Nuts and Bolts: Analyzing the Development of Technological Breakthroughs through Clustering

Jayden Huang
Institute for Computing in Research

Abstract

Patent citations are a strong indicator of technological innovations and breakthroughs. This article employs unsupervised learning algorithms alongside gamma distributions to analyze the ways in which breakthrough patents evolve over time in regards to their forward citations. The results demonstrate that there are three distinct groups of patent breakthroughs that yield intriguing implications for future technological development.

1 Background

Technological development is the essence of human progress. Advancements in technology have yielded drastic improvements in human lifespan and quality of life. Only very recently has technological intellectual property been protected by patents. There is a wide variety of literature on the impact that patents have had on technological development (Takalo & Kanniainen 2000). Inventors are economically and intellectually motivated to file patents, and in doing so they cite other patents. Trajtenberg correctly identifies that patents vary widely in terms of their innovative contributions to their respective fields (Trajtenberg 1990). He also notes that there is a close association between citation-based patent indices and the real social value a given patent produces. As such, this would seem to suggest that highly-cited patents are also very likely breakthrough patents, those that can generate some form of social capital. Overall, the literature-base seems to conclude that patent citations have an association with the perceived or actual value of a given patent.

However, that is not to say the the number of forward citations can serve as the sole determining factor of a patents worth. It can be noted that care need to be taken in over duly interpreting patent citations as a metric for measuring the economic and social success of patents. This is especially important considering the shifting value of a patent, often alongside its citation rates (Marco 2007). The changing nature of the perceived value, and thus the number of forward citations it receives, can be an interesting factor to study. This is especially pertinent when it comes to discussing patents that have acquired an especially high number of forward citations.

2 Methodology

The data used for this study was acquired from the Google Cloud Patent database. Data was retrieved using a SQL-query. A sample of patents with filing dates between 1990 and 2000 was selected for closer examination. This date range was selected because it gave requisite time for the patents examined to mature to fruition relative to the time in present-day. The forward citations from this subset of the database were also retrieved and accounted for. The patents that exceeded a threshold of 1500 forward citations were selected for further examination. While this threshold was somewhat arbitrary, it was still designed with intent. The average patent rarely exceeded 50 forward citations based on preliminary testing, meaning that the threshold of 1500 was sufficient enough to ensure that only the most successful patents were selected for analysis.

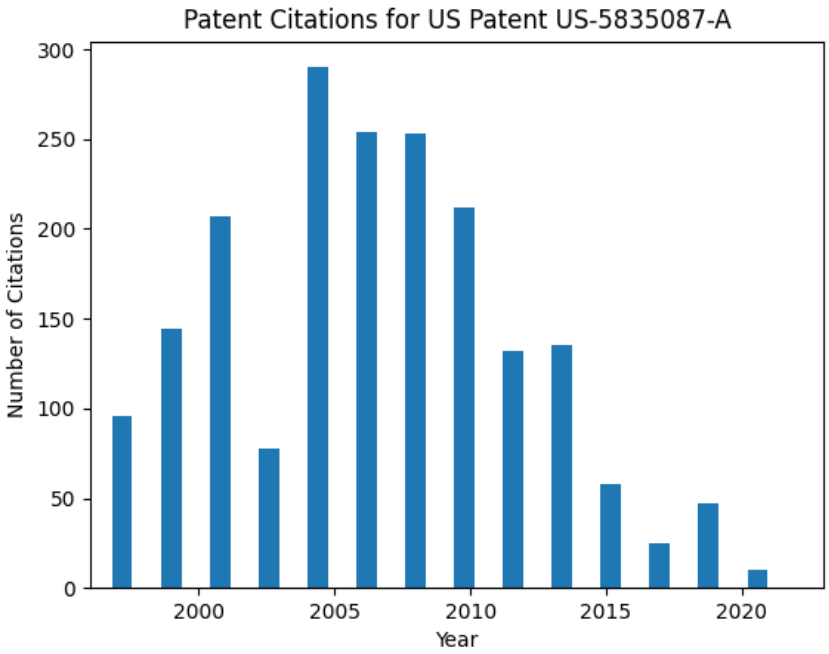

Figure 1: Histogram tracking patent forward citations over time

The analysis began with manual examination of histograms generated from individual patent data. This was followed by a K-means clustering algorithm. The clustering algorithm used was a part of the Scikit-Learn package, and this algorithm employed a sum-of-squares criterion shown here: $\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$

The K-means algorithim was selected because it is efficient at reducing inertia, especially given that the data contains few to no outliers and the relatively spherical distribution of the data (Pedregosa et al. 2011). In this case, $k = 3$ was used as a starting place for the algorithm based on loosely-identified groupings of patents discovered during manual examination. A multitude of features were used, including the time that it took for the given citation to reach 1500 citations. This was coupled with the following features: time to reach the first 10th percentile of citations, time to reach the first 90th percentile of citations, time it took to acquire the last 100 citations.

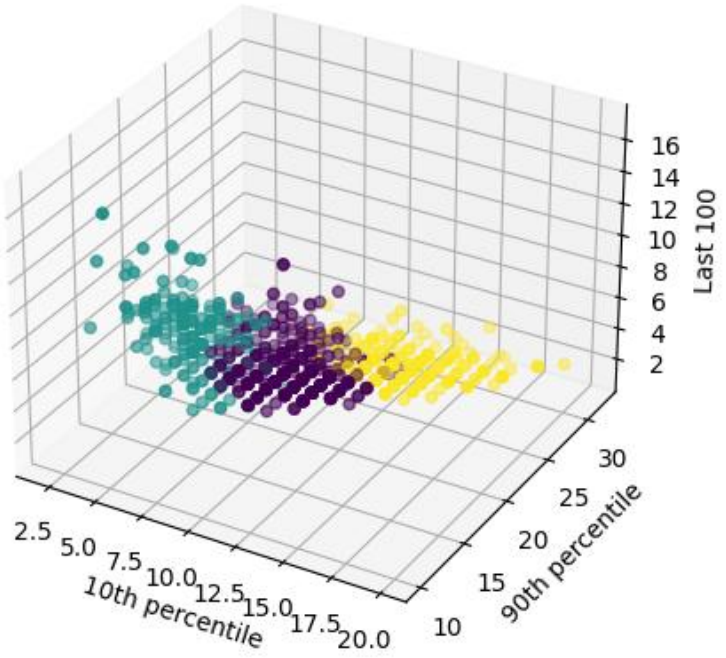

Figure 2: Three-dimensional Scatterplot Visualizing Clustering Algorithm Features and Results

As the figure above can demonstrate, the clustering algorithm has created recognizable clusters of data. However, in order to verify the veracity of the data beyond mere examination, further methods are necessary. To further verify the validity of the clustering algorithm, a gamma distribution was found for all patents within a given cluster. While the gamma distribution is typically used to model waiting times or disasters, it can be employed in this project to approximate probabilities for the peak of the citations for given patents. The gamma distribution has three main parameters: shape, scale, and threshold. Shape and scale parameters for the gamma distribution θ and k were determined using the conventional methods employing the variance and mean of individual patent data set. These parameters were taken from α and β where $\alpha = \frac{\mathrm{E}^2[X]}{\mathrm{Var}(x)}$ and $\beta = \frac{\mathrm{E}[X]}{\mathrm{Var}(x)}$.

The corresponding and resulting gamma distributions are then plotted based on the cluster they are a part of. The package used for calculating the gamma distribution probability density function is SciPy, and the fitting of the parameters was also done using this package (Virtanen et al. 2020). The threshold parameter was set based on the subset of data that was being analyzed. Each plot contains all of the gamma distributions of patents filed within a two-year time period that belong to one of the three clusters derived from the aforementioned clustering algorithm. In order to optimize the calculation in the gamma distribution, the data was normalized to a range between 0 and 1.

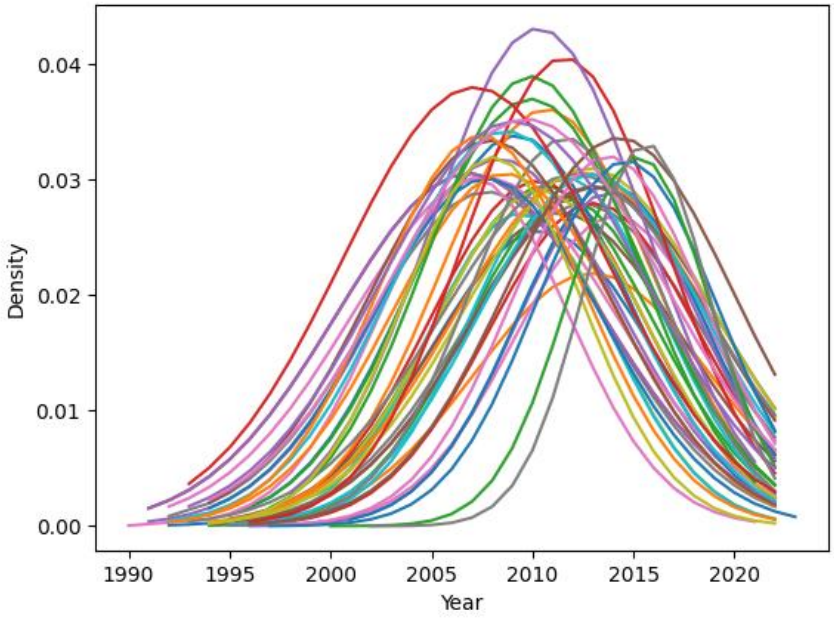

Figure 3: Visualization of Gamma Distributions for Patents within Cluster 1

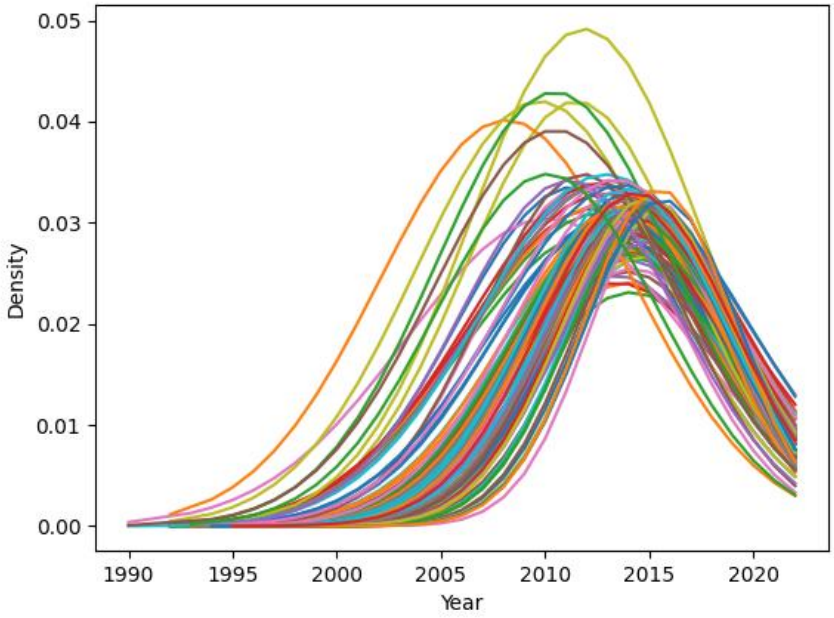

Figure 4: Visualization of Gamma Distributions for Patents within Cluster 2

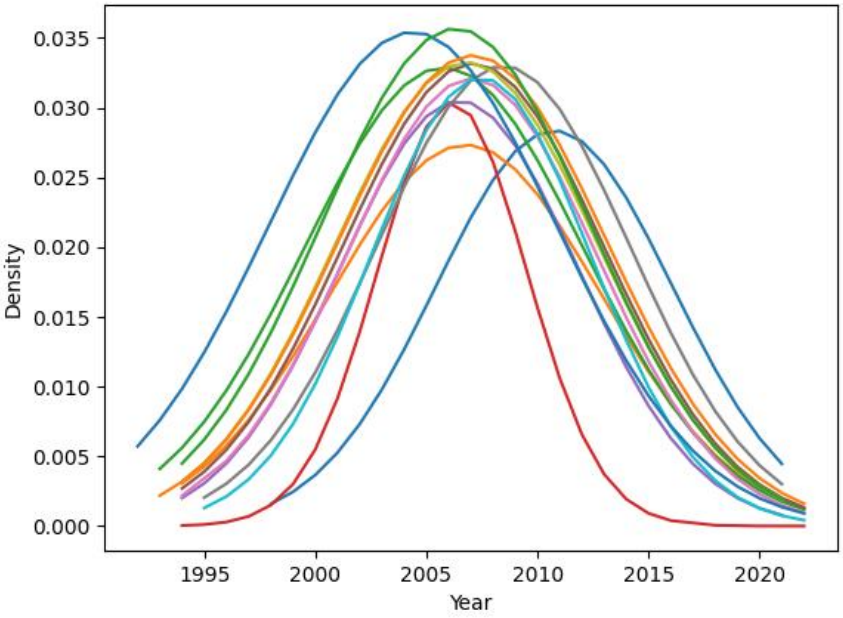

Figure 5: Visualization of Gamma Distributions for Patents within Cluster 3

A small disclaimer that the gamma distribution should not be interpreted as a perfect replication of the trends displayed in the histogram representing the same data. For example, the figure below charts the histogram for Apple's multi-touch patent alongside the corresponding gamma distribution. It is evident that the peaks of these two representations do not necessarily align, which is a necessary evil due to the nature of probability density functions.

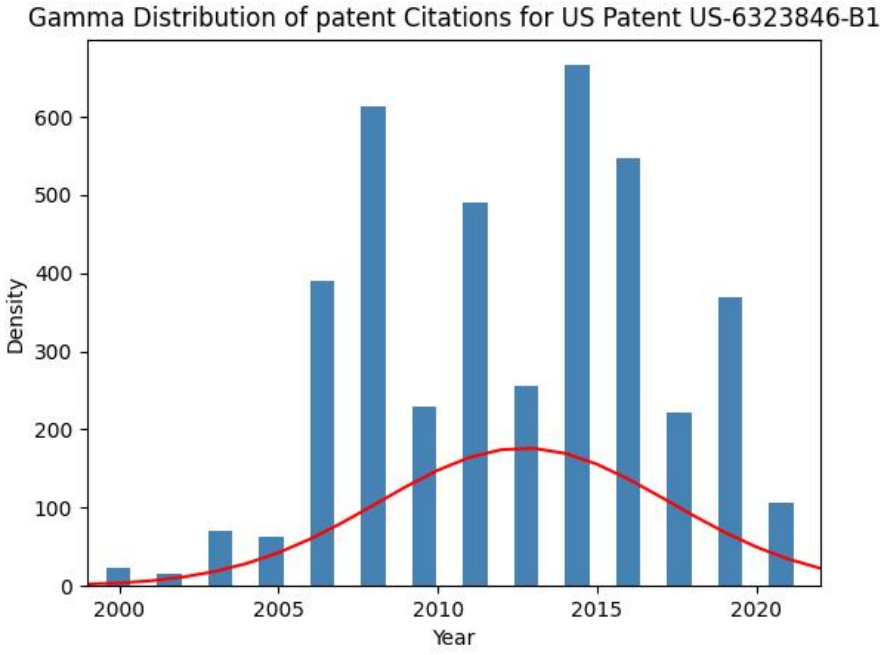

Figure 6: Combination Plot of Histogram and Gamma Distribution

3 Results

The integrity of the clustering algorithm appears to be held up considering the results of the plotted gamma distributions. The gamma distributions for different clusters are visually distinct from each other. What is interesting to note is the relative sizes of each of these clusters. When examining the breakthrough patents from 1992-1994, it is very easily apparent that there is a wide disparity in how many patents belong to each cluster. The second cluster has more patents than the other two clusters combined with a total of 112. This is easily seen upon visual examination where the density of the distribution is substantially greater than the graphs of the other two clusters. The third cluster is the least populated with only 13 patents. These results can be interpreted in a variety of ways, but the most interesting is the implications it has for a conclusion on how technologies develop. If the vast majority of patents belong to the second cluster, the same cluster that has the most left-skewed distribution, it appears that most technologies mature rather slowly.

The relative size of the third cluster corroborates this conclusion. The relative scarcity of patents that are able to achieve early success, in other worlds a more right-skewed distribution, is indicative of an overall sluggishness in technological development.

This result can have a variety of interpretations. Perhaps many of the patents lay in dormancy, waiting for a critical mass of other developments in their respective fields before achieving widespread success. It could be possible that many of these patents required pre-requisite technologies that did not exist at the time of filing in order to gain prominence or relevance.

Another possible interpretation is that fast breakthroughs like those in the third cluster may not exist at all. The massive burst in citations may not be indicative of an actual breakthrough technology, but rather of intra-company or intra-field citations that do not represent actual advancements. However, this theory is only a preliminary one that requires further research to validate or reject.

In terms of future research, this project shows great promise. The research that has been done in this field has been relatively minimal, leaving a vast expanse of topics to research. For example, the process of this study could be replicated but with the family classification code of each patent included as part of the clustering algorithm. This approach could unveil a potential correlation between the family classification of a patent and the way its citations develop.

4 Acknowledgements

I would like to thank Taylor Blair and Andrew J. Ouderkirk for their invaluable contributions to this project. This project would not be possible without their mentorship.

I would also like to acknowledge the Institute for Computing in Research for providing me with the opportunity to conduct this project and contribute to a rich history of scientific discovery.

References

Marco, Alan C. 2007. The dynamics of patent citations. *Economics Letters* 94(2). 290–296. doi:https://doi.org/10.1016/j.econlet.2006.08.014. https://www.sciencedirect.com/science/article/pii/S0165176506002837.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Takalo, Tuomas & Vesa Kanniainen. 2000. Do patents slow down technological progress?: Real options in research, patenting, and market introduction. *International Journal of Industrial Organization* 18(7). 1105–1127. doi:https://doi.org/10.1016/S0167-7187(98)00049-6. https://www.sciencedirect.com/science/article/pii/S0167718798000496.

Trajtenberg, Manuel. 1990. A penny for your quotes: Patent citations and the value of innovations. *RAND Journal of Economics* 21. 172–187. doi:10.2307/2555502.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt & SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17. 261–272. doi:10.1038/s41592-019-0686-2.