Sean Walker

Professor Xiyin Tang

CPSC 184

8 May 2019

Final Project: An Examination of Patent Complexity and its Impact on Trolling

INTRODUCTION

Although we didn't spend too much time discussing patent trolling explicitly in class, I decided to explore the idea and trends of its occurrence in my final project. I mainly decided to do this after finding out about Stanford Law School's newly released database of non-practicing entity (NPE) litigation cases in the 21st century.[1] The full database contains more than 43,000 fully coded and described patent lawsuits from 2000 to 2018. The Stanford team released a paper alongside the initial database announcement which performs some analysis about trends in patent trolling, performed by breaking down and coding different stakeholders as NPEs and patent assertion entities (PAEs), as well as classifying different stakeholders into 13 specific classes of patent holders (*e.g.* individuals, startups, governments, non-profits).[2]

---

* All source code available at https://github.com/seankwalker/cpsc-184-project. I'd like to extend my sincere thanks to Shawn Miller and his team at Stanford Law as well as Eren Orbey, a fellow Yale senior, whose thesis I happened across. Seeing projects by these authors greatly inspired this project. Also, I'd like to (of course) thank Xiyin and Allison for a wonderful semester of IP law.

[1] The database's project website is https://law.stanford.edu/projects/stanford-npe-litigation-database/. From there, one can submit a form to receive a random sample of 20% of the data in the full database. The full database was just made accessible this month and can be found at http://npe.law.stanford.edu/. Note that an account must be registered (it took under an hour for mine to be approved) for full usage of the latter website.

[2] Miller, Shawn P. "Who's Suing Us: Decoding Patent Plaintiffs since 2000 with the Stanford NPE Litigation Dataset." *Stan. Tech. L. Rev.* 21 (2018): 235.

PROJECT GOALS

My goal was to investigate how the patent trolling trends found in the Stanford Law database and paper overlap with trends in the patents themselves. Specifically, I decided to use text readability as a measure of patent complexity, and see how that correlates with patent trolling, volume of patent-related litigation, and so forth. My hypothesis, or at least intuition, is that patents are becoming more complex with the passage of time. I would also posit this makes patent trolling and litigation more common, because patent holders may have a harder time distinguishing subtleties which would qualify something as a patent violation or not. The same applies to courts—if courts are faced with more complex patents to parse and comprehend (in order to evaluate patent lawsuits), I expect that patent trolling may be seen as more lucrative, or at least more feasible. In fact, even if the case is settled rather than going to court to receive judgment, one might consider more complex patents a reason behind the alleged offender deciding to settle.

DESIGN

With the goal of evaluating patent readability, I first sought a data source to examine. The US Patent and Trademark Office (USPTO) has a variety of online tools for viewing historical patent data in bulk. Additionally, Google hosts a mirror of patent data until 2015, due to an agreement made with the USPTO. Data since 2015, however, is only available directly through the USPTO.

Dataset Selection

I decided to use the USPTO's Granted Patents dataset, which is a collection of granted patents from the 1970s to 2018.[3] Other research in this area seems to compile (via web scrapers) individual XML

[3] Available at http://www.patentsview.org/download/, under "patents."

files released on the Google mirror until 2015; because I wanted to compare the analysis I do to that of the Stanford paper, I needed data from a time period broader than 2006-2015.[4]

Methodology

With my dataset selected, I began to think about how I wanted to parse the data. As I was going through preliminary stages, it became clear that much of the data provided in the file from the USPTO was faulty; either values for fields in various rows were missing, the wrong datatype, or one of various methods used to represent a lack of data, none of which were listed or standardized. For instance, many entries for what should have been patents lacked an abstract, and yet across all the data, that either meant the row simply skipped that cell and moved on to the next immediately (meaning it disobeyed the header format and put other data in the abstract's place), was an empty string, was an entirely-whitespace string, was "NULL" or similar, etc. Thus, it was necessary to do a round of significant pre-processing on the data to clean this; as a result, the end analysis does not include *every* granted patent in the time period.[5]

Once that was done, the methodology for doing analysis was fairly straightforward. I decided to use the Flesch-Kincaid readability ease as a metric for abstract readability, as it is the most intuitive of various popular metrics used in NLP tasks. I applied the formula for said metric (see Equation 1 below) to every abstract in the dataset, then took the mean of this for each year represented. The results of this are displayed in the results section.

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$$

Equation 1. Flesch Reading Ease; higher numbers represent less complex, more readable text.

---

[4] See, *e.g.*, Stein, Benno, Dennis Hoppe, and Tim Gollub. "The impact of spelling errors on patent search." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012. See also Eren Orbey's senior thesis for computer science, available on Yale networks at http://zoo.cs.yale.edu/classes/cs490/18-19b/orbey.eren.eo235/ (or by request).

[5] It is still a very significant number, at 12,423,873 patents total in the dataset out of the original 18,676,479.

RESULTS

Year	Avg. Flesch Reading Ease	Year	Avg. Flesch Reading Ease
2000	24.331705	2009	21.644634
2001	24.250116	2010	21.13108
2002	23.205761	2011	20.0394
2003	23.451397	2012	19.279927
2004	23.48121	2013	18.804739
2005	24.419158	2014	18.57852
2006	23.716361	2015	18.419324
2007	23.008513	2016	18.361424
2008	22.308101	2017	17.932704
		2018	17.495761

Figure 1. Table of average Flesch Reading Ease for patents from 2000 to 2018.

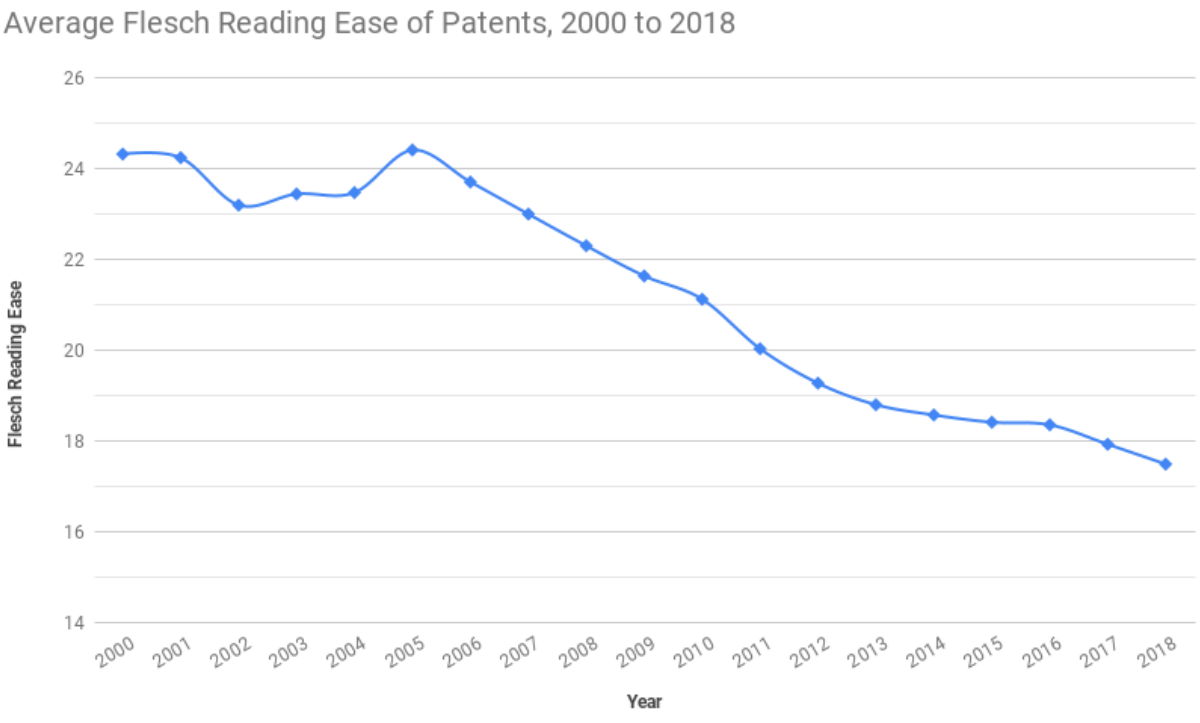

Figure 2. Plot of year vs. average Flesch Reading Ease from 2000 to 2018.

As the figures show, indeed the readability of patent abstracts decreased significantly from 2000 to 2018. This confirms the results of earlier work done on a more constrained time period. Indeed, since 2005, there has been a strict decline in the readability metric. This makes intuitive sense: as technology grows more and more complex and esoteric, so too must the patents which give IP rights to inventors and the abstracts which describe them. It's worth noting that all years' average patent readabilities are under 30, which corresponds to the reading level expected of college graduates.[6]

CONCLUSIONS

First, I briefly recap the conclusions of Miller for background. Evidence was found that patent trolling has been increasing. In particular, it was found that "licensing firms asserted their patents in less than 5% of distinct defendant-plaintiff disputes, but this number has risen to encompass more than 30% of all disputes."[7] This suggests that the use of patents in lawsuits by licensing firms "that acquired their patents from third parties,"[8] a prime suspect in patent trolling, has increased more than 25%.

Taking the results found in this project along with those from Miller, there's certainly a correlation between decreased patent abstract readability and an increase in patent trolling. I make no claims of causation or offer any methods which could potentially give weight to any such claims other than revisiting the intuition offered earlier in this paper. Although a difficult-to-understand abstract is hardly a perfect measure of the technical complexity of a patented technology, it certainly offers some suggestion of it. Moreover, the more technical a patent is, the more likely that its use in

---

[6] c.*f.* Flesch, Rudolf. *How to write plain English: A book for lawyers and consumers*. New York, NY: Harper & Row, 1979.
[7] Miller 274.
[8] *Ibid.*

lawsuits could seem intimidating or threatening to an alleged offender, and the more "wiggle room" a firm might have to argue that somewhere in the technical, nigh-unreadable specification of their patent lies a violation on the part of the alleged offender.