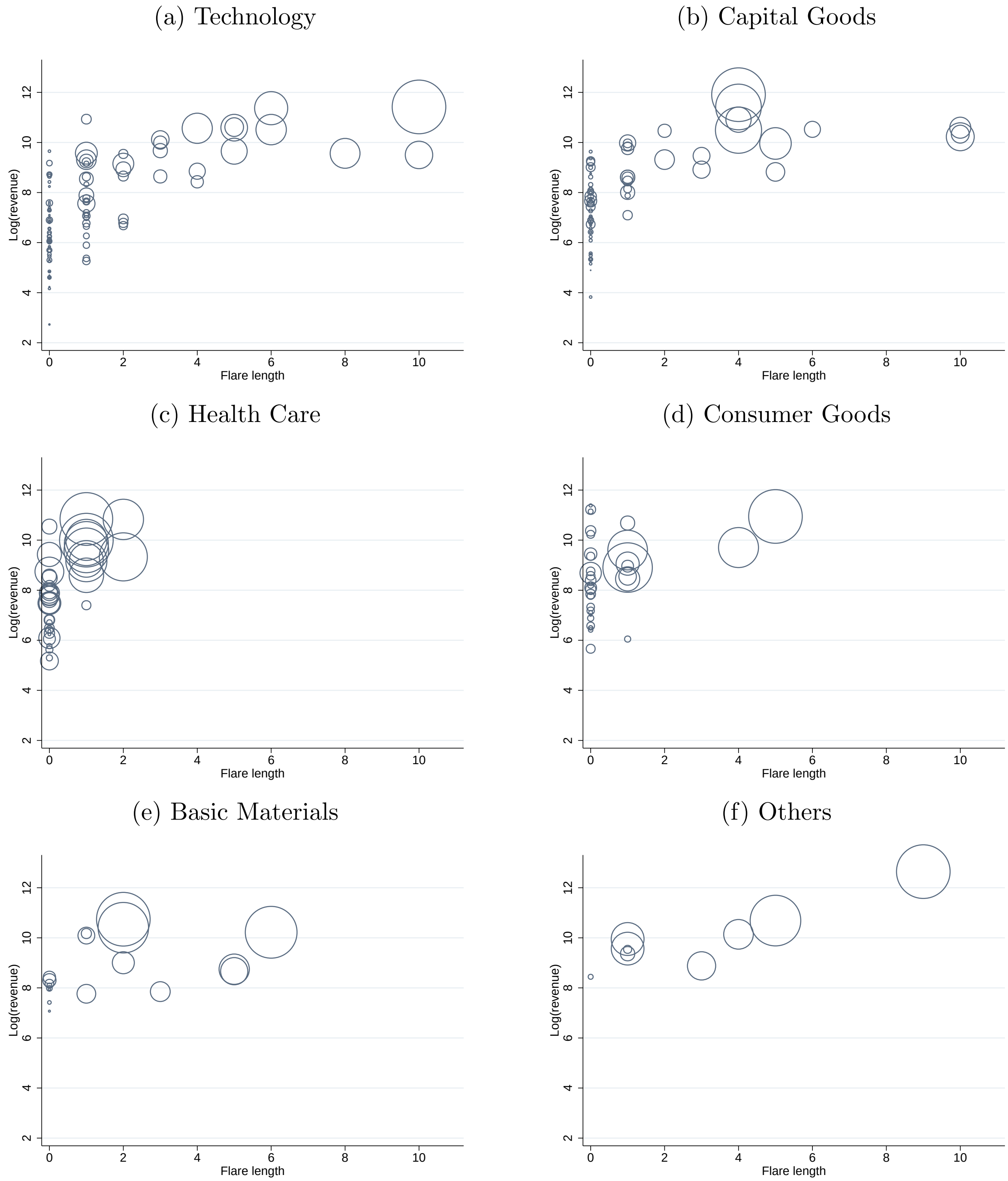
## 5.3 SIC Code-Level Comparisons: What Makes Portfolios Unique?

Figure 8 is similar to Figure 7 but focuses on more specific industries than broad sectors. Our purpose is to gain insights into what makes portfolios unique, by comparing firms in the same industry with different flare lengths. Appendix D shows firm-level bubble charts (similar to Figure 2 in section 2.5) for all the firms we discuss in this subsection.

Panel (a) plots semiconductor firms' revenues, flare lengths, and patent count. Texas Instruments and Intel have exceptionally large portfolios that resemble IBM's. Both the

[18]Upon closer examination, one would notice their business models do not rely on patents for obvious reasons (e.g., Walt Disney's main products are copyrighted characters and contents, not patented inventions).

Figure 7: Revenues and Flares by Sector

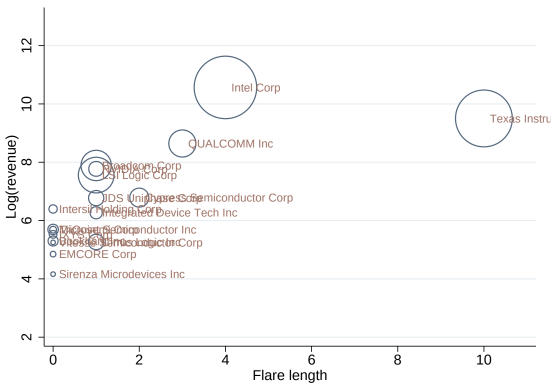



*Note*: “Consumer goods” include the S&P consumer cyclicals and consumer-staples sectors. “Others” include the S&P energy, communication services, transport, and utilities sectors.

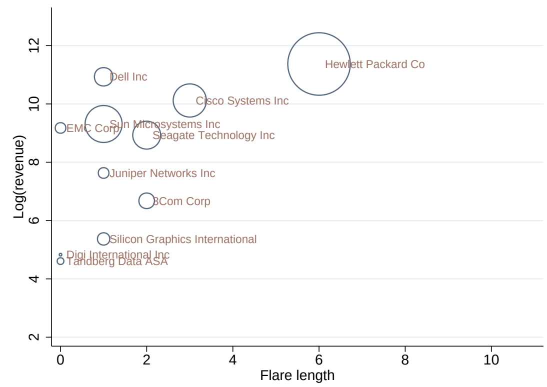
quantity and variety of patents seem to help make their portfolios unique (see the top panels of Figure 12 in Appendix D). However, note our definition of flares is based on *the graph*

Figure 8: Revenues and Flares by SIC Code

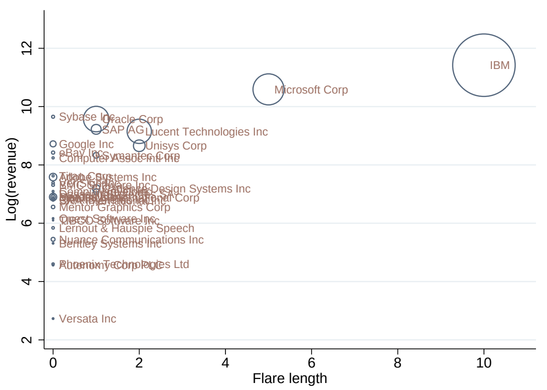
(a) Semiconductors



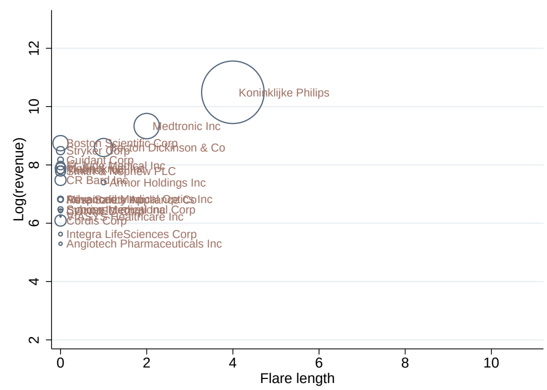

(b) Computers and Peripherals




(c) Programming and Software




(d) Medical Devices




*Note*: For semiconductors, we use SIC code 3674 (semiconductors and related devices). For computers and their peripherals, we use 3570 (computer and office equipment), 3571 (electronic computers), 3572 (computer storage devices), 3575 (computer terminals), and 3576 (computer communications equipment). Programming and software firms are in 7370 (services - computer programming, data processing, etc.), 7372 (services - prepackaged software), and 7373 (services - computer integrated systems design). For medical devices, we use 3841 (surgical and medical instruments and apparatus), 3842 (orthopedic, prosthetic and surgical appliances and supplies), 3843 (dental equipment and supplies), and 3845 (electromedical and electrotherapeutic apparatus). See Appendix D for the firm-level raw data (Figures 12, 13, and 14).

*of all firms in all years* and their distances from each other (e.g., Figure 5). Hence, Intel's proximity to Hewlett Packard and Lucent Technologies in that graph makes its flare length considerably shorter than Texas Instruments'. Qualcomm presents a good example of a unique portfolio despite having relatively few patents and seemingly simple distribution across classes.[19] By contrast, LSI Logic's portfolio is larger and more diverse, but its flare length is only 1, because other semiconductor firms have similar portfolios at multiple points in time. Finally, a comparison of NVIDIA (length 1) and TriQuint Semiconductor (length

[19]Its focus on telecommunication chips helped it become a dominant firm in this area in the 2010s.

0) highlights the role of time evolution. NVIDIA is active for only 11 years in our data, but its focus is clear and consistent over time. TriQuint's activities are less consistent in terms of both count and class, which leads to a fragmented subgraph that is connected to other firms in many nodes.

Panel (b) of Figure 8 shows the manufacturers of computers and their peripherals.[20] Hewlett Packard (length 6) has a massive portfolio, whereas Dell's (length 1) is much smaller. Both of them are among the largest computer makers, and their main patent classes are similar, but their approaches to R&D are different. Hewlett Packard is a traditional computer maker, whereas Dell's success is usually attributed to its unique business model in which the company sells directly to consumers and most of the manufacturing is outsourced to third-party suppliers in Asia, but such "business-model innovations" do not represent patentable inventions in most cases. Hence, patent statistics (and their topological representations) do not reflect Dell's "uniqueness" in this sense.

Panel (c) plots software firms.[21] Both Microsoft (length 5) and Oracle (length 1) sell packaged software, and mostly patent in classes that correspond to software (the early 700s). Many firms are active in this area, which tends to make their portfolios undifferentiated. However, Microsoft shows a considerable level of patenting in the 300s as well. This pattern visibly distinguishes the company from others in the trunk of the graph. More recent firms, such as Google and eBay, do not show flares. Certain algorithms represent patentable inventions, and Google's founders patented ones about their page-ranking methods in their initial years. However, their track records in our data are too short to generate flares.[22]

Panel (d) focuses on medical devices.[23] Figure 2 in section 2.5 shows Medtronic's patent portfolio (length 2), which looks broadly similar to Becton Dickinson's (length 1). Most other firms do not show flares. A visual inspection of the portfolios of Boston Scientific, CR Bard, and Cordis suggests their R&D activities are confined to "surgery" classes (the early 600s), which limits the extent to which they can differentiate.

# 6 Conclusion

We find our topological method offers a useful way to understand firms' locations in technological space. The high dimensionality of technological space, as well as its unstructured

[20]See panels (a) and (b) of Figure 13 in Appendix D for their firm-level raw data.

[21]See panels (c) through (f) of Figure 13 in Appendix D for their firm-level raw data.

[22]We are currently conducting the same analysis for only the last 5–10 years with shorter moving windows, so that the activities of more recent firms become visible.

[23]See Figure 14 in Appendix D for their firm-level raw data.

nature, has been a major obstacle to answering basic and important economic questions, such as "where do firms innovate?" and "how do firms' technological positions interact with each other?" The Mapper graph captures salient patterns of firms' patenting behaviors in an intuitive manner. Most importantly, it creates a global map of all firms' positions in all years, rather than a snap shot of each firm or industry in isolation. Because technological space is high dimensional and sparsely populated, precise locations with *absolute* addresses are not necessarily useful for describing firms' technological positions and their interactions. Hence, mapping their *relative* locations in a graph seems a valuable alternative.

The Mapper graph also provides a "map" to guide our subsequent, more detailed analysis. The definition, detection, and measurement of flares are an example of such investigations. Our analysis suggests these uniqueness measures contain relevant information that relates to these firms' underlying activities as well as financial performances. Moreover, part of this information is "new," in the sense that conventional measures, such as total patent count (i.e., the size of R&D activities), does not necessarily carry the same information, as our statistical analysis shows.

This research raises more questions than it answers, as it represents only a first step toward analyzing firms' behaviors in technological space more formally. In terms of methodology, we are currently investigating the possibility of characterizing other features of the Mapper graph. In terms of applications, we aim to develop a follow-up research to target more specific questions concerning competition and innovation. Application of TDA to other economic activities would seem promising as well.

# Appendix (For Online Publication)

## A. Proofs

First, we show the boundary $G_i \setminus F_i$ indeed serves as a "boundary" for $F_i$: to get outside of $G_i$, one always needs to go through the boundary.

**Lemma 12.** *Let $u \in F_i$ and $w \in G \setminus G_i$, and let $p$ be a path from $u$ to $w$. Then, the path $p$ passes through some node $v \in G_i \setminus F_i$.*

*Proof.* Let $p$ be such a path from $u \in F_i$ to $w \in G \setminus G_i$, which passes through the nodes

$$u = v_0, v_1, v_2, \dots v_{n-1}, v_n = w$$

in that order.

Suppose, to the contrary, that all $v_j$ are not in the boundary $G_i \setminus F_i$. We show by induction that $v_j \in F_i$ for all $j \in \{0, \dots, n\}$. First, $v_0 = u \in F_i$ is clear. Suppose $v_j \in F_i$. Because $v_{j+1} \in B_1(v_j) \subseteq G_i$ by definition of the interior $F_i$, and because $v_{j+1} \notin G_i \setminus F_i$ by assumption, we see $v_{j+1} \in F_i$. Thus, by induction, $v_j \in F_i$ for all $j \in \{0, \dots, n\}$. In particular, $v_n = w \in F_i$, which is a contradiction, because $w \in G \setminus G_i \subseteq G \setminus F_i$.

Therefore, some $v_j$ exists in the boundary $G_i \setminus F_i$. □

For $i$, a firm, and $u \in F_i$, recall the exit distance of $u$ in $F_i$ was defined to be

$$e_i(u) = \min\{d(u, w) \mid w \in G \setminus F_i\}$$

in Definition 6. Here, we reproduce Theorem 7 and provide a proof.

**Theorem 7.** *Let $u \in F_i$. Then,*

$$e_i(u) = \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\},$$

*where $d_{G_i}(u, v)$ is the distance between $u$ and $v$ in $G_i$.*

*Proof.* It is clear that

$$\min\{d(u, w) \mid w \in G \setminus F_i\} \leq \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\}.$$

Suppose the minimum of the left-hand side is achieved by a $w \in G \setminus F_i$, and let $d(u, w) = \ell(p)$, the length of a minimum path $p$ in $G$ from $u \in F_i$ to $w \in G \setminus F_i$. Let $v$ be the first node $v \in G_i \setminus F_i$ that $p$ passes through. Note such $v$ exists by Lemma 12.

In the case in which $v \neq w$, truncate $p$ to the path $p'$ from $u$ to $v$. By choice of $v$, $p'$ is fully contained in $G_i$, and $\ell(p') \leq \ell(p)$ because we only have positive weights and $p'$ has strictly fewer edges than $p$. It follows that

$$\min\{d(u, w) \mid w \in G \setminus F_i\} = \ell(p) > \ell(p') \geq \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\},$$

because $p'$ is a path from $u$ to $v$ that is contained in $G_i$. This is a contradiction.

Thus, $v = w$, and it follows that

$$\min\{d(u, w) \mid w \in G \setminus F_i\} = \ell(p) \geq \min\{d_{G_i}(u, v) \mid v \in G_i \setminus F_i\},$$

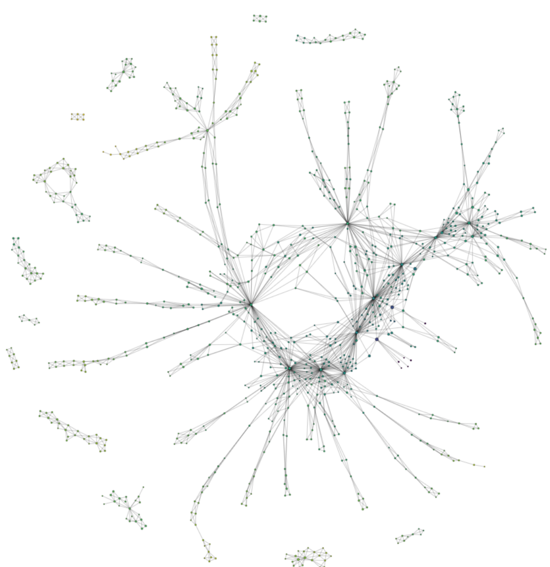
which shows the required equality. $\square$

# B. Additional Mapper Graphs

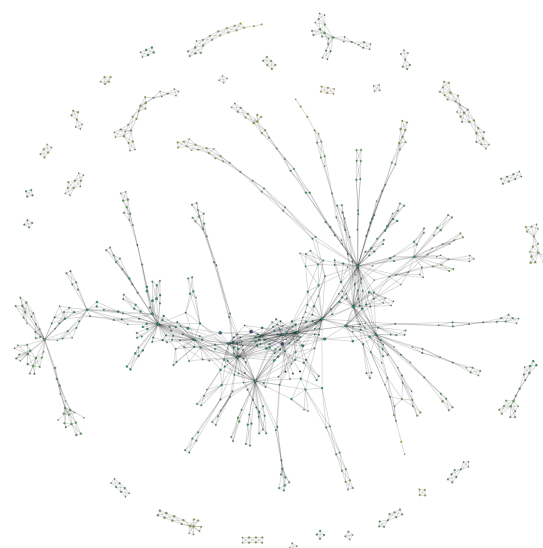
We report additional Mapper graphs as a sensitivity check. Figure 9 shows the results under $n = 15$ and $n = 25$, respectively. Figure 10 shows the results based on Euclidean distance (instead of correlation distance) under $n = 20$.

Figure 9: The Mapper Graphs ($n = 15$ and $n = 25$)

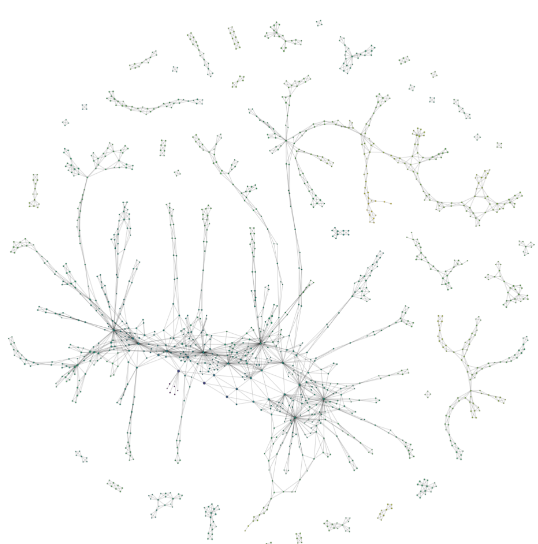
(a) In-House R&D Only ($n = 15$)



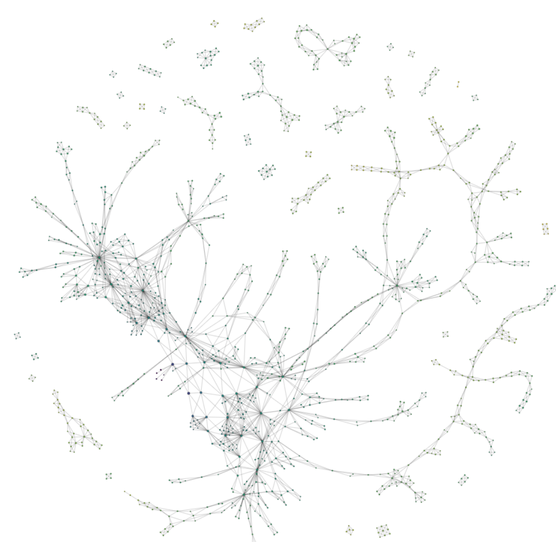
(b) Both R&D and M&A ($n = 15$)



(c) In-House R&D Only ($n = 25$)



(d) Both R&D and M&A ($n = 25$)

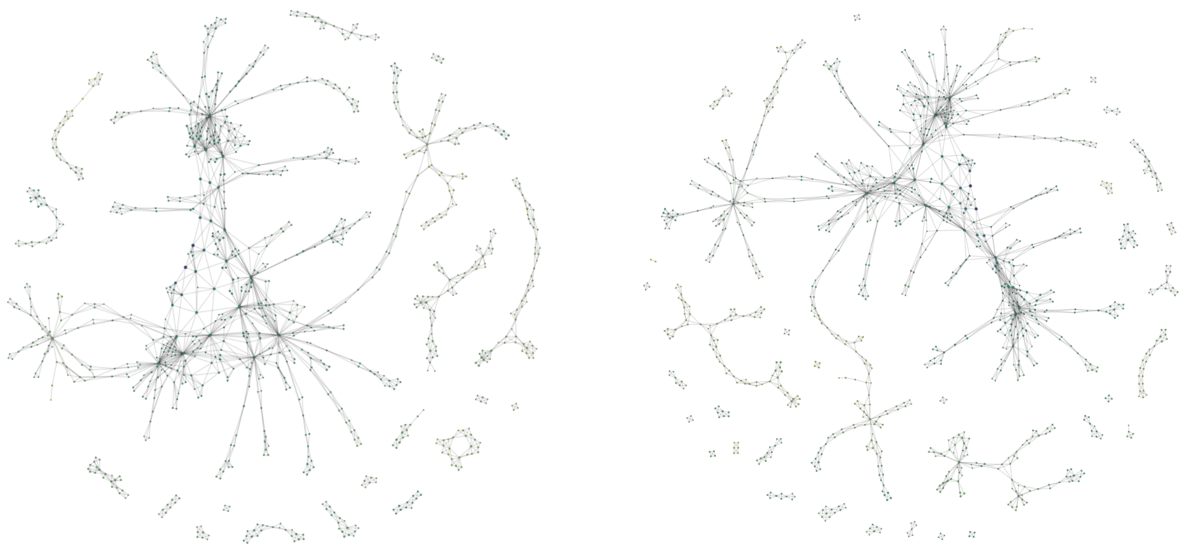


*Note*: These graphs use correlation distance.

Figure 10: The Mapper Graphs (Euclidean Distance, $n = 20$)

(a) In-House R&D Only

(b) Both R&D and M&A

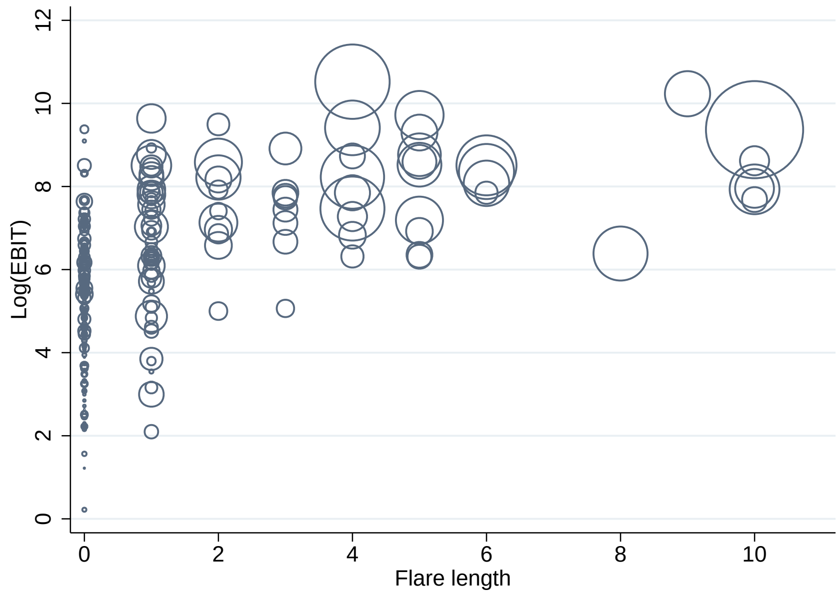


*Note*: These graphs use Euclidean distance.

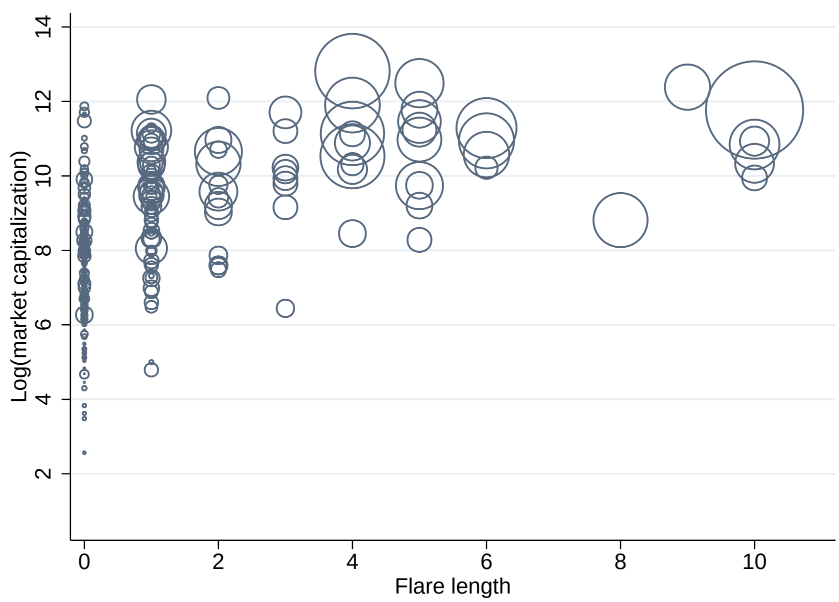
## C. Additional Scatter Plots and Regression Tables

Figure 11: EBIT, Market Capitalization, and Flares

(a) EBIT




(b) Market Capitalization




*Note*: The center of each circle represents the firm's EBIT (or stock-market capitalization) in 2005 and the flare length of its patent portfolio in 1976–2005 (based on both R&D and M&A patents, correlation distance, and $n = 20$). The circle size reflects the total patent count across all classes and all years. Infinitely long flares (i.e., islands-only type) are shown at length 10 for illustration purposes.

Table 5: EBIT (Earnings before Interest and Taxes) and Flares ($n = 20$)

| LHS variable: | Log(EBIT in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.583 | 0.397 | – | – | 1.939 | 0.189 | – | – |
| | (0.206) | (0.263) | (–) | (–) | (0.205) | (0.267) | (–) | (–) |
| Flares and islands | 1.838 | 0.135 | – | – | 1.832 | −0.695 | – | – |
| | (0.495) | (0.527) | (–) | (–) | (0.442) | (0.480) | (–) | (–) |
| Islands only | 3.958 | 1.784 | – | – | 3.515 | 0.799 | – | – |
| | (0.707) | (0.733) | (–) | (–) | (0.674) | (0.664) | (–) | (–) |
| Flare length = 1 | – | – | 1.076 | 0.310 | – | – | 1.305 | 0.093 |
| | (–) | (–) | (0.215) | (0.262) | (–) | (–) | (0.227) | (0.273) |
| Flare length = 2 | – | – | 2.351 | 1.294 | – | – | 2.492 | 0.720 |
| | (–) | (–) | (0.487) | (0.517) | (–) | (–) | (0.449) | (0.487) |
| Flare length = 3 | – | – | 2.698 | 1.453 | – | – | 2.443 | 0.752 |
| | (–) | (–) | (0.539) | (0.580) | (–) | (–) | (0.525) | (0.541) |
| Flare length = 4 | – | – | 2.047 | 0.737 | – | – | 3.156 | 0.861 |
| | (–) | (–) | (0.550) | (0.595) | (–) | (–) | (0.496) | (0.565) |
| Flare length = 5 | – | – | 2.825 | 1.407 | – | – | 2.859 | 0.684 |
| | (–) | (–) | (0.692) | (0.728) | (–) | (–) | (0.501) | (0.558) |
| Flare length = 6 | – | – | 3.434 | 1.908 | – | – | 3.428 | 1.089 |
| | (–) | (–) | (0.593) | (0.653) | (–) | (–) | (0.722) | (0.745) |
| Flare length = 8 | – | – | – | – | – | – | 1.896 | −0.667 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (1.426) | (1.361) |
| Flare length = 9 | – | – | 3.653 | 2.056 | – | – | 3.672 | 1.355 |
| | (–) | (–) | (1.513) | (1.489) | (–) | (–) | (1.470) | (1.390) |
| Flare length = 10 | – | – | 3.212 | 1.509 | – | – | – | – |
| | (–) | (–) | (1.045) | (1.064) | (–) | (–) | (–) | (–) |
| Flare length = $\infty$ | – | – | 4.050 | 2.348 | – | – | 3.548 | 1.240 |
| | (–) | (–) | (0.672) | (0.737) | (–) | (–) | (0.647) | (0.682) |
| Log(Patents) | – | 0.360 | – | 0.276 | – | 0.612 | – | 0.516 |
| | (–) | (0.055) | (–) | (0.058) | (–) | (0.069) | (–) | (0.075) |
| Adjusted $R^2$ | 0.301 | 0.402 | 0.369 | 0.420 | 0.355 | 0.507 | 0.407 | 0.500 |
| Number of observations | 262 | 262 | 262 | 262 | 264 | 264 | 264 | 264 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included.

Table 6: Stock-Market Capitalization and Flares ($n = 20$)

| LHS variable: | Log(Market capitalization in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.543 | 0.241 | – | – | 1.882 | 0.062 | – | – |
| | (0.202) | (0.253) | (–) | (–) | (0.202) | (0.260) | (–) | (–) |
| Flares and islands | 2.532 | 0.627 | – | – | 2.399 | −0.277 | – | – |
| | (0.483) | (0.507) | (–) | (–) | (0.436) | (0.473) | (–) | (–) |
| Islands only | 3.895 | 1.466 | – | – | 3.458 | 0.576 | – | – |
| | (0.719) | (0.730) | (–) | (–) | (0.691) | (0.674) | (–) | (–) |
| Flare length = 1 | – | – | 1.188 | 0.217 | – | – | 1.448 | 0.048 |
| | (–) | (–) | (0.215) | (0.255) | (–) | (–) | (0.225) | (0.266) |
| Flare length = 2 | – | – | 1.805 | 0.489 | – | – | 1.957 | 0.028 |
| | (–) | (–) | (0.448) | (0.470) | (–) | (–) | (0.416) | (0.443) |
| Flare length = 3 | – | – | 2.819 | 1.243 | – | – | 2.331 | 0.384 |
| | (–) | (–) | (0.559) | (0.582) | (–) | (–) | (0.543) | (0.544) |
| Flare length = 4 | – | – | 2.148 | 0.480 | – | – | 3.462 | 0.809 |
| | (–) | (–) | (0.571) | (0.599) | (–) | (–) | (0.514) | (0.567) |
| Flare length = 5 | – | – | 3.052 | 1.246 | – | – | 2.935 | 0.425 |
| | (–) | (–) | (0.720) | (0.735) | (–) | (–) | (0.519) | (0.560) |
| Flare length = 6 | – | – | 3.320 | 1.386 | – | – | 3.434 | 0.736 |
| | (–) | (–) | (0.614) | (0.654) | (–) | (–) | (0.749) | (0.751) |
| Flare length = 8 | – | – | – | – | – | – | 1.548 | −1.406 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (1.478) | (1.377) |
| Flare length = 9 | – | – | 3.478 | 1.438 | – | – | 3.180 | 0.492 |
| | (–) | (–) | (1.575) | (1.511) | (–) | (–) | (1.526) | (1.410) |
| Flare length = 10 | – | – | 3.472 | 1.304 | – | – | – | – |
| | (–) | (–) | (1.087) | (1.076) | (–) | (–) | (–) | (–) |
| Flare length = $\infty$ | – | – | 3.973 | 1.805 | – | – | 3.490 | 0.822 |
| | (–) | (–) | (0.698) | (0.742) | (–) | (–) | (0.670) | (0.687) |
| Log(Patents) | – | 0.401 | – | 0.355 | – | 0.652 | – | 0.599 |
| | (–) | (0.054) | (–) | (0.057) | (–) | (0.069) | (–) | (0.074) |
| Adjusted $R^2$ | 0.306 | 0.424 | 0.347 | 0.428 | 0.351 | 0.509 | 0.390 | 0.508 |
| Number of observations | 284 | 284 | 284 | 284 | 286 | 286 | 286 | 286 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included.

Table 7: Revenues and Flares ($n = 15$)

| LHS variable: | Log(Revenue in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.778 | 0.901 | – | – | 1.708 | 0.528 | – | – |
| | (0.172) | (0.206) | (–) | (–) | (0.168) | (0.203) | (–) | (–) |
| Flares and islands | 3.609 | 1.924 | – | – | 2.969 | 0.940 | – | – |
| | (0.832) | (0.813) | (–) | (–) | (0.412) | (0.436) | (–) | (–) |
| Islands only | 3.181 | 1.601 | – | – | 4.200 | 1.489 | – | – |
| | (0.534) | (0.549) | (–) | (–) | (0.597) | (0.617) | (–) | (–) |
| Flare length = 1 | – | – | 1.374 | 0.842 | – | – | 1.374 | 0.533 |
| | (–) | (–) | (0.190) | (0.209) | (–) | (–) | (0.178) | (0.204) |
| Flare length = 2 | – | – | 1.978 | 1.080 | – | – | 2.478 | 0.951 |
| | (–) | (–) | (0.312) | (0.346) | (–) | (–) | (0.267) | (0.329) |
| Flare length = 3 | – | – | 1.979 | 0.967 | – | – | 2.679 | 1.076 |
| | (–) | (–) | (0.461) | (0.483) | (–) | (–) | (0.667) | (0.656) |
| Flare length = 4 | – | – | 3.617 | 2.537 | – | – | 2.701 | 0.891 |
| | (–) | (–) | (0.652) | (0.658) | (–) | (–) | (0.528) | (0.551) |
| Flare length = 5 | – | – | 2.907 | 1.679 | – | – | 1.780 | 0.187 |
| | (–) | (–) | (0.585) | (0.608) | (–) | (–) | (0.905) | (0.864) |
| Flare length = 6 | – | – | 3.499 | 2.219 | – | – | 3.289 | 1.133 |
| | (–) | (–) | (0.749) | (0.758) | (–) | (–) | (0.895) | (0.880) |
| Flare length = 7 | – | – | – | – | – | – | 4.322 | 2.119 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (1.256) | (1.200) |
| Flare length = 8 | – | – | 5.145 | 3.492 | – | – | – | – |
| | (–) | (–) | (1.288) | (1.272) | (–) | (–) | (–) | (–) |
| Flare length = 9 | – | – | 4.292 | 2.912 | – | – | 4.698 | 2.453 |
| | (–) | (–) | (1.288) | (1.260) | (–) | (–) | (0.895) | (0.885) |
| Flare length = 10 | – | – | 4.239 | 2.872 | – | – | – | – |
| | (–) | (–) | (1.288) | (1.259) | (–) | (–) | (–) | (–) |
| Flare length = $\infty$ | – | – | 3.183 | 1.880 | – | – | 4.128 | 1.750 |
| | (–) | (–) | (0.511) | (0.550) | (–) | (–) | (0.574) | (0.629) |
| Log(Patents) | – | 0.288 | – | 0.235 | – | 0.480 | – | 0.429 |
| | (–) | (0.043) | (–) | (0.046) | (–) | (0.056) | (–) | (0.061) |
| Adjusted $R^2$ | 0.406 | 0.488 | 0.457 | 0.504 | 0.439 | 0.557 | 0.481 | 0.559 |
| Number of observations | 286 | 286 | 286 | 286 | 288 | 288 | 288 | 288 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included.

Table 8: Revenues and Flares ($n = 25$)

| LHS variable: | Log(Revenue in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.497 | 0.588 | – | – | 1.421 | 0.096 | – | – |
| | (0.175) | (0.206) | (–) | (–) | (0.175) | (0.191) | (–) | (–) |
| Flares and islands | 2.775 | 1.263 | – | – | 2.639 | 0.216 | – | – |
| | (0.359) | (0.394) | (–) | (–) | (0.369) | (0.381) | (–) | (–) |
| Islands only | 2.895 | 1.268 | – | – | 3.712 | 0.635 | – | – |
| | (0.655) | (0.646) | (–) | (–) | (1.008) | (0.891) | (–) | (–) |
| Flare length = 1 | – | – | 1.131 | 0.537 | – | – | 0.968 | 0.073 |
| | (–) | (–) | (0.192) | (0.211) | (–) | (–) | (0.187) | (0.194) |
| Flare length = 2 | – | – | 1.769 | 0.793 | – | – | 1.898 | 0.278 |
| | (–) | (–) | (0.343) | (0.369) | (–) | (–) | (0.316) | (0.334) |
| Flare length = 3 | – | – | 2.741 | 1.470 | – | – | 2.880 | 0.732 |
| | (–) | (–) | (0.378) | (0.425) | (–) | (–) | (0.349) | (0.393) |
| Flare length = 4 | – | – | 3.067 | 1.719 | – | – | 2.429 | 0.539 |
| | (–) | (–) | (0.565) | (0.588) | (–) | (–) | (0.571) | (0.548) |
| Flare length = 5 | – | – | 2.390 | 1.163 | – | – | 2.789 | 0.296 |
| | (–) | (–) | (0.677) | (0.678) | (–) | (–) | (0.566) | (0.575) |
| Flare length = 6 | – | – | 2.127 | 0.882 | – | – | 2.691 | 0.267 |
| | (–) | (–) | (0.566) | (0.581) | (–) | (–) | (0.679) | (0.659) |
| Flare length = 7 | – | – | 3.289 | 1.862 | – | – | – | – |
| | (–) | (–) | (0.777) | (0.780) | (–) | (–) | (–) | (–) |
| Flare length = 8 | – | – | 4.082 | 2.543 | – | – | – | – |
| | (–) | (–) | (1.336) | (1.297) | (–) | (–) | (–) | (–) |
| Flare length = 9 | – | – | 3.070 | 1.571 | – | – | – | – |
| | (–) | (–) | (1.336) | (1.295) | (–) | (–) | (–) | (–) |
| Flare length = 11 | – | – | – | – | – | – | 3.590 | 0.944 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (0.959) | (0.893) |
| Flare length = $\infty$ | – | – | 2.966 | 1.549 | – | – | 3.806 | 0.916 |
| | (–) | (–) | (0.637) | (0.656) | (–) | (–) | (0.959) | (0.908) |
| Log(Patents) | – | 0.315 | – | 0.269 | – | 0.581 | – | 0.534 |
| | (–) | (0.045) | (–) | (0.048) | (–) | (0.054) | (–) | (0.061) |
| Adjusted $R^2$ | 0.378 | 0.472 | 0.415 | 0.475 | 0.349 | 0.543 | 0.412 | 0.541 |
| Number of observations | 286 | 286 | 286 | 286 | 288 | 288 | 288 | 288 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included.

Table 9: Revenues and Flares ($n = 20$, Euclidean Distance)

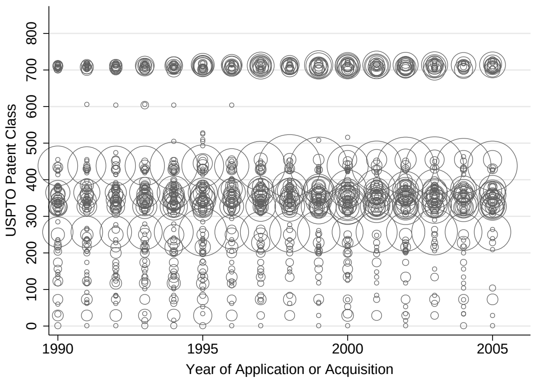
| LHS variable: | Log(Revenue in 2005) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patents acquired by: | In-house R&D only | | | | Both R&D and M&A | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Flares only | 1.666 | 0.568 | – | – | 1.720 | 0.189 | – | – |
| | (0.180) | (0.230) | (–) | (–) | (0.173) | (0.222) | (–) | (–) |
| Flares and islands | 2.479 | 0.946 | – | – | 2.646 | 0.356 | – | – |
| | (0.509) | (0.521) | (–) | (–) | (0.370) | (0.403) | (–) | (–) |
| Islands only | 3.786 | 1.878 | – | – | 4.023 | 0.800 | – | – |
| | (0.699) | (0.703) | (–) | (–) | (1.050) | (0.976) | (–) | (–) |
| Flare length = 1 | – | – | 1.138 | 0.410 | – | – | 1.296 | 0.167 |
| | (–) | (–) | (0.206) | (0.238) | (–) | (–) | (0.193) | (0.227) |
| Flare length = 2 | – | – | 1.962 | 0.913 | – | – | 2.189 | 0.468 |
| | (–) | (–) | (0.355) | (0.389) | (–) | (–) | (0.306) | (0.355) |
| Flare length = 3 | – | – | 2.626 | 1.427 | – | – | 2.344 | 0.639 |
| | (–) | (–) | (0.401) | (0.441) | (–) | (–) | (0.447) | (0.461) |
| Flare length = 4 | – | – | 2.720 | 1.317 | – | – | 2.984 | 0.555 |
| | (–) | (–) | (0.574) | (0.604) | (–) | (–) | (0.485) | (0.540) |
| Flare length = 5 | – | – | 2.576 | 1.448 | – | – | 2.652 | 0.426 |
| | (–) | (–) | (0.621) | (0.626) | (–) | (–) | (0.477) | (0.519) |
| Flare length = 6 | – | – | 3.615 | 2.064 | – | – | 3.414 | 1.038 |
| | (–) | (–) | (0.946) | (0.943) | (–) | (–) | (0.765) | (0.758) |
| Flare length = 7 | – | – | 3.557 | 2.035 | – | – | 3.623 | 1.048 |
| | (–) | (–) | (0.946) | (0.942) | (–) | (–) | (0.930) | (0.905) |
| Flare length = 8 | – | – | – | – | – | – | 3.180 | 0.572 |
| | (–) | (–) | (–) | (–) | (–) | (–) | (1.306) | (1.230) |
| Flare length = 9 | – | – | 3.318 | 1.868 | – | – | – | – |
| | (–) | (–) | (1.365) | (1.325) | (–) | (–) | (–) | (–) |
| Flare length = 10 | – | – | 3.150 | 1.505 | – | – | – | – |
| | (–) | (–) | (1.329) | (1.299) | (–) | (–) | (–) | (–) |
| Flare length = $\infty$ | – | – | 3.850 | 2.246 | – | – | 4.216 | 1.085 |
| | (–) | (–) | (0.674) | (0.705) | (–) | (–) | (1.018) | (1.006) |
| Log(Patents) | – | 0.325 | – | 0.268 | – | 0.561 | – | 0.517 |
| | (–) | (0.047) | (–) | (0.049) | (–) | (0.059) | (–) | (0.067) |
| Adjusted $R^2$ | 0.376 | 0.467 | 0.422 | 0.478 | 0.398 | 0.544 | 0.438 | 0.539 |
| Number of observations | 286 | 286 | 286 | 286 | 288 | 288 | 288 | 288 |

*Note*: Standard errors are in parentheses. S&P economic sector dummies are included.

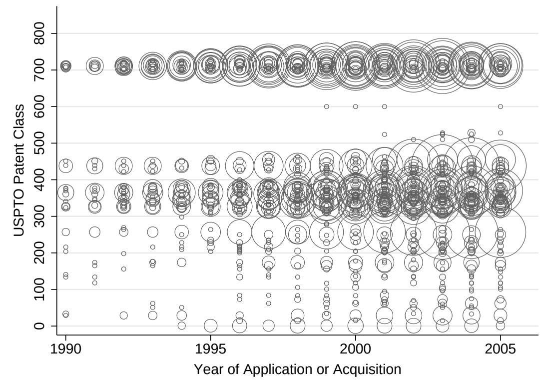
# D. Additional Plots of Firm-Level Patent Portfolios

Figure 12: Patent Portfolios of Semiconductor Firms

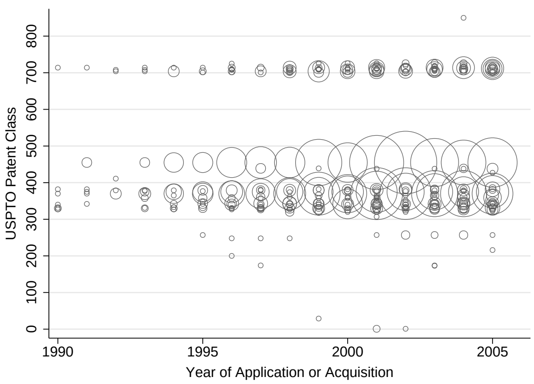
(a) Texas Instruments



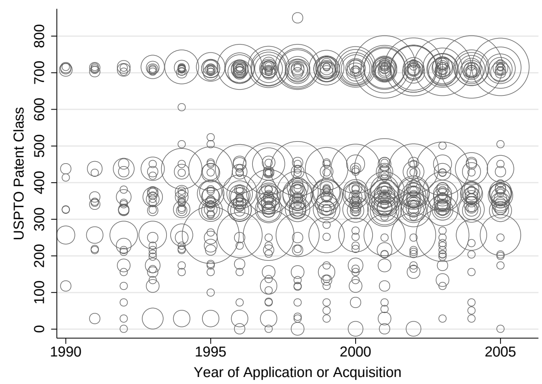

(b) Intel



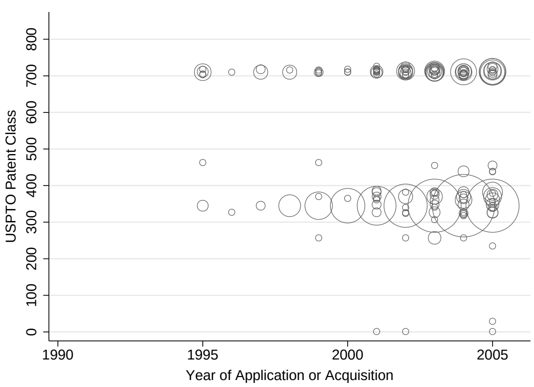

(c) Qualcomm



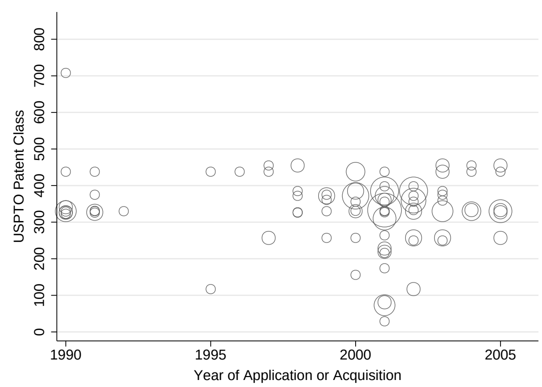

(d) LSI Logic




(e) NVIDIA



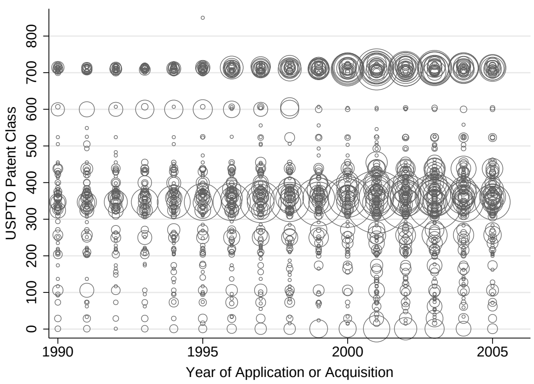

(f) TriQuint Semiconductor



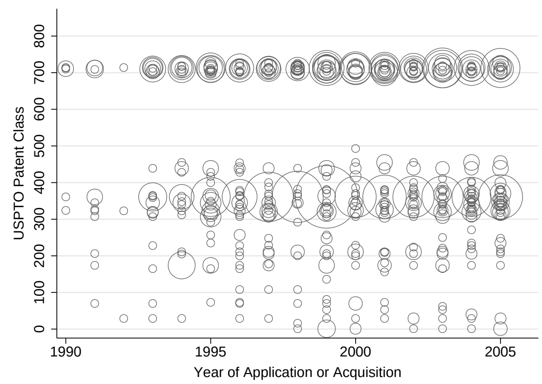

*Note*: The circle size represents the number of patents in each class-year. Based on our method and analysis in sections 3 and 4, the "flare lengths" of these firms' portfolios are: $\infty$ (Texas Instruments), 4 (Intel), 3 (Qualcomm), 1 (LSI Logic), 1 (NVIDIA), and 0 (TriQuint).

Figure 13: Patent Portfolios of Computer, Software, and Internet Firms

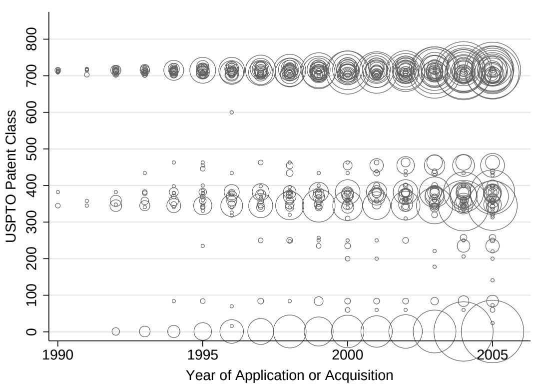
(a) Hewlett Packard



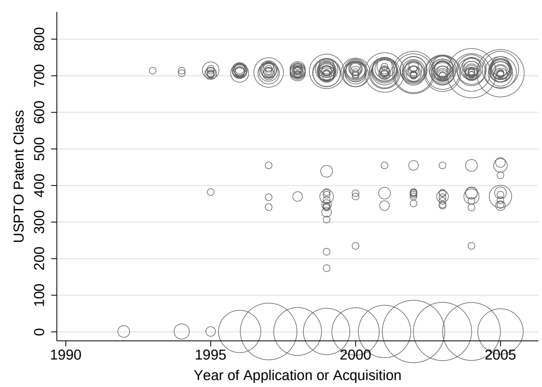

(b) Dell



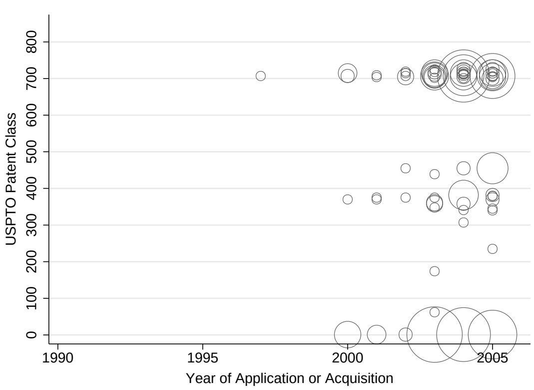

(c) Microsoft



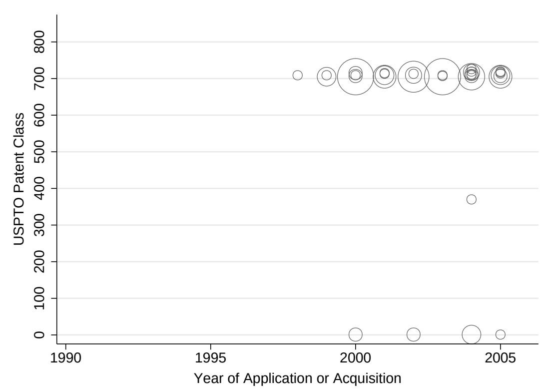

(d) Oracle




(e) Google



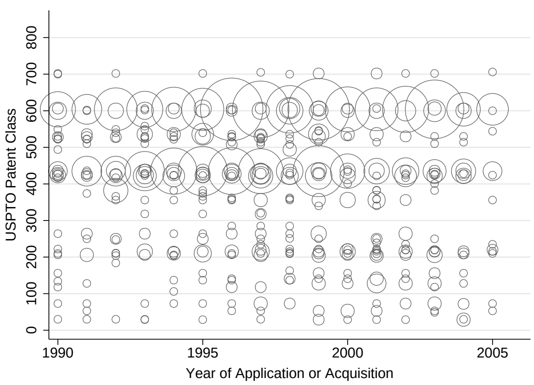

(f) eBay



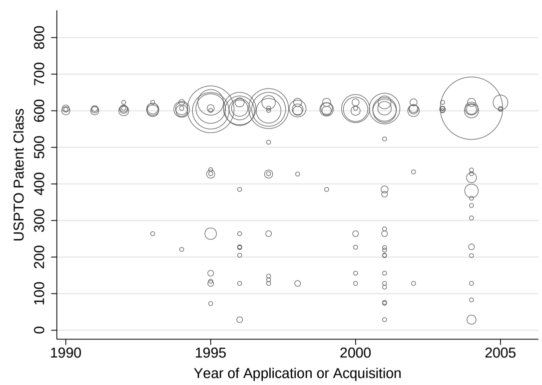

*Note*: The circle size represents the number of patents in each class-year. Based on our method and analysis in sections 3 and 4, the "flare lengths" of these firms' portfolios are: 6 (Hewlett Packard), 1 (Dell), 5 (Microsoft), 1 (SAP), 0 (Google), and 0 (eBay).

Figure 14: Patent Portfolios of Medical Device Firms

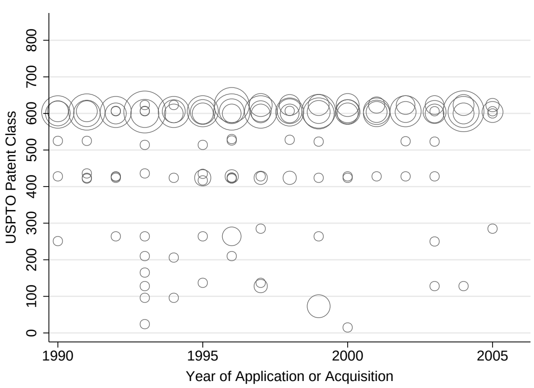
(a) Becton Dickinson



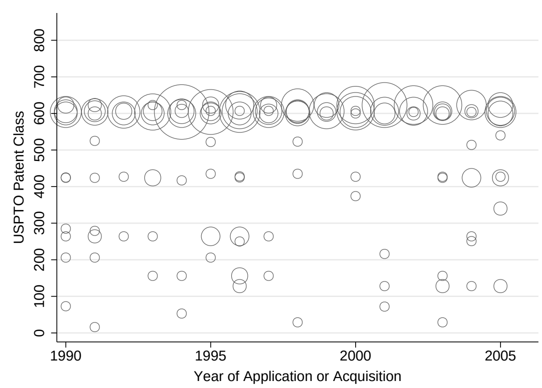

(b) Boston Scientific




(c) CR Bard




(d) Cordis




*Note*: The circle size represents the number of patents in each class-year. Based on our method and analysis in sections 3 and 4, the "flare lengths" of these firms' portfolios are: 1 (Becton Dickinson), 0 (Boston Scientific), 0 (CR Bard), and 0 (Cordis).

## References


Bar, Talia and Aija Leiponen (2012). "A measure of technological distance". In: *Economics Letters* 116(3), pp. 457–459.

Benner, Mary and Joel Waldfogel (2008). "Close to you? Bias and precision in patent-based measures of technological proximity". In: *Research Policy* 37 (9), pp. 1556–1567.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen (2013). "Identifying Technology Spillovers and Product Market Rivalry". In: *Econometrica* 81(4), pp. 1347–1393.

Cohen, Wesley M. (2010). "Fifty Years of Empirical Studies of Innovative Activity and Performance". In: *Handbook of the Economics of Innovation*. Ed. by Bronwyn H. Hall and Nathan Rosenberg. Vol. 1. Elsevier, pp. 129–213.

Edelsbrunner, Herbert, David Letscher, and Afra Zomorodian (2000). "Topological persistence and simplification". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE, pp. 454–463.

Griliches, Zvi (1990). "Patent Statistics as Economic Indicators: A Survey". In: *Journal of Economic Literature* 28(4), pp. 1661–1707.

Igami, Mitsuru and Jai Subrahmanyam (2019). "Patent Statistics as an Innovation Indicator? Evidence from the Hard Disk Drive Industry". In: *Japanese Economic Review* forthcoming.

Jaffe, Adam B. (1989). "Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers". In: *Research Policy* 18(2), pp. 87–97.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson (1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations". In: *Quarterly Journal of Economics* 108(3), pp. 577–598.

Lerner, Josh and Amit Seru (2017). "The Use and Misuse of Patent Data: Issues for Corporate Finance and Beyond". In: *NBER Working Paper 24053*.

Lum, Pek Y et al. (2013). "Extracting insights from the shape of complex data using topology". In: *Scientific reports* 3, p. 1236.

Nagaoka, Sadao, Kazuyuki Motohashi, and Akira Goto (2010). "Patent Statistics as an Innovation Indicator". In: *Handbook of the Economics of Innovation*. Ed. by Bronwyn H. Hall and Nathan Rosenberg. Vol. 2. Elsevier, pp. 1083–1127.

Nicolau, Monica, Arnold J Levine, and Gunnar Carlsson (2011). "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival". In: *Proceedings of the National Academy of Sciences* 108(17), pp. 7265–7270.

Ozcan, Yasin (2015). “Innovation and Acquisition: Two-Sided Matching in M&A Markets”.

Pakes, Ariel and Zvi Griliches (1984). “Patents and R&D at the Firm Level: A First Look”. In: *R&D, Patents and Productivity*. Ed. by Zvi Griliches. University of Chicago Press: Chicago, Illinois.

Rizvi, Abbas H et al. (2017). “Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development”. In: *Nature biotechnology* 35(6), p. 551.

Singh, Gurjeet, Facundo Mémoli, and Gunnar Carlsson (2007). “Topological methods for the analysis of high dimensional data sets and 3d object recognition.” In: *SPBG*, pp. 91–100.

Van Veen, Hendrik Jacob and Nathaniel Saul (2019). *KeplerMapper*. URL: `http://doi.org/10.5281/zenodo.1054444`.

Yao, Yuan et al. (2009). “Topological methods for exploring low-density states in biomolecular folding pathways”. In: *The Journal of chemical physics* 130(14), 04B614.