

Draft in preparation for Peter S. Menell & David L. Schwartz (eds.), *Research Handbook on the Economics of Intellectual Property Law* (Vol. II -- Analytical Methods) (Edward Elgar Publishing, forthcoming 2016).

EMPIRICAL STUDIES OF COPYRIGHT LITIGATION

MATTHEW SAG*

ABSTRACT

This essay summarizes the current state-of-the-art in the application of observational empirical tools to the study of copyright litigation. As a field, Empirical Legal Studies (or ELS) is relatively young, and the subfield of ELS applied to copyright litigation is in its infancy. The existing literature can be classified into three subcategories: studies of judicial behavior, studies of filing data and the day to day life of copyright litigation, and studies of particular aspects of copyright doctrine. Because the field is so small, this essay summarizes and evaluates all of the extant literature. It concludes with three guidelines for researchers planning to undertake an empirical study of copyright litigation. These guidelines or best practices relate to (1) the importance of open data, (2) the need to take selection effects seriously without abandoning empirical studies of litigation altogether (3) understanding that the future is out of sample—i.e., that empirical analysis may well identify patterns in a given data set, but that predictions about the future are *arguments* based on empirical conclusions, they are *not empirical conclusions* themselves.

* Professor, Loyola University Chicago School of Law; Associate Director for Intellectual Property, Institute for Consumer Antitrust Studies. © Matthew Sag 2015. Comments on this draft essay should be emailed to msag@luc.edu.

Draft in preparation for Peter S. Menell & David L. Schwartz (eds.), *Research Handbook on the Economics of Intellectual Property Law* (Vol. II -- Analytical Methods) (Edward Elgar Publishing, forthcoming 2016).

EMPIRICAL STUDIES OF COPYRIGHT LITIGATION

MATTHEW SAG*

INTRODUCTION

This essay summarizes the current state-of-the-art in the application of observational empirical tools to the study of copyright litigation. As a field, Empirical Legal Studies (or ELS) is relatively young, and the subfield of ELS applied to copyright litigation is in its infancy.

Broadly speaking, the existing literature can be classified into three subcategories: studies of judicial behavior, including the effect of judicial ideology on decision making (Sag, Jacobi & Sytch 2009), the effect of judge assignment in the European Court of Justice and modes of analysis associated with particular case outcomes in that context (Favale, Kretschmer & Torremans 2016), and assessments of judicial expertise in copyright cases (Ford 2006); studies of the day to day life of copyright litigation from the time cases are filed to their ultimate disposition (Landes 2004, Sag 2013, Cotropia & Gibson 2014, Sag 2015 and Sag 2016); and studies of particular aspects of copyright doctrine—this category can be further subdivided into event studies tracing the impact of significant Supreme Court decisions (Barnes 2000, Liu 2012) and broader studies the workings of copyright doctrine in practice, focusing particularly on fair use (Beebe 2008, Netanel 2011 and Sag 2012), substantial similarity (Lippman 2013, Rogers 2013, Netanel & Sag (unpublished)), and publication (Gerhardt 2011).

* Professor, Loyola University Chicago School of Law; Associate Director for Intellectual Property, Institute for Consumer Antitrust Studies. © Matthew Sag 2015. Comments on this draft essay should be emailed to msag@luc.edu.

Judicial Behavior – The Effect of Ideology in Copyright Cases

Sag, Jacobi & Sytch (2009)

Sag, Jacobi & Sytch (“SJS”) (2009) examined the effect of ideology on the voting behavior of individual Justices of the US Supreme Court in IP cases from 1954 to 2006. Contrary to the common view, they found that judicial ideology is a significant determinant of voting in IP cases; specifically the more conservative a justice is, the more likely he or she is to vote in favor of recognizing and enforcing rights to intellectual property.

Although bucking conventional wisdom among IP scholar to some extent, SJS found that the size of the effect of ideology in IP cases was significantly lower than in some other categories of cases involving prominent social issues. They also found that although this effect of ideology was discernable for all types of IP, the justices were significantly more likely to vote for the IP owner in copyright cases compared to the default category of patent cases.

SJS’s finding that the outcomes of IP cases are influenced by judicial ideology as measured on the traditional liberal-conservative scale was limited to Supreme Court cases. Examining a set of over 200 fair use determinations made between 1978 and 2005, Beebe (2008) found no significant relationship between a judge's ideology and his or her adjudication of the fair use defense. Beebe’s study of the application of the *Polaroid* factors in trademark cases likewise failed to find a significant effect (Beebe 2006). The null results on narrowly focused studies of lower court decision making do not negate the general findings made by SJS with respect to the Supreme Court. They do suggest, however, that other factors, such as the demands of case management, may make the effect of ideology less apparent in lower court decision-making, compared to the Supreme Court.

Judicial Behavior – Judge Assignment and Reasoning

Favale, Kretschmer & Torremans (2016)

A forthcoming paper by Marcella Favale, Martin Kretschmer and Paul C. Torremans (FKT) evaluates the reasoning of copyright cases decided by the European Court of Justice between 1992 and 2012. FKT undertake a systematic content analysis of the 40 cases relating to copyright and related rights and 9 additional database right cases in this period. They also note the non-random assignment of Advocates General and Judges to these cases and conclude that although private law and intellectual property law expertise is almost entirely missing from the EJC, the Court has nonetheless “developed a mechanism for enabling judicial learning through the systematic

assignment of cases to certain Judges and AGs. FKT observe that, contrary to the literature claiming that the ECJ is pursuing an activist, harmonizing agenda in copyright and related rights, about half of the judgments in this period narrowed rather than widened the scope of copyright protection. Going deeper into the text, the discern certain patterns in the arguments used by the Court to justify its rulings. Concepts such as ‘fair balance’ of rights and interests, free competition, fostering technological development, ‘adequate’ rightholder protection, and ‘minimal harmonization’ are associated with outcomes against the rightholder. In contrast, pro-rightholder rulings, tended to invoke unmodified calls for harmonization as well as ‘high protection for the copyright holder’.

Influence of Appellate Decisions in Copyright Textbooks

Ford 2006

William Ford (2006) undertook an extensive comparison of the representation of cases decided by different US Courts of Appeal in major copyright law textbooks in an attempt to measure the comparative expertise of each circuit in the field of copyright. Ford traced the citation of copyright cases from 1894 to 2004 in casebooks published from 1940 to 2003. Ford concludes that the Second and Ninth Circuits attract the most appellate level copyright litigation and that, even normalizing for caseload, Second Circuit cases are overrepresented in copyright casebooks. Whether, as Ford suggests, this reflects the Second Circuit’s “experience and influence” is unclear.

Docket Studies – Nature of Suit Coding

Sag 2013

The first step in any empirical study of copyright litigation is to identify which cases are, and are not, copyright cases. Studies focusing on written opinions need to develop criteria for making this determination that is consistent with their research question. For some purposes this might involve excluding opinions on procedural matters, but for other purposes – e.g. a study of judicial ideology in different contexts – such exclusion would be unwarranted. Researchers analyzing litigation filings usually rely on the Nature of Suit (“NOS”) coding in the Public Access to Court Electronic Records (“PACER”) database or the Administrative Office of the United States Courts (“AO”) data and hosted by the Inter-University Consortium for Political and Social Research. Datasets such as Bloomberg Law are derived from the PACER data. This raises the question, how reliable are the NOS codes? The two most obvious potential problems are that the litigants themselves are responsible for indicating the relevant NOS code at the time of filing and that the filing party is only

allowed to pick one subject matter classification. An unpublished paper by Matthew Sag (2013) investigates this question and concludes that relying on the NOS coding is satisfactory for many purposes, but that it may lead to a systematic bias and undercounting for others. Sag worked backwards from written opinions available in Lexis to the case filings which lead to those opinions to show that the NOS code for copyright captured about 80% of district court written opinions that had something to do with copyright.¹ Of the remaining 20%, almost half were filed under the NOS code for trademark and a quarter were filed under the NOS code for patent. The remaining quarter were filed under almost every category imaginable including: Contract, Cable/Sat TV, Other Statutory Actions, Insurance, Assault, Libel, & Slander, Other Personal Property Damage, Civil Rights, Fraud, Personal Injury and even some criminal filings.

Docket Studies – Who, What, When, Where?

Cotropia & Gibson 2014

Cotropia and Gibson’s 2014 investigation of copyright dockets and complaints provides a number of invaluable insights and attempts to answer questions such as “Who files copyright cases? What kinds of works are involved—software, books, music, film? What claims are made? How many cases go to trial? What remedies are awarded? Are some courts more favorable to claimants?” Cotropia and Gibson reviewed the complaints and the docket history of a random sample of almost one thousand copyright cases filed between 2005 and 2008. Cotropia and Gibson’s diligent research makes it possible to say something about party characteristics, associated industries, litigated subject matter, the statutory basis for infringement claims and remedies sought in filed cases, not just cases that end up being reported in Westlaw or Lexis. To see just how important this is, consider Table 1 (below).

¹ The universe of cases was determined by searching the Lexis district court database for all cases within a specific date range with word “copyright” and then reviewing each case to determine whether it actually addressed a claim of copyright infringement. *Id.*

Table 1 The Termination of Copyright Cases Filed Between 2005 and 2008

Nature of Termination	Prevailing Party Ascertainable		Prevailing Party Not Ascertainable		Total
	Plaintiff	Defendant	Neither Party	Agreed Judgment	
Trial	8	4			12
Summary Judgment	4	7			11
12(b)(6)		9			9
12(c)		2			2
Default Judgment	32				32
Lack of Jurisdiction		2			2
Settlement	1		127		128
Agreed judgment				74	74
Voluntary Dismiss			149		149
Other Dismissal	2	4	13		19
Total	47	28	289	74	438

Data Source: Cotropia and Gibson (2014) (<http://copyrightlawdata.com>). Additional calculations by the author. Excludes filesharing and unterminated cases.

Table 1 is based on Cotropia and Gibson’s publicly available data, excluding file-sharing cases and 7 cases that had not yet terminated as of 2014. As the table shows, only 12 out of these 438 cases ended in a trial; and only a further 11 ended in summary judgment. The upper-left quadrant of Table 1 accounts for almost all of the case in which it was possible to ascertain whether the plaintiff or the defendant was the prevailing party—this amounts to 68 cases out of 438 (or 15.5%). The lower-right quadrant of Table 1 accounts for all those cases in which the identity of the prevailing party is opaque, i.e., cases terminated by settlement, agreed judgment, voluntary dismissal or some other kind of dismissal—this amounts to 370 cases out of 438 (or 84.5%). It should be evident that factoids such as ‘plaintiffs won two of three trials’ or ‘defendants won seven of eleven summary judgment motions’ tell us nothing about the real world success or failure of copyright litigants. The ratio of known to unknown is too small and there is no reason to think that cases that ended in a trial or summary judgment are representative of the broader universe of cases filed. I will return to this problem of the selection of disputes for litigation (also referred to as the ‘selection effect’ or ‘selection bias’) throughout this essay.

By delving deep into the actual dockets of copyright cases, Cotropia and Gibson are able to provide a much richer view of the life-cycle of copyright cases than had been previously available. In fact the only previously published empirical data on the termination of copyright cases consisted of annual summary data extracted from the AO database for the period 1987 to 2000 (Landes 2004). It would be redundant to

repeat all of Cotropia and Gibson findings in this essay. However, three notable conclusions are worth reprising.

First, copyright cases vary significantly in intensity. The majority of cases do not appear to go much further than the initial filing and never give rise to a single dispositive motion by either side. However, only a small percentage of cases contained only one dispositive motion. The implication of this variation in intensity is that simply counting case filings is not a great measure of how much copyright litigation occurs in a given district or a given year.

The second major takeaway from Cotropia and Gibson's study is that copyright litigation in the sense of taking a case all the way to trial is not the 'sport of kings', it is mostly the province of small firms. Cotropia and Gibson found that, setting aside internet file-sharing cases, small firms were "far and away the most common litigants, constituting the largest plaintiff 55.73% of the time and the largest defendant 73.93% of the time." They found that half of the cases they refer to as 'commonplace' featured small firms as both the biggest plaintiff and the biggest defendant.

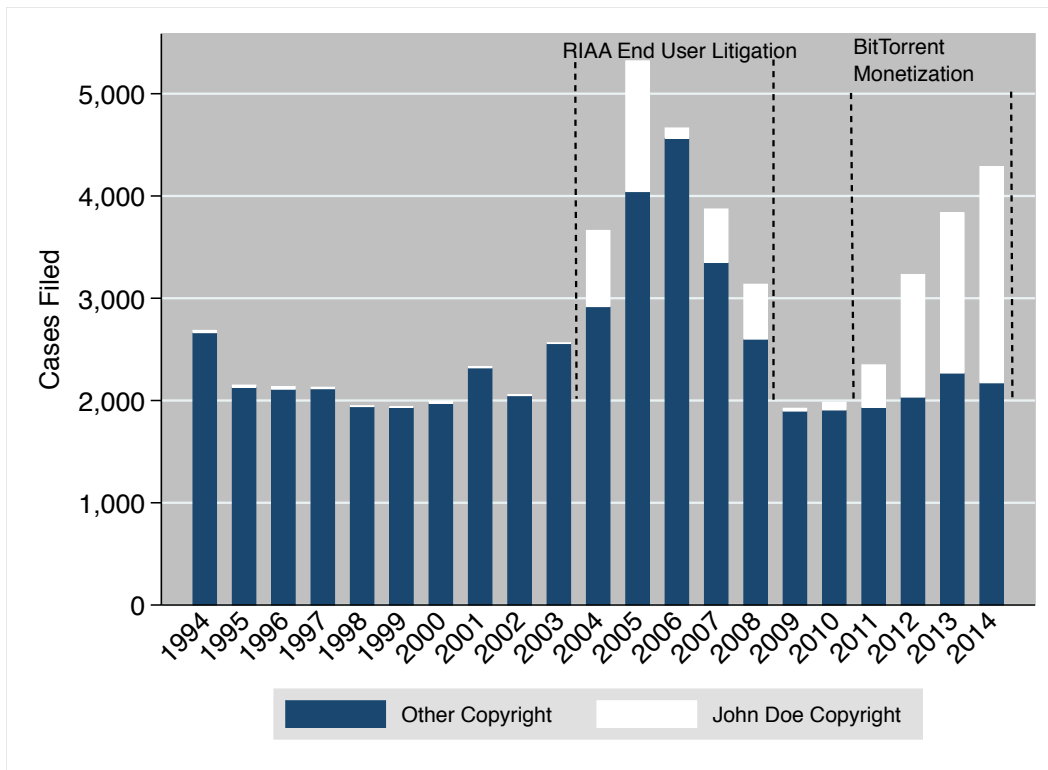
The third significant observation made by Cotropia and Gibson is that low-IP industries and works account for a surprisingly high proportion of copyright litigation. The authors give a detailed account of the industry classifications of the parties. Excluding filesharing cases, the most common pairing of plaintiff and defendant in the Cotropia and Gibson sample was 'music' versus 'food and drink'—i.e., music publishers suing restaurants and bars for unauthorized public performance—these pairings account for almost 14%. The next most common pairing consisted of cases where members of the apparel, fashion or textiles industries were both the lead plaintiff and the lead defendant. This pattern describes just over 10.5% of the non-filesharing cases. This is somewhat surprising because copyright law offers such thin protection for these industries. The next most common pairs were 'software versus software', 'architecture versus architecture', 'music versus music', 'publishing versus publishing', 'film and television versus film and television' and 'industrial design versus industrial design', accounting for almost 9%, almost 8%, almost 7%, almost 6%, almost 5% and about 4.5% of the non-filesharing cases, respectively. Determining whether these patterns are representative of other time periods will require further investigation. Understanding how different industries use copyright and use litigation is also a priority for research.

Pornography, Filesharing and John Doe Litigation

Sag 2015

Matthew Sag (2015) demonstrated how the rise of Internet filesharing has transformed copyright litigation in the United States. Copyright infringement actions against anonymous illegal file sharers typically begin as “John Doe” lawsuits. Sag systematically reviewed all copyright owner lawsuits against “John Doe” or “unknown” or otherwise unidentified defendants between 1994 and 2014 to determine which of those cases related to pornography. A task that was neither trivial, nor edifying. As depicted in Figure 1 (below) The John Doe phenomena can be segmented into two distinct phases: first, the Recording Industry Association of America (RIAA) End-User Litigation era, 2004—2008; and second, the BitTorrent Monetization era, 2010—2014.

Figure 1: Copyright Cases Filed in U.S. District Courts (1994—2014)



Source: Administrative Office of the U.S. Courts, PACER records, 1994—2014. Figure originally published in Sag (2016).

MATTHEW SAG, EMPIRICAL STUDIES OF COPYRIGHT LITIGATION

The RIAA announced its intention to target the end-users of filesharing software in 2003, however the litigation only began in earnest in 2004. By the end of 2008 the RIAA's campaign had largely subsided. The "BitTorrent Monetization" era began in 2010 and continues into the present day. In the BitTorrent monetization era, John Doe litigation is closely associated with pornography, however the percentage of pornography cases has varied over time, from 70% in 2010, 93% in 2011, 85% in 2012, 69% in 2013 and most recently, 88% in 2014.

The data shows that the nature of John Doe litigation has changed significantly since the beginning of the "BitTorrent Monetization" era in 2010. From 2010 to 2012 these cases relied on permissive joinder and were typically filed in the form "*XYZ Copyright Owner v. John Does 1 to 1000*". District Court judges appear to have grown more skeptical of the propriety of litigation in this form over time, and consequently the average number of John Does per suits has been declining. By 2014 the model had shifted almost entirely away from mass-joinder to suits against individual unnamed defendants.

Sag concludes that although illegal filesharing is obviously a widespread problem for the film, television, music and software industries, filesharing litigation is no longer a broad-based phenomenon. The trend from 2012 to 2014 is one of increasing concentration of plaintiff activity. In fact, the pornography producer Malibu Media is such a prolific litigant that in 2014 it was the plaintiff in over 41.5% of all copyright suits nationwide. Sag's interpretation of this data is that John Doe litigation is not a general response to Internet piracy; it is a niche entrepreneurial activity in and of itself.

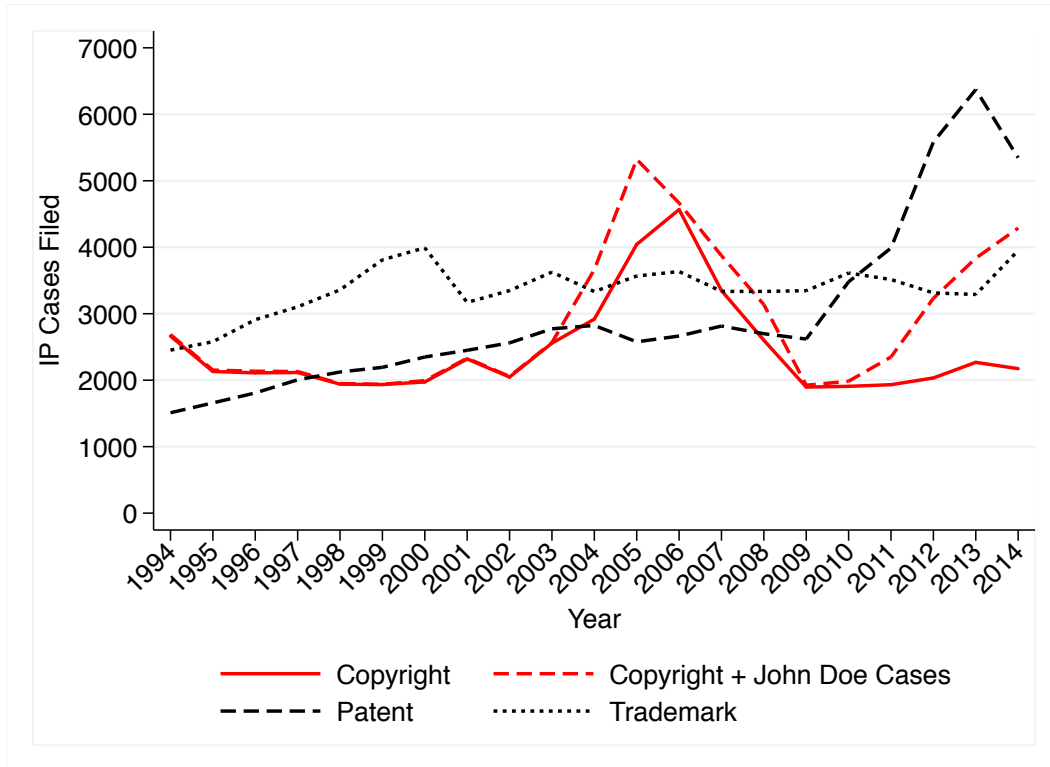
Docket Studies – Broader Trends in Copyright Litigation

Sag 2016

In a study that compliments Cotropia and Gibson discussed above, Matthew Sag (2016) reviews broad trends in federal IP litigation over a longer period. Sag's data is derived from PACER and contains entries for more than 190,000 individual copyright, patent and trademark cases filed in United States District Courts from 1994 to 2014. Sag examines the subject matter, geographical and temporal variation in IP litigation in a way that places trends in copyright litigation in the context of all federal intellectual property litigation. For example, Sag shows that, but for litigation against the uses of Internet filesharing software (the John Doe cases), the annual rate of copyright litigation would have been in slight decline in the period 1994 to 2014. Figure 4 (below), reproduced from Sag's original article, compares the number of cases filed in federal copyright, patent and trademark suits between 1994 and 2014. The trends can be summarized as follows: at the same time as patent litigation has

been sharply increasing, trademark litigation has increased only slightly (it actually peaked in 2000); meanwhile copyright litigation peaked in 2005 during the RIAA end-user litigation campaign and is peaking again due to BitTorrent monetization lawsuits.

Figure 2 Copyright, Patent and Trademark Filings 1994—2014 (Cases)



Source: Administrative Office of the U.S. Courts, PACER records, 1994—2014.
Figure originally published in Sag (2016)

Sag’s 2016 study provides a detailed analysis of the geographic distribution of copyright litigation in US District courts, by state and by district. Sag emphasizes the changing hierarchy of which districts attract the most litigation and suggests that future empirical studies need to be cautious about basing any conclusions about copyright litigation on small selections of time or limited geographic coverage.

Doctrinal Studies — Event studies

Barnes (2000) and Liu (2012) each explore the impact of significant US Supreme Court decisions in what can be termed ‘event studies’. Barnes examines the effect of the Court’s decision in *Fogerty v. Fantasy, Inc.* 510 U.S. 517 (1994) which supposedly

leveled the playing field between plaintiffs and defendants in terms of attorney's fee awards in copyright cases. Lui explores the implications of *eBay Inc. v. MercExchange, L.L.C.*, 547 U.S. 388 (2006), a decision on the injunction standard in patent law, for injunctions in copyright cases.

The Court's *Fogerty* decision held that attorney's fee awards under Section 505 of the Copyright Act of 1976 should be available to the prevailing party, regardless of whether that party was the plaintiff or the defendant. Barnes examined 262 rulings on motions for attorney's fees in cases with at least one claim of copyright infringement decided in the four and half years before and after *Fogerty* (i.e., between 1989 and 1998). Prevailing plaintiffs' fee requests were granted at a rate of 89 percent in both the pre- and post-*Fogerty* era. In contrast, prevailing defendants recovered attorney's fees on only 16 percent of cases before *Fogerty*, but 61 percent of cases following the Supreme Court's ruling. Barnes concludes that *Fogerty* has had a significant effect, but that it has not entirely leveled the playing field between plaintiffs and defendants. Specifically he notes that whereas fees for the prevailing plaintiff is par for the course, a prevailing defendant still must show that a plaintiff's claims or conduct during litigation are blameworthy in order to shift the fees.

The Court in *eBay* rejected the Federal Circuit's 'general rule' that a permanent injunction will issue once infringement and validity have been adjudged in favor of the traditional equitable four-factor balancing framework that typically governs the award of injunctive relief in US federal courts. In effect, *eBay* gives district courts substantially more discretion as to whether to award an injunction in patent cases. Based on a population of just over 500 decisions relating to permanent and interim injunctions between 2006 and 2010, Liu concludes that the majority of post-*eBay* decisions on copyright injunctions have "totally ignored the *eBay* decision as well as the four-factor test advocated therein."

Not every Supreme Court decision is well suited to an event study. The best candidates give rise to an abrupt change in the law, in at least some jurisdictions; are not widely anticipated such that litigants and lower courts would be expected to change their behavior in advance of the decision; and allow for meaningful pre- and post-decision comparisons. *Fogerty* appears to meet at least two of these criteria. One potential distortion is that given a more favorable standard for awarding attorney's fees to defendants, more defendants should be expected to ask for attorney's fees. The data reported in the Barnes study would be more persuasive if it could be shown that the rate at which prevailing defendants asked for fees under the old standard was not significantly lower than under the new one.

Lui's study is more problematic. Obviously, *eBay* was not likely to have been cited before it was decided, so there is no baseline against which to judge the significance of the fact that it is only cited in 11.3 percent of copyright injunction cases post-*eBay*.

Lui does not compare the rate at which copyright owners either asked for or were granted injunctions pre-and post-*eBay*. The possibility that the plaintiff's decision to seek an injunction is endogenous to the decision standard raises questions that Lui's study does not answer.

Doctrinal Studies — Fair Use

The fair use doctrine is so intriguing that it has been subjected to at least three, and arguably as many as five, empirical studies. David Nimmer's self-consciously "unscientific" study of 60 arbitrarily selected and idiosyncratically coded fair use cases in 2003 should probably be credited as the inspiration for the three core empirical studies, Beebe (2008), Netanel (2011) and Sag (2012), and indirectly for Pamela Samuelson's (2009) comprehensive taxonomy of fair use decisions into policy relevant clusters. Samuelson, Netanel and Sag each explicitly build on Beebe's dataset of reported federal opinions that made substantial use of the section 107 four-factor test from the January 1, 1978, effective date of the Copyright Act through 2005.

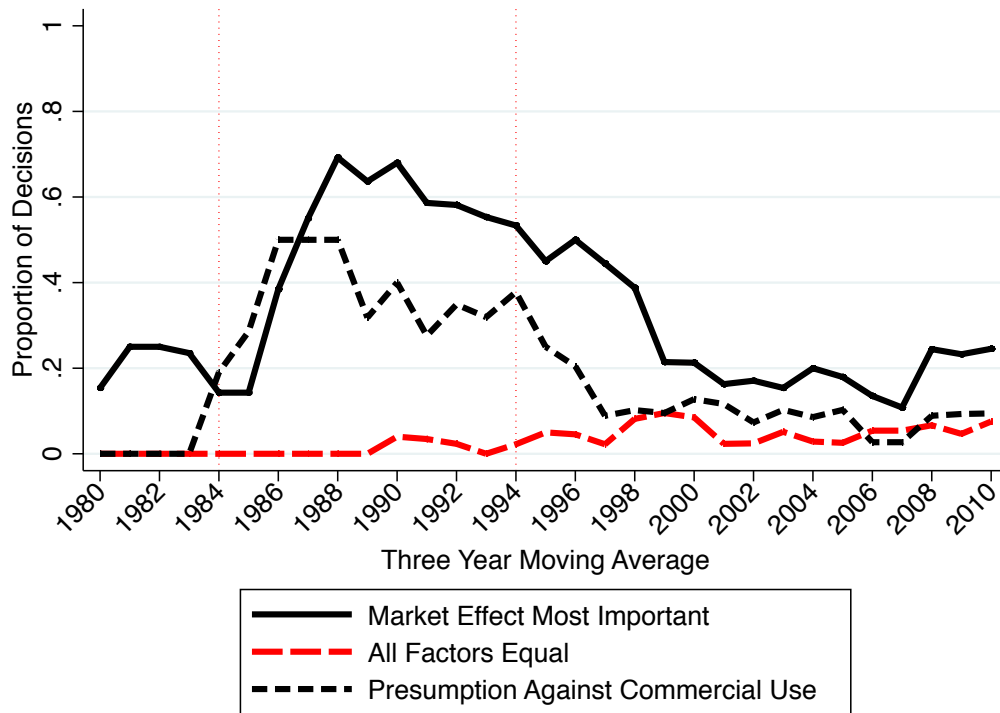
Beebe 2008

Beebe's groundbreaking study of the application of the fair use doctrine showed that much of the conventional wisdom about how judges decide fair use cases was simply wrong. For example, in spite of salient cases indicating the contrary, it turns out that district court fair use decisions are not especially liable to be reversed, nor is it obvious that the malleability of the fair use factors leads judges to twist every factor to align with the ultimate result.

In his dataset, Beebe coded variables for each of the four statutory fair use factors as well as many sub-factors for over 300 fair use decisions by district court and court of appeals judges. Whether it makes sense to throw district and appellate court decisions all in together depends on the nature of the question being asked. If the research question is directed to case outcomes in the context of a particular doctrine, the best practice is to focus on decision-making at the District Court level and then treat the fate of a decision on appeal as an additional fact about the District Court case. Another acceptable approach would be to treat district court cases and appellate court cases as two distinct data sets. Examining the average win rates of a mixed bag of district and appellate decisions makes no sense unless you believe that cases are appealed on a more or less random basis. However, where the research question is directed to charting the influence of particular memes, citations or maxims over time it might be acceptable to group district and appellate level decisions. Beebe's study is, in part, a good illustration of this latter approach.

Beebe traced the rise and fall of certain doctrinal memes across time and demonstrated the alarming extent to which subsequent courts continue to invoke notions from earlier cases that were all but renounced in the Supreme Court’s most recent teachings in *Campbell*. Beebe found that district courts continued to invoke the broad *Sony* presumption regarding commercial use, a presumption that *Campbell* repudiated. It seems that lower courts have been quick to embrace new concepts, such as transformative use, but slower to let go old aphorisms, such as the *Harper & Row* Court’s declaration the fourth factor was “undoubtedly the single most important element of fair use.”

Figure 3 Fair Use Memes



Three year moving average of selected memes in district court and court of appeals fair use decisions between 1978 and 2010. Data Source: Beebe (2008) as supplemented by Netanel (2011) (http://webshare.law.ucla.edu/Faculty/bibs/netanel/Netanel_FairUse_Master_030711.accdb). Additional calculations by the author.

Figure 3 (above) shows the decline in references to *Harper & Row*’s dictum of fourth factor supremacy and *Sony*’s presumption that commercial uses are not fair use and/or cause market harm. This figure may actually overstate the impact of *Harper*

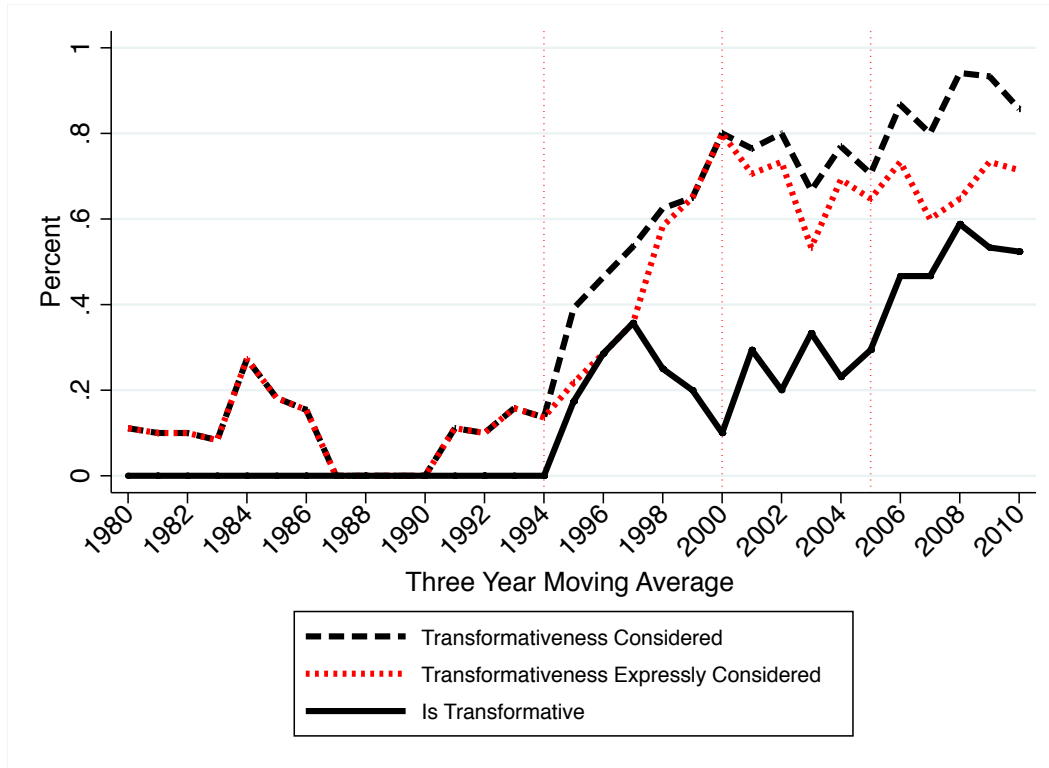
Row because, of the 26 opinions citing the decision in the period 2001-2010, 11 did so while expressly finding no market harm.

Netanel 2011

Neil Netanel (2011) extended Beebe's dataset with an additional five years of fair use cases and focused the rise to prominence of "transformative use". Based on his review of cases decided between 1978 and 2005, Beebe concluded that commentators had exaggerated the influence of the transformativeness of the use on fair use doctrine. Beebe noted that despite the Supreme Court's express adoption of transformative use as the touchstone of fair use, over 40% of reported district court opinions failed even to refer to transformative use, at least by name. Conversely, Netanel shows that since 2005 the transformative use paradigm has come overwhelmingly to dominate fair use doctrine. Netanel's result differs so dramatically from Beebe's for two reasons. First, as evident in Figure 4 (below) the trend is simply much clearer over the last ten years than it was previously. Based on Netanel's data, transformative use was considered by district courts in just over 47 percent of cases decided between 1991 and 2000, but by 80 percent of cases in the decade from 2001 to 2010. The same data also shows that the rate at which the defendant's use was found to be transformative was a mere 16 percent in the 1991-2000 period, but 43 percent in the 2001-2010 period.²

²These calculations only include cases where the fair use issue arose in a preliminary injunction, temporary restraining order, crossed motions for summary judgment by plaintiff and defendant or a bench trial. The figures are not substantively different if uncrossed motions for summary judgment by plaintiff and defendant are included.

Figure 4 Transformative Use in District Court Decisions (Moving Average)



Three year moving average of the application of ‘transformative use’ in district court fair use decisions between 1978 and 2010. Data Source: Beebe (2008) as supplemented by Netanel (2011) (http://webshare.law.ucla.edu/Faculty/bibs/netanel/Netanel_FairUse_Master_030711.accdb). Additional calculations by the author.

Second, Netanel looked beyond whether the court expressly adopted transformative use by name and also included those cases that could be fairly read to apply the concept of transformative or some functional equivalent thereof. This broader count of transformative use cases is indicated by the darker dashed line in Figure 4 (above). Specifically, Netanel includes the following: decisions that quote and apply the Supreme Court’s definition of transformative use in *Campbell*; decisions that ask whether the defendant has used the plaintiffs work for a different expressive purpose than that of the work’s creator; decisions that assess whether the defendant’s use is of a type, such as parody or criticism, that Judge Leval enumerated as a likely example of transformative use in his article and that courts have typically held to fall within that rubric as well; and decisions where the court finds that the use before them is analogous to a use defined as ‘transformative’ in a prior case, without repeating that designation in its own opinion.

Analysis of Netanel's data shows that although there is no significant correlation between merely considering transformativeness and fair use, there is a significant correlation between judicial findings of transformativeness (or its equivalents) and fair use outcomes favoring the defendant. This should be no surprise to anyone familiar with the fair use case law. That correlation is .57 in the five years beginning 1996, .65 in the period 2001-2010 and .66 in the period 2006-2010.

Sag 2012

Beebe's study and Netanel's follow-up study are excellent examples of the use of empirical tools to enable systematic content analysis. Systematic content analysis can provide an illuminating synthesis of explicit judicial reasoning; however, what judges say they do and what judges actually do may be two entirely different things. Correlation and regression analysis can help clarify the relationships between variables derived from systematic content analysis, but any inference of causation has to be treated with some skepticism.

Consider, for example the fourth fair use factor enumerated in Section 107 of the Copyright Act, "the effect of the use upon the potential market for or value of the copyrighted work." Beebe found that of the 141 judgments in his sample finding that this factor disfavored the defendant, all but one also ruled against the defendant on the ultimate issue of fair use. Thus, if we take written opinions at face value, then adverse market effect is an almost perfect predictor of a denial of fair use. But this correlation is so high that it suggests that the fourth factor is no factor at all. Rather, as Beebe suggests, "The fourth factor essentially constitutes a metafactor under which courts integrate their analyses of the other three factors and, in doing so, arrive at the outcome not simply of the fourth factor, but of the overall test". This highlights one of the most significant limitations of empirical legal studies in the realm of doctrine – it is hard to be certain whether a stated factor leads to a particular decision or is merely symptomatic of that decision.

To overcome the limitations of systematic content analysis, Sag (2012) extended Beebe's data with additional coding of externally identifiable characteristics that could be predicted to influence the outcomes of fair use cases. Sag reviewed the literature and case law relating to fair use to determine to what extent abstract doctrinal propositions could be reduced to specific hypotheses that were empirically testable using available data. Motivated by a desire to assess the predictability of the fair use doctrine from the perspective of potential litigants, Sag investigated the impact of externally identifiable characteristics on the outcomes of fair use cases. For example, without regard to what judges actually said about whether a use was commercial or non-commercial, Sag and his research assistants simply coded this variable according to a rigorous codebook definition. Whereas Beebe's systematic content analysis was a self-conscious effort to map what judges said about the cases

they decided, Sag's research question related to the predictability of the application of the fair use doctrine and thus had to focus on case facts which existed prior to any judicial determination. By focusing on characteristics of disputed uses that would have been apparent to litigants before their cases went to trial, Sag was able to study the ex ante predictability of the application of the fair use doctrine.

However, relying on externally verifiable data has some important limitations. For example, Sag was unable to code directly for whether a use was 'transformative' as that term is used in the fair use cases because transformativeness is usually a question of degree and is intensely disputed by the parties in any fair use case. Sag was however able to construct a proxy for transformativeness that captured at least part of the concept. Sag reasoned that

To the extent that transformative use means making a new work out of an old one, ... stark differences between the work allegedly copied and the defendant's work should be indicative of transformation. This general concept is operationalized by the variable *Creativity Shift*. *Creativity Shift* is set to 1 in cases where the plaintiff's work is creative and the defendant's is informational, or vice versa. In such cases, the defendant has not just created a new work, she has also created a work in a different category. This shift in category should almost always entail a fundamental change in purpose, which is the hallmark of transformative use."

Specifically, using multivariate regression analysis, Sag concluded that 'Creativity shift' makes a finding a fair use about twice as likely as otherwise. He also concluded that although commercial use in general had no discernable value as a predictor of fair use, 'Direct Commercial Use' – i.e., any use of the plaintiff's copyrighted work in a product or service sold to the public without the injection of additional labor or creativity by the defendant – made a finding of fair use less likely. Creativity Shift and 'Direct Commercial Use' can be interpreted, respectively, as proxies for the presence and absence of transformative use. These variables overlap with many conceptions of transformative use, but Sag does not claim that Creativity Shift is synonymous with transformative use or that 'Direct Commercial Use' is always an antonym of transformative use. Sag also concluded that fair use is more likely in cases where the defendant uses only part of the plaintiff's work. Finally, Sag used factors such as whether the parties were individuals versus corporations and the quality scores of the attorneys who represented them to try to test the theory that fair use favors the underdog. He found no evidence supporting that theory.

Whereas Beebe and Netanel trace the development of fair use doctrine and suggest correlations with case outcomes, Sag's study investigates the operation of fair use from the perspective of potential litigants. Sag demonstrates that, to the extent that

fair use jurisprudence and theory suggest actual testable hypotheses, the data confirms those hypotheses with respect to transformative use (factor one) and the amount and substantiality of the portion used (factor three). These findings, in conjunction suggest that – frequent assertions to the contrary notwithstanding – fair use is not an especially unpredictable legal doctrine. Sag’s empirical findings are consistent with Pamela Samuelson’s (2009) comprehensive taxonomy of the fair use cases in which she argues persuasively that, although not devoid of hard cases, fair use is not indeterminate. Rather, as Samuelson explains, the fair use case law falls into coherent and, essentially, predictable categories of uses.

Doctrinal Studies — Substantial Similarity

Lippman & Rogers (2013)

In 2013 Lippman and then Rogers published separate studies using the same dataset of published court of appeals decisions relating to substantial similarity. As far as their empirical contributions go, the Lippman and Rogers papers are cumulative. The Lippman-Rogers data comprises 234 circuit opinions issued between 1923 and 2011. The authors focused on the percentage of Court of Appeals opinions finding that the copyright owner had established substantial similarity and how that percentage changed over time and varied depending on subject matter, jurisdiction and the test articulated by the court.

The Lippman-Rogers study also demonstrates some of the central difficulties with empirical studies of litigation. Most acutely, by choosing to study only circuit level cases, Lippman and Rogers have virtually guaranteed that all their results are dictated by the selection effect. As Priest and Klein (1984) famously observed, win rates at trial are no reflection of whether a particular decision standard is plaintiff- or defendant-friendly because the decision to go trial is itself something of an oddity. The time-consuming and costly process of litigation does not generate a random sample of all potential disputes; rather, litigation acts as a filter, selecting only those cases where uncertainty about the law, asymmetric stakes, divergent expectations, or other quirks of human behavior have prevented the parties from settling their dispute.

In all likelihood, any knowledge gained from studying a doctrine such as substantial similarity purely at the Court of Appeal level has no application to district court cases. For example, Rogers notes that copyright plaintiffs only win about 33% of cases that are appealed on the issue of substantial similarity, but this does not say anything about the broader set of district court decisions that were not appealed. Also, Rogers concludes that the three main groupings of substantial similarity tests (Observer, Intrinsic/Extrinsic and Abstraction/Filtration) do not differ significantly

in terms of plaintiff win rate, but again, it is impossible to tell whether this is due to the irrelevance of the tests or because parties are only inclined to appeal relatively close cases. Because the tests are associated with particular circuits, it is very hard to claim any kind of causal relationship between the tests and case outcomes – the effects ascribed to tests could easily be due to some other circuit level characteristic. Finally, conclusions such as “the odds are overwhelmingly against plaintiffs in substantial similarity cases.” (Rogers 920) are very likely to be simply an artifact of only looking at decisions that were appealed.

The Lippman-Rogers study is a cautionary tale. In their own words, the authors “decided that the circuit court decisions would be reviewed over trial decisions, simply due to the massive number of trial court substantial similarity decisions.” (Rogers) This is a trap to be avoided. Certainly, some research questions lend themselves to narrow studies of appellate litigation, however if the research question relates to the effect of particular doctrines, tests, or subjects on case outcomes, there is no getting around the fact that the day to day application of copyright law takes place in district courts.³ Researchers who decide that there are too many cases to code should select a random sample of relevant cases rather than a complete but arbitrary sample.

Doctrinal Studies — Publication

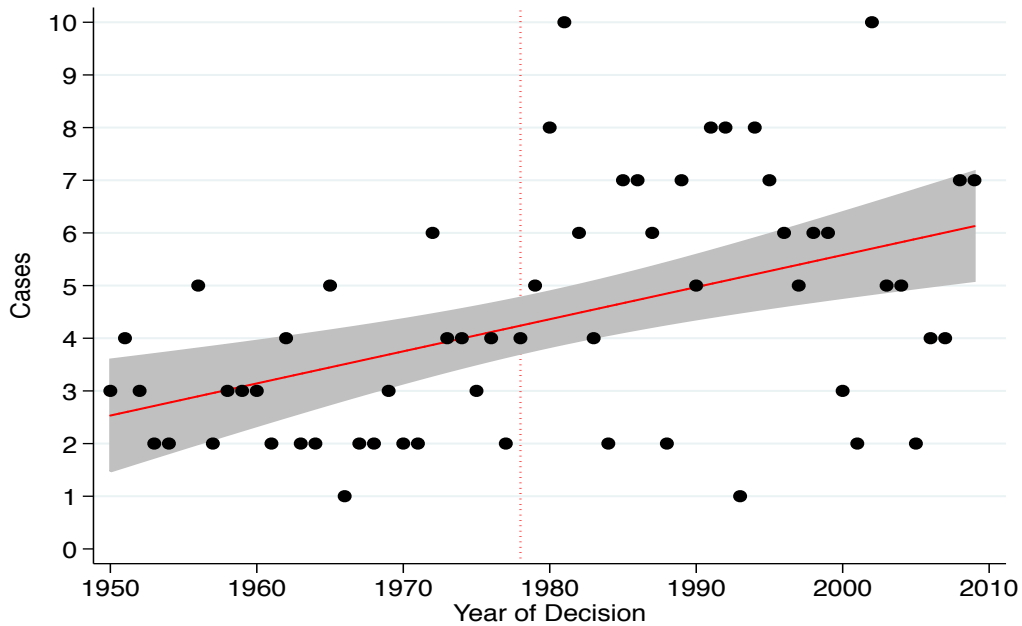
Deborah Gerhardt’s study of copyright publication decisions analyzes various factors that may be thought to influence a court’s decision as to whether a work is published or unpublished. Gerhardt carefully coded a large set of decisions spanning 1832 to 2009. Gerhardt’s study is significant because an erroneous understanding of the doctrine leads to the misclassification of works as copyrighted when they are in fact in the public domain and vice-versa.

Gerhardt selection of such a long time horizon is an interesting choice. Lippman and Rogers do something similar in relation to substantial similarity. Certainly, more data is usually better, and given that Gerhardt’s dataset is now publicly available this is a great resource to future scholars. However, a timeline this long is not necessary to compare the pre- and post-1978 worlds and the reader is left with the sense that comparing the substantive factors that determine publication outcomes in cases from the late 1800s to the early 21st Century is a bit like comparing the factors leading to battlefield mortality in Gettysburg (1863) and Fallujah (2004).

³ An additional reason that Lippman and Rogers may have focused on substantial similarity cases at the appellate level is that substantial similarity decisions at the district court level are very difficult to code. District courts do not articulate or apply the law in this area with anything approaching the clarity of fair use decisions.

Gerhardt shows that contrary to assertions to the contrary, since 1976, federal courts have seen an increasing number of copyright cases that involve publication issues. Even focusing just on those cases where publication was raised to determine whether the work was in the public domain, Gerhardt’s data shows an increasing trend, albeit with very low numbers.

Figure 5 Public Domain Publication Cases (1950–2009)

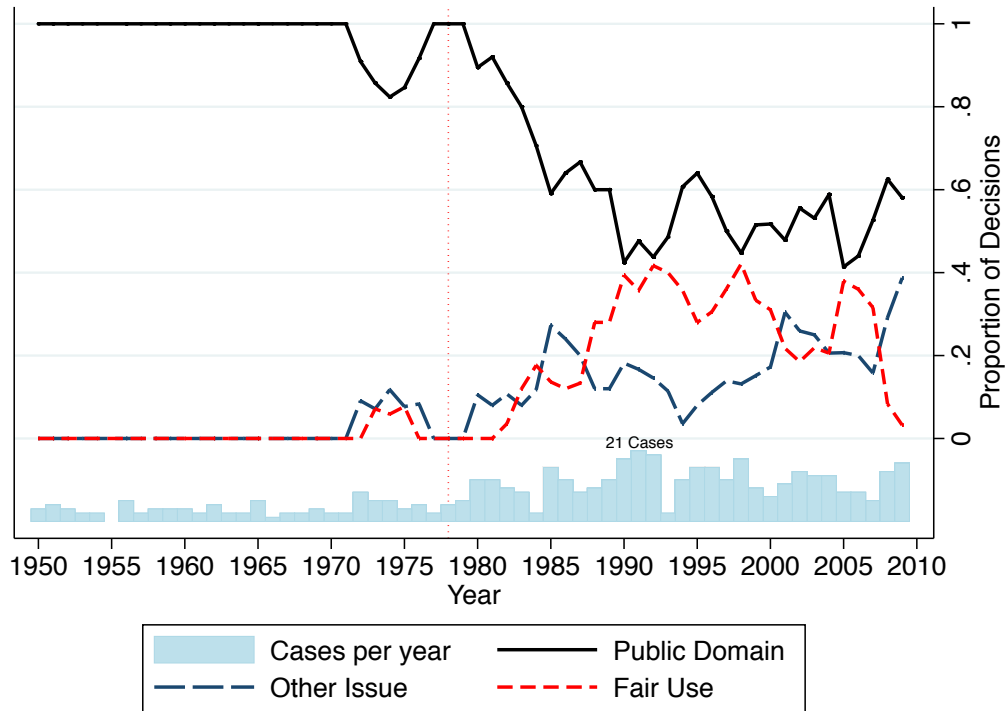


Yearly average of public domain publication cases with linear line of best fit and 95% confidence interval. The year 1978 is indicated by a dotted vertical line. Data Source: Gerhardt (2011) (<http://www.law.unc.edu/documents/faculty/gerhardtdataset.zip>). Additional calculations by the author.

One of the main trends Gerhardt’s study highlights is the changing relevance of publication since 1978. As illustrated in Figure 6 (below), between 1950 and 1978 the virtually every time the issue of publication was decided in a written opinion, it was to determine whether a work had been published prior to registration. In the United States until 1978, publication without adherence to copyright formalities placed the work in the public domain. Although the law changed with respect to new works as of January 1, 1978 the issue continues to be litigated for works created under the old regime. After 1978 the issue of publication was often raised in different contexts, especially in the application of the fair use doctrine. In the fair use context, the fact that a work was unpublished is said to weigh against a finding of fair use. Gerhardt concludes that courts still employ the limited/general publication distinction for

determining whether works are in the public domain, but do not use this distinction in fair use cases where the term publication is used as a proxy for whether the right of first publication was unfairly usurped.

Figure 6 The Relevance of the Publication Inquiry (1950–2009)



Three year moving average of the application of the reason for the publication inquiry in all federal cases between 1950 and 2009. The number of decisions in each year is indicted by the bars at the bottom of the graph. Data Source: Gerhardt (2011) (<http://www.law.unc.edu/documents/faculty/gerhardtdataset.zip>). Additional calculations by the author.

Gerhardt’s study is consistent with the view that the doctrine of publication operates differently depending on the consequences. In District Court cases since 1950, where the publication issue was raised with respect copyrightability courts found that work was published 68.7% of the time. In fair use publication cases, the comparable figure is 53.6%. There is a similar difference in Court of Appeal cases decided since 1950, where the issue is publication of the work was found to be published just over 74% of the time and where the issue was the use it was found to be published just over 63% of the time. These differences don't actually prove that the law is different in

the fair use context or that is applied differently. These differences could be solely attributable to the selection effect.

Gerhardt's study draws a number of interesting conclusions, many of which contradict the pronouncements by noted copyright experts. For example, the Nimmer Treatise concludes that:

publication occur[s] when, *by consent of the copyright owner*, the original or tangible copies of a work are sold, leased, loaned, given away, or otherwise made available to the general public, or when an authorized offer is made to dispose of the work in any such manner, even if a sale or other such disposition does not in fact occur.⁴

Gerhardt's data suggests the opposite; that a finding of publication is "less likely if the court reports that (1) the original was sold, (2) the work was publicly performed, (3) the work was rented, leased, or loaned, or (4) the work was deposited in a government archive." This finding is quite surprising given that Section 101 provides that these distributions amount to publication.

Gerhardt's findings do not mean that Nimmer is necessarily wrong, it could be the case that these findings are driven by a very strong selection effect, such that the cases that are litigated far enough to produce a written opinion are the exceptions to the general rule. In everyday terms, factors that take the work outside the control of the copyright owner are more likely to result in events that would be considered publication. However, when cases are litigated on the issue of publication and a court bothers to spell out such factors, they are, likely as not, being raised to distinguish the events that transpired from what the courts term "general publication". Gerhardt's data is not dispositive, but it does suggest a closer look is required.

CONCLUSION

Empirical studies of copyright litigation appear to be quite unlikely to displace traditional doctrinal studies. However, they have the potential to augment those studies by injecting some rigor into casual empirical observations and by identifying patterns of behavior and patterns of judicial decision-making that might otherwise go unobserved.

⁴ 1 Nimmer & Nimmer, § 403 [A], at 4-24 (emphasis added) (footnotes omitted).

It is almost certain that copyright scholars will continue to seek to answer critical questions of policy and theory with empirical tools. This is a positive development. Indeed, the Copyright Office itself has called for greater development of empirical research – although, given that the office currently employs exactly zero economists, how they plan to evaluate such research is unclear. However, empirical tools need to be applied with some sensitivity to their limitations. I conclude this chapter with three guidelines for researchers planning to undertake an empirical study of copyright litigation that represent current ‘best practices’.

1. Open data

The data referred to in empirical studies should be downloadable in commonly used formats by default. The only reason to make people ask for your data is to discourage them from looking at it, or to withhold permission. Of all the works surveyed in this chapter, Cotropia & Gibson (whose data is available on a dedicated website in an excel file) deserve recognition for setting the gold standard in this regard. Researchers who fail to make their data available upon publication should clearly explain their reasons. Data not involving human subjects that is not available for replication should be regarded as suspicious. Researchers may hesitate to open up their carefully collected data files because they fear that others might misuse, misinterpret or misrepresent their data. That is a risk worth taking, open data encourages accountability and transparency, it also benefits authors by allowing others in the academic community to assess the significance or insignificance of methodological choices.

2. Taking Selection Seriously

Copyright disputes are common, but most do not end in litigation. Disputes about authorship in television and motion pictures are usually resolved by the arbitration system of the Writers Guild of America, Internet filesharing and other contested online uses lead to DMCA takedown notices and other similar removal and blocking requests, but these rarely lead to litigation. In November of 2015, for example, 2,491 reporting organizations representing 5,795 copyright owners asked Google to remove or block over 70,000 internet domains and 65 million specific URLs from Google search;⁵ only a handful of these takedowns are likely to lead to the filing of a claim for copyright infringement in federal court. Given the problems raised by the selection effect, some may wonder whether there is any point to empirical studies of litigation at all.

⁵ Google Transparency Report, November 30, 2015 (available at <https://www.google.com/transparencyreport/removals/copyright/>) Significantly, the does not include requests directed at YouTube and other non-search properties. *Id.*

Every study of copyright litigation in the real world is conducted in the shadow of the selection effect. It is very difficult to make inferences about the broader universe of legal disputes by observing the subset of cases that are filed, the fraction of those that generate some form of written opinion or clear verdict, and the even smaller fraction of those that are appealed. The selection effect between each of these stages throws a wrench into almost every assertion causal inference.

Although no observational study of litigation is immune to the problem of selection bias, this fact should not be taken as a reason to abandon the field entirely, but nor should it be seen as a license to simply ignore the problem. We should not simply give up in empirical legal studies of copyright litigation, because even though litigated cases are not representative, the fact remains that they are important and they are constantly subject to ad hoc empirical assessments. Disputes that culminate in written decisions are the primary source of information for lawyers and judges attempting to discover the content of the law. Written opinions are particularly important because they provide analogies and reasoning that can be extended to future cases. Lawyers and academics constantly call upon their assessment of “what really happens” in fair use cases to inform their understanding of the law. It is easy to overlook the fact that these explanations are also prone to the very same selection bias.

Every empirical study of litigation should acknowledge the selection effect and address what it might mean in that particular context. In addition to this very general guidance, I offer two practical observations. First, researchers should understand that because the selection effect at the district and appellate levels are quite different, appeals cases should not be treated as equivalent to district court cases for most purposes. Second, those of us who rely on commercial databases need be sure that we understand those sources. For example, researchers should be aware that an increase in the apparent availability of written opinions on a particular subject over time could very well be a product of changing norms of opinion writing and changing criteria for database inclusion.

3. The Future is Out of Sample

Finally, researchers need to understand that the future is out of sample. In other words, an empirical analysis within a given timeframe might suggest patterns or trends within that timeframe that are significant, but there is no guarantee that such results have any predictive value. The maxim that correlation does not equal causation is often invoked merely as a plea for absolution. My own view is that correlation is science, and causation is argument. We should not shy away from these arguments, but we need to recognize them for what they are, subjective interpretations of the data. Researchers should feel free to argue that the patterns of yesterday will be the patterns of tomorrow, but they should recognize that causal

claims and predictions about the future are arguments based on empirical conclusions, they are not empirical conclusions themselves.

BIBLIOGRAPHY

Barnes, Jeffrey Edward. “Comment: Attorney’s Fee Awards in Federal Copyright Litigation after *Fogerty v. Fantasy*: Defendants Are Winning Fees More Often, but the New Standard Still Favors Prevailing Plaintiffs.” In 47 *UCLA L. REV.* 1381, 2000.

Beebe, Barton. “An Empirical Study of the Multifactor Tests for Trademark Infringement”, 94 *CALIF. L. REV.* 1581, 2006.

Beebe, Barton. “An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005”, 156 *U. PENN. L. Rev.* 549, 2008.

Cotropia, Christopher A. and James Gibson. “Copyright's Topography: An Empirical Study of Copyright Litigation.” 92 *TEXAS LAW REVIEW* 1981, 2014.

Favale, Marcella and Martin Kretschmer and Paul C. Torremans, “Is There A EU Copyright Jurisprudence? An empirical analysis of the workings of the European Court of Justice” (CREATE Working Paper Series, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2643699), Forthcoming *MOD. L. REV.*, 2016.

Ford, William K. “Judging Expertise in Copyright Law.” 14 *J. INTELL. PROP. L.* 1, 2006.

Gerhardt, Deborah R. “Copyright Publication: An Empirical Study.” 87 *NOTRE DAME L. REV.* 135, 2011.

Rogers, Eric. “Substantially Unfair: An Empirical Examination of Copyright Substantial Similarity Analysis among the Federal Circuits.” 2013 *MICH. ST. L. REV.* 893, 2013.

Landes, William M. “An Empirical Analysis of Intellectual Property Litigation: Some Preliminary Results.” 41 *HOUS. L. REV.* 749, 2004.

Lippman, Katherine. “The Beginning of the End: Preliminary Results of an Empirical Study of Copyright Substantial Similarity Opinions in the U.S. Circuit Courts”, 2013 *MICH. ST. L. REV.* 513, 2013.

MATTHEW SAG, EMPIRICAL STUDIES OF COPYRIGHT LITIGATION

- Liu, Jiarui. "Copyright Injunctions After Ebay: An Empirical Study." 16 *LEWIS & CLARK L. REV.* 215, 2012.
- Netanel, Neil Weinstock. "Making Sense of Fair Use." 15 *LEWIS & CLARK L. REV.* 715, 2011.
- Nimmer, David. "Fairest of Them All and Other Fairy Tales of Fair Use." 66 *LAW & CONTEMP. PROBS.* 263, 2003.
- Priest, George L. & Benjamin Klein. "The Selection of Disputes for Litigation." 13 *J. LEGAL STUD.* 1, 1984.
- Sag, Matthew. "Predicting Fair Use", 73 *OHIO ST. L.J.* 47, 2012.
- Sag, Matthew. "Empirical Studies of Copyright Litigation: Nature of Suit Coding 7." *Loyola Univ. Chi. Sch. of Law Pub. Law & Legal Theory*, Research Paper No. 2013-017, 2013), available at <http://ssrn.com/abstract=2330256>, 2013.
- Sag, Matthew. "Copyright Trolling, An Empirical Study." In 100 *IOWA L. REV.* 1105, 2015.
- Sag, Matthew. "IP Litigation in United States District Courts: 1994 to 2014" *Iowa Law Review*, Forthcoming. Available at SSRN: <http://ssrn.com/abstract=2570803>, 2016.
- Samuelson, Pamela. "Unbundling Fair Uses", 77 *FORDHAM L. REV.* 2537 (2009).