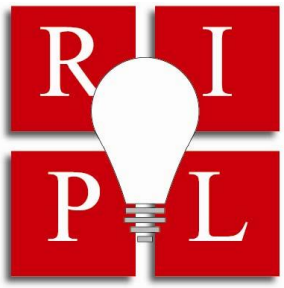
THE JOHN MARSHALL
REVIEW OF INTELLECTUAL PROPERTY LAW



AUTOMATIC DISCOVERY OF PRIOR ART: BIG DATA TO THE RESCUE OF THE PATENT SYSTEM

AMIR H. KHOURY & RON BEKKERMAN

ABSTRACT

In this research, we offer a fresh approach as to determining prior art. We do this by using Big Data methods. More specifically, we apply a model which constructs the semantic space of patents, in which *all* published patents and patent applications are arranged according to *semantic* similarities between each other. Our model provides a clear indication of how closely patents stand in relation to existing technologies, which we refer to as Near Inventions ("NI"). Our model exposes a certain level of deficiency when it comes to the disclosure, by patent applicants, of NIs. One conclusion which we draw from this approach is that there is no consistency among applicants when it comes to citing NIs. Another conclusion is that the more "densely populated" the semantic neighborhood of an invention is, the more rigorous the examination needs to be regarding its patentability.

Cite as Amir H. Khoury & Ron Bekkerman, *Automatic Discovery of Prior Art: Big Data to the Rescue of the Patent System*, 16 J. MARSHALL REV. INTELL. PROP. L. 44 (2016).

AUTOMATIC DISCOVERY OF PRIOR ART: BIG DATA TO THE RESCUE OF THE PATENT SYSTEM

AMIR H. KHOURY & RON BEKKERMAN

AUTOMATIC DISCOVERY OF PRIOR ART: BIG DATA TO THE RESCUE OF THE PATENT SYSTEM

AMIR H. KHOURY & RON BEKKERMAN*

I. INTRODUCTION

Disclosure of the relationship between the invention and existing technology has been inherent to the patent system. This continues on today. But, there has not been sufficient examination of the scope of that disclosure, that is, which existing technology is being disclosed and which is not being disclosed in patent applications. Here the modern Big Data Science toolset comes to the rescue. We model the semantic space of IP, in which all published patents and patent applications are arranged according to *semantic* similarities between each other. Our model provides a clear indication of how closely patents stand in relation to existing technologies, which we refer to as Near Inventions ("NI"). NIs are those inventions that are identified by our model as being in the semantic vicinity of a given patent application. We present a wide range of results obtained on a set of 34 multi-billion-dollar companies, each owning thousands of patents. We empirically prove that some companies are better than others when it comes to citing NIs in their patents. Our findings may have far-reaching consequences, especially for the companies on the low end of the NI citing spectrum. However, the goal of this article is not to point a blaming finger at any company. Our goal is to suggest a significant improvement in the system of patent application examination in a way that ensures that existing technologies and new inventions do not overlap. We believe that our Big Data model can make the patent system more efficient, more exact, and ultimately less costly in registration and potential litigation.

Our research is based on the concept that when examining a patent application, one should look at the semantic vicinity of the patent. That is to say, effective patent examination needs to inspect patents that are deemed to be of relevance to the examined technology, even if they do not necessarily directly overlap with it. We refer to these as Near Technology. In this context, we need to note that the "semantics," as it is referred to in this article is the analysis of words and their contextual meaning in order to discover relationships between patents which are essentially dealing with related subject matter that needs to be factored in when considering prior art. In this regards, Dratler alludes to the need to fix the broken patent system, part of should be "changing the substantive focus of patent law from abstract semantics to practical economic and commercial criteria amenable to adjudication."[1] Indeed, our approach is about utilizing Big Data in order to detect semantic proximities between patents so as to discover connections between patents and in so doing to identify NIs. Thus, in this research, we present our model, our findings and their implications on the patent

 Dr. Amir H. Khoury, Senior Lecturer, Faculty of Law, Tel Aviv University. Dr. Ron Bekkerman, Senior Lecturer, Faculty of Management, University of Haifa. Thanks to Olga Donin, Research Associate at the University of Haifa, for building the data infrastructure used in this research work.

[1] Jay Dratler, *Fixing Our Broken Patent System*, 14 MARQ. INTELL. PROP. L. REV 47 (2010).

system. We focus on the disclosure requirement, and the ramification of non-disclosure of NIs.

This article is comprised of two chapters. In the first chapter, we explain the importance of the prior art disclosure and why it is a crucial component in a viable patent system. In the second chapter, we describe our model and how it facilitates the mapping of patents in relation to their respective NI through the *CandorMap* system. We then apply our model for deep investigation of the way patents cite prior art.

II. ON THE SIGNIFICANCE AND DEFICIENCY OF PRIOR ART DISCLOSURE IN PATENT APPLICATIONS

Over the past three decades, it has become self-evident that patents are complex legal constructs, which are expensive to obtain and even more so to protect and litigate.[2] These problems plague the patent system not only in the United States but around the world.[3] This persistent and pressing situation is primarily owed to the structure of patents (and especially the patent claims section therein) and the way that patents interact in the technological space.[4] Simply stated, it has become exceedingly difficult to tell where one patent begins and where another ends. As such, patent registration, enforcement, and related litigation remain complex, costly and its outcomes are cast in doubt. Indeed, the cost of patents, in prosecution and litigation, is not a trivial issue. The empirical data provides evidence to the staggering costs of the patent system as far as inventors and patent owners are concerned. A 2013 survey by the American Intellectual Property Law Association, relating to the median litigation costs for patent infringement suits, reveals that the costs of patent litigation for claims in patents which were valued at under $1 million are over $800,000.[5] Furthermore, according to that survey, the median costs for patent litigation involving patents which are valued in the range of $1 million to $25 million, rose to $2.5 million.[6] In that survey, it was found that the median legal costs for patent litigation in patents

---

[2] James E. Bessen and Michael J. Meurer, *The Private Costs of Patent Litigation*, (2008), Boston University School of Law Working Paper No. 07-08, 2nd Annual Conferene on Empirical Legal Studies Paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=983736.

[3] Prof. Dietmar Harhoff, *Economic Cost-Benefit Analysis of a Unified and Integrated European Patent Litigation System*, Ludwig-Maximilians-Universität ("LMU") München Institute for Innovation Research, Technology Management and Entrepreneurship (INNO-tec) Final Report 26 February 2009 Tender No. MARKT/2008/06/D.

[4] In this regard, the most important section of any patent application and (patent) registration is the patent claims section. That section, which defines what is claimed by the inventor, is, essentially, the legal 'fence' that the inventor erects in order to protect their invention (and innovation). This fence metaphor is widely used in literature. *See, e.g.*, ALAN L. DURHAM, PATENT LAW ESSENTIALS: A CONCISE GUIDE, (2013) ("The function of patent claims is to identify the subject matter covered by the patent. If patent infringement can be compared to trespassing, the claims serve as the boundary markers that define what is, or what is not, an encroachment on the inventor's exclusive territory.").

[5] *See 2013 Report of the Economic Survey*, AMERICAN INTELLECTUAL PROPERTY LAW ASSOCIATION, *available at* http://www.patentinsurance.com/custdocs/2013aipla%20survey.pdf. It is important to note that the survey focused on the actual cost of fighting over the patent i.e. both as a defendant and as a plaintiff. However, the survey excluded the damages that a defendant would have to bear if s/he is not able to repel the case.

[6] *Id.*

valued in access of $25 million were over $5 million.[7] What is striking is that patent litigation is twice as costly as litigation pertaining to trademarks, copyright and trade secrets.[8] The cumulative sum of these costs is almost unimaginable. In this regard, a Podcast entitled techdirt reports that "patent litigation cost U.S. business about a trillion dollars in a quarter century."[9]

In our opinion, this reality is unacceptable—simply due to the fact that, by design, patents were supposed to be a tool for sharing knowledge and were never about excessive controls which sometimes seem to account for hoarding science.[10] Patents were intended to facilitate the sharing of knowledge and knowhow. They were intended to be an inclusive incentive-driven system and not what they have become—an exclusive cost-quelling construct.[11] Thus, in order to resolve these problems of complexity and cost, it is first imperative to bring to mind the basic rationales that underlie the patent system.

In its essence, a patent is a contract between the state and an inventor whereby if the inventor shares his knowledge with the world, and the world—or state—will reward him with a right over his invention for a limited period of time.[12] But this contract is not limited to its immediate parties, i.e. the inventor and the state. And its impact extends to encompass others that are not a formal party to said contact.[13] These 'external' aspects include the users, or consumers, of the technology as well as the competitors in the field. Notwithstanding their formal status, both of these 'silent' parties, meaning users and competitors, have an interest to get access to the technology with minimum costs attached. Thus, the patent contract is one that has repercussions beyond the formal two contracting, the inventor and the state. This state of affairs where many parties have a stake in any given patent application renders the patent contract a very complex endeavor that maintains a delicate social

---

[7] *Id*.

[8] For the full and detailed numbers in the survey see American Intellectual Property Law Association (AIPLA) 2013 Report of the Economic Survey, http://www.aipla.org/learningcenter/library/books/econsurvey/2013EconomicSurvey/Pages/. *See* IP Litigation Costs, WORLD INTELLECTUAL PROPERTY ORGANIZATION (Feb. 2010), *available at* http://www.wipo.int/export/sites/www/wipo_magazine/en/pdf/2010/wipo_pub_121_2010_01.pdf

[9] Glyn Moody, *Patent Litigation Cost US Business About A Trillion Dollars In a Quarter Century, Outweighing Benefits* (Apr. 27, 2014), https://www.techdirt.com/articles/20140416/04183626928/patent-litigation-cost-us-business-about-trillion-dollars-quarter-century-outweighing-benefits.shtml.

[10] Consider patent trolls as the most vivid reflection of the ugly side of the patent system.

[11] Andrew Grosvenor, *Why 'Patent Trolling' by High-Tech Companies is Stifling Competition & Innovation–And What we Should Do About It* (2011), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1923989. Grosvenor asserts that: "The purpose of the patent system is to encourage innovation and to reward inventors by protecting the fruits of their labor. Abuse of this sanctioned monopoly is helping to consolidate the tech marketplace to the few large companies that are winning the patent arms race.'"

[12] Shubba Ghosh, Patents and the Regulatory State: Rethinking the Patent Bargain Metaphor After Eldred, 19 BERKELEY TECH. L.J. 1315, 1349 (2004): "The metaphor of the patent bargain rests on a native view of social contract theory, based on questionable assumptions about private orderings that reduces patent law to a tool for protecting property rights."

[13] The most explicit case of the social obligation that entails patents relates to the question of access to patented medicines. *See* Amir H. Khoury, *The 'Public Health' of the Conventional International Patent Regime & the Ethics of 'Ethicals'*, 26 CARDOZO ARTS & ENT. L.J. 25, 25-70 (2008).

balance.[14] In the classic two-party contract, the parties are at liberty to draw the terms of the agreement and to assign to each other certain rights or obligations. But in the case of the patent contract, the state is basically called to act not only in a technical capacity—registering the invention—but, also as an entity whose task is to draw the line of distinction between the inventor's private domain and the public domain of the external parties.[15] Thus, patents involve an ongoing tug-of-war between the inventor that is seeking to maximize returns by expanding his monopoly over the technology and the silent parties who have a vested interest in ensuring access to the invention.[16] In between these polar interests (of rewarding the inventor and of ensuring access to technology), there is the never-ending endeavor to maintain the primary purpose of patent law: To promote the progress of science and innovation.[17] The patent system is not about dominating technology through overlapping technologies, but rather advancing innovation through filling the gaps in the innovation space. With that in mind, the importance of the disclosure of prior art is paramount.

Disclosure of prior art constitutes one of the primary obligations of the patent applicant. It is a precondition to receiving a valid patent.[18] Indeed, a lack of sufficient disclosure might effectively lead examiners into granting patents over inventions that otherwise should not have been granted.[19] This is not only unfair towards other innovators but also constitutes fertile ground for long and costly legal battles over the innovation precedence. The idea that accurate and thorough prior art should be disclosed to prevent the grant of "bad patents" sits well with the novelty condition wherein: Inventions that are not new are not patentable.

The inherent challenge of prior art disclosure in patents in not a secret. According to Kesan and Banik, in high technology areas such as biotechnology and computer software, the U.S. Patent and Trademark Office ("USPTO") is poorly informed about the relevant prior art.[20] In order to rectify this deficiency, Kesan and Banik proposed that the USPTO provide incentives to the patentee to perform a comprehensive prior

---

[14] Duncan Matthews, *Intellectual Property Rights, Human Rights and the Right to Health in Edward Elgar*, INTELLECTUAL PROPERTY RIGHTS AND HUMAN RIGHTS: A PARADOX, W. Grosheide, ed., 2009.

[15] *See id*.

[16] *See id*.

[17] This rationale was spelled out in the Constitution of the United States of America, wherein Article I, Section 8, Clause 8 of the Constitution, the United States Congress: "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."

[18] Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539 (2009). According to Fromer, Patent law is premised on the onward march of science and technology. Patent law encourages cumulative innovation, both by dangling the patent before the inventor as an incentive to invent in the first instance and by requiring him to disclose to the public his invention so that science can progress by building on the divulged knowledge. Patent disclosure is essential. It indirectly stimulates others' future innovation by revealing to them the invention so that they can use it fruitfully when the patent term expires and so that they can design around, improve upon, or be inspired by the invention both during and after the patent term.

[19] Jeffrey M. Kuhn, *Information Overload at the U.S. Patent and Trademark Office: Reframing the Duty of Disclosure in Patent Law as a Search and Filter Problem*, 13 YALE J.L. & TECH. 89, 90-139 (2010).

[20] Jay P. Kesan & Marc Banik, *Patents as Incomplete Contracts: Aligning Incentives for R&D Investment with Incentives to Disclose Prior Art*, 2 WASH. U. J.L. & POL'Y 023 (2000).

art search.[21] In fact, the picture in terms of disclosure of prior art is truly bleak, in this regard, Corinne Langinier and Philippe Marcoul shed light on the strategic non-revelation of information by patent applicants. They also highlighted the probability that patent applicants may conceal information. They explained that examiners tend to "make their screening intensity contingent upon the received information."[22]

Indeed, previous research indicates that applicants may cite related patents but do not always cover all relevant inventions. It is because of the failure that, technology that needs to be brought to the attention of the patent examiners never gets on their radar.[23] Although applicants are required to disclose all prior art that they are aware of, they are not obligated to conduct a predetermined type of search pertaining thereto.[24] In this regard, Atal and Bar attempted to classify the patentee's incentive to search for prior art by drawing a distinction between early state of the art search—conducted before R&D investment—and novelty search, conducted right before applying for a patent. Their research shows that search intensity increases with R&D cost, the examiners' expected search effort, and with patenting fees.[25] But while this might apply in an ideal patent system, the fact remains that it is a more rational option for applicants to cite less. Indeed, as Richardson points out: "applicants who systematically under-cite prior art stand to benefit."[26] That is because the examiners appear to invest less time on researching for prior art than on considering prior art that has been disclosed. As such, Richardson concludes that "applicants who cite less stand to have less time spent by the examiner during the application process on substantive evaluation, than on researching prior art."[27]

That is why, despite the importance of prior art and the role that should be played by the patentee, it is not possible to depend on the patentee to conduct an expansive search and disclosure of prior art. The system needs to devise a new mode of looking at data in order to ensure a more exact and relevant exploration of prior art. Another

---

[21] *See id*. Kesan and Bank argue that such incentives could accord a specific, high presumption of validity to the prior art (in post issuance litigation) that has been disclosed by the patentee (during patent prosecution), thereby limiting the use of the disclosed prior art for invalidation of the patent; *see also* Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539 (2009) (further discussion on the disclosure of prior art). According to Fromer, Patent law is premised on the onward march of science and technology. Patent law encourages cumulative innovation, both by dangling the patent before the inventor as an incentive to invent in the first instance and by requiring him to disclose to the public his invention so that science can progress by building on the divulged knowledge. Patent disclosure is essential.

[22] Corinne Langinier and Philippe Marcoul, *The Search of Prior Art and the Revelation of Information by Patent Applicants*, Review of Industrial Organization (2009).

[23] In Fromer's view, the disclosure of prior art indirectly stimulates others' future innovation by revealing to them the invention so that they can use it fruitfully when the patent term expires and so that they can design around, improve upon, or be inspired by the invention both during and after the patent term.

[24] Vidya Atal and Talia Bar, *Prior Art: To Search or Not to Search*, 28 INT'L J. OF INDUS. ORG. 5 (2010). Atal and Bar discuss the issue of patentability of innovations against the backdrop of existing prior art. They allude to the reality whereby while innovators have a duty to disclose any prior art (that they are aware of), they but have no obligation to conduct search for the same.

[25] *See id*. at 19-20.

[26] James H. Richardson, *Are Prior Art Citations Determinative of Patent Approval?: An Empirical Analysis of the Strategy Behind Citing Prior Art*, *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2557716.

[27] *Id*.

related element pertains to the method in which an applicant determines which inventions (and prior art) he needs to cite. Here too, the law is unclear and the mechanisms in place are insufficient.

In the next chapter, we shall explain why we think that NIs need to be explored and factored into a search seeking to discover prior art search and disclosure. We believe that NIs can harmonize the prior art search mechanism, and can bring to the examiner's attention many relevant prior inventions that need to be considered when conducting a patent examination. By applying patent mapping of NIs, the examiner would be able to better determine whether, to grant a patent, and the applicant would be able to better predict the chances of a clash with potential competitors or parent holders. In a nutshell, it is our conviction that our proposed method of mapping NIs can greatly improve how we discover prior art and novelty in patents.

III. MAPPING NEAR INVENTIONS

As we have showed above, what is missing from the patent examiners and applicant's tool box is a coherent and expansive model, that can process and predict which inventions need to be cited, and thus, brought to the attention of examiners. In this chapter we describe the NIs' discovery tool which can fill this void by providing better information about the innovation landscape that is in the semantic vicinity of new invention. Our model is intended to provide personalized 'maps' of the innovation landscape around new inventions. We believe that our model can be used by examiners to better determine the existence and impact of prior art. It can also be used by the inventor in predicting clashes with existing inventions. As such the NIs model boost the patent system's ability to identify relevant prior art. That is because, while prior art is a loose term that is open to interpretation, the NIs model provides a clear-cut, visual and contextual, tool for mapping of the innovative landscape around a new invention subject to a patent application.

A. The Idea of Employing a Recall-Oriented Search

We will start describing the NI model with presenting the analytical tool that underlies it, namely *CandorMap*.[28] This article is the first publication to introduce CandorMap, which is a Big Data analytics platform developed specifically for the Intellectual Property related analysis. CandorMap takes the approach that all the knowledge that is accumulated in Intellectual Property is in essence a large-scale dataset, suitable for automatic processing. Naturally, this also applies to patent data. Indeed, the text of patent registrations and patent applications published by USPTO over the past 40 years is a free-of-charge publicly available bulk dataset.[29] The dataset consists of textual records for each patent registration in a semi-structured format

---

[28] A commercial application of technology developed at the University of Haifa as a part of the CandorMap academic project is available *generally* CANDORMAP, http://www.candormap.com (last visited Aug. 24, 2016).

[29] *Patent Grant Red Book*, REED TECH, http://patents.reedtech.com/pgrbft.php (last visited Aug. 24, 2016).

composed of free-text fields[30] and metadata fields.[31] Overall, the CandorMap dataset consists of about 5.6 million patents and 1.7 million patent applications, all of which were reformatted into a unified, compact JSON representation.[32]

The majority of existing search engines, such as Google, Yahoo!, Baidu, Yandex, aim to efficiently find information most relevant for a search query. They do not, however, provide guarantees for the completeness of information found because of two reasons. The first reason is commercial: facing the tradeoff of providing most precise or most complete information, leading search engines choose precision over completeness as their typical user is interested in finding just *one piece* of relevant information. To access the CNN.com website, for example, the modern user searches for "cnn" on Google—just to save time on typing ".com". Needless to say, it is crucial for Google to present CNN.com as the first search result for the query "cnn," however, it is not economical to show *all* the results related to CNN.

The second reason for reluctance of leading search engines to invest in completeness of provide search results in technological: completeness-oriented search needs to deal with complex theoretical, practical, and pragmatic challenges that precision-oriented search technologies prefer to avoid. Suffice here to mention the notion of "relevance" which is subjective in nature wherein whatever is relevant for someone is not necessarily relevant for another. When a precision-oriented search engine deals with a query, it provides search results whose relevance was asserted by many previous searchers. Completeness--oriented technology faces a more daunting task: it needs to deal with the "subjectivity" of search, as it is supposed to provide not only most relevant results but also *less relevant* or marginally related—thus achieving a more comprehensive answer to the question at hand. In a technical parlance, completeness-oriented search technology is called *recall-oriented search*.[33]

In this reality of search, CandorMap is a pioneering, recall-oriented search system. In the patent domain, CandorMap focuses on searching for identical or similar technology. For this purpose, all the patents or patent applications relevant to a specific invention have to be identified, retrieved, and analyzed, in order to truly meet the disclosure obligation relating to prior art.

CandorMap comes in against the backdrop of current situation wherein it is practically impossible for an inventor, even if acting with the highest degree of diligence and good faith, to identify all previously patented related technologies. Consequently, the patent system is a constant state of flux between the formal obligation of disclosure and the practical limitations in executing said obligation. In other words, patent applications fail to cite a large portion of prior art despite the fact that patent data is publicly available and centrally organized. Notwithstanding

---

[30] These fields include title, abstract, brief description, detailed description, and claims.

[31] These fields include original assignee, filing date, publication date, U.S. classification, etc. The data format changed twice over the course of the 40 years: from 1976 to 2000 the data was stored in a "rich text" format which was specifically developed for the patent data. In 2001, the UPSTO adopted the SGML format that became obsolete and replaced by the XML format in 2005. The XML format has been in use by the USPTO ever since.

[32] Douglas Crockford, *The application/json media type for JavaScript Object Notation (JSON)* (July 2006), *available at* https://tools.ietf.org/html/rfc4627.

[33] *See, e.g.*, Walid Magdy & Gareth J.F. Jones, *PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications*, Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010).

applicants' failures, the worrying reality is that inventions get patented without a comprehensive prior art assessment which leads to an enormous amount of intellectual property related litigation, and in some cases, hesitation to file some applications.

CandorMap revolutionizes the concept of search by assuring comprehensiveness of search results. This is achieved due to its novel Big Data technology. The entire domain is organized into a *SemanticMap* wherein documents are like "towns" on the map, while "roads" represent semantic connections between the documents. Two documents are connected with an edge, or road, if they discuss the same topic[34]. As such, CandorMap is not a typical search engine. It is not based on an inverted index of documents that allows fast detection of query terms in the documents. Indeed, unlike typical search engines that interact with the users through short keyword-based queries, CandorMap is most effective when the query is a full-length document that describes the area of user's interest. In the intellectual property domain, the query document can be an existing patent, or any detailed textual description of the technology the user intends to investigate.[35]

A recall-oriented system such as CandorMap is a more suitable system for the research of prior art due to the following three attributes of the patent repository:

1. **Self-contained documents.** In the patent context, unlike other domains like Twitter or instant messaging, documents usually contain a sufficient amount of information to make an educated relevance judgment.
2. **Well-structured text.** Patents and other technical texts are not easy for a human reader. While their goal is clarity and unambiguity of presentation, their dry, monotonic style, extensive notation, and the complexity of grammatical constructions are serious obstacles for non-professionals. All this is not a problem for a computer system. Moreover, computers may be even better than humans at "understanding" the patent language because it maintains structure (such as *claims*, *abstract*, and *description* of a patent) that is usually consistent in terminology and notation. And it is mostly clean of misspellings and grammatical shortcuts, and practically never uses literary elements such as irony, metaphor, and allegory.
3. **Closed domain.** Recall-oriented search would not be possible in highly dynamic domains where millions of documents can be instantly added or removed. Patents are not such an arena: U.S. patents are issued on a weekly basis, which gives the system plenty of time for updating itself.

---

[34] Note that a typical document covers a few topics. In order to be considered similar, two documents do not have to share all their topics—they would rather have at least one topic in common.

[35] Here a recall-oriented search system cannot be implemented as an inverted index, primarily because the notion of a query is fundamentally different in recall-oriented search. If a user seeks all the available information on a particular topic, a search query of a few keywords would not help as those words cannot define the scope of the topic the user is interested in. As such, Intellectual Property is not the only domain in which CandorMap technology can prove extremely useful. Other domains include scientific publications (in the flavor of "automatically filling up the related work section of my paper"), legal and medical documents, insurance claim files, etc.

B. The Mechanics of Recall-Oriented Search

Having explained the idea behind a recall-oriented search of patents, we now turn to show how that can actually contribute towards better identifying prior art in the patent registration process. The process of searching is a matter of positioning the query document on the Semantic Map, or the domain of the search. These documents are already connected to other documents in the Semantic Map. This creates a "semantic neighborhood" around the query document which then needs to be traversed by the user who would decide which neighboring documents are relevant for the query document and which are not.

Conversely, if a given document is not in the semantic neighborhood of the query document, it simply cannot be relevant for the query document. This is deduced with high certainty given the topical locality property of the natural language: it is impossible to describe two similar concepts using two completely different vocabularies. For example, it is impossible to describe a new polymer exclusively in baseball terminology. Although the extended semantic neighborhood of the query document is supposed to contain all the relevant documents, by no means is every document in the neighborhood supposed to be relevant. Since the notion of relevance is subjective, the system cannot decide which aspect of a domain document would make it relevant for a specific query document. Moreover, the system cannot learn a high-quality relevance model from the usage data because even the same user can consider different aspects of relevance while looking at the same semantic neighborhood at different projects.

Thus, instead of proposing a one-size-fits-all machine learning solution, which does not appear to be a viable option, CandorMap takes a data visualization approach. It offers a novel Graphical User Interface ("GUI") that lets the user traverse the semantic neighborhood of the query document, while making ad-hoc relevance judgments. Since the semantic neighborhood is naturally organized in a topological composition of document clusters, it would often be enough to make a relevance judgment of a representative document from each cluster, to form an opinion about all clusters. If a cluster is considered non-relevant, it would be unnecessary to traverse other clusters located further down from it—away from the query document—as they are very likely to be non-relevant as well.

The construction of the Semantic Map may be extremely time-consuming. Given 7.3 million patents and patent applications published by the USPTO over the past 40 years, the construction of the Semantic Map in the most straightforward manner would have to involve checking semantic similarity of every pair among the 7.3 million documents (i.e. over 26 trillion pairs). Assuming that each document is represented as its Bag-Of-Words, and taking into account that the documents can be as long as 25 megabytes of text, it is safe to assume that a similarity comparison can take on average one millisecond on a modern machine. To complete performing over 26 trillion similarity comparisons at a one millisecond rate will take almost 850 years. At such a scale, the most sophisticated cloud computing infrastructures are unlikely to help. CandorMap is able to construct the entire Semantic Map of 7.3 million documents in under twenty hours on a cluster of fifty high-performance machines. This became possible due to a novel algorithmic solution developed for this task. CandorMap algorithms are based on two main principles:

1. **A document representation should be compact.** The standard Bag-Of-Words representation, while being one of most compact representations available, is not compact enough. CandorMap maintains a terminology-based document representation that compresses a megabyte-long document into single kilobytes.[36]
2. **Not every pair of documents should be checked for similarity.** An aggressive filtering protocol allows the system to compare significantly fewer pairs while trading off an insignificant loss of quality.[37]

To summarize, CandorMap is the first commercial tool to solve the fundamentally difficult problem of recall-oriented search while implementing the following three innovative approaches:

1. Domain representation as a Semantic Map, instead of an inverted index.
2. Interactive GUI for traversing the Semantic Map, instead of an application of a pre-built relevance model.
3. Focused similarity comparison protocol for constructing the Semantic Map, instead of an exhaustive all-pair comparison scheme.

C. Application of CandorMap to Patents: Empirical Results

Thus far, we have showed the problem of the inherent weakness of prior art citations and we have described the possibility of invoking recall-oriented search utilizing the CandorMap system. In this section, we will show that the proposed model does in fact have value to the patent system. Indeed, CandorMap is not only a useful tool in revamping prior art searches, but it is useful for revealing the deficiencies in patent citations.

Our starting presumption is that since CandorMap maintains a list of most semantically related patents for *each* patent issued by the USPTO over the last forty years, it would be reasonable to conclude that the USPTO patents should cite at least some of their semantically related patents as suggested by CandorMap. This assumption, on its own, might have been inherently weak had it not been for the fact that patentees do cite patents from the semantic neighborhoods constructed by CandorMap. That is to say, the CandorMap results indicating sematic relevance do indeed generate results that are deemed relevant by the applicants themselves. For example, Patent No. 9,112,724 (issued by Cisco) cites 11 U.S. patents, all of which are located in CandorMap's semantic neighborhood of Patent No. 9,112,724.[38] From the outset we would also like to acknowledge the fact that no matter how effective it is, CandorMap is very unlikely to generate a list identical to the list of patents' forward and backward citations, because patents tend to cite well-cited patents, instead of citing most closely related patents.

---

[36] Evgeniy Gabrilovich & Shaul Markovitch, *Computing semantic relatedness using Wikipedia-based explicit semantic analysis*, Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07) (2007).

[37] Elsayed, Tamer, Jimmy Lin, & Douglas W. Oard, *Pairwise document similarity in large collections with MapReduce*, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (2008).

[38] U.S. Patent No. 9,112,724 (filed Dec. 1, 2009).

With this in mind, we now turn to show how in fact the results produced by CandorMap do indeed reflect a deficiency in the citation of patents that are deeded to be part of the NIs and should be cited as a matter of course. To show this empirically, we designed the following experiment.

We constructed a list of over a million patent owning entities, which included companies and inventor groups. For each entity, we constructed a list of all patents that the company issued, as their original assignee, over the past forty years, as reported by the USPTO. For each such patent, we constructed a list of all its backward citations, and then compared it with the list of semantically related patents generated by CandorMap. Having done so, we then narrowed the list of patent owning entities to 34 companies that meet the following two criteria:

1. Each company being an original assignee on a large number of patents with a threshold at 3,000 patents per company. The main reason for concentrating on large patent holders only is diminishing the stochastic effect and obtaining statistically significant results.[39]
2. For each company, CandorMap is "confident" about what the company needs to cite in its patents, where the confidence is defined as follows. CandorMap is deemed to be *confident* about what a patent (*p)* needs to cite if CandorMap managed to construct a *confidence list (*C_p*)* of at least ten patents each filed earlier than p and each having a high similarity to p. The similarity is considered high if it is above a certain threshold chosen such that for each two patents with the similarity score above this threshold, human examiners would likely agree that those two patents disclose similar technologies. We determined that CandorMap is confident about what a company needs to cite in their patents, if CandorMap is confident about more than 50% of the company's patents.

It is important to note that only the chosen thirty-four companies satisfied the two criteria discussed above. No additional filtering was done on the list of companies.

We say that a patent p is *successful* about citing its NIs if p cites at least one NI (i.e. a patent from the C_p that CandorMap constructed for p). For each company, we report on the percentage of patents successfully citing NIs out of all its patents that CandorMap is confident about what they need to cite. Note that patents can cite other patents for a variety of reasons, and not all patents' citations have to belong to the CandorMap's confidence list. On the other hand, all patents from the confidence list C_p are closely semantically related to p, and may thus be cited in p. We cannot claim that every patent from C_p should be cited in p, because we have no way to prove it. Therefore, we take a statistical approach: since C_p consists of minimum ten patents, chances are tenfold that at least one of them should be cited in p.

[39] We can easily add more companies to our pool. In fact, if we lower the threshold on the patent portfolio size from 3000 to 300, we expand our pool from 34 to 313 companies, for each of which we have high-confidence results.

Figure 1, below, is a visual illustration of the design choices we made in this experiment described above. Note that for some patents of a company, CandorMap is not confident about what they need to cite. This does not necessarily mean that CandorMap failed to come up with good suggestions—it just means that CandorMap came up with fewer than ten such suggestions. Since we cannot make a strong case for those patents, we will ignore them in the rest of this analysis.

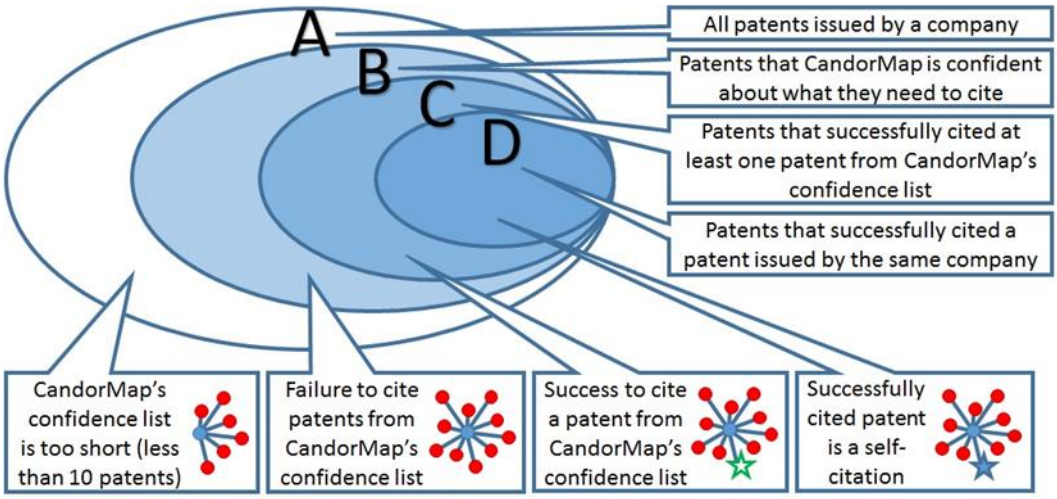


Figure 1: Design choices made in experiment[40]

[40] We chose companies with Size(A) > 3000, and Size (B)/Size (A) > 0.5.

Figure 2, below, shows the results of our experiment. What follows is an analysis of the main results obtained in our experiment.

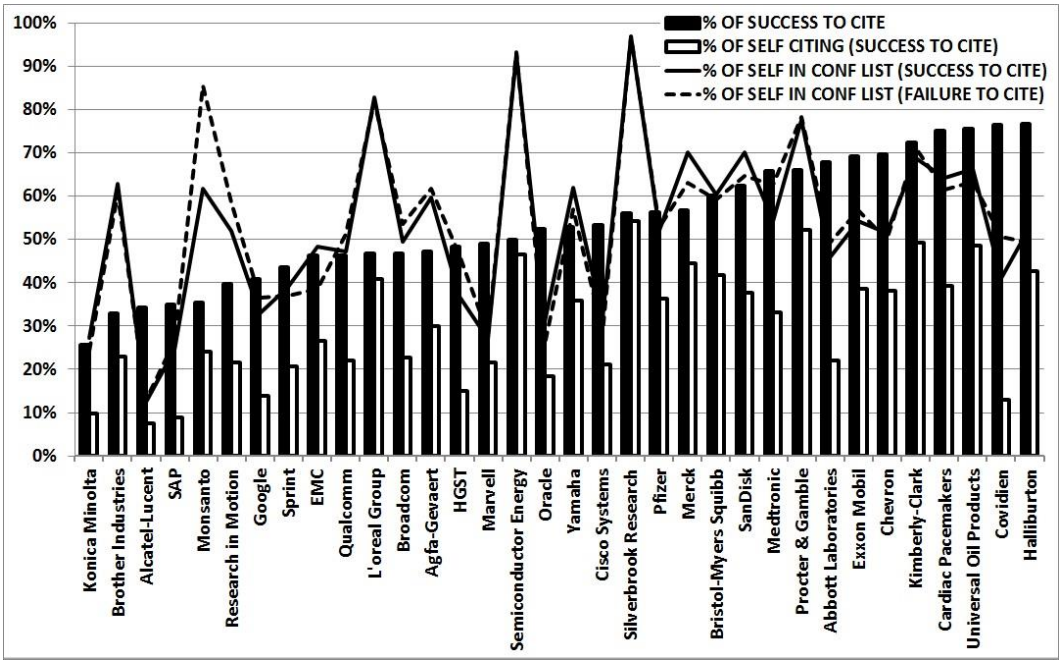


Figure 2: Results[41]

The black bars on Figure 2 represent our main result: for each of the thirty-four companies we chose, we show the percentage of patents that succeed to cite at least one patent from the confidence list constructed by CandorMap. It is striking that no company managed to get close to 100% success. The highest percentage was achieved by Halliburton; 76.8% of its patents managed to cite at least one patent from CandorMap's confidence list. On the other hand, no company demonstrated poor results. The lowest percentage of success is 25.6%, obtained by Konica Minolta; a quarter of its patents are in agreement with CandorMap's suggestions.

Before we move on to discussing other results represented in Figure 2, let us keep considering the black bars. Crucially, the wide range their spectrum (from one fourth to three fourths) leaves no doubt as to CandorMap's credibility. As alluded to earlier, if CandorMap had produced completely irrelevant suggestions, the degree of synergy between its finding and those of the patents that were actually cited by the thirty-four companies would have been nearly zero, and the overlap between their patents' citations and CandorMap suggestions would have been negligible. But that is not the case. In fact, each one of the selected companies shows at least 25% success. This clearly demonstrates that CandorMap does indeed produce relevant suggestions. Moreover, the fact that some companies stay in over 75% agreement with CandorMap

[41] The success of citation is depicted in black bars, which is Size(C)/Size(B) in Figure 1. The self-citation success rate can be seen in white bars, which is Size(D)/Size(B) in Figure 1.

strongly implies that CandorMap's suggestions are of remarkably high quality. Four companies, working independently from each other and from CandorMap and each implementing its own intellectual property strategy, managed to agree with CandorMap in at least 75% cases. Nineteen companies agreed with CandorMap in at least 50% cases.

Furthermore, it is very interesting to note that on the left side of Figure 2 (companies with low success rate) is mostly populated by high-tech companies (electronics and software), while the right side of Figure 2 (companies with high success rate) is mostly heavy-tech, like pharmaceutical, medical equipment, and oil companies. This is not surprising though, as heavy-tech companies are supposed to invest more resources in issuing patents than high-tech companies do. In high-tech, if a patent gets disputed, the company is likely to survive. In heavy-tech, in contrast, if a patent gets disputed, the company might get out of business.[42] The right-most high-tech company in Figure 2 is SanDisk with a success rate of 62.5%. SanDisk is a computer storage company known for its well-defined intellectual property strategy and the company has recently been acquired by Western Digital for $19B. [43] The left-most non-high-tech company is Monsanto, a 45-billion-dollar agrichemical corporation, with a 35.3% success rate.

In order to further establish this clear rift between high-tech companies vis-à-vis heavy tech companies, we adopt the hypothesis that companies on the right side of Figure 2 are indeed investing more resources into research and analysis of the prior art for the technologies they develop, comparing to companies on the left side. To support this hypothesis, we need to exclude other factors that might affect the higher success rate of right-side companies. One of those factors may be the percentage of self-citations. It is substantially easier for a company to cite prior art created by itself rather than by any other company. The reason for this is quite obvious: the company is supposed to be aware of its own patents, while it might not be familiar with patents belonging to another entity which develops similar technologies. To cite someone else's patents, the company needs to perform extensive patent search which takes time and resources, while self-citation is fairly cheap.

It may be that companies on the right side of Figure 2 have high success rates because they are mostly citing their own patents. To assess this statement, we show the white bars on Figure 2, which correspond to the percentage of self-citing patents among those that CandorMap is confident about. Note that if a patent cites a few patents from CandorMap's confidence list, we consider it as self-citing if at least one of those citations is a self-citation. As we can see, there is no visible correlation between the amount of self-citations and the success rate of companies. Indeed, the Pearson correlation coefficient is 0.12 which indicates no correlation at all.[44] Our conclusion is

---

[42] For example, Paragon Trade Brands, one of the leading diaper producers in the 1990s, went bankrupt in 1998 following a patent dispute loss.

[43] *See, e.g.*, Tomoko H Ogura, *Intellectual property strategy: analysis of the flash memory industry*, Massachusetts Institute of Technology (2006).

[44] A *correlation coefficient* is a coefficient that illustrates a quantitative measure of correlation and dependence, which is a statistical relationships between two or more random variables or observed data values. The Pearson correlation coefficient is a correlation coefficient that measures the strength and the direction of the linear relationship between two random variables.

that right-side companies are more successful with citing NIs regardless of the amount of self-citations in their patents.

With that being said, two companies provide a surprising exception, namely, Abbott Laboratories and Covidien, both operating in the space of medical devices. Despite the fact that both of these manifest a very low rate of self-citation they are still very successful in citing near inventions. Since the majority of their successfully cited patents do not belong to themselves, we believe that both companies are very particular about searching and studying prior art before disclosing their inventions.[45] Our findings can be used to rationalize this conclusion, as we see that Covidien apparently maintains an exceptionally strong patent portfolio.

Another thing that stands out pertains to Semiconductor Energy Laboratories and Silverbrook. Both companies appear to mostly self-cite. Remarkably, these are the only two companies among the thirty-four companies we chose that are invention licensing companies (i.e. they do not develop their inventions but rather license their intellectual property). Not surprisingly, the two invention licensing companies file patent applications that are semantically similar to each other, aiming at building large patent portfolios in the areas of interest, so the vast majority of NIs of their patents belong to themselves.

This observation led us to infer that if the semantic neighborhood of a patent mostly consists of patents of the same company, then citing NIs is an intrinsically simpler task. Thus, after we have checked whether the existing citations are self-citations, we also considered whether a company's patents have a higher chance to cite patents from CandorMap's confidence lists. Indeed, a patent p will have a higher chance to cite NIs if most of them belong to the owner of p.

Our results are depicted by the solid line and dashed line in Figure 2 above. These two lines show the average percentage of same-owner NIs of a company's patents that were successful (solid line) or unsuccessful (dashed line) in citing NIs. In this regard, the percentage of same-owed NIs provides an indication of "friendliness" of semantic neighborhoods around patents of a company: the higher their percentage of same-company patents is, the "friendlier" the neighborhoods are for the company's patents – and thus the easier it should be for the company to cite those NIs (as they are more aware of their own patents).

Indeed, there appears to be a very high correlation between the semantic neighborhoods' friendliness (solid line in Figure 2) and the level of patents' self-citation (white bars). In fact, their Pearson correlation coefficient is 0.91. In light of this, we had initially expected to see the solid line being mostly above the dashed line, as "friendly" semantic neighborhoods would likely mean the success in citing NIs. However, we did not see this phenomenon manifested in the results. In fact, in most cases, both lines practically overlay each other, which suggests that the success in citing NIs does not depend on the "friendliness" of semantic neighborhoods.

The friendliest semantic neighborhoods (above 90%) are of patents that belong to Semiconductor Energy Laboratories and Silverbrook Research, for the reasons discussed previously. What is less expected though, is that the second place in the semantic neighborhoods' friendliness (about 80%) is taken by L'oreal Group and Procter & Gamble, both well-known personal care product manufacturers. Since every

[45] It is worth mentioning that Covidien has recently been acquired by Medtronic (which is one of the 34 companies we chose) in a 50-billion-dollar deal.

four out of five patents in the semantic neighborhoods of their patents belong to themselves, we can conclude that both companies built tightly connected patent portfolios, which is always good for protecting the company's intellectual property. What is striking is that despite the similar strength of their portfolios, Procter & Gamble, with a success rate of 66.2%, is significantly more successful at citing NIs than L'oreal whose success rate is is 46.8%.

The level of friendliness of semantic neighborhoods (solid line) positively, though very slightly, correlates with the companies' success in citing NIs (black bars), with a Pearson correlation coefficient of 0.35. Some companies, however, do not follow this trend. Compare Brother Industries and Yamaha, both Japanese electronics manufacturers. Semantic neighborhoods of their patents are fairly friendly for both companies (about 60%). Nevertheless, while Yamaha is strictly in the middle of Figure 2, Brother Industries is too far on the left with only a 32.9% success rate.

Furthermore, we observe a remarkable similarity between the patent portfolios of Marvell, Oracle, and Cisco. All three are large high-tech corporations. although Marvell is rather smaller than the other two. Their patents seem to exist in unfriendly neighborhoods below 30%. But, all three are good at citing NIs (around 50% success rate). In fact, all three of them are on the far right side of the high-tech spectrum. This implies that the three corporations invest a similarly successful effort in researching prior art of their inventions.

Two other large high-tech companies, Qualcomm and Broadcom, show similar characteristics as well. However, both of them are slightly below Marvell, Oracle, and Cisco in terms of citing NIs (around 46% success rate), while their semantic neighborhoods are substantially friendlier (around 50%). It appears, according to our data, that these two companies have some room to improve their prior art citations.

Another striking similarity is between Exxon Mobil, Chevron, and Halliburton – all leading heavy-tech (oil) corporations. The friendliness of their patents' semantic neighborhoods is pretty much the same (about 50%), while their citation success rate is excellent: it is almost identical for Exxon Mobil (69.2%) and Chevron (69.7%), while the smallest of the three enterprises, Halliburton, shines with 76.8%.

Curiously, the far left edge of the spectrum is comprised of five companies that are quite unfortunate to have the citation success rate below 36%, creating a substantial gap in citation quality from the other companies depicted in Figure 2. In this group of low citing, Konica Minolta, shows the worst citation performance, at the 28% relative gap down from the next company in line. Alcatel Lucent, one of those five low citing corporations, apparently has the most unfriendly semantic neighborhoods (12.5%).

The fifth company from the left is Monsanto, the lowest scoring company among all non-high-tech companies in our survey. Monsanto's performance appears anomalous. We can see a significant gap between the solid and the dashed lines, while the dashed line is way above the solid one. This implies that the semantic neighborhoods of patents that managed to cite NIs have significantly fewer Monsanto patents (61.7% on average) than the semantic neighborhoods of patents that failed to cite NIs (those contain on average 85.6% Monsanto patents). Monsanto only cites NIs in 35.3% of cases, which means that in 64.7% cases the semantic neighborhoods of Monsanto patents almost exclusively consist of Monsanto patents, a large portion of which do not appear to be cited.

D. Validation of the Results

In the previous section, we demonstrated through our analysis that many corporations manifest a very serious deficiency when it comes to citing prior art, even when that prior art relates to self-owned patents. But in order to make sure that this data does indeed provide a trustworthy result, we need to perform one crucial validation test. We consider lengths of the respective confidence lists that CandorMap provides, in order to see whether or not the lengths of said lists correlate with the rate of NI citations. In other words, could the differences in NI citation rate amongst the thirty-four companies emanate from the fact that confidence lists for some companies are simply longer? Indeed, Figure 2 shows that there are more heavy-tech companies on the right side of the spectrum, which prompts the question: What if heavy-tech companies all build tighter patent portfolios which prompt CandorMap to construct longer confidence lists for them, subsequently increasing the changes of citing at least one patent from the longer confidence lists?

Figure 3 shows that our model continues to hold water here. Black and white bars on Figure 3 show the average lengths of confidence lists per company, where the companies are sorted exactly as in Figure 2. The black bars are the confidence list lengths for those patents of the company that succeed at citing NIs, while the white bars are confidence list lengths for those patents that fail. Figure 3 provides us with a validation of our model, through three observations:

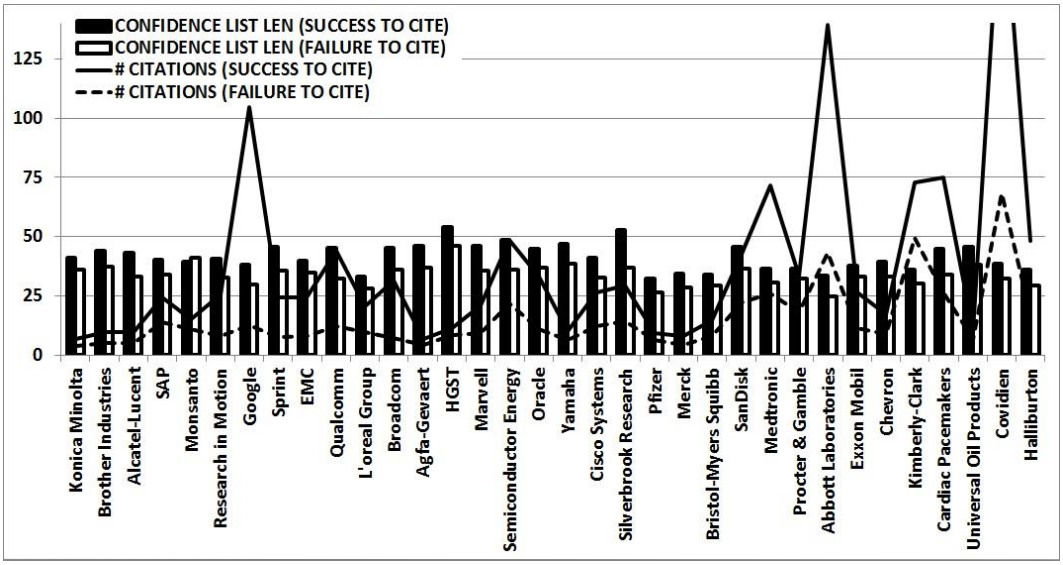


Figure 3: Lengths of confidence lists and numbers of citations, per company

1. Quite predictably, the white bars are lower than the black bars in almost all cases, which means that confidence lists of patents that succeed to cite NIs are on average longer than those of patents that fail to cite NIs.[46]
2. The confidence lists are all in the range of between 26 patents (for Pfizer) to 54 patents (for HGST) on average—longer than the minimum threshold of 10 patents that we predefined.
3. There is no visible increase in the confidence list lengths towards the right side of the plot. In fact, there is a very slight negative correlation between the confidence list lengths and the success rates (black bars from Figure 2), which means that there is a very slight decrease in the in the average confidence list lengths towards the right side of the plot. Indeed, the Pearson correlation coefficient between the black bars on Figure 2 and the black bars on Figure 3 is -0.25, and it is -0.32 between the black bars on Figure 2 and the white bars on Figure 3. This finding contradicts the hypothesis that patents of right-side companies are just by pure chance more successful at citing NIs.

A further test for validating the results can be presented through the following question: Is the success rate of citations contingent on the number of citations in a company's patents? Intuitively, the more patents a company cites, the higher probability is to cite relevant NIs. Let us clarify though, that there is nothing unfair in citing a lot of prior art. The USPTO does not impose any upper limit on the number of citations in an invention disclosure, as long as those citations are relevant for the disclosure. In certain cases, a disclosure has hundreds or even thousands of relevant prior works to cite.[47]

Solid and dashed lines on Figure 3 show the average number of citations per patent (out of the patents that CandorMap is confident about)—for the thirty-four companies we chose. The solid line is the average number of citations in patents that succeed in citing NIs, and the dashed line is the average number of citations in patents that fail to cite NIs. Quite predictably, the dashed line is always below the solid line. This time the variance is substantial, meaning there is obviously more chance to see NIs among longer lists of citations.

Here then, we observed a moderate positive correlation between the success rate (black bars in Figure 2) and the average number of citations. The Pearson correlation coefficient between the black bars in Figure 2 and the solid line in Figure 3 is 0.46, and it is 0.56 between the black bars in Figure 2 and the dashed line in Figure 3. The correlation is surprisingly higher for patents that fail to cite NIs over those that succeed. This can be explained by an unexpected hump in the solid line on the left side created by Google (over 104 citations on average in patents belonging to Google). A close investigation revealed that this hump is somewhat contingent on a group of 56 Google patents, each of which citing over 1,500 other patents, which substantially, and disproportionately, increases Google's average citation rate.

Another surprise relates to Abbott Laboratories and Covidien—patents of both companies (those that succeed to cite NIs) have a very high citation rate: it is 139

---

[46] The only exception is the confidence lists of Monsanto—we discussed its anomalous performance above.

[47] U.S. Patent No. 8,892,495 (filed Nov. 11, 2014), for example, cites over 5,800 U.S. patents and patent applications.

citations on average for Abbott, and striking 205 citations on average for Covidien—this number is so high that it went 1.5 times above the range on Figure 3. As we discussed before, both companies have a surprisingly low self-citation rate, so both of them are mostly citing patents of other companies, which requires a deep research of prior art technologies.

Another point worth mentioning is an extremely low citation average for Universal Oil Products. There were only ten citations on average in patents that succeeded to cite NIs. This is very surprising given Universal Oil's very high success rate (75.7%). And as we will see in Figure 4 below, they managed to cite not only one NI, but three NIs on average.

The final validation of our results comes by answering the following questions:

1. We decided that a patent is successful at citing near inventions if it cites at least one Near Invention, but do patents actually cite more than one NI?
2. For each patent, CandorMap constructs a confidence list of NIs sorted by similarity to the patent. So far, we ignored the similarity scores, while simply saying that they are high enough. However, the way NIs are ranked on the confidence lists is an important piece of information: obviously, the higher the NI is ranked, the more relevant it is supposed to be to the patent. Do companies manage to cite highly-ranked NIs?

Figure 4 clarifies those points, and provides a positive response for both.[48] The bars on Figure 4 show the average number of NIs cited in patents of the thirty-four companies, sorted exactly as in Figure 2. The line shows the average rank (in the confidence list) of the top-ranked NI, per company.

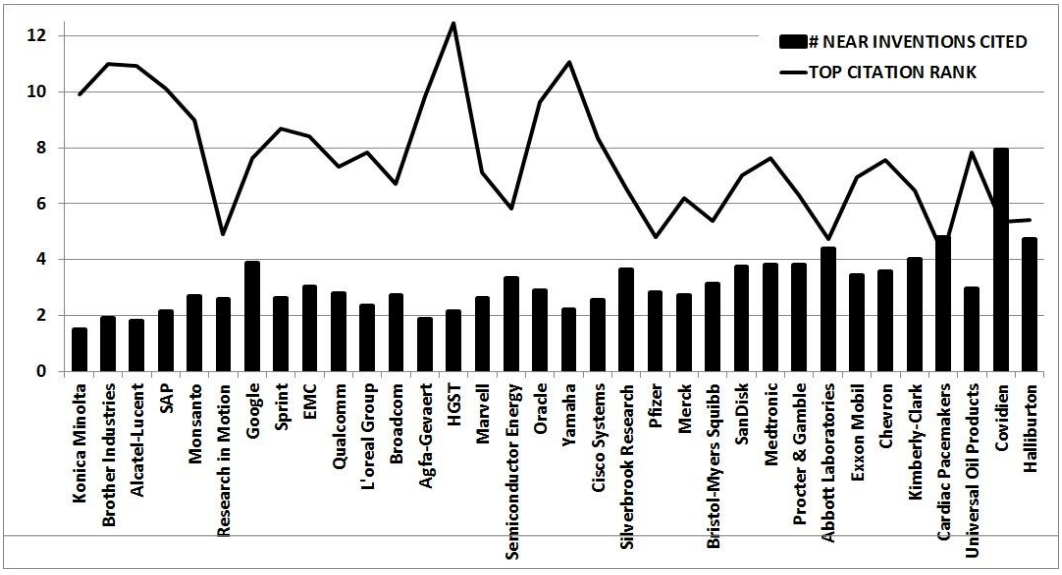


Figure 4: Statistics on citing near inventions

[48] Note that Figure 4 considers only patents that successfully cite Near Inventions.

Apparently, bars become longer towards the right side of Figure 4. Indeed, the Pearson correlation coefficient between the average number of NIs cited (bars on Figure 4) and the success rate (black bars on Figure 2) is 0.73. Notably, patents of most companies cite on average between two and four NIs, with a few exceptions, most prominent of which is, of course, Covidien, with a stunning eight NIs cited on average by each of their patent. This is not very surprising though, given the average of 205 citations Covidien patents have. Four companies on the left side of Figure 4 keep their NI citation averages under 2, with the minimum of 1.5 citations belonging to Konica Minolta.

The line in Figure 4 shows the average rank of the top-ranked NI cited by a company's patents. The closer this rank is to one, the more relevant the cited NI is supposed to be. If a patent cites a NI that is the first in the confidence list constructed by CandorMap, this would mean that the patent manages to cite the most relevant NI, as suggested by CandorMap.

In fact, our expectations are validated here as well, wherein we observe the negative correlation of the top rank of cited NI (the line in Figure 4) and the success rate (black bars of Figure 2): the Pearson correlation coefficient is -0.59. The negative correlation means that patents of companies on the right side of the spectrum tend to cite NIs ranked higher in CandorMap's confidence lists, so the right-side companies are in a higher agreement with CandorMap about which NIs are relevant. A prominent example is Cardiac Pacemakers, the fourth company from the right, with more than four NIs cited by its patents on average, out of which the top-ranked NI is at the average fourth place in the confidence list.

IV. CONCLUSION

Operative trends that appear to stem from our analysis are that:

1. There is a clear deficiency in the citation of prior art,
2. The deficiency is apparent even if the prior art is self-owned,
3. Heavy-tech companies are up to three times more successful at citing NIs than high-tech companies,
4. Heavy-tech companies tend to have more citations on average (up to thirty-one times more) than high-tech companies,
5. Heavy-tech companies tend to cite more NIs on average (up to five times more) than high-tech companies,
6. Heavy-tech companies tend to cite more relevant NIs than high-tech companies, and
7. Heavy-tech companies do not tend to self-cite more than high-tech companies do.

These finding provide a glaring indication that patent filings and citations therein are not similar across the board and that companies cannot be treated in a similar manner. Indeed, all the above aims to highlight the deficiencies in the way that some companies cite prior art and to call for revamping the disclosure mechanism for prior art with a view to simplifying the patent landscape by evading the needless overlapping of technology due to the lack of sufficient disclosure of near technology. The persisting reality remains that there is a deficiency in the citation of Near

Inventions that are themselves a significant part of prior art. What is of grave concern here is that said deficiency is also manifested in that fact that many corporations are not citing relevant prior art that belongs to the same company. With that being said, this research cannot provide the reasons for this deficiency. This remains beyond the scope of our research.

Thus, in conclusion, we believe that NIs should be considered in the context of searching for prior art. Absent this inclusion, the patent system will remain inefficient, because prior art limited in its scope can potentially miss relevant segments of related technology that needs to be considered. We argue that by factoring in these NIs by way of disclosure, it is possible to create a more precise process of patent examination which ultimately would be more beneficial to the progress of innovation. By applying the NI concept, the patent system would be rendered more efficiently and less costly and thus it can retain its relevance as a tool for prompting innovation and for sharing knowledge.

We would like to emphasize that all the insights presented in this paper were obtained by applying novel Big Data methodology to publicly available data. As much as this work appears unique and non-canonical, the availability of the USPTO data allows application of other Big Data methodologies, which will undoubtedly lead to other interesting insights. We believe, that our work opens the door for the utilization of Big Data research in the intellectual property domain. As such, our research is only an initial demonstration of Big Data capabilities.