Patent Disclosure—An Economic Analysis Using Computational Linguistics

Shupeng Sun [1] Nancy Kong [2] Uwe Dulleck [2] Adam Jaffe [23]

[1]Queensland Treasury [2]Queensland University of Technology
[3]M.I.T; Brandeis University; Motu Research

October 28, 2019

Overview

- **Research question: do patent disclosure levels vary by business models of patent applicants?**
- Hypothesis: Universities (income mainly from licensing innovation) disclose more than firms (profit from in-house commercialisation) in their patent applications.
- Method:
  - Computational linguistic program: readability, a proxy for disclosure.
  - Statistical analysis: OLS and PCA.
- We find firms' patents are 0.37 s.d. more difficult to read, and require 1.2 years more formal education to comprehend.

Overview

- **Research question: do patent disclosure levels vary by business models of patent applicants?**
- Hypothesis: Universities (income mainly from licensing innovation) disclose more than firms (profit from in-house commercialisation) in their patent applications.
- Method:
  - Computational linguistic program: readability, a proxy for disclosure.
  - Statistical analysis: OLS and PCA.
- We find firms' patents are 0.37 s.d. more difficult to read, and require 1.2 years more formal education to comprehend.

Overview

- **Research question: do patent disclosure levels vary by business models of patent applicants?**
- Hypothesis: Universities (income mainly from licensing innovation) disclose more than firms (profit from in-house commercialisation) in their patent applications.
- Method:
  - Computational linguistic program: readability, a proxy for disclosure.
  - Statistical analysis: OLS and PCA.
- We find firms' patents are 0.37 s.d. more difficult to read, and require 1.2 years more formal education to comprehend.

Overview

- **Research question: do patent disclosure levels vary by business models of patent applicants?**
- Hypothesis: Universities (income mainly from licensing innovation) disclose more than firms (profit from in-house commercialisation) in their patent applications.
- Method:
  - Computational linguistic program: readability, a proxy for disclosure.
  - Statistical analysis: OLS and PCA.
- We find firms' patents are 0.37 s.d. more difficult to read, and require 1.2 years more formal education to comprehend.

Previously: Incentivising and measuring innovation

- The patent system is used to incentivize innovation (Solow, 1957)
  - A bargain between society and the inventors.
  - Disclosure of their "secret" in exchange for protection.
  - Higher disclosure has a positive effect on the following inventions.
- Patent counts are used as a measure of innovation (Griliches, 1990).
- Few studies examine the extent of disclosure that actually occurs in patent documents.

Previously: Incentivising and measuring innovation

- The patent system is used to incentivize innovation (Solow, 1957)
  - A bargain between society and the inventors.
  - Disclosure of their "secret" in exchange for protection.
  - Higher disclosure has a positive effect on the following inventions.
- Patent counts are used as a measure of innovation (Griliches, 1990).
- Few studies examine the extent of disclosure that actually occurs in patent documents.

Previously: Incentivising and measuring innovation

- The patent system is used to incentivize innovation (Solow, 1957)
  - A bargain between society and the inventors.
  - Disclosure of their "secret" in exchange for protection.
  - Higher disclosure has a positive effect on the following inventions.
- Patent counts are used as a measure of innovation (Griliches, 1990).
- Few studies examine the extent of disclosure that actually occurs in patent documents.

Patent disclosure

- Patents "shall contain... full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains... to make and use the same...invention" (Title 35 of the U.S.Code, S112).
- In the real world, technical information contained in patent documents is often inadequate and unclear (Roin, 2005; Devlin, 2010; Lemley, 2012).
- Agents involved in innovation or technology commercialisation behave strategically in choosing different disclosure levels in patent applications.
- Universities focus on generating income from licensing of innovations, $\uparrow$ disclosure.
  - High moral requirement of university research.
  - Low technology transfer.
- Firms focus on in-house production, $\downarrow$ disclosure.

Patent disclosure

- Patents "shall contain... full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains... to make and use the same...invention" (Title 35 of the U.S.Code, S112).
- In the real world, technical information contained in patent documents is often inadequate and unclear (Roin, 2005; Devlin, 2010; Lemley, 2012).
- Agents involved in innovation or technology commercialisation behave strategically in choosing different disclosure levels in patent applications.
- Universities focus on generating income from licensing of innovations, $\uparrow$ disclosure.
  - High moral requirement of university research.
  - Low technology transfer.
- Firms focus on in-house production, $\downarrow$ disclosure.

Patent disclosure

- Patents "shall contain... full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains... to make and use the same...invention" (Title 35 of the U.S.Code, S112).
- In the real world, technical information contained in patent documents is often inadequate and unclear (Roin, 2005; Devlin, 2010; Lemley, 2012).
- Agents involved in innovation or technology commercialisation behave strategically in choosing different disclosure levels in patent applications.
- Universities focus on generating income from licensing of innovations, $\uparrow$ disclosure.
  - High moral requirement of university research.
  - Low technology transfer.
- Firms focus on in-house production, $\downarrow$ disclosure.

Patent disclosure

- Patents "shall contain... full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains... to make and use the same...invention" (Title 35 of the U.S.Code, S112).
- In the real world, technical information contained in patent documents is often inadequate and unclear (Roin, 2005; Devlin, 2010; Lemley, 2012).
- Agents involved in innovation or technology commercialisation behave strategically in choosing different disclosure levels in patent applications.
- Universities focus on generating income from licensing of innovations, $\uparrow$ disclosure.
  - High moral requirement of university research.
  - Low technology transfer.
- Firms focus on in-house production, $\downarrow$ disclosure.

Patent disclosure

- Patents "shall contain... full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains... to make and use the same...invention" (Title 35 of the U.S.Code, S112).
- In the real world, technical information contained in patent documents is often inadequate and unclear (Roin, 2005; Devlin, 2010; Lemley, 2012).
- Agents involved in innovation or technology commercialisation behave strategically in choosing different disclosure levels in patent applications.
- Universities focus on generating income from licensing of innovations, $\uparrow$ disclosure.
  - High moral requirement of university research.
  - Low technology transfer.
- Firms focus on in-house production, $\downarrow$ disclosure.

Readability as a proxy for disclosure

- Readability measures have already been used in finance and accounting to measure whether readers can extract information in financial reports (Li, 2008; Miller, 2010; You and Zhang, 2009; Lawrence, 2013).
- Our fine-grained readability measures from a computational linguistics program provide a proxy for the extent of patent disclosure.
- We use 106 linguistic features designed by Vajjala and Meurers (2014).

Readability as a proxy for disclosure

- Readability measures have already been used in finance and accounting to measure whether readers can extract information in financial reports (Li, 2008; Miller, 2010; You and Zhang, 2009; Lawrence, 2013).
- Our fine-grained readability measures from a computational linguistics program provide a proxy for the extent of patent disclosure.
- We use 106 linguistic features designed by Vajjala and Meurers (2014).

Readability as a proxy for disclosure

- Readability measures have already been used in finance and accounting to measure whether readers can extract information in financial reports (Li, 2008; Miller, 2010; You and Zhang, 2009; Lawrence, 2013).
- Our fine-grained readability measures from a computational linguistics program provide a proxy for the extent of patent disclosure.
- We use 106 linguistic features designed by Vajjala and Meurers (2014).

Linguistic features

The 106 indicators can be categorized into four features:

- **Lexical richness**: how complicated words and grammar are.
  - e.g. Adverbs/total sentences.
- **Syntactic complexity**: Sentence structure.
  - e.g. Average sentence length.
- **Discourse feature**: cognitive processing.
  - e.g. local noun overlap count.
- **Word characteristics**: how common the words and expressions are used.
  - e.g. Average word familiarity rating (MRC Psycholinguistic Database, 1988).

Linguistic features

The 106 indicators can be categorized into four features:

- **Lexical richness**: how complicated words and grammar are.
  - e.g. Adverbs/total sentences.
- **Syntactic complexity**: Sentence structure.
  - e.g. Average sentence length.
- **Discourse feature**: cognitive processing.
  - e.g. local noun overlap count.
- **Word characteristics**: how common the words and expressions are used.
  - e.g. Average word familiarity rating (MRC Psycholinguistic Database, 1988).

Linguistic features

The 106 indicators can be categorized into four features:

- **Lexical richness**: how complicated words and grammar are.
  - e.g. Adverbs/total sentences.
- **Syntactic complexity**: Sentence structure.
  - e.g. Average sentence length.
- **Discourse feature**: cognitive processing.
  - e.g. local noun overlap count.
- **Word characteristics**: how common the words and expressions are used.
  - e.g. Average word familiarity rating (MRC Psycholinguistic Database, 1988).

Linguistic features

The 106 indicators can be categorized into four features:

- **Lexical richness**: how complicated words and grammar are.
  - e.g. Adverbs/total sentences.
- **Syntactic complexity**: Sentence structure.
  - e.g. Average sentence length.
- **Discourse feature**: cognitive processing.
  - e.g. local noun overlap count.
- **Word characteristics**: how common the words and expressions are used.
  - e.g. Average word familiarity rating (MRC Psycholinguistic Database, 1988).

Data

- Full text of patent applications:
  - US patent applications since 2000.
  - Nanotechnology (USPC 977), battery (320) and photoelectric (136).
  - Processed in the linguistic program.
- Metadata:
  - Application date, priority numbers, earliest priority date and number, applicants, inventors, cited by patent counts, simple and extended family sizes, sequence count, NPL citation count, and NPL resolved citation count.
- Merged the metadata with the linguistic indicators.
- 31164 observations.

Data

- Full text of patent applications:
  - US patent applications since 2000.
  - Nanotechnology (USPC 977), battery (320) and photoelectric (136).
  - Processed in the linguistic program.
- Metadata:
  - Application date, priority numbers, earliest priority date and number, applicants, inventors, cited by patent counts, simple and extended family sizes, sequence count, NPL citation count, and NPL resolved citation count.
- Merged the metadata with the linguistic indicators.
- 31164 observations.

Data

- Full text of patent applications:
  - US patent applications since 2000.
  - Nanotechnology (USPC 977), battery (320) and photoelectric (136).
  - Processed in the linguistic program.
- Metadata:
  - Application date, priority numbers, earliest priority date and number, applicants, inventors, cited by patent counts, simple and extended family sizes, sequence count, NPL citation count, and NPL resolved citation count.
- Merged the metadata with the linguistic indicators.
- 31164 observations.

Data

- Full text of patent applications:
  - US patent applications since 2000.
  - Nanotechnology (USPC 977), battery (320) and photoelectric (136).
  - Processed in the linguistic program.
- Metadata:
  - Application date, priority numbers, earliest priority date and number, applicants, inventors, cited by patent counts, simple and extended family sizes, sequence count, NPL citation count, and NPL resolved citation count.
- Merged the metadata with the linguistic indicators.
- 31164 observations.

Universities and firms

According to the top 100 applicants:

- Universities are identified with the name “univ”, “inst” and “college”.
- Firms are identified as “INC”, “LTD”, “CORP”, “LLC”, and “CO”.

There are 3657 universities (11.7%) and 17035 firms (55%).

Descriptive analysis

Significant differences between firms and uni

Table 1: Summary statistics

	Firms Mean	SD	Uni Mean	SD	Difference Coefficient	T-stat
Categories:						
uspc136	0.36	0.48	0.23	0.42	0.12***	(13.12)
uspc320	0.33	0.47	0.04	0.19	0.29***	(55.69)
uspc977	0.33	0.47	0.76	0.43	-0.43***	(-46.37)
Characteristics:						
Cited_by_Patent_Count	14.33	26.32	10.96	18.22	3.37***	(8.01)
Simple_Family_Size	7.41	9.64	5.50	5.12	1.91***	(14.91)
Extended_Family_Size	11.82	29.56	6.71	9.33	5.11***	(17.05)
Sequence_Count	1.41	58.20	9.31	305.85	-7.89	(-1.28)
NPL_Citation_Count	0.36	1.25	1.10	2.31	-0.74***	(-15.56)
NPL_Resolved_Citation_Count	0.18	0.82	0.80	1.84	-0.63***	(-16.75)
Linguistic measures:						
Adverbs/total sentence	0.03	0.01	0.02	0.01	0.00***	(4.12)
Average sentence length	33.26	7.70	30.09	6.82	3.17***	(21.15)
Local noun overlapping	0.62	0.11	0.54	0.10	0.08***	(38.07)
Average word familiarity rating	3.92	0.19	3.88	0.18	0.03***	(8.70)
Observations	15870		2492		18362	

Note: The sample excludes applications submitted jointly by both firms and universities.

Econometric model

OLS regression:

$$Y_{ij} = \alpha + \beta_1 Firm_{ij} + \beta_2 Both_{ij} + \beta_3 Other_{ij} + \beta_4 \mathbf{X}_{ij} + \delta_j + \epsilon_{ij}$$

- Y_{ij} is linguistic indicators for application i in sub-classification j.
- $Firm_{ij} = 1$ if the patent application is filed by a firm. *Uni* is the base.
- X_{ij} is a vector of various citation counts, simple and extended family size.
- δ_j is US classification and sub-classification fixed effect.
- ϵ_{ij} is the error term.

Hypothesis: Firms have less readable patents:
i.e. β_1 to be significant and positive.

Results

Firm's applications are more difficult to read

Table 2: OLS estimates of representative linguistic features

VARIABLES	(1) Adverbs/sentence	(2) Avg sentence length	(3) Local noun overlap	(4) Avg word familiarity
Firms	0.00105***	1.838***	0.0475***	0.0206***
	(0.000196)	(0.175)	(0.00229)	(0.00422)
Observations	31,111	31,106	31,121	31,111
R-squared	0.056	0.085	0.174	0.068
Citation counts	Y	Y	Y	Y
Family size	Y	Y	Y	Y
Classification FE	Y	Y	Y	Y

More intuitively: Fog index

Firm's applications require 1.2 years more education to read

Gunning-Fog index (1968): a well established "hard to read" indicator:

$$Fog = 0.4(Words/Sentences + ComplexWords/Words)$$

E.g. $Fog = 7$ means 7 years of formal education a person needs to understand the text on the first reading.

Table 3: OLS estimates using Fog Index

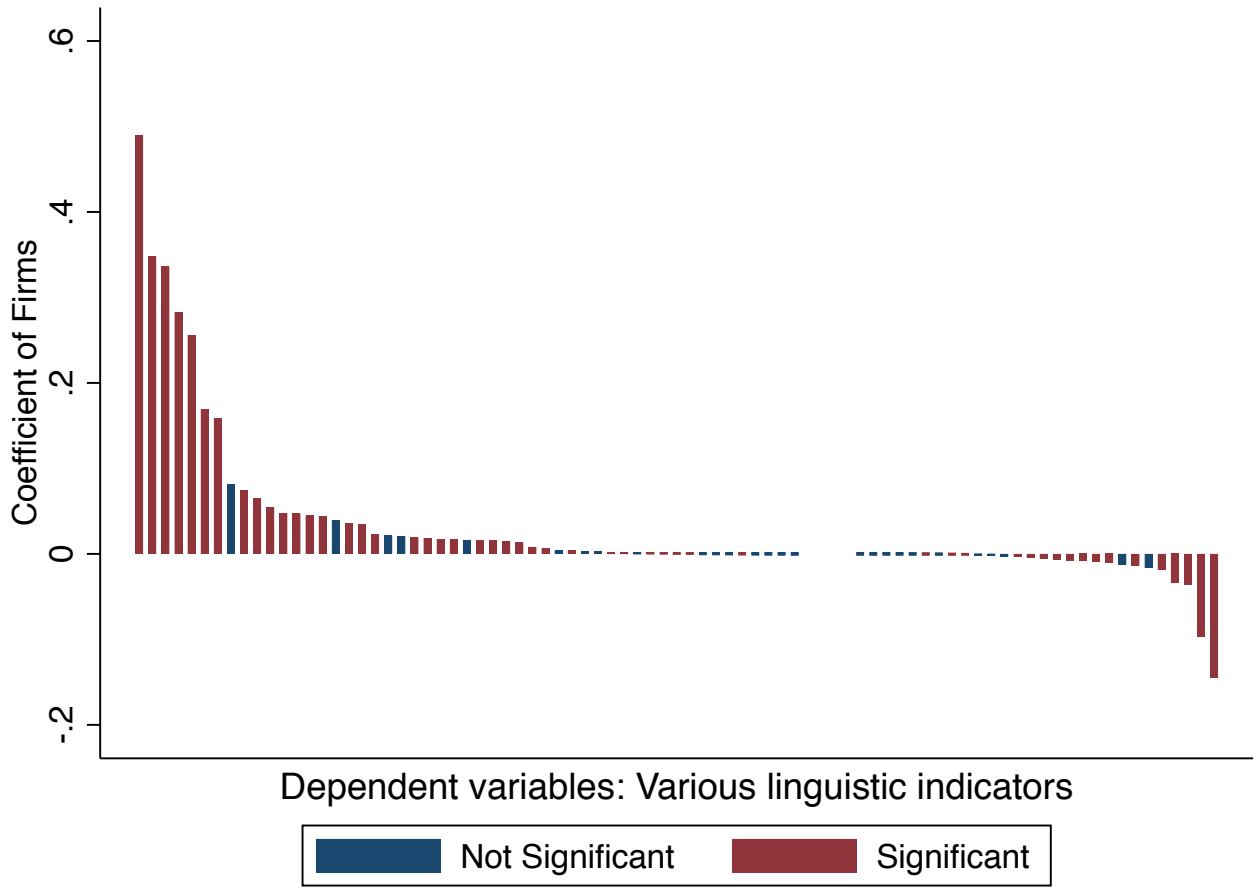
VARIABLES	(1) FOG
Firms	1.211*** (0.0772)
Observations	31,102
R-squared	0.079
Citation counts	Yes
Simple & extended family size	Yes
Classification fixed effects	Yes

Standard errors clustered at the US classification level in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

In total:

Among 106 linguistic features, 72 (68%) features are significantly different between firms and universities. Among these, 51 (71%) features show negative correlations between firms and patent readability.

Figure 1: Estimates of firms plotted with significance

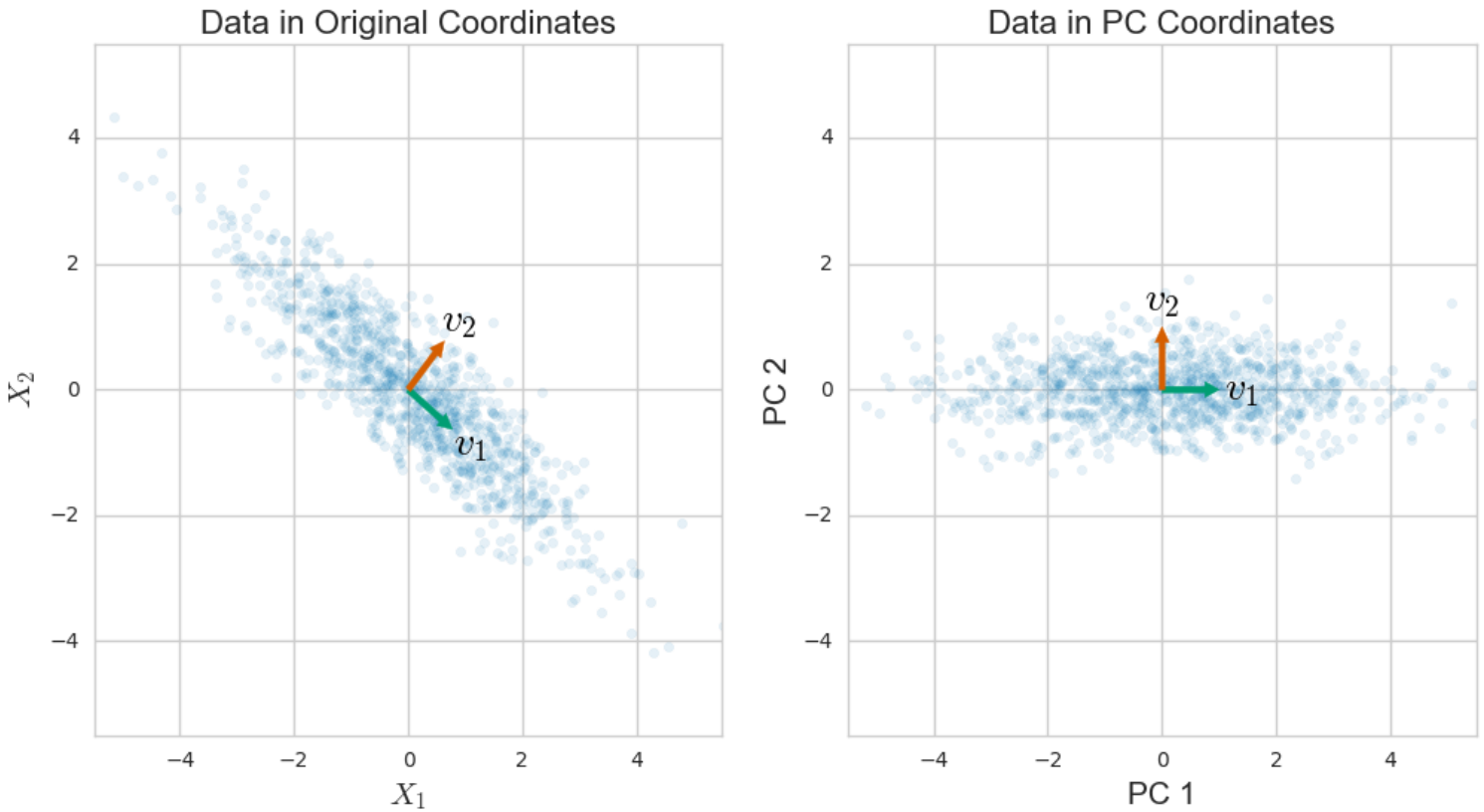


Machine learning: Principal Component Analysis (PCA)

PCA is an unsupervised, non-parametric statistical technique primarily used for reducing dimensions in machine learning.

$$\arg\max_{\boldsymbol{w}} \{||\boldsymbol{Yw}||^2\} \qquad \text{s.t. } \boldsymbol{w}^2 = 1$$

Figure 2: PCA illustration



Credit: https://intoli.com

PCA results

Firms' patents are 0.37 s.d. more difficult to read

Table 4: Estimates using principal component analysis

VARIABLES	(1) Standardized values of P1
Firms	0.368***
	(0.0128)
Observations	31,095
R-squared	0.289
Citation counts	Yes
Simple & Extended family size	Yes
Classification fixed effects	Yes

Standard errors clustered at the US classification level in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Summary

- We find strong evidence that firms' patent applications are more difficult to read in terms of lexical richness, syntactic complexity, discourse feature, and word characteristics.
- Gunning-Fog index shows firms' applications require 1.2 years more formal education to read.
- Principal Component Analysis shows that firms' patents are 0.37 s.d more difficult to read using a synthetic variable.

Future plans

- Eliminate other channels that could result in the differences in readability.
- Validate the measure by patent attorney's evaluations.
- Distinguish between domestic applicants and foreign applicants.
- Difference between practicing firms and non-practicing firms (patent trolls).