Measuring Growth Potential of Communities in the Patent Citation Network

Roger Chen, Jichan Park, Alexander Smith

December 10, 2014

1 Introduction

The United States Patent and Trademark Office (USPTO) requires inventors who submit patents to include citations to preexisting patents containing similar solutions to the technical problem the patent is solving. This has resulted in a massive network of connected technological ideas spanning decades, and serves as an invaluable proxy to measure and model technological innovation [3]. Such investigations have practical value, offering insights, for example, into the relative importance of an individual patent, how certain factors such as patent age and inventor predict patent importance, and the emergence and evolution of new technological fields. These insights can in turn be used to determine the economic value of patent portfolios and drive research and development decisions by firms.

Modeling the patent citation network as a graph allows us to uncover insights on the processes of technological innovation using graph properties alone. Node degree, for example, can be used in evaluating a patent's relative importance and predicting its success in terms of the number of citations it receives; the out degree. In our preliminary analysis, we hypothesized that some of the predictors of a patent's importance might include both in- and out- degrees, age, how similar a patent is to other patents, and some measure of how "original" a patent is based on its citations to communities not containing or related to that patent. Finally, community detection and tracking gives us a precise method of determining technological fields and their evolution.

In this paper we will focus on detecting communities within the patent network and tracking their growth over time, then evaluating which network properties predict long-term growth of the communities. The paper is divided into four sections: a review of existing literature on the subjects of community detection and tracking and prior investigations into citation networks; our procedure, including data definition and cleansing and the models, algorithms, and techniques used; our results; and finally, a discussion and suggestions for further research.

2 Prior Work

Structure and Evolution of the Patent Citation Network

Valverde, et al. analyzed the USPTO citation network to identify the factors that affect its evolution [3]. They uncovered a power-law degree distribution and strong negative correlation between the average clustering coefficient and node degree, suggesting a hierarchical organization of the network. They then created a model for edge creation that emphasized patent age and number of existing citations as factors that predict whether a patent receives a citation; a newer patent with more citations is more likely to receive a new citation, and thus has more utility. They fitted their model with functions of preferential attachment and a Weibull distribution, and discovered that the real network closely followed this model.

The findings in this paper gave us insights into the factors that contribute to the growth of a community. We disagree with the authors defining utility as the number of direct citations for a patent. We felt that this was too simplistic and did not accurately capture a patent's value. For example, the number of patents that a patent cites could be an indicator of how impactful a patent is; a patent that makes no citations, or a patent that cites patents from a diverse set of fields may be truly originally and open up an idea space for new patents. Refining the measure of utility will enable us to more accurately predict growth of patent networks.

Community Structure in Networks

There are several algorithms of detecting tightly-knit groups within a network, each varying in effectiveness and efficiency. In a groundbreaking paper [5], Newman introduced a fast community detection algorithm by optimizing modularity, a measure that quantifies the strength of a particular division of a network by subtracting the found fraction of edges within the division and the expected fraction of edges when placed at random.

$$Q = \frac{1}{4m} \sum_{ij} (A_{ij} - \frac{k_j j_i}{2m})(s_j s_i + 1) \tag{1}$$

Maximization of modularity corresponds closely to particularly satisfactory splits, and helps us to determine if the communities that we have defined are optimal. The paper proposes construction of a modularity matrix and running eigenvalue decomposition of the matrix, effectively dividing into subgraphs, until the matrix has no positive eigenvalues. Application of the algorithm to both social and biological real-world networks indicated that it yielded intuitively reasonable divisions when compared to prior knowledge on the community structure of networks, and was supported by quantitatively better divisions as measured by modularity.

Two addition algorithms, one building upon the modularity maximization strategy, offer significant performance improvements. Clauset, Newman, and Moore, in their "CNM" algorithm, improved the running time of modularity maximization by storing a matrix for the changes in modularity for each division merger step, rather than recomputing the changes at each step. [6] Another algorithm, BigCLAM, finds overlapping communities in a hierarchy. [7]

Tracking and Quantifying Growth of Evolving Communities in Dynamic Social Networks

Tracking dynamic communities over time is a difficult problem. This is not just because techniques appear to have limitations in performance, but also because communities themselves may be hard to define as they change membership and overlap with related communities. A common approach to tracking is to compute communities from static detection algorithms at particular points in time, or "snapshots", and relate these communities across time intervals using some sort of similarity measure. The static snapshot approach is advantageous as it decouples the tracking process from the chosen static community detection algorithm. Greene et al. describe such an approach by employing the Jaccard similarity measure to relate pairs of static communities across time. Consider a static community C_1 in time 1, and community C_2 in time 2. C_1 and C_2 become part of the same dynamic community if the following equation holds for some threshold θ:

$$\theta > Jsim(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \tag{2}$$

Each pair of static communities across a time unit interval is considered, resulting in a category of possible community events: merge, split, birth, and death. [8]

In a class paper presented in 2013, Britz et al. recognized limitations with the Jaccard measure; namely, fast-growing communities will not meet the threshold. They introduced an alternative similarity measure, again for some threshold θ:

$$\theta > Nsim(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1|} \tag{3}$$

In addition, they modeled each connection between two communities across a unit of time as an edge in a directed acyclic graph (DAG) in order to facilitate the detection of merging, splitting, birth, and death events. Merge, split, birth and death events can be counted by in degree, out degree, zero in degree, and zero out degree of the DAG respectively. They tested their approach on the AngelList network using CNM and BigCLAM as snapshot algorithms, and by plotting the quantity of events to different values of θ. [9]

3 Methodology

For each selected subcategory and year between 1975 and 1999, we construct a network that only contains patents belonging to the chosen subcategory and issued at or before the selected year, and apply a static clustering algorithm on each network in order to identify communities within each subcategory at each time point.

We then take these networks in aggregate and apply a dynamic tracking algorithm in order to link discovered communities between successive years. A list of hypothetical predictor metrics that could potentially be indicative of community growth rate are defined, and the values for these metrics, as well as the growth rates of networks, are calculated from the networks of each subcategory at each year.

Finally, we check for correlations between each metric and growth rate, and isolate the metrics that exhibited significant correlation as defined by R^2 value. We train a regression model on these metrics, and evaluate the predictive power of the model with testing data from three different subcategories.

4 Data

4.1 Initial Dataset

We used the pairwise citation set for all utility patents granted by the USPTO between 1975 and 1999, provided by the National Bureau of Economic Research (NBER) [1]. As described on the site, there are 3,923,922 patents with 16,522,438 total citations. In addition, a separate data set contains metadata including date, subcategory, class, and number of claims for each patent granted between 1963 and 1999. The metadata set will be associated with the pairwise citation set when appropriate in order to provide supporting information to help create the data set.

4.2 Data Cleansing

We have discarded 1,018,903 patent nodes in the citation network that are missing corresponding metadata since this additional information is required for our approach. Furthermore, we have decided to exclude 636,444 patents that were invented before 1975. This is because we wanted to increase computational efficiency and eliminate patents that are too old to be predictive of future growth. We believe that patents can be influential for a maximum of two decades; any patents that are older might create noise and thus are better removed. Finally, we removed 12,066 nodes that have zero in-degree and zero out-degree. Since clustering is based on network connections, patents that do not have any citations cannot be placed in any cluster, and to leave them in our data set could possibly impact the runtime of clustering algorithms. After nodes were discarded, the network is left with a total of 2,119,421 nodes and 10,557,536 edges with a clustering coefficient of 0.096 and an average degree of 5. The maximum in-degree of the network is 776, and the maximum out-degree of the network is 686.

4.3 USTPO Classification System and Level of Granularity

Initially, we planned to analyze the entire patent citation network. However, we found this to be problematic because: (1) The entire patent citation network is too large for community detection algorithm to complete in a reasonable amount of time (CNM takes 24+ hours for a network with 2M+ nodes) (2) We wanted to prevent detecting communities that are too diverse and incoherent. When community detection is applied to the entire network, communities consisting of patents from vastly different fields may arise, undermining clustering accuracy. At the same time, we did not want to be too rigid and wanted to still allow for some flexibility so patents of different classes could be placed in the same functional cluster. Had we chosen to force communities to contain patents of the same class, 45.6% of total citations would have been left unused.

The USPTO classification system consists of three levels: categories, subcategories, and classes. Each patent is assigned to one of six categories, which are divided into 37 subcategories, which are further divided into 417 patent classes. For each classification level, we calculated its average size. We also calculated fraction of citations that are between different categories, subcategories, and classes. (i.e. citations that would go unused when we force communities to be of the same category/subcategory/class):

	Network	Category	Subcategory	Class
Size	2,119,421	353,236	57,281	5,082
Percentage of unused citations	0%	23.2%	34.5%	45.6%

To reach a balance between keeping size down and preserving citations across different fields, we decided to detect communities at the subcategory level. We arbitrarily chose three subcategories, organic compounds (14), computer hardware/software (22), and transportation (55), for the training set and chose three additional subcategories to be used as the test set for our growth rate regression models: drugs (31), electrical devices (41), and agriculture (61).

5 Community Detection and Dynamic Tracking

5.1 Static Community Detection

We chose the Clauset-Moore-Newman (CNM) algorithm to cluster communities within a given subcategory in a given year. CNM optimizes modularity to demarcate boundaries among communities [6]. The following summary statistics were obtained from the output:

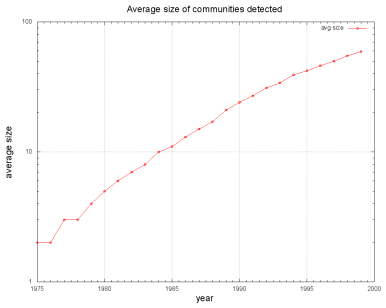

Figure 1: Average size of communities found over the years

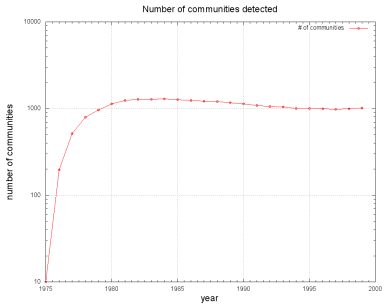

Figure 2: Number of communities found over the years

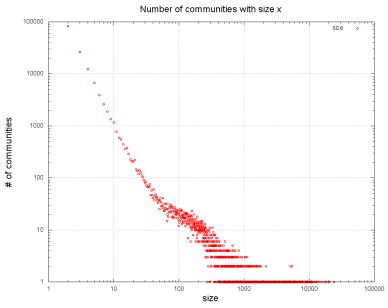

Figure 3: Number of communities of size x

We note that average size of communities within subcategories increases expontentially over the years (Figure 1). Although earlier years (1975-1980) suffer from lack of citations due to data truncation as described earlier, we find that the number of communities detected by CNM remains fairly consistent across the years (Figure 2). In addition, we observe that size distribution of communities follows a power law with $x_{min} = 176$ and $\alpha = 1.95$ (Figure 3). These suggest that as time evolves, new patents are added to existing communities, preferentially to those that are already sizeable.

In addition, to give a rough evaluation of how accurate the clustering is, we chose to evaluate homogenetiy of clusters generated by CNM. We leveraged USPTO patent class classification (e.g. 101 for printing, 353 for image projectors, 512 for perfume compositions) from the patent metadata. Although we do expect communities to contain patents of multiple classes, we believe that the classes of patents within the community will be largely similar if clustering was accurate.

We quantified community heterogeneity using Shannon entropy metric, which is defined as:

$$H(x) = \sum_i P(x_i)I(x_i) = -\sum_i P(x_i)log_b P(x_i) \tag{4}$$

where x_i is the patent class for node i and $P(x)$ is the probability of a patent belonging to class x in a given cluster (i.e. frequency of a patent). Entropy, which ranges from 0 to 1, measures the unpredictability of information content; the more diverse the classes are and the more infrequently each class appears in a cluster, the higher the entropy is and the more heterogeneous the community is.

We see that majority of clusters are fairly homogenous, with < 0.4 entropy (Figure 4) and containing patents of < 5 distinct classes (Figure 5). This shows that citation-based CNM clustering outputs reasonably clustered communities.

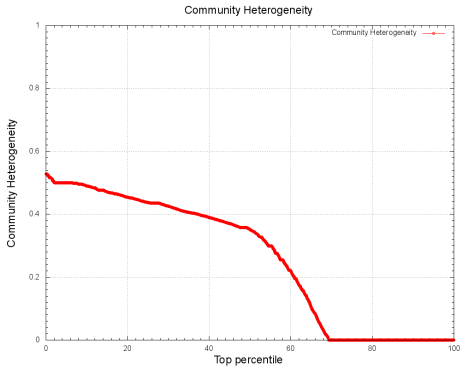

Figure 4: Community heterogeneity of each community, in increasing order

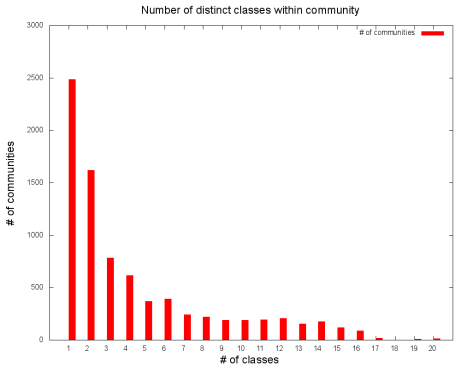

Figure 5: Number of classes within community

5.2 Tracking community evolution over time

The goal here is to uncover more heterogeneous, yet meaningful dynamic communities that would extend across long periods of time and contain a substantial number of patents. We ultimately decided to expand upon the approach for tracking communities in dynamic networks as proposed by Greene [8] and utilized by Britz [9]. We found Britz's similarity measure to be reasonable because it would allow us to capture communities that grow rapidly while still preserving the homogeneity of that single dynamic community as it grew. We tested values of $\theta = 0$ and $\theta = 0.5$, and noticed that for the CNM algorithm, split events were eliminated after $\theta = 0.5$. We felt this was a weakness of $Nsim$. A community which receives a large portion of nodes from an ancestor community, where those nodes only constitute a small portion of the ancestor community, will not receive an edge from the parent community. That means our analysis would show that the receiving community grew faster than the natural process of patent addition and citation, and the ancestor community may have shrunk more than the natural process. Therefore we extended the $Nsim$ measure to be

$$\theta_0 > \frac{|C_1 \cap C_2|}{|C_2|} \quad \wedge \quad \theta_1 > \frac{|C_1 \cap C_2|}{|C_1|} \tag{5}$$

Values of $\theta_0 = 0.8$ and $\theta_1 = 0.8$ were used to produce the dynamic community DAG. We then found the weakly connected components (WCC) of the DAG. For each WCC, we collapsed the merge and split events into a single dynamic community by aggregating the communities in each time step. We thought this was reasonable as the commmunities spun off (split) from an older, larger community are essentially densified components of the larger community, and can be considered subcommunities. Communities that have not merged yet were also considered as part of one dynamic community. Finally, we discarded dynamic communities whose end sizes were smaller than 50 patents, and whose length in years was less than 10.

Table 1: Subcategory dynamic community statistics

Subcategory	Number	$\bar{Length}$	$\bar{StartSize}$	$\bar{EndSize}$	Splits	Mergers	Non-events
55	82	14.4 yrs	3.3 patents	1019.5 patents	98	1012	40791
22	6	14.2 yrs	8.3 patents	14262.7 patents	76	320	8190
14	268	16.6 yrs	3.6 patents	385.4 patents	63	1714	104185

(a) Events counted before filtering smaller communities. Non-events are degree 1

Algorithm 1 Dynamic community tracking algorithm

```
D: set of dynamic communities
T: time range of network
initialize dynamic community directed graph G
for t, t + 1 ∈ T do
    for C_i ∈ t do
        for C_j ∈ t + 1 do
            if Nsim(C_i, C_j) > θ_0 ∧ Nsim(C_i, C_j) > θ_1 then
                Add edge (C_i, C_j) → G
            end if
        end for
    end for
end for
find weakly connected components WCC of G
for w_i ∈ WCC do
    d: new dynamic community
    y: time range of WCC, d
    if y < 10 then
        continue
    end if
    for t ∈ y do
        c: set of static communities in w_t
        Add union of communities, ∪c → d_t as one community
    end for
    if |d_{t=y−1}| > 50 then
        add dynamic community to set of all dynamic communities
        d → D
    end if
end for
```

6 Measuring Growth Potential

6.1 Quantifying Growth

We are interested not in the absolute size of growth, but rather percentage of growth. Annual growth rate of a cluster A between two successive years is simply defined as $GR = \frac{|A|_{t+1}}{|A|_t} - 1$. However, we have noticed that for many clusters, GR can fluctuate from year to year, sometimes very significantly (i.e. GR plummets in one year, then rises rapidly in the next). Also, we believe that there exists some latency until the effect of current patent environment is fully realized. For example, it might take several years until some influential innovation spawns other children patents. To smooth out fluctuation and to take into account latency, we decided to focus on longer-term growth rate, using 5-year and 10-year CAGR (Compound Annual Growth Rate)

$$CAGR(t_0, y) = (|A|_{t_0+y}/|A|_{t_0})^{\frac{1}{y}} - 1 \tag{6}$$

For each community we found, we calculated CAGR(t_0, 5) for [birth-year of community] $\leq t_0 \leq 1994$ and CAGR(t_0, 10) for [birth-year of community] $\leq t_0 \leq 1989$.

6.2 Extracting predictor metrics from communities

We first hypothesized a list of cluster network features that we believe are predictive of high growth rate in the future. Rather than feeding a laundry list of features into the regression model, we checked correlation between each of the features and community growth rate and eliminated those that have no correlation with growth. For metrics that we believe dto have exponential relationship with growth rate (i.e. size of in-component and age), we used log-transformed values. Here are a number of relevant features that we have tested:

- Average relative age of patents in the community. The relative age of a patent is determined as the difference between the current year and the year the patent was approved. We believe that the influence of a patent on innovation decays exponentially as time passes. Therefore, clusters with lower average relative age will grow less than those with higher average relative age.
- Size of in-component of the community normalized by community size. In-degree only measures direct citations that a patent receives. Measuring the total size of the in-component for a community will gauge indirect influence of the patents in the community (i.e. patent leading to another patent

that leads to another patent). The larger the size of in-component, the more innovative patents are in the community, and the higher the growth. We believe that the relationship between the metric and growth rate is exponential, following preferential treatment model.

- Community heterogeneity. Heterogeneity is determined by the number and frequency of distinct classes that are represented within a community as indicated by metadata (see 5.1). We believe that heterogeneous communities exhibit greater potential for growth because patents in this community are able to take advantage of a wide variety of ideas and skills across different fields, and are thus uniquely positioned for growth in an interdisciplinary manner.

- Fraction of citing patents that are foreign. For each patent X in the community, we looked at whether patents that cite X is of the same community as X or of a different community. We predicted that communities that receive more citations from foreign communities (rather than having more citations among themselves) will exhibit higher growth since it means that patents in the community were influential and inventive enough to be referenced by other communities.

$$\frac{|In(C)|}{|In(C)| + |C|} \tag{7}$$

- Number of foreign communities cited by the community. We looked at the number of distinct communities that contain at least one patent that is cited by the given community's patents. We predicted that if the community takes more ideas from other communities, interdisciplinary synergy would occur and the community will grow faster.

- Average PageRank score. PageRank scores of each patent in the subcategory were computed. For each community, we averaged its patents' PageRank scores. We believe that the higher PageRank is, the higher the growth potential is. This is a similar metric to size of in-component, reflecting frequency of both direct and indirect citations.

- Clustering coefficient. A community with high clustering coefficient implies citation patterns of its member patents are similar. This high level of unity among patents may suggest that patents in the community might be unoriginal. Therefore, we expect clustering coefficient to be negatively correlated with growth rate.

- Average number of claims. Number of claims is part of the metadata for each patent. A claim refers to the subcomponent, or "building block" of a patent. The greater the number of claims, the larger the scope of the invention, and presumably the higher the growth potential. We used mean number of claims for patents in the community.

6.3 Correlation between features and growth rate

After extracting a set of features and future growth rates for each community in each year, we aggregated all data points and plotted correlation between individual features and growth rate. We observed that communities that were originally tiny often exhibit extremely variable growth rates (i.e. 100%+ annual growth) due to their small absolute size, hence causing noise in the data. Therefore, we discarded data points where community size < 100 and focused on communities that are reasonably big and established.

The following correlation plots were made with $N = 534$ for 5-year CAGR (red dots) and $N = 229$ for 10-year CAGR (green dots). R-values for correlations using 5-year and 10-year growth rates are denoted as r_5 and r_{10}, respectively.

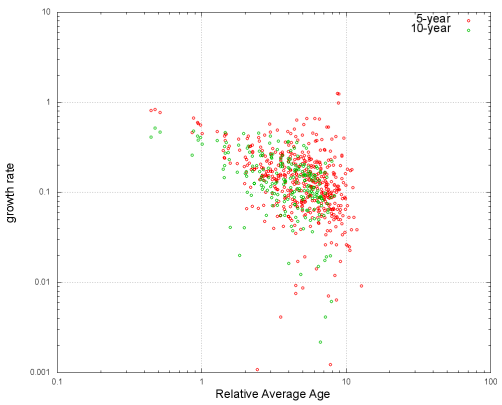

Figure 6: Average relative age of patents (log-log scale), $r_5 = -0.3968, r_{10} = -0.4813$

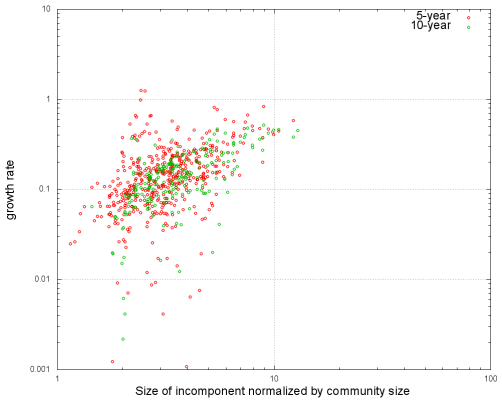

Figure 7: Size of in-component normalized by community size (log-log scale), $r_5 = 0.4290, r_{10} = 0.6033$

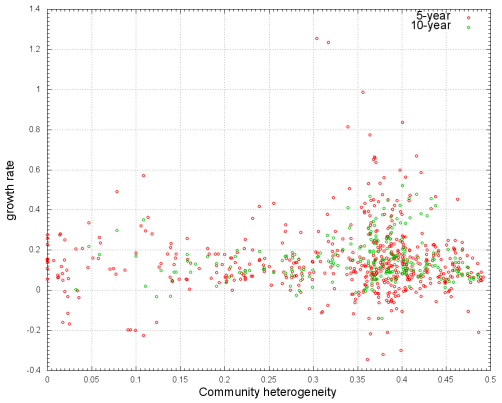

Figure 8: Community heterogeneity, $r_5 = 0.0318, r_5 = 0.1329$

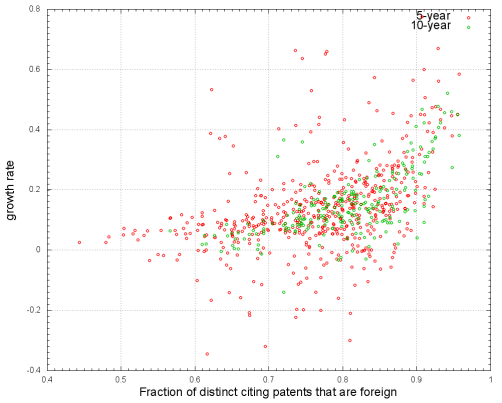

Figure 9: Fraction of citing patents that are foreign, $r_5 = 0.3838, r_{10} = 0.6536$

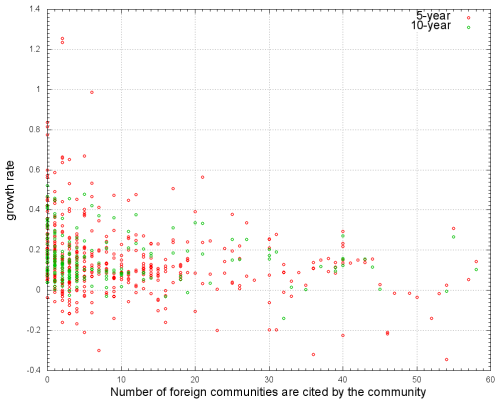

Figure 10: Number of distinct foreign communities that are cited by the community, $r_5 = -0.1012, r_{10} = -0.0396$

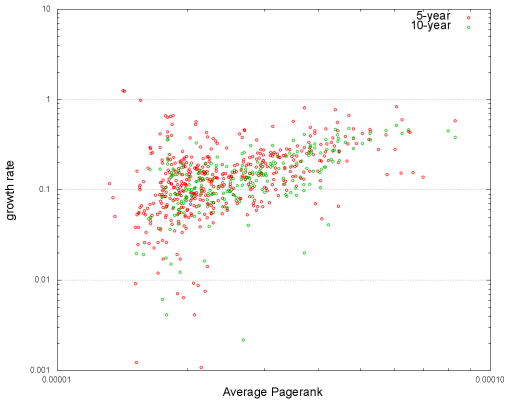

Figure 11: Average PageRank score (log-log scale), $r_5 = 0.4462, r_{10} = 0.5688$

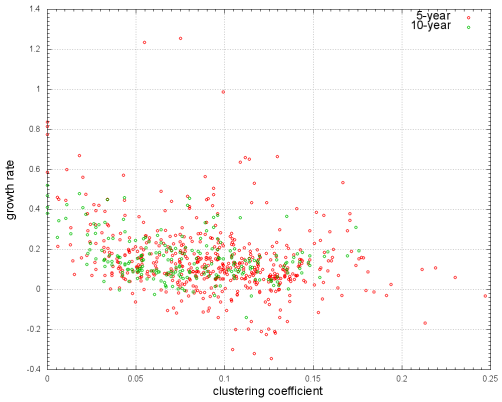

Figure 12: Clustering coefficient, $r_5 = -0.2977, r_{10} = -0.3429$

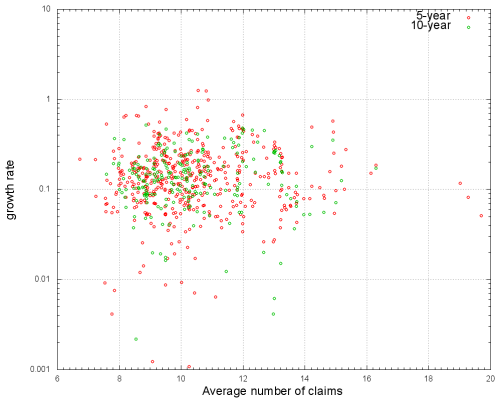

Figure 13: Average number of claims, $r_5 = 0.0098, r_{10} = 0.0493$

Results show that the following metrics had a moderate ($r^2 > 0.23$) correlation with growth rate. The directions of relationships were consistent with our expectations. These four metrics were chosen as features for our regression model:

- Fraction of citing patents that are foreign (positively correlated)
- Size of in-component normalized by community size (positively correlated)
- Average PageRank score (positively correlated)
- Average relative age of patents (negatively correlated)

The remaining metrics had insignificant correlation with growth rate and were discarded.

7 Regression model for predicting growth rate

We experimented with two regression models: linear regression and support vector regression. For each regression model, we trained it based on any features extracted from the previous section that had a positive correlation between features and growth rate, which were the fraction of citing patents that are foreign, the size of the in-component normalized by community size, the average PageRank score, and the average relative age of patents. We evaluated the models by attempting to predict growth rates for a new set of communities purely based on features and the generated models, and then calculating R^2 and RMSE values for each of the models.

Data set	Regression model	R^2	RMSE
5year	Linear	0.761164321029556	0.017577168370117868
5year	Support vector	0.76442561386179264	0.02097234985300641
10year	Linear	0.77931944330264635	0.016241037845776587
10year	Support vector	0.78470176288222593	0.019167236411349209

Table 2: R^2 and RMSE values for linear and support vector regression models on the 5year and 10year data sets.

Our results indicate that these features in aggregate can be used to reasonably predict growth rate of communities, and that it is appropriate to use learning algorithms to train a model for performing the aforementioned prediction task.

8 Conclusion

For dynamic community tracking, tweaking the similarity formula to handle split events and increasing the similarity thresholds to 0.8 greatly improved both the homogeneity across time within communities and hetereogeneity between communities.

In addition, fraction of foreign citing patents, size of in-component normalized by community size, and average PageRank score of patents were found to be positive indicators of growth potential, while average relative age of patents was a negative indicator of growth potential. These features can be used to train a regression model that is capable of predicting future growth rate with moderate success.

9 Future Work

We believe there is a strong need for an algorithm that simultaneously detects and tracks dynamic communities; perhaps an extension of CNM or BigCLAM into time, or even better, a real-time algorithm that updates upon each edge and node addition. Ideally, this algorithm should consider the amorphous and hiearchical nature of dynamic communities.

Another interesting topic would be to evaluate the full patent citation network and understand the interactions between different categories (i.e. meta-communities). Our current method is limited in that it fails to take into account inter-categorical interaction. We ran community detection within each subcategory in isolation; also, other than the size of in-component, most of the predictors we used were independent of other patents outside the subcategory. If we could improve the running time of dynamic community detection and explore the full citation network in its entirety, a lot of interesting observations can arise.

For instance, we tabulated counts of distinct (subcategory, subcategory) relationships, where the first item represents the subcategory of the citing patent and the second item represents the subcategory of the cited patent. Counts are normalized on a log-log scale and converted to a number on a scale of 0 to 255. A density map is visualized with this information in order to help identify interesting subcategories that are exhibiting high growth rates over time. Segregating the dataset into years based on the year in which the citing patent was approved and then making heatmaps from the data for each year shows evolution of inter-category relationships. It would be interesting to answer questions such as "Do communities in two categories that have a strong exchange grow in tandem (similar growth rate)?", "Do communities in categories that are isolated from other categories grow slower than average?", and "Does inter-category relationship get stronger over the years (positive feedback)?"

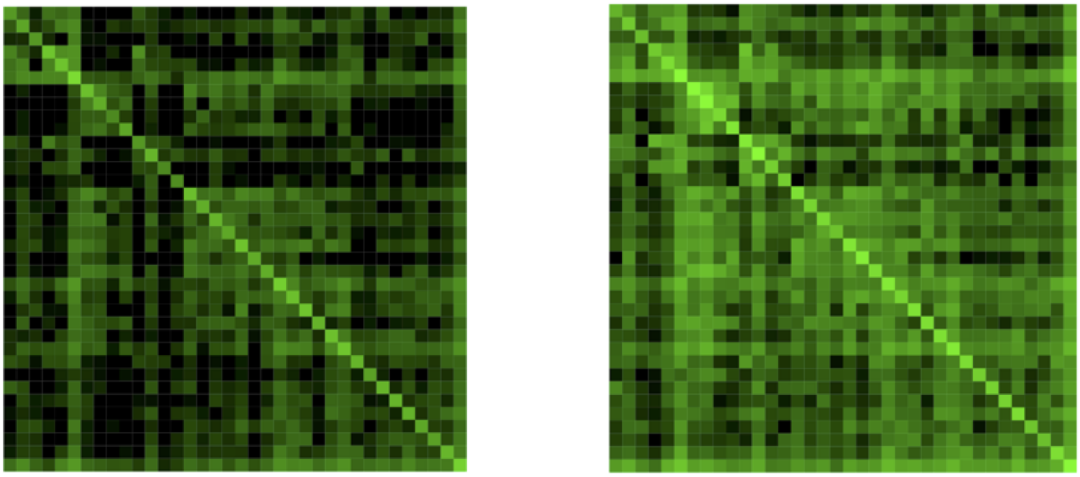

Figure 14: Plots representing frequency of citations between two categories normalized on a log-log scale, shown for 1975 (left) and 1999 (right). Each square represents a distinct (category, category) relationship. Squares are colored on a scale of 0 (darkest) to 255 (brightest) with 255 representing the most frequent relationship.

10 Individual Contributions

We feel that our contributions to the project were approximately equal.

- Roger: Background research, implementation of regression training and testing for identified features, future work, inter-subcategory citation visualization, writing sections 3, 8, 9, making poster
- Jichan: Background research, running static clustering, cleaning data, defining and extracting predictor features and growth rates, measuring and plotting correlation, writing section 4.3, 5.1, 6
- Alex: Background research, dynamic community tracking (improving algorithm and running it), proposed predictor features, writing sections 1,2,5.2

References

[1] Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools.* NBER Working Paper 8498.

[2] Chang, Shann-Bin, Kuei-Kuei Lai, and Shu-Min Chang. *Exploring technology diffusion and classification of business methods: Using the patent citation network.* Technological Forecasting and Social Change 76.1 (2009): 107-117.

[3] Valverde, Sergi, Ricard V. Solé, Mark A. Bedau, and Norman Packard. *Topology and evolution of technology innovation networks.* Physical Review E 76.5 (2007): 056118.

[4] Lancichinetti, Andrea, and Santo Fortunato. *Community detection algorithms: a comparative analysis.* Physical review E 80.5 (2009): 056117.

[5] Newman, Mark EJ. *Modularity and community structure in networks.* Proceedings of the National Academy of Sciences 103.23 (2006): 8577-8582.

[6] Clauset, Aaron, Mark EJ Newman, and Cristopher Moore. *Finding community structure in very large networks.* Physical review E 70, no. 6 (2004): 066111.

[7] Yang, Jaewon, and Jure Leskovec. *Overlapping community detection at scale: A nonnegative matrix factorization approach.* In Proceedings of the sixth ACM international conference on Web search and data mining, pp. 587-596. ACM, 2013.

[8] Greene, D.; Doyle, D.; Cunningham, P., *Tracking the Evolution of Communities in Dynamic Social Networks,* Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on , vol., no., pp.176,183, 9-11 Aug. 2010

[9] Britz, Denny, Caiyao Ma, and Chuan Xu. *Quantifying Community Growth in Dynamic Social Networks.* Course Project in CS 224W, Stanford University 2013.