JOURNAL OF
EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 9, Issue 3, 482–510, September 2012

Predicting Securities Fraud Settlements and Amounts: A Hierarchical Bayesian Model of Federal Securities Class Action Lawsuits

*Blakeley B. McShane, Oliver P. Watson, Tom Baker, and Sean J. Griffith**

This article develops models that predict the incidence and amount of settlements for federal class action securities fraud litigation in the post-PLSRA period. We build hierarchical Bayesian models using data that come principally from Riskmetrics and identify several important predictors of settlement incidence (e.g., the number of different types of securities associated with a case, the company return during the class period) and settlement amount (e.g., market capitalization, measures of newsworthiness). Our models also allow us to estimate how the circuit court a case is filed in as well as the industry of the plaintiff firm associate with settlement outcomes. Finally, they allow us to accurately assess the variance of individual case outcomes revealing substantial amounts of heterogeneity in variance across cases.

I. Introduction

Understanding and predicting the outcomes of class action lawsuits is a topic of theoretical and practical importance. A particular area of focus is on securities fraud litigation as such suits comprise approximately 35–40 percent of all class actions (Eisenberg & Miller 2010; Fitzpatrick 2010) and approximately 75 percent of all settlement awards (Fitzpatrick 2010). Consequently, scholars have examined a number of factors associated with these settlements.

A topic of considerable investigation has been the effects of the U.S. Private Securities Litigation Reform Act of 1995 (PLSRA; Pub. L. No. 104-67, 109 Stat. 737 (1995)). This Act was designed to remedy the widely perceived belief that the "merits" of a case do not matter in class action securities fraud lawsuits (Alexander 1991) and the principal goal of the legislation was to make case outcomes depend more on the evidence of fraud and less on the avoidance of defense costs. To accomplish this, the PLSRA introduced a number of

*Address correspondence to Blakeley B. McShane, Assistant Professor, Kellogg School of Management, Northwestern University, 2001 Sheridan Rd., Evanston, IL 60208; email: b-mcshane@kellogg.northwestern.edu. Watson is Principal and Vice President at Juridigm, Inc.; Baker is William Maul Measey Professor of Law and Health Sciences, University of Pennsylvania Law School; Griffith is T.J. Maloney Chair in Business Law, Fordham University School of Law.

We are grateful to Adam Savett for providing the data examined in this article and for several helpful discussions.

changes to the law, including (1) presuming that the investor with the largest financial stake in the suit (in practice, often an institutional investor such as pension fund or mutual fund) would be appointed lead plaintiff and class representative, (2) heightening pleading requirements, (3) prohibiting discovery prior to a decision on the defendant's motion to dismiss, (4) requiring particularity and "strong inference" of fraud, and (5) replacing joint and several liability with proportionate liability in general (Cox et al. 2008).

Consistent with the theory underlying the enactment of the PSLRA's lead plaintiff provision, prior research has shown that the presence of an institutional investor increases the dollar amount of settlements (Cox et al. 2006, 2008). Other important post-PSLRA factors include provable losses (Cox et al. 2006) and the presence of an SEC enforcement action (Cox et al. 2008). However, the lead plaintiff provision has also led to controversial "pay-to-play" practices, with researchers finding evidence of "lawyers making campaign contributions to public pension funds' political leadership in order to gain favorable consideration by the funds for appointment as class counsel" (Johnson-Skinner 2009).

Returning to the merits of the case, Pritchard and Sale (2005), Johnson et al. (2006), and Baker and Griffith (2007, 2009) have quantitatively and qualitatively demonstrated the importance of merits variables in the post-PLSRA period. Pritchard and Sale (2005) have empirically demonstrated the importance of merits variables such as earnings restatements, Generally Accepted Accounting Principles (GAAP) violations, insider trading allegations, and transactions-related variables such as whether the case pertains to an initial public offering (IPO) or a merger. Johnson et al. (2006) have argued that not only do the merits matter, they actually matter more in the post-PLSRA period. Finally, Baker and Griffith (2007, 2009) focus on how the merits matter, showing, for example, that corporate governance and thus merit variables are important factors in Directors' & Officers' liability underwriting and settlement practices.

The above studies demonstrate that settlement amounts are tied to the evidence (i.e., merits and damages) of cases in the post-PLSRA period. Although this is one path to providing the correlation between case outcomes and evidence desired by lawmakers, there are at least two others: a tighter correlation would also obtain (1) if fewer "frivolous" lawsuits were filed or (2) if these frivolous lawsuits were dismissed. Empirically, the number of class action securities fraud lawsuits filed now exceeds pre-PLSRA levels (Buckberg et al. 2005), thus casting doubt on the former. Nonetheless, there have been a greater percentage of dismissals in class action securities fraud lawsuits in the post-PLSRA period, providing some evidence for the latter (Foster et al. 2000).

Another source of evidence on filings and motions to dismiss comes from Stephen J. Choi and his colleagues. Notably, Choi et al.'s (2009) results "do not show that Congress's efforts to discourage frivolous litigation have succeeded." They also demonstrate the presence of a "screening effect" whereby certain pre-PLSRA nonnuisance claims would be less likely to be filed post-PLSRA, an effect that is particularly strong for claims lacking strong evidence of fraud or insider trading. On the other hand, Choi (2003) suggests that "existing literature on filings and settlements in the post-PSLRA time period provide[s] evidence that frivolous suits existed prior to the PSLRA and that a shift occurred in the post-PSLRA period toward more meritorious claims," while Choi (2007) shows that nonnuisance IPO

cases lacking strong evidence are less likely to be filed, more likely to be dismissed, and more likely to be awarded lower settlement values in the post-PLSRA period.

Rather than examining the effects of the PLSRA, the present article confines itself to the post-PLSRA period and aims to turn a corner in empirical legal scholarship by creating a predictive model to forecast case outcomes based exclusively on information available at the time a lawsuit is filed. Using an extensive data set on nearly 1,200 class action securities fraud lawsuits (constructed primarily from Riskmetrics' securities litigation data and augmented with publicly available data on securities returns, industry groupings, and news sources), we create predictive models that estimate (1) the probability of settlement versus dismissal of a securities class action and (2) the amount for which the class action will settle conditional on settlement.

Our model improves on those reported by consulting firms (e.g., National Economic Research Associates (NERA 1999) and Cornerstone Research (Ryan & Simmons 2009)) in two major respects. First, by using exclusively variables known at the time of filing, our model can be used for early forecasting. Second, by separately estimating the probability of dismissal or settlement and the value of settlement conditional on settlement, our model allows for more accurate and automated identification of high versus low variance lawsuits.

Beyond these contributions, we also (1) use a data set that is more extensive both in terms of the number of cases and the number of covariates as compared to prior academic studies, (2) employ a hierarchical Bayesian model that estimates the cross-effects of the federal circuit in which the lawsuit is filed and the industry group of the defendant, (3) use the Google News Archive to create a measure of the notoriety of the defendant (Baker & Griffith 2009) that, as expected, has a positive relationship with settlement amounts, and (4) identify previously unreported weaknesses in the highly regarded Riskmetrics data widely used in industry and academia (Johnson-Skinner 2009; Fitzpatrick 2010) that, unless properly addressed, could potentially bias estimates based on that data.

Among the relationships identified by our model, the most interesting are the following. All the merits variables coded by Riskmetrics have the expected positive sign in both the settlement/dismissal model and settlement amount model, with two noteworthy exceptions. First, there is a change in the direction of the sign of the Rule 10b-5 variable. Cases coded as Rule 10b-5 class actions are less likely to settle but, if they do settle, the amounts are somewhat higher; this is consistent with the more generous damages available in Rule 10b-5 cases and suggests that the Rule 10b-5 variable might be better considered a damages variable than a merits variable. Second, there is the opposite change in the direction of the sign of the GAAP violations variable. Cases coded as alleging GAAP violations are more likely to settle but, if they do settle, the amounts are no larger. This latter result suggests that because a GAAP violation increases the likelihood of surviving a motion to dismiss, plaintiffs' lawyers are willing to invest in such lawsuits even if the potential damages award is relatively low.

All the damages variables (some collected by Riskmetrics, others by us) have the expected positive sign in the settlement amount portion of the model, while many of those same variables have a negative sign in the settlement/dismissal portion of the model, thus indicating that lawsuits selected for filing may be biased in the direction of larger potential damage awards. Among the damages variables exhibiting this latter shift in sign are the

market capitalization of the defendant, the company return during the class period,[1] the number of Google News Archive hits, and the presence of an institutional investor as a lead plaintiff (which, because of the lead plaintiff rules, we regard as a proxy for damages predictors not otherwise captured in our model). All these results are consistent with rational behavior on the part of those selecting lawsuits for filing.

The remainder of our article is organized as follows. In Section II, we discuss in depth the construction and merging of our various data sources. With a final data set in hand, we present summary statistics in Section III and our hierarchical Bayesian model in Section IV. In Section V, we present our coefficient estimates, predictions, and model evaluation. Finally, in Section VI, we provide a brief summary and discussion of our results.

II. Data Construction

A. Introduction

Constructing our database of securities fraud class action lawsuits proved to be a non-trivial endeavor requiring merging data from several different sources. In this section, we discuss those sources, the variables derived from them, and several interesting difficulties encountered.

Our principal source of data is the Riskmetrics Group's Securities Class Action Services Division, which tracks securities fraud class action lawsuits on a commercial basis. Data from this group have been previously validated in other research (Johnson-Skinner 2009; Fitzpatrick 2010) and are considered to be among the most high-quality of the several extant proprietary sources of class action securities litigation data. The Riskmetrics dataset has two components: a case settlement database giving the case details for each securities fraud class action lawsuit and a case securities database giving financial details for the securities involved in each case. Since each case is identified by a unique Case ID variable, it is theoretically possible to easily merge the two data sets; in practice, there are several difficulties involved in doing so, which we discuss in greater depth below.

In addition to the Riskmetrics data, we use four additional sources. We use data from Yahoo! Finance (2011) and the Center for Research in Security Prices (CRSP; University of Chicago Booth School of Business (2011)) to externally validate the securities return data contained in the Riskmetrics case securities database (see Section II.C for details). We also use the Yahoo! data to obtain the return of the S&P 500 over the class period of each defendant company—an important variable for our model (see Section II.D). In addition, we use Kenneth French's Data Library (French 2011) in order (1) to associate each defendant company with an industry group and (2) to obtain industry group returns over the class period (see Section II.D). Finally, as the literature suggests that case notoriety may be important for predicting securities fraud settlements (Baker and Griffith 2009), we use

---

[1]The class period is defined as the period between the alleged fraud or misstatement (the class period start date) and the corrective disclosure (the class period end date), with the class consisting of those who transacted in securities between those dates. Thus, it is the time period during which any monetary loss incurred by the plaintiffs as a result of the alleged illegal activity of the defendants took place.

the Google News Archive (2011) search functionality to construct numerical proxies for how well-known a company was prior to the case filing (see Section II.E).

We discuss each of these sources in greater depth below.

B. Case Data

The Riskmetrics case settlement database is among the most comprehensive datasets of securities fraud class actions cases gathered to date, containing information on 6,084 such settlements from 5,898 unique cases (we note that 136 cases have two or more partial settlements associated with them). In addition to a unique Case ID that identifies each case, the variables available to us came in two basic flavors: legal features of the case as well as potential merits variables. As these variables pertain to the cases themselves, they are of course identical across all partial settlements associated with a given case.

The legal features available in the data set are (1) the status of the case (i.e., settled, dismissed, or other), (2) the court (e.g., federal circuit or district), and (3) the name of the lead plaintiff and whether it was an individual or an institution.

Our potential merits variables are seven binary variables indicating whether or not (1) the case was an IPO case, (2) GAAP violations were alleged, (3) the allegation mentions that the company's financial statements were restated, (4) the case was a Rule 10b-5 case, (5) the case was an Securities Act Section 11 case, (6) insider trading was alleged, and (7) the case was transactional (i.e., involving a deal or merger).

Finally, we also have a variable that gives the total settlement if the case settles and that is zero otherwise.

C. Company Financial Data

The second database provided by Riskmetrics is the case securities database. Whereas the case settlement database contains one row for each settlement (i.e., 6,084 in total), the case securities database contains multiple rows per settlement (i.e., 32,068 in total). The "many-to-one" matching of securities to cases is not completely straightforward and thus we outline our matching procedure below.

The reason for the much greater number of securities than cases is the fact that many cases have a very large number of derivative securities associated with them in addition to simple common stock; these securities of various other classes (or types) may render the holder eligible for compensation if the case settles. The problem that faces us in preparing the data for modeling is to (1) identify the principal securities that pertain to the case and (2) combine these securities into covariates for our statistical model.

The first thing we did was to calculate (1) whether a given case had any securities associated with it in the case securities database and, if so, (2) how many classes of securities were associated with it. Of our 6,084 settlements records from 5,898 unique cases, 5,842 of the cases have at least one security and 4,530 have one and only one associated security. The case with the greatest number of securities of various types attached was the case filed on 09/13/2002 against ABN AMRO Holding N.V. (Callable CDs), which has 4,803 securities types associated with it.

To identify the principal securities, we restricted ourselves to securities with data on the five following variables: (1) the price at the beginning of the class period, (2) the price

at the end of the class period, (3) the number of shares outstanding at the beginning of the class period, (4) the number of shares outstanding at the end of the class period, and (5) the SIC code of the company. This restriction effectively guarantees that the securities under consideration are equity securities. More importantly, it also guarantees that we have enough information to calculate each security's return over the class period.

In restricting ourselves to this subset, we are left with 2,798 cases available for modeling. Of these, only 211 cases have more than one security type with defined values for all the fields above. The case with the largest number of such securities types is the IPO Securities Litigation Master case filed on 08/09/2001, which is associated with 88 security types.

For each of the cases, we sought a measure of the economic return of holding the associated equity securities over the class period. Ideally, we would take the buy-and-hold return of the common stock over the class period (or, in cases where there are multiple common stock securities associated with a case, we would take the average return of these securities over the class period where, in computing the average, each security would be weighted by its market capitalization). However, there was a significant difficulty associated with this endeavor: the price and shares outstanding data provided by Riskmetrics appeared to have some inaccuracies.

Consequently, we attempted to validate the Riskmetrics price and shares data against external sources, namely, Yahoo! Finance and CRSP. Unfortunately, Riskmetrics provided only a ticker for each security (i.e., as opposed to a unique security identifier). Since tickers change when companies are de-listed and since they are often later reassigned to new stocks, we could only unambiguously obtain matched external data for approximately 70 percent of the securities; not surprisingly, smaller companies were disproportionately unmatched.

For this subset of securities, the Riskmetrics data were generally equivalent to those provided by our external sources. We therefore viewed this as establishing the accuracy of the Riskmetrics data for all companies, apart from some smaller ones. It is plausible that data errors are more likely for these small companies, and we therefore used a statistical screen to remove data points that seemed erroneous (i.e., companies with class period returns that seemed implausibly large; see Section II.F for details).

Before proceeding, we note that we cannot compute the full economic return using Riskmetrics data; rather, we can only compute the change in market capitalization. The full return and change in market capitalization are, however, very similar with differences between the two normally coming from cash dividends. Thus, the change in market capitalization should adequately serve as a proxy for the economic return. We therefore define the company return during the class period as the percentage change in market capitalization of all securities associated with a given case.

D. Market and Industry Group Benchmark Return Data

For each security satisfying the conditions in Section II.C, we created two benchmarks against which we could compare the company return. As the company return was computed over the class period, we also compute our benchmarks over the class period. The first benchmark is the market return over the class period. In particular, we take the percentage

Table 1: Industry Group Descriptions

Industry	*Description*
1. BusEq	Business Equipment: Computers, Software, and Electronic Equipment
2. Chems	Chemicals and Allied Products
3. Durbl	Consumer Durables: Cars, TVs, Furniture, Household Appliances
4. Enrgy	Oil, Gas, and Coal Extraction and Products
5. Hlth	Healthcare, Medical Equipment, and Drugs
6. Manuf	Manufacturing: Machinery, Trucks, Planes, Office Furniture, Paper
7. Money	Finance
8. NoDur	Consumer Nondurables: Food, Tobacco, Textiles, Apparel, Leather, Toys
9. Other	Other: Mines, Construction, Transportation, Hotels, Entertainment
10. Shops	Wholesale, Retail, and Some Services (Laundries, Repair Shops)
11. Telcm	Telephone and Television Transmission
12. Utils	Utilities

change in the S&P 500 index (as provided by Yahoo! Finance) over the class period. This allows us to assess how the company's return over the class period compared to the market as a whole during the same period.

A potentially more relevant benchmark is to compare the return of the firm in question over the class period to the returns of similar firms rather than the market as a whole. For this purpose, we turned to Kenneth French's Data Library (French 2011), which provides (1) a daily index for the 12 different industry groups presented in Table 1 as well as (2) mappings from SIC codes to the 12 industry groups. Using this, we can compute the industry group return over the class period. When multiple securities were associated with a single case and these securities were associated with different industry groups, we set our industry group return variable to the market capitalization-weighted average of the distinct industry group returns over the class period.

E. Google Data

As company and case notoriety (e.g., newsworthiness) could be associated with settlement incidence and amount, we constructed a numerical proxy for notoriety via the Google News Archive. In particular, we counted the number of news stories returned when the company name was entered as the search term and the results were restricted to the one year prior to the filing date of the case.

As a robustness check, we also looked at the number of news matches in the one year subsequent to the filing date when the company name was entered along with "class action" as the search term. This variable was highly correlated with the number of hits for the original search, thus confirming our initial variable. Due to the high correlation, we omitted the second variable from our model because it was unlikely to add explanatory power. Furthermore, it had substantially larger number of zero hits. Finally, and most importantly, since our goal is to produce a model that is predictive of settlement incidence and outcome at the time of the company filing (i.e., a model that predicts the likelihood of dismissal and expected settlement amount using only variables whose values can be

known on the day of the filing), the number of news stories prefiling can be used for this purpose while any postfiling variable cannot.

F. Other Data Concerns

We have already discussed the major difficulties involved in merging our various data sources; here, we briefly discuss some additional minor difficulties that arose. We do so to indicate the issues that arise when dealing with legal data sets, even those of high quality such as that provided by Riskmetrics.

We have already noted that multiple securities were associated with 211 of our Case IDs. We note that the vast majority of these cases were the result of a merger (e.g., when a company with a case filed against it was bought out by another company and that company took on the liabilities of the initial company). As noted, for these cases, we aggregated the market capitalization of the underlying companies and used a market capitalization-weighted average for the company and industry group returns over the class period. While this is a reasonable approach, it is not clear that it is optimal and it may lead to overestimation of the true market capitalization. However, since our models use the natural logarithm of the market capitalization, this potential overestimation would likely not substantially alter our conclusions—particularly since so few cases were affected by this problem.

As noted, after screening cases that lacked adequate data on the associated securities (i.e., prices, shares outstanding, and SIC code; see Section II.C for details), we were left with 2,798 cases. From this set, we first removed all cases that did not settle in full in one settlement (i.e., we removed those cases with multiple partial settlements) thus mitigating any potential difficulties and subjective judgments involved in aggregating across the multiple partial settlements. There were only 84 such cases and Enron was the most prominent, with eight partial settlements listed in the database. Removing these cases left us with 2,714 unique case records and their corresponding 2,714 unique settlements.

Next, we wanted to confine our study to those cases governed by the PLSRA. Consequently, we removed the 412 cases that were filed before January 1, 1996 as well as eight additional cases for which no filing date was available in the Riskmetrics database.

Next, we addressed survivorship bias. Ultimately, we wish to build a model that predicts both (1) the likelihood of the case surviving the motion to dismiss and (2) the expected value of the final settlement. Because of the screening effect of the motion to dismiss and the fact that exceedingly few cases are resolved by summary judgment or trial, the duration of cases that settle is generally longer than those that do not. According to qualitative research, this is because cases that survive the motion to dismiss generally settle (Baker & Griffith 2009). Consequently, recent cases in the database suffer from severe selection bias. We thus eliminated the 677 cases that were filed on or after January 1, 2005. We chose this date because it was five years before our data set end date of January 1, 2010 and because ≈ 90 percent of cases are resolved by the five-year mark (Plancich & Starykh 2009).

Further, we noticed that the Riskmetrics case securities data implied that several cases had unusually large increases in market capitalization during the class period (e.g., one case

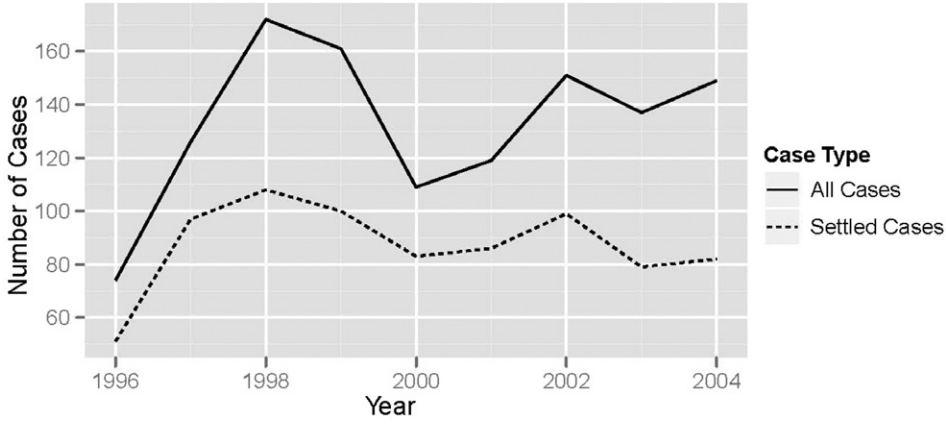
had a 13,898 percent increase), indicating incorrect financial data as discussed in Section II.C. To correct for these outliers, we assumed a maximum average daily volatility of 3 percent. If the underlying return associated with the case was positive and more than three standard deviations greater than the equivalent 3 percent daily volatility over the total length of the class period (i.e., if the return was greater than $0.09\sqrt{T}$ where T is the length of the class period in days), we inferred that the Riskmetrics price and share data were incorrect and thus excluded the case. This led to the removal of 53 cases.

Finally, we made several other minor reductions to our data set. First, we only considered the cases that had a case status of "Dismissed" or "Settled," leading to the removal of 36 cases; although manual inspection revealed these cases to be still active, it is generally safe to assume this given the extreme rarity of trials in class actions. Second, we removed all 159 cases in the database that were associated with the IPO Securities Litigation Master case based on the judgment that, because this large set of cases was resolved as a unit, the individual cases would not be sufficiently representative or predictive of other, more typical cases. Third, we removed the 10 SEC cases because they were not securities class actions. Fourth, we removed the 150 cases that were state court rather than federal court cases. Fifth, we removed the seven cases that were listed as settled but that had zero listed as the total settlement amount, and a further seven cases that were listed as settled but that had NA listed as the total settlement amount. Finally, we removed the five cases that had a filing date that was prior to the class start date.

Having begun with 5,898 cases, we are left with 1,198 cases for statistical modeling after having eliminated 3,100 cases for missing financial data (see Section II.C) and a further 1,600 cases for various other reasons (see the above paragraphs of this section). Before proceeding, we comment on the potential biases that could be induced by the screens we imposed on the data. For the latter set of screens (i.e., those described in this section that led to the removal of 1,600 cases), we note that the largest reductions came from the timing screens; these screens were necessary to ensure all cases were governed by the PLSRA and to avoid survivorship bias in the results. None of the other categories of excluded cases appear to be biased with regard to settlement versus dismissal or the amount of settlement.

Of the 3,100 cases excluded because of missing financial information (see Section II.C), only 562 cases would otherwise have been included in the final data set (i.e., 2,538 of the cases would have been eliminated because they were outside the 1996–2004 period, were state court cases, were associated with the IPO Securities Litigation Master case, or were subject to the other screens identified in this section). We thoroughly reviewed these 562 cases to determine whether there were any systematic differences between them and our set of 1,198 cases, finding only two. First, 44.1 percent of the 562 cases settle as compared to 65.5 percent of the 1,198 cases in the final data set. Second, the 562 cases are much more likely to have an empty plaintiff variable. Both these differences are consistent with Riskmetrics' greater incentive to fully code cases that settled than those that did not. On the other hand, the settlement amounts of the 562 cases and the 1,198 cases are very similar when examined both marginally and by date; furthermore, the ratio of the number of cases by year from the two sets of cases is roughly constant whether one examines all cases or just the settled cases.

Figure 1: Number of cases by year.



NOTE: There are 1,198 cases in total; 785 of these cases settled.

In sum, our screening process hinted at no major potential biases with the exception of the possibility that cases with missing data reflect Riskmetrics' incentive to more fully code the more valuable cases.

III. DATA

Using the process described above, we prepared our final two databases: (1) the database of all 1,198 cases that we use to model the probability that a case settles or is dismissed and (2) the database of the 785 settled cases that we use to predict the total settlement amount conditional on a case being settled. The latter database is thus a proper subset of the former. In this section, we define each of our variables and present summary statistics.

First, in Figure 1, we present the filing years of our cases, both for the full set of 1,198 cases and for the subset of 785 cases that settled. As can be seen, the number of cases in the full set and reduced set is roughly uniformly distributed over the nine years, with the exception of there being somewhat fewer cases in 1996.

Next, we examine our two response variables. The settlement variable is a binary variable that indicates that 785 of the 1,198 cases settled. Our settlement amount variable is the natural logarithm of the total settlement amount that we plot in Figure 2. The mean settlement amount is 15.7 with a standard deviation of 1.5. As can be seen, there appears to be no substantial trend to either the mean or the variance of the settlement values over time and the histogram shows a roughly bell-shaped distribution.

Next, we present the federal circuits and industry groups for each of our cases in Figure 3. As can be seen, the Second and Ninth Circuits account for the lion's share of our cases with the Third, Fifth, and Eleventh also accounting for a substantial fraction. As for the industry groups (see Table 1 for full descriptions), it appears business equipment is

Figure 2: Log settlement amount by filing date and histogram of log settlement amount for the 785 settled cases.

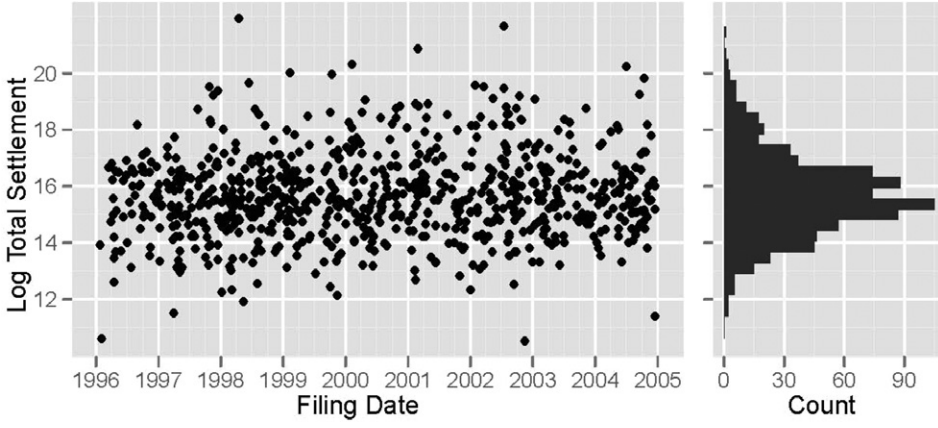
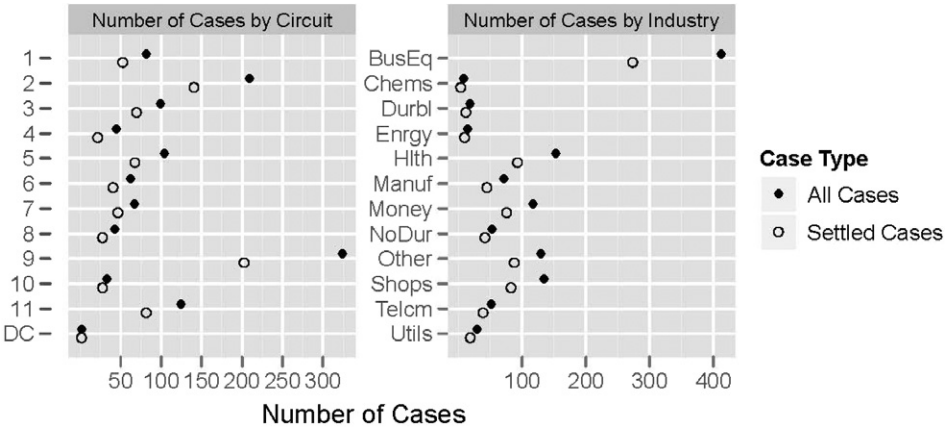


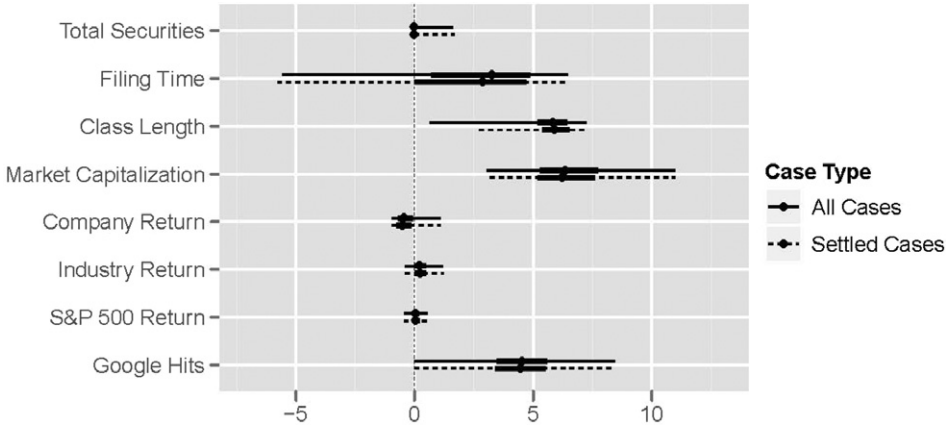
Figure 3: Number of cases by federal circuit and by industry group.



dominant in terms of number of cases. Healthcare, money, shops, and other also account for comparably large numbers of cases.

We now discuss our 18 covariates, of which eight are continuous, nine binary, and one categorical. In Figure 4, we present the distributions of our eight continuous covariates: (1) the total number of securities types associated with each Riskmetrics Case ID and contained in the Riskmetrics case securities database (Total Securities), (2) the length of time from the class end date to the filing date (Filing Time), (3) the length of the class period as obtained from the start and end dates in the Riskmetrics case settlement database (Class Length), (4) the market capitalization at the start of the class period as computed

Figure 4: Distributions of continuous covariates.



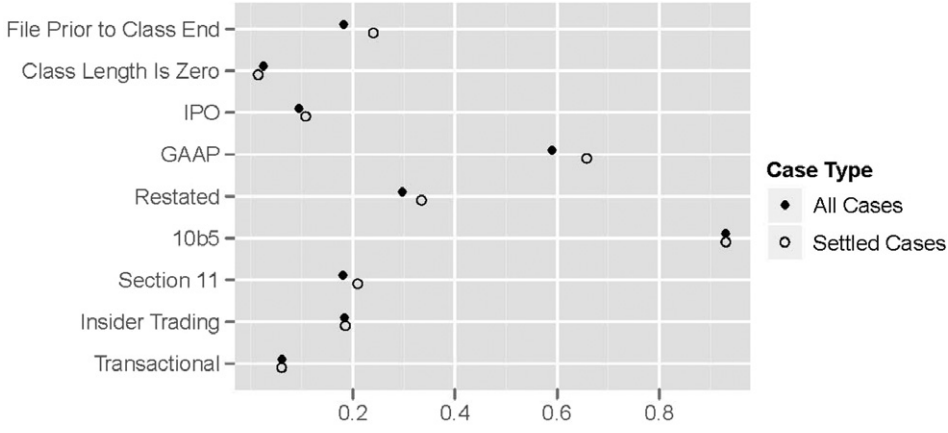
NOTE: The median of each covariate is given by the dot and 50 percent and 95 percent intervals (based on the 25 and 75 percentiles of the data and the 2.5 and 97.5 percentiles of the data, respectively) are given by the thick and thin lines, respectively. The units for each variable are given as (1) natural logarithm of the number of securities, (2) signed natural logarithm of the filing time in days, (3) natural logarithm of the class length in days, (4) natural logarithm of the market capitalization in millions of dollars, (5) company return as a fraction, (6) industry group return as a fraction, (7) S&P 500 return as a fraction, and (8) natural logarithm of the number of Google hits.

from the Riskmetrics case securities database as outlined in Section II.C (Market Capitalization), (5) the company return as computed from the Riskmetrics case securities database as outlined in Section II.C (Company Return), (6) the industry group return as obtained from the French Data Library as outlined in Section II.D (Industry Return), (7) the S&P 500 return as obtained from Yahoo! data as outlined in Section II.D (S&P 500 Return), and (8) the number of news stories associated with the company name based on a Google News Archive search in the 365 days prior to the filing date as outlined in Section II.E (Google Hits).

There are several noteworthy features visible in Figure 4. First, a majority (≈ 81 percent) of our cases have only a single security type associated with them (i.e., zero on the log scale). Nonetheless, some cases have up to 53 different types associated. Second, filing times are typically under six months from the end of the class period, while class periods typically last about six months to two years. Third, when only the 785 settled cases are examined, the distributions of the covariates do not differ greatly. Finally, since the distributions of several of our covariates possessed rather long tails, we used the natural logarithm of these covariates in modeling (see the note to Figure 4 for details).

We present our binary covariates in Figure 5. Our nine variables are whether or not (1) the case was filed prior to the class end date (File Prior to Class End), (2) the class start date is equal to the class end date (Class Length is Zero), (3) the case was an IPO case (IPO), (4) Generally Accepted Accounting Principles (GAAP) violations were alleged (GAAP), (5) the allegation mentions that the company's financial statements were restated (Restated), (6) the case was a Rule 10b-5 case (10b5), (7) the case was a Securities Act

Figure 5: Mean of binary covariates.

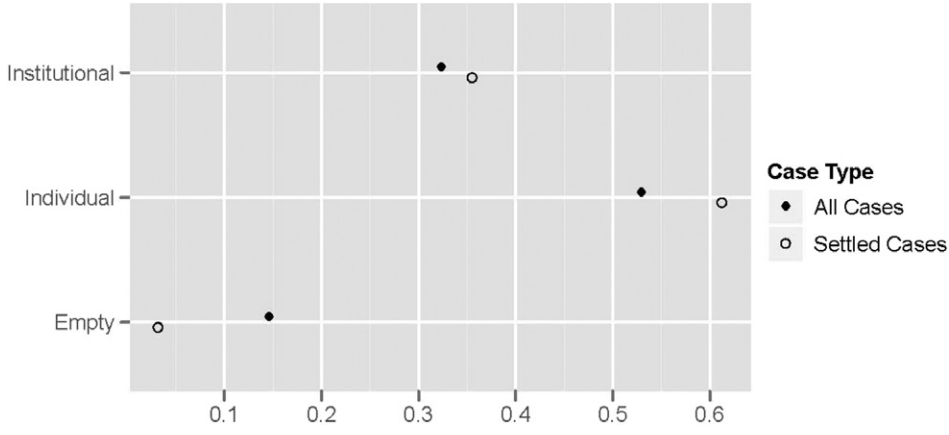


Section 11 case (Section 11), (8) insider trading was alleged (Insider Trading), and (9) the case was transactional (i.e., involving a deal or merger) (Transactional). All were obtained from the Riskmetrics case settlement database and are potentially related to the merits of the case. Thus, they may have substantial bearing on whether or not a case settles and for how much it settles.

We note that two of these binary variables were constructed by us from the Riskmetrics data. First, we included a variable for whether the case was filed prior to the class end date. In such cases, plaintiffs typically contend that the misrepresentations were not fully revealed until after the lawsuit was filed and this might have some bearing on the incidence or amount of settlement. We also included a covariate for whether or not the class period was of length zero (i.e., class period start date equals class period end date) because this likely indicates an IPO, secondary offering, or merger case where the alleged event took place on a single day (and that allegedly affected, e.g., those who received an allocation on the date of an IPO or who held common stock on the date of a merger).

Finally, in Figure 6, we present our single categorical covariate, which is based on the Riskmetrics case settlement variable describing the plaintiff. If one or more institutions are listed as the plaintiff, we set our plaintiff variable equal to "Institutional." If no institutions are listed but one or more individuals are, we set it equal to "Individual." Finally, if nothing is listed in the database, we set it equal to "Empty." Of course, this does not mean there is no plaintiff in the case; rather, it means Riskmetrics has not obtained the information for this variable. This is potentially informative for whether or not a case settles and for how much it settles if it does settle. In particular, given the nature of Riskmetrics' business, it is has the highest incentive to collect complete data for cases that settle and especially those that settle for large amounts; consequently, we would a priori expect Empty plaintiff cases (1) to be less likely to settle and (2) to settle for less when they do settle.

Figure 6: Fraction of cases in each level of the categorical plaintiff covariate.



IV. Model

A. Introduction

We model securities fraud class action lawsuits using a Bayesian hierarchical model. The Bayesian paradigm provides a principled method for (1) sharing information across cases that belong to the same "group" (e.g., industry group or federal circuit) and (2) for balancing this group-level information with the individual-level information provided by our 18 covariates (e.g., market capitalization, company return).

Throughout, we let Z_j denote whether the case was settled ($Z_j = 1$) or dismissed ($Z_j = 0$), where j indexes all 1,198 cases, and we let Y_j denote the settlement amount for case j where j indexes the 785 cases that settled. We will also let X_j denote the (column) vector of covariates for case j; continuous covariates are scaled by two standard deviations (Gelman & Hill 2006). We first discuss the model for whether a case is settled or dismissed and then the model for settlement amounts conditional on a settlement.

Before proceeding, however, we note that all models are estimated using Markov chain Monte Carlo (MCMC) methods (Metropolis et al. 1953; Hastings 1970; Geman & Geman 1984). We ran 10 chains of 10,000 iterations for each model, discarding the first 5,000 as burn-in and thinning every 10 iterations; criteria such as R demonstrated that convergence of the iterative simulation was achieved (Gelman et al. 2003).

B. Settlement/Dismissal Model

As settlement incidence is a binary variable (i.e., cases are settled or dismissed), we employ a Bernoulli likelihood

$$Z_j \sim Bern(p_j)$$

$$\text{logit}(p_j) = \alpha_j + X_j^{\intercal}\boldsymbol{\beta}$$

where α_j is a case-specific intercept term and $\boldsymbol{\beta}$ is the vector of slopes for our covariates.

With only one observation on each case, clearly the terms α_j are not identified without further assumptions. We consequently model them as

$$\alpha_j = \bar{\alpha} + \alpha_{c[j]} + \alpha_{i[j]} + \alpha_{c[j],i[j]}$$

where $c[j]$ is the circuit in which case j was filed and $i[j]$ is the industry group of the plaintiff firm for case j. That is, we decompose the case-specific intercept into various additive terms: (1) the overall "average" intercept given by $\bar{\alpha}$, (2) the overall effect of circuit $c[j]$ given by $\alpha_{c[j]}$, (3) the overall effect of industry $i[j]$ given by $\alpha_{i[j]}$, and (4) the interaction of circuit $c[j]$ and industry $i[j]$ given by $\alpha_{c[j],i[j]}$.

In studying securities fraud class action lawsuits, we will be principally interested in how (1) circuits and industries associate with settlement versus dismissal (i.e., the $\alpha_{c[j]}$, $\alpha_{i[j]}$, and $\alpha_{c[j],i[j]}$) and (2) covariates associate with settlement versus dismissal (i.e., $\boldsymbol{\beta}$).

Given our likelihood, we must also specify a set of priors for our parameters. Simply put, we used standard noninformative priors for both our nonhierarchical terms ($\bar{\alpha}$, $\boldsymbol{\beta}$) and our hierarchical terms (the various subscripted α terms α_c, α_i, and $\alpha_{c,i}$). Our priors for the former are given by

$$\bar{\alpha} \sim N(0, K^2) \qquad \beta_i \sim N(0, K^2)$$

where we set $K = 100$ and each β_i is an element of the vector $\boldsymbol{\beta}$. We note that our *a priori* independent prior on the β_i allows for *a posteriori* dependence (i.e., it does not imply that each β_i is a *posteriori* independent).

Our Bayesian approach allows for a principled sharing of information between the parameters for our hierarchical terms. For example, the average effect of circuit c is given by α_c. We pool each of these α_c terms toward a common mean in order to stabilize our estimates; circuits with little data will be pulled more strongly toward this common mean whereas circuits with much data will be pulled less strongly. In particular, we let

$$\alpha_c \sim N(0, \tau_c^2) \qquad \tau_c \sim Unif(0, M_1)$$

for $c = 1, \ldots, 12$ (where 12 represents the DC circuit). In this equation, the τ_c parameter controls how much each $\alpha_{c[j]}$ is pulled toward the common mean. When τ_c is large, there is correspondingly little shrinkage toward the common mean, and, when it is small, there is correspondingly greater shrinkage. As noted above, of course, the rate of shrinkage is also dependent on the sample size for each circuit.

We follow an identical procedure for industry groups, setting

$$\alpha_i \sim N(0, \tau_i^2) \qquad \tau_i \sim Unif(0, M_1)$$

for the $i = 1, \ldots, 12$ industries represented in Table 1. Finally, our prior for the interaction terms $\alpha_{c,i}$ is quite similar

$$\alpha_{c,i} \sim N(0, \tau_{c,i}^2) \qquad \tau_{c,i} \sim Unif(0, M_2).$$

As of yet, our priors are not completely specified because we have not given values for M_1 and M_2. Indeed, we tried three different choices, each of which has different implications about the nature of securities fraud class action litigation. First, we tried $M_1 = M_2 = 0$. This has the effect of setting all the subscripted α terms to zero, forcing complete shrinkage to the common intercept $\bar{\alpha}$. If circuits or industry groups had zero effect beyond that captured by the covariates X_j, this would be our preferred model.

Our second set of priors fixes $M_1 = 100$ and $M_2 = 0$. Setting $M_1 = 100$ allows the α_c and α_i terms to vary, being pulled to a common mean at the data-determined rate of τ_c and τ_i, respectively. Nonetheless, setting $M_2 = 0$ forces the $\alpha_{c,i}$ terms to zero. This means that the effect of a given industry is the same across all circuits or, alternatively, that the effect of a case being tried in a given circuit is the same for all industries.

This assumption may be false. For example, a telecommunications case specifically may be more likely to settle (or be dismissed) in the Second Circuit compared to the Ninth Circuit beyond whatever differences exist on average between the Second and Ninth Circuits. This is indeed an empirical claim and must be determined by the data. Thus, our third and final set of priors sets $M_1 = M_2 = 100$ and allows all subscripted α terms to vary.

C. Settlement Amount Model

Our model for settlement amounts conditional on a settlement follows a structure almost identical to that for the settlement amount model. However, since our response Y_j is a continuous variable, we use a Gaussian likelihood. In particular, we let

$$(Y_j | Z_j = 1) \sim N(\alpha_j + X_j^{\intercal}\boldsymbol{\beta}, \sigma^2).$$

Thus, for cases that settle we assume the average settlement amount is given by $\alpha_j + X_j^{\intercal}\boldsymbol{\beta}$ and there is an error component ε_j that is distributed normally with mean zero and homogeneous variance σ^2. In a slight abuse of notation, we again let α_j be the case-specific intercept term and $\boldsymbol{\beta}$ be the vector of slopes for our covariates. In all presentations of results, we will make it clear whether we are referring to the parameters of the settlement/dismissal model or those of the settlement amount model.

Our prior specifications for the settlement amount model are identical to those for the settlement/dismissal model and we again set $K = 100$ and use (1) $M_1 = M_2 = 0$, (2) $M_1 = 100$, $M_2 = 0$, and (3) $M_1 = M_2 = 100$. We further require a prior for σ in the settlement amount model, which we set to

$$\sigma \sim Unif(0, K).$$

V. Results

A. Model Selection

Before turning to the results of our settlement/dismissal and settlement amount models, we must select which of the three prior specifications performs "best." We recall our three specifications and the modeling assumptions implied by them.

1. **Prior Specification 1**: $M_1 = M_2 = 0$. This model sets all $\alpha_c = \alpha_i = \alpha_{c,i} = 0$, thus forcing each case to have the same intercept $\bar{\alpha}$. This would be the "correct" model if there were no variation in the outcome variables Y_j and Z_j due to circuits or industries beyond that captured by the X_j.
2. **Prior Specification 2**: $M_1 = 100$, $M_2 = 0$. This model sets all $\alpha_{c,i} = 0$ but allows the α_c and α_i terms to vary, pulling them to a common mean at the data-determined rates of τ_c and τ_i, respectively. Thus, each case has intercept $\bar{\alpha} + \alpha_{c[j]} + \alpha_{i[j]}$. This would be the "correct" model if there were variation in the outcome variables Y_j and Z_j due to circuits or industries beyond that captured by the X_j but that this variation decomposes into a fixed additive "circuit" component and a fixed additive "industry" component.
3. **Prior Specification 3**: $M_1 = M_2 = 100$. This model allows the α_c, α_i, and $\alpha_{c,i}$ terms to vary, pulling them to a common mean at the data-determined rates of τ_c, τ_i, and $\tau_{c,i}$, respectively. Thus, each case has intercept $\bar{\alpha} + \alpha_{c[j]} + \alpha_{i[j]} + \alpha_{c[j],i[j]}$. This would be the "correct" model if there were variation in the outcome variables Y_j and Z_j due to circuits or industries beyond that captured by the X_j and that this variation does not decompose into additive terms.

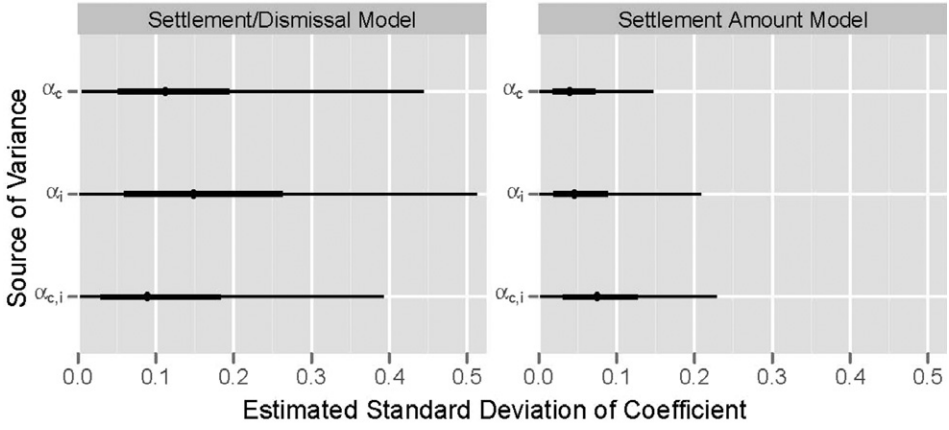
We note that Specification 3 is by far the most realistic. Furthermore, from a modeling perspective, rather than picking a single "discrete" model or, indeed, averaging over several such models, we prefer to expand the model so that it estimates all parameters. Specification 3 expands Specifications 1 and 2 by not restricting any α terms to zero; further, because these α parameters are pooled, they are essentially a weighted average of the overall value (i.e., set to zero) and the observed value (i.e., as in classical regression with interactions) where, as noted, the weights are determined by the parameters τ.

Indeed, an analysis of variance (Gelman 2005) for Specification 3 shows that the $\alpha_{c,i}$ terms contribute to both the settlement/dismissal model and the settlement amount model (see Figure 7). Indeed, in both cases, the standard deviation of the magnitude of the variation due to the $\alpha_{c,i}$ is estimated to be similar to that due to the α_c and α_i. Thus, both subjective and empirical considerations suggest that we should not fix the $\alpha_{c,i}$ at zero (i.e., we should not fix M_2 at zero).

B. Settlement/Dismissal Model Coefficients

We begin the presentation of our results by focusing on the covariate coefficient estimates for the settlement/dismissal model, which we present in the top panel of

Figure 7: Analysis of variance for the settlement/dismissal and settlement amount models using Prior Specification 3.



NOTE: The posterior median of the finite population standard deviation of each component is given by the dot. 50 percent and 95 percent posterior intervals are given by the thick and thin black lines, respectively.

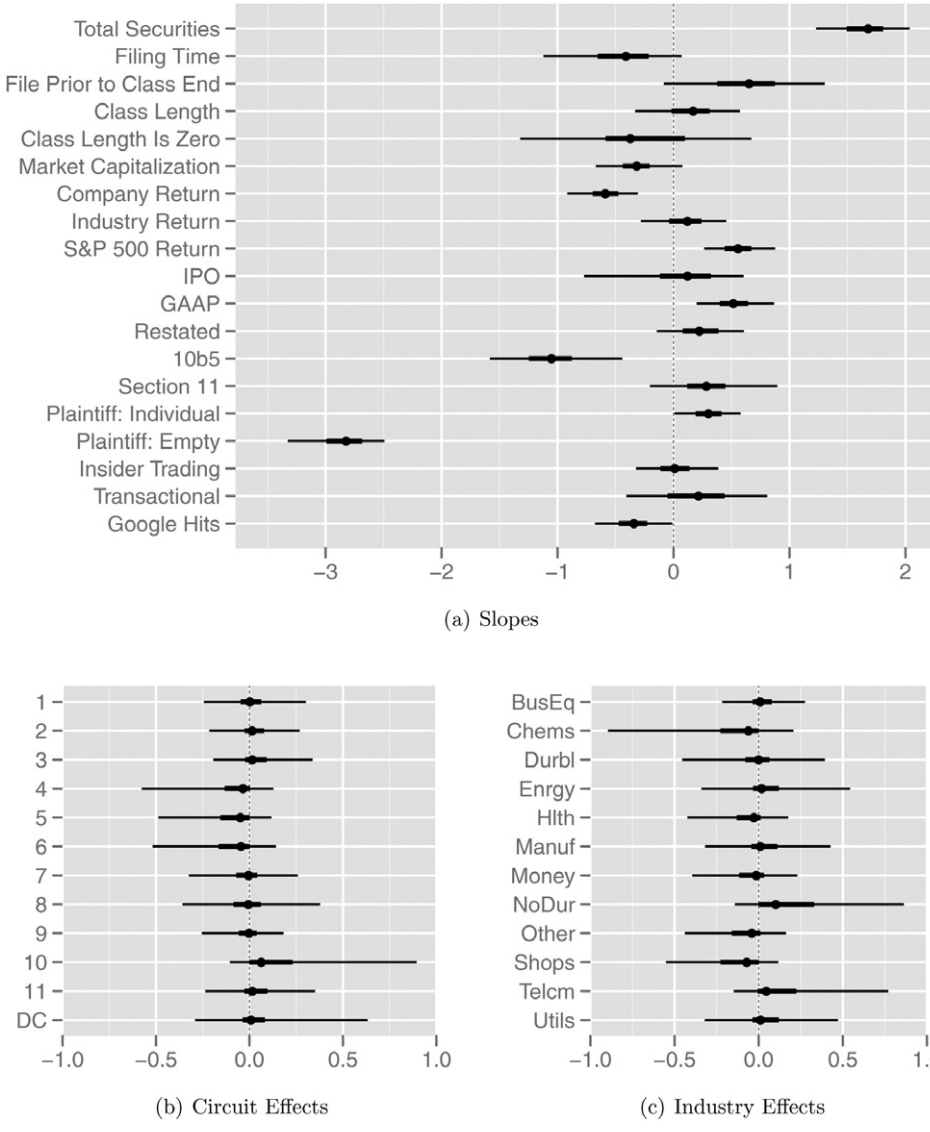
Figure 8.[2] We note that these results are on the logit scale; at the maximum, a one unit increase on the logit scale is roughly equivalent to a 25 percent increase on the probability scale and, hence, a given coefficient divided by four represents an upper bound on the change in the probability of settlement associated with a one unit change in a given predictor.

Not surprisingly, a great number of our coefficients have very high posterior probability of being nonzero (i.e., a great number of our coefficients are "statistically significant"[3]) for predicting whether a securities fraud class action lawsuit is settled or dismissed. Indeed, a larger number of securities types associated with a case, a higher return on the S&P 500 during the class period, whether or not GAAP violations were alleged, and having an individual (as opposed to an institutional) plaintiff listed all make a case more likely to settle. On the other hand, longer filing times, higher market capitalizations, a higher company

[2]We reiterate that all continuous variables have been transformed as indicated in the note to Figure 4 and then scaled by two standard deviations. Consequently, the slope estimates for the continuous variables presented in the top panel of Figure 8 are on the transformed and scaled scale; slope estimates on the raw transformed variable scale can be obtained by dividing each value by twice the standard deviation of the underlying raw transformed variable. For completeness, we note that the standard deviations are 0.48 (Total Securities), 3.41 (Filing Time), 1.34 (Class Length), 1.94 (Market Capitalization), 0.54 (Company Return), 0.39 (Industry Return), 0.24 (S&P 500 Return), and 2.05 (Google Hits).

[3]For Bayesian models, one typically does not make statements about statistical significance, but about the posterior distribution of a given parameter or a set of parameters. For example, when the 95 percent posterior interval of a given parameter (typically formed by taking the 2.5 and 97.5 percentiles of the posterior distribution of that parameter) does not overlap zero, there is very high posterior probability that the parameter is different from zero. Further, when noninformative priors are used (as they are here), the conclusions one would make by examining the posterior distribution of a given parameter (e.g., whether or not the 95 percent posterior interval overlaps zero) generally coincide with those one would make by using classical tests of statistical significance in a non-Bayesian setting.

Figure 8: Settlement/dismissal model coefficient estimates.



NOTE: The posterior median is given by the dot and 50 percent and 95 percent posterior intervals are given by the thick and thin black lines, respectively. On the logit scale, a given coefficient divided by four represents an upper bound on the change in the probability of settlement associated with a one unit change in a given predictor (see main text for details).

return during the class period, being a Rule 10b-5 case, having no plaintiff listed, and a larger number of Google hits in the year prior to filing all serve to increase the likelihood of a dismissal.

There are many plausible drivers of these phenomena. The merits of the case, for instance, should drive settlement versus dismissal and we see variables like the allegation of GAAP violations doing exactly that. On the other hand, the outcome of the motion to dismiss should also be impacted by the "plaintiff selection effect"; namely, while plaintiffs are likely to seek out cases that have a high probability of surviving a motion to dismiss, they are also likely to make rational tradeoffs by, for example, taking on cases with lower probability of settlement when such cases are likely to have large settlement amounts conditional on surviving the motion to dismiss. This selection effect is most tied to variables like market capitalization and the number of Google hits; consistent with the selection effect, the results in the top panel of Figure 8 show that these variables are associated with increased likelihood of dismissal, while the results presented in the next section show that they are also associated with larger settlement amounts conditional on surviving the motion to dismiss. The same is true for cases with an institutional plaintiff listed, suggesting that institutional plaintiffs choose to become involved in cases with larger damages potential.

In the bottom panels of Figure 8, we present our respective estimates for the effect of each circuit and industry. As can be seen, our hierarchical model engages in principled sharing across the various parameters for each group (i.e., circuit and industry). For instance, circuits like DC, which have relatively few cases, are pulled strongly to the overall mean of zero. That said, if a circuit has relatively few cases but a strong enough effect in the data, its coefficient will be moved from the overall mean; for instance, the Fourth and Tenth Circuits have relatively few cases but, those they do have, are substantially less likely and more likely, respectively, to settle. Nonetheless, the posterior intervals for the circuits with few cases are generally much wider than those with a large number. For example, the Second and Ninth Circuits, which each account for a large fraction of the cases, have the narrowest intervals of all. Turning to the industries, it appears that suits brought against nondurables companies and telecommunications companies may be somewhat more likely to settle, whereas cases brought against chemicals and shops may be somewhat more likely to be dismissed.

Rather than presenting each of the 144 $\alpha_{c,i}$ coefficients individually, we present the bottom 10 and top 10 in Table 2. As can be seen, most of the bottom 10 coefficients (i.e., circuit-industry combinations for which a case is most likely to be dismissed) are either from the Fourth or Sixth Circuit and/or the chemicals or shops industry. On the other hand, cases that are most likely to settle are those in the Tenth Circuit or nondurables industry.

C. Settlement Amount Model Coefficients

Turning to the slope estimates for the settlement amount model, which are given in the top panel of Figure 9,[4] we again see a number of parameters with high posterior probability of

---

[4]As noted in footnote 2, slope estimates on the raw transformed variable scale can be obtained by dividing the value for each continuous variable in the top panel of Figure 9 by twice the standard deviation of the underlying raw transformed variable. Again, for completeness, we note that the standard deviations are 0.55 (Total Securities), 3.69

Table 2: Settlement/Dismissal Model: Bottom and Top Intercept Coefficients

Bottom Ten Combinations				*Top Ten Combinations*			
		$(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$				$(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$	
Circuit	*Industry*	*Mean*	SD	*Circuit*	*Industry*	*Mean*	SD
5	Chems	1.49	0.73	10	NoDur	2.08	0.78
6	Chems	1.49	0.73	10	Telcm	2.05	0.79
5	Shops	1.49	0.69	2	NoDur	2.00	0.73
4	Chems	1.50	0.74	3	NoDur	1.98	0.75
6	Shops	1.51	0.68	DC	NoDur	1.97	0.77
4	Shops	1.53	0.69	10	BusEq	1.97	0.72
6	Other	1.53	0.69	11	NoDur	1.97	0.75
4	Other	1.53	0.71	10	Enrgy	1.96	0.73
9	Chems	1.55	0.72	10	Utils	1.94	0.76
6	Hlth	1.55	0.67	1	NoDur	1.94	0.73

Population Level $\bar{\alpha}$: Mean = 1.73, $SD = 0.65$

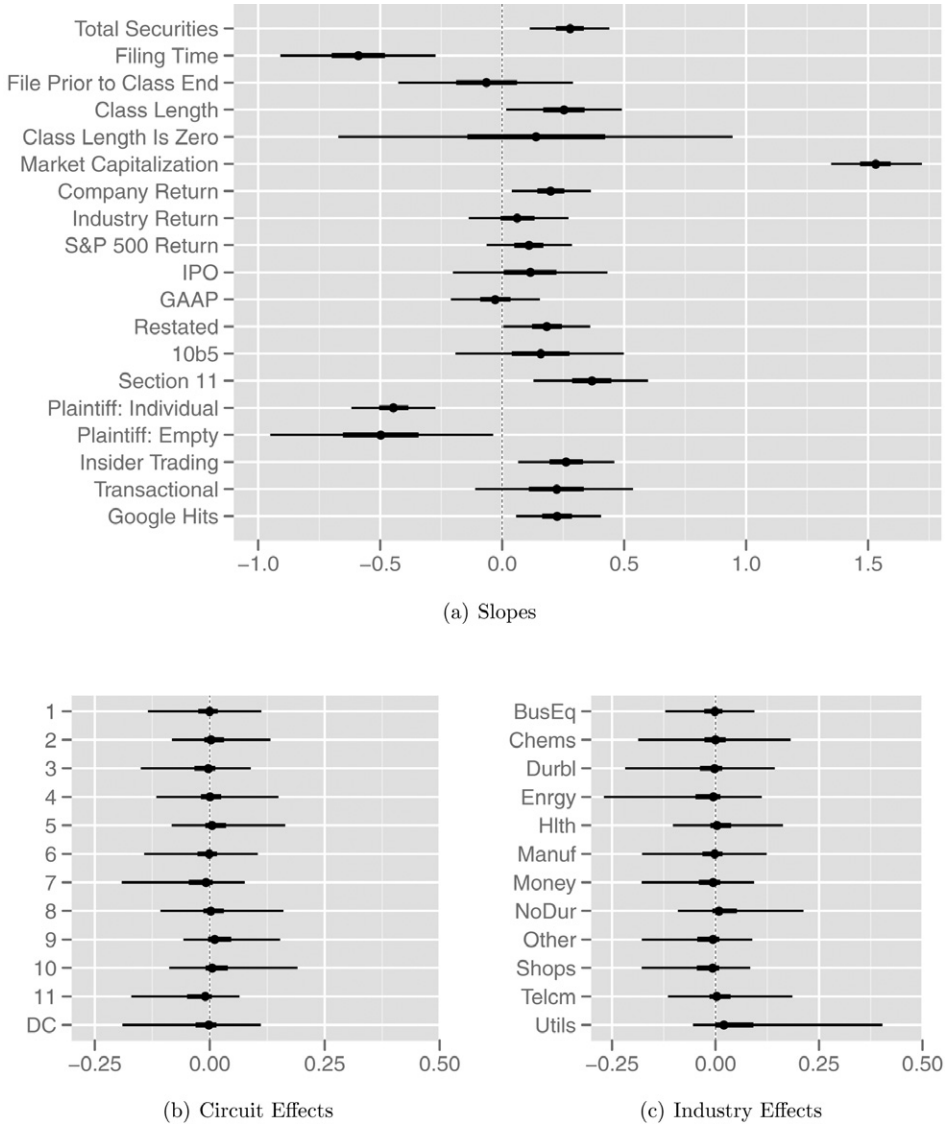
NOTE: For each circuit-industry combination, we provide the posterior mean and posterior standard deviation for the intercept term $(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$. The posterior mean and standard deviation of $\bar{\alpha}$ is also provided for comparison.

being nonzero. In particular, it appears the total number of security types, the length of the class period, the market capitalization, the company return during the class period, whether or not earnings were restated, whether or not the case was a Securities Act Section 11 case, whether or not insider trading was alleged, and the number of Google hits all positively impact settlement amounts; on the other hand, longer filing times and having no institutional plaintiff listed (i.e., having only an individual plaintiff listed or having no plaintiff listed) are associated with lower settlements. The latter association is consistent with the institutional plaintiff selection effect noted above (i.e., institutional plaintiffs are more likely to be involved in cases with higher damages potential).

Many of these results are quite plausible. For instance, firms with higher market capitalizations and Google hits are typically larger firms and this serves as a proxy for how much damage can be done and how large a settlement can be extracted. Interestingly, some merits variables such as Restated and Insider Trading, which in theory should only affect whether a case settles or is dismissed, also impact the settlement amounts, thus suggesting that decisions over whether or not there were damages versus how great those damages were may not be entirely independent. Indeed, qualitative interviews suggest that merits variables like Restated and Insider Trading are sufficiently "sexy" that they impact both settlement amounts as well as the likelihood of surviving the motion to dismiss (Baker & Griffith 2009); notably, the interviews of Baker and Griffith (2009) also suggest that GAAP violations are not sexy but boring, consistent with our model results that such violations impact settlement/dismissal but not settlement amount.

(Filing Time), 1.20 (Class Length), 1.93 (Market Capitalization), 0.56 (Company Return), 0.42 (Industry Return), 0.25 (S&P 500 Return), and 2.10 (Google Hits).

Figure 9: Settlement amount model coefficient estimates.



NOTE: The posterior median is given by the dot and 50 percent and 95 percent posterior intervals are given by the thick and thin black lines, respectively.

In the bottom panels of Figures 9, we present our circuit and industry effect size estimates, respectively. Again, the model engages in principled sharing across circuits and industries and effect sizes are pooled toward the common mean. Nonetheless, it appears that, above and beyond any effects of our covariates X_j, there are no massive variations from

Table 3: Settlement Amount Model: Bottom and Top Intercept Coefficients

Bottom Ten Combinations				*Top Ten Combinations*			
		$(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$				$(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$	
Circuit	*Industry*	*Mean*	SD	*Circuit*	*Industry*	*Mean*	SD
6	Money	12.25	0.39	5	Utils	12.46	0.41
7	Money	12.27	0.39	9	Utils	12.45	0.40
7	Manuf	12.27	0.39	6	Utils	12.44	0.41
3	Other	12.27	0.39	8	Utils	12.43	0.41
6	Enrgy	12.27	0.40	10	Utils	12.43	0.41
11	Other	12.28	0.38	2	Hlth	12.42	0.39
11	Manuf	12.28	0.39	9	NoDur	12.42	0.38
11	Shops	12.28	0.38	2	Utils	12.42	0.40
3	Shops	12.28	0.38	4	Utils	12.42	0.40
11	Enrgy	12.29	0.39	1	Utils	12.41	0.41

Population Level $\bar{\alpha}$: Mean = 12.35, $SD = 0.36$

NOTE: For each circuit-industry combination, we provide the posterior mean and posterior standard deviation for the intercept term $(\bar{\alpha}+\alpha_i+\alpha_c+\alpha_{c,i})$. The posterior mean and standard deviation of $\bar{\alpha}$ is also provided for comparison.

circuit to circuit; the Eleventh Circuit appears to have modestly lower settlement amounts, whereas the Ninth and Tenth Circuits have modestly higher settlements amounts. Similarly, utilities firms have somewhat higher settlement amounts.

As we did for the settlement/dismissal model, rather than presenting each of the 144 $\alpha_{c,i}$ coefficients of the settlement amount model individually, we present the bottom 10 and top 10 in Table 3. As can be seen, most of the bottom 10 coefficients (i.e., circuit-industry combinations that have the lowest settlements) come from the Eleventh Circuit. On the other hand, cases that have the highest settlements appear to be utilities firms.

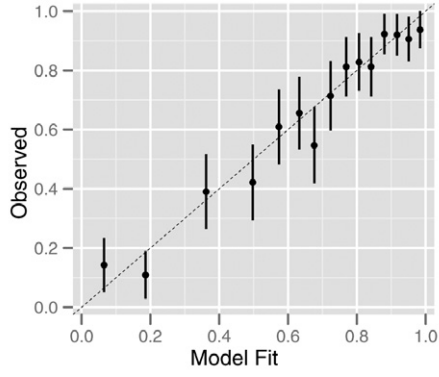
D. Model Evaluation and Application

In Figure 10, we evaluate our two models against the observed data. To assess the models' ability at prediction, we held out a random 25 percent of our observations, thus leaving 899 cases as in-sample and 299 as out-of-sample. All 899 observations were used to fit the settlement/dismissal model, whereas the 592 cases that settled were used to fit the settlement amount model. Of the 299 out-of-sample cases, 193 settled.

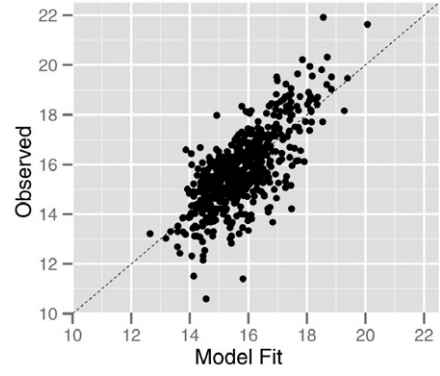
In the left panels of Figure 10, we put each of the 899 in-sample cases (top) and each of the 299 out-of-sample cases (bottom) into one of 15 equally-sized "bins" based on the model's prediction of settlement. For each bin, we calculate the fraction of cases that actually did settle along with ± 2 standard errors. As can be seen, the model predictions appear to be well-calibrated with little degradation in performance on the out-of-sample cases.

In the right panels of Figure 10, we plot the predicted versus observed settlements for each of the 592 in-sample cases (top) and each of the 193 out-of-sample cases (bottom) that settled. Again, the model generally fits quite well, with most points lying on or near the 45-degree line and no violations of linearity. The root mean square error of the model is 1.03 in-sample and 1.10 out-of-sample; similarly, the median absolute error is 0.64 in-sample

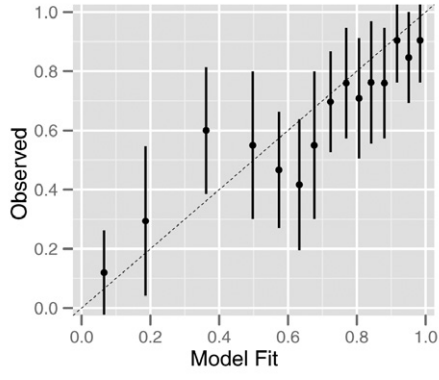
Figure 10: Model evaluation.



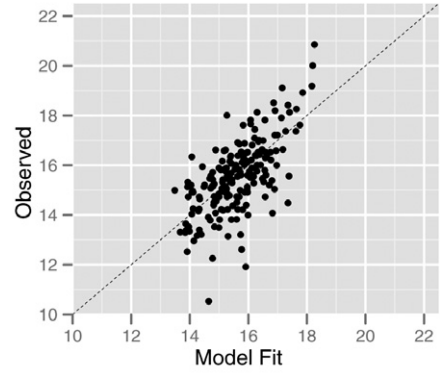
(a) In-Sample Settlement/Dismissal Model Fit



(b) In-Sample Settlement Amount Model Fit



(c) Out-of-Sample Settlement/Dismissal Model Fit



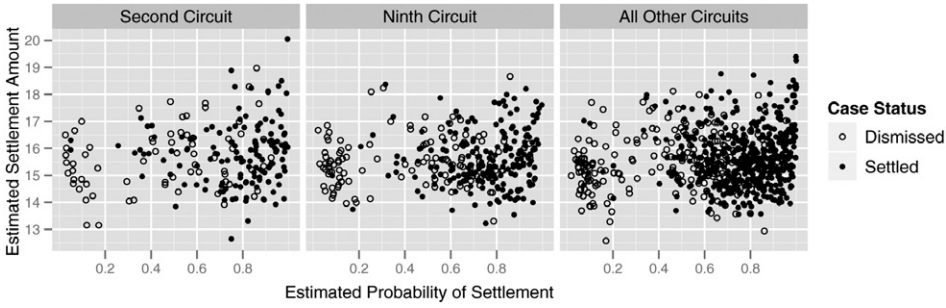
(d) Out-of-Sample Settlement Amount Model Fit

NOTE: Model evaluation in-sample (top) and out-of-sample (bottom). On the left, we evaluate the settlement/dismissal model by placing our model estimated probabilities of settlement into 15 equally-sized "bins" and calculating the observed probability of settlement for each bin along with ± 2 standard errors. On the right, we evaluate our model for settlement amounts conditional on settlement by plotting our model estimated settlement amount versus the observed settlement amount.

and 0.68 out-of-sample. The forecasts are also calibrated as 74 percent (95 percent) of the predicted settlements are within one (two) standard deviation(s) of the true settlement for the in-sample cases; 72 percent (94 percent) of the predicted settlements are within one (two) standard deviation(s) of the true settlement for the out-of-sample cases. We can again conclude the model predicts well with little degradation in performance on the out-of-sample cases.

In addition to the tests described above, we also performed a more difficult out-of-sample evaluation. In particular, we held out the 286 cases that were filed in either 2003

Figure 11: Comparison of estimated probability of settlement and estimated settlement amount by circuit.



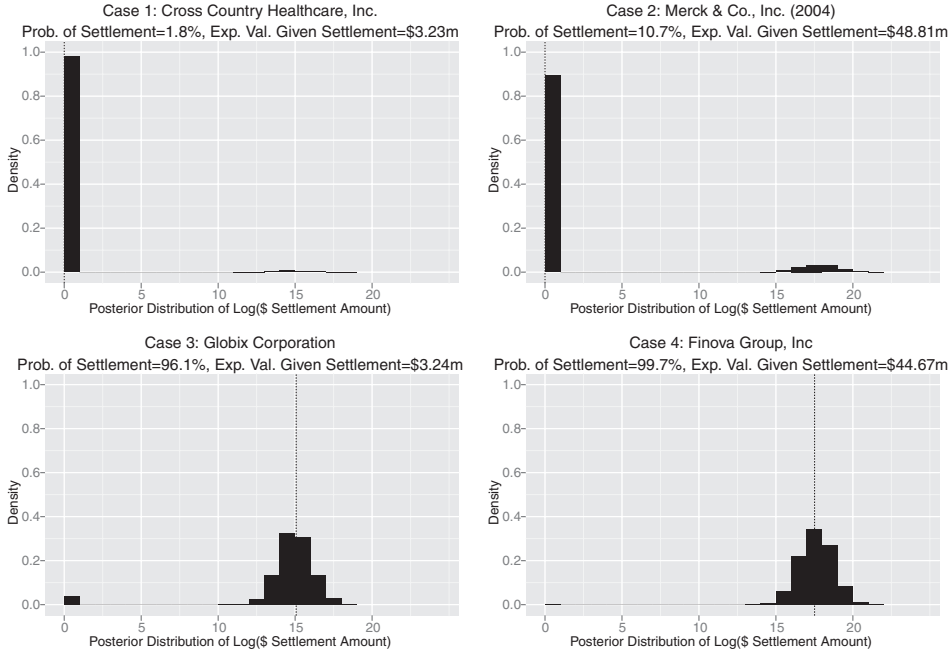
NOTE: We plot our estimated probability of settlement by the estimated settlement amount conditional on settlement for the Second Circuit (left), Ninth Circuit (middle), and all other circuits (right).

or 2004 (i.e., the last two years of our data; these cases account for 24 percent of the data). As above, all 912 of the remaining cases were used to fit the settlement/dismissal model, whereas the 624 cases that settled were used to fit the settlement amount model. Out-of-sample results under this more difficult hold-out schema remained strong. In particular, the diagnostic plots and fit statistics for this hold-out schema differed minimally from those discussed above and presented in Figure 10 for the random hold-out schema.

Given that our models fit well both in- and out-of-sample, for the remainder of this section we return to using the models as estimated on the full sample of cases. In Figure 11, we plot the predicted probability of settlement versus the predicted settlement amount conditional on settlement for each of the 210 Second Circuit cases (left), 324 Ninth Circuit cases (middle), and 664 remaining cases (right). As can be seen, the model predicts settlement incidence quite well, with the majority of settled cases appearing on the right of the plot and a majority of the unsettled cases appearing on the left. Interestingly, the predicted probabilities of settlement and predicted settlement amounts appear uncorrelated in all three panels of the plot. That is, cases that are predicted to be more likely to settle are not predicted to settle for greater amounts. This is potentially quite important and has implications for "selecting" which cases to pursue from a plaintiff standpoint. Finally, it is worth noting that, on the whole, the predicted settlement probabilities and amounts for the important (i.e., largest in terms of number of securities fraud class action cases) Second and Ninth Circuits appear quite similar to those for the other circuits.

In Figure 12, we demonstrate a principal value of our two models by showing the full posterior distribution of settlement outcomes as predicted by our models for four selected cases along with the actual settlement amount. Case 1 is an archetypal low impact case: it is unlikely to settle and, if it does settle, the expected settlement amount is relatively low. Case 4, on the other hand, is an archetypal high impact case: it is highly likely to settle and for a substantial amount. Cases 2 and 3, however, are much more interesting and disparate in terms of their impact. Though these cases have similar expected settlements, and therefore would appear similar based on relatively unsophisticated analyses of expected settlements,

Figure 12: Posterior distribution of settlement amounts of four cases.



NOTE: The actual settlement amount is given by the vertical dashed line. For the purpose of this figure, we set $\log(0) = 0$, where 0 denotes a dismissal.

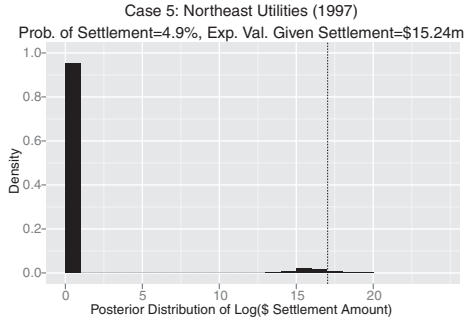
our model shows that these two cases are actually quite different. Case 3 is very likely to settle and for a relatively low amount; in other words, it is a relatively known quantity. On the other hand, Case 2 is a "black swan" (Taleb 2007): it is relatively unlikely to settle but, if it does, it will settle for a large amount. A principal benefit of our model is the ability to identify such cases so that precautions can be taken from the standpoint of the defense (or, alternatively, so that these cases can be capitalized on from the standpoint of the plaintiff).

We further emphasize this notion with the example of the Northeast Utilities litigation presented in Figure 13. Black swan cases like this one do not just settle in some "hypothetical" world; indeed, they sometimes settle in the real world, thus creating low probability, high variance, high impact events. The Northeast Utilities case had only a 5 percent chance of settling. However, the expected settlement amount if it did settle was quite large. In fact, this case did settle for $25 million (i.e., 17.0 on the log scale), thus demonstrating the large risk associated with black swan cases.

VI. DISCUSSION

This article develops a model that, to our knowledge, is the first predictive model of securities fraud class action lawsuits. Our model is predictive of settlement incidence

Figure 13: Posterior distribution of settlement amounts for the Northeast Utilities case.



Note: The actual settlement amount is given by the vertical dashed line. For the purpose of this figure, we set log(0) = 0, where 0 denotes a dismissal.

(i.e., likelihood of dismissal) and outcome (i.e., expected settlement amount) at the time a case is filed. As such, it only uses variables whose values can be calculated on the day of the filing, distinguishing it from other models such as those estimated by Cornerstone Research and NERA that use variables that are known only after the filing date. Overall, both of our models provided accurate fits to the data.

We found that many of our variables are predictive of whether a securities fraud class action is settled or dismissed. Factors that indicate a case will most likely settle include a greater number of classes or types of securities associated with the case, a higher return on the S&P 500 during the class period, whether or not GAAP violations were alleged, and having an individual plaintiff listed. Factors that indicate a case is less likely to settle include longer filing times, higher market capitalization, a higher company return during the class period, having an institutional plaintiff listed, and greater public notoriety (as measured by the number of Google hits in the year prior to filing).

Our analysis also revealed a number of variables that are predictive of the settlement amount. Factors found to positively impact the settlement amount include the total number of securities, the length of the class period, the market capitalization of the company, the company return during the class period, whether or not earnings were restated, whether or not the case was a Securities Act Section 11 case, whether or not insider trading was alleged, the existence of an institutional plaintiff, and the number of Google hits. Factors associated with lower settlement amounts include longer filing times and not having an institutional investor listed (i.e., having only an individual plaintiff listed or having no plaintiff listed).

Interestingly, we found that those cases predicted to be more likely to settle were not predicted to settle for greater amounts. Cases coded as alleging GAAP violations are more likely to settle but those that do settle do not have higher settlement amounts. This result is likely due to the fact that an allegation of a GAAP violation significantly bolsters the merits of the case. This in turn increases the plaintiff's chances of surviving a motion to dismiss, making it more appealing for the plaintiff to take on such a lawsuit even if the potential damage award is relatively low.

A principal benefit of our model is this ability to identify cases that are relatively unlikely to settle but, upon settling, will settle for a large amount. In such cases, precautions can be taken by the defense or an opportunity can be capitalized on by the plaintiff. Cases coded as Rule 10b-5 class actions were found less likely to settle but those that do settle have higher settlement amounts. This result is likely attributable to the greater damages available to Rule 10b-5 plaintiffs. Our findings are also consistent with the plaintiff selection effect. That is, plaintiffs will generally attempt to pick cases that will survive a motion to dismiss but they are rationally more willing to pursue cases with lower likelihoods of settlement when such cases are likely to have large settlements (provided they were to survive a motion to dismiss). Because institutional plaintiffs are given priority under the PSLRA, they appear to be able to choose cases with larger settlement potential. Cases without an institutional plaintiff are more likely to survive the motion to dismiss. It is possible that this pattern results not only from institutional plaintiffs selecting the high potential value cases but also from plaintiffs' lawyers exercising more care regarding the merits of cases with only an individual plaintiff.

Additionally, using our principled Bayesian strategy, we were able to estimate separate effects for each combination of circuit and industry—even with a paucity of data for some combinations. These estimates revealed a number of differences among the various circuits and industries that are interesting in their own right and suggest further work to uncover them at a deeper level.

In closing, we wish to mention several areas of future research that could lead to improvements in fit and more substantive understanding of the determinants of case outcomes. This first would be to build a model that allows the slopes to vary rather than just the intercepts. Such a model would allow one to determine whether the effect of, say, market capitalization on settlement incidence and amount varies across circuits and industries. Another research idea, which would allow for more substantive understanding of the differences between the different circuits and industries, would be to further model the intercepts (i.e., the α_c, α_i, and $\alpha_{c,i}$) in terms of circuit- or industry-level variables (e.g., a measure of average ideology of the judges in a given circuit). Finally, one could examine the distribution of settlement outcomes to see if there is a heterogeneous variance component that could be modeled, whether in terms of circuits, industries, or some other grouping. These ideas are generalizations of the existing model and thus could be compared to it in order to assess the importance of these potentially important innovations.

References

Alexander, J. C. (1991) "Do the Merits Matter? A Study of Settlements in Securities Class Actions," 43(3) *Stanford Law Rev.* 497.

Baker, T., & S. J. Griffith (2007) "Predicting Corporate Governance Risk: Evidence from the Directors' and Offers' Liability Insurance Market," 74 *Univ. of Chicago Law Rev.* 487.

—— (2009) "How the Merits Matter: Directors' and Officers' Insurance and Securities Settlements," 157(3) *Univ. of Pennsylvania Law Rev.* 755.

Buckberg, E., T. S. Foster, R. I. Miller, & S. Plancich (2005) *Recent Trends in Shareholder Class Action Litigation: Bear Market Cases Bring Big Settlements*, Technical Report, National Economic Research Associates (NERA).

Choi, S. J. (2003) "The Evidence on Securities Class Actions," 56(6) *Vanderbilt Law Rev.* 1465.

—— (2007) "Do the Merits Matter Less After the Private Securities Litigation Reform Act?" 23(3) *J. of Law, Economics, & Organization* 598.

Choi, S. J., K. K. Nelson, & A. C. Pritchard (2009) "The Screening Effect of the Private Securities Litigation Reform Act," 6(1) *J. of Empirical Legal Studies* 35.

Cox, J. D., R. S. Thomas, & L. Bai (2008) "There Are Plaintiffs and . . . There Are Plaintiffs: An Empirical Analysis of Securities Class Action Settlements," 61(2) *Vanderbilt Law Rev.* 355.

Cox, J. D., R. S. Thomas, & D. Kiku (2006) "Does the Plaintiff Matter? An Empirical Analysis of Lead Plaintiffs in Securities Class Actions," 106 *Columbia Law Rev.* 1587.

Eisenberg, T., & G. P. Miller (2010) "Attorney Fees and Expenses in Class Action Settlements: 1993–2008," 7(2) *J. of Empirical Legal Studies* 248.

Fitzpatrick, B. T. (2010) "An Empirical Study of Class Action Settlements and Their Fee Awards," 7(4) *J. of Empirical Legal Studies* 811.

Foster, T. S., D. N. Martin, V. M. Juneja, & F. C. Dunbar (2000) *Trends in Securities Litigation and the Impact of the PSLRA*, Techical Report, National Economic Research Associates (NERA).

French, K. (2011) *Kenneth French Data Library*. Available at <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html>.

Gelman, A. (2005) "Analysis of Variance—Why it is More Important than Ever," 33(1) *Annals of Statistics* 1.

Gelman, A., J. Carlin, H. Stern, & D. Rubin (2003) *Bayesian Data Analysis*, 2d ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & J. Hill (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Geman, S., & D. Geman (1984) "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," 6 *IEEE Transaction on Pattern Analysis & Machine Intelligence* 721.

Google News Archive (2011) *Google, Inc.* Available at <http://news.google.co.uk/archivesearch>.

Hastings, W. (1970) "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," 57 *Biometrika* 97.

Johnson, M. F., K. K. Nelson, & A. C. Pritchard (2006) "Do the Merits Matter More? The Impact of the Private Securities Litigation Reform Act," 23(3) *J. of Law, Economics, & Organization* 627.

Johnson-Skinner, D. T. (2009) "Paying-to-Play in Securities Class Actions: A Look at Lawyers' Campaign Contributions," 84(6) *New York Univ. Law Rev* 1725.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, & E. Teller (1953) "Equation of State Calculations by Fast Computing Machines," 21 *J. of Chemical Physics* 1087.

NERA (1999) *What Do NERA Experts Do for Clients in Securities Fraud Litigation?* Technical Report, National Economic Research Associates (NERA).

Plancich, S., & S. Starykh (2009) *Recent Trends in Securities Class Action Litigation: 2009 Year-End Update*, Technical Report, NERA Economic Consulting.

Pritchard, A. C., & H. A. Sale (2005) "What Counts as Fraud? An Empirical Study of Motions to Dismiss Under the Private Securities Litigation Reform Act," 2(1) *J. of Empirical Legal Studies* 125.

Ryan, E. M., & L. E. Simmons (2009) *Securities Class Action Settlements: 2009 Review and Analysis*, Technical Report, Cornerstone Research.

Taleb, N. N. (2007) *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.

University of Chicago Booth School of Business (2011) *Center for Research in Security Prices*. Available at <http://www.crsp.com/index.html>.

Yahoo! Finance (2011) *Yahoo, Inc.* Available at <http://finance.yahoo.com>.