Akron Law Review

Volume 55
Issue 2 *Intellectual Property Issue*

Article 4

2022

Information Theory and Patent Documents

W. Michael Schuster

INFORMATION THEORY AND PATENT DOCUMENTS

W. Michael Schuster[*]

I. INTRODUCTION

Recent scholarship has expanded the scope of analytical tools available to patent law researchers. Examples include the use of textual analysis to determine the similarity of two patents[1] and the use of network analysis to assess patent value.[2] This essay continues that trend by proposing a theoretical application of information theory to analyze

[*]Assistant Professor of Legal Studies, University of Georgia, Terry College of Business with a courtesy appointment at the University of Georgia School of Law.

1. *See, e.g.*, Sam Arts, Bruno Cassiman & Juan Carlos Gomez, *Text Matching to Measure Patent Similarity*, 39 STRATEGIC MGMT. J. 62, 64–65 (2018); *see also* W. Michael Schuster & Kristen Green Valentine, *An Empirical Analysis of Patent Citation Relevance and Applicant Strategy*, 59 AM. BUS. L.J. (forthcoming Summer 2022) (manuscript at 231) (using the approach from *Text Matching to Measure Patent Similarity* to analyze backward patent citation relevance).

2. Guan-Can Yang, Gang Li, Chun-Ya Li, Yun-Hua Zhao, Jing Zhang, Tong Liu, Dar-Zen Chen & Mu-Hsuan Huang, *Using the Comprehensive Patent Citation Network (CPC) to Evaluate Patent Value,* 105 SCIENTOMETRICS 1319 (2015).

textual ambiguity and identify particularly original disclosures in patent documents.

Mathematician and engineer Claude Shannon published the foundations of information theory in 1948.[3] His research focused on analyzing the amount of information per second that could be transmitted and how to encode messages for efficient transmission.[4] As discussed in Section II, Shannon surmised that a message's information content is a function of the uncertainty (also called "surprise") of the message.[5] Highly unlikely messages convey a greater deal of information,[6] and the probability of a message can be determined by reference to earlier messages and the current context.

For example, if a message thus far consists of the following characters: "I N F O R M A T I O," then it is highly likely that the next character will be an "N."[7] Thus, the receiver obtains very little new information from the letter "N." Similarly, if an English language message contains a "Q," then very little information (surprise) is received when the next character is a "U," as a Q will be followed by a U in the vast majority of instances in English communications.[8] In contrast, we receive a great deal of information when "Q" is followed by an "F" because it is very improbable.

From this recognition that not all characters (or words) convey the same amount of information, Shannon quantified the expected information content of any message through its *entropy*.[9] This metric (largely unrelated to the thermodynamics metric of the same name) quantifies the amount of information we expect to receive from the next part of a message, given (a) what we know about the message's current context and (b) statistical trends in earlier bodies of collected messages.

---

3. C. E. Shannon, *A Mathematical Theory of Communication*, 27 BELL SYS. TECH. J. 379, 623 (1948), reprinted in CLAUDE E. SHANNON & WARREN WEAVER, THE MATHEMATICAL THEORY OF COMMUNICATION 29 (Univ. of Ill. Press Urbana ed. 1964) (10th prtg. 1964); Jeanne C. Fromer, *An Information Theory of Copyright Law*, 64 EMORY L.J. 71, 76–77 (2014); Alan L. Durham, *Copyright and Information Theory: Toward an Alternative Model of "Authorship,"* 2004 BYU L. REV. 69, 73 (2004).

4. *See* Thomas M. Cover & Joy A. Thomas, ELEMENTS OF INFORMATION THEORY 1 (2d ed. 2006).

5. John R. Pierce, AN INTRODUCTION TO INFORMATION THEORY: SYMBOLS, SIGNALS, & NOISE 23–24 (2d rev. ed. 1980); Dan L. Burk, *The Problem of Process in Biotechnology*, 43 HOUS. L. REV. 561, 584 (2006).

6. Ian Goodfellow, Yoshua Bengio, & Aaron Courville, DEEP LEARNING (ADAPTIVE COMPUTATION AND MACHINE LEARNING SERIES) ILLUSTRATED EDITION 73 (2016).

7. Durham, *supra* note 3, at 76–77.

8. Pierce, *supra* note 5, at 49.

9. Ernesto Estevez-Rams, Ania Mesa-Rodriguez & Daniel Estevez-Moya, *Complexity-Entropy Analysis at Different Levels of Organisation in Written Language*, 14 PLOS ONE 1 (2019).

While seminal to modern digital technology, researchers have widely applied Shannon's work and information theory, including within several legal studies articles.[10] In this paper, I extend that work by presenting a theoretical application of information theory to quantify several aspects of patent law, including lexical ambiguity and originality in innovation. To this end, Section II introduces Shannon's ideas of quantifying a message's information content and related entropy measures. Section III recognizes that to ascertain a message's information content, one must identify how likely or unlikely that particular message is. This section discusses how to quantify a message's probability.

Section IV looks to prior applications of information theory to identify textual ambiguity and proposes how to apply these lessons to patent law. For instance, Section IV(B) discusses several instances where firms might employ information theory-centric approaches to identify and avoid (or court) textual ambiguity in patent documents. Finally, Section V analyzes the literature on identifying novel[11] text using information theory, then discusses application of this literature to quantify patent law phenomena such as groundbreaking patents or patent thickets.

II. QUANTIFYING INFORMATION

A. Information of an Event

Shannon defined the *self-information* of a particular event or outcome—such as a single character or word—as the amount of information disclosed when that event or outcome occurs.[12] In mathematical terms, self-information equals the log of one over the probability of the event.[13] Important to the current study, that definition shows that *self-information is a function of probability that the event will occur*.[14] According to the above definition, the lower an event's expected probability, the greater the information disclosed.[15] In other words, a high information event will have a high degree of surprise (as it was largely unexpected).

10. Oren Bar-Gill & Omri Ben-Shahar, *An Information Theory of Willful Breach*, 107 MICH. L. REV. 1479 (2009); Durham, *supra* note 3, at 76–77.

11. The term "novel" is used here to mean "standing out from its peers," as opposed to the usage commonly associated with 35 U.S.C. § 102.

12. Darrel Hankerson, Greg A. Harris, & Peter D. Johnson, Jr., INTRODUCTION TO INFORMATION THEORY AND DATA COMPRESSION 26 (2d ed. 2003).

13. *Id.* at 25.

14. *Id.* at 25–26.

15. *Id.* at 26.

Where a message is 100% likely to occur, nothing new is communicated, and the information conveyed (surprise) is at its minimum—zero.[16] This is not remarkable, as the receiving party obtains nothing new from a message they already knew would convey one specific message. On the contrary, the information contained in a message selected out of a predetermined set of possible messages is at its maximum when all messages are equally likely to occur.[17] Further, the quantum of information conveyed by a message increases when the number of possible (equally likely) messages increases.[18]

Shannon chose to quantify the amount of information conveyed by a message in terms of *bits*—the number of yes/no questions that an analyst must ask to identify the message given a known probability of possible messages.[19] For example, the outcome of a fair coin toss conveys one bit of information to the receiver (i.e., the party viewing the coin toss). Restated, to identify the amount of information from a coin toss, we must ask one yes/no question, namely "Does the coin end up on heads?"

Mathematically, the number of bits conveyed by a particular message (e.g., "the coin came up heads") is generalized as:

$$\text{Information of a specific message } x = I(x) = \log_2 (1 / p(x)) \quad (1)$$

where I(x) is the number of bits of information conveyed by a specific message x, and p(x) is the probability of that particular message being sent. For the fair (50/50) coin toss example, p(x) is .5, and application of Equation 1 finds the information conveyed to equal the expected 1 bit (one yes/no question).

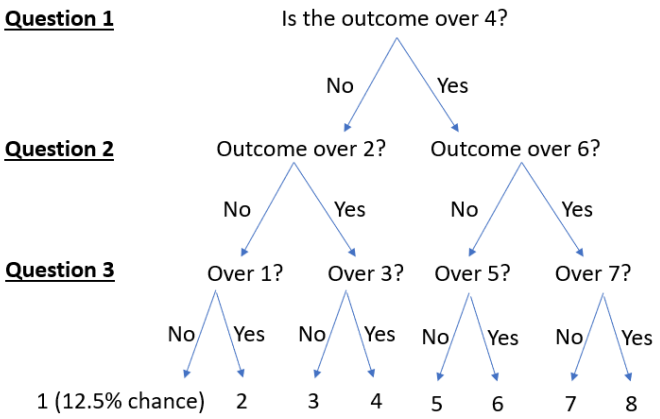
16. Shannon, *supra* note 3, at 51 ("[O]nly when we are certain of the outcome does H vanish."); Hankerson et al., *supra* note 12, at 41.

17. Hankerson et al., *supra* note 12, at 40.

18. Pierce, *supra* note 5, at 23.

19. Shannon, *supra* note 3, at 380; Steven T. Piantadosi, Harry Tily, & Edward Gibson, *The Communicative Function of Ambiguity in Language*, 122 COGNITION 280, 282–83 (2012); Hankerson et al., *supra* note 12, at 27.

As a second example, imagine an eight-sided die with each of the eight sides equally likely to come up 12.5% of the time. To calculate the number of bits conveyed by any particular outcome, we must identify how many yes/no questions (bits) we must ask to identify the message.[20] We can do this through just three questions, as demonstrated by the following flowchart of three yes/no questions:



Again, application of equation 1 gives us the expected outcome of 3 bits (i.e., 3 yes/no questions) for any particular outcome (1 or 2 or … 8), given that each number has an equally likely 12.5% chance of occurring. The $\log_2$ of (1 / .125) equals 3 (bits).

The above assumes that every possible message (e.g., each side of a coin or die) has an equal chance of occurring. This is not the way the real world works: patterns exist such that a single message may be more or less likely to occur relative to all others. For example, if we picked a random letter in a random book on a random shelf in a library, the odds that the letter is an "E" is approximately 171 times greater than it being a "Z."[21] According to one dataset of 40,000 words, an "E" will come up 12.02% of the time, and a "Z" will come up .07% of the time. Applying Equation 1 (to calculate the number of bits of information conveyed as a function of the message's likelihood) thus finds that an "E" conveys 3.06 bits of information and a "Z" conveys 10.48 bits. Consistent with our prior

20. It is of note that while Shannon chose to use base 2 (*e.g.*, how many yes/no questions must be answered to identify a message), this is not mathematically necessary. Hankerson et al., *supra* note 12, at 27. Bases are essentially units used to quantify information, whereby a user can change units (*e.g.*, to base 10) without changing the quantification of the information analyzed. *Id.*; *see also* Shannon, *supra* note 3, at 380.

21. *English Letter Frequency*, CORNELL, http://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html [https://perma.cc/CML5-EZFG].

discussion, the occurrence of a relatively rare message (such as "Z") will convey a greater deal of information than a relatively common message (such as "E").

To this point, we have considered the information conveyed by particular communications (coin tosses, dice rolls, or single letters) that are X% probable to occur. However, the definition of *message* is not so limited. Any message or event that has a particular likelihood of occurrence conveys quantifiable information. Germane to the current study, discrete words can constitute information-conveying messages.

Consider an example where you receive the following word-by-word message: "For lunch, I will eat a peanut butter and." We almost invariably read the next word of the expected message as "jelly." This is due to the particularly high frequency that the word "jelly" occurs after the words "peanut butter and." The probability of "jelly" as the next word is further increased because people commonly eat peanut butter and jelly sandwiches for lunch. Thus, the information communicated when you receive "jelly" is low. Assuming that jelly will be the next word in this string 99% of the time, Equation 1 tells us that "jelly" as the last word communicates .01 bits of information. In contrast, if the next word is "pickles"—which we assume to be highly unlikely at a .01% probability—it will convey a significantly greater amount of information. Calculated using Equation 1, "pickles" would convey 13.3 bits of information.

In each instance described above (i.e., words, letters, coins, or dice), the amount of information conveyed is a function of the probability that a specific unit (e.g., the letter "H") will be conveyed. The following addresses the average expected information conveyed by a currently unknown message that will be drawn from a set of probable outcomes.

B. Expected Information of an Unknown Event (Entropy)

In an uncertain world (where we do not know the actual message conveyed in a single event), we may want to know an event's *entropy*: its average expected information content (a single variable of multiple possible outcomes). Entropy is the sum of the information content of each possible outcome in bits, multiplied by the likelihood that it will occur. [22] It can be calculated using Equation 2: [23]

22. FRED ATTNEAVE, APPLICATIONS OF INFORMATION THEORY TO PSYCHOLOGY: A SUMMARY OF BASIC CONCEPTS, METHODS, AND RESULTS 7–8 (1959); Paul H. Edelman, *The Dimension of the Supreme Court*, 20 CONST. COMMENT. 557, 559 (2004).

23. Shannon, *supra* note 3, at 393; Hankerson et al., *supra* note 12, at 40.

$$H(\mathcal{E}) = -\sum_{i \in I} P(E_i) \log P(E_i) \quad (2)$$

where H is the entropy (expected information content) and $P(E_i)$ is the probability of a particular outcome E_i. A higher entropy is indicative of a higher degree of average surprise in the information received.

As an example, assume a message has three possible outcomes: Banana (60% likely), Taco (30%), and Ice Cream (10%). The entropy (the average expected information content) of this message is:

$$.6 * -\log_2(.6) + .3 * -\log_2(.3) + .1 * -\log_2(.1)$$

because we expect banana 60% of the time with an information content of .74 bits ($- \log_2 (.6)$), taco 30% of the time with an information content of 1.74 bits, and ice cream 10% of the time with the information content of 3.32, we expect that any particular message drawn from this distribution will have an average information content of 1.30 bits, per equation (2).

Just like information content, entropy is at its minimum—zero—when the message has only a single outcome ($1 * \log_2(1 / 1)$).[24] The greatest entropy for a set of X possible outcomes occurs when each outcome is equally likely to occur.[25] Moreover, all else being equal, a message with greater possible outcomes will convey more information than a message with fewer potential outcomes.[26]

An additional example will further clarify the idea of entropy. On average, we would expect to receive very little information from flipping a weighted coin that will land heads up 99.99% of the time. Restated, the coin toss conveys little information because it is almost certain to land on heads (with tails occurring only once every 10,000 flips). Using Equation 2, the entropy (expect information content) of flipping the weighted coin is .0015 bits, compared to the 1-bit maximum entropy of a fair (50/50) coin toss .[27] The higher entropy, fair-flip coin has a significantly greater element of surprise on any given coin toss relative to the weighted coin.

24. Hankerson et al., *supra* note 12, at 41.

25. Ali Mehri & Amir H. Darooneh, *The Role of Entropy in Word Ranking*, 390 PHYSICA A 3157, 3157 (2011); Hankerson et al., *supra* note 12, at 40.

26. Pierce, *supra* note 5, at 23.

27. I note that at least some researchers argue that certain coin tosses are not actually 100% random, as (for example) a heavier side of a coin may face down a disproportionately large percentage of the time. *Think a Coin Toss Is a 50/50 Shot? Think Again!*, RIPLEY'S BELIEVE IT OR NOT (May 3, 2018), https://www.ripleys.com/weird-news/coin-toss-or-not/ [https://perma.cc/JPH4-25B5].

III. MESSAGE PROBABILITY

To this point, we have defined entropy and information content as a function of a message's probability (or the amount of "surprise" received with the message). However, no discussion of how to calculate the relevant probabilities has yet been given. This section addresses this point by discussing different methods of calculating relevant probabilities and the underlying information used to calculate these probabilities.

A. Entropy of Language using First, Second, Third, etc. Order Analysis

Information received by any specific message is a function of how the probability of receiving that message is calculated. For example, assume that we receive a series of letters with no spaces. If we assume that the next letter received is completely random and that each letter is equally likely to occur, the probability any letter in the alphabet is 3.85% (1 / 26). Under Equation 1, any of these messages conveys 4.70 bits of information. This expectation of completely random letter messages is obviously unrealistic as certain letters are more/less likely to occur relative to others in common usage.

From this recognition, we can calculate the likelihood that a given letter comes up in actual usage of the English language as shown through some specific text (with the selection of text discussed further later). In this situation, the probability of any specific character being an "E" (12.02%) is much higher than "W" (2.09%) as "E" is a much more common letter.[28] This is a first-order approximation of what letter we expect next. That approach represents an improved method of identifying message probability, but the model could still be improved.

For example, we can evaluate the likelihood of a particular letter being transmitted as a function of what the preceding letter was (i.e., the context). In English, certain letter pairs occur more often than others. For example, a "T" followed by an "H" is common, whereas a "T" followed by a "Q" is relatively rare.[29] We call these multi-unit communications N-grams, where N is the number of units (e.g., letters) considered. The two-letter communication is a 2-gram or bigram and is considered a second-order analysis. To the extent that probability data is available, the N-gram approach can be expanded to an Xth order analysis (e.g., a fifth-order

28. *English Letter Frequency*, *supra* note 21.

29. "TH" is the most common bigram in the English language, occurring approximately 168 times per thousand words. JAMES GLEICK, THE INFORMATION: A HISTORY, A THEORY, A FLOOD 227 (2011).

analysis could tell us the likelihood that an "O" follows the letters "NACH").

We can also extend this approach to words as discrete messages. A first-order analysis gives the probability of receiving any word as a message based on how common that word is in English. For instance, the highest probability words are "the," "be," and "to" (in descending order).[30] A second-order analysis would consider the likelihood that a particular word is received after one prior word.[31] If the first word is "Thank," the likelihood that the next message is "you" is much greater than "octopus." Again, this analysis can be extended to an Xth order analysis if data is available.

B. The Reference Corpus of Information

The prior subsection described *how* probabilities underlying information content are calculated but did not discuss the information used to derive those probabilities. This data is ascertained from a large corpus of communications related to the information of interest. For example, to identify the information content of verbal communications, a body of prior conversation transcripts could disclose relevant probabilities. This information could include word occurrence data (for first-order analysis), word occurrence probability after any given term (second-order analysis), and so on.

Relevant to the current discussion, the selection of a particular corpus of earlier messages corresponds to the reader's expected knowledge. Thus, calculating entropies associated with technology-field-specific terminology would favor drawing probabilities from in-field patents or technical journals. In contrast, if the expected application of entropy focuses on general (e.g., nonfield specific) terminology, the favored corpus should be a large body of general (language specific) text.

IV. AMBIGUOUS PATENT LANGUAGE

This section presents a proposed use of information-theoretic analysis within patent law. It begins by discussing a potential manner to identify ambiguous language and ascertain a set of potential meanings ascertained. From there, we quantify the likelihood that a particular use of

30. *Most Common Words in English*, WIKIPEDIA, https://en.wikipedia.org/wiki/Most_common_words_in_English [https://perma.cc/6UVC-SS76].

31. *See* Joseph Scott Miller, *Reasonable Certainty & Corpus Linguistics: Judging Definiteness After* Nautilus *&* Teva, 66 U. KAN. L. REV. 39, 86–89 (2017) (discussing possible uses of n-gram analysis in claim construction).

a term is intended to convey a particular meaning. Then in possession of a specific set possible meanings (outcomes) and the probabilities that each outcome will occur, we can determine the entropy of any given word. The literature explains that a higher entropy corresponds with greater ambiguity.

Part B explores the application of this insight to patent law and lexical ambiguity in patent documents. It addresses different situations wherein this methodology may be beneficial, including patent acquisition and prosecution.

A. Using Information Theory to Identify Ambiguities

An ambiguous word may be intended to convey one of several meanings that require additional information (e.g., context) to properly identify.[32] In information theory, these different meanings can be viewed as different outcomes (like a coin coming up heads) with each occurring at a given probability. Where a word can have *many* different meanings, entropy related to *which* meaning is intended is high, and more bits of information are required to ascertain the intended meaning.[33] Restated, where entropy is high, more "yes/no" questions must be asked to identify which of many meanings are intended.[34]

Given the need for additional information to disambiguate uncertain terminology, one might argue that ambiguous language hinders communication. However, the literature posits that ambiguity is actually a good thing from an efficiency perspective. Initially, assuming that context provides information about a word's intended meaning, there is no need to create new words to disambiguate meaning where that goal is already achieved by context.[35] Second, short and simple words are easier to communicate and understand, such that re-use of these terms is efficiency enhancing where any disambiguation is achieved via context.[36] If we avoided the re-use of short and efficient terms, we would have to

32. Piantadosi et al., *supra* note 19, at 280 ("Ambiguity is a pervasive phenomenon in language which occurs at all levels of linguistic analysis. Out of context, words have multiple senses and syntactic categories, requiring listeners to determine which meaning and part of speech was intended.").

33. Peter McMahan & James Evans, *Ambiguity and Engagement*, 124 AM. J. SOCIO. 860, 872 (2018) ("As a measure of ambiguity, information-theoretic entropy concisely summarizes both the probability and diversity of meaning distributions. The entropy of a word's possible meanings efficiently models a reader's uncertainty about its sense in a given context.").

34. Piantadosi et al., supra note 19, at 283 ("[W]hen the entropy is high, more bits of information are needed to disambiguate which of the possible meanings was intended.").

35. *Id.* at 281.

36. *Id.*

invent increasingly long and communicatively labor-intensive words to convey particular meaning.

Piantadosi, Tily, and Gibson empirically address the position that language has evolved such that short and easy-to-communicate words tend to have multiple meanings. Their research found that relatively short words have more different meanings than longer words (i.e., greater ambiguity).[37] This analysis, however, only used the aggregate number of different meanings as a dependent variable.[38] Unfortunately, that variable omits important nuance, as *it does not consider how often each of the different meanings are used*.

To account for this, they propose using an entropy-based measure of ambiguity to calculate a single word's entropy using information about how many different definitions a word has and how often the term is used for each distinct definition.[39] They were unable to employ such an entropy-based measure in their work because they lacked data about how often a particular definition of a word was intended in communication.

McMahan and Evans continued in this general area by attempting to identify a word's different meanings *plus* the probability that the speaker meant to convey a specific meaning. To this end, they presented an entropy-based methodology to quantify a term's ambiguity (identifying potential intended meanings and probability of a specific meaning) using word occurrence data (i.e., "quantifying the uncertainty of meaning imparted by any given word as encountered in a text.").[40]

They initially identified all possible meanings that a word might have. To this end, they consider "[a] term with n synonyms [as shown via a thesaurus to be] associated with n + 1 distinct meanings." Specifically, the word could take on the discrete dictionary meaning of each of its synonyms—which each provide a slightly different connotation—or the dictionary definition of the word itself. Each of these different words represents a generally interchangeable term that may convey a slightly different (and thus, slightly ambiguous) meaning.[41] This provides the different meanings (i.e., outcomes) of a specific word, but it does not provide the probability that any given meaning is intended.

McMahan and Evans quantified the probability that a word is intended to mean its dictionary definition or the (slightly different) definition of one of synonyms by counting the number of times the word

37. *Id.* at 285-86.
38. *Id.* at 286–87.
39. *Id.* at 283, 286.
40. McMahan et al., *supra* note 33, at 871–72.
41. *Id.* at 874.

or one of its synonyms appear in a corpus text. For example, the ambiguity of the term "run" could be ascertained by identifying the number of occurrences of that particular word over the total number of occurrences of the word "run" and each of its synonyms.

For example, assume (unrealistically) that "run" only has four thesaurus synonyms: jog, sprint, dash, and trot. We then look at how often the reference corpus uses these terms. Let's assume that "run" appeared 22 times, jog (6), sprint (2), dash (5), and trot (11). From this, we can calculate that run appeared 47.8% of the time among these five synonymous terms, jog 13.0%, and so on. Then using Equation 2, we can calculate the entropy of run's ambiguity as 1.93 bits. Remember that a high degree of entropy equates to high ambiguity. This calculation does not, however, consider the context in which the words are used.

It is possible to analyze a particular word and its synonyms in a specific, relevant context. For example, in a hyper-specific instance, McMahan and Evans consider the word "hibernate" (and its synonyms: slumber, kip, rest, nap, sleep, bundle, and estivate) in the phrase "pikas don't *hibernate* through winter." Here, we could analyze how often "hibernate" and its synonyms are used in this exact phrase in the relevant corpus and calculate the probability that each such term would be used in that phrase. This gives us a more nuanced analysis of when a given term is meant to be used in a specific way, but it would require a large corpus to provide a large enough data set.

It would likewise be possible to identify a term and its synonyms' use in broader situations, such as the use of "hibernate" near the word "winter." This would largely include the use of the term for animals sleeping through the cold months but exclude uses where someone is described as eating too much and then hibernating on the couch (a distinct usage).

Employing the above methodology, the researchers validated their approach by comparing their measured entropy-based ambiguity measurement to the human participant's "individual uncertainty on the basis of whether or not they were confident that they understood what a given term meant in the context of a displayed sentence." [42] Using this approach, the researchers found "a strong support for our association between measured ambiguity and individually perceived semantic uncertainty."[43] Restated, their entropy-centric ambiguity metric correlated with human identification of ambiguity. While this only represents one

42. *Id.* at 877.
43. *Id.* at 907.

(of potentially many) manners to employ information theory to identify ambiguity, their results show that the general approach can effectively identify uncertainty in language.

B. Ambiguity and Patent Law

McMahan and Evans's methodology has applications within patent law and patent claim analysis. Initially, their approach could automate the identification of specific words in patent claims that are particularly ambiguous. Such a process could be used to intentionally create or diminish uncertainty when drafting a claim. Likewise, certain firms will target patents for purchase depending on the level of their claim ambiguity. These behaviors affect claim scope, validity, and the innovation sphere.

As an example, claim ambiguity is a benefit for those seeking to extract rents via patent litigation or the threat thereof.[44] Initially, unclear claim language hinders the ability to determine what constitutes infringement ex-ante.[45] This uncertainty incentivizes settlement to avoid unpredictable claim construction and the possibility of having to alter product design to avoid infringement in the face of a detrimental Markman opinion.[46] Claim ambiguity is thus beneficial for certain litigants seeking quick settlements but harmful to the subjects of these lawsuits.

Further, uncertainty has significant effects in patent prosecution. Claims must provide "full, clear, concise, and exact terms as to enable" one having ordinary skill in the art to practice the invention.[47] Failure to do so can lead to claim rejection, necessitating amendment or potentially

44. *See* Jeremiah Chan & Matthew Fawcett, *Footsteps of the Patent Troll*, 10 INTELL. PROP. L. BULL. 1, 4 (2005) ("[T]housands of ambiguous and dubious patents are issued every year, leading to confusion in the scope and coverage of any one patent. For patent trolls, these ambiguous or 'bad' patents are effective weapons." (citation omitted)); *see also* Dargaye Churnet, *Patent Claims Revisited*, 11 NW. J. TECH. & INTELL. PROP. 501, 509 (2013) ("The ambiguity of patent claims has contributed to the emergence of patent trolls. This group, often referred to as 'non-practicing entities,' acquires patents with no intention of practicing the invention. Instead, the troll simply waits for a manufacturer to sufficiently commercialize a product that could arguably read on the troll's patent and then seeks to extract exorbitant licensing fees." (citations omitted)).

45. U.S. GOV'T ACCOUNTABILITY OFF., GAO-13-465, INTELLECTUAL PROPERTY: ASSESSING FACTORS THAT AFFECT PATENT INFRINGEMENT LITIGATION COULD HELP IMPROVE PATENT QUALITY 28 (2013).

46. *Id.* at 32 ("Some economic literature we reviewed suggests that accused infringers have an incentive to settle quickly to avoid the uncertainty of claim construction and high litigation costs, particularly if they face very high costs of changing their products to avoid infringement.").

47. 35 U.S.C. § 112(a).

causing a failure to secure a patent.[48] At a minimum, such delays in patent prosecution impose additional costs on the applicant.

To this point, I have only addressed the potential application of McMahan and Evans's methodology to identify ambiguous terms in a claim. This approach is, however, generalizable to identify the average information content (and thus ambiguity) of an entire claim. For example, Keller calculated sentence-level information content averages on a per-word basis.[49] Specifically, he calculated the per unit entropy (i.e., the per word entropy) for each word in a sentence and then averaged those amounts. For current purposes, this is simply using Equation 2 for each word in a sentence (or paragraph or patent claim) and averaging those amounts.

The validity of this approach is measurable through comparison to several objective metrics. Initially, it would be possible to compare the measured ambiguity of claims in an application to § 112 rejections from the Patent Office's *OCE Office Actions* database. Similarly, given the preference of nonpracticing entities (NPEs) to employ ambiguous patents, our objective metric could be compared to the claim language of NPE-asserted patents from the Lex Machina NPE database.[50] Lastly, claims invalidated on § 112 grounds in litigation grounds could be analyzed for ambiguous content to verify the above-presented metric.

The McMahan Evans metric is, however, subject to several methodological choices and potential limitations of note. It relies on a thesaurus to define all possible meanings that a word may have. This analysis is limited because it is only as good as the thesaurus in use, which raises two distinct issues. First, multi-word terms of art may have a distinct meaning apart from the constituent words' discrete definitions

48. U.S. PAT. & TRADEMARK OFF., U.S. DEP'T COM., MPEP § 2173 (9th ed. Rev. 10.2019, June 2020) ("[C]laims that do not meet this standard must be rejected under 35 U.S.C. 112(b) or pre-AIA 35 U.S.C. 112, second paragraph as indefinite. Such a rejection requires that the applicant respond by explaining why the language is definite or by amending the claim, thus making the record clear regarding the claim boundaries prior to issuance. As an indefiniteness rejection requires the applicant to respond by explaining why the language is definite or by amending the claim, such rejections must clearly identify the language that causes the claim to be indefinite and thoroughly explain the reasoning for the rejection.").

49. Frank Keller, *The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation Against Eye-Tracking Data*, PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 317, 318–19 (2004). Keller bases his approach off of Dmitriy Genzel & Eugene Charniak, *Entropy Rate Constancy in Text*, PROC. 40TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS, at 199 (2002); Dmitriy Genzel & Eugene Charniak, *Variation of Entropy and Parse Trees of Sentences as a Function of the Sentence Number*, PROC. CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING 65 (2003).

50. LEXMACHINA, https://lexmachina.com/ [https://perma.cc/42C3-UC3U].

(and synonyms in the standard thesaurus). For example, "freezer burn" has a distinct meaning from "freezer" and "burn."[51] If the McMahan Evans methodology analyzes these terms separately (instead of looking at synonyms of "freezer burn"), it will introduce noise into the analysis. Using (or automating the creation of) a thesaurus comprising multi-word terms of art can mitigate that noise.

Second, using a standard thesaurus may be over- or under-inclusive regarding the patent's field of art. A word may have nonstandard, field-specific synonyms that a standard thesaurus would not include. That would be under-inclusive. In contrast, a standard thesaurus may include synonyms of a word inapplicable to a given field (i.e., the word would never be used in a particular way in a particular field). That would be over-inclusive. Again, these phenomena may introduce noise into the analysis.

To the extent that these potential noise sources exist, they should not inhibit the ability to evaluate large-scale trends. Noise should cancel itself out and allow the signal (i.e., relevant information) to remain. Further, the McMahan Evans methodology is given only as an example of using information theory to quantify textual ambiguity. Other methods may be employed as appropriate.

Beyond identifying potential sources of noise in the analysis, a note on data preparation is warranted. Words used in a patent may differ in their morphological and inflexional endings but share a common basic meaning (stem). For example, "compute, computes, computed, computing, computer, computation, computerize, or computational" all share a common basic meaning but would be analyzed as discrete terms.[52] Depending on the specific goal of a given project, it may improve accuracy to *stem* each word in a patent and entry in a thesaurus by removing the morphological and inflexional endings (i.e., turn each of the above computer-related terms into "comput").[53]

Furthermore, to the extent that a given project is specifically interested in analyzing potential ambiguities in substantive terms, it may

51. *See*, *e.g.*, U.S. Patent No. 6,020,013 col. 41. 35 (filed Mar. 1, 1999) ("The method of claim 1 wherein the food in the triple seal storage bag is stored in a frozen condition over an extended period of time without the ingress of ambient air which causes freezer burn.").

52. Shannon Brown, *Peeking Inside the Black Box: A Preliminary Survey of Technology Assisted Review (Tar) and Predictive Coding Algorithms for Ediscovery*, 21 SUFFOLK J. TRIAL & APP. ADVOC. 221, 248 (2016).

53. *Id.*; Wenjuan Luo, Fuzhen Zhuang, Qing He, & Zhongzhi Shi, *Exploiting Relevance, Coverage, and Novelty for Query-Focused Multi-Document* Summarization, 46 KNOWLEDGE-BASED SYS. 39 (2013).

be prudent to remove nonsubstantive "stop words."[54] These terms include words like "you," "because," and "will" that are necessary to complete a sentence but lack substantive meaning.[55] While not necessary, this approach may remove unimportant yet ambiguous terms before data processing.

V. QUANTIFYING ORIGINALITY IN PATENT DOCUMENTS

Patented technologies exist on a continuous scale between incremental innovation and groundbreaking technologies.[56] A groundbreaking technology may represent an economically important, drastic innovation.[57] Identifying such technologies and the patents that describe them is valuable. Likewise, identifying incremental innovations may be important to the study of patent thickets and follow-on innovation. The information theory literature provides a stepping-stone toward identifying highly innovative patents.

Dasgupta and Dey propose an information theory-based method to identify highly original documents from a large number of texts.[58] They base their approach on the idea that "a document having high information content is potentially a [highly innovative] document."[59] With this in mind, they proposed that a particularly innovative document is relatively more likely to employ a *unique* vocabulary.[60]

The Dasgupta Dey method quantifies a document's originality through its entropy. They quantify entropy (E_T) of a document as:

$$E_T(p_1, \ldots, p_n) = \frac{1}{\lambda} * \sum_{i=1}^{n} p_i \, (\log_{10} \lambda - \log_{10} p_i) \quad (3)$$

where λ equals the number of words in the document, and p_i measures a particular term's probability of occurring within the corpus of relevant

54. Sam Arts, Bruno Cassiman & Juan Carlos Gomez, *Text Matching to Measure Patent Similarity*, 39 STRAT MGMT J. 62, 64–65 (2018).

55. W. Michael Schuster & Kristen Valentine, *supra* note 1.

56. *See generally* Ron A. Bouchard, Jamil Sawani, Chris McLelland, Monika Sawicka & Richard W. Hawkins, *The Pas De Deux of Pharmaceutical Regulation and Innovation: Who's Leading Whom?*, 24 BERKELEY TECH. L.J. 1461, 1520 (2009); Viral V. Acharya, Ramin P. Baghai & Krishnamurthy V. Subramanian, *Labor Laws and Innovation*, 56 J.L. & ECON. 997, 1007 (2013).

57. Acharya et al., *supra* note 56, at 1007.

58. Tirthankar Dasgupta & Lipika Dey, *Automatic Scoring for Innovativeness of Textual Ideas*, THE WORKSHOPS OF THE THIRTIETH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE KNOWLEDGE EXTRACTION FROM TEXT: TECHNICAL REPORT (2016).

59. *Id.* I replaced the word "novel" in this quote to avoid its patent law-specific connotations, which were not intended by Dasgupta and Dey.

60. *Id.*

documents.[61] Equation 3 essentially averages the information content of each word in a document relevant to a body of reference documents.

Their team applied this methodology to approximately 1,500 entries from a "real world innovation contest" and then compared their results to (1) innovativeness[62] rankings (0–5) made by human experts and (2) industry-standard, automated benchmarks (e.g., Cosine Similarity and Kullback-Leibler divergence) for identifying innovative text.[63] The results found their entropy-based method to outperform the benchmarks in matching the results from the human expert ratings.[64]

Dasgupta and Dey's results show that entropy-based metrics can effectively identify originality in text documents. This finding has a variety of patent-centric applications. Parties may attempt to identify groundbreaking technologies for investment or follow-on innovation. To this end, firms would be wise to identify patent documents with relatively original content (e.g., high-entropy as per the Dasgupta-Dey method). This proposition, however, necessitates a second query: firms should look for original content *relative to what*? This calls into question the textual corpus used to identify the unique vocabulary indicative of originality.[65]

To this end, a reference corpus could be amassed comprising a significant body of earlier-filed patent documents from a related United States Patent Classification or Cooperative Patent Classification field.[66] This corpus could be supplemented or replaced with prior academic literature in a similar field. However, an automated comparison of a patent document against these earlier documents in a related field only identifies original contributions. An original patent document is not necessarily the same as valuable or groundbreaking patent filings.

Textual originality *may* suggest value or original ideas, but it is unlikely to be sufficient to establish these new ideas as valuable. For instance, if I were to claim, "A tasty treat consisting of a turtle-shaped popsicle comprising Xylene and moon rocks," it would likely prove very original but also very unsuccessful as a product and, thus, not valuable.

61. *Id.* To calculate the probability of a particular term occurring within the corpus of relevant documents, Dasgupta and Dey employ a metric called *Inverse Document Frequency*. This metric's calculation is fully described in their article, but a complete presentation is beyond the scope of this essay.

62. Again, Dasgupta and Dey use the term "novelty," which is avoided here because of its patent-specific connotations.

63. *Id.*

64. *Id.*

65. *Id.*

66. The analyzed patent document and corpus of reference documents may be stemmed and have stop words removed as per footnotes 52–55 and related text.

Again, a high degree of originality may be indicative of (or be a necessary component of) high value, but originality is not sufficient to show value. Thus, a secondary metric for identifying novel *and* valuable patents is necessary.

To this end, it will be beneficial to run a second analysis on highly original patents, except this time to compare the patent to a corpus of *later-filed* patent documents. Research shows that continued research in a particular field demonstrates the value of any earlier-filed patents.[67] Thus, if a patent has a *low* originality score against later-filed patents, this shows value because others are continuing in that field of research. Accordingly, a patent with high originality versus earlier-filed patent documents and low originality against later-filed patent documents is likely to indicate value *and* originality.

This method of identifying original disclosures in patent documents is testable in several manners. The literature instructs that groundbreaking or particularly novel patents are relatively more valuable—all else equal.[68] Thus, our originality metric from Dasgupta and Dey could serve as an independent variable to predict patent value. Consistent with past research, patent value could be measured by forward citations,[69] through stock market analysis,[70] or maintenance fee payment.[71]

In contrast to the above-proposed use of information theory-centric metrics to identify original patent filings, the same approaches may prove valuable in identifying fields with minimal originality. Significant modern research focuses on identifying and remedying *patent thickets*[72]—a

67. *See*, *e.g.*, ORG. FOR ECON. CO-OPERATION & DEV. [OECD], *Patent Statistics Manual*, at 138 (2009); Peter A. Malaspina, *Patent Citation Analysis and Patent Damages*, 18 CHI.-KENT J. INTELL. PROP. 232, 234 (2019).

68. Sannu K. Shrestha, *Trolls or Market-Makers? An Empirical Analysis of Nonpracticing Entities*, 110 COLUM. L. REV. 114, 142 (2010) (Originality indicates patent value); John R. Allison Mark A. Lemley, Kimberly A. Moore & R. Derek Trunkey, *Valuable Patents*, 92 GEO. L.J. 435, 450 (2004) (describing the literature's use of originality to indicate patent value).

69. Dietmar Harhoff, Francis Narin, F.M. Scherer & Katrin Vopel, *Citation Frequency and the Value of Patented Inventions*, 81 REV. ECON. & STAT. 511, 515 (1999); *see also* Jean O. Lanjouw & Mark Schankerman, *Characteristics of Patent Litigation: A Window on Competition*, 32 RAND J. ECON. 129, 129–151 (2001); Tania Bubela, E. Richard Gold, Gregory D. Graff, Daniel R. Cahoy, Dianne Nicol & David Castle, *Patent Landscaping for Life Sciences Innovation: Toward Consistent and Transparent Practices*, 31 NATURE BIOTECHNOLOGY 202, 205 (2013) ("Studies have shown that the number of citations made to a patent is related to the private economic value of that patent.") (internal citations omitted)).

70. Leonid Kogan, Dimitris Papanikolaou, Amit Seru, & Noah Stoffman, *Technological Innovation, Resource Allocation, and Growth*, 132 Q.J. ECON. 665, 666 (2017).

71. Gregory R. Day & W. Michael Schuster, *Patent Inequality*, 71 ALA. L. REV. 115, 122 (2019).

72. Bronwyn H. Hall, Georg von Graevenitz & Christian Helmers, *Technology Entry in the Presence of Patent Thickets [Our Divided Patent System?]*, 73 OXFORD ECONOMIC PAPERS 903

situation where many firms own many overlapping patents within a discrete technological field.[73] These thickets are believed to impair innovation and competition.[74] The current originality metrics can identify specific technological sectors with very low patent originality, which may indicate largely redundant/related patents documents and patent thickets. Specifically, where a significant portion of recent patents granted in a particular field are largely unoriginal, patentees may be obtaining the many related and overlapping patents indicative of a thicket. This approach to identifying patent thickets could be verified through comparison to earlier metrics on the topic.[75]

VI. CONCLUSION

This essay proposes new manners of analyzing patent text through information theory-centric metrics. Based on prior scholarship, new methods of analyzing claim ambiguity and originality have been proposed. This discussion is, however, only a starting point. Empirical research should ascertain the value of the proposed metrics. Further, future methodologies presented in the information theory literature should be reviewed for relevance to patent analysis.

(2021); Georg von Graevenitz, Stefan Wagner & Dietmar Harhoff, *How to Measure Patent Thickets—A Novel Approach*, 111 ECONOMICS LETTERS 6 (2011).

73. Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 INNOVATION POL'Y & ECON. 119, 119 (2000); R. Polk Wagner, *Information Wants to Be Free: Intellectual Property and the Mythologies of Control*, 103 COLUM. L. REV. 995, 997 n.6 (2003). Similar situations have been referred to as "anticommons." Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCI. 698, 698 (1998).

74. Day et al., *supra* note 71, at 154.

75. *See, e.g.*, von Graevenitz et al., *supra* note 72 (discussing a citations-based approach to identify thickets).