Does Product Market Competition Drive CVC Investment? Evidence from the U.S. IT Industry

Statistical Appendix

1. Detailed Description on TNIC-based Measures

1.1 TNIC Industry Classification

TNIC industry classifications are based on the foundation that firms in the same industry use many of the same words to describe their products in the business description section of firm 10-Ks. Using 10-K text offers many advantages including improved signal strength, the ability to measure competition dynamically even when the product market changes rapidly, and the ability to use a generalized intransitive network structure. Existing standard classifications including SIC and NAICS offer little if any parallel capabilities. These features have proven useful in other settings including research on mergers, asset prices, payout policy, and firm organizational form. They are particularly well suited for use in IT industry, which is known for rapid change.

In order to construct TNIC industries, Hoberg and Phillips (2010, 2015) web crawl all available 10-Ks from the SEC's Edgar website and link each 10-K to the Compustat database in each year using the central index key (CIK) as the unique firm identifier. A link table from CIK to Compustat gvkey is provided by the Wharton Research Data Service (WRDS) through the SEC Analytics package. Once business description sections are parsed from each 10-K, common words (those appearing in more than 25% of all 10-Ks in a given year) are discarded. Additionally, any word that does not appear as a noun or proper noun is also discarded. The typical firm uses 200 unique words, and any firm using fewer than 20 is omitted.

Words are then mapped to numerical vectors and firm pairwise cosine similarity scores are computed for every pair of firms in each year (cosine similarities are a popular tool in computational linguistics, see Sebastiani (2002) for example). In particular, each firm *i*'s vocabulary can be represented by a vector P_i, which has a length equal to the number of unique words, with each element being populated by the number one if firm *i* uses the given word, and zero if it does not. These vectors are then normalized to have unit length (which is done by dividing all elements of the vector by the square root of the vector's dot product with itself). We denote the normalized vectors as V_i. Because all firms are thus represented by vectors of length one in the same space, it follows that TNIC industries imply that firms have unique locations on a high dimensional unit sphere, and that they move along the surface of the sphere as their products are revised from year to year.

To identify peers in the TNIC network, we calculate the cosine similarity between pairs of firms *i* and *j* as follows:

$$Cosine\ Similarity_{i,j} = (V_i \cdot V_j)$$

The full TNIC network is thus fully described by an *NxN* square matrix that is populated with cosine similarities between all firms in each year, where *N* is the number of firms. Due to the properties of cosine similarities, all entries are real numbers in the interval [0,1]. Because firms update their 10-Ks annually, the network (the entire square matrix) is time-varying. In the current article, the level of coarseness of TNIC-3 matches that of three digit SIC codes, as both classifications result in the same number of firm pairs being deemed related (Hoberg and Phillips 2015). For example, if one picks two firms at random from the CRSP/COMPUSTAT universe, the likelihood of them being in the same three digit SIC code is 2.05%. Analogously, when the TNIC-3 cutoff is specified using our approach, the likelihood of two randomly drawn firms being deemed related in their TNIC-3 is also 2.05%. Hence, TNIC-3 is constructed to be 'as coarse' as are three digit SIC codes.

TNIC industries have many important features. First, tests in Hoberg and Phillips (2015) illustrate that they are more informative than other frequently used industry classifications such as SIC or NAICS

code. Second, the entire classification is dynamically updated every year because firms file updated 10-Ks every year as required by regulation S-K. Third, TNIC industries are part of a generalized intransitive network, providing flexibility to consider customized sets of rivals for each firm. In contrast, classifications such as SIC codes are less powerful, rarely updated, and are constrained to be transitive.

Hoberg and Phillips (2010, 2015) document that TNIC is more informative than SIC codes along many dimensions. We note that these comparisons of TNIC and SIC are done while holding granularity fixed, as the baseline TNIC classification has as many firm pairs being in the same industry as do SIC-3 industries. These validations are statistical, and are in addition to the aforementioned technical advantages. The first validation relates to regressions in which a number of accounting variables are regressed on industry controls, either based on TNIC or SIC-3. By examining the adjusted R^2 in both regressions, the results show that TNIC industries absorb considerably more variation in firm characteristics than do equally granular SIC-3 industries. These characteristics include more primitive variables such as sales growth, market betas, and profitability as well as policy variables such as leverage, cash holdings, and dividend policy. The increased explanatory power of TNIC is substantial. For example, when explaining firm profitability (operating income over sales), SIC-3 industry controls generate roughly 30% adjusted RSQ, which compares to 45% for TNIC controls.

The second major validation test considers measures of competition based on TNIC variables versus those based on SIC-3. The objective is to first define a dependent variable as the extent to which the firm mentions "high competition" in its management's discussion and analysis section of its 10-K. The test reveals that TNIC based measures of competition are considerably more informative in predicting managerial mentions of high competition, consistent with TNIC providing more informative measures of competition. In all, the higher informativeness of TNIC found in these validations, along with the key technical advantages of TNIC including its dynamic updating, indicate why TNIC industries are particularly well suited to the IT industry, which is the focus of the current paper.

1.2 TNIC Total Similarity

TNIC similarity is based on the primitive concept of product differentiation. A firm in a high TNIC similarity market has low product differentiation and thus faces a high degree of competitive pressure due to its inability to separate itself from its rivals through product choice. Analogously, a firm with a low TNIC similarity effectively holds a local product monopoly in its market and faces little pressure. The concept of product differentiation dates back to seminal papers in the economics literature including Chamberlin (1933) and Hotelling (1929), with the latter studying modeling product differentiation in a spatial model. We note that TNIC also has a spatial interpretation, making the parallel with Hotelling (1929) quite direct and apposite. Although it has roots in the early literature, differentiation continues to be an active area of research in more recent studies including Seim (2006), Mazzeo (2002), and Hoberg and Phillips (2015) among others.

TNIC similarity is a measure of product differentiation based on a dynamic network that is based on 10-K text and the network has the benefit of being fully redrawn in each year as 10-Ks are updated.

1.3 Product Market Fluidity

Product market fluidity is based on the foundation that the intensity at which rival firms are revising their 10-K product market vocabulary measures the extent to which product markets surrounding a firm are in flux (or are "fluid"). In turn, this implies the extent of product market threats a firm faces, and also the speed at which products are evolving in a particular market (Hoberg et al. 2014). Because fluidity focuses on product market change, it is distinct from measures of static competition including those based on TNIC industry definitions such as total similarity, rival counts, or HHI measures.

To compute product market fluidity, let J_t denote a scalar equal to the number of all unique words used in the product descriptions of all firms in year *t*. Let W_{it} denote a boolean vector of length J_t

identifying which of the J_t words are used by firm *i* in year *t*. Element *j* of W_{it} equals one if firm *i* uses word *j* in its product description and zero otherwise. We normalize W_{it} to unit length and define the result as N_{it}. We then define the aggregate word change vector $D_{t-1,t,}$as

$$D_{t-1,t} \equiv \left| \sum_j (W_{j,t} - W_{j,t-1}) \right|$$

A firm's product market fluidity is then the dot product between its own word vector N_{it} and $D_{t-1,t}$, suitably normalized.

$$Product\ Market\ Fluidity_i \equiv \langle N_{i,t} \cdot \frac{D_{t-1,t}}{\|D_{t-1,t}\|} \rangle$$

Fluidity is thus a cosine similarity between a firm's own word usage vector N_{it} and the aggregate change vector $D_{t-1,t}$. Because this dot product is based on nonnegative vectors, fluidity thus lies in the interval [0,1].

Product market fluidity is higher when a firm's words overlap more with $D_{t-1,t}$, the vector that reflects rival changes. In such cases, the interpretation is that competitive threats and the rate of product change are higher. We also note that product market fluidity reflects product market threats measured from competitor actions, and not own-firm actions. The notion that rival threats are important, perhaps even more so than static measures of market share, is consistent with theories of contestable markets in industrial organization (Baumol et al. 1983). Fluidity is an empirical construct that captures this intuition.

2. **Additional Robustness Tests**

In this section, we provide additional robustness tests to establish the validity of the results presented in the paper. These include the interaction of product market competition with product market fluidity, the use of additional instrumentation based on the TNIC, estimation of a panel-data Heckman analysis, additional analyses to account for technological and financial shocks, and the use of different dependent variables. In each case, we show that our central relationships of interest are robust.

2.1 Interaction of Product Market Competition with Product Market Fluidity

Even though the main focus of the analysis presented in the paper is on the level of competition faced by focal firms in their product markets, other characteristics of competitive environments, such as the dynamism of competition and the cost of entry, may also play a role in driving CVC investments. The TNIC captures more of the dynamism present in IT markets relative to SIC codes but we go one step further to characterize the dynamic nature of competition. Consider a particularly turbulent environment in which rapidly changing customer needs and technologies create severe uncertainties for the IT-producing firm (Pavlou and El Sawy 2006). In such environments, flexible CVC investments are likely to be more attractive. Traditional measures of environmental turbulence and change have included survey measures (Jaworski and Kohli 1993, Pavlou and El Sawy 2006). Given our focus on text analytics, we consider a novel product market fluidity measured proposed by Hoberg et al. (2014) as an alternative measure of the dynamic nature of the firm's market space. Product market fluidity measures product market turbulence, i.e, when the product descriptions of a focal firm's rivals are changing rapidly from one year to the next. Thus, product market fluidity is higher when competitive threats and the rate of product change are higher for the focal firm. This construct is defined more formally in the next section of this Appendix.

We split the sample of firms into quartiles based on the product market fluidity they face in a given year and then interact the competition variables with each quartile dummy. We regress these variables on both the sales-normalized CVC spending variable as well as the ratio of CVC to CVC plus R&D spending variable; these estimates are presented in Table A1. We find that the effect of competition on CVC is significant mainly for firms in the high fluidity group. The result show that high turbulence of product markets from rival actions *increases* the incentive of IT firms to use CVC investment to escape competition. This finding also provides some evidence for why IT-producing firms are major CVC

investors (Carnoy and Castells 1997); the IT sector is particularly notable for high levels of turbulence and volatility overall in its product markets (McAfee and Brynjolfsson 2008). Therefore, it is not surprising that open innovation models like CVC appear particularly attractive here.

Table A1: Interaction of Competition with Product Market Fluidity

	DV: Ln(CVC/sales)			*DV: Ln(CVC/(RD+CVC))*		
Competition:	*TNIC total similarity /1000*	*TNIC number of firms /1000*	*Ln(TNIC Industry LI)*	*TNIC total similarity /1000*	*TNIC number of firms /1000*	*Ln(TNIC Industry LI)*
	(1)	(2)	(3)	(4)	(5)	(6)
*Competition*Top25% in level of fluidity*	1.468**	5.830**	0.868***	1.194*	4.792	0.734**
	(0.582)	(2.727)	(0.318)	(0.658)	(3.212)	(0.322)
*Competition*25-50% in level of fluidity*	0.536	2.319	0.308	0.121	1.158	0.213
	(0.492)	(2.245)	(0.378)	(0.519)	(3.174)	(0.409)
*Competition*50-75% in level of fluidity*	0.829	2.881	0.531	1.006	3.371	0.277
	(0.516)	(2.221)	(0.420)	(0.623)	(2.284)	(0.482)
*Competition*Bottom25% in level of fluidity*	0.357	2.678	0.157	-0.159	-4.261	0.008
	(1.577)	(7.169)	(0.672)	(1.850)	(7.312)	(0.624)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Hansen	0.97	0.98	0.92	0.45	0.45	0.37
Diff Hansen	1.00	1.00	1.00	1.00	1.00	1.00
AR(1)	0.00	0.00	0.00	0.00	0.00	0.00
AR(2)	0.61	0.62	0.65	0.79	0.81	0.82
Observations	1185	1185	1185	1185	1185	1185

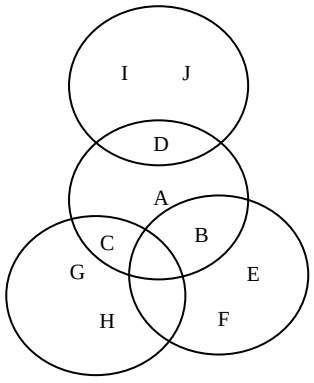
Note: The table reports system GMM regressions using lagged independent variables. The system GMM estimates are Windmeijer corrected for robust standard errors. The values reported for Hansen test are p-values for the null hypothesis of instrument validity. The Diff Hansen reports the p-values for the validity of the additional moment restrictions required by the system GMM. The values reported for AR(1) and AR(2) are the p-values for first and second-order auto correlated disturbances in the first differenced equations. To construct instruments in Ln(CVC/sales) (Ln(CVC/(RD+CVC))), we use lags 2-3 (2) of the levels for competition and patent stock and lag 2 for the others of all the control variables for the transform equation, respectively, and lag 1 of the same variables in differences for the levels equation. Product market fluidities are ranked at a given year for all firms. Year dummies are assumed to be exogenous in this specification. For ease of presentation, we omit results for control variables. *** significant at 1%; ** significant at 5%; * significant at 10%

Additional Instrumentation Based on the TNIC

Our main approach in the paper to addressing endogeneity is based on the Arellano-Bond estimator, reported in Table 4. Here we explore the effects of additional instruments that use the unique network structure of the TNIC. We instrument for the competition faced by a focal firm by the *average competition encountered by rivals of the firm* in that year. For example, consider firm A in Figure A1. The endogenous variable here is the total product similarity faced by firm A in its TNIC, comprising of three other firms B, C and D. We propose that a valid instrument for firm A's total similarity is the average total similarity faced by the firms B, C and D in their own TNICs. The reasoning here is that the average total similarity of firm A's rivals is likely correlated with the similarity faced by firm A, whereas it is unlikely to directly affect the CVC investment of the firm A. The competition faced by rivals of firm A, determined by examining firms E, F, G, H, I and J, is not chosen by firm A and thus not likely to be related to firm A's contemporaneous CVC investments except through its effect on firm A's competition. This additional-degree-of-separation variable is often used in network analysis as an instrumental variable.

To improve the efficiency of our analysis, we also include *product market fluidity* by rivals of the firm as a second instrument for competition (see the earlier section on product market fluidity in this document). Product market fluidity measures the extent to which rivals can move in the product market space and thus the extent to which they pose a competitive threat to a focal firm. Because the extent to which rivals pose a competitive threat is not under the control of a focal firm, it can only affect the focal firm's CVC decisions through its implications for changes in competition, thereby serving as a reasonable instrument. The second-stage results of an IV regression based on these two instruments for competition are presented in Table A2. The control variables are the *standard deviation of monthly stock returns, Ln(Firm size), Ln(Patent stock), Ln(Free cash), Ln(R&D), Ln(Sale growth rate),* and *CVC experience*, (definitions provided in Table 1). The results are qualitatively similar to our main findings. Increases in competition are associated with higher investments by IT firms in CVC.

Figure A1: IV Construction for Industry-level CVC Variable



Each circle represents a given firm's industry. For example, firm A has three firms, B, C, and D, in its own industry. In addition, firm A's direct competitors also have its own industry. Firm B has firms C, E, and F, while firm D has firms I and J, and firm C has firms G and H. Thus, firms can also be indirectly linked to other firms in the product market network.

Table A2: Two-stage Least Squares Model

Panel A: Second-stage regression						
DV	Ln(CVC/sales)			Ln(CVC/(RD+CVC))		
	(1)	(2)	(3)	(4)	(5)	(6)
TNIC total similarity/1000	14.807***			14.505***		
	(4.595)			(4.715)		
TNIC number of firms		23.690***			21.129***	
		(5.409)			(5.649)	
Ln(TNIC Ind LI)			1.268**			1.171*
			(0.599)			(0.630)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1182	1182	1182	1182	1182	1182

Panel B: First-stage regression			
DV	*TNIC total similarity/ 1000*	*TNIC number of firms*	*Ln(TNIC Ind LI)*
Average TNIC total similarity of rivals/1000	0.054		
	(0.039)		
Average TNIC number of firms of rivals		0.616***	
		(0.055)	
Average Ln(TNIC Ind LI) of rivals			0.259***
			(0.025)
Product market fluidity of rivals	0.025***	-0.001	0.077***
	(0.008)	(0.002)	(0.013)
Hansen-Sargen test (p-value)	0.618	0.058	0.001
Controls	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes

Note: The table reports results from 2SLS using lagged independent variables. Standard errors are clustered by firms. For 2SLS, we use means of each competition variable and the product market fluidity of rivals of a focal firm as an IV for each competition variable. For example, to instrument total similarity of a given firm, we use the means of total similarity and product market fluidity of rivals of the firm. For ease of presentation, we omit results for control variables. *** significant at 1%; ** significant at 5%; * significant at 10%

2.2 Using a Selection Model in Panel Data

Since CVC investments are optional and represent an additional investment made by the focal firm in in external innovation spending, we need to account for self-selection. In other words, it is possible that firms decide to invest in CVC in a given year, and then decide the investment amount. To account for this structure, we estimate a sample selection model adapted for panel data (Wooldridge (2010), p. 835). Since this methodology is different from the typical two-stage selection model, we provide a brief explanation. In the first stage, we estimate a Probit model of the CVC decision (CVC yes/no) based on all of our control variables for each year. We include our CVC firms and all other SIC4-based public IT firms in the analysis, totaling 1770 firms. In the second stage, we run the pooled OLS regression using only observations with positive CVC investment. We include the same control variables, their time average, the inverse Mills ratio and interactions of the inverse Mills ratio and year dummies. The time average

refers to the average of each control variable during our study period. The results from the second stage of this analysis are presented in Table A3, and the results obtained are robust, showing the influence of the competition variables on CVC investment.

Table A3: Two-stage Sample Selection Model

	Ln(CVC)			Ln(CVC/(RD+CVC))		
	(1)	(2)	(3)	(4)	(5)	(6)
TNIC total similarity/1000	1.019***			0.899**		
	(0.285)			(0.451)		
TNIC number of firms		4.837***			4.566***	
		(1.256)			(1.747)	
Ln(TNIC Ind LI)			0.180			0.114
			(0.141)			(0.141)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	499	499	499	499	499	499

Note: The table reports results from a selection model in a panel data using lagged independent variables. Standard errors are clustered by firms. For the selection model in a panel data, we first estimate a probit of CVC investment decision on the control variables for each year and then save the inverse Mills ratio for a given firm and year. Next, using only observations with positive CVC investments we run the pooled OLS regression of CVC investment on the same control variables, the inverse Mills ratios, and interactions of inverse Mills ratios and year dummies. For ease of presentation, we omit results for control variables. *** significant at 1%; ** significant at 5%; * significant at 10%

Accounting for Potential Confounding Effects of Technological Regime Changes

One potential confounding effect in our analysis is the possibility that a change in technological opportunities might simultaneously influence competition and CVC activity (Dushnitsky and Lenox 2005a, Levin et al. 1985). It is well known that new technological opportunities or a change in technology regimes can determine innovative activity (Cohen 2010, page 172). In addition, industry subclasses are likely to differ in their level of technological and scientific knowledge (Klevorick et al. 1995). This inter-industry difference can in part explain variation in innovation productivity across industry subclasses. For example, if an industry subclass has greater technological opportunities, it is more likely to attract prominent entrepreneurs starting new ventures. Thus, when there are more innovative ventures in the marketplace, investors are likely to be attracted to and invest more in CVC. Although GMM and IV regressions reported in the paper mitigate this concern, we directly include additional control variables from the economics of innovation literature (Cohen 2010) to test the robustness of our results.

It is likely that industries have different innovative opportunities over time. To control for technological opportunities at the industry level, we first include *industry by year* fixed effects, as shown in columns (1)-(3) of Table A4. Our results again remain robust. In addition, we include a control for the technological opportunities within an industry reflected through innovative output - average citation-weighted patents within a firm's SIC industry (Dushnitsky and Lenox 2005a). As shown in (4)-(6) of Table A4, our results are robust to the addition of these control variables as well, indicating the stability of the results.

We further examine whether our results hold when we account for a financial shock that might affect both competition and CVC investment. Prior studies document that CVC investments, especially by IT firms, peaked around 2000 (Dushnitsky and Lenox 2005b) and that following the Internet bubble era, the IT industry experienced consolidation through corporate restructuring (Park and Mezias 2005). Thus, our results might be driven by these boom and bust periods in financial markets. Year dummies control for financial shocks within the entire IT industry. As a further test, we include a control for the NASDAQ index at the end of each year as a proxy for financial shocks. Our results do not change when we additionally control for the NASDAQ index, as shown in columns (7)-(9) of Table A4.

Table A4: Accounting for Potential Confounding Effects

DV: Ln(CVC/sales)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TNIC total similarity/1000	1.029**			1.210***			1.235***		
	(0.464)			(0.410)			(0.402)		
TNIC number of firms		3.798**			4.554***			4.720***	
		(1.839)			(1.706)			(1.673)	
Ln(TNIC Ind LI)			0.651***			0.653***			0.615***
			(0.220)			(0.190)			(0.191)
Ln(Industry-level number of patents)				0.133	0.116	0.192*			
				(0.108)	(0.109)	(0.108)			
NASDAQ index							0.000	0.000	0.000
							(0.002)	(0.002)	(0.002)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 dummies	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
SIC3*Year dummies	Yes	Yes	Yes	No	No	No	No	No	No
DV: Ln(CVC/RD+CVC)									
TNIC total similarity/1000	0.855**			0.944**			0.973**		
	(0.428)			(0.384)			(0.378)		
TNIC number of firms		3.121*			3.384**			3.589**	
		(1.793)			(1.644)			(1.625)	
Ln(TNIC Ind LI)			0.506**			0.478**			0.439**
			(0.217)			(0.204)			(0.207)
Ln(Industry-level number of patents)				0.156	0.143	0.200*			
				(0.114)	(0.114)	(0.113)			
NASDAQ index							0.001	0.001	0.001
							(0.002)	(0.002)	(0.002)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 dummies	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
SIC3*Year dummies	Yes	Yes	Yes	No	No	No	No	No	No
Observations	1185	1185	1185	1185	1185	1185	1185	1185	1185

Note: The table reports results from OLS using lagged independent variables. Standard errors are clustered by firms for OLS. For ease of presentation, we omit results for control variables. *** significant at 1%; ** significant at 5%; * significant at 10%

2.3 Tests Using Different Dependent Variables

In addition to sales-normalized CVC spending, we also consider three other DVs – the unadjusted annual amount of CVC spending, the number of individual deals by a firm in a year, and a dummy variable for whether the firm invests in CVC or not in a given year. In the first case, it is possible that CVC investments require a certain baseline investment, regardless of the size of the investing firm. Therefore, normalizing the CVC investment by firm sales may not be necessary. In the second case, it is possible for firms to treat CVC investments not in terms of dollar terms but in the number of deals, thereby increasing the odds of successful knowledge. From an options perspective, it is preferable for the investor firm to have multiple investments in several firms, rather than one large investment. We test for this possibility. Finally, it is possible that the firm's propensity to invest in CVC is affected by product market competition, rather than the actual investment amount. This can be tested by using a suitably defined dummy variable as the dependent variable.

The results from Table A5 show consistent results across all three DVs. We observe significant coefficients for competition on the annual amount of CVC, as well as the number of deals, showing that competition induces greater number of deals, consistent with theory in options, as well as supporting the notion that normalizing by sales may not be necessary here. Additionally, we see that competition does indeed enhance the odds of the focal firm investing in CVC, given that the firm is reasonably familiar with the CVC model (recall that the sample includes only those firms that have displayed prior investments in CVC). Broadly, we see support for the argument that product market competition leads to increased CVC investments in the IT industry.

Table A5: Analysis Using Different DVs

DV	Ln(CVC amount)			Ln(number of CVC round financing)			CVC dummy		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TNIC total similarity/1000	0.543***			0.315***			0.148**		
	(0.198)			(0.112)			(0.061)		
TNIC number of firms		2.010**			1.504***			0.552**	
		(0.948)			(0.548)			(0.271)	
Ln(TNIC Ind LI)			0.230***			0.130***			0.041
			(0.080)			(0.042)			(0.030)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1185	1185	1185	1185	1185	1185	1185	1185	1185

Note: The table reports system GMM regressions using lagged independent variables. We divide the data into technology leaders and laggards based on the median of patent stock in a given year within our sample. The system GMM estimates are Windmeijer corrected for robust standard errors. The estimates also pass the Hansen test for instrument validity. To construct instruments, we use lags 2-6 of all the control variables for the transform equation and lag 1 of the same variables in differences for the levels equation. Year dummies are assumed to be exogenous in this specification. For ease of presentation, we omit results for control variables. *** significant at 1%; ** significant at 5%; * significant at 10%

2.4 Using Propensity Score-based Matching Sample

As another way of addressing a potential sample selection bias, we augment our sample with a matched sample of IT firms that are *similar* to those in the sample. We use the propensity score matching methods developed by Rosenbaum and Rubin (1983) and Heckman et al. (1998) to identify the matched sample for the main CVC sample used in the analyses reported in the paper. The matched sample includes public IT firms (using SIC-4 classification) that are similar to the firms in the CVC sample but have not invested in CVC in our observation period. 1770 public IT firms were thus used as potential matches to the CVC sample firms. We use a probit regression with the following firm variables as independent variables - firm size, patent stock, R&D investment, free cash flow, sales growth rate, SIC-3 code, and the standard

deviation of monthly stock returns. The estimation of the probit model (with DV = 1 if the firm invested in CVC in that year and 0 otherwise) provides the propensity score for each firm in a given year. Since the CVC sample is significantly different from the sample of non-CVC IT firms, we use single nearest-neighbor matching with replacement to reduce the bias from the matching procedure (Dehejia and Wahba 2002). This procedure matches each CVC firm with one control firm but allows some control firms to be matched multiple times to firms in the CVC sample. The final sample consists of 1118 firm-year observations for the CVC and 600 firm-year observations for the control sample. The control sample is smaller due to matching with replacement.

We use the system GMM estimation and the results, presented in Table A6, show significant effects of TNIC similarity and TNIC number of firms on CVC spending as well as on the ratio of CVC to (CVC + R&D). The results are robust to augmenting with the propensity score-based matching sample.

Table A6: Using Propensity Score-based Matching Sample

	DV: Ln(CVC)			DV: Ln(CVC/(RD+CVC))		
	(1)	(2)	(3)	(4)	(5)	(6)
TNIC total similarity /1000	0.616***			1.286**		
	(0.183)			(0.614)		
TNIC number of firms /1000		2.348***			3.732*	
		(0.767)			(2.271)	
Ln(TNIC Industry LI)			0.166**			0.174
			(0.073)			(0.277)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Hansen	0.26	0.36	0.15	0.15	0.18	0.11
Diff Hansen	0.79	0.60	0.98	0.84	0.71	0.99
Observations	1718	1718	1718	1718	1718	1718

Note: The table reports system GMM regressions using lagged independent variables. The system GMM estimates are Windmeijer corrected for robust standard errors. The values reported for Hansen (Diff Hansen) test are p-values for the null hypothesis of instrument validity. To construct instruments, we use lags 2-6 of all the control variables for the transform equation and lag 1 of the same variables in differences for the levels equation. Year dummies are assumed to be exogenous in this specification. *** significant at 1%; ** significant at 5%; * significant at 10%

References

Baumol, W. J., J. C. Panzar, R. D. Willig. 1983. *On the theory of perfectly contestable markets*. Bell Telephone Laboratories.

Carnoy, M., M. Castells. 1997. Labour markets and employment practices in the age of flexibility: A case study of Silicon Valley. *International Labour Review*. **136**(1) 27.

Chamberlin, E. H. 1933. *The Theory of Monopolistic Competition: A Re-orientation of the Theory of Value*. Harvard Business Press, Boston.

Cohen, W. M. 2010. Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*. **1** 129-213.

Dehejia, R. H., and Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* **84**(1) 151–161.

Dushnitsky, G., M. J. Lenox. 2005a. When Do Firms Undertake R&D by Investing in New Ventures? *Strategic Management Journal*. **26**(10) 947-965.

Dushnitsky, G., M. J. Lenox. 2005b. When do incumbents learn from entrepreneurial ventures? Corporate venture capital and investing firm innovation rates. *Research Policy*. **34**(5) 615-639.

Heckman, J. J., H. Ichimura, P. Todd. 1998. Matching as an econometric evaluation estimator. *The review of economic studies*. **65**(2) 261-294.

Hoberg, G., G. Phillips. 2010. Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *Review of Financial Studies*. **23**(10) 3773-3811.
Hoberg, G., G. Phillips. 2015. Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy, forthcoming.
Hoberg, G., G. Phillips, N. Prabhala. 2014. Product market threats, payouts, and financial flexibility. *The Journal of Finance*. **69**(1) 293-324.
Hotelling, H. 1929. Stability in Competition. *The Economic Journal*. **39**(153) 41-57.
Jaworski, B. J., A. K. Kohli. 1993. Market Orientation: Antecedents and Consequences. *The Journal of Marketing*. **57**(3) 53-70.
Klevorick, A. K., R. C. Levin, R. R. Nelson, S. G. Winter. 1995. On the sources and significance of interindustry differences in technological opportunities. *Research Policy*. **24**(2) 185-205.
Levin, R. C., W. M. Cohen, D. C. Mowery. 1985. R & D Appropriability, Opportunity, and Market Structure: New Evidence on Some Schumpeterian Hypotheses. *The American Economic Review*. **75**(2) 20-24.
McAfee, A., E. Brynjolfsson. 2008. Investing in the IT That Makes a Competitive Difference. *Harvard Business Review*. **86**(7/8) 98-107.
Park, N. K., J. M. Mezias. 2005. Before and after the technology sector crash: the effect of environmental munificence on stock market response to alliances of e-commerce firms. *Strategic Management Journal*. **26**(11) 987.
Pavlou, P. A., O. A. El Sawy. 2006. From IT leveraging competence to competitive advantage in turbulent environments: The case of new product development. *Information Systems Research*. **17**(3) 198-227.
Rosenbaum, P. R., D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*. **70**(1) 41-55.
Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*. **34**(1) 1-47.
Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.