How Widespread and Predictable is Stock Broker Misconduct?*

Craig McCann, PhD, CFA† Chuan Qin, PhD‡
Mike Yan, PhD, CFA, FRM§

Abstract

In this paper we reconcile widely diverging recent estimates of broker misconduct. Qureshi and Sokobin report that 1.3% of current and past brokers are associated with awards or settlements in excess of a threshold amount. Egan, Matvos, and Seru find that 7.8% of current and former brokers have financial misconduct disclosures including customer complaints, awards, and settlements. Qureshi and Sokobin arrive at their low estimate by excluding 85% of all brokers, including those brokers most likely to have engaged in misconduct. Applying Qureshi and Sokobins restrictive definition of potential misconduct to all brokers, we find that misconduct is much more widespread.

Qureshi and Sokobin find that the top quintile of brokers sorted by estimated likelihood of being subject of a customer complaint contain 55% of the brokers subsequently subject to complaints. We demonstrate that more sophisticated data mining techniques can sort brokers so that the highest risk quintile contains 75% of the brokers subsequently subject to complaints.

We also evaluate Qureshi and Sokobin's claim that FINRA's BrokerCheck website provides helpful information to investors seeking to avoid bad brokers and answer the question posed by Egan, Matvos, and Seru: If BrokerCheck data can predict broker misconduct, why don't investors use that data to protect themselves? We find that BrokerCheck is worthless in its current hobbled form, but could easily be modified so that investors could protect themselves and market forces would substantially reduce broker misconduct.

1 Introduction

FINRA is a self-regulatory organization tasked with policing registered representatives of brokerage firms ("brokers"). It maintains a database of investor complaints and disciplinary and employment

†President, Office: (703) 539-6760 or Email: CraigMcCann@slcg.com.
‡Senior Financial Economist, Office: (703) 539-6778 or Email: ChuanQin@slcg.com.
§Principal, Office: (703) 539-6780 or Email: MikeYan@slcg.com.

history for over 1,200,000 current and past brokers and publishes some of this information on its BrokerCheck website, www.brokercheck.finra.org.

Two recently published research reports have reported wildly different estimates of the extent of investment fraud perpetrated by brokers based on the BrokerCheck data. These reports also evaluated the potential of BrokerCheck data to predict future broker misconduct, thereby allowing investors to discern good brokers from bad brokers.

Qureshi and Sokobin [2015] (hereafter "Qureshi and Sokobin") analyze publicly and non-publicly available BrokerCheck data on 181,133 brokers and find that 2,349 of the brokers, or 1.3% of the total studied, had at least one customer complaint during the period from 2000 to 2013 which resulted in an award or settlement above a $10,000 threshold before May 18, 2009 and above a $15,000 threshold thereafter. Thus, it appears from Qureshi and Sokobin, that misconduct is rare in the brokerage industry.

Egan, Matvos, and Seru [2016] (hereafter "Egan, Matvos and Seru") using the BrokerCheck data covering a slightly different time period - 2005 to 2015 - found that that 7.8% of brokers have misconduct disclosures on their record and that brokers remain in the industry despite repeated misconduct. While Egan, Matvos and Seru use a more expansive definition of broker misconduct than Qureshi and Sokobin, this difference doesn't explain the dramatically different assessment of brokerage industry misconduct.

Qureshi and Sokobin only report the incidence of awards and settlements by 15% of brokers who were registered between 2000 and 2015. The 85% who were also registered during this period and which are excluded by Qureshi and Sokobin have a much higher incidence of customer awards or settlements than the brokers they included. Including all brokers increases the number of brokers with disclosed awards or settlements more than ten-fold, from 2,349 to 27,494.

Qureshi and Sokobin investigated whether BrokerCheck provides investors with information that can predict future investor harm and concluded that information available on BrokerCheck significantly predicts future meritorious investor claims:

> Overall, our results suggest that BrokerCheck provides valuable information to investors, thereby allowing them to discriminate between brokers with a high propensity for investor harm from other brokers. [p.4]

Egan, Matvos and Seru also find information on past broker misconduct can be used to predict future misconduct. They find stock brokers with recent customer complaints are more likely to be terminated by their employer, subsequently have longer spells of unemployment, and are thereafter re-employed at lower compensation and by less prestigious firms than brokers who do not have customer complaints. Egan, Matvos and Seru's results confirm that, rather than weeding bad brokers out of the industry, the regulatory environment and labor market sifts bad brokers down the quality ladder over time into brokerage firms with loose hiring practices and lax compliance ethics and that these bad brokerage firms specialize in preying on unsophisticated investors.[1]

[1] Dimmock et al. [2016] in a related study find that financial fraud is contagious. They find that a broker's

We use the same BrokerCheck data as Qureshi and Sokobin and Egan, Matvos and Seru to reassess whether BrokerCheck provides information to retail investors that helps them avoid bad brokers. We fit two regression models, a *probit* model and a *gradient boosting machine* model, to the BrokerCheck data and evaluate the models' predictive performance. Gradient boosting machine models generally achieve much better predictive performance than probit models, demonstrating the importance of selecting appropriate statistical models to make the most of the vast amount of BrokerCheck data.

We consider both characteristics of the individual broker (available from the BrokerCheck website, one broker at a time) and characteristics of those working with the broker at the same brokerage firm, (calculated using individual broker characteristics) as our models input variables. While the models using only individual broker characteristics have power to discriminate brokers with a high propensity for investor harm from others brokers, adding coworker characteristics significantly improves our models' predictive performance. As we explain below, the results of our analysis - and of the analyses performed by Qureshi and Sokobin and Egan, Matvos and Seru - do *not* support the conclusion that BrokerCheck provides *any* useful information to investors.

FINRA promotes a perception of BrokerCheck that is a classic example of the *fallacy of composition.* It suggests that since information on each of the 1.2 million brokers is accessible, the information on all 1.2 million brokers is accessible. FINRA could provide investors with the results of statistical modeling of all the BrokerCheck data on which Qureshi and Sokobin base their study. While this step would greatly enhance the usefulness of BrokerCheck, a much simpler solution is obvious: FINRA could simply make BrokerCheck information truly publicly available and allow ratings companies such as Lipper and Morningstar and news outlets like US News and World Report and BusinessWeek to rank brokerage firms on the risk of fraud.

The remainder of this paper is organized as follows. We first summarize and reconcile the estimates of potential brokerage misconduct reported in Qureshi and Sokobin (1.3%) and in Egan, Matvos and Seru (7.8%). Then we replicate the main Qureshi and Sokobin results on the predictability of investor harm using data subject to some limitations. We also apply a more sophisticated data mining technique, gradient boosting machine, to the BrokerCheck data and demonstrate that BrokerCheck data could be even more useful than suggested by Qureshi and Sokobin if the data were truly made publicly available. Next we apply similar techniques to predicting investor harm using the entire BrokerCheck data on 1.2 million brokers. We conclude with an explanation of why BrokerCheck data in its current form is not helpful to investors trying to protect themselves from bad brokers and how it could be dramatically improved at little or no cost.

propensity to commit financial fraud is significantly influenced by his or her coworkers' propensity to commit fraud after controlling for firm culture, branch atmosphere, market conditions and state regulatory environment.

2 How Widespread is Stock Broker Misconduct?

Brokers and investment advisers make recommendations and take orders. While the mix of activities varies from relationship to relationship, a broker or investment adviser can engage in misconduct that harms their customers. Some instances of broker misconduct are followed by customer complaints, arbitration filings or both. These customer complaints or arbitration filings may allege that unsuitable recommendations were made, important risks were not disclosed, accounts were churned or some other bad acts occurred. Broker financial misconduct might not directly involve a retail investor. For instance, a broker might have unsatisfied liens or personal bankruptcies which reflect on the broker's fitness to manage or guide other people's investments.

Additionally, some brokers are disciplined by regulators such as the Securities and Exchange Commission, FINRA, state securities regulators, and state Attorneys General. The conduct underlying these regulatory actions may have already been subject to customer complaints or arbitration filings but regulators have the ability to enforce larger systemic remedies on brokerage firms than investors can accomplish by filing individual complaints.

FINRA maintains a database of registration, employment, complaint and disciplinary history for each brokerage firm and broker, the Central Registration Depository, or CRD. The CRD includes each brokers involvement in customer disputes, financial, disciplinary and criminal events, employment history, and qualifications. FINRA makes a portion of the information in the CRD public through its BrokerCheck website.

2.1 Qureshi and Sokobin [2015]

Qureshi and Sokobin analyzed BrokerCheck information on 181,133 brokers registered with FINRA between 2000 and 2013 and found only 2,349, or 1.3%, of these brokers have been associated with an award or settlement above a threshold dollar amount.

Qureshi and Sokobin analyze the BrokerCheck data only for brokers first registered in 2000 or later. Thus, a broker first registered in 1998 and still active in 2008 is excluded by Qureshi and Sokobin. Of the 552,016 brokers with at least one current state registration in BrokerCheck on December 31, 2015, 57% or 314,652 were first registered in 2000 or later and 43% or 237,364 were first registered sometime earlier. Thus, Qureshi and Sokobin excluded nearly half of the currently registered brokers from their study.

Qureshi and Sokobin also only include brokers that were registered with four or more states for at least half of their careers because such brokers are asserted to be more likely to have client interactions. Of the 552,016 brokers with at least one current state registration in BrokerCheck as of December 31, 2015, only 53.6% are registered with four or more states. Combined, the two restrictions on the sample imposed by Qureshi and Sokobin limit their sample to only 15% of the 1.2 million brokers on BrokerCheck.

Qureshi and Sokobin define the initial filing of a grievance that subsequently results in an arbitration award in favor of the customer or in a settlement in excess of $10,000 prior to May

18, 2009 and in excess of $15,000 thereafter as an investor harm event. Their definition of an investor harm event assumes that settlements below these thresholds do not evidence any investor harm but are, in fact, entered into by brokerage firms to avoid further litigation costs in meritless cases. This assumption is overly restrictive. Low settlements and awards may reflect the low wealth of harmed investors not the lack of merit of their claims. Many meritorious claims are not brought because the potential recovery is too small and too uncertain to warrant investors and their attorneys expending the effort to prosecute a case or because the investor does not know about the opportunities for redress.

2.2 Egan, Matvos, and Seru [2016]

Egan, Matvos and Seru analyze BrokerCheck data for all 1.2 million brokers registered at some point in time between 2005 and 2015 without regard for when the broker was first registered and for how many states he or she has been registered with. They include as misconduct disclosures arbitration filings resulting in awards and lower settlements and other reported events as indicative of broker misconduct.

Egan, Matvos and Seru using BrokerCheck data from 2005 to 2015 find that 46,900 currently registered brokers have misconduct disclosures and nearly as many brokers no longer registered have disclosed misconduct compared to the 2,349 current and past brokers Qureshi and Sokobin find to have been associated with awards and settlements above their thresholds. Egan, Matvos and Seru also find that misconduct in the brokerage industry is persistent; in any given year 0.60% of active brokers report a misconduct disclosure in the current year and 7.8% have a misconduct disclosure at some point in their career.[2]

Egan, Matvos and Seru find that broker misconduct can be predicted by disclosures of the broker's past misconduct. They find 38% of brokers that reported financial misconduct in a given year had previous misconduct disclosures. That is, brokers who report misconduct are not drawn randomly from brokers with clean or checkered pasts. They document that brokers with a misconduct disclosure at some point in their career previously are approximately five times as likely to have a misconduct disclosure in the current year. They find recidivist rate in the first year after a misconduct disclosure is nearly 20 times the average rate of misconduct and remains more than five times the average rate five years after the most recent misconduct disclosure.

Egan, Matvos and Seru find that the labor market partially but not fully penalizes financial misconduct. Brokers with recent customer complaints are more likely to be terminated by their employer, have longer spells of unemployment, and are re-employed at lower compensation than brokers who do not have customer complaints. Their results confirm that, rather than weeding bad brokers out of the industry, the regulatory environment and labor market sifts bad brokers down

[2] Egan, Matvos and Seru group six CRD disclosures in FINRA classifications, Customer Dispute - Settled, Regulatory - Final, Employment Separation After Allegations, Customer Dispute - Award, and Civil - Final, as a priori "misconduct" disclosures.

the quality ladder over time into brokerage firms with loose hiring practices and compliance ethics. Supplementing the BrokerCheck data with Census Bureau data, they find the bad brokerage firms which accumulate these bad brokers specialize in preying on unsophisticated investors.

2.3 Investor Harm is Substantially More Widespread than Qureshi and Sokobin's Estimate

Qureshi and Sokobin report that only 1.3% of brokers had been associated with an award or a settlement in excess of the $10,000/$15,000 thresholds described above. Adopting Qureshi and Sokobin's definition of an investor harm event, but including the brokers excluded by Qureshi and Sokobin results in a much higher estimate of the prevalence of brokers with customer settlements and awards above the thresholds.

Table 1 reports the number of brokers, number of investor harm events and the number of brokers associated with investor harm events in the period of 2000-2014 by number of state registrations as of December 31, 2015. 27,494 current and former brokers have customer complaints resulting in settlements or awards meeting the definition of investor harm used by Qureshi and Sokobin between 2000 and 2014. 14,351 or 4.85% of brokers currently registered with 4 or more states have a disclosed investor harm event. Qureshi and Sokobin's conclusion that only 2,349 current and former brokers registered with at least 4 states during half of their career had investor harm events is only 8.5% (i.e. 2,349/27,494) of the brokers who meet their definition investor harm events in the BrokerCheck database.

Table 1: Investor Harm Events in 2000-2014 by State Registrations as of December 31, 2015

# of State Registrations	Brokers	Investor Harm Events	Brokers Associated with Investor Harm
0	648,657	19,464	10,676 (1.65%)
1	166,567	1,407	1,012 (0.61%)
2	57,893	1,055	767 (1.32%)
3	31,786	894	688 (2.16%)
>= 4	295,770	20,010	14,351 (4.85%)
Total	1,200,673	42,830	27,494 (2.29%)

The very low incidence of investor harm reported by Qureshi and Sokobin is primarily the result of their exclusion of brokers who were first registered before 2000. Qureshi and Sokobin excluded these brokers because the pre-2000 data available in electronic format was converted from a paper-based legacy system and may not be completely reliable. This might justify excluding the brokers first registered before 2000 from the regression analysis which includes length of time in the industry but does not justify excluding these brokers - 48.6% of all brokers active in the

2000-2014 period - from the estimation of how widespread broker misconduct was in 2000-2014. Since the likelihood a broker will have a disclosed settlement or award increases with time in the industry, Qureshi and Sokobin significantly understates the incidence of brokers with customer complaints by excluding brokers first registered before 2000.

Table 2 reports the number of brokers and the number of brokers associated with investor harm events in 2000-2014 by number of state registrations divided into brokers first registered in 1999 or earlier and those first registered in 2000 or later. Our dataset only includes the number of state registrations for each broker as of December 31, 2015, not the number of state registrations throughout a broker's career used by Qureshi and Sokobin so we can not exactly replicate their sample. Nonetheless, the left panel of Table 2 closely tracks Qureshi and Sokobin's results. They found that 1.3% of the 181,133 current and previously licensed brokers first registered after 1999 and with 4 or more state registrations for more than half their career were associated with a customer award or settlement above a dollar threshold. Consistent with their estimate, we find 1,943, or 1.35%, of the 144,178 brokers first registered after 1999 with 4 or more current state registrations have been associated with a customer award or settlement above Qureshi and Sokobin's threshold.

Table 2: Investor Harm Events in 2000-2014 by State Registrations as of December 31, 2015

	First Registered In or After 2000 (Included by Qureshi and Sokobin)		First Registered Before 2000 (Excluded by Qureshi and Sokobin)	
# of State Registrations	**Brokers**	**Brokers Associated with Investor Harm**	**Brokers**	**Brokers Associated with Investor Harm**
0	436,583	1,847 (0.42%)	212,074	8,829 (4.16%)
1	116,888	235 (0.20%)	49,679	777 (1.56%)
2	35,599	149 (0.42%)	22,294	618 (2.77%)
3	17,987	139 (0.77%)	13,799	549 (3.98%)
>= 4	144,178	1,943 (1.35%)	151,592	12,408 (8.19%)
Total	751,235	4,313 (0.57%)	449,438	23,181 (5.16%)

In total, 4,313 brokers (0.57%) of the 751,235 brokers first registered after 1999 have reported a settlement or award in excess of Qureshi and Sokobin's thresholds. On the other hand, 23,181 brokers (5.16%) of the 449,438 brokers first registered before 2000 have reported a claim in 2000 or later which resulted in an award or settlement above the threshold amounts. Brokers registered before 2000 are therefore nine times as likely to have reported a claim in the 2000 to 2014 period as brokers first registered after 1999.

The currently registered brokers with four or more state registrations are nearly evenly divided between 151,592 brokers first registered before January 1, 2000 and 144,178 brokers first registered after January 1, 2000. 12,408 (8.19%) of the brokers first registered before 2000 and currently registered with four or more state have at least one investor harm event compared to only 1,943

(1.35%) of the brokers first registered after 1999. Thus, the currently active brokers with four or more state registrations excluded by Qureshi and Sokobin are approximately six times as likely to have been associated with investor harm events as the currently active brokers Qureshi and Sokobin include (8.19% vs. 1.35%).

As Table 2 shows, the brokers excluded by Qureshi and Sokobin are between six and nine times as likely to have had a settlement or award in excess of their thresholds as those brokers included in their study, regardless of the number of state registrations the broker had as of December 31, 2015.

Figure 1 presents the numbers of active brokers in each year from 2000 to 2014 who were first registered before 2000 and those who were first registered in 2000 or later. The number of active brokers first registered in 2000 or later exceeded the number of those first registered before 2000 starting in 2008.

Figure 1: Number of Brokers by Year

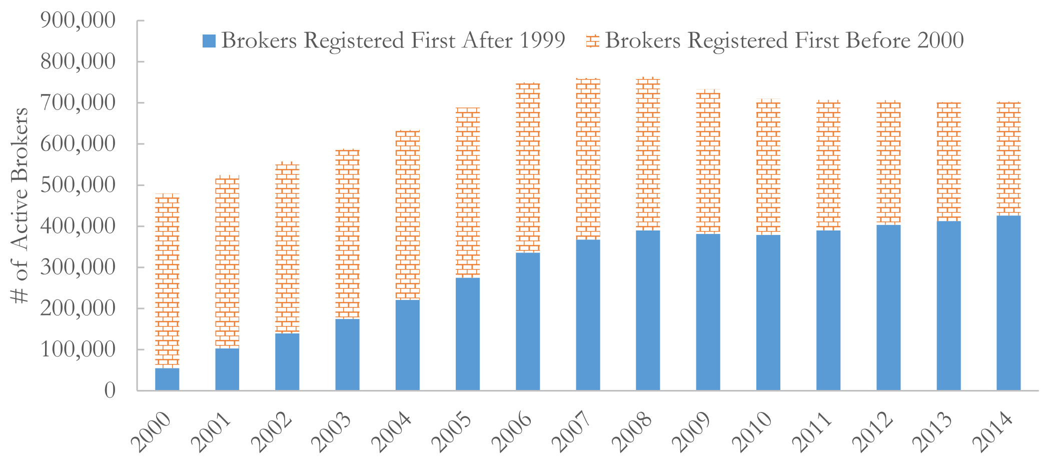
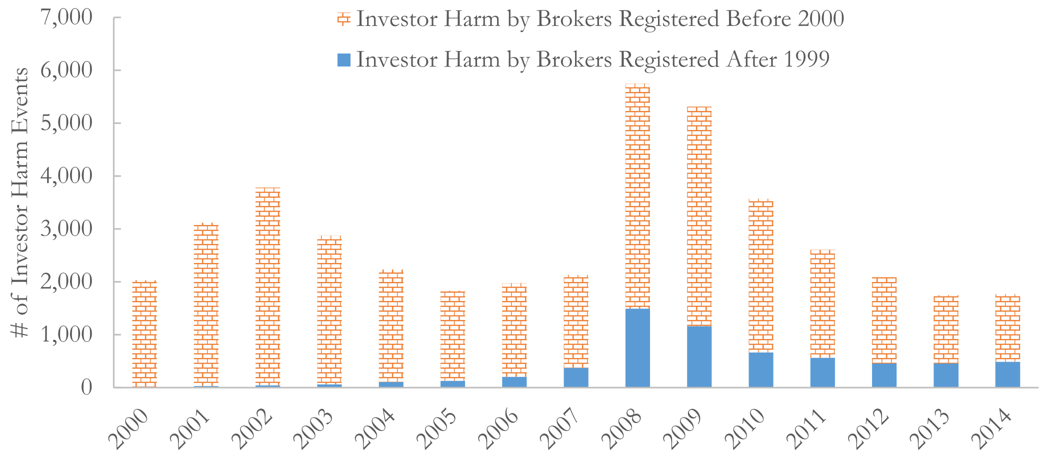


Figure 2 contrasts the numbers of investor harm events associated with brokers in these two groups by year. Many fewer of the investor harm events over the entire period of 2000-2014 were attributable to brokers first registered in 2000 or later than were associated with brokers first registered before 2000. This is not simply a result of brokers first registered earlier having a longer time period to accumulate customer complaints. Even though the brokers registered in 2008 are roughly equally divided between those first registered before 2000 and those first registered in 2000 or later, the brokers registered first in the earlier time account for 75% of the investor harm events in 2008.

Figure 2: Number of Investor Harm Events by Year



3 Can Stock Broker Misconduct Be Predicted?

In this section, we replicate the Qureshi and Sokobin results on the predictability of investor harm subject to some data limitations. Qureshi and Sokobin use data on bankruptcies within 10 years prior to each year's observances of broker misconduct. Since personal bankruptcies are only available on BrokerCheck for 10 years, we can't observe bankruptcies prior to 2005 in the data we downloaded as of December 31, 2015. Also, Qureshi and Sokobin limit their sample to brokers registered in four or more states for at least half their career. The BrokerCheck website only lists current state registrations so we approximate this filter by requiring brokers to be currently registered in four or more states. As a result, we exclude brokers who are no longer in the industry from our analysis in this section of the paper. Also, Qureshi and Sokobin use the broker's gender as an explanatory variable even though it is not one of the data items available on BrokerCheck. As we show below, these and other data limitations do not seem to matter; we are able to closely replicate Qureshi and Sokobin's results.

3.1 Data and Model Features

We aggregate the characteristics (disclosures, employment, qualifications, etc.) of each broker in each calendar year of the broker's tenure over the 2000-2014 period. This gives rise to an annual panel of broker-year observations spanning 15 years. We do not include information customer claims filed in 2015 and 2016 because the outcome-based-filter requires enough time to pass after the filing of an arbitration claim to observe whether it resulted in an award or settlement in excess of their thresholds. For a disclosure event associated with more than one date, we assign the event with the earliest date when the underlying event was reported. For example, if a customer complaint was filed in August 2013 and arbitrated or settled in January 2015, we add the complaint to the 2013 panel. Thus, our sample is an annual panel dataset including the complete career history through 2014 of all brokers who first registered with FINRA in 2000 or later and were registered with four or more states on December 31, 2015. There are 144,178 unique brokers and 1,163,927 broker-year observations in our sample.

We predict the occurrence of investor harm events in each year by building statistical models on BrokerCheck historical data. The model inputs, called *features*, are constructed from the raw data to reflect the characteristics of each broker and the brokerage firms where they were employed in each year. We expect these features to contain useful information for predicting investor harm.

Customer complaints filed against a broker may result in a settlement or award, or remain unresolved, or they may be denied. Following Qureshi and Sokobin, we assume that arbitration filings that fail to lead to an award or to a settlement above $10,000 before May 18, 2009 and above $15,000 thereafter do not reflect investor harm or broker misconduct. The time when the investor harm event occurs is approximated by the arbitration filing year. We associate each year in a broker's career with an indicator variable which equals 1 if the broker discloses an investor harm event that year and 0 otherwise.

Table 3 summarizes the distributions of investor harm events (that is, settlements and awards satisfying the $10,000/$15,000 threshold) and general customer dispute disclosures regardless of outcome during the 2000-2014 period. A small portion of brokers (1.35%) in our sample are associated with investor harm events as defined in Qureshi and Sokobin, while the percentage of brokers that were ever involved in any type of customer dispute events is more than three times as high (4.42%). The majority of the brokers associated with either investor harm or general customer disputes only had one disclosure event.

Table 3: Summary of Settlements and Awards and Customer Disputes

Settlements and Awards (SA)	**Brokers with Settlements and Awards**	**Customer Disputes (CD)**	**Brokers with Customer Disputes**
1	1,627	1	5,040
2	226	2	886
3 or more	90	3 or more	448
Total	1,943	Total	6,374

We are interested in predicting the occurrence of investor harm events associated with a broker in a given year using BrokerCheck information prior to that year. The disclosure history of a broker will have significant power to predict future investor harm caused by the broker if past offenders are more likely to commit similar offenses in the future. The six broker disclosure features and five qualifications and employment features listed in Table 4 are compiled from the BrokerCheck data in each year of each broker's career.

Table 4: Description of Broker Features

Feature	Symbol	Description of Feature
Customer Dispute - Settlements and Awards	SA	Number of costumer disputes that led to an award or to a settlement against the broker above $10,000/$15,000
All Customer Disputes	CD	Number of costumer disputes against the broker regardless of outcome
Disclosed Judgements and Liens	JUDG	Number of unsatisfied judgments and liens[3]
Disclosed Bankruptcies	BKCY	Number of bankruptcies and compromises currently disclosed
Disclosed Disciplinary Events	DPRY	Number of regulatory actions, investigations, civil judicial actions, and employment terminations available on BrokerCheck
Criminal Events	CRIM	Number of criminal disclosures
Exams Passed	EXAM	Number of exams (S6, S7, S63, S66) passed
Association with Expelled Firm	EXPEL	An indicator that equals 1 if the broker had been registered with any firm that has been expelled from the industry
Prior Employers	PREMPL	Number of firms the broker worked at and separated from
Employment Years	EMPLYR	Number of years registered as a broker
Dual Registration	DUAL	An indicator that equals 1 for brokers registered with SEC as an investment adviser

The previous studies (Dimmock et al. [2016], Egan et al. [2016], and Qureshi and Sokobin [2015]) found that firm culture and coworker misconduct influenced on the amount of fraud committed by individual brokers. For each year in a broker's career, we construct the nine coworker characteristics listed in Table 5 which describe the disclosure and employment profile of the broker's coworkers.

Table 5: Description of Coworker Features

Feature	Symbol	Description of Feature
Harm Associated with Coworkers	HAC	Average number of investor harm events per coworker-year for all coworkers, averaged over all firms the broker was employed by during the year
Customer Disputes Associated with Coworkers	CDAC	Average number of customer disputes per coworker-year for all coworkers, averaged over all firms the broker was employed by during the year
Average Number of Coworkers	CONUM	Average number of coworkers, averaged over all firms the broker was employed by during the year

[3] Satisfied judgments and liens events and Bankruptcies more than 10 years old are not disclosed on BrokerCheck.

Coworker Disclosed Judgements and Liens	COJUDG	Average number of unsatisfied judgments and liens per coworker-year, averaged over all firms the broker was employed by during the year
Coworker Disclosed Disciplinary Events	CODPRY	Average number of disclosed disciplinary events per coworker-year, averaged over all firms the broker was employed by during the year
Coworker Criminal Events	COCRIM	Average number of criminal disclosures per coworker-year, averaged over all firms the broker was employed by during the year
Coworker Affiliation with Expelled Firm	COEXPEL	Percentage of coworkers that were previously registered with an expelled firm, averaged over all firms the broker was employed by during the year
Average Number of Prior Employers for Coworkers	COPREMPL	Average number of prior employers per coworker for all coworkers, averaged across all firms the broker was employed by during the year
Average Number of Employment Years for Coworkers	COEMPLYR	Average number of years registered with FINRA per coworker for all coworkers, averaged across all firms the broker was employed by during the year

Table 6 presents the average values of broker features in Table 4 and coworker features in Table 5 for broker-years subsequently associated with investor harm events and for broker-years not subsequently associated with investor harm events. The extremely small p-values for the two-sample t-tests suggest that brokers associated with investor harm have a higher average number of past customer complaints that led to an award or settlement above the threshold, general customer disputes, judgments and liens, disciplinary events, and criminal events. The brokers associated with investor harm on average passed more exams and had more previous employers and a longer registration history.[4] Brokers associated with investor harm events are also more likely to be SEC-registered investment advisers and are more likely to have been previously affiliated with an expelled firm. The only feature that does not appear to be significantly different across the two subsets of brokers is the incidence of prior personal bankruptcies.

Table 6: Differences in Broker and Coworker Features

[4] These differences in means across the subsets of brokers are interesting but should not be overly interpreted. For example, brokers associated with investor harm events have passed more exams on average than brokers who are not associated with investor harm. This difference is statistically significant due to the very large sample size but it may not be meaningful. The brokers associated with investor harm on average have passed only 5.5% more exams than other brokers but 18.6 times as many have been associated with prior settlements and awards. Also, as we show below in the regression analysis, after controlling for other differences across brokers, differences in the number of exams passed provides no useful information for predicting which brokers will be associated with an investor harm event.

Feature	**All Brokers**	**Brokers Associated with Investor Harm**	**Brokers Not Associated with Investor Harm**	**Difference**	***p*-value**
Individual Broker Features:					
SA	0.0100	0.1800	0.0097	0.1703	0.0000***
CD	0.0369	0.4308	0.0360	0.3948	0.0000***
JUDG	0.0080	0.0415	0.0079	0.0336	0.0003***
BKCY	0.0223	0.0310	0.0222	0.0088	0.1939
DPRY	0.0060	0.0244	0.0060	0.0184	0.0000***
CRIM	0.0226	0.0449	0.0226	0.0223	0.0002***
EXAM	2.0854	2.1996	2.0852	0.1144	0.0000***
EXPEL	0.0083	0.0530	0.0082	0.0448	0.0000***
PREMPL	0.8742	1.4255	0.8731	0.5524	0.0000***
EMPLYR	5.4154	7.1533	5.4119	1.7414	0.0000***
DUAL	0.6865	0.8434	0.6862	0.1572	0.0000***
Coworker Features:					
HAC	0.0012	0.0031	0.0012	0.0019	0.0000***
CDAC	0.0046	0.0106	0.0046	0.0060	0.0000***
CONUM	2359.40	2262.29	2359.60	-97.31	0.0824
COJUDG	0.0018	0.0039	0.0018	0.0021	0.0000***
CODPRY	0.0012	0.0020	0.0012	0.0008	0.0000***
COCRIM	0.0071	0.0081	0.0071	0.0010	0.0059**
COEXPEL	0.0089	0.0413	0.0088	0.0325	0.0000***
COPREMPL	0.9544	1.1910	0.9540	0.2370	0.0000***
COEMPLYR	5.3717	5.5721	5.3713	0.2008	0.0000***

*Note: The symbols ***, ** and * denote significance at the 0.001, 0.01 and 0.05 level, respectively.*

Table 6 also presents average values of coworker features for brokers associated with investor harm and for brokers not associated with investor harm in the current year. The extremely small p-values reflect the significant difference in the disclosure and qualification history between the coworkers of the brokers associated with investor harm and the coworkers of those without such association. Brokers associated with investor harm are more likely to have coworkers who had more past settlements and awards above the thresholds, all customer disputes, judgements and liens, disciplinary events, criminal charges, and prior affiliations with expelled firms than brokers not associated with investor harm. The coworkers of brokers associated with investor harm also tend to have more previous employers and a longer employment history than the coworkers of brokers not associated with investor harm. All the coworker features seem to be significantly different between the two groups of brokers except for the average number of coworkers. This indicates that the coworker features listed in Table 5 may contain valuable information for distinguishing brokers who are more likely to be associated with future investor harm events from other brokers and therefore should be incorporated in statistical models that predict investor harm.

3.2 Probit Models

We use the following probit regression model to estimate the propensity of any given broker to cause investor harm:

$$\mathrm{P}(Y_{i,t} = 1|X_{i,t-1}) = \Phi(\beta_0 + \beta_1 X_{i,t-1}), \quad i = 1, 2, \ldots, N, \quad t = 1, 2, \ldots, n_i,$$

where $P(Y_{i,t} = 1|X_{i,t-1})$ denotes the probability of investor harm in year t given $X_{i,t-1}$, a vector of model features containing BrokerCheck information prior to year t, and Φ is the cumulative distribution function ("CDF") of a standard normal random variable. The dependent variable $Y_{i,t}$ equals 1 if at least one investor harm event caused by broker i occurred in year t. The model relates the probability of broker i causing investor harm in each year t of broker i's career to a linear combination of model features via the standard normal distribution function. Since the broker features in each year are compared to the occurrence of investor harm in the subsequent year, the brokers first registered in 2014 are effectively excluded from the sample. The number of broker-year observations used in the regression is 1,020,707, corresponding to 133,556 unique brokers. After obtaining the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can calculate the predicted probability of broker j causing investor harm in year t (denoted by $\hat{Y}_{j,t}$) as the following

$$\hat{Y}_{j,t} = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_{j,t-1}) \tag{1}$$

We explore alternative specifications by fitting three probit regression models on different sets of features to assess the importance of each feature in predicting investor harm. Our reproduction of Qureshi and Sokobin's baseline model includes eleven features: prior settlements and awards, disclosed judgments and liens, bankruptcies within ten years, disciplinary and criminal events, exams passed, associations with expelled firms, dual registration and the previous year's annual index return on the S&P 500 index (denoted by MKRN)[5]. Our second model adds harm associated with coworkers as well as coworkers' disclosed judgements and liens, and disciplinary and criminal events. We apply an exponent of 1/3 to the coworker disclosure features to better fit the data. We obtain the third model by replacing awards and settlements above the threshold (SA) with all customer disputes (CD) and coworker awards and settlements above the threshold (HAC) with all coworker customer disputes (CDAC).

Table 7 summarizes the estimated model coefficients (β_1) with corresponding Z-scores and p-values. As explained above our sample differs slightly from that in Qureshi and Sokobin but our reproduction's parameter estimates are quite similar. The main difference in our baseline models is that Qureshi and Sokobin included gender, an important feature, which is not available on BrokerCheck.

[5] The annual market return acts as a control for the macroeconomic conditions in the year under consideration.

Table 7: Probit Regression Model Results

Feature	Qureshi and Sokobin	Baseline Reproduction	Add Coworker Features	Replace SA with CD
SA	0.332***	0.3550***	0.3126***	
	(8.936)	(17.593)	(15.400)	
CD				0.2422***
				(21.076)
JUDG	0.112***	0.1114***	0.0657**	0.0588*
	(5.748)	(5.310)	(2.757)	(2.377)
BKCY	0.0226**	0.0124	0.0178	0.0175
	(2.324)	(0.453)	(0.647)	(0.633)
DPRY	0.230***	0.1690***	0.1181**	0.0558
	(5.038)	(3.801)	(2.597)	(1.157)
CRIM	0.170***	0.1371***	0.1073***	0.0919**
	(6.240)	(4.427)	(3.302)	(2.733)
EXAM	0.00663	0.0040	0.0146	0.0058
	(0.507)	(0.314)	(1.122)	(0.444)
EXPEL	0.432***	0.5578***	0.1804***	0.1402**
	(6.779)	(13.286)	(3.394)	(2.581)
PREMPL	0.0861***	0.0253***	0.0007	-0.0017
	(15.91)	(4.599)	(0.118)	(-0.281)
EMPLYR	0.0222***	0.0424***	0.0587***	0.0559***
	(16.59)	(18.086)	(20.238)	(18.936)
DUAL	0.279***	0.2911***	0.3186***	0.2929 ***
	(18.81)	(14.933)	(14.810)	(13.294)
GENDER	0.292***			
	(14.28)			
MKRN	-0.514***	-0.6214***	-0.4489***	-0.4598***
	(-14.84)	(-17.790)	(-12.070)	(-12.404)
CONUM			-0.00003***	-0.00002***
			(-8.655)	(-6.993)
COEXPEL			0.5013***	0.2734*
			(4.273)	(2.239)
COPREMPL			0.0750***	0.0774***
			(4.282)	(4.307)
COEMPLYR			-0.0903***	-0.0950***
			(-13.793)	(-14.206)
$HAC^{1/3}$			2.548***	

			(17.935)	
$CDAC^{1/3}$				2.333***
				(19.064)
$COJUDG^{1/3}$			0.3530**	0.2783*
			(2.871)	(2.212)
$CODPRY^{1/3}$			0.4239**	0.1713
			(3.011)	(1.194)
$COCRIM^{1/3}$			-0.0684	-0.2634*
			(-0.583)	(-2.174)
Model Chi-Square	2303.9***	1769.74***	2498.80***	2796.41***
Observations	1,014,873	1,020,707	1,020,707	1,020,707

*Note: The symbols ***, ** and * denote significance levels at 0.001, 0.01 and 0.05, respectively.*

In each model, most features are statistically significant. The broker and coworker past disclosures (namely, SA, CD, HAC, and CDAC) achieve high Z-scores relative to other features in their respective models. The signs and significance levels of the coefficients shed light on the value of information contained in each model feature. For example, since the coefficient for SA is positive and highly statistically significant, the past settlement and award information is useful in predicting future investor harm. Also the coefficients are economically significant. For example, in the baseline reproduction, the impact of an additional previous expelled firm record on the predicted probability is equivalent to the impact of 1.57 additional settlements and award, or 3.30 additional disciplinary events, or 4.07 additional criminal records, or 5.01 additional disclosed judgements and liens.

The log-likelihood chi-square statistics indicate that the model incorporating coworker features and using all customer disputes and not just settlements and awards is the most statistically significant model. This result has important implications. Even if all you are trying to predict is filings that result in settlements and awards in excess of the threshold, the best set of explanatory features includes all customer complaints. Thus, as an empirical matter, cases that are dismissed or which result in settlements below the threshold contain important information for differentiating good brokers from bad brokers.

We use the estimated coefficients to calculate the predicted probability of investor harm for each year in each broker's career. Following Qureshi and Sokobin, we rank all broker-year observations and allocate these observations into quintiles according to their predicted probabilities. In our baseline reproduction, 51.29% of the investor harm events are associated with the broker-year observations allocated to the highest risk quintile, and only 3.72% of the investor harm events are attributed to the broker-year observations allocated to the lowest risk quintile. Qureshi and Sokobin found 55.5% of investor harm events were by brokers in the highest risk quartile and 3.8%

in the lowest risk quartile.[6] Thus, as with the incidence of harm found, we are able to replicate the Qureshi and Sokobin regression results. If the model had no predictive ability, roughly 20% of the investor harm events should have been attributed to the broker-year observations in each of the lowest risk quintile and the highest risk quintile. Hence our findings show that the probit model is effective in discriminating brokers associated with investor harm from those without such association, and that the BrokerCheck information is useful for predicting investor harm *if all the data is available and subjected to sophisticated statistical analysis.*

Continuing to replicate Qureshi and Sokobin, we compare the predicted probability of each broker causing investor harm in a given year to the unconditional probability of investor harm, defined as the ratio of the number of broker-year observations associated with investor harm to the total number broker-year observations. The number of broker-year observations associated with investor harm is 2,094, and the unconditional probability of investor harm equals $2,094/1,020,707 = 0.21\%$. We predict broker j to cause investor harm in year t if and only if the predicted probability $\hat{Y}_{j,t}$ is greater than 0.21%. Using this criterion, 1,458 (69.63%) of the 2,094 broker-year observations associated with investor harm are correctly predicted to have investor harm, while 337,475 (33.13%) of the 1,018,613 broker-year observations without investor harm are erroneously predicted to have investor harm. The baseline probit model at cutoff of 0.21% has a *true positive rate* of 69.63% and a *false positive rate* of 33.13%. This provides further evidence that the baseline model is effective in predicting investor harm.

Table 8 reports true positive rates, false positive rates, and distribution of investor harm events among quintiles of broker-year observations for each model. The best probit model sorts brokers so that the highest risk quintile captures 60.17% of investor harm events and the lowest risk quintile is associated with 1.77% of the investor harm events.

Table 8: Within-Sample Predictive Performance of Probit Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	69.63%	33.13%	3.72%	8.17%	11.99%	24.83%	51.29%
Add Coworker Features	71.30%	31.27%	2.34%	5.21%	13.80%	20.15%	58.50%
Replace SA with CD	72.87%	31.10%	1.77%	5.54%	11.13%	21.39%	60.17%

The within-sample measures might overstate the predictive performance since a model that fits one dataset well may not necessarily predict well when applied to a different data set (called "over-

[6] The difference in our results from the baseline model reproduction and Qureshi and Sokobin's results appears to be two-fold. First, they have access to the brokers' gender which is not available in the data on BrokerCheck and on which we rely. Second, the sample we consider is slightly different from their sample because they have access to each broker's state registration continuously throughout their careers while we can only observe state registrations as of December 31, 2015.

fitting"). To gauge the true predictive efficacy of the models we perform 5-fold cross-validations on the entire data. We randomly partition all the broker-year observations into five groups. Each group is left out and a probit model is fit to the remaining groups combined. The estimated model is then used to predict the occurrence of investor harm in the held-out group. The prediction accuracy measures for all five models are averaged to generate the accuracy measure for one cross-validation procedure. We repeat the procedure ten times and average the resulting ten groups of accuracy measures to get the final measures, which are reported in Table 9.

Table 9: Cross-Validated Predictive Performance of Probit Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	69.18%	32.90%	3.72%	7.98%	12.41%	25.41%	50.48%
Add Coworker Features	71.14%	31.21%	2.43%	5.27%	13.76%	20.45%	58.09%
Replace SA with CD	72.58%	31.09%	1.83%	5.65%	11.21%	21.31%	60.00%

The true positive rates and the numbers of investor harm events captured in the highest quintile in Table 9 are, at worst, only slightly smaller than those in Table 8 and follow a similar pattern across the models. The models including both broker and coworker features assign more than 58% of the investor harm events in the hold-out samples to the highest quintile and less than 2.5% of investor harm events to the lowest quintile, which confirms that the BrokerCheck database contains valuable information for discriminating potentially harmful brokers from non-harmful ones. The improvement in predictive performance from Qureshi and Sokobin's baseline model indicates that coworker features play a significant role in predicting investor harm. The improved performance of the model with all customer disputes suggests that not only the brokers disputes leading to award or settlement above a threshold amount, but also those pending, denied, or closed without action are useful in determining the likelihood of future investor harm.

3.3 Gradient Boosting Machine Models

The probit model combines features in a linear fashion and may have poor predictive performance on data with complex nonlinear structures. To predict investor harm more effectively we consider a more sophisticated statistical model, the *gradient boosting machine* (GBM) model. The GBM combines the outputs of many "weak" predictive models (a model that predicts marginally better than random) to produce an ensemble model with superior predictive performance. The algorithm seeks to minimize the expected value of a *loss function* in an iterative fashion where a weak model is fit to the data-based negative gradient vector of the loss function in each iteration. The GBM procedure has many desirable properties as an effective "off the shelf" tool for data mining, including immunity to outliers in the response variable and robustness against the inclusion of irrelevant input variables (see Friedman [2001]).

A GBM model has three tuning parameters: the number of terminal nodes in each regression tree (i.e., the weak model), a shrinkage parameter that prevents overfitting, and the number of boosting iterations. In our application of GBM, we set these parameters to be 4, 0.05, and 500, respectively[7]. The dependent variable and input feature vectors are the same as those defined above. We build three GBM models using the same sets of features as used in the probit models above except that the coworker disclosure features are directly input to the models without being taken to 1/3 power[8].

Analogous to probit regressions, the GBM model outputs a predicted probability of subsequent investor harm for each broker-year observation. The predicted probabilities can then be used to classify investor harm events and to sort broker-year observations into quintiles. Table 10 reports the within-sample predictive measures for each model. Compared to their probit counterparts, each model has a lower false positive rates and captures a higher proportion of investor harm events in the top quintile of broker-year observations. Similar to probit models, the proportions of investor harm events allocated to the top quintile by the three GBM models increase when adding coworker features and replacing settlements and awards with general customer disputes.

Table 10: Within-Sample Predictive Performance of GBM Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	72.45%	26.45%	1.91%	5.83%	9.26%	16.86%	66.14%
Add Coworker Features	72.21%	21.32%	1.77%	4.35%	7.69%	15.57%	70.63%
Replace SA with CD	72.06%	19.61%	1.72%	3.96%	6.97%	14.80%	72.54%

The out-of-sample predictive measures of the models obtained from 5-fold cross-validation procedures are reported in Table 11. The distributions of investor harm events over the five quintiles are not much different from those in Table 10, indicating that the GBM models do not seriously overfit the data and the prediction results in Table 10 are robust. By comparing the out-of-sample predictive measures in Table 8 and Table 10, we conclude that the GBM model specifications do a much better job than the probit models at differentiating high risk brokers from low risk brokers.

[7] The choice of tuning parameters may have a significant impact on the predictive performance of GBM models. Here we experiment with a range of tuning parameter combinations and select the one that gives optimal prediction results on hold-out samples in cross-validation procedures.

[8] A remarkable strength of the GBM model is that there is usually little need to pre-process the input features due to the model's robustness to monotone transformations of individual features.

Table 11: Cross-Validated Predictive Performance of GBM Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	70.79%	25.38%	2.24%	5.57%	10.12%	16.90%	65.17%
Add Coworker Features	70.85%	21.70%	2.06%	4.86%	8.02%	16.08%	69.87%
Replace SA with CD	71.07%	19.09%	1.97%	4.46%	6.48%	15.10%	72.00%

While it is impossible to evaluate the strength or relevance of features in a GBM model by Z-scores and p-values, the model has its own built-in measurements of feature importance. The model can return a "relative importance score" for each feature, which measures the feature's relative contribution to the overall predictive power of the model. We report the importance scores for the features used in three GBM models in Table 12. The order of feature importance scores for the GBM is similar to the order of absolute values of Z-scores for the probit model including the same features (Table 7), which is possibly due to the additive nature of the GBM algorithm[9]. The predictive ability of GBM models largely stems from the broker and coworker history of customer disputes. Broker settlements and awards (SA) and customer disputes (CD) achieve the highest scores in their respective models. All the seven coworker features in the second and third models are influential to some extent, with coworker settlements and awards (HAC) and coworker disputes (CDAC) ranked the second most important. The broker features JUDG, BKCY, DPRY, EXAM, and EXPEL appear to be unimportant compared to the settlements and awards and coworker disputes features.

Table 12: Relative Importance Scores for GBM Models

Feature	**Baseline Model**	**Add Coworker Features**	**Replace SA with CD**
SA	35.60	19.82	
CD			22.07
JUDG	5.36	1.30	0.88
BKCY	0.24	0.57	0.02
DPRY	2.16	0.17	0.25
CRIM	2.99	1.25	1.71
EXAM	1.85	0.47	0.11
EXPEL	6.43	0.44	0.21
PREMPL	9.83	2.70	2.70

[9] The weak regression tree models are combined in an additive fashion to produce the final ensemble model. See Friedman [2001] for more details.

EMPLYR	15.05	4.92	5.20
DUAL	5.72	0.85	0.72
MKRN	14.78	4.17	3.36
CONUM		5.24	4.92
COEXPEL		8.53	6.62
COPREMPL		5.58	9.89
COEMPLYR		10.23	10.74
HAC		17.40	
CDAC			15.50
COJUDG		6.63	6.25
CODPRY		3.85	4.90
COCRIM		5.89	3.95

4 The Full BrokerCheck Data

The BrokerCheck data set analyzed by Qureshi and Sokobin excludes brokers first registered before 2000. These brokers were associated with 75% of all investor harm events between 2000 and 2014 (see p.8). Qureshi and Sokobin also restrict the sample only to "public-facing" brokers, defined as those registered with more than 3 states in at least half of their career. This rule of thumb for identifying brokers with public dealings[10] leaves out public-dealing brokers that were registered with fewer states. In this section we drop the above restrictions and perform probit and GBM regressions on the entire BrokerCheck database.

4.1 Data and Model Features

The data considered hereafter is an annual panel consisting of the career history of all the brokers in our sample between 2001 and 2014.[11] This includes 10,009,600 broker-year observations and 1,200,673 unique brokers, 27,494 (2.29%) of which are associated with investor harm events. Table 13 reports the distributions of investor harm events and all customer disputes associated with these brokers between 2000 and 2014.

[10] See footnote 22 of Qureshi and Sokobin [2015].
[11] The brokers with no registration history in 2001-2014 are not included.

Table 13: Summary of Settlements and Awards and Customer Disputes

Settlements and Awards (SA)	Brokers with Settlements and Awards	Customer Disputes (CD)	Brokers with Customer Disputes
1	20,963	1	45,458
2	3,909	2	11,897
3	1,274	3	4,218
4 or more	1,348	4 or more	4,317
Total	27,494	Total	65,890

For each broker we construct broker and coworker features for each year in the broker's tenure between 2000 and 2013, which are used to predict the propensity of investor harm in the subsequent year. Hence the investor harm events that occurred in 2000 and the brokers first registered in 2014 are not used in the regressions. The inputs to the regression models are 9,306,361 broker-year observations, corresponding to 1,153,362 unique brokers. The number of broker-year observations associated with investor harm is 31,280 (0.34%), attributed to 25,245 (2.19%) brokers.

Table 14 presents the averages of broker features (Table 4) and coworker features (Table 5) for broker-years subsequently associated with investor harm events and for broker-years not subsequently associated with investor harm events. While the average feature values are higher than those in Table 4 and Table 5, the signs of differences in feature averages between the two groups remain the same with the exception of only CONUM.

Table 14: Differences in Broker and Coworker Features

Feature	All Brokers	Brokers Associated with Investor Harm	Brokers Not Associated with Investor Harm	Difference	p-value
Individual Broker Features:					
SA	0.0246	0.4663	0.0231	0.4432	0.0000***
CD	0.0733	0.9138	0.0705	0.8433	0.0000***
JUDG	0.0147	0.0426	0.0146	0.0280	0.0000***
BKCY	0.0156	0.0201	0.0156	0.0045	0.0033**
DPRY	0.0294	0.1722	0.0289	0.1433	0.0000***
CRIM	0.0246	0.0518	0.0245	0.0273	0.0000***
EXAM	1.9504	2.0984	1.9499	0.1485	0.0000***
EXPEL	0.0238	0.0788	0.0237	0.0551	0.0000***
PREMPL	1.7973	3.0476	1.7931	1.2545	0.0000***
EMPLYR	11.146	16.565	11.128	5.4370	0.0000***
DUAL	0.4330	0.7371	0.4320	0.3051	0.0000***
Coworker Features:					
HAC	0.0015	0.0035	0.0015	0.0020	0.0000***

CDAC	0.0049	0.0097	0.0049	0.0048	0.0000***
CONUM	7984.1	9434.1	7979.3	1454.8	0.0000***
COJUDG	0.0022	0.0026	0.0022	0.0004	0.0000***
CODPRY	0.0024	0.0041	0.0024	0.0017	0.0000***
COCRIM	0.0055	0.0065	0.0055	0.0010	0.0000***
COEXPEL	0.0251	0.0614	0.0250	0.0364	0.0000***
COPREMPL	1.8886	2.3572	1.8870	0.4702	0.0000***
COEMPLYR	11.0672	12.3082	11.0630	1.2452	0.0000***

*Note: The symbols ***, ** and * denote significance levels at 0.001, 0.01 and 0.05, respectively.*

4.2 Probit Models

We fit three probit regression models to the full BrokerCheck data: the baseline model, the “add coworker features” model, and the “add coworker features with SA and HAC replaced by CD and CDAC” model. All the features are defined the same as before except that the five broker disclosure features (namely, HAC, CD, JUDG, DPRY, and CRIM) are raised to a power of 1/3. We find that performing this power transformation on these features not only improves model fitting but also resolves the issue of *quasi-complete separation*[12], which would arise if disclosure counts were directly input as model features.

The regression coefficients and chi-square statistic of the probit models are given in Table 15. All broker features but $BKCY^{1/3}$ are highly statistically significant, with $SA^{1/3}$ and $CD^{1/3}$ attaining the highest Z-scores in their respective models. Compared to the Z-scores for similar models built on the restrictive data (Table 7), the distinctively high Z-scores of $SA^{1/3}$ and $CD^{1/3}$ in Table15 indicate that the information about past customer complaints has a larger influence on the probit model built on the full data. The coworker disclosure features $HAC^{1/3}$ and $CDAC^{1/3}$ are also seen to be significant, obtaining the fourth highest Z-scores in the second model and third model, respectively. This confirms the significant impact of coworkers’ prior disclosure record on the model’s predictive power.

Table 15: Probit Regression Model Results

Feature	**Baseline Model**	**Add Coworker Features**	**Replace SA with CD**
$SA^{1/3}$	0.7220***	0.6540***	
	(128.603)	(114.180)	
$CD^{1/3}$			0.5717***

[12] Quasi-complete separation in a two-class probit regression occurs when there exists a hyperplane in the feature space that almost perfectly separates the observations in one class from those in the other. The estimated coefficients of a probit model suffering from quasi-complete separation tend to have unusually large standard errors, and hence are spurious.

			(132.309)
$JUDG^{1/3}$	0.1666***	0.1436***	0.1201***
	(12.949)	(10.837)	(9.006)
$BKCY^{1/3}$	-0.0165	-0.0002	-0.0237
	(-0.964)	(-0.013)	(-1.343)
$DPRY^{1/3}$	0.2733***	0.2241***	0.2062***
	(37.837)	(29.932)	(27.547)
$CRIM^{1/3}$	0.1674***	0.1335***	0.1141***
	(16.233)	(12.638)	(10.701)
EXAM	0.0475***	0.0566***	0.0398***
	(15.129)	(17.405)	(12.048)
EXPEL	0.2784***	0.0867***	0.0750***
	(30.863)	(8.245)	(7.068)
PREMPL	0.0142***	0.0082***	0.0081***
	(16.368)	(8.621)	(8.497)
EMPLYR	0.0112***	0.0125***	0.0104***
	(46.955)	(48.450)	(39.272)
DUAL	0.3570***	0.3232***	0.2904***
	(77.466)	(65.566)	(57.577)
MKRN	-0.5898***	-0.5589***	-0.5898***
	(-59.857)	(-55.213)	(-57.984)
CONUM		0.0000***	0.0000
		(5.838)	(1.638)
COEXPEL		0.2494***	0.3101***
		(8.013)	(10.079)
COPREMPL		-0.0069	-0.0090*
		(-1.899)	(-2.485)
COEMPLYR		-0.0129***	-0.0088***
		(-14.125)	(-9.739)
$HAC^{1/3}$		2.8730***	
		(52.749)	
$CDAC^{1/3}$			1.9810***
			(45.630)
$COJUDG^{1/3}$		-0.9913***	-1.2660***
		(-22.793)	(-28.264)
$CODPRY^{1/3}$		1.071***	1.2380***
		(18.905)	(22.216)
$COCRIM^{1/3}$		0.6684***	-0.6607***
		(13.004)	(12.810)

Model Chi-Square	43,686.02***	50,491.94***	55,306.84***
Observations	9,306,361	9,306,361	9,306,361

*Note: The symbols ***, ** and * denote significance levels at 0.001, 0.01 and 0.05, respectively.*

The predicted probabilities of investor harm obtained from the probit regression can be used to distinguish broker-year observations associated with ensuing investor harm, rank observations into quintiles and assess the model's predictive power by the true positive rate, false positive rate, and proportion of broker-year observations with subsequent investor harm allocated to each quintile. Table 16 presents the within-sample predictive measures of the three probit models.

Table 16: Within-Sample Predictive Performance of Probit Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	71.73%	27.79%	1.82%	5.00%	9.83%	22.36%	60.99%
Add Coworker Features	75.98%	27.42%	1.24%	3.58%	8.63%	21.09%	65.46%
Replace SA with CD	73.18%	23.24%	1.24%	3.11%	7.87%	19.12%	68.65%

Table 17 reports the out-of-sample prediction results, where the out-of-sample measures are obtained using 5-fold cross-validation repeated 10 times.

Table 17: Cross-Validated Predictive Performance of Probit Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	71.74%	27.81%	1.82%	4.98%	9.84%	22.36%	61.00%
Add Coworker Features	75.95%	27.41%	1.25%	3.57%	8.63%	21.09%	65.45%
Replace SA with CD	73.16%	23.25%	1.25%	3.12%	7.88%	19.10%	68.66%

The baseline model captures the lowest proportion (60.99%) of within-sample investor harm events in the top quintile, and the "replace SA with CD" model captures the highest (68.65%). Each model gives more accurate predictions than the corresponding model built on the restrictive data (Table 8) in the sense that they allocate more investor harm events to the top quintile and fewer investor harm events to the bottom quintile. The predictive power of the three models increases in order, confirming that coworker disclosure and employment history and all types of customer dispute records are useful in predicting investor harm.

The within-sample and out-of-sample measures are extremely close, showing that they characterize the models' true predictive power. The best-performing model, "Replace SA with CD", allocates an average of 68.66% investor harm events in the hold-out samples to the top quintile, compared to an average of 60.00% investor harm events captured by the same model built on the restrictive data. While this improvement in prediction accuracy may simply be a result of larger training sample size, which generally reduces the *bias* of regression models, the larger number and higher incidence of investor harm events[13] may also have played a significant part. Adding investor harm events associated with brokers first registered before 2000 or currently registered with no more than 3 states likely endows the model with additional power to distinguish brokers associated with investor harm from others.

4.3 Gradient Boosting Machine Models

We build three GBM models on the full data using the same sets of features as described above without any power transformation. The number of terminal nodes in regression trees, the shrinkage parameter, and the number of iterations are chosen to be 4, 0.05, and 2000, respectively. We calculate the in-sample predictive measures of the GBM model using the predicted probabilities and report the results in Table 18.

Table 18: Within-Sample Predictive Performance of GBM Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Baseline Reproduction	76.08%	26.11%	1.21%	3.45%	8.25%	18.48%	68.61%
Add Coworker Features	79.04%	24.73%	0.68%	2.19%	6.66%	16.91%	73.56%
Replace SA with CD	79.02%	22.18%	0.64%	2.10%	5.87%	14.91%	76.48%

Each GBM model has better predictive performance than the probit model using the same set of features, allocating a higher proportion of investor harm events to the top quintile and a lower proportion to the bottom quintile. The discriminative power of the three models increases in order with the third model allocating 76.48% investor harm events to the top quintile and 0.63% to the bottom quintile, the best performance among all the models discussed in this paper. A comparison of Table 18 to Table 10 shows that the GBM models built on full data have more predictive power than those built on restrictive data, a pattern also observed for the probit regressions.

Table 19 reports the out-of-sample measures for the "Replace SA with CD" model by performing 5-fold cross-validation 10 times and averaging over hold-out samples. The model allocates an

[13] The number of broker-year observations associated with investor harm is 31,280 (0.34%) in the full data, in contrast to 2,094 (0.21%) in the restrictive data.

average of 76.04% investor harm events in the hold-out samples to the highest-risk quintile and an average of 0.66% to the lowest-risk quintile, which are quite similar to the in-sample measures in Table 18. This indicates that the GBM models only slightly overfit the data and the prediction results recorded in Table 18 are robust and indicative of the GBM models' true predictive power. The out-of-sample measures for the other two models are omitted here due to computational burdens[14].

Table 19: Out-of-Sample Predictive Performance of GBM Models

			Investor Harm Events				
Model	**True Positive**	**False Positive**	**1st Quintile**	**2nd Quintile**	**3rd Quintile**	**4th Quintile**	**5th Quintile**
Replace SA with CD	78.66%	22.23%	0.66%	2.14%	6.01%	15.14%	76.04%

We record the feature importance scores for the GBM models built on the full data in Table 20. The importance scores between different features are similar to those in the models built on restrictive data (Table 12). Broker settlements and awards (SA) is the most influential feature in both the first and second models, and broker customer disputes (CD) is the most important feature in the third model, achieving a score higher than the next four features combined. The broker qualification and disclosure features EXAM, CRIM, JUDG, and EXPEL seem to have negligible impact on the model. Similar to the Z-scores for probit models (Table 15), the importance scores for SA and CD are relatively more extreme compared to those in Table 12, showing that the GBM models built on full data rely more heavily on individual history of customer complaints to derive their predictive power than those built on restrictive data.

Table 20: Relative Importance Scores for GBM Models

Feature	**Baseline Model**	**Add Coworker Features**	**Replace SA with CD**
SA	47.15	31.99	
CD			34.72
JUDG	1.27	0.35	0.47
BKCY	0.58	0.20	0.33
DPRY	5.39	2.46	2.05
CRIM	1.28	0.52	0.41
EXAM	2.39	0.81	0.60
EXPEL	2.10	0.18	0.11

[14] On a desktop with Intel i7 CPU and 32GB RAM, it takes about four days to finish the 5-fold cross-validation with 10 repetitions.

PREMPL	6.96	2.29	2.55
EMPLYR	13.93	5.64	5.09
DUAL	4.52	2.88	2.30
MKRN	14.42	7.54	7.44
CONUM		6.64	6.04
COEXPEL		5.57	5.25
COPREMPL		3.65	4.13
COEMPLYR		6.27	5.03
HAC		10.31	
CDAC			9.95
COJUDG		3.99	3.95
CODPRY		4.44	4.66
COCRIM		4.24	4.91

5 Discussion

5.1 Current BrokerCheck Provides Little Useful Information

Qureshi and Sokobin conclude that BrokerCheck information "has significant power to discriminate between brokers associated with investor harm events and other brokers [because] [t]he 20% of brokers with the highest ex ante predicted probability of investor harm are associated with more than 55% of the investor harm events in our sample." While this statement may accurately reflect their statistical analysis of more than a dozen variables for each of the 181,133 brokers in their study, it says nothing about whether BrokerCheck provides retail investors with any useful information. An investor in Dallas going to the BrokerCheck website to research a broker who just cold-called them from an impressive sounding Long Island brokerage firm cannot determine whether the broker is one Qureshi and Sokobin determined to be one of "the 20% of brokers with the highest ex ante predicted probability of investor harm."

Qureshi and Sokobin's analysis and our analysis start with BrokerCheck records on over 1.2 million brokers. We both apply sophisticated statistic modeling. Investors need all the data and our sophisticated modeling to glean the information Qureshi and Sokobin attribute to the BrokerCheck data. Retail investors preyed upon by bad brokers do not have access to the all the ostensibly public BrokerCheck data or to our analytical capabilities. Retail investors can only observe an infinitesimal portion of the BrokerCheck data. Investors querying BrokerCheck only see information on one broker at a time and so do not know whether a broker's reported characteristics are unusual or not and whether those characteristics portend a higher likelihood that the broker they are querying will engage in fraud.

Imagine that BrokerCheck displays data on a wall eight feet high and running the 120 yard length of a football field, from the back of one end zone to the back of the far end zone. Now imagine that FINRA places a black-out drape over the entire length of the wall. Qureshi and Sokobin find patterns using all the data on the wall but when investors want information from BrokerCheck, FINRA opens the drapes the thickness of two sheets of copier paper. If investors were to review the information on 100 brokers it would still only be less than 1/2 an inch of data on FINRA's shrouded 120 yard wall of data. Even if investors had the analytical capabilities we have they could never infer the patterns of broker characteristics and investor harm from that vanishing small sliver of the data what Qureshi and Sokobin derive from the BrokerCheck data.

5.2 Current BrokerCheck Information is Insufficient to Differentiate High Risk from Low Risk Brokers

Our analysis, and the analyses conducted by Egan, Matvos and Seru and Qureshi and Sokobin, show that association with past customer complaints and disciplinary events is a good indicator of higher propensity for future investor harm. While avoiding brokers with disclosure events may be a good rule of thumb for unsophisticated investors who have access to nothing more than public BrokerCheck information, it is not sufficient. Even at the highest risk firms, 80% of brokers don't have customer complaints. The 20% of brokers at these firms with a history of customer complaints do, though, increase the likelihood that other brokers at the same firm with a clean record will cause investor harm in the future. Investors need to know the disciplinary history of a broker's co-workers.

To illustrate, consider two brokers with the same time in the industry and identically clean records at the end of 2014 - no customer complaints, no judgments or liens, no bankruptcies, no disciplinary events, no criminal record. Both have passed two exams, were never affiliated with an expelled, had only one prior employer and were dually registered. An investor using BrokerCheck to make an educated choice between the brokers would be at a loss. BrokerCheck could not help this investor select the more trustworthy broker. Perhaps she would toss a coin.

One of these two brokers ("Broker A") had a customer complaint in 2015 that is still pending as of 2016 while he maintained a clean disclosure record in 2014. The broker with a complaint in 2015 is currently employed at Aegis Capital Corp, and the broker without any customer complaint is employed at Pyramis Distributors Corporation LLC.

Noticeably, our statistical models are able to provide some additional insights into the quality of these two brokers. The probabilities of inflicting investor harm in 2014 predicted by the "Replace SA with CD" probit model[15] for the Aegis broker was 0.50% and for the Pyramis broker was 0.007%. The probit analysis places the Aegis broker in the highest risk quintile on December 31, 2014 and the Pyramis broker in the lowest risk quintiles on December 31, 2014. The discriminative power of the "Replace SA with CD" model results from its use of eight coworker features, none of which is accessible through BrokerCheck. The different coworker features of the two brokers on

[15] The probit model is built on the restrictive data containing brokers who first registered after 1999 and have 4 or more current state registrations.

December 31, 2014 are presented in Table 21.

Table 21: Coworker Input and Coefficients of Two Brokers

	CONUM	COPREMPL	COEMPLYR	$CDAC^{1/3}$	$COJUDG^{1/3}$	$CODPRY^{1/3}$	$COCRIM^{1/3}$	COEXPEL
Broker A	58	2.73	6.62	0.3895	0.337	0.2267	0.1595	0.4398
Broker B	62	1.4	6.34	0	0.1136	0	0.2103	0
Δ	-4	1.33	0.28	0.3895	0.2234	0.2267	-0.0508	0.4398
C3 Coef	-0.00002	0.0774	-0.0950	2.333	0.2783	0.1713	-0.2634	0.2734
$\Delta \times$ Coef	0.00008	0.10294	-0.0266	0.90870	0.06217	0.03883	0.01338	0.12024

On average, the Aegis broker had a much more harmful group of coworkers than the Pyramis broker. Although there was no definitive evidence of a causal relationship between having distinctive coworker profile in 2014 and causing investor harm in 2015, adding coworker information has clearly helped the statistical model make a more educated (in this case, likely correct) guess about the likelihood of future investor harm. Among all the coworker features in this example, CDAC or Customer Disputes Associated with Coworkers has the higher impact, 74.50% on the final predicted probability.

While Qureshi and Sokobin also noticed that "investors may benefit from information about harm associated with brokers' coworkers" (which they also denoted by HAC), their analysis has left an impression that including coworker harm variable can only lead to a marginal increase in efficiency on top of the current BrokerCheck information[16]. The above example shows that information about coworkers, in particular CDAC is tremendously useful for investors to make wise decisions in choosing between brokers with clean disclosure records.

The importance of publishing relevant coworker information on BrokerCheck and making all the public information truly public is further confirmed by regression analysis on the subset of broker-year observations that are not associated with any prior personal disclosure events. We construct a sub-sample containing the broker-year observations for which all the broker disclosure features (i.e., CD, JUDG, BKCY, DPRY, and CRIM) equal zero, and fit to this sub-sample the probit model "Replace SA with CD" with the five broker disclosure features excluded. The model built on the sub-sample created from the restrictive data allocates 52.68% and 1.92% of investor harm events to the highest and lowest risk broker-year quintiles, respectively. The model built on the sub-sample created from the full data captures 58.39% and 1.74% of investor harm events in the top and bottom quintiles, respectively. The effectiveness of the regression models shows that the coworker disclosure and employment history contains valuable information for predicting the *first incidence* of investor harm event in a broker's career, and this information, if carefully compiled and explained, may protect investors from potential misconduct by brokers with clean disclosure record.

[16] Their baseline probit model allocated 55.5% of the investor harm events to the highest quintile, compared to 58.9% of the investor harm events captured in the top quintile in the "Baseline + HAC" prediction.

5.3 Fixing BrokerCheck and Reducing Misconduct Is Easy

Our results above confirm the findings of Qureshi and Sokobin and of Egan, Matvos, and Seru that the risk a broker will commit misconduct is significantly increased if he or she works with co-workers who have previously committed misconduct. In fact, investors would be as well informed to know the average misconduct history of a brokers coworkers as they would be knowing the brokers own misconduct history.

FINRA could easily make the public-facing BrokerCheck data available in bulk to anyone interested in analyzing the data. FINRA and the SEC have already determined that this information is not confidential and should be disseminated to the public. If the publicly available BrokerCheck information were truly publicly available researchers, third party vendors, ratings companies like Lipper and Morningstar, and news outlets like US News and World Report and BusinessWeek could rank brokerage firms on the risk of fraud. These rankings would generate substantial publicity and bad brokerage firms would no longer be able to prey on unsophisticated investors with relative impunity. We rank the 210 brokerage firms with 400 or more registered brokers by the percentage of their brokers associated with investor harm events as defined by Qureshi and Sokobin as of December 31, 2015. Table 22 excerpts the 30 firms with the highest percentage of brokers associated with investor harm events. We have identified the firms with more than 1,000 brokers in bold font. These are the firms considered in Egan, Matvos, and Seru's Table 6. There are six firms with a higher percentage of brokers associated with investor harm events than Oppenheimer, the highest risk firm with more than 1,000 brokers in the Egan, Matvos, and Seru study. We also compile firm rankings by the percentage of brokers with misconduct disclosures as defined by Egan, Matvos, and Seru. The top six firms in Table 22 are the same whether we rank them based on the Qureshi and Sokobin investor harm measure or Egan, Matvos, and Seru's financial misconduct measure.

These six firms - Aegis Capital, Summit Brokerage Services, National Securities, Centaurus Financial, Independent Financial Group and Kovack Securities employ a far higher percentage of brokers associated with investor harm events than other firms. These six highest-risk firms are also among the top ten firms ranked by percentage of current brokers who were previously fired by other firms after customer allegations of misconduct. 7.71% of the registered brokers in these six high risk firms have been fired at least once by a previous employer after allegations of misconduct, 10 times the average of 0.78% of the remaining 204 brokerage firms. Given their coworkers' disclosure record as of 2014, 83.7% of the brokers at these six firms would be in the highest risk quintile as defined by Qureshi and Sokobin and should be avoided by investors.

Table 22: Top 30 Firms with 400 or More Registered Brokers Ranked by Percentage of Brokers with Investor Harm Events Defined by Qureshi and Sokobin

Investor Harm Ranking	CRD	Firm Name	Registered Brokers	Investor Harm Brokers	Investor Harm Rate	Brokers Previously Fired	Previously Fired Rate
1	15007	AEGIS CAPITAL CORP.	444	107	24.10%	21	4.73%
2	34643	SUMMIT BROKERAGE SERVICES	676	129	19.08%	65	9.62%
3	7569	NATIONAL SECURITIES CORP.	760	137	18.03%	42	5.53%
4	30833	CENTAURUS FINANCIAL, INC.	602	98	16.28%	39	6.48%
5	7717	INDEPENDENT FINANCIAL GROUP	638	90	14.11%	50	7.84%
6	44848	KOVACK SECURITIES INC.	434	58	13.36%	57	13.13%
7	**249**	**OPPENHEIMER & CO. INC.**	**2,217**	**276**	**12.45%**	**92**	**4.15%**
8	877	WEDBUSH SECURITIES INC.	634	77	12.15%	32	5.05%
9	30613	INVESTORS CAPITAL CORP.	641	72	11.23%	39	6.08%
10	2543	WUNDERLICH SECURITIES, INC.	459	51	11.11%	21	4.58%
11	**8174**	**UBS FINANCIAL SERVICES INC.**	**12,555**	**1,377**	**10.97%**	**94**	**0.75%**
12	**32444**	**FIRST ALLIED SECURITIES, INC.**	**1,179**	**122**	**10.35%**	**49**	**4.16%**
13	46214	NEXT FINANCIAL GROUP, INC.	796	75	9.42%	21	2.64%
14	14503	VSR FINANCIAL SERVICES, INC.	511	48	9.39%	7	1.37%
15	18456	STERNE AGEE FINANCIAL SERVICES	580	53	9.14%	35	6.03%
16	**11025**	**WELLS FARGO ADVISORS FINANC**	**1,993**	**176**	**8.83%**	**30**	**1.51%**
17	18487	AMERICAN PORTFOLIOS FINANC	838	73	8.71%	36	4.30%
18	**149777**	**MORGAN STANLEY SMITH BARNEY**	**23,782**	**2,065**	**8.68%**	**151**	**0.63%**
19	**705**	**RAYMOND JAMES & ASSOCIATES**	**5,812**	**488**	**8.40%**	**101**	**1.74%**
20	**463**	**JANNEY MONTGOMERY SCOTT**	**1,369**	**114**	**8.33%**	**29**	**2.12%**
21	**793**	**STIFEL, NICOLAUS & COMPANY**	**4,588**	**380**	**8.28%**	**113**	**2.46%**
22	14303	SIGMA FINANCIAL CORPORATION	678	55	8.11%	19	2.80%
23	7684	INVESTACORP, INC.	500	39	7.80%	17	3.40%
24	**19616**	**WELLS FARGO ADVISORS, LLC**	**26,319**	**1,998**	**7.59%**	**309**	**1.17%**
25	**10205**	**SECURITIES AMERICA, INC.**	**2,662**	**191**	**7.18%**	**61**	**2.29%**
26	20804	UNITED PLANNERS' FINANC SERV	510	35	6.86%	19	3.73%
27	18697	GIRARD SECURITIES, INC.	477	31	6.50%	19	3.98%
28	**10299**	**CETERA ADVISORS LLC**	**1,618**	**105**	**6.49%**	**48**	**2.97%**
29	**29604**	**NATIONAL PLANNING CORP.**	**1,815**	**117**	**6.45%**	**38**	**2.09%**
30	**35747**	**PURSHE KAPLAN STERLING**	**1,229**	**78**	**6.35%**	**33**	**2.69%**

If FINRA unshackled BrokerCheck, researchers would come up with innovative ways to reach and inform unsophisticated investors about high risk brokers and brokerage firms. Releasing the potential of BrokerCheck to protect investors would also benefit some brokers and brokerage firms. Brokers with clean CRDs would have an incentive to move to firms with a lower proportion of bad brokers so they would not be penalized in the rankings for associating with bad brokers. Brokerage firms would compete to hire better brokers and fire brokers with prior settlements and awards to improve their quality rankings.

Continuing with the analogy above, proposals to supplement data items available on BrokerCheck or add a search term miss the mark badly. These proposals amount to adding a few inches of height to the wall *and* to the drapes. If FINRA continues to only allow investors the benefit of a glimpse at 0.01 inch of the 120 yard long wall it wont matter if the wall of data is 8 feet tall or 8 feet, 2 inches tall. The only way to empower investors to protect themselves is for

FINRA to take down the drapes.

References

S. Dimmock, W. Gerken, and N. Graham. "Is Fraud Contagious? Co-Worker Influence on Misconduct by Financial Advisors". Working paper, April 2016.

M. Egan, G. Matvos, and A. Seru. "The Market for Financial Adviser Misconduct". Working paper, March 2016.

J. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". *Annals of Statistics*, 29(5):1189–1232, 2001.

H. Qureshi and J. Sokobin. "Do Investors Have Valuable Information About Brokers?". Working paper, August 2015.