What a firm does is more revealing than how much it makes, but firms are often described with metrics for economic size. Instead, we characterize what firms know in terms of what they read, the information footprint, using a data set of hundreds of millions of records of news articles accessed by employees in millions of firms. We discover that the reading habits of firms are of limited diversity. This observation suggests that information constraints act on firms. To understand how, we relate a firm’s information footprint with economic variables, showing that the former grows superlinearly with the latter. This suggests that information constraints act on firms. To understand how, we relate a firm’s information footprint with economic variables, showing that the former grows superlinearly with the latter.

Second, we reconstruct the topic space firms inhabit, finding that the space resembles a tangled “hairball” with a few dense knots of topics and many specialized strands sticking out. Some of the topics are ubiquitous, indicating inescapable demand regardless of firm size. Finally, we connect these pieces in a model of how firms grow in the space of topics. We show that diversity in firm reading habits can be explained by a mixed strategy of local exploration and recurrent exploitation on the topic graph. This shows that the constraints that the space of ideas imposes on firm growth provide a useful and new perspective on firm development.

The knowledge of a group of individuals is not just the sum of its parts (1). The notion of “collective knowledge” and how it accumulates has become the topic of intense study in research on the importance of social skills (2), coworker complementarities (3), teamwork (4), and cultural evolution (5). Here, we focus on the collective knowledge held by firms. Firms differ tremendously in their size and economics, management scholars, and physicists have employed a variety of quantitative metrics that capture the size of a firm in terms of its assets, sales, or workforce (6–9). However, the size of the firm does not provide any information on what the firm does and even less on what the firm knows. However, a firm’s competitive advantage (10) is typically related to the unique resources and capabilities it commands, i.e., to its collective knowledge. This means that understanding what firms know is key to understanding how well they perform (11).

We contribute to the understanding of collective knowledge of firms by extracting the information footprint from how individual employees collect, process, and generate information. To do so, we analyze an extensive data set, or “intent data” (12, 13), consisting of hundreds of millions of records of content accessed by firm employees within a large universe of publishers including The Wall Street Journal, Bloomberg, Forbes, Business Insider, and CBSi, along with more specialized groups of sites such as 1105Media, ITCentral Station, and Questex. Most are anonymous but span technology, marketing, legal, biotech, manufacturing, and a wide range of business services (12). We focus on a two-week period between the dates of June 10, 2018 and June 23, 2018, which we expect is generally representative of the data set as we detail further in Appendix A in the Supplementary Information. The limited time window also ensures that proprietary preprocessing steps used to generate the data remain consistent. In principle, the data would allow us to determine which news article a particular employee at a firm accessed and when. For each article, we have up to ten associated topics that have been identified with a proprietary topic modeling algorithm (more details in reference (12)). These different properties permit us to analyze employee reading at different scales of resolution from the individual article, which then belong to sets of content distributors or “sources,” and that may overlap in topic as diagrammed in Figure 1. Importantly, the comprehensive nature of the data set allows us to obtain a multiscale portrait of how firms seek out information down to the individual acts of information acquisition.

**Limited diversity of reading**

We begin with several remarkable patterns in the way that firms acquire information as a function of how often they read in Figure 2. For each firm in the data set, we calculate the total number of times the firm accessed content, or left a “record,” with the number of unique types of information accessed by the firm from articles, sources, to topics. This Heaps’ plot leads to several curious observations (14).

First, we know that a single employee might read the same...
article several times or share it with his colleagues such that this plot would at maximum trace the black, dashed, one-to-one line. While the least read firms saturate the diversity of articles, the maximally diverse firms fail to saturate this curve. This occurs beyond about $R \sim 10^2$ records per firm per day for articles, $R \sim 10^5$ records for sources, and $R \sim 10$ records for topics in Figure 2. This raises the question of how the cost of acquiring new information imposes a ceiling on its diversity and why it does not seem to matter below some amount. Second, the maximum given the number of records scales with a sublinear power law above this lower cutoff with exponent given by the slope of the straight line on the log-log axes. This pattern suggests that active firms are limited to repetitive reading perhaps because of some underlying constraint on reading dynamics even for the most read firms. Third, the distance between the maximally and minimally diverse firms shrinks with number of records as if more prolific firms become more deterministic in their behavior. This is especially intriguing because many less read firms often are extremely monotonous, revisiting exclusively the same sources of information — although it is possible that some of the most repetitive readers are automated, or bots. By connecting these curious patterns to one another, we aim to investigate what these tell us about firm information acquisition.

We consider three main parts in this paper. First, we demonstrate that firms display power law scaling in the information dimension that is related to but different from known scaling along traditional economic measures. Importantly, we find that the amount that firms read is a proxy for firm size. Then, we characterize the topic space in which firms live to show that it resembles a tangled “hairball.” Firms grow in this topic space as they grow in economic size, and their reading habits are shaped by the features of the hairball topic network. Third, we connect the aforementioned features to one another in a simple model of how firm growth is accompanied by expansion in the space of topics. We show that firm extent in the space of ideas is a combination of topic space structure and bimodal information strategies indicating the role of exploration and exploitation.

![Fig. 1. Information falls across multiple layers of resolution from articles (gray circles) provided by content providers, or sources (blue circles), which can be related by topic (dashed circles). Articles form the most unstructured view of the data and each can belong to multiple sources or topics and so form effective scales for comparative analysis. Here, we consider a copy of the same article on a different content provider as a separate article.](image1)

![Fig. 2. Heaps’ plots for firm reading. Scaling of (a) article, (c) source, (e) topic with number of records. Each graph presents one measure of the rate at which firms collect new information. While the total number of articles and sources much greater than what any single firm accesses, the number of topics is bounded at $T = 4,338$. Maximally diverse readers (orange) fall below a 90% of linear information collection at about $R = 3,500$ records for articles, $R = 80$ records for sources, and $R = 15$ records for topics. Mean diversity in blue and minimum values greater than one in green. (b,d,f) Deviation from linearity as ratio of plotted markers to 1:1 line. Plots show data for June 10, 2018 through June 23, 2018 on a 10% subsample of the data.](image2)

**Economy of scale in information**
As a first step to characterizing how firms seek out information, we consider the relationship between the simplest measure of information acquisition, the number of times a firm is recorded in the database, or “records,” and conventional measures of firm capital. In Figure 3, we plot for public firms recorded in the COMPUSTAT database firm assets, PPE (plants, property, equipment), employees, and sales against the number of records. A few, well-known example firms are shown including Microsoft, Apple, Zendesk, and BioNTech. Building on previous work on scaling in firms (15), we consider a power law relationship between economic measure $Y$ and records $R$,

$$Y = AR^\beta,$$

for some positive constant $A$ and positive exponent $\beta$. Importantly, the exponent $\beta$ is independent of the units of $X$ and $Y$, which are separately captured in $A$. We find that all economic measures scale sublinearly with the number of records with scaling exponents of about $\beta \approx 3/4$ except for assets which scale somewhat slower as $\beta = 0.65 \pm 0.02$ (Table S3). Furthermore, it tends to be the case that service firms (as indicated by Standard Industrial Classification codes 7000-8999, red points falling above the fit line) have more records per unit economic measure as compared to firms in mining (SIC codes 1000-1499, green points falling below the fit line). This is not the case for the scaling with employees, but employees is known to be poorly estimated. While “services” includes a broad swathe of industries and thus displays wide variation along the vertical,
First, we note that a power law distribution describes well the distribution of information quantity if it is the case that the distribution of firm receipts, here given from Axtell that the distribution of firm receipts, here given.

Importantly, the scaling between economic and information measures is consistent with the heavy-tailed distributions of property, equipment (PPE), employees, and sales. Each blue point is a firm in COMPUSTAT and our data set. Mining (green), service (red), and all other (gray) firms. Orange line shows a power law fit to $Y \sim X^\beta$ with exponents (a) $\beta = 0.65 \pm 0.02$, (b) $\beta = 0.74 \pm 0.02$, (c) $\beta = 0.73 \pm 0.01$, and (d) $\beta = 0.72 \pm 0.01$ using one standard deviation from bootstrapped fits as error bars. Fitting range $R \geq 9$ given by the power law fit in Figure 4. We highlight well-known firms Microsoft, Apple, Zendesk, and BioNTech. BioNTech had no reported sales in 2018.

Cost of diverse information constrains firms

As firms read more, they may be reading about the same or about new topics. The distinction between the two possibilities manifests in the rate at which new articles, sources, or topics are accessed for a given number of records, which is the Heaps’ plot displayed in panels a, c, and e of Figure 2, respectively. The three Heaps’ plots all show qualitatively similar patterns of two distinct regimes. In particular, topics are extracted from a proprietary model and are bounded to a maximum of $T = 4,338$, whereas articles and sources are closer to the raw data and more numerous. We find that initially, small firms with the most diverse reading tendencies saturate the maximum number of articles they could read per record. At a larger size, the most diverse readers peel away from linear growth, indicating that the same articles $A$, sources $S$, and topics $T$ are reread. When the inflection point is defined as the number of records at which the quantity in question first reaches 90% of linear growth, it occurs at $R = 2,200$ records for articles, $R = 90$ records for sources, and $R = 15$ records for topics over the two-week period of study. Using our scaling relations above, the points correspond to publicly listed assets and annual sales of typically $5$ billion and $700$ million, $600$ million and $60$ million, and $200$ million and $20$ million, respectively. That the variation of the inflection point maps to firms of different sizes suggests that firms may pass critical sizes at which the costs of obtaining new types of information begin to matter.

One might anticipate that if there were some effective cost of per unit of information that it would increase with information granularity; after all, a firm searching for diverse topics is likely to be structurally different from one reading new articles but on the same topic. We can consider the implications of this hypothesis by imagining that the diversity inflection point occurs at the same threshold cost $C^*$ for some effective unit

$$\Delta \equiv \beta - \alpha + 1 = 0.$$  

First, we note that a power law distribution describes well the information footprint above some minimum quantity $X_{min}$. We fit the histogram of firms by the number of records, articles, and topics in Figure 4 using a standard method (17). We then find maximum likelihood exponents of $\alpha = 1.819 \pm 0.004$, $\alpha = 1.854 \pm 0.003$, $\alpha = 1.912 \pm 0.004$, and $\alpha = 1.747 \pm 0.001$ for records, articles, sources, topics, respectively. We note that the fit to topics goes through the data, but is statistically distinguishable from a power law and is limited in its range. Perhaps because of these limitations, we find that the exponent scaling relation in Eq 2 is least well satisfied for topics, $\Delta = 0.15 \pm 0.02$, but for the remaining quantities $\Delta = -0.10 \pm 0.01$, $\Delta = -0.11 \pm 0.02$, and $\Delta = -0.06 \pm 0.02$ for records, articles, and sources, respectively. That the exponents are close to satisfying the exponent relation Eq 2 — despite the fact that $\beta$ has been extracted for a small subset of public firms in COMPUSTAT but $\alpha$ from all firms in the reading data — indicates that our scaling approximation is reasonably self-consistent.

In contrast with the economic footprint, which are presumably constrained by physical (18), or even metabolic, limitations, the largest firms can have a disproportionately larger presence in the information dimension. Heavy-tailed distribution of firm information search reveals strong inequality in how firms read, suggesting that the benefits of new information are worth the costs.

Fig. 3. Scaling of record count with measures of capital (a) assets, (b) plants, property, equipment (PPE), (c) employees, and (d) sales. Each blue point is a firm in COMPUSTAT and our data set. Mining (green), service (red), and all other (gray) firms. Orange line shows a power law fit to $Y \sim X^\beta$ with exponents (a) $\beta = 0.65 \pm 0.02$, (b) $\beta = 0.74 \pm 0.02$, (c) $\beta = 0.73 \pm 0.01$, and (d) $\beta = 0.72 \pm 0.01$ using one standard deviation from bootstrapped fits as error bars. Fitting range $R \geq 9$ given by the power law fit in Figure 4. We highlight well-known firms Microsoft, Apple, Zendesk, and BioNTech. BioNTech had no reported sales in 2018.

These valuations and record counts correspond to public firms like The Geo Group, Inc., Chicago Board Options Exchange, and CAE in the first group, NCS Multistage, Casa Systems, and College Pharmaceutical in the second, or transportation and logistics company Grupo TMM in the smallest.
of information $I$, such as that accessed in a single record,

$$C^* \sim I^\gamma.$$  \[3\]

In addition, we expect that different aggregations of information obey different scalings with $I$ because they will contain redundant information. We write the scaling with information as $A \sim I^{\mu A}$ for number of articles $A$ as an example, where we would anticipate larger aggregations to have smaller scaling exponents $\mu$, reflecting correlated information as articles are grouped into sources and topics. The scaling then implies that $C^* \sim A^{\gamma/\mu A}$. If the inflection points all occur at the same threshold unit cost $C^*$, then for any pair of articles, sources, and topics denoted as $X$ and $Y$ we have the scaling $X \sim Y^{\mu X/\mu Y}$. Note that the scaling exponent $\gamma$ does not matter for the ratio. Importantly, we must put articles, sources, and topics into the same units, which we can do by measuring the inflection point in terms of records. From the scaling relations and the inflection values of $R$ cited earlier, we find the ratios $\mu_A/\mu_S \approx 1.9$ and $\mu_S/\mu_T \approx 1.6$. This suggests that the change in granularity from articles to sources and from sources to topics is comparable though the latter is less dramatic. Reassuringly, this implies that larger aggregates show slower scaling with $I$, implying that the information units in articles are correlated given a particular source and that sources are correlated given a particular topic. Furthermore, this implies that the effective information cost $C^*$ increases for a firm of fixed size with information granularity. That it is more expensive to be diverse per topic than per unit source than per unit record presumably reflects organizational or environmental changes that influence how firms are exposed to new, distinct information.

If total information costs are increasing with firm size, then information constraints may matter more as firms become larger. Consistent with this hypothesis, we find that whether we consider articles, sources, or topics the spread amongst firms narrows with total information consumed. The collapse in the distribution is especially prominent around a fixed firm size of $10^4$ records;\(^1\) this suggests that large public firms are increasing confined to a deterministic trajectory in terms of information diversity, echoing observations in economic growth rates (8, 15, 19). As we discuss in the next sections, the narrowing in the distribution is explained by constrained growth trajectories of large firms in the space of topics.

**The topic space**

To track firm trajectories in the space of topics, we reconstruct the graph of the underlying space of topics by using topic co-occurrence in the universe of articles. As a measure of relatedness, we count the number of joint appearances $n(A, B)$ of topics $A$ and $B$ with their individual frequencies $n(A)$ and $n(B)$ across the set of accessed articles,

$$r(A, B) \equiv \frac{2n(A, B)}{n(A) + n(B)}. \quad [4]$$

Eq 4 returns $r(A, B) = 0$ when the topics never appear together and $r(A, B) = 1$ when the topics only appear together and never alone. By adjusting the threshold value of $r^*$ at which we draw an edge between two topics, we find a sudden transition where a small change in the threshold leads to a single connected component of nearly all topics. We define a percolation point $r^*$, where a giant connected component constituting half of all topics emerges. The exact threshold value depends on the particular day on which data was collected because the set of articles read changes, but it is always

\(^1\)Such firms typically have assets of about $1.4$ billion and annual sales of about $2$ billion dollars. This includes firms like Infosys, Ryder, and Labcorp when considering all firms within 40% of these values.

---

Fig. 4. (a-d) Scaling of information variables with sales. Mining (green), service (red), and all other (gray) firms. Black X’s are medians for logarithmic bins, and orange line is the scaling fit. Scaling exponents are $\beta = 0.72 \pm 0.01$, $\beta = 0.75 \pm 0.02$, $\beta = 0.85 \pm 0.02$, and $\beta = 0.89 \pm 0.02$ using bootstrapped error bars. Lower limits of fit are given by best fits to distributions below except for topics, where the lower cutoff can be set $a$ priori because the topic relevancy vector is of size 10. (e-h) Distribution of the number of records $\alpha = 1.82$, articles $\alpha = 1.85$, sources $\alpha = 1.91$, and topics $\alpha = 1.75$ accessed by firms show power law scaling beyond a lower cutoff. A standard fitting procedure involving maximum likelihood for the exponent $\alpha$ with the Kolmogorov-Smirnov statistic for the lower bounds returns $x_{\min} = 9$, $x_{\min} = 6$, and $x_{\min} = 6$ (17). For topics, we do not fit the lower bound given the strictly limited range of topics, but this also makes clear that our simplified scaling model only captures scaling with records in panel d in spirit. Arrows indicate where Microsoft, Apple, Zendesk, and BioNTech would fall on the power law curve. See Table S3 for further exponents.
close to $r^* \sim 10^{-2}$. Regardless, many properties of the giant component are preserved, and it serves as the best-mapped portion of the topic space in which firms live.

The topic graph appears to be a “tangled hairball,” where the vast majority of topics are connected to a few other topics and a small number are connected to many like the knots in the eponymous analog. This feature is reflected in the degree distribution plotted in Figure S4, where we compare the distributions for a changing threshold (above and below the halfway point) and a Poisson random graph model, where every pair of topics is connected with some small probability equal to the average degree of the nodes in the data. The degree distribution of the topic graph indicates that there are dense clusters of topics that are well-connected to each other, but that many topics form relatively separate chains of nested topics usually of increasing specialization within some industry.

A closer inspection of the graph reveals that the clustering we find in the graph to be qualitatively sensible. We show examples of the most well-connected topics in Table S7, which have to do with general topics of mundane but universal use such as shipping or security including “Ring Central,” “Application Aware Network,” “Packing Supplies,” “Patient Care,” etc. Isolated clusters of topics that form a “hair strand” out from the more densely connected core dealt with groups of specialized topics related to, for example, retirement investment, office equipment, or voice over internet protocol (VoIP) as shown in the inset of Figure 5. These are distinct from the most frequently accessed topics that have much closer relationships to consumer interests, which include “Social Media,” “Discipline,” “Vacation,” “Trade Notes,” “Health Promotion/Recreation/Wellness Benefits,” etc. On the other hand, a handful of topics are ubiquitous across firms, appearing in over 90% of firms with at least $10^2$ records as in Table S1. While many of these topics are again mundane and overlap with the previous categories, we also have the emergence of social issues like “Sex Discrimination” and “Conflict Resolution” and when going down to 80% ubiquity “Immigration,” “Globalization,” and “Generational Difference.” Ubiquitous topics are “furniture” topics — ones that serve a necessary, everyday purpose — that reflect both information that any firm must consume as part of the overhead of employing people and topics that are repetitively refreshed because they are relevant. As a result, firms fill out the space of topics as they grow as we show for the examples Apple, Microsoft, Zendesk, and BioNTech on the right side of Figure 5 rather than moving from one exclusive sets of topics to another. Thus, the topic graph reflects how similar topics are connected because of common information.

Fig. 5. (left) Largest connected component of topic graph at threshold point representing 50% of all topics in the dataset. Most common topics are reproduced in Table S7. Produced from data on June 10, 2018. Projection of example firms Apple, Microsoft, Zendesk, and BioNTech (using data from June 12, 2018) on topic graph. Topics read by the firm are blue, whereas gray markers are unread.
needs and thus reasons for why firms come to encompass an increasing diversity of topics over time.

Information search constrained by topic graph

The topics that a firm spans on the graph from Figure 5 reflect the diversity of ideas within a firm at a given point in time, but they do not seemingly reveal more than a snapshot alone. Related research, however, shows that firms typically grow in economic size as they age such that we expect older firms to be larger economically and, in the context of scaling relations from Figures 3 and 4, more topically diverse (15). The corresponding explanation from the management literature is that firms run on two parallel tracks, focusing on their core competencies and less often looking for opportunities to expand them (20). This observation suggests that a model for firm growth should incorporate how they expand into neighboring regions of idea space to connect the scaling patterns (21–23).

The way that firms expand is determined by the dynamics of how they choose and the underlying structure that constrains their choices (24, 25). As a way of separating out these factors, we start with the simplest, null explanation that firms are unconstrained by the topology of the topic space and so can jump from any topic to additionally encompass any other. Since some firms concentrate exclusively on a single or a few topics, we also consider an affinity probability p with which the firm refrains from choosing a random topic to add to its portfolio. Consequently, with probability 1 − p, the firm explores by sampling again from the set of all topics. To test the predictive power of the model, we scan the range of affinity p while fitting to only the mean of the Heaps’ plot for the limited range of R ≤ 10^2 against the number of topics T. This range is below the point at which the strong curvature imposed by the upper bound on topics manifests. After selecting the parameter values that optimize the limited fit, we check whether it does well for the remaining curves. We compare the null model with optimal fit parameter p^* with the data in Figure 6. Unsurprisingly, the structureless null model fails to match the Heaps’ plot, overestimating the maximum in panel b, exceeding the sublinear scaling of the minimum in panel d, and predicting hardly any gap between the two.

At the other extreme, we might hypothesize that firm search is governed completely by the structure of the topic graph. In other words, firms randomly walk in the graph such that their exploration choices are uncorrelated in time and apparent correlations in topics are only given by the structure of the graph. As before, we allow for the possibility that some firms have an affinity p for returning to a topic that they have already incorporated and 1 − p for randomly moving to any adjacent topic, which may or may not already be in its portfolio. This dynamic entails a tendency for firms to focus on their core competencies but otherwise explore randomly the topic space as if without foreknowledge of whether or not exploration leads to a new topic.

By systematic exploration of the parameter space, we find that the structured diffusion model predicts closely the Heaps’ patterns in the data for p^* = 1/20 (see SI Section D for other models). We find close agreement with the maximum (defined as 95th percentile) and mean for the topics Heaps’ plot in Figure 6 and, compared to the jump null, significantly closer alignment with the minimum as defined by the 5th percentile. While highly repetitive readers pose a complication for all models we consider, we note that some of these data points likely represent bots. Furthermore, a mixed strategy that combines both jumping and local exploration, inspired by exploration vs. exploitation dynamics found in ecological search dynamics (26, 27), is worse (Figure S10). Thus, the typical trajectory of small firms predicts large firm diversity and maximally diverse firms of all sizes, which is consistent a mixed exploitation and local exploration strategy in the population that is constrained by the structure of the topic space.

Discussion

The information footprint of a firm represents a complementary dimension to its economic size, which we study here with a high-resolution dataset of employee news reading habits. As a starting point, we show that the volume of firm online reading is related to its diversity in the form of three Heaps’ plot for read articles, sources, and topics in Figure 2. The plots present several curious features that shed light on the meaning of the information footprint and how firms explore the space of ideas.

Our first mystery from the Heaps’ plots was that the most diverse firms fail to saturate the maximum theoretical possible diversity above a critical size. We note that the point at which firms deviate from the theoretical maximum depends on the granularity of the information source. As we increase the granularity from articles to sources to topics the size

\[ N \sim 10^p \]
of the typical corresponding firm shrinks from $R = 2.200$, $R = 90$, and $R = 15$ records, or annual sales of $700$ million and $60$ million to $20$ million estimated for public firms. We might expect the decrease in transition point reflects structural constraints on how firms gather information.

In order to make this connection, we show how the information footprint of a firm is related to its economic size in Figures 3 and 4. Interestingly, we measure several of the scaling exponents with records and articles to be notably close to $3/4$ (see SI Table S3). The reason such scaling is conspicuous is because quarter-power scaling is a fundamental characteristic of metabolic scaling in biology from cells to whales, which derives from energetic limits given the geometry of the circulatory system (28, 29). In particular, Kleiber’s Law is an example of remarkable convergence between organisms across taxa, where metabolic rate scales sublinearly with mass to the $3/4$ power (30). Similar, if different, energetic principles have been argued for the development of cities (18, 31). Here, by entertaining the analogy to biology, we would think of the number of records as a measure of “mass,” while the economic measure would reflect “metabolism.” Importantly, reading amount is not simply proportional to employee number, which indicates that the “mass” consuming information is not employees but most likely internal organizational structures that make up the collective knowledge of a firm (32). It remains unclear, however, how the internal organization of a firm distributes economic resources throughout the “body” of the firm, and it is also unclear why it would adhere to physical constraints akin to that of a circulatory system. This presents an intriguing and open line of inquiry about how the internal organizational structure of a firm determines its information uptake and vice versa.

One implication of the sublinear scaling of economic with information size is that the effective cost of an additional unit of information shrinks, signaling an economy of scale. This observation aligns with the intuition that the tools of the knowledge economy enhance firm productivity (34, 35). As a result, large firms consume more articles per employee, which implies that their employees are more information productive on the whole. This leads to an information inequality that is an exaggerated version of the economic inequality between firms given by Axtell’s classic finding of Zipf’s law in US Census data (7). While the economy of scale holds of all levels of information granularity, its degree varies. We show this by starting with a fundamental unit of information that turns out to be aggregated into larger, increasingly redundant supersets as we go from articles and sources to topics. Assuming that the atomic unit of information always has the same cost, we find that the reason a firm of a given information volume is relatively smaller in source and topic diversity is because the costs of diversity for an aggregate unit increases. This is summarized in the scaling exponents $\beta$ in Figure 4, which are the smallest for records and articles and the largest for topics. One possible reason is that processing diverse information at larger scales implies physical restructuring of the firm such as another division or a new mission statement, whereas reading a new article on the same topic is trivial. This would mean that firm growth in information space at the coarsest levels is more tied to economic growth, echoing the role of intangible aspects in determining firm costs (36). These observations suggest two different ways in which information costs may modulate firm growth, delivering an increasing economy of scale and affecting firm diversification in the information space.

As a closer look at what effects determine how firm diversity scales with size, we build a model of how firms grow in the space of topics. To do this, we first map out the space of ideas as is represented by the topics we extract from the articles by using the fact that there is a sharp similarity threshold at which the topics cohere into a single large connected component. As we show in Figure 5, we find that the underlying space at that point resembles a tangled hairball, where a few topics connect many but the vast majority are only connected to a few others in long strands of specialization. How such structure may be similar to a physical network of inputs and outputs such as the product space of nations presents another direction for future work (22).

In an echo of how economically complex regions cumulatively build on the product portfolio (37), we find topics that are ubiquitous amongst firms. Though ubiquitous topics can be central to the network as measured by eigenvector centrality, non-central topics are never ubiquitous. Ubiquitous topics are indicated by a central location in the core of the network but also by non-structural factors such as social relevance in the cases of “Sex Discrimination” and “Immigration.” Larger firms thus do not outgrow the ubiquitous topics, which we show are basic and mundane but seem to be necessary just like “furniture,” that come with the basic needs of firm organization such as employing human labor.

Using the topic graph, we consider how firms grow and what that reveals about constraints and strategies for expansion. By considering a simple model with minimal search dynamics that are either unconstrained or constrained by the graph topology, we show that the topic space is essential. We find that a population of firms with a weak tendency to focus on the existing portfolio, or “affinity,” successfully predicts how firms scale in topic diversity both in terms of the mean and maximum scaling. Our model makes the assumption that firms tend to spread to nearby areas of the topic graph in form of local exploration (20). As a result, our model also implies that firms starting out in highly specialized topics are slower to spread to new regions of the graph compared to firms that seed their growth in highly connected regions of the graph, a kind of competitive advantage (38). Here, we do not have the longitudinal data to test such a hypothesis, but this suggests a potentially interesting question of whether the information content of a firm earlier in the life cycle of a firm is predictive of its later success.

These macroscopic trends suggest ways to distinguish firms from one another or to highlight unusual ones. While one has to be careful with startups and the smallest firms which show much more variability in information and economic behavior (15), our scaling models establish a basis for comparison relative to an expected trend. Large deviations from patterns extracted over many millions of firms are likely to represent surprising activity that demands further attention. Our work suggests a way forward for a calibrated metric of firm performance by going beyond economic measures and following instead the information footprint.
Acknowledgements
E.D.L. thanks the CSH theory group including Tuan Pham, Jan Korbel, and Stefan Thurner and others including Chris Kempes and Geoffrey West for useful discussions. E.D.L. acknowledges funding from BMBWF, HRSM 2016 (Complexity Science Hub Vienna), F.N. acknowledges financial support from the Austrian Research Agency (FFG), project #873927 (ESSENCSE).

Data availability
Code for analysis will be made available in a code repository to be determined upon publication, but data presented are proprietary, and we will consider requests for data access for reproducibility on a case-by-case basis.
A. Data preprocessing

The data set consists of access records of news content that is within the universe of publishers monitored by a firm specializing in “intent” data, hereafter called the “Company.” As we specify in reference (12),

‘Intent’ refers to a recent strain of data analytics aiming to gauge a prospective business customer’s buying interest based on patterns of reading on the internet. If a given business customer’s reading on a particular topic increases at a rate relative to a baseline, one might presume the business customer has a greater ‘intention’ of transacting in a related service as this increase in reading is indicative of the research that occurs prior to a purchase.

Each record, in principle, does not identify who is accessing the content, only the origin of the request, or the IP address. This information, however, can be linked to known servers operated by firms inferred from their domains, which then allows us to obtain the set of access records for linked to individual devices associated with the firm. The Company analyzes the textual content and uses proprietary topic model implementation to extract a set of hand-labeled, customer-relevant topics associated with each page of content. Thus, we have a database of who is accessing what and information about the content in question.

The data is proprietary and the analyses are run on big data sets, so we cannot feasibly rerun our own preprocessing. However, we can take some steps to mitigate potential limitations of the data set.

1. One consideration is that the topic list is updated over time to account for changing data sources and customer interests. To limit this time-varying effect, we focus on a two-week period between June 10, 2018 and June 23, 2018, for which we do not have any reason to believe is unusual relative to other points in the data set, which represents a large number of labeled topics, and that does not have changes in topic preprocessing. Within this time frame, the preprocessed data consists of 4,338 unique topics with more than 3.5 million firms as identified by their domains, and day-to-day statistics are similar.

2. The articles are not identified uniquely. In order to determine them uniquely for a given source, we combine the URL along with information about the first two topics (names and relevancy scores), which defines a new article ID. We can check the accuracy of the new article ID by using the fact that no article can be labeled with more than 10 topics. If we then aggregate articles by this new ID and count the number of articles that have more than 10 distinct topics, we find that they compose below 0.03% of articles for every day or that our article ID is highly accurate.

3. Some of the extracted topics may not describe well the content at hand, so we impose a lower threshold on the topic relevance value. We first calculate the typical relevance value. For firms whose average relevancy falls in the bottom 5th percentile, we remove firm from consideration as we show in Figure S1. This allows us, in our cross-firm comparison, to only consider firms that are well-represented within our given, bounded universe of topics.

4. We exclude Amazon from fitting our power law models because it is a clear outlier by over an order of magnitude in terms of records, which most likely indicates that customer behavior was incorrectly flagged as employee behavior such as through Amazon Web Service. This is also a possible source of error in other cloud or internet service provider firms like Comcast and Microsoft, but there is no clear indication for how to treat them as outliers as is the case with Amazon.

As for other preprocessing steps,

1. For our analysis, we subsample the full data set to a representative 10% to reduce computational load using PostGreSQL’s TABLESAMPLE function. Since this entails sample sizes of $10^6$, some results such as exponents for power-law distributions do not change significantly.

Once the above steps have been taken, we have a selection of firms with some basic statistics plotted in Figure S2. Thus, we are considering roughly $10^6$ to $10^7$ firms per day and the distribution of unique topics accessed by any given firm displays a power law tail.

Besides businesses, the data set contains government agencies, non-governmental organizations, academic institutions, and entities registered with a domain. While we could filter out some of the organizations, it is unclear if such pruning can be done in a consistent way across all countries especially when considering that firms in many countries are closely tied to the government such as Saudi Arabia’s Aramco and Russia’s Gazprom to name a couple. As a check of how large an effect government organizations might have on the data set, we took a large list of 9,040 US government and government-affiliated URLs from https://github.com/GSA/govt-urls in addition to top-level “gov” and “mil” domains to find that of over 550,000 domains from the 10% subsample on June 11, 2018, only 4,007 were US government affiliated. In terms of records, they constitute about 4.8% of the daily subsample. As for academic institutions, we consider them to be firms for the purposes of our analysis, but domains ending containing a .edu suffix (either alone or followed by country-specific domain) also constitute about 5,475 and about 12% of the day’s records. Thus, we expect that the vast majority of the data we analyze corresponds to non-government, non-educational entities.
B. Poisson graph model

In order to fit the Poisson graph model of the degree distribution of a firm in topic space, we must account for the fact that the subgraph, by definition, does not include nodes of zero degree. Yet, the number of nodes of zero degree are essential for calculating the mean degree and thus for fitting the model.

By accounting for the possibility of such nodes, we obtain a self-consistency condition that we can use to solve for both the expected number of nodes of degree zero \( n_0 \) and the mean degree \( \bar{d} \) that accounts for those nodes. The condition is

\[
\bar{d} = \frac{1}{n_> + n_0} \sum_{i} d_i. \tag{5}
\]

We know how many nodes of degree greater than zero \( n_> \) from the data to calculate the mean degree \( \bar{d} \) from the model, and we must sum over all nodes to get the total outgoing degree. A result of one such fit is shown in Figure S4b.

C. Topic graph

For the simulation that we show in the main text, the topic graph is constructed from the statistics on a single day for computational efficiency. This means that we take all the articles that are read on that day and use them to calculate the similarity measure defined in Eq 4. While the graph is not exactly the same should we use other days, it is the case that its statistical properties are conserved. As an example, we construct the graph using our 10% subsample of the entire two-week period of observation, and we find that features like the degree distribution, average connectivity, etc. are conserved.

D. Model specification

We simulate firm growth as a random walk on the topic graph, where at each time point we select a random topic either from
Table S1. Ubiquitous topics that appear in over 90% of firms with at least $10^2$ records on June 10, 2018. These represent hand-labeled topics to identify clusters of business-relevant publications, i.e. “Vacations” label articles relevant to the vacation industry such as economic studies of travel or popular places to vacation, “Best Places to Live” could be about work-at-home or housing policies. Popular topics can also indicate intensive marketing such as for “Call of Duty (COD),” a popular video game that had a major release in mid-October of 2018 including a major news release on June 11. Importantly, the topics represent information flow into firms.

| Topic                                      |
|--------------------------------------------|
| Health Promotion / Recreation / Wellness Benefits |
| Best Places to Live                        |
| Interpol                                   |
| Out of Home                                |
| Call Of Duty (COD)                         |
| Tractors                                   |
| South By Southwest (SXSW)                 |
| Group Calendars                           |
| GoFundMe                                   |
| Blu-ray                                    |
| Vacations                                 |
| Discipline                                 |
| Faucets                                    |
| Trade Notes                               |
| Conflict Resolution                       |
| Pattern Recognition                       |
| BBC                                        |
| Home Decor                                |
| Holiday Season                            |
| Live Streaming                            |
| Sex Discrimination                        |
| Social Media                              |

Table S2. Optimal simulation parameters from minimizing the distance for Heaps’ law for $R \leq 10^2$.

| Simulation Type   | $p^*$ |
|-------------------|-------|
| Jump              | 0.33  |
| Walk + Jump       | 0.44  |
| Walk              | 0.05  |
| Self-loop         | 0.19  |

In Figure S10, we show an overview of the different models that we consider and their deviations from the main text, maximun, and minimum Heaps’ curves. We find that the random walk model generally does better than the others and the self-loop model a close second. The logarithmic errors are shown in Figure S10.

1. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–688.
2. Deming DJ (2017) The growing importance of social skills in the labor market. *The Quarterly Journal of Economics* 132(4):1593–1640.
3. Neffke FM (2019) The value of complementary co-workers. *Science Advances* 5(12):eaax3370.
Fig. S6. Most connected topics in topic graph at percolation point as found for each day separately. Topic labels indicate that the content of the article has to do with the indicated label. Topics have to do with business-relevant articles about "Vacation" related industries and not necessarily about where one might vacation.

| Topic Label | Connected Topics |
|-------------|-----------------|
| Hotel       | Hotel, Vacation |
| Restaurant  | Restaurant, Vacation |
| Cruise      | Cruise, Travel |
| Resort      | Resort, Vacation |
| Vacation    | Vacation, Travel |
| Travel      | Travel, Cruise |
| Hotel       | Hotel, Vacation |
| Vacation    | Vacation, Travel |
| Cruise      | Cruise, Travel |
| Resort      | Resort, Travel |
| Travel      | Travel, Cruise |

Fig 5b: shows connected topics in topic graph at percolation point as found for each day separately. Topic labels indicate that the content of the article has to do with business-relevant articles about "Vacation" related industries and not necessarily about where one might vacation.
### Fig. S7. Top 10 most popular topics on each day by number of records containing topic.

| 2018-06-10 | 2018-06-11 | 2018-06-12 | 2018-06-13 | 2018-06-14 | 2018-06-15 | 2018-06-16 | 2018-06-17 | 2018-06-18 | 2018-06-19 | 2018-06-20 | 2018-06-21 | 2018-06-22 | 2018-06-23 |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Social Media | Discipline | Social Media | Discipline | Social Media | Discipline | Social Media | Discipline | Social Media | Vacations | Vacations | Vacations | Vacations | Vacations |
| Discipline | Social Media | Discipline | Social Media | Vacations | Disciplin | Social Media | Vacations | Social Media | Social Media | Discipline | Discipline | Discipline |
| Health Promotion / Recreation / Wellness Benefits | Health Promotion / Recreation / Wellness Benefits | Discipline | Health Promotion / Recreation / Wellness Benefits | Discipline | Health Promotion / Recreation / Wellness Benefits | Discipline | Health Promotion / Recreation / Wellness Benefits | Discipline | Health Promotion / Recreation / Wellness Benefits | Social Media | Social Media | |
| 3 Vacations | Vacations | Vacations | Health Promotion / Recreation / Wellness Benefits | Vacations | Health Promotion / Recreation / Wellness Benefits | Vacations | Health Promotion / Recreation / Wellness Benefits | Discipline | Social Media | Social Media | Health Promotion / Recreation / Wellness Benefits | |
| 4 3D Animation Software | Trade Notes | Trade Notes | Trade Notes | Home Decor | Home Decor | Home Decor | Checking Account | Home Decor | Home Decor | United States Border Patrol | Globalization | GoFundMe | Home Decor |
| 5 Trade Notes | Blu-ray | Home Decor | Home Decor | Checking Account | Trade Notes | Out of Home | Retirement Planning | Blu-ray | Blu-ray | Globalization | Trade Notes | Trade Notes | Blu-ray |
| 6 Home Decor | Home Decor | Blu-ray | Blu-ray | Trade Notes | Out of Home | Nimble | Home Decor | Checking Account | Trade Notes | Trade Notes | Home Decor | Home Decor | Holiday Season |
| 7 Blu-ray | Checking Account | Interpol | Interpol | Retirement Planning | Nimble | Waterfall | Blu-ray | Trade Notes | Globalization | Home Decor | United States Border Patrol | Blu-ray | GoFundMe |
| 8 Checking Account | 3D Animation Software | GoFundMe | GoFundMe | Out of Home | Blu-ray | Word of Mouth | Feuoste | Retirement Planning | Interpol | GoFundMe | Globalization | Trade Notes | |
| 9 Retirement Planning | Best Places to Live | Checking Account | Checking Account | Blu-ray | Waterfall | Medical Aesthetics | Trade Notes | Interpol | Immigration | Blu-ray | Holiday Season | Holiday Season | Best Places to Live |
Fig. S8. Random walk model with self-loops to account for firm topic repetition. Fit is not as close as with random walk model with random return to a previously explored site with affinity probability as in Figure 6.

Fig. S9. Random walk model with jumps to any other node. Fit is not as close as with random walk model with random return to a previous step in the walk with affinity probability as in Figure 6.

Table S3. Scaling exponents for economic against information footprint. Error bars represent one standard deviation over bootstrapped fits. Exponent estimates overlapping with 3/4 are bolded.

| asset   | PPE | employees | sales   |
|---------|-----|-----------|---------|
| records | 0.65 ± 0.02 | 0.74 ± 0.02 | 0.73 ± 0.01 | 0.72 ± 0.01 |
| articles| 0.68 ± 0.01 | 0.77 ± 0.02 | 0.76 ± 0.01 | 0.75 ± 0.01 |
| sources | 0.76 ± 0.02 | 0.89 ± 0.02 | 0.86 ± 0.01 | 0.85 ± 0.01 |
| topics  | 0.85 ± 0.02 | 0.92 ± 0.02 | 0.88 ± 0.01 | 0.89 ± 0.01 |

4. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036–1039.
5. Henrich J, Boyd R (2002) On modeling cognition and culture. Journal of cognition and culture 2(2):87–112.
6. Barro RJ, Sala-i-Martin X (2003) Economic Growth. (MIT Press), Second edition.
7. Astell RL (2001) Zeta Distribution of U.S. Firm Sizes. Science 293(5536):1818–1820.
8. Stanley MH, et al. (1996) Scaling behaviour in the growth of companies. Nature 379(6568):804–806.
9. Lee Y, Amaral LAN, Canning D, Meyer M, Stanley HE (1998) Universal Features in the Growth Dynamics of Complex Organizations. PHYSICAL REVIEW LETTERS 81(15):4.
10. Barney J (1991) Firm resources and sustained competitive advantage. Journal of management 17(1):99–120.
11. Kogut B, Zander U (1992) Knowledge of the firm, combinative capabilities, and the replication of technology. Organization science 3(3):334–345.
12. Kwan A, Zhu C (2019) Does Internet Research Activity by Sophisticated Investors Lead to Heightened Adverse Selection? SSRN Journal.
13. Kwan A, Liu Y, Matthews B (2022) Institutional Investor Attention.
14. Heaps HS (1978) Information Retrieval: Computational and Theoretical Aspects. (Academic Press).
15. Zhang J, Kemps CP, Hamilton MJ, West GB (2021) Scaling laws and a general theory for the growth of companies. arXiv:2102.03547 [physics].
16. Powell WW, Snellman K (2004) The Knowledge Economy. Annu. Rev. Sociol. 30(1):199–220.
17. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. SIAM Rev. 51(4):661–703.
18. Bettencourt LMA (2013) The Origins of Scaling in Cities. Science 340(6139):1438–1441.
19. Perline R, Axelrod R, Totebaum D (2006) Volatility and Asymmetry of Small Firm Growth Rates Over Increasing Time Frames. (SBA Office of Advocacy), Technical Report 285.
20. Wernerfelt B (1984) A resource-based view of the firm. Strat. Mgmt. J. 5(2):171–180.
21. Teo CE, Rumelt R, Dosi G, Winter S (1994) Understanding corporate coherence. Journal of Economic Behavior & Organization 23(1):1–30.
22. Hidalgo CA, Klinger B, Barabási AL, Hausmann R (2007) The Product Space Conditions the Development of Nations. Science 317(5837):482–487.
23. Lee ED, Kemps CP, West GB (2022) Idea engines: A unified theory of innovation and obsolescence from markets and genetic evolution to science. arXiv:2203.14617 [physics].
24. Perrotte ET (2009) The Theory of the Growth of the Firm. (Oxford University Press, Oxford ; New York), 4th ed., rev. ed. edition.
25. Staufler G, Aharonov A (1994) Introduction to Percolation Theory. (CRC Press).
26. Schlesinger MF, Klafter J (1986) Lévy walks versus Levy flights in Growth and Form, eds. Stanley HE, Ostrowski N. (Martinus Nijhof Publishers, Amsterdam), pp. 279–283.
27. Reynolds AM, Rhodes CJ (2009) The Levy flight paradigm: Random search patterns and mechanisms. Ecology 90(4):877–887.
28. Savage VM, Deeds EJ, Fontana W (2008) Sizing Up Allometric Scaling Theory. PLoS Comput. Biol. 4(9):e1000171.
29. West GB, Brown JH, Enquist BJ (1997) A General Model for the Origin of Allometric Scaling Laws in Biology. Science 276(5309):122–126.
30. Kleiber M (1947) Body Size and Metabolic Rate. Physiological Reviews 27(4):511–541.
31. Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proc. Natl. Acad. Sci. 104(17):7301–7306.
32. Kogut B, Zander U (1992) Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. Organ. Sci. 3(3):383–397.
33. West GB (2017) Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. (Penguin, New York).
34. Stroh KJ (2022) Information Technology and the U.S. Productivity Revival: What Do the Industry Data Say? The American Economic Review p. 19.
35. Black SE, Lynch LM (2004) What’s Driving the New Economy? The Benefits of Workplace Innovation. The Economic Journal p. 20.
36. Black SE, Lynch LM (2001) How to Compete: The Impact of Workplace Practices and Information Technology on Productivity. The Review of Economics and Statistics 83(3):434–445.
37. Hidalgo CA, Hausmann R (2009) The building blocks of economic complexity. Proc. Natl. Acad. Sci. U.S.A. 106(26):10570–10575.
38. Barney J (1991) Firm Resources and Sustained Competitive Advantage. Journal of Management 17(1):99–120.