## Appendix B: Methodology

We applied a natural-language Python[1] algorithm to identify whether a claim is independent or dependent, and to parse each individual claim for the full text of each claim. To do so, we assumed that all dependent claims contain some dependency language referring to (and thereby incorporating limitations from) earlier claims, rather than actually reciting the language of limitations of the claims from which they depend.

In the `document_stats` datasets, we aggregated the individual claims-level data into patent/application-level summary statistics.[2] Each observation contains, for each application at publication and for each patent at grant, the number of independent and dependent claims, the average number of words in all independent claims, and a count of the number of words in the shortest independent claim. Since this paper's principal focus is the analysis of patent application and granted patent claims and filing characteristics, the dissemination and analysis of other patent-prosecution-related characteristics, such as data on RCE filings, numbers and types of continuations generated, appeals, etc., will be left to future dataset releases and analyses.

This Appendix details the data sources, methodology, descriptive statistics, and some general trends that can be observed in the `claims_stats`, `claims_fulltext`, and `document_stats` datasets. It is our hope that researchers will be able to use this data to enhance understanding of the examination process, including but not limited to assessing patent scope and how it changes during examination.

### Data Sources

Our primary data sources for the claims-level datasets include the Patent Application Publication Full-Text and Patent Grant Full Text files provided by the U.S. Patent and Trademark Office (USPTO).[3] The Patent Application Publication Full-Text data, provided in XML format and disseminated as separate files by years or ranges of years, contains the full-text of all patent applications published from December 2000 to December 31, 2014. The Patent Grant Full-Text files, provided in multiple file formats (XML, SGML, and APS), contain the full-text of all patents issued from 1976 to December 31, 2014. These files were cleaned, parsed, and appended to create the `claims_fulltext` datasets (one each for PGPubs and patents), which includes the patent or application number, the full-text of each claim, and an indicator variable to distinguish between independent and dependent claims, in a STATA® data file format.[4] In the `claims_stats` datasets (again, one each for PGPubs and patents), we include claim-level statistics (e.g. word count, number of "or"s, etc.) but not the full text of each claim. This allows researchers to analyze claim-level data in a more manageable dataset size.

---

[1] The Python code used to generate the USPTO's Patent Claims Research Datasets will be made available soon on GitHub.

[2] The data were obtained from USPTO Electronic Bulk Data Products (http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products)

[3] Full-text of patents and patent applications is available at http://patft.uspto.gov/. Bulk data is available at http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products.

[4] Cancelled claims were identified in claims_fulltext but were not included in independent claim count and length summary statistic calculations.

For our analysis, but not included in our data release, we merged an in-house USPTO patent application database with the `document_stats` datasets to link certain filing and prosecution information at the application/patent-level for publicly available (published and/or granted) applications with our measure of patent scope. This information includes the nature of any parent application for the subject application (e.g., having a parent that was a foreign or PCT application) and the relationship to the parent of the subject application if the parent is a regular utility application (e.g., the subject application is a divisional application of that parent) and any filing priority information relating to the parent application. The USPTO in-house database includes various post-filing prosecution characteristics such as disposal type (`disp_ty`) and disposal date (`disp_dt`), among others.[5] We also used certain prosecution characteristics. For example, we use the disposal date for an application (which includes the time evaluating any requests for continued examination (RCEs) in the same application) to determine total pendency from filing to abandonment or grant ("disposal"[6]), and post-first-action pendency to measure the time from first-action to disposal. While the dataset does not include claim counts or claim lengths at the time of an abandonment, the our merged data on publications included a variable to distinguish whether the application matured into a granted patent or ultimately went abandoned (`disp_ty`). Accordingly, many of our analyses distinguish characteristics at publication of applications that result in grants from applications that result in abandonments. Of course, abandonment (or grant) of a particular application does not mean that prosecution ended on the invention described in the abandoned (or granted) application, as various forms of continuation applications may have been filed prior to final disposition of any particular application.

**Data Limitations**

Relying on publicly available information on claims as captured from existing databases limits our sample in several ways. First, we can observe the claim text only at the time of publication and at the time of grant. This reliance also restricts the time period, because pre-grant publication of patent applications has been practiced by the USPTO only for applications filed after November 29, 2000.[7] Since that time, and without a non-publication request (which requires foregoing international protection on the patented innovation), publication has been required by statute 18 months after the filing priority date requested in relation to the earliest related parent application.[8] Applications filed prior to November 29, 2000 are unpublished. Thus, although our source patent dataset (grants) extends back to 1976, the bulk patent application data contains applications filed only during and after 2000. We have calculated that since November 29, 2000, approximately ten percent of filed applications have opted out of publication.

Further, in contrast to the captured data on claims from granted applications (at publication and at issue), machine-readable claim text is not readily available for abandoned applications (after publication). That is, we cannot observe the change in claims between publication and abandonment. Consequently, we limit our analysis of difference variables (`dif_wrd_min`,

[5] For a fuller description of all of the prosecution characteristic variables that were available for coding, please see the variable descriptions in Appendix C.

[6] There are two types of disposals: abandonment or grant. For more information on disposals and patent prosecution, please see http://www.uspto.gov/patents-getting-started/general-information-concerning-patents#heading-22. Please note that abandoned applications can sometimes be reinstated.

[7] *See* 35 U.S.C. 122(b).

[8] *See* 35 U.S.C. § 122(b)(2)(B); 37 C.F.R. § 1.213(a)(1)-(4).

`dif_wrd_avg`, and `dif_clm_ct`) to publication-patent pairs (i.e., to applications that resulted in granted patents).

Although it is possible for claims in a particular application to change between filing and publication, we believe this is a relatively infrequent event. Our analysis shows that only 8.11 percent of total applications in the dataset have a preliminary claims amendment filed after their actual (not priority) filing date but before the publication date. Normal office practice is to incorporate preliminary amendments into the claims when they are published, and thus these claim amendments (except for the possible few that are filed too close to publication to be incorporated) are reflected in the publication data. Since the percentage of applications with preliminary amendments submitted between filing and publication is relatively small, we have treated for analysis the claims at publication as a reasonable approximation of the claims at filing.[9]

As can be expected in any dataset of this size, the source data files (the Patent Application Publication Full-Text and Patent Grant Full Text files) have some errors. Specifically, some claim language was excluded from the text and word length counts. The general (introductory) claiming language (e.g., "I claim" or "What is claimed is") has been excluded from the `claims_fulltext` datasets.[10] Similarly, we have not included the numeral associated with any claim in the claim length counts; rather, we have included only the language following the numeral for any particular claim (although the numeral is included in the dataset).[11] For example, U.S. patent 4,788,349[12] was issued with three claims of word lengths fourteen, two, and two, respectively. Excluding the general claiming language – which is not included in the datasets and consists of the words, "I claim:" – and the numeral assigned to the claims thus allows for one word claims such as chemical compounds. The exclusion of the general claiming language and numerals from the claim counts slightly biases the individual, average, and minimum independent claim length downwards.

**Claim Identification and Measurements**

As stated above, we used full-text claims data for patents and applications (`claims_fulltext`) to create patent-level summary statistics for both PGPubs and patents. We computed the summary statistics by applying a Python-based algorithm developed to distinguish independent claims from dependent claims and to compute various measures of claim length and claim count, among other variables.[13] The algorithm identifies independent from dependent claims by assuming that dependent claims will reference independent and other dependent claims, but not vice-versa.[14] Specifically, if claim language contains a direct reference to another claim or a group of claims, we designated the referring claim as a dependent claim (and coded it as such in the database). If the claim contains no such language, we designated it as

[9] It should be noted that not all preliminary amendments are included in an application's publication. See MPEP 1121.

[10] There are exceptions in the claims_fulltext data set: (1) the first claims of twenty-two utility patents begin with the general (introductory) claiming language, "I claim"; and (2) claims in ten patents, such as patent 6,901,209, begin with the words, "I Claim." For example, claim 5 states, "I claim the access system of claim 4 characterized by the addition of data manager means to allow a user to access the program." This list is not exhaustive.

[11] The Claim number can be found as a separate field in the claims_fulltext data set (claim_no).

[12] See https://www.google.com/patents/US4788349

[13] See Python code in Appendix D

[14] It may be the case that a claim will contain referents to other claims that do not incorporate the other claims' limitations. However, we believe this to be a rare event.

an independent claim. We repeated this process for all applications at publication and for those that are granted at issue. To measure the independent claim length (ICL), we used a simple count of the number of words in each independent claim.[15] To create a patent-level metric, we measured ICL by using the minimum claim length among all independent claims of an application or granted patent. Our metric for the number of independent claims is a simple independent claim count (ICC). [16] We did not include in `document_stats` the minimum claim length for dependent claims. [17]

Following our assumption that patent scope depends on the length and number of independent claims, it is important to provide the arithmetic difference in the length and number of independent claims between publication and grant. These differences from publication to grant provide an approximation of the changes in breadth of the independent claims from filing to grant and thus of the change in the scope of the applications during prosecution. For example, as a direct result of our assumption on patent scope, if the change in independent claim length (ICL) from publication to grant is positive, then it follows that the patent scope at grant should (generally) be narrower than at publication (and filing). If the change in independent claim count (ICC) is positive, then the scope of the patent should (generally) be broader at grant than at publication (and filing).

[15] Because the algorithm uses natural language processing, claims that separate portions of words with spaces are automatically read as including separate words, which may thereby artificially increase the claim's word count. For example, chemical formula sometimes are written as a single word without spaces, but occasionally may contain many spaces, which would artificially increase the word count by as many spaces as are added. See US Patent 3,262,977, claim 4 ("N – [1' -phenyl-propyl-('1)] – 1,1 diphenyl-propyl-(3)-amine").

[16] Our algorithm also identifies specific words or phrases (e.g., "or" and "selected from") that are more likely to have the potential to broaden the scope of an independent claim by addition of other words, to permit robustness checks.

[17] To measure the dependent claim length (DCL), we would need to start with a simple count of the number of words in each dependent claim, and then add the count of the limitations language of the claim(s) from which the dependent claim depends and eliminate the count of the referential language in the dependent claim (as such language would then become duplicative and unnecessary). Nevertheless, the data in `claims_fulltext` are coded with the claim number(s) from which each dependent claim directly depends. Accordingly, some automated counts to approximate the number of words of dependent claims are possible to perform, e.g., by tracing the chains of dependency and adding the simple count of the words of each dependent claim and of the claim(s) from which it depends. (Such simple counts would be slightly over-weighted, by including counts of both the referential language and of the full text of the claim(s) to which those dependent claims refer). Some dependent claims, moreover, reference multiple independent or dependent claims that may have different lengths, which makes it more difficult to provide a count that is an accurate length for any such dependent claim. (Of course, each such multiply dependent claim could be decomposed into separate claims for further analysis.)