The first part of this appendix describes the process of collecting covenant violation data from 10-K and 10-Q SEC quarterly filings for the universe of Compustat non-financial firms from 1996 to 2009. The second defines all main variables used in the analysis.

We are providing two data sets, described below, for public use. We only make three requests of researchers that use the data. First, please read this document carefully. We have worked hard to describe exactly how these data are collected. Second, please acknowledge somewhere in your research the source of the data. Third, please refer readers to this appendix for more information on the data.

A. Covenant Violation Data

The initial sample of firm-quarter observations are derived from a September 2009 extract of the fundamentals quarterly table from Compustat. We include any U.S. firm (fic = “USA”) outside of the financial industry (sic outside of 6000 to 6999) and all firm-quarter observations with non-missing information on total assets (atq), total sales (saleq), common shares outstanding (cshoq), closing share price (prccq), and the exact calendar quarter (datacqtr) of the observation. We make these deletions because observations without these any of these five variables are difficult to match to a corresponding SEC filing; conversely observations with all five of these variables nearly always can be matched with a corresponding SEC filing.

1. Matching Compustat quarterly observations to Edgar websites

The first step in our data collection process is matching each quarterly Compustat observation to the SEC filing that generated the Compustat data. Our starting point is the SEC Edgar website that contains indices of every filing submitted to the commission. It is located here: http://www.sec.gov/cgi-bin/edgar_archive_indices.

Using these index files, we create a list of every 10-Q and 10-K filing by any firm. We use a Perl script to pull identifying information from each filing from the corresponding EDGAR website where the filing is located. Every SEC filing has a standard header which contains important information including firm name, firm address, the central index key (CIK), the IRS tax number of the firm, and the reporting date of the filing. We extract all of this information to form an SEC matching file.

We then match Compustat observations to this file. We do three iterations. First, we merge based on the central index key (CIK), which is in both data sets. For any unmatched observations, we match on the IRS tax identification number. This variable is ein in Compustat and is included only in the fundamentals annual table. Finally, we hand match
the remaining observations. In all cases, matching is done by firm-quarter, using the
calendar quarter of the Compustat observation and the reporting date of the SEC filing.
Using this process, we are able to match 98% of Compustat firm-quarter observations,
resulting in a match for 286,437 firm-quarters.

The resulting file, $CSTATSEC\_NSS\_20091005.dta$, can be matched to Compustat using the
two variables $gvkey$ and $datacqtr$. It is available at:
http://faculty.chicagobooth.edu/amir.sufi/$CSTATSEC\_NSS\_20090701.dta$. This data set
includes several valuable pieces of information: the website of the filing associated with the
quarterly observation, the exact date of the report and the date of the filing, the exact name of
the firm at the time of the filing, and the filing type (i.e., 10-K or 10-Q). As opposed to the
firm name variable in Compustat ($conm$), the company name in this filing is not back-filled.
It represents the exact firm name at the time of the filing. This can be very useful for
matching historical data to other data sets based on name.

Using a perl script, we then download all of the 10-K/10-Q filings for observations in
$CSTATSEC\_NSS\_20091005.dta$. The perl script visits the website listed in the variable
websiteSEC and downloads the filing. Some filings are in .html format, which makes text
searching difficult. We utilize an html to text converter to get rid of html tags. The result is
a set of text files matched to Compustat observations. We use the text files to create
additional data to augment that which is available from Compustat.

2. Searching for violations

Our goal is to search the SEC filings to identify the occurrence of a covenant violation. Our
methodology reflects a desire to minimize the number false negatives while limiting the
amount of false positives that must be corrected manually. We have purposely been
conservative, meaning that our methodology generates a lot of false positives and requires
abundant manual coding.

In order to create the best text-searching algorithm to find reported violations in SEC filings,
we first created a random sample of 1,000 firm-year observations using the SEC 10-K filings.
We manually read these filings to find any mention of a violation. For these 1,000
observations, we know exactly whether the firm is in violation of a covenant or not. We then
used this “true” violation data set to test our text-search algorithm. For the 1,000 firm-year
observations, we find 105 violations. The incidence of violations in this sample is larger than
in the sample employed in the paper mainly because we examine only the 10-K filings,
which have a higher incidence of reported violations.

After several iterations, we find that the best text-search algorithm for finding violations is
the following. If the filing contains the word “covenant,” then our algorithm searches for the
following five terms within three lines above or below the line containing “covenant”: “waiv,”
“viol,” “in default,” “modif,” and “not in compliance.” This particular search
methodology finds 94 of the 105 violations in our test dataset. For comparison, the
methodology used in Roberts and Sufi (2009) and Nini, Smith, and Sufi (2009), when
applied to this sample, finds only 66 of the 105 violations. The algorithm used in those
papers had a more limited set of search terms, resulting in a significant increase in false negatives.

Although the text-search algorithm finds roughly 90% of all actual reported violations, it also produces a large number of false positives. In the sample of 1,000 firm-year observations, we find 117 false positives. Given the large number of false positives, when we apply the text-search algorithm to the universe of all filings, we must manually inspect the paragraphs around each “hit” to ensure the proposed violation is an actual violation. Our text search algorithm outputs the five lines before and after the hit to allow for manual reading.

The vast majority of covenant violations are handled by a contractual waiver, meaning the lender voluntarily relinquishes the rights granted following the default, perhaps in exchange for concessions from the borrower. In our coding methodology, we count waivers as violations. This is important, because as discussed in Dichev and Skinner (2004) and Roberts and Sufi (2009), information on firm covenant violations is available given SEC Regulation S-X, which requires that “any breach of a covenant of a[n] … indenture or agreement which … exist[s] at the date of the most recent balance sheet being filed and which has not been subsequently cured, shall be stated in the notes to the financial statements” (SEC (1988), as quoted by Beneish and Press (1993)). As Sufi (2007b) notes, the SEC has reinforced this requirement in recent interpretations: “companies that are, or are reasonably likely to be, in breach of such covenants must disclose material information about that breach and analyze the impact on the company if material (SEC (2003)).” Since waivers typically expire in a fairly short period, violating firms will again being in violation of the contract, absent any change in circumstances. This encourages companies to report violations, even if granted a waiver prior to the filing date. However, firms do have some discretion when reporting waivers of violations.

One possible concern is that a number of firms choose not to report violation waivers for relatively minor waivers, leading to the reported waivers representing, on average, more serious violations. However, as noted in Sufi and Roberts (2009), “a comparison of observable measures of credit quality and investment around the initial reported covenant violation in our sample versus the initial violation in previous studies reveals very similar patterns” For example, cash flow and capital expenditures show patterns around the first reported violation in our sample that are almost identical to those found in studies by Dichev and Skinner (2002) and Chava and Roberts (2008), which suggests that initial reported violations in our sample correspond closely to initial actual violations.”

Covenant violations through 2008 are available in CSTATVIOLATIONS_NSS_20091005.dta, which can be matched to Compustat using the two variables gvkey and datadate. The variable of interest is viol, which is an indicator variable that is equal to one if we find a covenant violation. The data set is available at: http://faculty.chicagobooth.edu/amir.sufi/CSTATVIOLATIONS_NSS_20091005.dta

3. **Additional notes**
There are a few additional notes worth mentioning. First, we collect the covenant violation data for the 10-Ks and 10-Qs separately. When we collect the data for the 10-Ks, we record as a violation any violation that occurred at any point during the fiscal year. We do this because the 10-K filing is often a “catch-all” where the firm reports information that it avoids reporting in the typically shorter 10-Q filing. Violations on 10-Qs are only recorded if the violation takes place in the fiscal quarter in question.

Second, the incidence of repeat covenant violations is quite high in the data. This is due to two factors. First, the violation data for 10-Ks represents a violation at any point in the fiscal year. If the firm repeats information on the violation in both a 10-Q and the 10-K in the same fiscal year, it will be counted twice. Second, there is high serial correlation in violations given that waivers expire and often must be granted repeatedly. For both of these reasons, we believe the most useful information in the data is for new violations, which we define as violations by a firm that has not reported a violation for the past four quarters.

Third, we do not collect violations of non-financial covenants, such as limits on capital expenditures or acquisitions.

Fourth, we strongly advocate that all users of the data investigate the exact language firms use when reporting violations. We would advocate picking 5 to 10 violations and examining the 10-K or 10-Q filing that corresponds to the violation. This will give the user a better sense of the advantages and disadvantages of the data.
B. Compustat Variable Definitions

Total assets = atq
Average assets = ((Total assets) + (lagged Total assets)) / 2
Market-to-book-ratio = (Market value) / (Total assets), where

Market value = Market value of equity – book value of equity + total assets,
Market value of equity = prccq * cshoq,
Book value of equity = Total assets – ltq + txditcq

Total debt = dltcq + dlttq
Leverage ratio = (Total debt) / (Total assets)
Net worth to assets ratio = seqq / (Total assets)
Current ratio = actq / lctq
PPE scaled by assets = ppentq / (Total assets)
Cash scaled by assets = cheq / (Total assets)
Shareholder payout = prstkcq + dvq
Operating income scaled by average assets = oibdpq / (Average assets)
Interest expense scaled by average assets = xintq / (Average assets)
Capital expenditures quarterly = capxy adjusted for fiscal quarter accumulation
Cash acquisitions quarterly = aqcy adjusted for fiscal quarter accumulation
Capital expenditures scaled by average assets = (Capital expenditures) / (Average assets)
Cash acquisitions scaled by average assets = (Cash acquisitions) / (Average assets)
Net debt issuance = (Total debt – Total lagged debt) / (Lagged total assets)
Sales = saleq
Operating costs = Sales – (Operating income)
Sales scaled by average assets = Sales / (Average assets)
Operating costs scaled by average assets = Sales / (Average assets)