Beyond Patent Citations as Measures of Innovative Search and Success

Project Summary

Hundreds, possibly thousands, of research papers have used patent citation data to study innovative search and success. Recent refinements in this literature (Alcacer and Gittleman 2006; Cotropia, Sampat and Lemley 2012) have raised substantial questions about citation measures. Using big data techniques including LASSO content labels, Jaccard similarity of labels, and computational linguistics analysis of communication between a patent applicant and their examiner, we: 1) review and critique citation measures, 2) propose very different measures of innovative search and success, 3) outline a plan for their validation, 4) make the measures back to 1975 available for all researchers, and 5) offer examples of substantive purchase, mainly in more comprehensive theory and conceptualizations of technological trajectories, search and innovative success, and knowledge diffusion. The techniques open up additional applications in inventor and assignee disambiguation, prior art search, and matching of treated and control patents for econometric estimation.

The intellectual merit of the proposal:

Citations in a patent to prior patent art are intended to demonstrate an awareness of extant work in the area, such that the current patent can clearly demarcate its additional contribution and worthiness of grant. Prior art citations are not intended to measure the cited patent's technical importance, social impact, or commercial value, yet that is how they are typically used in the research literature (widely cited patents are assumed to be breakthroughs, for the seminal paper, see Trajtenberg 1991). While citation counts correlate with social and economic impact, the relationship is indirect, noisy, and difficult to validate (Harhoff et al. 1999). The main goal of this proposal is to supplement future patent citation counts, as measures of an invention's importance; the method is to consider how many future patent applications are "blocked" by a particular patent, and the assumption to be validated is that "blocking patents" are valuable pieces of technology space.

Prior art citations are also used as measures of local vs. radical search, by considering the proportion of prior art that is owned by the patentee (Sorensen and Stuart 2000) or the age of the prior art cited (Fleming 2001). Once again, this was not the original intent of prior art citation, and even less than citation counts, these measures have not typically been validated. A second goal of this proposal is to supplement the usage of derivatives of prior art patent citations, as measures of local search; the method is to identify the most differentiating words for a given patent, relative to all other patents, since 1975, and the assumption to be validated is that similarly "tagged" patents are close to one another in technology space, and thus represent local search.

Broader impacts resulting from the proposal

The use of bibliometric data (patents and scientific papers) to study innovation and its social and economic impact has become increasingly common with the widespread availability of large databases. The advantages of such approaches are easily calculated and consistent measures across millions of documents and thousands of technologies. The disadvantages are that the measures (prior art references by patents, in this case) provide poor and indirect correlations with actual phenomena. Using computational social science, this research applies improved big data techniques, combined with survey validation, that could ameliorate the disadvantages, and made available to researchers, significantly improve research in the innovation and technology strategy literatures. In addition to enabling better research, this work provides wider societal benefit in helping inventors, lawyers, examiners, and investors in prior art search, improved disambiguation of the patent database, and faster identification of important patents.

Project Description

Introduction: This narrative will begin with 1) a critique of the use of prior art patent citations in the development of measures of innovative search and success and then 2) offer new approaches based on maximally differentiated content label ("*tags*") identified from each patent's claims data, and blocking actions issued by the USPTO (United States Patent and Trademark Office) in the application process of new patents. We will then 3) discuss validation of these measures, 4) new theory enabled by the measures, and finally 5) additional applications of interest to scholars, patent applicants, examiners, and lawyers and investors.

1) A review and critique of using citations to calculate measures of innovative search and success

Using backward prior art citations to calculate "local search." Many have described innovation as a search process, and most of those who use this analogy conceive of innovative search as a "local" vs. "distant" search process (March and Simon 1958; Nelson and Winter 1982). Local search occurs in close proximity to prior success, where proximity is similarity in technologies, approaches, applications, components, or architectures. Distant search occurs, in contrast, far away from current approaches, success, and extant productivity. Local search is incremental and "normal" (Kuhn 1962); distant search is radical and risky (March 1991).

Operationalizing these notions of local and distant search remained difficult, with a notable advancement coming from Stuart and Podolny (1996). Using the backward prior art citations [1] from a firm's portfolio of patents, they computed the overlap in citations between firms. (A patent must cite "prior art", or similar previously granted patents, and then establish its novelty, relative to that prior art). From this proportion of overlapping cites, they calculated a relative measure of technological "distance" between firms and a position of the firm in "technological space," relative to other firms. Repeating this method for a latter time period produced a very similar picture, and thus supported the claim that firms search locally and do not typically break out into radically new domains and areas of technology space.

Another advancement, more representative of the research frontier at the time, came from Sorensen and Stuart (2000). This work used the proportion of prior art citations made to a firm's previous patents, the argument being that firms that cited their own prior art searched more locally. Rather than "stick to their (known) knitting", firms that cited other firm's patents were assumed to be searching more distant and less familiar innovative opportunities. Many other papers have used similar metrics.

As the patent and innovation literatures have matured, however, scholars have uncovered some potential shortcomings in these measures. Alcacer and Gittleman (2006) exploited new data from the USPTO to demonstrate that patent applicants provide only half of their prior art citations on average. In other words, inventors (or their lawyers) are poorly informed as to the neighborhood in which they have searched. Consistent with this, Roach and Cohen (2013) find that citations underestimate knowledge flows and can be biased by firms' strategic interests. Finally, recent research in the law literature has cast doubt that overwhelmed patent examiners are even capable of adequately checking all pertinent prior art (Cotropia, Sampat, and Lemley 2012), given the limited approval time they have for each patent application. Given that

[1] We define backward prior art citations as those that cite backwards in time to an earlier patent, from a focal patent. Forward prior art citations are those received by a focal patent from future patents (referred to by the USPTO as "Referenced By".)

inventors appear uninvolved with the citation process, that managers see little correlation between their patent citations and sources of knowledge, that examiners confront vast search spaces without useful tools and enough time, these papers have cast doubt upon the meaningfulness of prior art citations, even for their original intent, let alone as measured or derived indicators of search and innovation.

Using forward prior art citations to calculate patent impact, value, or success. The study of citations to previous patents or publications has a long history in the bibliometric literature. Citations to papers provide a concrete example of Newton's "standing on shoulders" description of the social processes of science; researchers acknowledge prior work (generally favorably, though sometimes critically) by summarizing the prior contribution, partly in order to contrast their current work.

The number of citations that a patent receives has been widely used as a measure of its importance, impact, or value (Jaffe and Trajtenberg 2004). Trajtenberg (1990) was amongst the first to suggest that scholars of innovation use future prior art patent citations as a measure of a patent's social, technical, or commercial value. Campbell and Nieves (1979) had suggested earlier that patent citations were better indicators of importance than science paper citations, because they should have been vetted by an examiner. Harhoff et al. (1999) established the financial value of German patents and found an excessively skewed distribution, from (typically) nothing to hundreds of millions.

Though no definitive studies exist, probably hundreds and possibly thousands of peer-reviewed papers have used citations as a measure of a patent's importance. Very few of these studies have validated their use of citations through fieldwork or surveys, and most simply rely upon a citation to Trajtenberg (1990) or Harhoff et al. (1999). Yet patent citations do not directly measure non-technical or commercial impact at all, and correlations with these outcomes surely reflect some underlying but omitted variable. One plausible model is that areas of commercially important technology receive more future investment and research attention, such that more patents in the area are invented; earlier patents in the area are cited because the patent office receives and approves additional patents that need to cite something. Hence citations might provide a better measure of future effort in a promising technological trajectory, rather than a direct measure of the commercial and non-technical importance of a particular patent.

2) New approaches to measure the novelty of search and importance of a patent.

We will introduce and describe two approaches that can replace the use of prior art citations in measuring search and impact: 1) *tagging*: characterizing innovative search as local or distant, by looking at the similarity in unique words in their claims, and 2) *blocking*: characterizing the value of the impact, as measured by how many other inventors were denied in their attempt to claim the same innovation, or piece of technology space.

Using maximally differentiating content labels ("tags") identified from each patent's claims data. The LASSO (Least Absolute Shrinkage and Selection Operator) is a well-known sparse learning algorithm (Tibshirani 1996) and takes the form

$$\min_{\beta} \left\| X^T \beta - y \right\|_2^2 + \lambda \left\| \beta \right\|_1 \qquad (1)$$

where X is a $n \times m$ data matrix, with each column a specific feature (in this case a word from our dictionary of all patent words), each row a specific data point (in this case a particular patent), y is a m-dimensional response vector, and $\lambda > 0$ is a parameter. The l_1-norm penalty encourages the regression coefficient vector β to be sparse; this minimizes the bag of words that describes each patent, thus easing interpretability. If each column is a feature of word or term, then a zero element in β at the optimum of (1) implies that this particular feature (or word in our case) is absent from the optimal model. If λ is large, then the optimal β is very sparse, and the LASSO model then allows selection of only a few words that are the best predictors of the response vector (or particular patent).

This way of selecting features (so-called tags) has two advantages. For one thing, LASSO automatically eliminates non-distinguishing features (so-called stop words), hence avoids the step of manual compilation of stop word list, which can be domain specific. Secondly, LASSO does not assume independence of features as opposed to other machine learning algorithms, e.g., Naive Bayes, hence adding objectivity.

For example, U.S. patent 4243019 (Light-weight-trough type solar concentrator shell) has 28 maximally differentiating tags, when compared to 28,789 patents in clean technology (Nanda, Younge, and Fleming 2013): radiant vantine fek reflective with pleats cylindric parker polygons paraboloidal pleated minnesota aluminized focussing paul facet mining extremes facets reflecting corrugations race rivets striking polygon radii simulator search. The number of tags is variable, and is increased until the incremental increase in information passes some threshold (for example, and in our case, 5%). In contrast to this smaller set of 28,789 patents, we propose to develop maximally differentiating tags across all US patents back to 1975 (essentially, all patents for which we can access the text of claims data). We have scoped 2 million unique terms in the unique word dictionary for these approximately 5 million patents. This implies regression matrices of 2 by 5 million and requires the large hardware budget we request below.

Given a set of tagging words from each patent, we are now in a position to compare patents and determine their "distance" from one another. While a variety of distance measures exist, we choose a simple Jaccard measure [2]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \qquad (2)$$

This measure takes the number of same terms in both sets and divides it by the total number of unique terms in both sets. The measure ranges from 0 (no terms in common) to 1 (all terms in common). This measure is calculated for pairs of patents (and hence is fundamentally "relative" – it only exists in relation to some other patent or set of patents); we intend to do all pairs of US patents since 1975, using a fast processor and extra memory. Once the approximately 5 million Jaccard measures are calculated for a given patent, we can conceptually graph a histogram for each patent, which we expect to look something like the following Figure 1 (we observed this in developing the measure on the smaller dataset for Nanda, Younge, and Fleming 2013):

[2] http://en.wikipedia.org/wiki/Jaccard_index

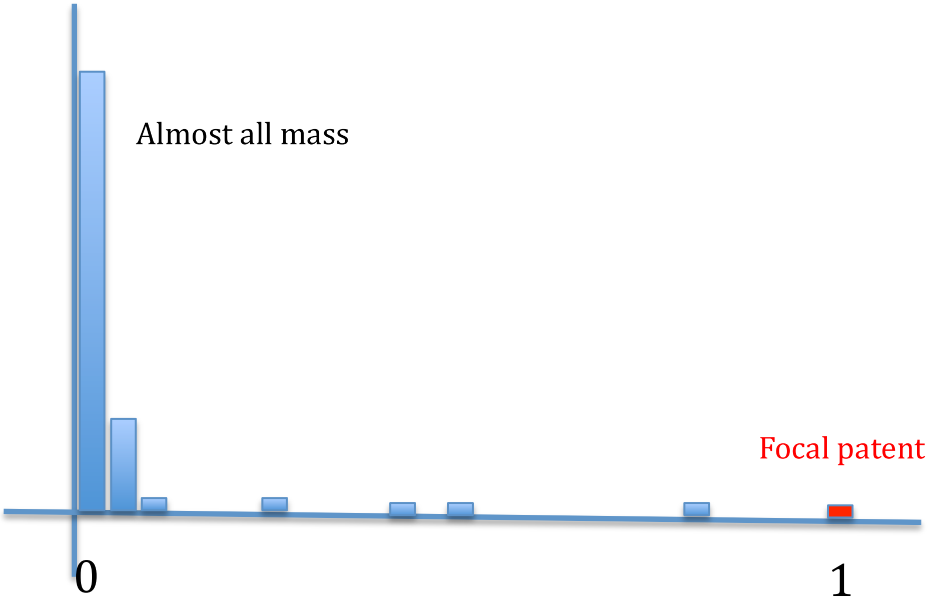


Figure 1: Hypothesized distribution of Jaccard or relative distances between a focal patent and other patents in the database.

Since we are considering so many patents, almost all of which will be in fields completely different than the focal patent, we anticipate the bulk of observations will have no overlap (no tags in common - the large mass to the far left at 0). The remaining observations with partial overlap will string out along the bottom, until the focal patent, which will be at 1 (exactly similar to itself). Some of these distributions will be more skewed than others and these differences provide us with the opportunity to calculate an absolute measure of "distance". We will calculate and characterize a variety of absolute distance measures, which could include: the proportion of patents that have any common tags (surely a very small number), nearest neighbor (the distance to the first patent to the left of the red focal patent above), or the distance to some percentile. [3]

Once we have a single measure of absolute distance for all patents, relative to some other set of patents (in our case, the patent record back to 1975), we can move to more useful measures, conceptualizations, and visualizations for innovation scholars. In particular, we can clearly operationalize and visualize where a patent falls within a technological trajectory (Dosi 1982). We define four phases: new, normal (the middle productive phase), dying, or failed (never got under way).

To determine the phase, we calculate two distance measures. By changing the set of comparison patents, and combining past and future comparisons, we can identify where in a technological trajectory a patent appears. When we compare to past patents, we determine whether the patent was invented in a "crowded" neighborhood, near other patents, or whether it was invented in the country, far away from other patents. When we compare to future patents, we determine whether a patent will have new neighbors arriving and crowding in, or whether they will stay or become increasingly lonely over time, as the space is deemed to be

[3] Some patents are essentially the same patent, with same name, inventor (s), assignee, and almost similar claims. We can remove those easily, using the Fung Institute patent database.

unattractive. Putting the “backward” and “forward” similarity together can be visualized in the Figure 2. We expand on the theoretical implications of this below.

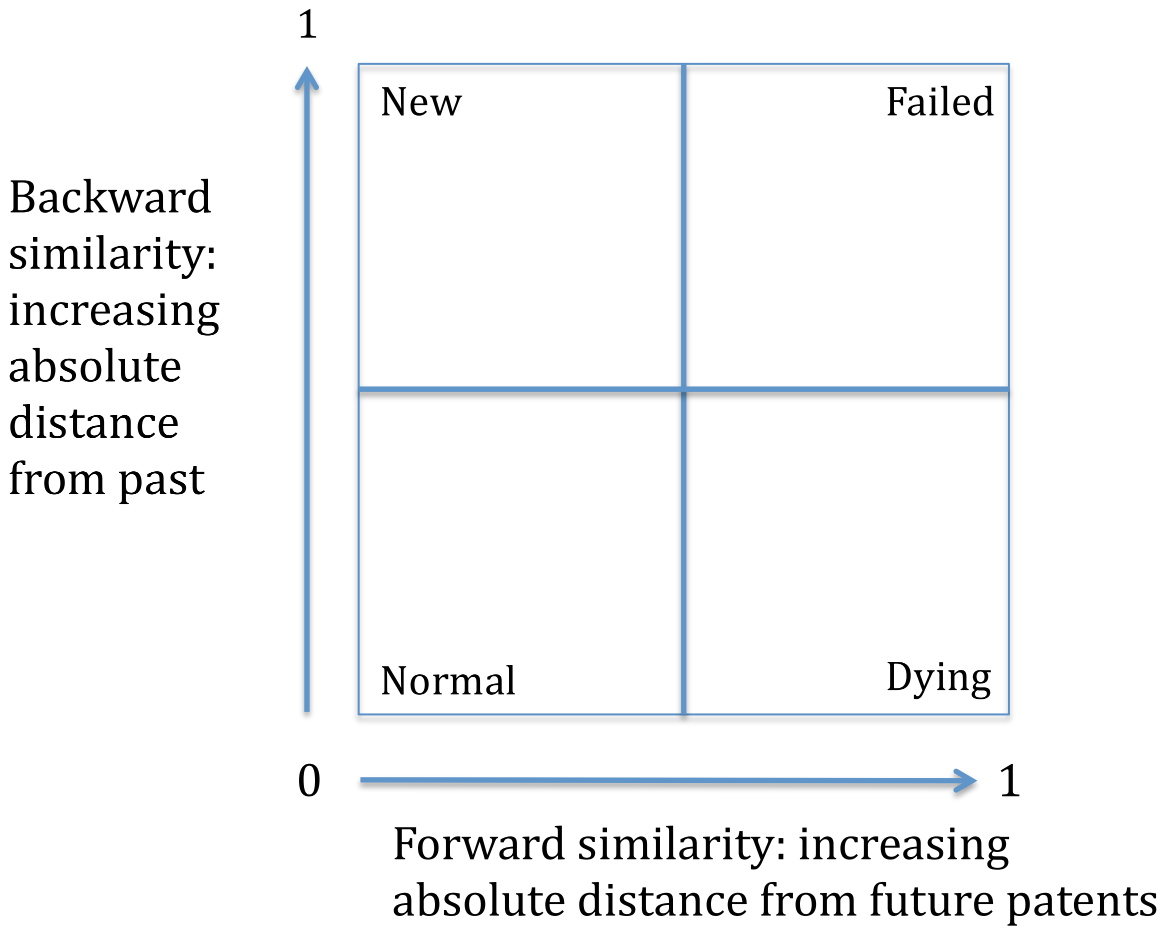

1
New
Failed
Backward similarity: increasing absolute distance from past
Normal
Dying
0
1
Forward similarity: increasing absolute distance from future patents

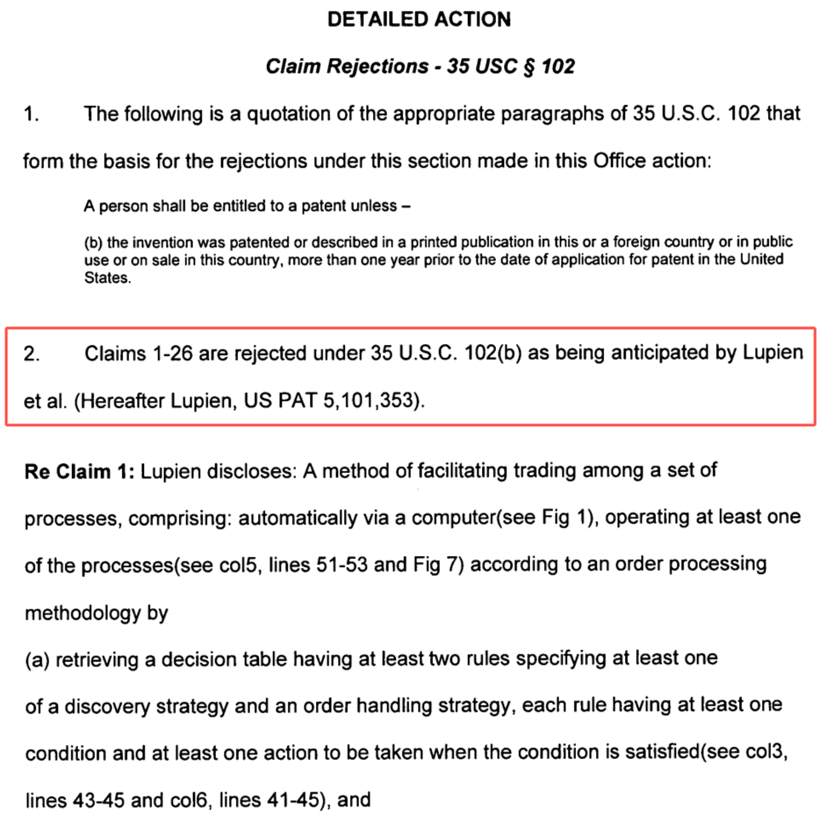
Figure 2: Illustration of proposed measure of placement of an invention within a technological trajectory, based on absolute measures of Jaccard distance between patents.

Using blocking actions issued by the USPTO in the application process to measure importance. The USPTO has recently begun making public the correspondence between the patent applicant and examiner (see the Patent Application Information Retrieval (PAIR) database, http://portal.uspto.gov/pair/PublicPair). While the data are typically in .pdf format and not well organized, they still provide many new research opportunities (though they require download and optical character recognition – OCR). We propose to exploit one particular feature of the data, namely, to identify what new patent applications are denied, because previous patents have already claimed that segment of technology “space.” Assumedly, previous patents which “block” future patents are much more valuable commercially, having claimed the important space first.

The PTO decisions of primary interest are called 102 and 103 actions. A 102 action rejects the application because it was found to not be novel at the time of application. 103 rejections cover those applications deemed obvious. Other actions exist, but these two are the bulk of the actions (Cotropia, Sampat, and Lemley 2012) and should be best at identifying prior patents that have already claimed valuable technology space.

The blocking material can be scientific literature, a US granted patent, a US application that has been published and is now considered as prior art, or foreign patents. Due to the inconsistent formats of examiner names and blocking materials, we will apply computational linguistics techniques to parse the English sentence structure, and extract the subject/object after a verb. After this step, we will construct a network of blocking materials and blocked applications.

The actual process of obtaining the data goes as follows. First, the correspondence data are accessed (the application's *Transaction History*), either at Google or directly from the USPTO. [4] Files with the words containing "Rejection" are collected and OCRed and text strings such as "XX U.S.C. XXX(X)" and "As anticipated by" are identified. For example, patent 5,101,353 that blocks the other patent application (09/801,583) by examiner Ankeeta Shah (please see http://funglab.berkeley.edu/pub/09801583-2005-04-28-00005-CTNF.pdf, page 3)



DETAILED ACTION

Claim Rejections - 35 USC § 102

1. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

2. Claims 1-26 are rejected under 35 U.S.C. 102(b) as being anticipated by Lupien et al. (Hereafter Lupien, US PAT 5,101,353).

Re Claim 1: Lupien discloses: A method of facilitating trading among a set of processes, comprising: automatically via a computer(see Fig 1), operating at least one of the processes(see col5, lines 51-53 and Fig 7) according to an order processing methodology by

(a) retrieving a decision table having at least two rules specifying at least one of a discovery strategy and an order handling strategy, each rule having at least one condition and at least one action to be taken when the condition is satisfied(see col3, lines 43-45 and col6, lines 41-45), and

Figure 3: Example of patent applicant-examiner communication, from USPTO PAIR data.

[4] David Kappos, former Director of the USPTO, and Stu Graham, former Chief Economist of the USPTO, have indicated that the data could probably be made available directly from the USPTO. We have learned from within the Google patent group that it is not clear how much longer Google will support these data. The Fung Institute estimates it will three months before all the source files could be downloaded and stored (storage alone if the source files are discarded looks to cost $1,500, if the source files are kept permanently, ~$10,000).

And from this we automatically extract:

09/801,583 2005-04-28 ...directed to Ankeeta Shah whose telephone ... Art Unit: 3628 35 U.S.C. 102(b) as being anticipated by Lupien et al. (Hereafter Lupien, US PAT 5,101,353). Re Claim 1: Lupien discloses: A method of facili...

This process is similar for all types of blocking art. Concurrent with parsing from the unstructured texts, we have also extracted the examiner names, USPTO Art Unit, and timestamp when the action is issued. Our characterization of the source documents remains preliminary, there may well be additional richness to be exploited. For a large example of extracted fields, please see: http://funglab.berkeley.edu/blocking-list-3.php.

Figure 4 illustrates a network visualization, from our very sparse (~1/50) sample of the blocking data for the proof of concept. We will program automatic identification of the blocker and blockee. The full graph should be much more dense; in this case, we were lucky to find degree higher than one. We suspect the ultimate graph will have a very skewed degree distribution, with the important patents illustrating very high degree and most patents of degree 0. The data and computational challenges for both measures and visualizations are quite substantial, hence our large budget request for computer hardware.

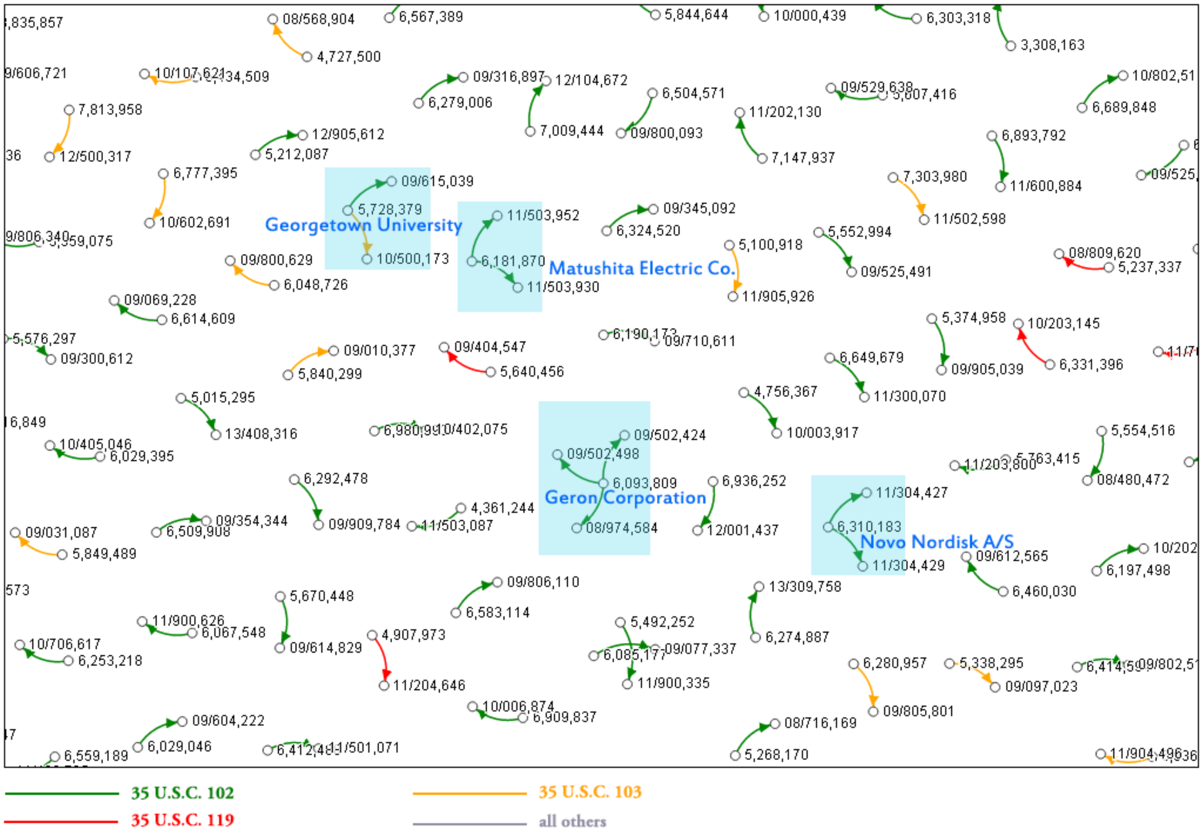


Figure 4: Graph of blocking actions in patent applications, from uniform random sampling (~1/50) of PAIR applications. Geron Corporation's patent number 6,093,809, Telomerase, appears to have blocked three other patent applications. The patent is co-owned by CU Boulder and was invented by Thomas R. Cech, of Boulder CO, and Joachim Lingner, of Epalinges, Switzerland.

3) Validation of tag novelty and blocking importance measures

We will run a pre-survey and interviews (probably with the clean technology dataset developed in NSF grant # 1064182, described below) and a second full survey of inventors to establish the validity of these measures. We will establish correlations between the following metrics: backward similarity, forward similarity, trajectory location, backward citations, forward citations, 102 and 103 actions, litigation (we may not need to ask inventors about this, given increasing availability of free data, please see: http://wiki.piug.org/display/PIUG/2013/08/28/Litigation+Databank%2C+Now+Free+for+Academicians), and financial value. We will base the estimated power needs and size of the second survey on the results of the pre-survey and interviews. The second survey will randomly sample across time and technologies.

Both surveys will be run under the auspices of the UC Berkeley Internal Review Board to protect the subjects' identities. While the initial data will be public, and will provide the basis for contacting the subjects, we will be collecting data on the legal, commercial, and financial value of their inventions. Because of the highly sensitive nature of those data, it will never be released, and only aggregate and non-identifying statistical measures will be reported in the characterization of the measures.

4) New theory enabled by the measures

These measures will enable us to revisit classical theory and results and forge new directions as well. Kuhn's theory of paradigms (Kuhn 1962) and Dosi's of technological trajectories (Dosi 1982) can now be developed empirically, with the ultimate goal of informing thinking on search behavior and strategic implications. The tagging technology will enable very close matching of the technology of a patent, the power of which we will demonstrate with a natural experiment.

RQ1: Who searches in which phase of technological trajectories? A first simple but important baseline question will be to determine the inventors and entities that work within each phase of technological trajectories. Many predictions can be developed from prior work. Large incumbents are most likely to do the bulk of invention in normal trajectory phases (Tushman and Anderson 1986). Older firms do a larger proportion of work in dying trajectories (Sorensen and Stuart 2000). Lone inventors invent more failed trajectories (Singh and Fleming 2009). New trajectories are more likely to be invented by universities and scientists (Bush 1945), brokers (Burt 2006), cross-disciplinary teams (Wuchty et al. 2007), and young inventors. The answers to these questions may vary depending on the particular technology and historical path.

RQ2: Who searches successfully in each phase of technological trajectories? Independent from the first question of who matures and pushes a technological trajectory in each phase is the question of who is likely to find the most important technology in each phase. This is a key thrust of the proposal and opportunity afforded by the new blocking measure as a dependent variable; rather than using the indirect and noisy measure of citation counts, the blocking measure will provide a much more direct observation of the legal, financial, and commercial value of an invention.

Citations are a popular but flawed measure. The problem for this particular question is that they conflate maturation of a trajectory with successful search. For example, considering the four quadrants proposed above in Figure 2, we would expect the typical patent within the top right (new trajectory) and bottom right (normal trajectory) to be more cited, independent of its importance. This occurs simply because many inventors are focusing on the promising new areas – many highly cited inventions are not breakthroughs, rather, they receive citations simply because they are part of a broad and concerted effort to explore a promising area of technology space. One approach to remedy this has been to consider similar patenting areas, as defined by technology classes, for example. Unfortunately, this approach relies upon government or expert classification schemes that are manual, quickly out of date, and sometimes arbitrary.

Another argument for clearly separating search effort from success is that strategic efficacy probably varies across different phases. For example, new trajectories probably on average rely more heavily upon science than normal trajectories, and much more heavily than dying trajectories. As another example, seizing valuable space during normal trajectories is probably a stronger function of organization and resources, than insight and risky search. With the tagging and blocking measures, we can unpack the search process from strategic efficacy in each phase of the technological trajectories.

RQ3: Can we explain non-monotonic results in the search literature? Much of the search literature has found a non-monotonic relationship between novelty and impact (Ahuja and Lampert 2001, Fleming 2001, Jones and Uzzi 2013). This "Goldilocks" hypothesis argues that a moderate amount of novelty is the best for breakthroughs. Such results would be more convincing if they did not depend on citations as a measure of impact. Developing the above criticisms of the conflation of citations and trajectory, the number of citations within a given area probably always follow a non-monotonic progression; first inventions are unevenly cited (breakthroughs are recognized, though after the fact and by the citation definition of breakthroughs, by increasing attention to the area), then all patents in the area are lifted by a flood of citations, followed by a decrease in interest and citations, as the area's fertility is exhausted and inventors focus their efforts elsewhere. This underlying process might swamp more subtle effects, such that incompletely specified models indicate a non-monotonic result. This literature should be revisited, controlling for the phase of a technological trajectory.

Recent work by Jones and Uzzi (2013) suggests that breakthroughs in science come from within technological trajectories – that most of their components of recombination are well-understood and well-used, leavened with some novelty. This is counter to the argument that breakthroughs come from distant and risky search (the PI remains agnostic on the controversy). These measures will enable fine-grained (and far simpler) measures to see if the Jones-Uzzi results hold in technology as well.

RQ4: Do noncompete agreements cause knowledge drain from states that enforce them? This research question will hopefully demonstrate a clean and sharp empirical benefit to the tagging measure. Marx, Singh, and Fleming (2013) demonstrate that an inadvertent flip in the enforcement of noncompetes causes a brain drain from states that enforce to those that don't, particularly for the most productive and collaborative inventors. If knowledge flow follows personnel flow, then states that enforce noncompetes should also suffer a knowledge loss. Acknowledging the problems with using patent citations as measures of knowledge flows (Thompson Fox Kean 2005), we would propose to set up a differences in differences model at the patent level, similar to Furman and Stern (2012).

The identification would come from Michigan's inadvertent change in enforcement (from proscribed to enforceable) in 1985. The treatment group would be all Michigan patents prior to the law change in 1985; the control group is all patents from other states that do not enforce noncompetes. One establishes a pre-treatment baseline ratio of citations between treatment and control patents, and then looks for a change in the ratio following the law change. Further purchase on the diffusion of knowledge can be gained by considering the location, assignee, and inventors of the citing patent. Inventor mobility probably mediates the knowledge flow, and by using the Fung Lab's database of disambiguated inventors, we can include measures for such mobility.

We would use a patent matching tool, described below, to find the most technically similar patents. This is the innovation that will enable this study to accurately identify the impact of the law, by more accurately matching treatment and control patents. It also provides an illustration of the matching tool that should be similarly useful in other studies (as discussed below, we will make it public).

5) Additional applications of interest to patent applicants, examiners, lawyers, and investors.

Our first objective in this work is to contribute to the innovation literature, in a small way with our own new theory and results, and more importantly, by publically providing comprehensive measures from current times back to 1975, for backward similarity, forward similarity, trajectory, and the number of 102 and 103 actions. We also foresee many other applications of this work.

For research, we will build a publically available tool that matches any patent with its nearest neighbors. This will facilitate research designs that wish to match treatment patents with control patents. We can foresee providing a number of additional variables to match on (building on the underlying database which is maintained at the Fung Institute): time proximity, geographical distance or state, size of inventive team, size or name of assignee, and even backward or forward citations.

For commercial and societal applications, we will explore the provision of improved prior art identification tools (based on technology similarity). Based on the sobering work of Cotropia, Sampat, and Lemley (2012), this tool is still needed. Using maximally differentiated tags should provide a great improvement over class based sampling, heuristics, or keyword searching. This tool would help inventers, examiners, applicants, firms, and patent lawyers. We will also explore provision of measures of value based on blocking, which should be of interest to the same groups.

Project management plans for the Fleming Lab

The PI is the Director of the Coleman Fung Institute for Engineering Leadership and in that capacity, has provided and will continue to provide physical space for researchers and the substantial computing infrastructure. The PI's Lab sits next door to the PI's office on the third floor of Blum Hall, thus facilitating daily interaction. The Lab has weekly meetings, with outside stakeholders often present. The Institute sits within the UC Berkeley College of Engineering, and physically next door to the Departments of Electrical Engineering and Computer Science (EECS), which are typically ranked in the top 3 worldwide. While the PI is in the Department of Industrial Engineering and Operations Research (and courtesy at the Haas School of Business), he works closely with EECS and can call on their resources for computing and data science challenges.

Limitations of the proposal and opportunities for future work

One obvious future possibility, if we can characterize the “tags” as legitimate measures of patent components, would be to see whether we can then identify the architecture that binds these components. If this is possible, then we can operationalize the notion of architectural innovation (Henderson and Clark 1990), across millions of patents. It should be possible to visualize the flow of technology over time, as it mixes and swirls, and converges and diverges.

Figure 5 illustrates how tags appear with “solar” energy over time, and while we can’t yet offer a characterization of it, it provides a potential illustration of the flow of technology. “Components”, measured as tags, enter technological trajectory at different times (RQ: When and why do new components enter the purview of inventors and their firms?). They are recombined with other components and over time, some components coalesce into a dominant design and others drop out (RQ: What predicts emergence of dominant design? Which inventors or firms are more likely to invent it? When does a technological trajectory die?) This graph can be extended to include other variables, such assignee, geography, or social networks.

The sample from Figure 5 was taken for convenience, but we will be developing more defensible samples that can also be analyzed (the graphs quickly blow up and become difficult to read). We intend initially to write case histories, in order to characterize the tool.

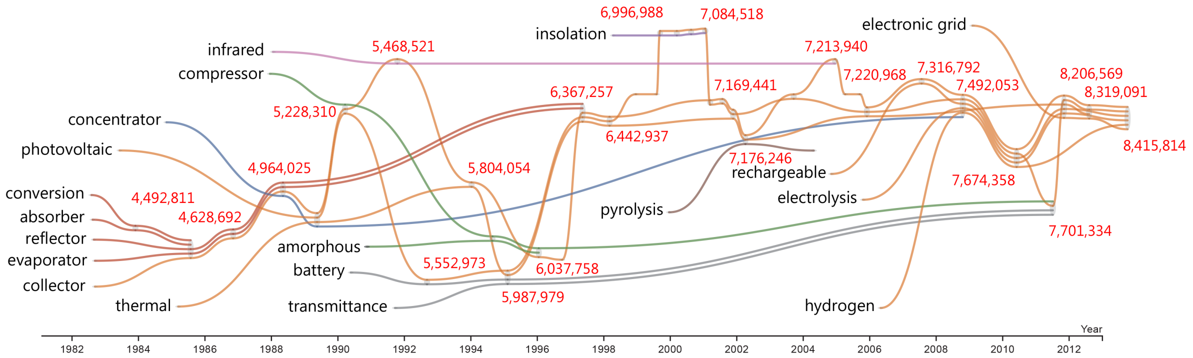


Figure 5: An illustration of the flow of technology in solar technologies, as illustrated by the appearance of tags and their co-occurrence in tagged patents over time. The vertical axis represents distance between tags, prior to the study.

One critique of the technological trajectory measure is that it incorporates future data; for those wishing to estimate purely predictive models, the future distance comparison will not be useable. One obvious risk in trying to replace a well-used measure is that the new measure will not justify the investment. Doing this will require characterization of the original measure, however, which has rarely been done, hence, we will learn something about citations, search, and impact, regardless of the outcome of this proposal.

Results from prior NSF support of PI Lee Fleming:

NSF 0830287: Creating a Patent Collaboration Network Database to Examine the Social Production of Knowledge ($375,063; 10/01/08-10/01/10)
This proposal disambiguated the USPTO database and provided data to visualize network variables for co-authorship.

1) Preliminary disambiguation posted Feb. 2009.

2) Prototype DVN network interface released Oct 3, 2011.
3) Improved disambiguation posted Mar 31, 2011.
4) Almost 22,000 downloads as of August 2013.
5) Working papers:
    1) "Disambiguation of the U.S. Patent Database," Lai, Doolin, Damour, Yu, Sun, and Fleming, invited revision at Research Policy.
    2) "The careers and co-authorship networks of U.S. patent-holders, since 1975," Lai, Damour, and Fleming
    3) "Regional Disadvantage: Non-competes and brain drain", Marx, Singh, and Fleming, invited revision at RESTAT.
6) Trained Alex Damour and Ron Lai in statistical methods, disambiguation algorithms, R, and network visualization algorithms (Damour matriculated in the doctoral program at Harvard in Statistics 9/10, Lai has begun working on the R&D dashboard, a project to quantify the impact of federal research dollars).

With Vetle Torvik, NSF (0965259): From grant to commercialization: an integrated demonstration database which permits tracing, assessing, and measuring the impact of scientific funding. ($750K: 7/01/10-6/01/12)
This proposal integrates disambiguations of the patent and Medline databases, by common people, citations, and organizations (please note that Vetle Torvik's outputs are not listed here, though they are part of the grant).

1) Preliminary integration performed as of 8/31/10.
2) Hired and trained two GSLIS graduate students, both with experience in statistical modeling/data mining, and computer systems (one with back-end database expertise and one with front-end web interface expertise). Hired and trained research assistant on advanced disambiguation techniques. Set up infrastructure: purchased and set up two server-class machines for extensive computational experiments, and one virtual machine for web interface. Wrote code that collects and organizes data on authors, inventors and grants and creates databases which will in turn serve data for planned statistical modeling and web tools.
3) Database available and documented: http://abel.lis.illinois.edu/cgi-bin/PAGe/search.pl.
4) Workshop held at UIUC in June of 2012.
5) Trained Edward Sun and Amy You in data science.

NSF (1064182, not completely spent, these are partial results): From bench to biosphere: a critical analysis of effective deployment and commercialization of clean energy technologies ($657K)

Spending on this grant was delayed until Aug. 1, 2012, due to the PI's move from Harvard to UC Berkeley. The intent of this proposal is to understand the commercialization process of clean energy technologies and identify differences across sectors.

1) A taxonomy of clean technology patents has been developed, based on a discrimination of clean and dirty patents with machine learning, of clean vs. dirty patents. We began with a database of all world-wide startups in solar, wind, and biofuels technology (approximately 4,000 firms, from the i3 cleantech database, see http://research.cleantech.com/). We identified the patents from these firms through matching of firm names and patent assignees. Using these data as the training set, we then discriminated between clean and dirty patents in the entire database and validated

with experts. Ken Younge, postdoc at the Fung Lab and now an Assistant Professor at Purdue, has promised to document this by December of 2013.

2) Working with Ramana Nanda, an Associate Professor at the Harvard Business School (and provider of the i3 database), we linked the clean tech patent data to a database of venture capital and large project investment in clean tech. The work demonstrates how VC investment in clean tech in the mid 2000s co-occurred with more impactful and novel patents (as measured by a cosine similarity measure of distance). It is appearing in a volume edited by Adam Jaffe and Ben Jones: Nanda, R. and K. Younge, L. Fleming, (forthcoming). “Innovation and Entrepreneurship in Clean Energy,” *Rethinking Science and Innovation Policy*, NBER. For a working draft, please see: http://www.funginstitute.berkeley.edu/sites/default/files/Renewable_Energy_0.pdf.
3) An explanation of the fundamental possibilities for geo-location of US patents, written with technical and data support from Google: http://www.funginstitute.berkeley.edu/sites/default/files/GeocodingPatent.pdf.
4) A clean technology geographical mapping tool, documentation is at: http://www.funginstitute.berkeley.edu/sites/default/files/CleanEnergyMapper.pdf. Tool is at: http://funglab.berkeley.edu/cleantechx/.
5) A more general (though less specific) geolocation mapping tool, documentation is at: http://www.funginstitute.berkeley.edu/sites/default/files/PatentMapper.pdf. Tool is not yet posted.
6) An inventor mobility mapping tool, documentation is at http://www.funginstitute.berkeley.edu/sites/default/files/Mobility_Mapper.pdf. Tool is at: http://funglab.berkeley.edu/mobility/.
7) A note describing how to parse US patent data formats: http://www.funginstitute.berkeley.edu/sites/default/files/Extracting_and_Formatting.pdf .
8) Trained David Doolin, Gabe Fierro, Guan-Cheng Li, Kevin Shen, Kevin Johnson, Bill Yeh, Jill Rabinowitz, and Minyoon Jung in data science.

Specific outputs from the currently proposed research:

1) A theory/empirical paper, targeted at an organizations journal (ASQ/Org Science/Research Policy) that develops classical theories of paradigms and trajectories and demonstrates their empirical operationalization. This paper would answer the question of who invents in what phase of a technological trajectory.
2) A strategy paper (Management Science, Strategic Management Journal, Strategy Science) that investigates what types of inventors, collaborations, firms, and strategies are most successful, in each phase of a technological trajectory.
3) A paper that investigates why so much research has found a non-monotonic relationship between local vs. distant search and citations. Appropriate outlets would include all of the above journals listed. This paper would probably contain the documentation of the underlying technology.
4) A shorter empirical piece that would demonstrate the matching tool and hopefully find a clean differences in differences result, that noncompetes cause states to lose their knowledge, as well as their inventors. This paper would also document the underlying tagging technology.
5) A series of technical notes from the Fung Institute, which would include documentation for the public tools.

We intend to make the following data files available to all researchers:

1) Tag words for all granted U.S. patents back to 1975. These are maximally differentiating bags of words for each patent, based on spare regression of the claims section of each patent, against all unique words in the patent corpus. This will be made available in .csv format (Excel spreadsheet), and if resources permit, possibly a database for SQL queries and/or and Application Programming Interface.

An example based on the clean tech sample described in the narrative:

4243019 radiant vantine fek reflective with pleats cylindric parker polygons paraboloidal pleated minnesota aluminized focussing paul facet mining extremes facets reflecting corrugations race rivets striking polygon radii simulator search

or in Excel format (etc. to the right):

patent no	tag1	tag2	tag3	tag4	tag5	tag6	
4243019	radiant	vantine	fek	reflective	with	pleats	…

2) The number of 102 and 103 blocking actions for all granted U.S. patents back to 1975. These will be count variables, probably in skewed distributions with modes of 0. They will be made available in .csv formats (Excel spreadsheet), and if resources permit, possibly a database for SQL queries and/or and Application Programming Interface.

An example (data are entirely made up):

Patent number	102 actions	103 actions
4243019	0	0
4243020	3	1
4243021	1	0
4240322	0	1
4240321	0	0

3) Results of a measure validation survey (which will only include aggregate measures, to protect subject response privacy).
4) A matching tool, so that any user can input a patent number and receive technologically similar patents in return. This tool should be helpful to scholars working to find counterfactual patents.