**2017 Patent Grant Analytics**

Looking for patterns to predict delays in patent prosecution based on bibliographic patent data.

Repository: <https://github.com/jayclare/2017_patent_grants_analytics>

**Introduction:**

In recent years, the United States Patent and Trademark Office (USPTO) has consistently received over 500,000 patent applications.  This number has risen steadily over the last 20 years and puts an added strain on the USPTO’s resources.  At present, there is a tremendous backlog of patents that are awaiting examination in the system. Our goal here is to explore whether there are any observable trends or correlations that are impacting the USPTO’s efficiency.

One measure of the USPTO’s efficiency is the mean prosecution time for a patent application - i.e., the amount of time it takes from the filing date until a patent is granted or “issued” on the application.  From a process standpoint, we want a system in which a patent will take more or less time to grant based solely on its merit (i.e., the novelty of its subject matter invention and how many similar inventions or references exist).  Put another way, we do not want a system where patents are taking more or less time based on other factors (e.g., how many drawings are included, how many claims are used to describe the invention, etc.). An ideal system is one that supports predictability and commonality for applicants in all industries.

For this analysis, we want to specifically explore any trends or correlations among characteristics of patents that are unrelated to the merits of the invention but nevertheless impact the mean prosecution time.

**Data Source**:

The data used for this analysis comes from the Bulk Data Storage System which is periodically updated and maintained by the USPTO.[[1]](#footnote-1)

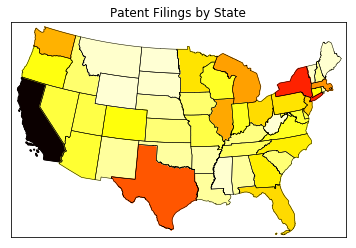
Each week, the USPTO publishes the bibliographic text information regarding all patents that were granted that week.  The data underlying this information is made available to the public in the form of 52 XML files uploaded to the USTPO’s bulk data storage site (see link to raw data above).  Many of these XML files are nearly 200 MB in size and contain over six million lines of data. In total, the XML data comprised more than 350,000 patent grants from 2017 that were analyzed in this project.

**Data Wrangling:**

In order to perform our exploratory data analysis in a jupyter notebook, an algorithm was needed to extract the data from each XML file and put it into a pandas DataFrame. In the data cleaning process, there were errors observed in each of the XML files in the form of extra and unnecessary “DOCTYPE” and “XML Version” tags.  Furthermore, each XML file lacked an appropriate root tag to wrap all of the data contained in the file. In order to clean and compile the data, the globbing module was used to develop a custom script that performed the following functions on all 52 XML files at once: (1) remove the extraneous tags, (2) wrap the data in an appropriate root tag and (3) extract all desired data and place into a single unified pandas DataFrame. The code for the data wrangling and cleaning can be found in the following repository: <https://github.com/jayclare/2017_patent_grants_analytics> specifically in the ‘/scripts and /jupyter\_notebooks’ folders.

**Geographic Overview:**

Before diving into the possible correlations and trends impacting the prosecution timeline, some time series and geographic plots were generated to visualize the spread and timing of the patent filings.  The first figure included here is a heat map that illustrates, with respect to the US applicants, where were the patents granted in 2017 filed:

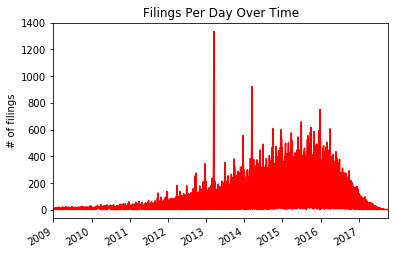


*Fig. 1 - heatmap of patent filings*

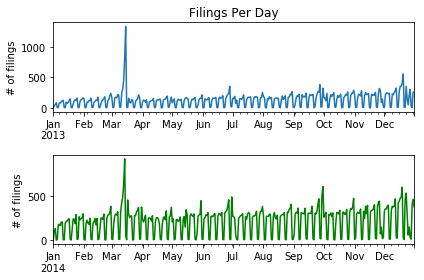
As Figure 1 illustrates, California is by far the most prolific state with respect to the patents granted in 2017.  Other states with a particularly high volume of filings include New York, Texas, Massachusetts and Washington.

**Timeseries Analysis:**

The next set of visualizations shows all of the filings over time for the patents issued in 2017 and then subset view of the patent filings from 2013 and 2014:



*Fig. 2 - all filings over time for 2017 grants*



*Fig. 3 - filings from 2013 & 2014 for 2017 grants*

Figure 3 was generated to further explore the pattern of spikes observed in Figure 2.  When we break out the filings from 2013 and 2014, we see a clear and repeated pattern of spikes in filing activity during early March, late June, late September and late December.  It is unclear what caused these clear repeated patterns. Perhaps these are observations of seasonal patterns, however, with respect to the March spike, it is such a dramatic increase that it seems likely that something else is going on.  More data and investigation is necessary to determine what is causing this repeated pattern of spikes.

**Correlations Investigated:**

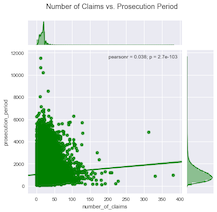
There were four essential questions that were investigated with respect to possible trends and correlations affecting prosecution time:

1. Is there a significant difference in prosecution period of utility patents versus non-utility patents (i.e., design and plant)?
2. Is there a significant difference in prosecution period between plant and design patents?
3. Is there a significant correlation between the number of claims and the prosecution period for utility patents?
4. Is there a significant correlation between the number of figures and the prosecution period for patents?

For question (1), a two sample hypothesis test was conducted to determine the statistical significance of the observed difference in mean prosecution time for utility patents versus non-utility patents (i.e., plant and design patents).  With our sample data, the mean prosecution time of the utility patents is approximately 1,067 days and the mean prosecution time for non-utility patents is approximately 584 days. Therefore, the observed difference in the means of the two samples is approximately 483 days.  Assuming the true difference in means between the respective populations is 0 (i.e., our *null hypothesis*, N0), the probability of observing a difference in mean prosecution period as extreme as ours (~483 days) is less than 0.000000001.  Therefore, we can reject N0 and conclude that there is a statistically significant difference in mean prosecution time of utility versus non-utility patents.  Based on our data, it seems utility patents take nearly two times as long to prosecute as non-utility patents.

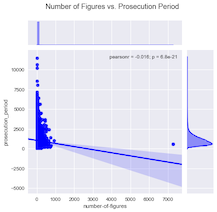
For question (2), a similar two sample hypothesis test was conducted to determine the statistical significance of the observed difference in mean prosecution time for plant patents versus design patents.  With our sample data, the mean prosecution time of the plant patents is approximately 648 days and the mean prosecution time for design patents is approximately 581 days. Therefore, the observed difference in the means of the two samples is approximately 66 days.  Assuming the true difference in means between the respective populations is 0 (i.e., our *null hypothesis*, N0), the probability of observing a difference in mean prosecution period as extreme as ours was computed to be 4.286e-44 (i.e., a very small p-value).  Therefore, we can reject N0 and conclude that there is a statistically significant difference between the population mean prosecution time for plants versus design patents.  Based on our sample data, it appears that plant patents take a bit longer to prosecute on average versus design patents.

Questions (3) and (4) dealt with testing the significance of observed correlation coefficients (i.e., Pearson’s r) for different variables versus the prosecution period.  In each case, a simple hypothesis test was used in which the null hypothesis (N0) assumed the true correlation coefficient was 0 (i.e., no correlation). For question (3), based on our sample data, a Pearson r value of 0.038 was calculated with a p-value of  2.7e-103. Therefore, we were able to reject our N0 and conclude that there is a statistically significant positive correlation between the number of claims and the prosecution period.  *However*, the observed Pearson r value here is *very low* at 0.038, indicating that the correlation is of very little *practical significance*.  The following visualization shows a joint plot (with histograms on each axis) and a regression line for question (3):



*Fig. 4 - number of claims vs. prosecution period joint plot*

For question (4), based on our sample data, a Pearson r value of -0.016 was calculated with a p-value of  6.8e-21. Therefore, we were able to reject our N0 and conclude that there is a statistically significant *negative* correlation between the number of figures and the prosecution period.  *However*, the observed Pearson r value here is *very low* at -0.016, indicating that the correlation is of very little *practical significance*.  The following visualization shows a joint plot (with histograms on each axis) and a regression line for question (4):



*Fig. 5 - number of figures vs. prosecution period joint plot*

**Linear Regression Analysis:**

Ordinary least squares multivariable linear regression was performed on the data set for utility patents. The experimental variables chosen for the regression analysis were the number of claims and the number of figures. The target variable was the prosecution period in days. A summary of the initial linear regression model is provided here:

|  |  |
| --- | --- |
| **R-Squared** | 0.0003 |
| **Adjusted R-Squared** | 0.0003 |
| **Coefficients:** |  |
| **Number of Claims** | -0.912 |
| **Number of Figures** | 3.0519 |

Additional models were trained on the following permutations: (a) number of claims as the sole experimental variable; (b) number of figures as the sole experimental variable; and (c) a repeat of the full model and single-variable models (a) and (b) with the a subset of the data. For (c), patents with prosecution times of greater than 5 years were disregarded. The results in the permutations were never greater than the original full model. In each case, the resulting R2 value was less than 0.0003, meaning that those models explained no more of the variance of the system than the original full model. Of note, the model generated from (b) above (focusing solely on the number of figures) did out-perform the model focusing on the number of claims. Therefore, based on these models, it appears that the number of claims in a patent application has *less* predictive value than the number of figures when it comes to prosecution time. Also, as noted above, the coefficient associated to the number of figures is *negative* which indicates that the prosecution period decreases slightly on average as the number of figures is increased. Again, given the extremely low Pearson coefficient reported above and the very weak R2 value of our model, the ultimate predictive value of these correlations is too weak to be useful.

To visualize the models performance across the quantiles of data, a quantile plot was used with the normalized residual values predicted by the model. Figure 6 below shows the quantile distribution of the sample data along with a 45° line which represents theoretical “normal distribution” of the data.



*Fig. 6 – quantile-quantile plot of data*

The distribution of residuals is a sign of the fit of a linear regression model. The more normally distributed the residuals, the better the fit. Based on Figure 6, it appears our model performs reasonably well (near the theoretical 45° between the -2.5 and 2.5 quantiles but is residuals are not normally outside of that range.

A histogram plot of the normalized residual values in Figure 7 below further confirms that the residuals are skewed significantly by a long right tail.



*Fig. 7 – histogram of normalized residuals*

**Conclusion:**

The review and analysis of 2017 patent grant bibliographic data provided some interesting insight into the temporal and geographic distribution of patent filings. From the heat map visualization, it is clear that certain states are far more prolific than others in terms of new filing and securing granted patents. These trends are consistent with what would be expected based on a high concentration of technology businesses in states like California, New York and Massachusetts.

With respect to filings over time, the repeated spikes in activity that occurs during the second week in March was an interesting and unexplained observation. Potential explanations may include change of law or policy, fiscal quarterly planning, or new developments in popular science. However, the fact that the spike in activity was specific to a single week and repeated over multiple years is worth further investigation. Other peaks are common throughout the year, but none so extreme was week 2 of March. Many possible explanations exist for increases in filing activity toward the end of the year. Further investigation would be needed to flesh out these hypotheses.

The correlations analysis provided a few interesting observations, but the ultimate magnitude of the correlation coefficients were simply too small to have significant predictive value. In a data set of this size, weak correlations are easier to find. One interesting observation was the *negative* correlation between the number of figures in a patent and the prosecution time – suggesting that including more figures in a patent may decrease the time it takes to be granted. This observation was repeated in the context of multivariate linear regression, which offered little insight into the predictive value of the number of claims in combination with the number of figures. In the case of our model, at best, our R2 was 0.0003, meaning the model explained very little of the variance. However, again, there was a negative coefficient observed for the number of figures versus prosecution time, implying that the weak correlation is a statistically significant observation.

Ultimately, the analysis shows that there are no signs of bibliographic information having a significant influence on the prosecution time period for a patent. There is more to investigate regarding temporal and geographic patterns in filing activity, however, it does not appear as though the design of the patent application (i.e., number of figures and number of claims) will significantly influence the amount of time it takes to secure the issued patent.

1. https://bulkdata.uspto.gov/ [↑](#footnote-ref-1)