Human Capital Acquisition in Response to Data Breaches

ONLINE APPENDIX

A1. Definitions of Cybersecurity, PR, and Legal SOC Occupations

Our paper defines three categories of jobs relevant to data breaches: cybersecurity, PR, and legal. To select the job postings that fall into these categories, we rely on the posting's Standard Occupation Classification (SOC) code. Table A1 provides a detailed list of the specific occupations included in each category. In cases where the SOC code has changed between the 2010 and 2018 classification, we note the year associated with the code.

Table A1 Definitions of Cybersecurity, PR, and Legal SOC Occupations

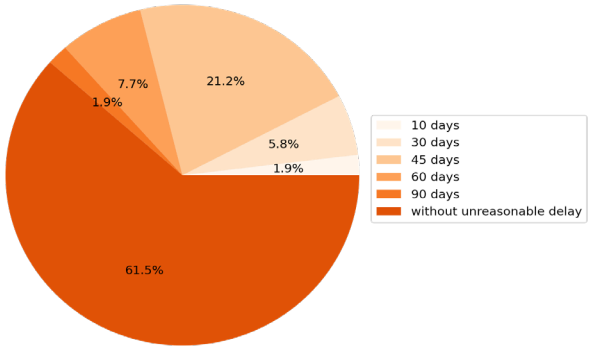
SOC Code	SOC Name	Notes
15-1120	Computer and Information Analysts	Cybersecurity; 2010 Code
15-1121	Computer Systems Analysts	Cybersecurity; 2010 Code
15-1122	Information Security Analysts	Cybersecurity; 2010 Code
15-1140	Database and Systems Administrators and Network Architects	Cybersecurity; 2010 Code
15-1141	Database Administrators	Cybersecurity; 2010 Code
15-1142	Network and Computer Systems Administrators	Cybersecurity; 2010 Code
15-1143	Computer Network Architects	Cybersecurity; 2010 Code
15-1152	Computer Network Support Specialists	Cybersecurity; 2010 Code
15-1210	Computer and Information Analysts	Cybersecurity; 2018 Code
15-1211	Computer Systems Analysts	Cybersecurity; 2018 Code
15-1212	Information Security Analysts	Cybersecurity; 2018 Code
15-1231	Computer Network Support Specialists	Cybersecurity; 2018 Code
15-1240	Database and Network Administrators and Architects	Cybersecurity; 2018 Code
15-1241	Computer Network Architects	Cybersecurity; 2018 Code
15-1244	Network and Computer Systems Administrators	Cybersecurity; 2018 Code
15-1245	Database Administrators and Architects	Cybersecurity; 2018 Code
11-2030	Public Relations and Fundraising Managers	PR
11-2032	Public Relations Managers	PR
27-3030	Public Relations Specialists	PR
27-3031	Public Relations Specialists	PR
23-1011	Lawyers	Legal
23-2011	Paralegals and Legal Assistants	Legal

A2. Data Breach Notification Laws and Data Breach Announcement Timing

As discussed in the Empirical Methods section, data breach notification laws in most states require firms to report such events without reasonable delay or within 30 to 60 days of noticing a breach of individual customer data. Therefore, the treatment time is not sharply identified at month zero as there could be strategic timing decisions by firms on when to announce breach events (Foerderer & Schuetz, 2022). We use the data breach announcement date as the treatment date for our baseline specification. While this is

likely neither the exact date of the breach nor the date on which the firm notices the breach, we consider it to be the most reasonable date available. However, in order to reduce noise due to the ambiguity in the treatment timing, we also test specifications that drop the observations from Quarter 0 (Months 0, -1, and -2) and present these results in Table A2. The results are largely consistent with our baseline results from column 1 in Tables 2, 3, and 4.

Figure A1 Timely Notification Requirements by States' Data Breach Notification Laws



Notes: Figure derived from Perkins Coie Security Breach Notification data – Revised June 2020, available at https://www.perkinscoie.com/en/newsinsights/security-breach-notification-chart.html. About 60% of US states require breached firms to notify the public as soon as they realize that they were breached. In total, 98% of all US states require firms to notify the public no longer than 60 days after suffering a data breach.

While our analysis is predicated on the change that occurs before and after the breach, it may be the case that firms strategically disclose their breach and hire prior to the breach disclosure. To mitigate the concern that strategic disclosure affects our results, we estimate the main regression, excluding quarter 0 while including an additional preceding quarter, e.g., quarter -4. This means we are excluding months -2, -1, and 0 from our estimation. These are the months most likely to be affected by strategic disclosure, as discussed above.

Table A2 Effect of Data Breaches on Hiring (Excluding Quarter 0)

Occupations	Cybersecurity	PR	Legal
Variables	(1) Full Sample	(2) Full Sample	(3) Full Sample
After × Breached	0.024*** (0.005)	0.011*** (0.003)	0.008** (0.003)
R-Squared	0.308	0.172	0.245
# of Firms	89,121	89,121	89,121

Notes: All results are based on the full analytic sample, similar to column 1 of Tables 2, 3, and 4. Each column estimates the difference-in-differences specification outlined in Equation (1) in the manuscript. Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A3. Alternative Time Window

Our main analysis uses a window of four quarters before and after the data breach. However, one concern is that this extended window may not effectively capture the transient nature of the breach's impact. Moreover, a longer time window may increase the probability that confounding factors are a source of bias. We therefore re-estimate our main specification restricting the sample period to only two quarters before and after the breach. The results are largely consistent with our baseline results from column 1 of Tables 2, 3, and 4.

Table A3 Effect of Data Breaches on Hiring (Two Quarters Before and After)

Occupations	Cybersecurity	PR	Legal
Variables	(1) Full Sample	(2) Full Sample	(3) Full Sample
After × Breached	0.019*** (0.005)	0.010*** (0.004)	0.004 (0.004)
R-Squared	0.307	0.171	0.244
# of Firms	89,063	89,063	89,063

Notes: All results are based on the full analytic sample, similar to column 1 of Tables 2, 3, and 4 with one key difference: we focus on a shorter time window (i.e., two quarters before and after the data breach events). Each column estimates the difference-in-differences specification outlined in Equation (1). Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A4. Alternative Model Specification

Linear probability models have two primary limitations: they can produce predicted probabilities outside of the zero-one range, and they possess constant marginal effects. As a result, a logit model may be more conducive when the dependent variable is binary, as in our case. We estimate the main regression in column 1 of Tables 2, 3, and 4, replacing the linear probability model with a logit model. Due to computational limitations and the incidental parameter problems associated with nonlinear models, we estimate the regression with only firm fixed effects. The results are displayed in Table A4.

Table A4 Effect of Data Breaches on Hiring (Logit Models)

Occupations	Cybersecurity	PR	Legal
Variables	(1) Full Sample	(2) Full Sample	(3) Full Sample
After × Breached	0.367*** (0.039)	0.238*** (0.050)	0.238*** (0.054)
# of Firms	59,370	32,521	26,214

Notes: Columns 1, 2, and 3 estimate the two-period difference-in-differences specification outlined in Equation (1). Each cell reports the coefficients of the independent variable from the Logit model. Only firm-fixed effects are included in the specifications because of computational limitations and the incidental parameter problem associated with fixed effects. Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A5. Quarterly Dynamics with Propensity Score Matching

Although our identification strategy exploits the fact that the timing of the data breach events is largely unanticipated by affected firms, post-event labor demand may be driven by other firm-level characteristics prior to the data breaches such as firm size or firms' prior hiring for cybersecurity and other personnel, which could lead to biased estimates. We thus perform propensity score matching to construct a sample that consists of control and treatment firms that are relatively more similar, based on pre-data-breach firm-level observables. More specifically, we match each firm using their cumulative number of job postings (as a proxy for firm size), cumulative number of job postings for cybersecurity, PR, and legal talent in the 12 months prior, as well as the industry of the firms. We match the treatment and control groups through the nearest neighbor matching method and impose common support with 5 neighbors and 0.001 caliper.[1]

Using the matched sample, we re-estimate our baseline specification of quarterly dynamics. We find consistent and robust results similar to the ones reported in Figures 1, 2, and 3.

Table A5 Quarterly Dynamics with Matched Sample using Propensity Score Matching

Models	Probability		
Variables	Cybersecurity	Public Relations	Legal
	(1)	(2)	(3)
Quarter -3	-0.009	0.005	-0.005
	(0.008)	(0.006)	(0.006)
Quarter (-2)	-0.009	-0.003	-0.000
	(0.007)	(0.006)	(0.005)
Quarter (0)	0.002	0.008	0.005
	(0.007)	(0.005)	(0.005)
Quarter (+1)	0.006	0.012**	-0.008
	(0.008)	(0.006)	(0.006)
Quarter (+2)	0.025***	0.010*	0.012**
	(0.009)	(0.006)	(0.006)
Quarter (+3)	0.010	0.002	0.003
	(0.009)	(0.006)	(0.006)
Quarter (+4)	0.012	0.010	0.004
	(0.009)	(0.006)	(0.006)
Observations	610,932	610,932	610,932
R-squared	0.424	0.282	0.338
# of firms	5,534	5,534	5,534
Mean(Y)	0.164	0.047	0.047

Notes: the results are based on the matched sample. The outcome variables for columns 1-3 are the indicators for firms that, respectively, posted jobs for cybersecurity, public relations, and legal occupations. Quarter -1 is the benchmark period. Firm, industry and time, and pair-fixed effects are controlled in the specification. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

[1] Robustness checks using different numbers of nearest neighbors generate qualitatively and quantitatively similar results. These results are available upon request.

A6. Placebo Test of Non-Relevant Occupations

Though we identify significant effects throughout the manuscript, one potential concern may be that these identified demand increases may coincide with other shocks to the firm that are not a function of a data breach. To address this concern, we replicate our main table for occupations not relevant to a data breach, i.e. any occupation not defined in Table A1 above.

Table A6 Effect of Data Breaches on Hiring for Non-Relevant Occupations

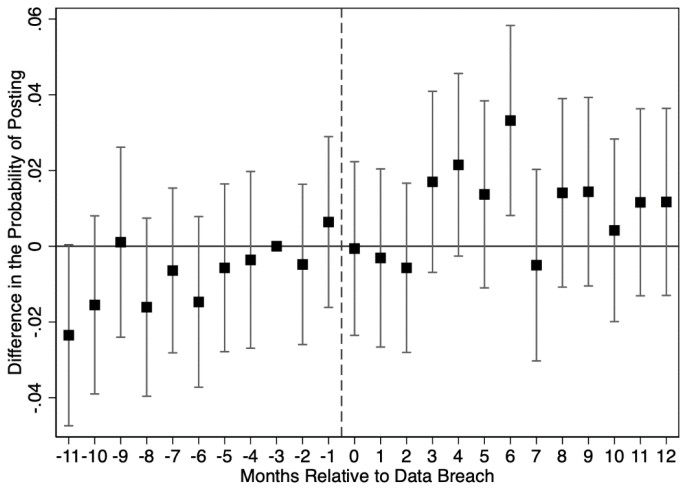
Models	Probability			Counts	
Variables	(1) Full Sample	(2) Matched Sample	(3) Matched Sample Sun & Abraham	(4) Full Sample	(5) Full Sample Poisson
After × Breached	-0.008	0.002	0.003	16.45	0.978
	(0.007)	(0.009)	(0.009)	(10.35)	(0.081)
R-Squared	0.311	0.346	0.352	0.383	-
# of Firms	89,121	5,077	5077	89,121	89,043

Notes: Results in columns 1, 4, and 5 are based on the full sample. Results in columns 2 and 4 are based on the matched sample. Each column estimates the difference-in-differences specification outlined in Equation (1) while column 3 further employs the S&A estimator. Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A7. Monthly Dynamics of Firm's Hiring in Response to Data Breaches

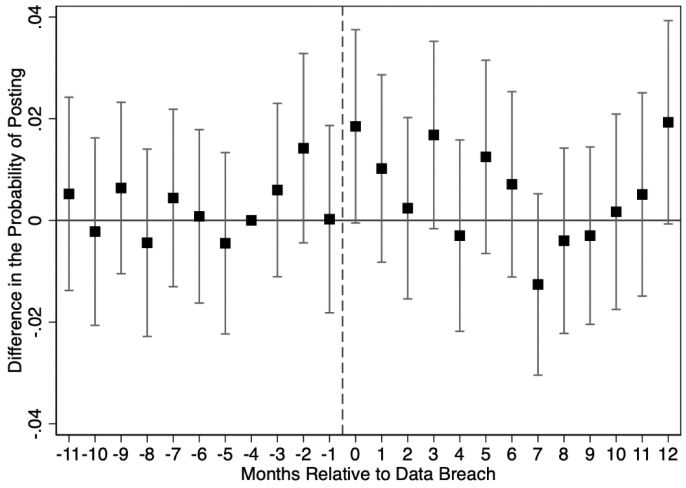
To further assess the monthly dynamics of firm's hiring in response to data breaches, we estimate a model with monthly indicators with propensity score matched sample.

Figure A2 Monthly Dynamic for Cybersecurity Hiring



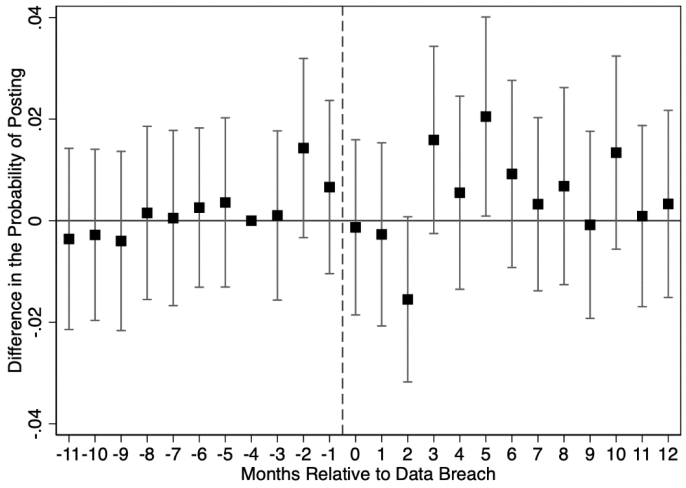
Notes: Figure displays the parallel trends test described in the Empirical Methods section, using the monthly coefficients on the matched sample. Quarter -3 is the omitted category. The reported 95 percent confidence intervals use standard errors clustered at the firm level.

Figure A3 Monthly Dynamic for PR Hiring



Notes: Figure displays the parallel trends test described in the Empirical Methods section, using the monthly coefficients on the matched sample. Quarter -3 is the omitted category. The reported 95 percent confidence intervals use standard errors clustered at the firm level.

Figure A4 Monthly Dynamic for Legal Hiring



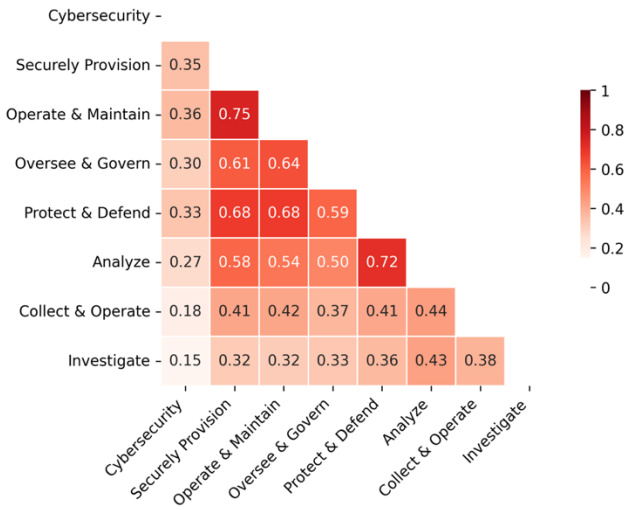
Notes: Figure displays the parallel trends test described in the Empirical Methods section, using the monthly coefficients on the matched sample. Quarter -3 is the omitted category. The reported 95 percent confidence intervals use standard errors clustered at the firm level.

A8. Correlation between NICE Categories and the Cybersecurity Occupations

Figure A5 shows the correlation between the cybersecurity-related indicator variables in our study: our definition of cybersecurity occupations, which is an occupation-based measure, and the Lightcast correspondence to the seven NICE categories. The cybersecurity occupation measure, which includes IT and computer networking-related occupations, is somewhat correlated with the NICE-based measures of cybersecurity functions, suggesting that the cybersecurity-related skills captured by the NICE framework cover a broader range of cybersecurity-related functions.

Moreover, while indicator variables for NICE categories are highly correlated with one another – for example, Securely Provision and Operate & Maintain functions show a correlation coefficient of 0.75 – others are less correlated, indicating meaningful differences across these functions. The findings thus emphasize the demand heterogeneity across occupations and skills and highlight the value of the NICE categories for identifying granular yet distinctive hiring responses.

Figure A5 Heatmap of Cybersecurity Variables



Notes: This heatmap visualizes the correlation coefficients of cybersecurity variables and NICE indicators, derived from the estimation sample at the firm-month level.

A9. Effect Heterogeneity of Data Breaches by NICE Cybersecurity Functions

The relative effect size (measured by the coefficient divided by the dependent variable mean) is greatest for "Oversee & Govern" at 168% (0.021/.013). This is consistent with firms hiring workers to manage and develop strategies for cybersecurity work. While CIOs and CTOs fall into this domain, this also includes non-executive workers, such as IT project auditors, project managers, program managers, and cyber policy and strategy planners. The second largest effect appears for "Protect & Defend" – roles that identify, analyze, and mitigate threats to IT systems and networks, such as vulnerability assessment analysts and cyber-defense analysts, among others. This effect demonstrates that firms that experience breaches respond substantively and in a targeted manner to incidents by strategically improving relevant infrastructure and/or assessing their systems and networks to understand their vulnerabilities. In contrast, we do not observe firms seeking cybersecurity roles that fall into the "Collect & Operate" or "Investigate" categories. "Collect & Operate" involves providing specialized denial and deception operations.

Table A7 Effect Heterogeneity by Cybersecurity Functions (Cyber Breaches Only)

Variables	Securely Provision	Operate & Maintain	Oversee & Govern	Protect & Defend	Analyze	Collect & Operate	Investigate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
After × Breached	0.019***	0.021***	0.021***	0.018***	0.011***	0.006**	0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)	(0.002)
R-Squared	0.136	0.137	0.116	0.120	0.102	0.093	0.082
# of Firms	88,078	88,078	88,078	88,078	88,078	88,078	88,078
Mean(Y)	0.015	0.017	0.013	0.012	0.007	0.004	0.002

Notes: Results are based on the full sample with cyber-related data breaches only since NICE is a cybersecurity workforce framework. Each column estimates the difference-in-differences specification outlined in Equation (1). The dependent variables are binary indicators that equal one for firms posting any job ads that correspond to each category from the NICE Cybersecurity framework. Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A10. Matching Lightcast Hiring Data with PRC Data Breach Events

Since the PRC and Lightcast databases do not share a common firm identifier, we take a multi-step approach to merge the two databases. Specifically, we first clean and standardize the firm name strings in both databases. Next, we use a combination of name and address fuzzy matching to construct a bridge between the PRC and Lightcast data. After algorithmic matching, we manually validate and check all possible high-quality matches to further increase the match rate. Overall, we identify and match over 50% of organizations from the PRC data in the Lightcast data. We further require that all organizations have at least 100 job postings in the Lightcast data (i.e., 2010 - 2020) to ensure sufficient quality of the job posting data (though our results are robust to different cutoffs). Overall, our main sample contains a total of over 87,000 organizations and over 1,400 data breach events. Many matched organizations in the PRC data are smaller local business (e.g., local clinics) and are hence dropped due to the 100-job posting cutoff we impose. On average, organizations in our sample have 12 job postings per month with about 0.32 of them related to cybersecurity occupations in each month.

A11. Wage Bill Estimation

To better understand the economic magnitude of these hiring responses, we estimate the implied wage bill increases in monetary terms. Since reliable wage information is often absent from job postings (Batra et al. 2023), we use average, occupation-level annual wages from the Bureau of Labor Statistics' Occupational Employment and Wage Statistics (OEWS) along with the number of job postings in the related occupations to calculate firms' implied additional wage bills that the firms intend to spend on the new hirings. We estimate Equation 1, our main regression specification, with the wage bills associated with cybersecurity, PR, and legal hiring. In addition, we also perform the specification on the total wage bill, combining the three categories.

Table A8 Effect of Data Breach on Wage Bill

Occupations	Total Wage Bill	Cybersecurity	PR	Legal
Variables	(1)	(2)	(3)	(4)
After × Breached	61,961***	52,010***	1,965	7,986***
	(14,701)	(13,192)	(1,836)	(1,883)
R-Squared	0.880	0.895	0.324	0.469
# of firms	89,121	89,121	89,121	89,121

Notes: All results are based on the full analytic sample, similar to the one used in column 1 of Table 2, 3, and 4 with the dependent variable as the imputed wage bill the firms intend to spend on the new hires. Each column estimates the difference-in-differences specification outlined in Equation (1). Standard errors are clustered at the firm level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.