Data Analytics Supports Decentralized Innovation Communities

Lynn Wu, Bowen Lou, Lorin Hitt

The Wharton School

Abstract

We examine the relationship between data analytics and innovation, focusing on how the benefits of analytics may differ depending on how firms organize their innovative activities. Our analysis draws on prior work that has measured firm analytics capability using detailed employee-level data and matches these data to metrics on innovation structure that are constructed by analyzing intra-firm inventor networks. We apply community detection algorithms to these inventor networks to determine whether a firm has a centralized or decentralized innovation structure. In a panel of large firms from the years 1988 to 2013, we find that firms with a decentralized innovation structure have a greater demand for analytics skills and receive greater productivity benefits from their analytics capabilities, consistent with a complementarity between analytics and decentralized innovation. Furthermore, we find the complementarity is strongest for innovation involving the recombination of existing technologies. This suggests that data analytics can alleviate a weakness in decentralized innovation structures by allowing firms to search for knowledge broadly and link distant ideas to create new ones, an advantage previously available primarily to centralized innovation.

The growth in data availability and associated analytics technology has facilitated a substantial increase in the use of data for managerial and operational decision-making (Brynjolfsson et al. 2011, Muller and vom Brocke 2016). However, not all firms have been able to take advantage of the opportunities created by the increased availability of data; a recent study revealed 59% of firms fail to use advanced analytics despite having the necessary data (Bradstreet 2017). One possible explanation is that not all firms have strategies or other organizational structures that are complementary to the use of analytics. Prior work has suggested that IT-organizational complementarities are a particularly important determinant of both the demand for and benefits from IT investment (e.g. Bharadwaj et al. (2007), Bresnahan et al. (2002), Nagle (2017), Pang (2016), Tafti et al. (2013)). This relationship between IT investments and benefits is likely to hold true for investments in analytics capabilities, although the complements critical to leveraging analytics appear to differ from those for leveraging general IT.

In this paper, we focus on the potential for analytics to foster innovation activity. Prior work has suggested that IT has a role in driving the innovation process (Gao and Hitt 2012, Kleis et al. 2012) and these relationships may be even more important going forward given the recent apparent slowdown in both innovation output and the associated economic growth (Bloom et al. 2017). There may be an especially close tie between data analytics and innovation because the acquisition of data can facilitate the creation of knowledge (Joshi et al. 2010, Majchrzak and Malhotra 2016). In addition, prior literature has documented a close relationship between the organization of the innovation process – particularly the choice of centralization versus decentralization (Argyres and Silverman 2004, Siggelkow and Rivkin 2006) – and the nature and outcome of innovation activity, suggesting an important role for organizational complementarities.

Our analysis is motivated by the observation that successful firms can structure their innovation activity in substantially different ways and that the choice of structure may potentially affect the ability to gain benefits from changes in the firm environment such as an increase in available data and associated analytics tools. For instance, Google and Apple, two of the most innovation-driven firms in the information

economy,[1] organize their innovation processes differently, as shown in their respective networks of patent co-authorship (Figure 1). Google displays a more decentralized innovation structure with many small groups of loosely connected inventors and some larger clusters with ties extending throughout firm. In such a structure where information is often hard to transfer across organizational boundaries (Von Hippel 1994), individual groups directly involved in firm operations may be better at understanding the nature of operational problems and at identifying and implementing solutions. By contrast, at Apple, much of the innovation output is centralized in a few tightly-linked clusters whose connections outside their groups are limited. This centralized structure has the advantage of enabling the conception and development of foundational technologies that are applicable beyond the confines of a specific group or the immediate needs of current customers and local markets. Accordingly, it can facilitate the search for external information and technologies which may not be of direct interest to any particular internal group (Argyres and Silverman 2004).

The structural difference between Apple and Google appears to be associated with differences in the results of innovation. Apple is known for the creation of novel, breakthrough products, while Google has been most successful in generating a steady stream of incremental improvements to their search and targeted advertising technology. As data availability has grown exponentially and data analytics is increasingly adopted by firms, we ask whether a centralized or decentralized innovation structure is better suited for leveraging the new capabilities that analytics can bring. We hypothesize that, although analytics may provide benefits to all firms in their innovation processes, decentralized structures, such as those found at Google are complementary to analytics capability. Therefore, firms that combine analytics with decentralized innovation are likely to receive greater benefits than firms that combine analytics with centralized innovation. A main driver for the complementarities relationship is that analytics capabilities can mitigate a central weakness associated with decentralization: the lack of search capabilities for

[1] According to 2016 data, Google and Apple both produced many patents, ranked 5th and 12th respectively in the number of patents granted by the US patent and trademark office in 2016. (http://www.ipo.org/wp-content/uploads/2017/05/2016_Top-300-Patent-Owners.pdf)

acquiring diverse knowledge from many different areas (Argyres and Silverman 2004). By collecting digital traces from a variety of business processes and user behaviors both inside and outside of the department or functional area, data analytics can help firms assimilate information from divergent sources that is necessary to generate novel solutions. This effort is accelerated by recent advances in data analytics such as deep learning and machine learning that have become increasingly effective in discovering hidden patterns and ideas across different domains. For example, IBM's Watson digested 23 million medical papers across many different disciplines to find information related to a tumor suppressor known as p53 that is associated with half of all cancers. In a short amount of time, Watson was able to identify six previously-unknown proteins that interact with p53, a feat that would have taken researchers more than 6 years to accomplish (Chen et al. 2016), a feat that can potentially generate substantial financial return (Zafar et al. 2013). Essentially, to find these proteins, Watson has employed data analytics to substantially reduce the cost of conducting a broad search and then linking distant innovation areas to create new solutions.

Similarly, it has been noted that as a product's design matures, informal communication channels that support the development process become deeper and narrower, reducing the likelihood of knowledge sharing outside each group and limiting opportunities to link innovative ideas across areas (Henderson and Clark 1990). Data analytics can mitigate this disadvantage to some extent by automating the search of patterns across innovations among communities outside of the immediate group, and thereby facilitate the transfer and use of knowledge across products and organizational boundaries. Thus, we expect data analytics and decentralized innovation structures are complementary and their joint adoption is more likely to generate innovations that recombine technologies which exist both inside and outside of one's immediate group.

The relationship between data analytics and centralized innovation is less clear than it is for data analytics and decentralized innovation. On the one hand, the preceding arguments suggest that analytics can support the gathering of external information which could benefit innovation output generally. On the other hand, there may be limits to the benefits of analytics for centralized structures for at least two reasons. First, centralized structures have existing mechanisms for sharing internal knowledge and gathering

external information, limiting the marginal benefit in these areas for any new technology (including analytics). Second, centralized structures are disproportionately used in firms engaged in novel innovation (Arora et al. 2014, Siggelkow and Rivkin 2006), which has less to gain from the current state of analytics technology since there is often little or no available data for novel or breakthrough products because they do not have existing producers or customers prior to their initial deployment. Despite these critical knowledge sharing constraints in novel innovation, it has been noted that much of the relevant information is tacit and shared through informal communications channels (Avery and Norton 2014), thus making it less amenable to digitization or automated analysis. This further supports the argument suggesting a complementarity between decentralization and data analytics: a strong connection between the centralization of the innovation process and novel innovation, and a weak connection between analytics and novel innovation.

We use network analysis of patent co-authorship relationships to construct an intra-firm patent network for each firm in our sample between 1988 and 2013 with one network for each year. A node in a network represents an employee inventor and an edge (link) indicates the presence of one or more coauthored patents between two inventors working in the same firm. To measure decentralization, we apply machine learning-based community detection algorithms (Multilevel) to these networks (Blondel et al. 2008) to identify distinct innovation communities and calculate a Herfindahl-based metric to measure how widely the innovations communities are dispersed for the firm (Appendix A). This dispersion metric differs from previous measurements in two important ways. Earlier approaches measure the degree of decentralization through formal hierarchies (e.g., organizational charts) but it is difficult to extend these approaches to large scale analyses because internal structures are extremely difficult to observe. Our approach also has advantages over other metrics based on organizational relationships because it is entirely data driven – the firm's innovation structure is identified by the interaction of inventors rather than (potentially) idiosyncratically-defined organizational units. Moreover, our metric is not subject to the criticism that the informal and formal structures of organizations are different, which has been an important observation in the knowledge management literature (Cross et al. 2001).

Our measures of innovation output are also based on patent data consistent with prior work on R&D productivity (Griliches 1990, Hall et al. 2001). To measure recombination in innovation, we use a Herfindahl concentration index, calculated based on the distribution of different three-digit main classes of prior art that are cited by the focal patent (Hall et al. 2001). For our analysis, it is important to distinguish reuse of existing combinations from the creation of new combinations (Akcigit et al. 2013). We measure this difference by segmenting innovations into existing combinations and new combinations. These can be measured at both the firm level (combination is new to the firm, but other firms have done similar combinations) or at a global level (an original combination that no one else have created). We also measure novelty: whether an innovation involves the creation of a new technology class that can be used as the basis for future combinations. This also can be measured at a firm or global level. Thus, we have a total of 6 ways to classifying innovation along two dimensions: (1) novel technology, reuse and new recombination, and (2) local (firm) versus global (Appendix B).

To measure a firm's data analytics capabilities, we use six million resumes from 1988 to 2007, and 3.7 million job reviews from 2008 to 2013 to calculate the total number of employees possessing data analytics skills. We apply natural language processing techniques on free-form text (when available) and job titles to identify the analytics skills of each employee and aggregating all employees with data analytics skills for each firm in each year after adjusting for a sampling rate (based on the fraction of employees found in the data for each firm) (Appendix C). We then link these data to financial metrics such as physical assets, employees, and sales using the Compustat Industrial Annual files. The summary statistics for the financial variables, data analytics and patent-related variables are shown in Table 1. Trends in our data analytics measure are shown in Figure 2.

Our primary analysis tests for complementarities between data analytics and innovation structure using (1) correlations (adoption or demand equations) and (2) performance differences (productivity equations). Although the demand equations have the advantage that they are relatively simple and provide the greatest power if firms are matching complementary practices optimally, they have the disadvantage that the simplicity makes them vulnerable to unobserved heterogeneity. In addition, such an analysis will tend to

understate the strength of complements if not all firms are endowed with or able to change to the optimal match between complements. The productivity test has the advantage of a direct tie to a relevant firm outcome (performance), and the effects are most powerful statistically when not all firms have found the optimal match, which may be likely for a relatively new change in business practices. Over time, as the complementarities system diffuses to other firms, the correlation would increase but the productivity premium would decrease because the relative advantage of using a complementary system diminishes as its adoption spreads.

For econometric identification, we take the view that data skills are rapidly changing (and potentially endogenous) while firm capabilities to pursue specific innovation practices are quasi-fixed (Hannan and Freeman 1984, Lam 2005) and therefore exogeneous, an assumption that has also been made previously for other types of organizational complements to technology (Autor et al. 2003, Bresnahan et al. 2002, Brynjolfsson and Hitt 1996, Milgrom and Roberts 1990). The rapid diffusion of analytics technology and a relatively recent, exogenous shift in the amount of data available to firms enable us to observe the effects of data analytics on different firm practices related to innovation, even those practices that are not optimally configured to benefit from analytics capabilities. Table 2 shows the demand for data analytics when firms have a decentralized innovation structure. After controlling for firm and year fixed effects, as well as the level of R&D and patent activity, the demand for data analytics is positively correlated with a decentralized innovation structure. Using four different popular community detection algorithms to measure decentralization, we find similar results in both magnitude and direction providing robust evidence that decentralization is complementary to the demand for data analytics skills. In the subsequent analysis, we use the Multilevel algorithm to measure decentralization.

If data analytics and the decentralization of innovation are complementary, firms that possess both should experience a greater return than firms that have only one of the two complements. We first estimate a baseline regression specification in firm-level fixed effects that shows that our regression analysis yields similar estimates to those found in prior work for IT labor, as well as other productivity inputs (Table 3, Column 1). The regression continues to show reasonable properties when we add metrics for analytics and

innovation (data analytics skills, number of inventors, and number of patents) (Columns 2-4). In general, we find data analytics is positively associated with productivity while the direct effect of patents and inventors is indistinguishable from zero. Our key result (Column 4) shows that while having decentralized innovation structures is negatively correlated with productivity (albeit not significantly so; it is perhaps due to the impact of fixed effects with a slow-changing quasi-fixed factor), the interaction with data analytics is positive. On average, a firm that is more decentralized than the average (one standard deviation above in our decentralization metric) receives a 2.7% increase in productivity for every standard deviation increase in data skills above the mean.

To examine whether our results are affected by the potential endogeneity of analytics skills we repeat our main analysis using 2SLS and GMM. We treat innovation structure in firms as quasi-fixed, and instrument data analytics with a metric based on the adoption of enterprise systems and the flow of analytics skills in neighboring firms in the labor supply network (i.e., firms that the focal firm hires from). Our instruments are motivated by the argument that adoption of large-scale IT innovations changes the relative availability of skills in the market which affects the cost of acquiring data talent. This shift, driven by choices external to the firm, is not directly reflected in free cash flow or management characteristics that would cause data skills to be influenced by prior performance (Appendix D). The first-stage regression is shown in Column 5 of Table 2 and the associated first-stage F-statistics are above the threshold needed to pass the weak instrument test. In the GMM analysis, we also add additional instruments derived from information within the panel. The results on the key coefficients continue to be consistent and, if anything, are somewhat larger in the 2SLS regression (although not statistically different) suggesting that potential endogeneity of data analytics is not leading to an upward bias of our complementarities measure. In addition, we also find some evidence (especially in the GMM regressions) of a negative relationship between a decentralized community structure and productivity consistent with the prior innovation literature (Argyres and Silverman 2004). Notably, our complementarities results suggest that the presence of complementary analytics skills may offset some or perhaps all of the disadvantages of decentralized innovation.

To explore the complementarities between different types of innovation, we use the same variables from our previous models (data analytics, decentralization, their interactions, and other controls) to determine their relationship with the production of different types of patents (see Table 4). First, we find that none of our metrics have a substantial effect on the production of completely new knowledge overall (Column 1), but our hypothesized complementarity is present for within-firm novel innovation, indicating that analytics can enable the decentralized innovation structure to better acquire knowledge about novel technologies outside the firm (Column 2). Next, we examine the effect of analytics-decentralization complementarities on innovation through the recombination of existing ideas. First, we do not distinguish new combinations from refinements of existing combinations, and we find that the combination of data analytics and decentralized innovation facilitate recombination in innovation (Column 3, Column 4). After distinguishing new combinations from the reuse of existing combinations, we find that our data-decentralization complementarity facilitates both combinations that are new to the firm (Column 5) and those new to the overall market (Column 6). Consistent with the prior literature, we also find that decentralization promotes innovations that constitute the reuse of existing combinations, but there appears to be no particular benefit of data analytics (direct or complementary) to this type of innovation (Columns 7 & 8).

To further explore the extent to which decentralization and data analytics can support the ability to integrate available external information, we use an alternative measure to capture the integration of new ideas into a firm's patent portfolio. Following Lin et. al. (2016), we compute the proportion of citations supporting a new patent that are made to a firm's own prior art. In Table 5, we repeat our analysis with a dependent variable that is defined with different thresholds for citing a firm's own prior art. Column 1 begins with innovations that entirely cite a firm's own patents, and repeat this analysis lowering the threshold by increments of 10% per column. We observe that complementarities between data analytics and decentralization have a modest effect when the patents cite the firm's own prior art heavily (as shown in Column 1 and 2) and that this complementary effect increases as the amount of new knowledge used in the patent rises (that is, when the share of prior art citations drops to about 60%). However, when the patents

substantially deviate from the firm's own prior art, the interaction effect drops and even turns negative when the percentage of prior art cited dips below 30% (Columns 6-10). Although each individual estimate for the interaction from Columns 6-10 is not statistically significant, the size of the effect is greater when the prior art cited is at 60% than when the prior art accounts for 10% ($p<0.05$). These results suggest that data analytics and decentralized innovation are complements for generating innovations that lie somewhere between incremental improvement in a firm's own stock of innovations and innovations that are radically different. This is consistent with our earlier results that show the complementarities are strongest when data has the highest potential marginal value—settings where there is potentially useful external information that can be combined with a firm's own stock of knowledge, but not when the innovation is completely new or mostly incremental.

Finally, all our prior metrics focus on analytics of the firm broadly across any business purpose, which raises the possibility that our results are driven by the characteristics of data-intensive firms rather than by the specific use of analytics to support the innovation process. To examine whether there exists a direct link between analytics and innovators, we repeat our main analyses focusing specifically on the data skills of named inventors. While named inventors are only a subset of all employees with data skills and matching limits the number that we can identify, we can still conduct an exploratory analysis where we divide analytics skills between inventors and other employees who also have data skills. Overall, we find that this narrower measure, despite the likely presence of considerable measurement error, behaves similarly to our broader data skill measure. The correlation test results are straightforward: decentralization is associated with a higher level of data skills in innovators (Table 6). The productivity test results are also similar. Furthermore, the interaction to data skills for non-inventors show a significantly positive effect, and the magnitude is substantially greater than the measure for inventors (see Table 7, column 2). This result suggest that while the complementarities associated with data analytics come from both types of employees, complementarities effect is stronger for employees who provide infrastructure and extensive support to the inventors than the inventors who have analytics skills themselves. As a final note, we repeat our analysis

of types of innovation using this more narrowly-defined measure of data skills and find directionally to be consistent results, but in general the coefficient estimates are too imprecise to formulate strong conclusions.

Overall, this pattern of results is consistent with the idea that data analytics enables firms engaged in decentralized innovation to better integrate external information and to reuse novel information from within the firm. This assistance from data analytics is significant because both integrating external information and reusing novel information are known disadvantages of decentralized structures. However, data analytics provide limited help in improving completely novel innovation or known combinations of existing technologies. This finding also appears reasonable because both of these situations are where data may have limited marginal benefit either because there is no data to integrate (in the case of a completely novel innovation) or such data already exists locally (in the case of reuse). Similarly, we find the optimal use of data analytics for a decentralized innovation structure is at the point where 60% of the prior art citations cite a firm's own prior art, consistent with the idea that analytics is especially useful when the value of combining existing information is high. When we separate data skills for the named inventors and other data-skilled employees, we find that the complementarities effect persists and is especially strong for data skills in non-inventors.

As data availability has grown exponentially and data analytics is increasingly adopted by firms, our results suggest that the diffusion of analytics skills in the short run will favor the innovation process for some firms and some types of innovations over others; in the longer term, it suggests that firms can potentially benefit by shifting to a more decentralized innovation structure if they can also benefit from the types of innovations these structures produce. With data analytics becoming even more important over time, especially from the advances in artificial intelligence and new ways of gathering data (e.g. using robotics to capture optimal machine toolpaths from the motion of human laborers), a firm's innovation structure would also likely evolve substantially to take advantage of the capabilities these technologies could bring. Those firms that can best match their existing capability or quickly adjust to acquire such capabilities would be likely to benefit substantially from using data analytics.

Table 1. Summary Statistics

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
ln(Capital)	11,033	5.212	2.306	-0.209	13.05
ln(R&D)	10,824	2.788	1.987	-3.997	9.513
ln(Sales)	11,033	5.890	2.186	-4.224	13.00
ln(Materials)	11,033	5.487	2.000	-2.315	12.74
ln(IT labor)	11,033	4.054	1.673	0	11.68
ln(Other labor)	11,033	7.301	2.090	0	14.60
ln(Data skills)	11,033	4.080	3.531	0	13.22
ln(Recombination)	11,033	0.326	0.109	0	0.693
ln(Novelty)	11,033	0.0118	0.00687	0	0.0937
ln(Community dispersion: multilevel)	11,033	0.526	0.158	0	0.690
ln(Number of inventors)	11,033	3.214	1.222	1.609	8.450

Correlations

	1	2	3	4	5	6	7	8	9	10
1. ln(Sales)	1									
2. ln(Materials)	0.958	1								
3. ln(Capital)	0.911	0.913	1							
4. ln(IT labor)	0.545	0.553	0.532	1						
5. ln(Other labor)	0.934	0.899	0.908	0.544	1					
6. ln(Dispersion)	0.0668	0.0787	0.0734	0.0411	0.0760	1				
7. ln(Inventors)	0.206	0.239	0.221	0.151	0.200	0.668	1			
8. ln(Data skills)	0.469	0.465	0.446	0.345	0.485	0.0220	0.126	1		
9. ln(R&D)	0.405	0.467	0.408	0.353	0.374	0.194	0.383	0.271	1	
10. ln(Recomb)	-0.0733	-0.0593	-0.0917	-0.0559	-0.0899	0.150	0.156	0.0726	0.0862	1
11. ln(Novelty)	-0.247	-0.208	-0.202	-0.0868	-0.211	0.0297	0.0368	0.0112	-0.0082	0.311

Table 2. The Correlation Test: Data Analytics Skills and Innovation Community Structure

DV: ln(Data skills)	(1)	(2)	(3)	(4)	(5)
Model	FE	FE	FE	FE	FE
ln(Sales)	0.553***	0.555***	0.551***	0.552***	0.573***
	(0.0881)	(0.0883)	(0.0880)	(0.0880)	(0.0976)
Patents	0.190	0.185	0.191	0.190	0.142
	(0.127)	(0.129)	(0.127)	(0.127)	(0.142)
ln(R&D)	0.121*	0.125**	0.120*	0.120*	0.111*
	(0.0638)	(0.0638)	(0.0638)	(0.0638)	(0.0643)
ln(Community dispersion: multilevel)	0.985**				0.979**
	(0.463)				(0.426)
ln(Community dispersion: infomap)		0.576			
		(0.393)			
ln(Community dispersion: leading eigenvector)			1.104**		
			(0.448)		
ln(Community Dispersion: fastgreedy)				1.049**	
				(0.452)	
Avg. Neighbor ERP					0.725**
					(0.356)
Avg. Neighbor HCM					0.699
					(0.457)
Avg. Neighbor Data					0.234
					(0.820)
Lagged ERP					-0.0660
					(0.342)
Lagged HCM					-0.181*
					(0.104)
Observations	11,033	11,033	11,033	11,033	9,329
R-squared	0.683	0.683	0.684	0.683	0.691
# of firms	1,423	1,423	1,423	1,423	1,338
Industry	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES

Note: Column 1-4 are results using different community detection algorithms. Column 5 includes instrumental variables for data analytics, and can thus be interpreted as the first stage results in 2SLS estimation used in Column 5 of Table 2. Robust (clustered by firm) standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3. The Productivity Test: Data Analytics Skills and Innovation Community Structure

DV: ln(Sales)	(1)	(2)	(3)	(4)	(5)	(6)
Model	FE	FE	FE	FE	2SLS	GMM/IV
ln(Materials)	0.594***	0.593***	0.593***	0.594***	0.564***	0.620***
	(0.0238)	(0.0239)	(0.0239)	(0.0242)	(0.0265)	(0.0334)
ln(Capital)	0.102***	0.102***	0.102***	0.104***	0.106***	0.00274
	(0.0202)	(0.0202)	(0.0202)	(0.0202)	(0.0215)	(0.0309)
ln(IT labor)	0.0110**	0.0108**	0.0112**	0.0114**	-0.00639	-0.0122
	(0.00549)	(0.00548)	(0.00551)	(0.00555)	(0.00855)	(0.0131)
ln(Other labor)	0.326***	0.325***	0.325***	0.325***	0.319***	0.439***
	(0.0244)	(0.0245)	(0.0245)	(0.0244)	(0.0315)	(0.0344)
ln(Emp w/ college+)	0.00215	0.00178	0.00180	0.00177	0.000403	-0.00939
	(0.00261)	(0.00265)	(0.00265)	(0.00265)	(0.00410)	(0.00770)
ln(Data)		0.00269*	0.00274*	-0.00443	-0.00455	-0.0123
		(0.00159)	(0.00159)	(0.00294)	(0.0407)	(0.00889)
Patents			-0.00592	-0.00552	-0.000703	1.19e-05
			(0.00833)	(0.00868)	(0.0104)	(0.00471)
ln(# of inventors)				-0.00998	-0.0117	0.0225
				(0.0103)	(0.00948)	(0.0140)
ln(Community dispersion)				-0.0380	-0.116	-0.324***
				(0.0588)	(0.102)	(0.105)
Data X dispersion				0.0266**	0.0581*	0.0716**
				(0.0109)	(0.0300)	(0.0309)
Observations	11,033	11,033	11,033	11,033	9,310	9,329
R-squared	0.988	0.988	0.988	0.988	0.798	
# of firms	1,423	1,423	1,423	1,423	1,319	1,338
Industry	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES

Note: (1) The Multilevel algorithm is used to measure community dispersion. We replicated Column 4 of Table 3 with other measurements of decentralization. They show similar results (0.0210 for infomap, .0260 for leading eigenvector and .0237 for fastgreedy). All are significant at p<.05 (against the null hypothesis of zero).

(2) Five instrumental variables are used: (1) The average data employees in neighboring firms; (2-3) the average number of neighbors that adopted an ERP or HCM systems; (4-5) lagged values for ERP and HCM variables in what is calculated for (2-3). The first-stage F-statics passes the weak instrument test: Cragg-Donaldson F(10, 7942)=17.40. The Sargan over-identification test also passed: (χ^2 =10.007, p <=.188).

(3) We use the Blundell and Bond (1998) SYS-GMM estimator which derives additional instrument using the lagged level and differences from production inputs. This method was originally developed specifically for micro-productivity applications, especially for measuring the productivity of R&D. We used the 3rd and 4th lags of the production inputs as instruments in addition to the external instruments used in 2SLS. The Arellano-Bond test for AR(2) autocorrelation in first differences suggests that serial correlation is not a concern in the first-differencing equations (p=0.908). Neither the Hansen test of over-identification restrictions (p=0.202) nor the difference-in-Hansen test of the system GMM instruments (p=0.206) rejects the null that the instruments are uncorrelated with the error term, ensuring the validity of the instruments used in the GMM estimation.

(4) Robust (clustered by firm) standard errors in parentheses

(5) *** p<0.01, ** p<0.05, * p<0.1

Table 4. Data Analytics Skills, Community Dispersion, and the Types of Innovation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DV	ln(new-tech)	ln(new-tech)	ln(recomb)	ln(recomb)	ln(new comb)	ln(new comb)	ln(reuse)	ln(reuse)
Model	Global FE	Local FE	Global FE	Local FE	Global FE	Local FE	Global FE	Local FE
ln(Data)	-0.00197	-0.0263**	-0.00305	-0.00371*	-0.0339**	-0.0258*	0.00402	0.00465
	(0.00227)	(0.0112)	(0.00208)	(0.00211)	(0.0144)	(0.0141)	(0.0119)	(0.0111)
ln(Community dispersion)	-0.00454	0.722***	0.119***	0.116***	1.007***	0.780***	0.830***	0.556***
	(0.0153)	(0.129)	(0.0208)	(0.0209)	(0.158)	(0.159)	(0.139)	(0.130)
Data X dispersion	0.00462	0.0992**	0.0125*	0.0138*	0.125**	0.0899*	0.0115	0.0248
	(0.00822)	(0.0419)	(0.00704)	(0.00717)	(0.0534)	(0.0523)	(0.0465)	(0.0443)
ln(Sales)	0.00213	0.0628**	-0.00196	-0.00136	0.0855***	0.0967***	0.106***	0.0981***
	(0.00355)	(0.0274)	(0.00408)	(0.00405)	(0.0323)	(0.0303)	(0.0289)	(0.0260)
ln(Emp)	0.0103*	0.150***	0.00497	0.00539	0.215***	0.211***	0.212***	0.194***
	(0.00617)	(0.0348)	(0.00502)	(0.00494)	(0.0467)	(0.0461)	(0.0396)	(0.0359)
ln(Emp w/ college+)	0.000190	0.00610	-0.000342	-0.000524	0.00749	0.00284	0.00121	1.51e-05
	(0.000577)	(0.00615)	(0.00101)	(0.00104)	(0.00761)	(0.00743)	(0.00638)	(0.00607)
Observations	11,033	11,033	11,033	11,033	11,033	11,033	11,033	11,033
R-squared	0.485	0.709	0.623	0.608	0.825	0.842	0.804	0.809
# of firms	1,423	1,423	1,423	1,423	1,423	1,423	1,423	1,423
Industry	YES	YES	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES	YES	YES

Note: The Multilevel algorithm is used to measure community dispersion. Robust (clustered by firm) standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Innovation Quality Compared to Firm's Own Patent Stock from the Previous 5 years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DV	Prior 100	Prior 90	Prior 80	Prior 70	Prior 60	Prior 40	Prior 30	Prior 20	Prior 10
Model	FE	FE	FE	FE	FE	FE	FE	FE	FE
ln(Data)	-0.0159	-0.0193*	-0.0268**	-0.0342***	-0.0402***	-0.00414	0.00156	0.00430	0.00838
	(0.0103)	(0.0106)	(0.0119)	(0.0125)	(0.0134)	(0.0136)	(0.0134)	(0.0128)	(0.0111)
ln(Community dispersion)	0.0807	0.0588	0.116	0.156	0.276*	1.224***	1.163***	0.942***	0.786***
	(0.139)	(0.140)	(0.150)	(0.148)	(0.156)	(0.157)	(0.155)	(0.149)	(0.138)
Data X dispersion	0.0704*	0.0851**	0.111**	0.133***	0.157***	0.00702	-0.0153	-0.0222	-0.0449
	(0.0373)	(0.0382)	(0.0430)	(0.0458)	(0.0487)	(0.0511)	(0.0502)	(0.0476)	(0.0423)
ln(Sales)	0.0155	0.0106	0.00462	0.00905	0.0161	0.110***	0.0891***	0.0809**	0.0734***
	(0.0210)	(0.0214)	(0.0235)	(0.0257)	(0.0286)	(0.0339)	(0.0335)	(0.0322)	(0.0281)
ln(Emp)	0.115***	0.124***	0.151***	0.158***	0.177***	0.246***	0.248***	0.230***	0.193***
	(0.0281)	(0.0292)	(0.0345)	(0.0367)	(0.0401)	(0.0460)	(0.0457)	(0.0429)	(0.0382)
ln(Emp w/ college +)	-0.00488	-0.00456	-0.00555	-0.00540	-0.00528	0.00764	0.00705	0.00833	0.00401
	(0.00521)	(0.00521)	(0.00584)	(0.00627)	(0.00650)	(0.00715)	(0.00712)	(0.00720)	(0.00675)
Observations	11,033	11,033	11,033	11,033	11,033	11,033	11,033	11,033	11,033
R-squared	0.791	0.789	0.792	0.796	0.804	0.813	0.809	0.807	0.798
Industry	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The Multilevel algorithm is used to measure community dispersion. Robust (clustered by firm) standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6. The Correlation Test using Data Analytics Skills of Inventor

ln(Inventor with data skills)	(1)
Model	FE
ln(Community dispersion)	0.111*
	(0.0602)
ln(Sales)	0.0327*
	(0.0180)
Patents	0.156
	(0.0991)
Observations	9,120
R-squared	0.592
Industry	YES
Year	YES

Note: The Multilevel algorithm is used to measure community dispersion. Robust (clustered by firm) standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table 7. The Productivity Test using Data Analytics Skills of Inventors

ln(Sales)	(1)	(2)
Model	FE	FE
ln(Materials)	0.542***	0.541***
	(0.0256)	(0.0256)
ln(Capital)	0.0952***	0.0956***
	(0.0188)	(0.0188)
ln(IT labor)	-0.000990	-0.00116
	(0.00577)	(0.00578)
ln(Other labor)	0.344***	0.342***
	(0.0236)	(0.0236)
ln(Employees with college +)	-0.00111	-0.00141
	(0.00222)	(0.00222)
ln(Data inventor)	-0.321	-0.295
	(0.196)	(0.197)
ln(Data non-inventor)		-0.00449
		(0.00296)
Patents	0.00987	0.0101
	(0.00713)	(0.00739)
ln(Number of inventors)	-0.00545	-0.00645
	(0.00976)	(0.00976)
ln(Community dispersion:)	-0.000377	-0.0593
	(0.0441)	(0.0533)
(Data skills & inventor) X dispersion	0.144*	0.132
	(0.0871)	(0.0873)
(Data skills & non-Inventor) X dispersion:		0.0268**
		(0.0116)
Observations	9,120	9,120
R-squared	0.990	0.990
Industry	YES	YES
Year	YES	YES

Note: The Multilevel algorithm is used to measure community dispersion. Robust (clustered by firm) standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

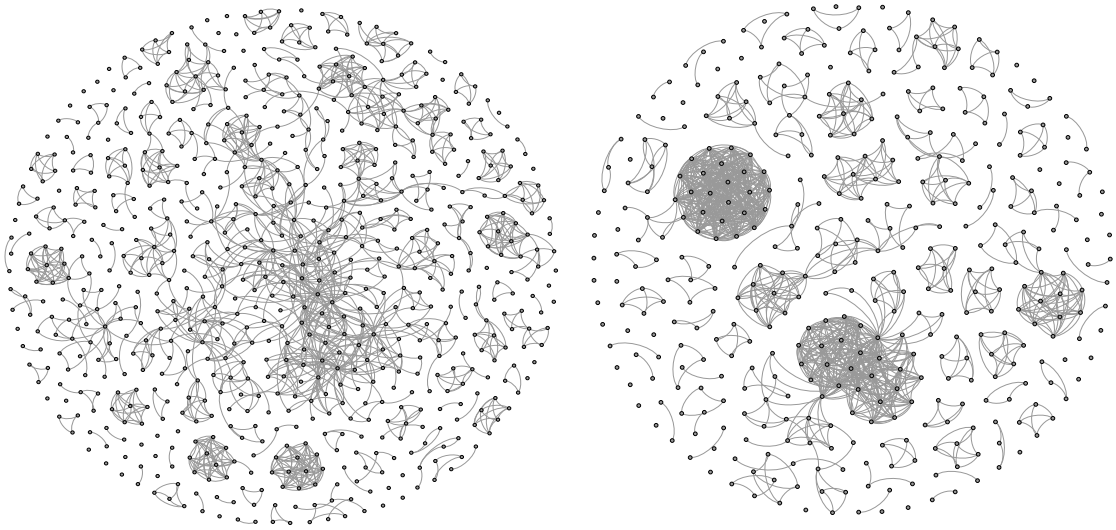


Figure 1. Comparing innovation structures between Google (left) and Apple (right). Each node on the graph is a particular inventor, while the edges are inventors appearing on the same patent. This graph is generated from the patent data in our sample.

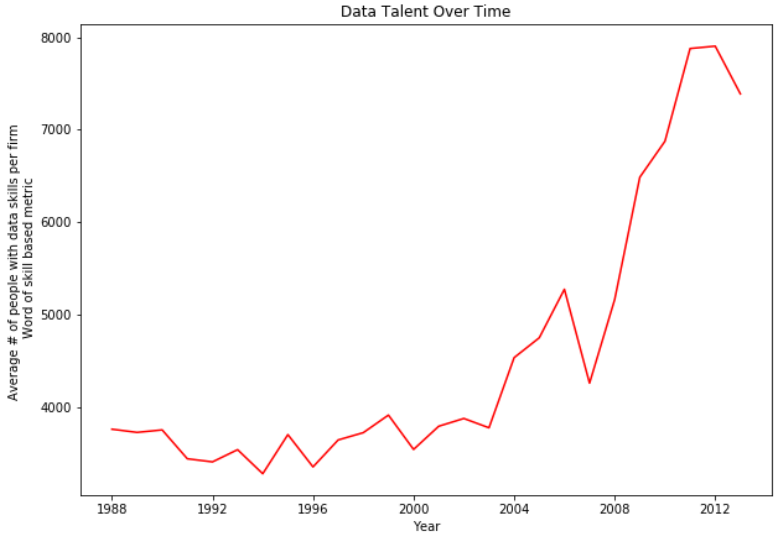


Figure 2. Average data talent over time

References:

Akcigit U, Kerr WR, Nicholas T (2013). *The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents*. Technical report, Working Paper.

Argyres NS, Silverman BS (2004). R&D, organization structure, and the development of corporate technological knowledge. *Strategic Management Journal*. 25(8-9) 929-958.

Arora A, Belenzon S, Rios LA (2014). Make, buy, organize: The interplay between research, external knowledge, and firm structure. *Strategic Management Journal*. 35(3) 317-337.

Autor DH, Levy F, Murnane RJ (2003). The Skill Content of Recent Technological Change: An Empirical Exploration*. *The Quarterly Journal of Economics*. 118(4) 1279-1333.

Avery J, Norton M (2014). Learning From Extreme Consumers.

Bharadwaj S, Bharadwaj A, Bendoly E (2007). The performance effects of complementarities between information systems, marketing, manufacturing, and supply chain processes. *Information systems research*. 18(4) 437-453.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*. 2008(10) P10008.

Bloom N, Jones CI, Van Reenen J, Webb M (2017). *Are ideas getting harder to find?* National Bureau of Economic Research.

Bradstreet D (2017). *Analytics Accelerates Into the Mainstream*. Dun & Bradstreet.

Bresnahan TF, Brynjolfsson E, Hitt LM (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*. 117(1) 339-376.

Brynjolfsson E, Hitt L (1996). Paradox lost? Firm-level evidence on the returns to information systems spending. *Management science*. 42(4) 541-558.

Brynjolfsson E, Hitt LM, Kim HH (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Chen Y, Elenee Argentinis JD, Weber G (2016). IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*. 38(4) 688-701.

Cross R, Parker A, Prusak L, Borgatti SP (2001). Knowing what we know:: Supporting knowledge creation and sharing in social networks. *Organizational dynamics*. 30(2) 100-120.

Gao G, Hitt LM (2012). Information technology and trademarks: Implications for product variety. *Management Science*. 58(6) 1211-1226.

Griliches Z (1990). *Patent statistics as economic indicators: a survey*. National Bureau of Economic Research.

Hall BH, Jaffe AB, Trajtenberg M (2001). *The NBER patent citation data file: Lessons, insights and methodological tools*. National Bureau of Economic Research.

Hannan MT, Freeman J (1984). Structural inertia and organizational change. *American sociological review* 149-164.

Henderson RM, Clark KB (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly* 9-30.

Joshi KD, Chi L, Datta A, Han S (2010). Changing the competitive landscape: Continuous innovation through IT-enabled knowledge capabilities. *Information Systems Research*. 21(3) 472-495.

Kleis L, Chwelos P, Ramirez RV, Cockburn I (2012). Information technology and intangible output: The impact of IT investment on innovation productivity. *Information Systems Research*. 23(1) 42-59.

Lam A (2005). *Organizational innovation*. Oxford University Press.

Majchrzak A, Malhotra A (2016). Effect of knowledge-sharing trajectories on innovative outcomes in temporary online crowds. *Information Systems Research*. 27(4) 685-703.

Milgrom P, Roberts J (1990). The economics of modern manufacturing: Technology, strategy, and organization. *The American Economic Review* 511-528.

Muller F, vom Brocke J (2016). The effect of big data analytics on firm performance: An econometric analysis considering temporal dynamics and industry characteristics.

Nagle F (2017). Open Source Software and Firm Productivity. *Management Science*. Forthcoming.
Pang M-S (2016). Politics and Information Technology Investments in the US Federal Government in 2003–2016. *Information Systems Research*. 28(1) 33-45.
Siggelkow N, Rivkin JW (2006). When exploration backfires: Unintended consequences of multilevel organizational search. *Academy of Management Journal*. 49(4) 779-795.
Tafti A, Mithas S, Krishnan MS (2013). The effect of information technology-enabled flexibility on formation and market value of alliances. *Management Science*. 59(1) 207-225.
Von Hippel E (1994). "Sticky information" and the locus of problem solving: implications for innovation. *Management science*. 40(4) 429-439 %@ 0025-1909.
Zafar SY, Peppercorn JM, Schrag D, Taylor DH, Goetzinger AM, Zhong X, Abernethy AP (2013). The financial toxicity of cancer treatment: a pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. *The oncologist*. 18(4) 381-390.

Appendix

A. Patent Data and Networks

Our measures of organization of innovation are based on network analysis of patent co-authorship relationships. We begin by constructing a list of patents attributable to each firm from 1988 to 2013.[2] We use the NBER Patent Citation Data File (Hall et al. 2001), which links patents to publicly-traded firms through 2006. To capture the recent change from the growth of data analytics from years 2007 to 2013, we obtain the original patent data from the USPTO and use the same methods as Hall et al. (2001) to match patents to firms. These methods correct for matching related issues, such as misspellings, mergers and acquisitions, or patents assigned to subsidiaries. Following the convention in the R&D literature (Griliches et al. 1986, Hall et al. 2005, Saunders 2011), we use the application year (as opposed to issue year) because this is likely closer to the time the technology is developed and deployed.

Using the authors listed for each patent, we construct an intra-firm patent network for each firm annually between 1988 and 2013, using a sliding window of 5 years.[3] A node in this network represents an employee inventor and an edge (link) indicates the presence of one or more coauthored patents between two inventors in the same firm. The sliding window approach allows the patent network to be more accurately reflected than one that uses only a single year of patents (Wu 2013), although our results using one year of patents for each network do not qualitatively change. We then apply machine learning-based community detection algorithms to these networks to identify distinct innovation communities within each firm: a community of inventors is identified when they work closely with each other but have limited collaboration outside the community. Our primary analyses rely on the Multilevel community detection algorithm to identify clusters of innovative activity (Blondel et al. 2008)[4]. As suggested by Yang et al.

[2] The 2013 end date allows for a 3-year window for patent reviews. Thus, our data include patents granted before 2016.

[3] Each intra-firm network for a particular year includes all inventors listed on the patents that were applied in the current year, plus two years prior and two years after the current year.

[4] We use a greedy clustering algorithm that begins by treating every node (inventor) as a separate community and then moves each node to adjacent communities to see if it improves the modularity score: a high score indicates dense connections within communities but sparse connections across communities, while a low score indicates the communities are not clearly delineated. The first stage of the algorithm terminates when the movement of a single

(2016), we confirm our main results using other popular community detection algorithms (Fastgreedy, Leading eigenvector, and Infomap), all of which generated results (both in the metrics and the results of analysis of these metrics) similar to the Multilevel algorithm. These results suggest limited potential biases arising from the algorithm choice.

Using the communities identified, we create a dispersion metric which is one minus the Herfindahl concentration index of the fraction of employees in each community. The value is always between 0 and 1. A low value indicates most inventors are concentrated in a few large communities, and a high value indicates that inventors are dispersed across many different communities. The higher the value on this metric, the more decentralized is the structure of innovation.

B: Innovative Output

Our measures of innovative output are also based on patent data consistent with prior work on R&D productivity (Griliches 1990, Griliches et al. 1986, Hall et al. 2001). Like these prior studies we exploited the fact that these data are public, consistent (defined by law and subject to a largely stable regulatory environment), and rich, including information on inventors and their affiliations, detailed records of citations that can be used to measure impact (e.g., citation count), general area of technology described in the patent (patent class), a listing of prior art that can be used to identify the extent to which the innovation is the result of combining existing technologies or creating new ones, and an extensive textual description of the innovation that can be used to rank the level of novelty. We recognize that we cannot measure unpatented innovation and thus the associated effects of data analytics and decentralization on such innovations (Nagle 2014, Saunders and Brynjolfsson 2013). However, R&D expenditure, which can approximate a firm's overall effort on innovation, is often found to be positively correlated with patents

node no longer leads to a material improvement in modularity. For the next stage, the nodes in each community are aggregated into a single node which will represent a community as opposed to a single inventor. The link between two communities in this new network is the sum of all links between inventors in one community to the another (before the aggregation occurred). The greedy algorithm is applied again by assigning nodes to adjacent communities until modularity cannot be improved anymore. The final set of clusters identified is the distinct communities identified in the co-authorship network.

(Kleis et al. 2012). Thus, the patent-based metrics could potentially underestimate the effect of data analytics and innovation structure on innovation output[5].

To measure recombination, we begin with a measure created by Hall et al. (2001) [also described as "technological diversity" (Kaplan and Vakili 2015)] that is a Herfindahl concentration index, calculated based on the distribution of different three-digit main classes of prior art that are cited by the focal patent. The exact formulation is $1 - \sum_j^{n_i} s_{i,j}^2$, where $s_{i,j}$ represents the percentage of citations that patent i cited that belong to patent class j, out of n_i different patent classes. Patents that cite a wider variety of classes will have a measure closer to 1, whereas patents citing a single technology class will have a score of 0. Thus, the technology diversity metric represents the level of recombination because a patent that cites many technology classes is more likely to represent a synthesis of existing technologies into a new technology than a patent that cites few or a single area of prior art. To account for potential bias in this metric for patents with small numbers of citations (that is, patents with fewer citations necessarily cite fewer different categories of prior art) we adjust this measure by citations as proposed by Hall et al, 2001, pp 44-46):

$$Recombination = \left(\frac{Cites}{Cites-1}\right)\left(1 - \sum_j^{n_i} s_{i,j}^2\right). \quad (1)$$

Finally, we aggregate this measure at the firm level by using a weighted average on citations. For our analysis, it is important to distinguish reuse of existing combinations from the creation of new combinations (Akcigit et al. 2013). Our approach is based on the idea that reuse is defined as recombination of technologies that have been previously used; more specifically, if the firm has previously created a patent that combined the same patent classes (technology areas) in the past we consider that patent reuse. If a firm combines patent classes in a new way, we consider this to be a new recombination of existing technical classes. We first examine all the citations in a patent as well as what these citations have cited to determine the patent classes associated with the focal patent. We then

[5] Nagle (2014) and Saunders and Brynjolfsson (2013) have detailed discussions on valuing unpatented innovation. While our work use patent output to approximate the overall innovation output, future work should examine these effects for innovations that are not patentable as well as other intangibles that are difficult to measure.

determine whether the patent classes represent an existing or a new combination. If the combination has appeared for the first time in the patent's primary class, the patent is viewed as a new combination. If it already exists, then it is viewed as a reuse recombination. When a patent creates a new patent class that did not exist before, we consider it to be novel technology. To obtain a firm level measure, we the aggregate all the patents (m) at the firm level in a single year and count how many patents are reuse recombination or new recombination or novel technology.

$$reuse\ recombination = \sum_{j=1}^{m} 1(existing\ combination_j) \quad (2)$$

$$new\ recombination = \sum_{j=1}^{m} 1(new\ combination_j) \quad (3)$$

$$novel\ technology = \sum_{j=1}^{m} 1(new\ technology_j) \quad (4)$$

This distinction for each type of innovation can be made at the firm level (e.g. a new recombination for the firm, even if other firms have done similar combinations) or on a global level (e.g. a recombination is new only if no other firm has ever combined that set of patent classes to create a new patent). This distinction between firm-specific prior knowledge (local) and global (all available knowledge) is useful because it enables discrimination between innovations that bring in external knowledge versus innovations based on existing firm knowledge. Thus, we have a total of 6 ways to classifying innovation along two dimensions: (1) novel technology, reuse and new recombination, and (2) local vs. global.

Lastly, we experiment with another alternative method to measure novelty. We note that patents contain a considerable amount of free-form text data describing the innovation, which can be used to represent the underlying inventive ideas. Prior research on patent innovation, including in information system research, used bibliometric methods and citation based analysis to study innovation impact and evolution (Hall et al. 2001, Kleis et al. 2012, Saunders 2011). However, the use of raw text from the patent has been less frequent with a few notable exceptions (Azoulay et al. 2007, Kaplan and Vakili 2015, Upham et al. 2010, Wu 2013). Our measure is based on the idea that novel ideas can often be detected through vocabulary shifts. In our case, we examine the appearance of new words or topics in the patent abstract,

which provides a summary of the innovation. We apply a "bag of words" model to identify patent terms and then calculate the age of each word by when it first appeared in the prior art. Each (non-stop) word in the patent is assigned an age. A word that has not appeared previously is assigned an age of zero. If the word has appeared in another patent before, each word in each patent is assigned an age that is the difference between the application date of the focal patent and the time that word first appeared in any patent. The overall novelty score is the average age of the words in a patent:

$$\text{Novelty} = \frac{1}{N}\sum_{w=1}^{N}\frac{1}{\text{Age}_w+1}.$$

The novelty of a firm's patent portfolio for a particular year is equal to weighted average of the novelty score of firm's own patents.[6]

We chose to use a "bag of words" technique because it is likely to be more sensitive to shifts in vocabulary than topic modeling methods such as Latent Dirichlet Allocation (LDA), which relies on identifying a stable set of topics for classification (Blei et al. 2003, Shaparenko and Joachims 2009); in fact, the shifting vocabulary likely makes it difficult to find a stable ranking of topics, which is essential for accurate classification of LDA (Greene et al. 2014). In addition, the large topic space (essentially the language of all scientific endeavor over 20 years) would make it difficult to establish a reasonable topic space without expert input, which defeats the benefits of automated classification. In contrast, "bag of words" models are relatively simple, have long been established to detect novelty at both sentence and document levels (Voorhees 2003), and they do not require auxiliary assumptions about the number of topics or degree of association between words needed to define a topic. As a check, we also ran an LDA model to identify topics in patents and measured the novelty of each topic in patents according to when it was first detected. The LDA results are directionally consistent. Using the vocabulary-based metric to measure novel technologies produced qualitative similar results to the novelty metric based on the technology class.

[6] We also used the sum of the novelty score and the results are very similar to using the average.

C: Data Analytics

We use two data sources—resumes and job reviews—to measure a firm's data analytics capabilities. First, we use resume data of employees at the firm to calculate the total number of employees possessing the relevant data analytics skills. The total universe of resumes used across all of the firms in this study is six million, collected in 2007. Since a resume provides a time stamp of employment, this provides a panel data structure with measures from 1988 through 2007. Prior work suggests that resumes are representative of employment generally due to their large size and the fact that many employees in information-intensive occupations engage in passive job seeking – posting their resumes publicly, making them available to potential employers even when they are not actively searching for a job – and thus reduce the concerns of selection bias (see further discussion of the advantages and disadvantages of resume data in Tambe and Hitt, 2014).[7] To capture the level of investment in data analytics skills for more recent time periods (2008-2013), we also gathered data on employee job titles (a subset of what is available in the resume data) from a large online job-posting site where employees can post employer reviews.

To classify data analytics skills from the resume data, we apply natural language processing techniques on free-form text and job titles to identify the skills of each employee from 1988 to 2007. We include direct matches on keywords like "analytics" as well as inferred matches on phrases such as "regression modeling," "Hadoop," or "Machine Learning." Table A1 shows the list of words used. For each employee in our data set, we identify whether a data-related skill is present in his or her resume and approximate the timing of when the skill was acquired from the job history in the resume. We then measure firm-level data analytics capabilities by aggregating all employees with data analytics skills for each firm in each year and adjust it by a sampling rate[8] (Tambe and Hitt 2013). This approach assumes that the observed employees are sampled from the underlying population of all workers. We also assume that firm and occupation-specific

[7] Our use of resumes is similar to other studies that have used resumes to study network relationships among individuals and firms or used them to identify the presence of particular skills within a firm. For instance, Tambe and Hitt, (2013b) used such an approach to measure information technology investment by counting the number of IT employees in the firm and Tambe (2014) used resume data to analyze the diffusion of specific IT skills, such as Hadoop, across firms and locations.

[8] The sampling rate is calculated annually for each firm.

factors with respect to the likelihood of posting career trajectory or job reviews are uncorrelated.[9] Thus, our firm-level sample rate could be estimated by $\theta_j' = \frac{x_j}{L_j}$, where x_j is the number of employees in all occupations in our sample for firm j, and L_j is the total number of employees at firm j, data we obtained from Compustat.

Table A1: Taxonomy used to classify data analytics skills	
General:	Business intelligence, Data analysis, Data center, Data driven, Data fusion, Data integration, Data mining, Data warehouse
System oriented:	Cassandra, Cloud computing, Distributed system, Hadoop, HBase, Google File System, MapReduce
Algorithms & Methodology oriented:	A/B test, Ensemble learning, Genetic algorithm, Machine learning, Natural language processing, Neural network, Network analysis, Optimization, Pattern recognition, Predictive model, Regression, Signal processing, Simulation, Supervised learning, Unsupervised learning, Visualization

Note: The Sovren resume parser (http://www.sovren.com) provides us with the text mining and skills-based semantic matching algorithms that map free-form text responses to a taxonomy of skills. It is also used by a number of online job sites to facilitate skills-based resume searches (among other activities). We then classify data analytics skills by the taxonomy.

To extend the data analytics metric from 2008 to 2013,[10] we use a second data source containing information from a large online job review platform that had 3.7 million job reviews. Each review has information on job title, employer name and length of employment. While these data are not as detailed as the resume data, they do provide job titles and employer names which can then be used to identify skills available at a firm at any point in time. Specifically, we use the job title classification from O*Net to identify analytics related positions, similar to how IT labor is distinguished from other employees in earlier work (Tambe and Hitt 2013, Tambe and Hitt 2013). If any of the skills listed under the job title is related to data

[9] To the extent such matching varies systematically across firms, we can address this problem by using fixed effect models to alleviate the bias from time-invariant firm characteristics.

[10] We search the word "data" in the Occupation Search box provided by https://www.onetonline.org/, and obtain a list of job titles ranked according to how well these job titles matched to the keyword "data," according to the matching process in https://www.onetonline.org/help/online/search#keyword. A total of 576 primary data-related job titles are generated, with detailed reports on the requisite technology skills and general skills. An additional 788 alternate occupation titles are found from the primary job titles. For example, the occupation "Database Administrator" has several alternate job titles, such as data engineer, data miner and data center manager. We count a person as having data analytics skills if any of the skills in that job title is related to data.

analytics, we assume anyone with such a job title to possess data analytics skills. We aggregated these individual level skills for each firm-year observation, tracing back to the year when the employment started and adjusted the raw aggregation by the underlying sampling rate as we did with the resume data. To check our title-based classification used in job posts, we also used a simple keyword-based classification that identifies three relevant phrases about data analytics ("data," "statistic," and "business analytics") and the results are consistent.

To ensure our results are not driven by the use of keyword matches in the full text in the resumes, we also apply a title-based classification to the resume data using the same method as we did for the job review data. This ensures the same title-based classification is uniformed applied to both resumes and job reviews. As shown in Table A2, results from the title-based classification for analytics skills are consistent with the main results.

We also appropriately scaled the two different data sources to ensure that the metrics are comparable between the sources. This involves scaling the job-title-based classification used in job reviews to the full-text-based classifications used in resumes. We first convert the full-text-based classifications in resumes to their corresponding job-title-based classifications. We then convert this job-title-based classification in resumes to the equivalent in job reviews. When a firm does not exist in the matching data, we use the industry average of that firm to calculate the appropriate scale. Thus, our data cover the critical period for the rising demand in data analytics ranging across more than 26 years from 1988 to 2013. Although we combined two different data sources to measure data analytics skills, the year dummies help to control for any idiosyncratic characteristics associated with each data source that may bias our results. Furthermore, we used the logarithm of employees with data skills which should alleviate the potential scaling differences in the underlying data. We also performed a Chow test on our main results that cannot reject the hypothesis that the analytics coefficients using the combined data are equal[11] and thus we can reasonably combine the two data sources. To ensure our results are not driven by the merging of two different data sources, we

[11] For instance, the Chow test for pooling our data in our key regression (see Column 4 of Table 3 in the body of the main paper) on the complementarities between data and type of innovation is $F(3,2994) = 0.62$, $p < 0.602$).

estimated our models in each data source separately and the results are consistent with the combined sample (Table A3) and observe the complementarities to grow stronger in later years using the job reviews.

Table A2: The Productivity Test: Using Title-based Data skills

DV: ln(Sales)	(1)	(2)	(3)	(4)
Model	FE	FE	FE	FE
ln(Materials)	0.594***	0.594***	0.594***	0.594***
	(0.0239)	(0.0239)	(0.0239)	(0.0243)
ln(Capital)	0.102***	0.102***	0.102***	0.104***
	(0.0202)	(0.0202)	(0.0202)	(0.0201)
ln(IT labor)	0.0111**	0.0111**	0.0115**	0.0114**
	(0.00548)	(0.00548)	(0.00551)	(0.00553)
ln(Non-IT labor)	0.326***	0.326***	0.326***	0.325***
	(0.0244)	(0.0244)	(0.0244)	(0.0245)
ln(Employees with college higher degree)	0.00221	0.00221	0.00223	0.00183
	(0.00260)	(0.00260)	(0.00260)	(0.00263)
ln(Data)	-0.000577	-0.000577	-0.000594	-0.00137
	(0.00154)	(0.00154)	(0.00154)	(0.00156)
Patents			-0.00548	-0.00574
			(0.00822)	(0.00868)
ln(number of inventors)				-0.00947
				(0.0103)
ln(Decentralization: multilevel)				-0.00956
				(0.0520)
Data Skills X Decentralization: multilevel				0.0128**
				(0.00572)
Observations	11,033	11,033	11,033	11,033
R-squared	0.988	0.988	0.988	0.988
# of firms	1,423	1,423	1,423	1,423
Industry	YES	YES	YES	YES
Year	YES	YES	YES	YES

Note: (1) The Multilevel algorithm is used to measure community dispersion.
(2) Robust (clustered by firm) standard errors in parentheses
(3) *** p<0.01, ** p<0.05, * p<0.1

Table A3: The Productivity Test: Sample Splitting

Sample	Resumes (1988-2007)	Job reviews (2008-2013)
DV: ln(Sales)	(1)	(2)
Model	FE	FE
ln(Materials)	0.567***	0.557***
	(0.0263)	(0.0857)
ln(Capital)	0.104***	0.116
	(0.0227)	(0.0769)
ln(IT labor)	-0.00521	0.210***
	(0.00602)	(0.0788)
ln(Non-IT labor)	0.323***	0.0554
	(0.0257)	(0.104)
ln(Employees w/ college+ degree)	0.00124	-0.00263
	(0.00298)	(0.00268)
ln(Data)	-0.00463	-0.0141
	(0.00308)	(0.00886)
Patents	0.000331	-0.00510
	(0.00826)	(0.0320)
ln(number of inventors)	-0.0101	0.0273
	(0.00990)	(0.0190)
ln(Decentralization: multilevel)	-0.0394	-0.0519
	(0.0663)	(0.0769)
Data Skills X Decentralization	0.0266**	0.0579**
	(0.0123)	(0.0288)
Constant	-0.218	1.145**
	(0.159)	(0.453)
Observations	9,329	1,704
R-squared	0.989	0.997
# of firms	1,338	478
Industry	YES	YES
Year	YES	YES

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

D: Method and Identification

Our primary analysis tests for complementarities between the deployment of data analytics and innovation structure. Two types of statistical tests have been developed to assess the existence of such complementarities (Arora and Gambardella 1994, Brynjolfsson and Milgrom 2013, Milgrom and Roberts 1990): correlations (adoption or demand equations) and performance differences (productivity equations).

The correlation test, which can be estimated in the form of conditional correlations or reduced-form demand equations, determines whether a cluster of practices is more likely to be adopted jointly rather than separately. Although the demand equations have the advantage that they are relatively simple and provide the greatest power if firms are optimally matching complementary practices, they have the disadvantage that the simplicity makes them vulnerable to unobserved heterogeneity. In addition, such an analysis will tend to understate the strength of complements if not all firms are endowed with or able to change to the optimal match between complements. An alternative approach to examining complementarities is to measure whether there are performance differences between firms that adopt complementary practices as a group and those that do not. These metrics have the advantage of a direct tie to a relevant firm outcome (performance), and the effects are most powerful statistically if not all firms have found the optimal match, which may be likely for a relatively new change in business practices. Over time, as the complementarities system diffuses to other firms, the correlation would increase but the productivity premium would decrease because the relative advantage for using the system diminishes as its adoption spreads.

We take the view that firm capabilities to pursue specific innovative practices are quasi-fixed because it is difficult to change them due to organizational inertia, path dependency, culture, norms, and prior strategic choices (Hannan and Freeman 1984, Lam 2005). The rapid diffusion of analytics technology and a relatively recent, exogenous shift in the amount of data available to firms enable us to observe the effects of data analytics on different firm practices related to innovation, even those that are not optimally configured to benefit from analytics capabilities. A similar assumption has been made previously in IT research for other types of IT-related complements (Autor et al. 2003, Bresnahan et al. 2002, Brynjolfsson and Hitt 1996, Milgrom and Roberts 1990).

We first explore the correlation between data analytics and the dispersion of innovation communities. According to our theory, we expect the two to be positively correlated. Thus, we relate data analytics capabilities to innovation dispersion after controlling for sales, innovation inputs, and firm (fixed effects) (γ_i) and year (y_t). This can be viewed as a reduced-form demand equation to the extent that the "price" of

data skills can be viewed as varying by firm and time, and we are treating innovation structure as quasi-fixed relative to data skills.[12]

$$Data\ Analytics_{it} = \beta_0 + \beta_1\, ln(sales)_{it} + \beta_2\, ln(R\&D)_{it} + \beta_3 patent_{it} + \beta_4 disp_{it} + y_t + \gamma_i + \epsilon_{it} \quad (1)$$

The productivity test examines whether the hypothesized system of complements is more productive if adopted together or separately; it is typically implemented by including the direct effects as well as the interaction effect in a productivity or performance model. A complementarities relationship is consistent with a significant interaction term but can also be estimated by sample comparisons that would also imply increasing benefits of one factor in the presence of increasing levels of another complement (see e.g., Brynjolfsson and Milgrom (2013)), a technique that has been applied empirically in a number of recent papers (see e.g., (Aral et al. 2012, Tambe et al. 2012)).

Although there are several approaches to measuring the marginal effect of information technology on firm performance, we choose the multi-factor productivity framework implemented with the Cobb-Douglas production function, which is the most commonly used framework in estimating IT value. Using this framework, we relate firm output such as sales or value-added to various firm inputs such as labor, capital and materials after controlling for temporal and firm-specific variations. The residual of this equation can be interpreted as the multi-factor productivity, and we are interested in exploring whether data analytics skills and their interaction with innovation structure are related to this measure of productivity.

Equation 6 below represents a standard function that has been used in the IT-productivity literature augmented with the additional terms for data analytics skills and innovation (and their interactions). Specifically, we relate sales to production inputs such as materials expense (m), physical capital stock (k) and labor (*IT and nonIT*, measured as the number of employees with and without IT skills) as well as controls for firm fixed effects (i) and year (t) dummy variables that address firm-specific unobserved heterogeneity, temporal productivity shocks and measurement error in price deflators. Here and throughout

[12] A typical factor demand equation relates the quantity of a factor to the prices of that factor and the levels of quasi-fixed factors. Sales can also appear in this equation if the demand system is assumed to arise from cost minimization (since this model treats sales as exogenous). In our specification, sales also serve as a control for firm heterogeneity in size.

we will use italics to designate specific regression measures (see Tables 1 for further detail). Our panel analysis includes the performance and labor metrics on a 27-year panel from 1998 to 2013. To this base specification, we add data analytics skills (*data)*, the innovation structure of the firm (*disp*) as well as their interactions (*data X disp)*. Of particular interest in this equation is β_7, since a positive value of this coefficient suggests the presence of complementarities.

$$\ln(sales)_{it} = \beta_0 + \beta_1 \ln(m)_{it} + \beta_2 \ln(k)_{it} + \beta_3 \ln(nonIT\ labor)_{it} + \beta_4 \ln(IT)_{it} + \beta_5 \ln(data)_{it} + \beta_6 disp_{it} + \beta_7 Data_{it} X\ disp_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (2)$$

Next, we explore the nature of innovation as enabled by these complementarities. We first investigate whether the interaction of data skills and innovation structure could help produce a novel technology element that can serve as a building block for future recombination in innovation or general recombination in innovation. We then examine the extent to which data can facilitate the generation of a new recombinant innovation or reuse of existing recombination.

$$Novelty_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 disp_{it} + \beta_3 data_{it} X\ disp_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (3)$$

$$All_recomb_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 disp_{it} + \beta_3 data_{it} X\ disp_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (4)$$

$$Reuse_recomb_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 disp_{it} + \beta_3 data_{it} X\ disp_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (5)$$

$$New_recomb_{it} = \beta_0 + \beta_1 \ln(data)_{it} + \beta_2 disp_{it} + \beta_3 data_{it} X\ disp_{it} + y_t + \gamma_i + \text{controls} + \epsilon_{it} \quad (6)$$

We also use a specification in which data, IT and other labor are disjoint. (i.e. that a person could either have IT skills or data skills but not both), and the results are similar to when IT and data skills are not disjoint.

Complementarity relies on the benefits of matching two or more investment decisions rather than the effectiveness of the decisions themselves. As a result, it is naturally robust for some types of endogeneity and reverse causality problems. While there are many explanations for why a particular investment might be correlated with unobserved shocks (e.g. firms spend unexpected free cash flow on new investment), it is difficult to explain why such a relationship would be present only when two investments are made together or not at all. In addition, the correlation test and the productivity test for complementarities are subject to

different sources of bias which provide some ability to cross-check results. Nonetheless, we estimate that there are at least two endogenous factors that could lead to a positive bias. First, high-performing firms can have resources available that can be redeployed to increase investment in data talent acquisition, which could lead to a reverse causality between performance and data analytics. By itself this is not a problem in interpreting our complementarities result, but it could introduce bias if this tendency depends on the organizational form. (This is not implausible given that human capital practices and organizational practices are often closely linked). Second, there could be omitted variables, such as higher management quality attracting higher quality labor, that could lead to an upward bias. We address this problem in two ways. First, selection bias from omitted variables is less likely in the complementarities framework if we find the use of data analytics is only complementary to some firm practices but not others. Second, we treat data analytics skills as endogenous and use three sets of instruments to correct potential biases: 1) data analytics skills in a firm's existing hiring network; 2) a neighbor's adoption of an enterprise resource planning (ERP) system and the average number of neighbors that started to use ERP systems; 3) a neighbor's adoption event and the average number of neighbors that started to use Human Capital Management (HCM) systems. All these instruments are derived from a firm's immediate hiring network, which can be used as a proxy to measure the ease of access to existing data analytics talent and therefore the "cost" of acquiring analytics skills. This hiring network is constructed by examining employee flows across firms using techniques similar to those in Wu et al. (2016).

Firms are represented by nodes and firm-to-firm employee movements are represented by edges in the network. A directed edge is identified between Firm A and Firm B if A hired someone from B in the previous five years[13] and the weight of that edge is expressed as the number of employees hired. The first instrumental variable is a Hausman-type instrumental variable, based on the neighboring firm's existing data talent. It measures the potential labor pool accessible to the firm through employee mobility, which reflects the cost of acquiring additional analytics talent. At the same time, the data analytics skills pool from

[13] This type of graph has been previously utilized to identify the position of a firm in the network of labor flows (Wu et al. 2016).

the neighboring firms is only the potential access a firm could have, and thus it should not affect the firm's own performance. The only point at which the talent pool could affect performance of the focal firm is when the firm hires from a neighboring firm, which the firm fixed-effect model can address because it can capture the effect from the change in data analytics skills between years. Accordingly, the potential for productivity spillovers through labor movement would not invalidate our instruments.

The second and third instrumental variables are derived from measures of the adoption of ERP systems broadly and the HCM module by firms in the same labor pool. Because these types of systems are likely complementary to the use of data analytics talent,[14] this can change the overall access to data analytics skill as firms acquire employees with these skills during implementation and employees transfer these skills through job-hopping. The ERP adoption event may have an especially large effect of increasing the availability of employees because of a transition from an implementation team to operations may free up some types of staff. We focus specifically on the point at which a firm implements these systems (rather than simply licenses, which could be several years earlier) to prevent common industry-level shocks from potentially generating an increased demand for these systems in nearby firms as well as a direct shock to productivity in the focal firm (see related arguments in (Hitt and Brynjolfsson 1996, Hitt et al. 2002)). Furthermore, our instrumental variables derived from hiring neighbors are inherently more protected against common industry shocks because a firm's hiring network is composed of a much more diverse set of firms than firms in the same industry. Thus, while a neighboring firm's adoption of ERP should not directly affect the focal firm's productivity, it should have an impact on the focal firm's access to data analytics-related talent. To illustrate this, we show in Figure A1 the impact on the access to data talent of the focal firm (Firm A) when its neighbor (Firm B), from which A hires people, adopted an ERP system. B now has a need to hire employees with ERP experience as well as experience working with the associated process change that often accompanies the ERP implementation, and starts to hire from new firms, such as

[14] ERP systems are off-the-shelf software packages that offer a variety of functions including finance, human resources, supply-chain management, consumer relationship management and business intelligence (Hitt and Brynjolfsson 1996; Hitt et al. 2002). These types of enterprise information systems often produce a large amount of data about firms' business processes, which can affect the demand for data skills.

C. Not only has B's hiring pool expanded, at the same time, this network change also increases to A's access to data analytics-related talent. As shown in Figure A1, A still hired from the same three firms but its access to data talent has expanded because of B's newly hired data talent that it acquired from C. This example shows that B's ERP implementation induces a change in A's access to talent even though A has not changed its hiring practices. Thus, we can use a neighbor's ERP adoption event as an instrument for the focal firm's access to data analytics talent. The first stage results are shown in Column 5 of Table 2 in the main body of the paper.

We have two available measures of ERP adoption. First, we can infer ERP adoption by examining whether employees in a firm have job titles or job descriptions in their resumes that suggest the use of ERP (e.g., employees stating in their past job description that they "used an ERP system to conduct inventory control"). We can infer the adoption of HCM systems using data from a major ERP vendor that provided the dates when clients purchased and implemented the HCM module. This is used to construct a binary variable to represent the firm's HCM module implementation status, which is 0 in the years before the firm started using the system and 1 afterward.

Innovation structure could also be endogenous. Here, we rely on the assumption that organizational structure is often quasi-fixed, an assumption which has been used in prior work on IT-organizational complementarities (Bresnahan 2001, Brynjolfsson and Hitt 1996, Levy and Murnane 1996, Milgrom and Roberts 1990). The quasi-fixity of our innovation structure measures is likely be correct given the literature that has documented the difficulty firms face in changing their innovation practices and portfolios, because changing them requires firms to substantially alter practices in many areas including finance, talent acquisition, monitoring policy, incentive pay and reporting relationships (Nagji and Tuff 2012, Wessel and Christensen 2012). We verified the quasi-fixed assumption by examining the extent to which the structure of the innovation process varies across time within the same firm. Overall, the standard deviations of the dispersion metric within firms are very small, suggesting that it is difficult for firms to change their innovation structure. Thus, our regression results can be interpreted as assessing whether pre-existing firm differences in innovation structure influence the benefits of acquiring data analytics capabilities. Even

though it is quasi-fixed, there is some temporal variation in this metric across so the coefficient of the direct term can be identified, but this variable should be interpreted with caution due to the likelihood of a low signal to noise ratio in our fixed effects regressions on the decentralization variable alone; this concern does not apply to the interaction term which is the focus of our analysis.

Finally, we address general endogeneity of production inputs with respect to output using standard panel-based instrumental variables methods (the SYS-GMM estimator of Blundell and Bond (1998)), that were developed to specifically address problems of endogeneity in micro-level productivity analysis. The combination of instrumental variables, SYS-GMM, and labor quality controls should substantially reduce the potential for bias in our estimates.

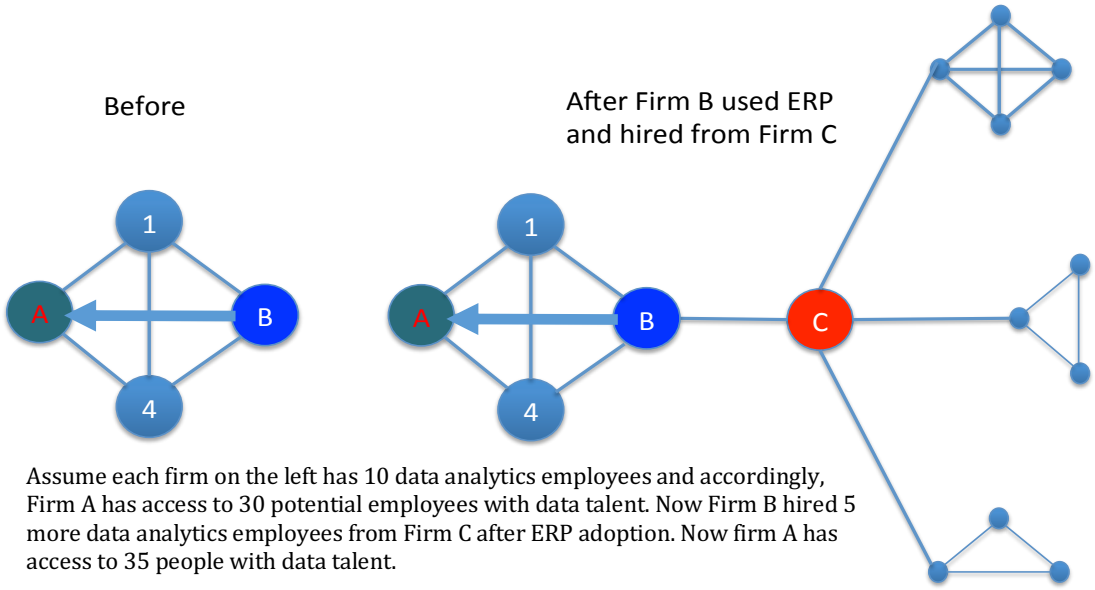


Figure A1: Instrumental Variables