Patent Co-Authorship Network Analysis

CS 224W Final Report

Stephanie Mallard, Thao Nguyen, Jeff Pyke
Computer Science Department
Stanford University

1 Introduction

This project aims to model the h-index of a firm as a function of several properties of the owning organization's patent co-authorship and citation networks. Studies have shown that the number of forward citations a patent receives strongly correlate with its market value [3], which is of great interest to potential investors. Unfortunately, future number of forward citations of particularly high-value patents can be difficult to predict, due to effects such as first mover advantage [7]. Therefore, instead of attempting to directly predict number of forward citations, we propose to predict h-index, a statistic derived from forward citations instead [4]. Additionally, while most previous studies examine only characteristics of patent citation networks as forecasters, we propose to use patent co-authorship networks. These can be generated with publicly available data from the US Patent and Trade Office (USPTO) database. Our hypothesis is that the underlying organization responsible for generating the patent is a strong predictor for future productivity. Because we wish to examine if the features of the underlying organization directly contribute to a high h-index, the interpretability of our model is a top priority.

2 Related Work

Several concepts, which have been studied in the following papers, are cornerstones to our project. First, that our success metric, or an organization's h-index, cannot be simply modeled as a function of one or a few simple variables. Secondly, that collaboration network structure can be used to predict innovative productivity or success later on. And finally, that collaboration networks exhibit unique properties which can be leveraged.

2.1 The First Mover Advantage in Scientific Publications

This paper discusses the first-mover advantage in scientific publication [7]. The first-mover advantage is the fact that the first papers in a field will receive citations at a higher rate than later papers published in the same topic. Derek de Solla Price wrote one of the first studies on patterns of citations, in 1965. This was the first time that the relation was modeled, under the Pareto distribution. Since then, several mathematical models have been presented to describe this pattern. The model that this paper uses is called the preferential attachment model, proposed by Barabasi and Albert in 1999. The model produces power-law distributions.

They begin by testing the preferential attachment model against real citation data, and discuss certain papers that buck the trend. They found the equations held up in most areas, but they found no first mover effect in other areas. An interesting finding in this paper was that the quality of papers is largely irrelevant compared

to the magnitude of citations for first mover advantage. It discusses the unusual value of certain high quality papers that attract more citations than the theory predicts. Since h-index is derived from the number of forward citations a company accumulates, this paper implies that predicting h-index will require more than one or two simple variables build a successful model.

2.2 The Structure of Scientific Collaboration Networks

This paper built collaboration networks from research publications in several scientific fields, with authors represented as nodes and co-authorship on a paper represented by edges [6]. After these networks were built and analyzed, they concluded that several scientific communities fit the definition of a small world as laid out by Watts and Strogatz [11] and that the networks are very highly clustered. Additionally, Newman observed that the structure of collaboration networks across different fields are statistically significantly different.

This paper provides strong evidence that the ”small world” phenomena occurs in co-authorship networks, at least within certain scientific fields. It also alerted us to the fact that collaboration networks can differ significantly from field to field, and that we will need to control for those potential differences.

2.3 Small Worlds and Regional Innovation

The author searched for evidence that ”small world” co-authorship network structure enhanced innovative productivity within certain geographic regions [1]. They were testing the hypothesis that certain geographic regions, such as Silicon Valley and Boston, are more innovative because of the multitude of connections formed simply due to geographic proximity. The main novelty of this paper is their development and study of the patent co-authorship network instead of the patent citation network. While they failed to find evidence that small worlds enhanced productivity level, they did find that other network structural properties such as shorter path lengths and larger connected component did significantly affect innovative productivity.

2.4 Patent Co-citation Networks of Fortune 500 Companies

The paper explores technological progress in several main technology groups by analyzing patent co-citation information of Fortune 500 companies indexed in Derwent Innovations Index database [10]. This is based on the assumption that Fortune 500 companies are representative of their fields and current technological frontiers, while patents speak volumes about the intensity of innovative activities going on in each company. Methods such as bibliometrics, social network analysis and information visualization are employed to study the co-citation networks.

By constructing and computing network statistics such as betweenness centrality and Jaccard coefficient matrix, the paper provides a perspective on which companies are the pivotal point in the network, as well as macro-level interactions among major industry fields. It divides the 30 years of data into three equal time spans, so as to compare the decade-by-decade evolution of the technology landscape. Another contribution of the paper is its search strategy to process the patent data, accounting for cases of merger and joint venture. This helps to reduce problems of incomplete and unreliable data.

2.5 Exploring technology diffusion and classification of business methods: Using the patent citation network

This work explores patent citations as indicators of technology forecasting. The frequency of citation of a patent indicates that it is diffused, i.e. widely applied and useful. The paper explores measuring relationships between patents, and classifying the patents into groups of technology diffusion. [9]

The paper defines two types of relationships. Lineal relationships are those of a direct citation, or an indirect citation. An indirect citation is like a grandparent/grandchild relationship. In contrast, a collateral relationship is one

that relates siblings, uncles, etc. This was not a part of the study. The paper created a patent citation matrix and calculated an indicator of the lineal linkage strength. Then it performed hierarchical cluster analysis. The study concluded by analyzing the businesses related to the groups of patents it found.

3 Dataset

The United States Patent and Trade Office (USPTO) has made all data related to patents from 1976 - 2017 publicly available in a series of tab separated files on patentsview.org [8]. All data necessary to generate the networks of interest (patent co-authorship, owning firms, date of patent issue, patent citations) was downloaded and converted to the networks of interest. Specifically, we generated and analyzed the co-authorship networks of several innovative firms from the years of 1990 - 2000, and counted the number of forward citations generated from 2000 - 2010.

The co-authorship networks were generated in a fairly straightforward way - nodes indicated inventors and edges indicate a co-authorship on a patent submitted between the years of 1990 and 2000. Separate networks were generated for each firm chosen, and the firms were filtered on the criteria that they had produced at least 100 patents in the 1990 - 2000 time frame. These networks were all simple and undirected graphs.

In addition to this, we generated citation networks from the same tab separated files. These were also generated separately for each company, and only included citations made between the years of 1990 and 2000. For these citation networks, each node represented a patent and each edge represented a citation to another patent. Citations made to patents outside of the company became dead ends in the network, but the node for the external patent was included. So the citation networks contain all of the patents from a company, and all of the patents that they cite.

4 Approach

4.1 Features

We chose several key network features for our prediction algorithm, including: node count, edge count, clustering coefficient, number of strongly connected components (SCCs), maximum SCC size, modularity, average patents per inventor (or number of internal citations for the citation networks), total number of patents, and the year of the oldest patent. We discarded data samples with missing fields and only used the networks that had all of those 9 features. While most of these features are self-explanatory, three bear further examination: oldest patent, clustering coefficient, and modularity.

4.1.1 Oldest Patent

The oldest patent feature was our attempt to factor in the first mover advantage into our predictive model, and its value p was calculated as follows:

$$p = 2000 - f \tag{1}$$

where f is defined as the year (as registered in USPTO database) of the oldest patent belonging to the company represented by the network.

4.1.2 Clustering Coefficient

In choosing features for our predictor, we hypothesized that how "tightly knit" the co-authorship networks were would be a significant feature. Clustering coefficient captures this property, and is defined for any given node i with neighborhood N_i and k_i neighbors, within a graph with edge set E as:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk \in E}\}|}{k_i(k_i - 1)} \tag{2}$$

For the purposes of this project, we chose to use the average clustering coefficient, as defined in the 1998 Watts and Strogatz paper [11]. This is due to the fact that we need a fixed number of features for each network, and therefore need a summary statistic.

4.1.3 Modularity

We used the Girvan-Newman community detection algorithm [2], based on betweenness centrality in each graph to compute modularity Q. Modularity for a network that has been partitioned into communities S, with individual communities $s \in S$, adjacency matrix A, and number of neighbors for a given node i k_i have been detected is defined as follows:

$$Q = \frac{1}{2m} \sum_{s \in S} \sum_{i \in s} \sum_{j \in s} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \qquad (3)$$

4.2 Forward Citations and H-Index

We computed total forward citation counts (from 2000 to 2010) as well as the h-indices for the 144 organizations selected. The distribution of future citation count ranges from 0 to 11610, with mean 960.7 and standard deviation 1351, while that of h-index ranges from 0 to 38, with mean 13.05 and standard deviation 6.59.

Note that our definition of h-index differs slightly from standard definition. In most of the literature, h-index is calculated on a person by person basis, and if calculated for an organization, it is normalized by the number of people who contributed. However, in our definition, h-index is calculated for each company without the number of contributers as a normalization factor. This is because we were interested in investigating, as a feature in our model, the effect of the number of inventors on the future h-index, which would have been difficult or impossible with the normalization factor.

The wide range and standard deviation of the forward citation count was the main reason why we decided to predict h-index instead. A variable with that wide of a distribution could be extremely challenging for a simple linear regressor to predict with any kind of accuracy. However, given that one of the main goals of this project was to determine which network features most contributes to a productive innovation network, we did not wish to use a complicated model and sacrifice interpretability in return.

4.3 Prediction

We used the aforementioned features as inputs into a linear regression model to predict h-index. We defined forward citation to be the number of citations that happened between 2000 and 2010 only. And an organization is said to have index of h if they have published h patents, each of which has been cited at least h times. Again, the same time period restrictions applied to both the cited patents belonging to the organizations of interest and the ones citing them.

4.3.1 Feature Selection

To preserve interpretability, we experimented with 2 feature selection techniques that picked/ removed features from the original pool, instead of using PCA (Principle Component Analysis) which would transform the original features:

- F-test for significance: compute F-regression for each individual feature and choose the top k highest-scoring features. In this case, F-regression is a univariate linear regression test that for each feature, compares an intercept-only regression model with a 1-feature model and sees the extent to which the addition of this feature increases the model's ability to fit the given data.
  We varied k from 4 (keeping half of the features) to 8 (keeping all but 1) and found that using 6 features gave the best result, even better than using the full set of all features.
- Recursive Feature Elimination (RFE): select k features by recursively removing features one by one. Given an intermediate set of features at each iteration, this method used an estimator (Huber Regressor) to fit a linear regression model to the given data, obtaining weights/ coefficients for each feature through cross-validation. Then the least important feature is pruned from the current set, leaving behind a smaller set of features each time.

Again, we varied k from 4 to 8 and found that using 7 features gave the best result.

The 2 feature selection methods yielded different sets of optimal features. We evaluated the performance of the selected feature sets using 8-fold cross-validation, and found that RFE produced better results. In the final predictive model, we removed node count and edge count, and only used the remaining 7 features.

To evaluate the effectiveness of our feature selection method, we tested for multicollinearity (feature redudancy). This was done by computing the correlation among the remaining features and finding the eigenvalues of the resulting covariance matrix. We found that none of the eigenvalues was zero, meaning that there is no exactly linear relationship between any 2 of the features. Moreover, the eigenvalues were also sufficiently above zero. This means that there is no strong correlation among the predictor variables used, or in other words, it's not possible to combine them in linear or simple ways but still preserve the amount of information that they individually capture.

4.3.2 Huber Regressor

We used a linear regression model with Huber loss [5] to be more robust to outliers, which are likely to be present in our data. Huber loss is a hybrid of quadratic loss (for relatively small errors) and linear loss (for relatively large ones), defined as follows:

$$X(m,n) = \begin{cases} \frac{r^2}{2} & \text{if } \left|\frac{r}{\sigma}\right| < \epsilon \\ \epsilon|r| - \frac{1}{2}\epsilon^2 & \text{if } \left|\frac{r}{\sigma}\right| \geq \epsilon \end{cases} \quad (4)$$

where $\epsilon = 1.35$ as commonly used and defined by Huber himself, $r = y - \hat{y}$ (the residual from our predictions) and σ is a parameter to make sure that if y was scaled by a certain factor, we would not need to rescale ϵ to achieve the same robustness. This means that the Huber loss is not heavily influenced by the outliers, but also not totally ignoring their effect.

Due to the limited data from the networks that we constructed, instead of using a held-out validation set, we used cross-validation (8 folds) to evaluate the performance of our model.

5 Results and Findings

After some experimentation, we decided to fit three regressors: one with features generated from the co-authorship network, one with features generated from the citation network, and one with features drawn from both networks. The results are enumerated in Table 1.

	Co-Authorship	Citation	Combined
Features Retained	7	6	11
Mean Absolute Error	3.75	3.54	3.50

TABLE 1
Results from predicting with different features

As shown, the citation network features did better in the regression task compared to the co-authorship network, but the combining features from both performed the best. Additionally, the co-authorship network seemed to benefit from retaining more features than the citation network after feature selection.

5.1 Co-Authorship Graph Structure

After visualizing the co-authorship graph structures, we found them to be incredibly diverse. Most were characterized by SCCs of various sizes, representing teams that had worked on patents together. Some only had isolated cliques with no bridges between teams, while others were very well connected and had minimal disconnected components. See Figure 1 for an example.

5.2 Citation Graph Structure

We found the citation graph structures vary greatly from company to company. The co-authorship graph and citation graph for a given company could also be very different from each other. For example, some companies had only a few authors produce patents, but cited many other patents. Some companies had highly connected co-authorship networks, but sparse citation networks.

5.3 Co-Authorship Feature Analysis

After fitting to our linear regression model, we examined the weights assigned to each feature to try to draw conclusions about their contribution to organizational success.

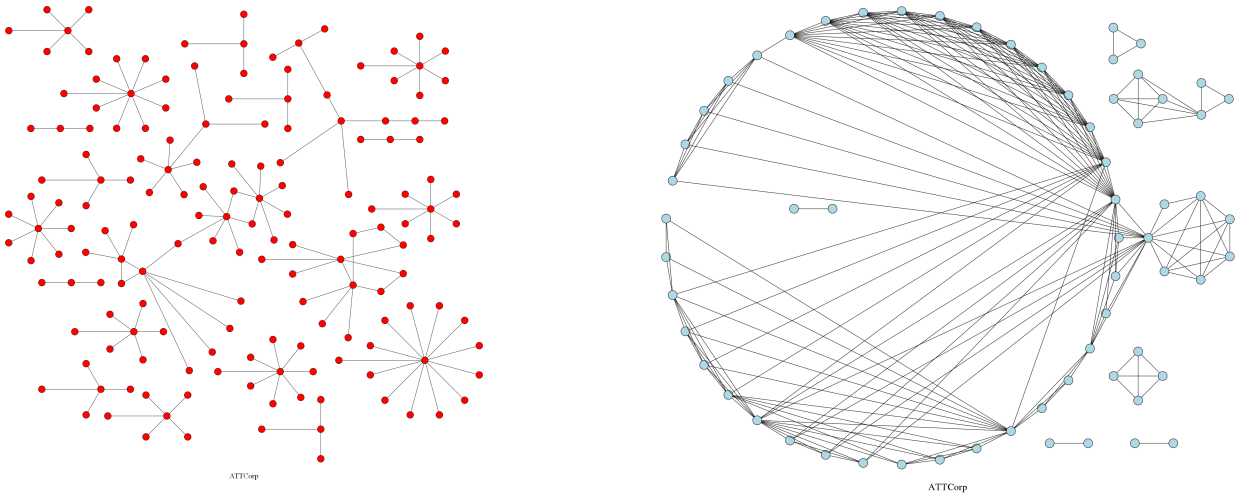

Fig. 1. Visualizations for the citation network (left) and co-authorship network (right) of AT&T Corporation. Note the very different structures of the two, despite being the same company.

5.3.1 *High Weighted Features*

Both the maximum SCC proportion and modularity features were by far the highest weighted after fitting the linear regression model. They were also positive, indicating correlation between these features and a high h-index.

The high weight given to the maximum SCC proportion is understandable: a company with a well connected innovation network seems more likely to produce quality patents. However, since the modularity is also highly weighted, this seems to indicate that this is only true to a point. In the context of our co-authorship network, this could mean that companies that produce patents with many authors, or companies whose innovators collaborate enough to form tightly knit communities are more productive. Thus, both factors make intuitive sense - patents with many authors are more likely to be higher quality, to a point, and similar to the conclusions drawn from the weight of the maximum SCC proportion feature, innovators that are highly collaborative are also more likely to produce quality patents.

5.3.2 *Low Weighted Features*

Surprisingly, the oldest patent feature had a moderately low weight, indicating it had relatively little importance in determining h-index. This was very interesting, given that we hypothesized that the first mover advantage would strongly influence the number of forward citations a company gained, and thus its h-index.

However, this low weight might also have been due to the fact that the oldest patent feature was almost constant across the 144 companies examined - only a few companies met our criteria that hadn't been producing patents for the entirety of the ten year timespan. Therefore, we decided not to pursue further compensation in our model for first mover advantage, like adjustments to number of forward citations.

Another interesting thing we noted was the negative (if low) weight of the average of patents per inventor feature. This implies an inverse relationship with high h-index, in turn meaning that companies with a few inventors that produce a large number of patents are more likely to produce low quality patents.

5.4 Citation Feature Analysis

The following is an analysis of the most contributing features from the citation network in the regression task:

5.4.1 *Modularity*

The modularity feature had the largest weight in the citation linear regression model. The

weight was positive, indicating a positive correlation with h-index. High modularity indicates more of a community structure, meaning that the citation network can be subdivided into many strongly connected pieces. Therefore, a modular network could indicate that the company has filed multiple patents in several different research areas, with many citations within that research area.

5.4.2 Maximum SCC Proportion

The maximum SCC proportion feature had the second highest weight. The weight was positive, at around half the magnitude of modularity. This indicates that having some level of connection in the citations of patents within an organization correlates with higher citation levels later on.

5.4.3 Internal Citation Proportion

Internal Citation Proportion was the third largest weight, and also the largest negative weight. This measure what proportion of patents in the network belongs to the company, out of the total number of patents in the citation network. This indicates that having more internal citations correlates negatively with a higher h-index. This is interesting because it indicates that authors who cite outside their company more often are more likely to receive a lot of citations themselves.

5.5 Combined Feature Analysis

We combined the features from both graphs and performed linear regression to try to predict the h-index. The maximum SCC proportion was most important from both graphs, followed by co-authorship clustering coefficient and Internal Citation Proportion.

5.5.1 Feature selection

The best performance we found with recursive feature elimination was with 11 features. The features that were removed were the edge count of the co-authorship network, the node count of the co-authorship network, the number of patents, the edge count of the citation network, and the number of SCCs in the citation network.

5.5.2 Maximum SCC proportion

The most important feature was the maximum SCC proportion from the co-authorship network. The second most important feature was the maximum SCC proportion from the citation network. This indicates that the community structures of both networks are important. Authors being connected is beneficial, as well as citations citing similar patents.

5.5.3 Co-authorship clustering coefficient

This was the third most important feature. This indicates that companies with patent authors who tend to work together are more highly cited than those who have independent groups of authors.

5.5.4 Internal Citation Proportion

This was the fourth most important feature. It was a negative predictor of h-index. This makes sense, because it means that networks that cite other external patents often will tend to have a higher h-index. This indicates having high awareness of the field through a large amount of research on the work that other people are doing.

6 Conclusion

The standard deviation of the ground truth h-index, which we were predicting, was 6.59. The baseline of using the one best feature - the clustering coefficient of the citation network - to predict gave a mean absolute error of 4.44. Our best result for the linear regression was a mean absolute error of 3.50. This was achieved by combining both the co-authorship and citation networks, using average patents per inventor, clustering coefficient, maximum strongly connected component (SCC) proportion, modularity, number of SCCs from the former, as well as internal citation proportion, clustering coefficient, maximum SCC proportion, modularity, and node count (total number of patents created and cited) from the latter.

Overall, while the citation network features were more successful at predicting future productivity, the co-authorship network features did provide extra information that refined

the predictions made. We concluded that co-authorship networks offered additional useful information in determining the future productivity of a company, as measured by h-index.

7 Future Work

We can continue experimenting with other ways to study the influence of the first-mover advantage on future productivity, such as by increasing the time span to 20 years and applying a discount factor to the number of future citations, depending on the average age of patents from the organization or how long the oldest patent had been published in that span, thus obtaining a completely different set of target labels for the prediction task.

In this project, one of our main goals is coming up with a set of identifiable, interpretable factors that can predict future performance (measured in h-index) sufficiently well. However, if accuracy is a higher priority, we could in our future work use more complex features, or introduce nonlinearities to make our predictive model more expressive and able to capture higher-order interactions among the features.

References

[1] Lee Fleming, Charles King III, and Adam I. Juda. Small worlds and regional innovation. *Organization Science*, 18:938–954, 2007.

[2] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[3] Bronwyn H. Hall, Adam Jaffe, and Manuel Trajtenberg. Market value and patent citations. *Rand Journal of Economics*, 36(1), 2005.

[4] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

[5] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[6] M. E. J. Newman. The structure of scientific collaboration networks. In *Proceedings of National Academy of Sciences of the United States of America*, pages 404 – 409, Santa Fe, NM, 2000. Santa Fe Institute.

[7] M. E. J. Newman. The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86, 2008.

[8] PatentsView. patentsview.org, 2017.

[9] Shu-Min Chang Shann-Bin Chang, Kuei-Kuei Lai. Exploring technology diffusion and classification of business methods: Using the patent citation network. *Technological Forecasting & Social Change*, 76:107–117, 2009.

[10] Xianwen Wang, Xi Zhang, and Shenmeng Xu. Patent co-citation networks of fortune 500 companies. *Scientometrics*, 88(3):761–770, 2011.

[11] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440 – 442, 1998.

Contributions

- Stephanie Mallard: dataset location, generation of co-authorship networks, original linear regression, graph visualizations, analysis of features for co-authorship network model
- Thao Nguyen: generation of forward citation count and h-index, feature selection, fine-tuning linear regression, evaluation of different steps in the machine learning pipeline (error analysis, feature importance, etc.)
- Jeff Pyke: generation of features, generation of citation networks, adaptation of linear regression to citation & combined networks incl. feature elimination, analysis of features for citation & combined networks