Measuring Productivity and User Engagement from Workplace Network Interactions

Sanjiv R. Das, Kailin Hu,
Mugdha Potdar, Preethi
Subbrayan Ranganathan,
Xiang Zhang

Santa Clara University
500 El Camino Real
Santa Clara, CA 95053

*srdas, khu, mpotdar,
psubbrayanranganatha,
xzhang11 @scu.edu*

Ivan Galea, Mark Scarr

Atlassian, Inc.
321 E Evelyn Ave
Mountain View, CA 94041

mscarr, igalea@atlassian.com

Anand Subramanian,
Hariharan Swaminathan

Box, Inc.
900 Jefferson Avenue
Redwood City, CA 94063

*asubramanian, hswaminathan
@box.com*

ABSTRACT

The productivity of a workforce is very hard to measure for non-routine types of work. This paper develops a new formula for measuring productivity based on non-social networks using unique data on file-sharing activities within a company. This formula integrates individual productivity with a measure of teamwork, using a rich set of graph-theoretic constructs. The resultant formula has many attractive properties and is easy to compute across time and companies, and may be used for a large-scale study of workplace productivity. We present results for a sample of 525 companies, followed for 16 weeks. An example application shows that the percent of weekly active users can be predicted with high levels of accuracy.

Keywords: productivity; networks; file-sharing; engagement

1. INTRODUCTION

Measuring *productivity* in the workplace has been a focus of companies since the industrial revolution. Whereas it is easy to measure work that is based on measurable individual outcomes, as in (physical) factory production tasks, or paper processing tasks such as (cognitive) paralegal work, etc., much of the output of today's workforce is team-based and not individual. Many modern workplace tools are designed to make coordinated work easier and efficient, yet the ability to measure the effect of coordinated work versus individual work is poor. A McKinsey Consulting report [5] highlights the role of collaboration in organizational productivity.

Organizations have many systems that measure individual productivity and performance, but determining the productivity of a company as a whole is hard to do, though aggregate financial measures based on accounting and stock price performance are commonly used. Another aspect of productivity that is not measured is how much an individual in a company contributes to the company's overall productivity. In this paper, we develop a novel productivity score based on a mathematical framework that employs unique data to generate (i) a score for both individual and company productivity, (ii) a decomposition of company productivity into that contributed by each employee, and (iii) a breakdown of company productivity into that which comes from individual versus collaborative effort.

The productivity score can help managers evaluate their company's historic performance. It is normalized and therefore, it answers questions such as "How is my company performing compared to other companies in the same industry?" or "How can we compare productivity across different companies or industries and across time?". Our standardized productivity score is measured using the same variables for all companies using data from a unique dataset of workplace interactions.

Our new metrics of company, individual, and collaborative productivity will provide companies a picture of their *growth* patterns, and enable them to measure the "physics" of their growth, see [9]. External measures of productivity, mostly financial, are used to assess growth, and non-financial measures such as innovation productivity measured by patent counts [4] abound as well. In contrast, we offer an internal measure of companies' growth and productivity. Also, there are many approaches [8],[7] for measuring collaboration[1], but the effect of collaboration on company productivity is hard to measure. Our framework offers an objective way to do so.

While there is some evidence that social networks enhance workplace productivity [12] (and also hamper it)[2], we examine non-social workplace networks using file sharing data. This enables us to capture the effect of collaboration in a graph-theoretic manner. File-sharing networks embody complex, nested input-output structures that feed on each other in enhancing productivity and our data enables us to quantify these effects, very much as input-output structures have been used to characterize economic productivity, see [10], [11],[1].

[1] http://broadleafconsulting.ca/uploads/3/4/0/8/3408103/tools_for_measuring_collaboration.pdf.

[2] https://sloanreview.mit.edu/article/does-social-media-enhance-employee-productivity/.

The paper proceeds as follows. Section 2 describes the data we use. Section 3 explains our mathematical approach to construction of the network of file-sharing interactions. Section 4 introduces our novel productivity metric, and its decomposition by employee and by individual versus collaborative quantities is explained in Section 5. A large-scale data analysis is undertaken in Section 6, and we offer some concluding discussion in Section 7.

2. DATA

The data comes from an online file sharing and content management service, hosted by a major provider in this industry. The company provides cloud storage and file hosting for personal accounts and businesses. A useful aspect of this data is that file sharing activity amongst users within an organization and with external users is recorded. These data-sharing relationships offer a useful lens through which productivity may be measured.

Productivity may be defined as a function of data production and collaborative sharing. The structure of interactions is an additional facet of the data that may be used to measure productivity. This structure is uncovered from the network of interactions. Networks foster the spread of productive information and some network structures offer greater flow than others. Quantifying this flow will generate an additional aspect of productivity.

File sharing data is extensive, running into exabytes. This paper focuses on an extensive and anonymized subset of companies for which this data is extracted. The data set contains $\sim$ 30 million rows of data, covering $\sim$ 525 unique companies, over 16 weeks. The dates covered in this study run from 03/04/2018 through 07/14/2018.

3. NETWORK CONSTRUCTION

For each company on the file sharing platform, the daily interactions among users for a given company are used to construct the network. For each week we use the original data to create several metrics from the network of interactions. Alternatively, other granularities can be chosen, such as daily or monthly data.

The networks we construct are directed, weighted, possibly cyclic graphs. The network may be thought of as a weighted "edge list", i.e., links between sender and receiver, where a pair (i,j) will have value greater than zero if i shared files with j, else the value would be 0. As we will make precise in ensuing sections, this is easily developed from the data, and offers some choices in the construction of network weights, such as: (i) Frequency of interaction; (ii) The number of files shared in the interaction; (iii) The size (bytes) in the interaction; and (iv) an indicator variable if a sender and receiver interacted in a given period (usually a day). This graph is normalized so that all edge weights are in $(0,1)$. Our metrics below produce a composite measure for the file-sharing network.

The directed network graph G for a given company c at time (week) t can be defined as:

$$G(c,t) = \{V(c,t), E(c,t)\} \tag{1}$$

where $V(c,t)$ is the set of vertices for company c at time t and $E(c,t)$ the set of ordered pairs of vertices i.e. edges.

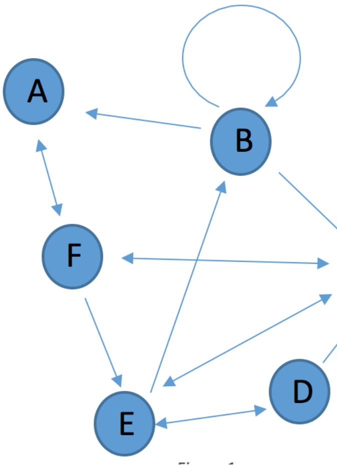


Figure 1: Example of a directed graph of user interactions.

We show an example of such a graph network of users for a company using Figure 1.

4. THE PRODUCTIVITY METRIC

A good measure of productivity should incorporate the quality, quantity, and transmission of information through the network. We propose the following novel metric, a single number P that describes the average productivity per person of a company derived from file-sharing data. This number is based on a flexible metric of file production undertaken by each person in the company, denoted by a vector Q, and the linkages between persons in the company given by a network matrix N, derived from file-sharing activity. The dimension of vector Q is n, the number of people (nodes) in the network, i.e., $Q_i, i = 1,2,...,n$. Correspondingly, the dimension of matrix N will be $n \times n$. In the next two sections we will describe the exact manner in which these two quantities, Q and N, are computed from the data.

We quantify the productivity per person in the company over time. The metric we use is defined as follows:

$$\begin{aligned} P &= \frac{1}{n} \cdot \sqrt{Q^\top \cdot N \cdot Q} \\ &= \sqrt{\frac{Q^\top}{n} \cdot N \cdot \frac{Q}{n}} \\ &= \sqrt{Q^{*\top} \cdot N \cdot Q^*} \end{aligned} \tag{2}$$

where $Q^* = Q/n \in \mathcal{R}^n$.

In our implementation, we compute Q and N weekly, for each company. So we may write

$$P(c,t) = \frac{1}{n} \cdot \sqrt{Q(c,t)^\top \cdot N(c,t) \cdot Q(c,t)}, \quad \forall c,t \tag{3}$$

where c indexes the company, and t indexes time. This equation implies that productivity, as denoted by scalar quantity P, increases if the elements of Q (*individual productivity*) increase, holding n and N constant. Likewise, ceteris paribus, if the elements of N (*collaboration connectivity*) increase, the metric also increases. This is intuitive, given that all values in Q and N are non-negative. We normalize the metric by dividing it by n, so that we measure productivity per person.

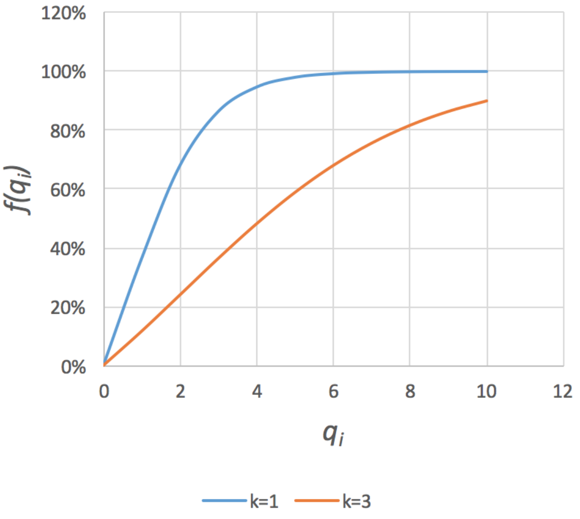


Figure 2: The specialized logistic function for quantity of information for any employee i, for two value of k. See equation 5.

The values in vector Q and matrix N are bounded in the range $(0, 1)$. Hence, P is a positive real number in $(0, 1)$. Because $P \geq 0$, cumulative P is a monotone increasing function over time.

In order to compare productivity over time, or across different entities, we make sure that P is normalized to the same scale. This is done by normalizing Q using the mapping function described in the next section.

4.1 Quantifying Individual Productivity

We compute the vector Q to quantify individual productivity of every employee (i.e., a node i in the file-sharing graph), by generalizing the standard logistic function, i.e., $Q_i = f(q_i) = 1/(1 + e^{-q_i})$, where q_i is the number of files generated by node i. This is a well understood and documented function that maps an unbounded q_i onto $(0, 1)$, with $f(0) = 0.5$. Our generalized form is

$$f(q_i) = a + m/(1 + c\, e^{-(q_i - q_0/k)}) \tag{4}$$

where a is the y-intercept, m is the curve's maximum value, k the steepness of the curve, c the asymmetry of the curve, and q_0 the value of the sigmoid's mid-point. Clearly, setting $a = 0$, $m = 1$, $k = 1$, $c = 1$, and $q_0 = 0$ produces the standard logistic function as a special case. Since we require positive numbers only and $f(0) = 0$, setting $a = -0.5$, $m = 1.5$, $c = 2$, and $q_0 = 0$ defines the functional form we use for any employee's entry in the Q vector as follows

$$f(q_i) = -0.5 + 1.5/(1 + 2e^{-(q_i/k)}) \tag{5}$$

Figure 2 shows the behavior of the plot for varying values of k. Clearly, varying the value of k controls how quickly $f(q_i)$ converges to 1.

Having defined our quantum of information mapping function, we have complete flexibility to adjust the shape of its curve to reflect different information profiles, in particular, differing magnitudes of information. The parameter k in equation can be estimated empirically from the data to achieve a specific shape for $f(q_i)$, or it may be set to a pre-determined value to meet certain characteristics. For example, in our data, if we desire $f(q_i)$ to approximately span as much of $[0, 1]$ as possible in a quasi-linear fashion, then we choose k as:

$$k = Q_{99}/5 \tag{6}$$

where Q_{99} represents the 99th percentile of the q values in the data. We used the 99th percentile (as opposed to the 95th percentile, for example) because we wanted to differentiate better the big companies from each other. This also ensures that the maximum remains the same for a long time period. Similar results are obtained if we use a simple mapping function, such as $f(q) = \ln(q)$.

4.2 Network matrix N

The matrix N quantifies file sharing. Given q_i is the total number of files generated by employee i and we let q_{ij} represent the number of files shared from node i to node j. We define the matrix N as follows:

$$\begin{aligned} N_{ii} &= 1 \\ N_{ij} &= \frac{f(q_{ij})}{f(q_i)} \in [0, 1] \end{aligned}$$

Therefore, N_{ij} is the normalized fraction of files generated by i that are shared with j. The values in matrix N are all positive and are less than or equal to one.

5. PRODUCTIVITY DECOMPOSITION

5.1 By Employee

The matrix N quantifies the standardized file-sharing metric $f(q)$, and defines the employee work flow network. We can break down the productivity measure for a company by employee. This decomposition of the scalar function P is possible because the function is linear homogenous in vector $Q^{*\top}$. Euler's theorem[3] applies and we have that

$$P = \frac{\partial P}{\partial Q_1^{*\top}} Q_1^* + \frac{\partial P}{\partial Q_2^{*\top}} Q_2^* + \ldots + \frac{\partial P}{\partial Q_n^{*\top}} Q_n^* \tag{7}$$

Each derivative $\frac{\partial P}{\partial Q_i^{*\top}}$ multiplied by Q_i^* is the productivity contribution P_i of node i. We can calculate all contributions P_i in closed form using the following vector derivative calculation:

$$P_i = \frac{\partial P}{\partial Q_i^{*\top}} = \frac{1}{2P}(N \cdot Q^{*\top} + N^\top \cdot Q^{*\top}) \cdot Q_j^* \tag{8}$$

which gives an $(n \times 1)$ vector of derivatives P_i. Once we know the amount of productivity that is contributed by each node, we can pinpoint the most productive users in the network.

5.2 Individual vs Group Productivity

We can also break down the productivity measure for a company by individual versus collaborative contribution. The diagonal of the Network Adjacency matrix represents the individual node's productivity when there is no collaboration between the nodes. Any productivity that occurs

[3] http://mathworld.wolfram.com/EulersHomogeneousFunctionTheorem.html

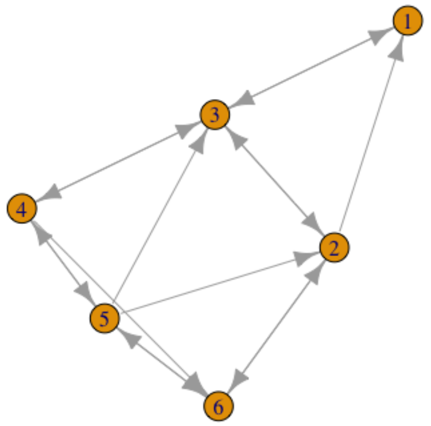


Figure 3: Network described by matrix N.

from collaboration will get captured in the non-diagonal elements of the matrix. Thus, productivity can be divided into two components as:

$$P = P_C + P_I \tag{9}$$

where P is total productivity, P_C is productivity due to collaboration, and P_I is productivity due to individual contribution. P_I is computed using

$$P_I = \frac{1}{n} \cdot \sqrt{\sum_{i=1}^{n} Q_i^2} \tag{10}$$

This is the result of using equation (2) where the network matrix contains no collaboration, i.e., is the identity matrix. Then, we can calculate productivity due to collaboration as a residual from equation (10), i.e., $P_C = P - P_I$.

5.3 Numerical Examples

A few examples will make the model clear. We start with the following base case. We set the number of nodes to be 6, the links between nodes are shown in Figure 3. The vector q is shown below and has nodes that are ordered in increasing order of number of files generated. (Even though we used an ordered set of integers from 1 through 6, this is not an index vector, but just the count of the number of files produced.) The network matrix N is also shown. Note that the diagonal of N is equal to 1, as we do need to capture the individual productivity of each node.

$$q = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

$$N = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0.50 & 1 & 1 & 0 & 0 & 0.50 \\ 0.33 & 0.67 & 1 & 0.67 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0.20 & 0.40 & 0.60 & 1 & 0.80 \\ 0 & 1 & 0 & 0 & 0.67 & 1 \end{pmatrix}$$

We will now bound Q in $[0, 1]$, set constant k equal to the 99-th percentile of q, i.e., $k = 5.95$. Using the quantum of information mapping function $f(q)$, our bounded Q is

$$Q = f(q) = \begin{pmatrix} 0.06 \\ 0.12 \\ 0.18 \\ 0.24 \\ 0.31 \\ 0.37 \end{pmatrix}$$

We compute the productivity metric as follows.

$$P = \frac{1}{n} \cdot \sqrt{Q^\top \cdot N \cdot Q} = 0.1615 \tag{11}$$

Now, if there is no collaboration, then all we get is individual productivity, as follows.

$$\begin{aligned} P_I &= 1/n \cdot \sqrt{\sum_{i=1}^{n} Q^2} = 0.0973 && (12) \\ P_C &= 0.0642 && (13) \end{aligned}$$

We may therefore define the percentage of “network” effect on productivity as

$$\frac{P_C}{P} = 0.3975 = 39.75\% \tag{14}$$

Finally, we may also calculate the productivity decomposition by node, calculated using equation (8), shown here:

Node	Decomposition	%age Decomposition
1	0.0022	1.34
2	0.0122	7.58
3	0.0180	11.15
4	0.0340	21.03
5	0.0438	27.10
6	0.0514	31.80
TOTAL	0.1615	100.00

5.4 Additional Properties

Here, we highlight some additional properties and intuition of the metrics we described in the previous subsection.

1. If we scale Q by $\alpha > 1$, then the value of P will also scale accordingly. That is if we increase file production by 20%, productivity will become αP. To see

this, suppose we assume a scaling factor $\alpha = 1.20$, and multiply Q by α. Then the new productivity value becomes:

$$P = \frac{1}{n} \cdot \sqrt{(\alpha Q)^\top \cdot N \cdot (\alpha Q)} = 0.1938 \qquad (15)$$

which is exactly $0.1615 \times \alpha$.

2. We also check that productivity per node is insensitive to the changes in the number of nodes, provided the structure of the system remains the same. We measure this by simulating many networks and see how the P measure varies. To begin, we randomly generate a network of 6 nodes a 100 times and compute P, and then look at the mean P and the summary statistics for the 100 trials. For this experiment, we use a different set of values so as to run a controlled experiment. We set all values in the Q vector to be equal to 0.35. We also assume that the probability of a directed link is 0.5. We then randomly generate the N matrix 100 times and compute P each time.

   We then executed the same experiment with network size increased from $n = 6$ to $n = \{10, 20, 100, 500, 1000\}$ nodes. We see that the productivity metric remains in the same ballpark as before, though it tends to decline mildly, and asymptote eventually. And as n grows the standard deviation reduces sharply, increasing the accuracy of this property as networks become larger.

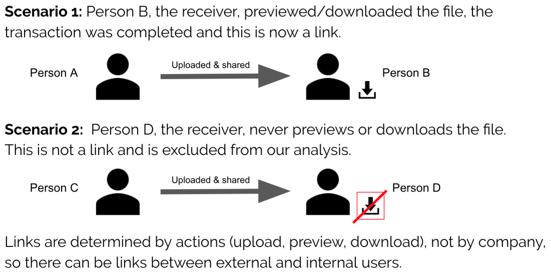
	Number of nodes n				
	10	20	100	500	1000
Mean P	0.2601	0.2519	0.2488	0.2477	0.2476
Median P	0.2596	0.2542	0.2489	0.2478	0.2476
Std dev P	0.0101	0.0058	0.0011	0.0003	0.0001

6. LARGE DATA ANALYSIS

6.1 Data Structure

We begin by describing the features of productivity across all companies in the sample. The sample we consider covers about 525 companies over the period early March to mid July 2018 (16 weeks), where daily file-sharing interactions between users are recorded, amounting to a total of 30 million records of data. If two users interacted on a file during the day it is counted as an "action" regardless of how many times the two users engaged that day on that file. An action requires that a user i sends a file to user j who then opens, previews, or downloads this file, else it does not count as an action. Interactions are captured in the network matrix N and the individual productivity vector Q. This is shown in Figure 4.

File sharing between any two users on a given day is denoted as a "transaction" and may involve any number of files and actions, though the most common number of files shared on a transaction are 1 or 2 files. Since the number of actions per file may be very large when a single file is sent to all users (such as with blast emails sent by human resources), the maximum number of actions is set to the 95th percentile value of the actions in order to trim such egregious outliers. (For example, in the case of one large company, the number of users who uploaded files in the sample period was 2,856, and the 99th percentile of the number of actions is 26.)



Scenario 1: Person B, the receiver, previewed/downloaded the file, the transaction was completed and this is now a link.

Scenario 2: Person D, the receiver, never previews or downloads the file. This is not a link and is excluded from our analysis.

Links are determined by actions (upload, preview, download), not by company, so there can be links between external and internal users.

Figure 4: How are links determined?

Since a user may send a file to more than one receiver, the number of actions may be greater than the number of files. The data file also contains details of the Sender and Receiver IDs, their company IDs, file type, file count and the number of actions. File sharing for a given company involves users who are part of the same company (i.e., internal) and those who are not, i.e., external users.

6.2 Network Metrics

We use the data to construct the network adjacency matrix N and productivity vector Q as discussed in Sections 4 and 5. We may choose any granularity for network construction, such as daily, weekly, monthly. We chose to present all metrics using non-overlapping weekly blocks of data, where a week is defined as Monday through Sunday.

For each week we construct the following measures for each user in the network (vectors of size n, the number of users): (i) A vector of eigenvalue centrality [3] scores, which quantifies the importance of user position in the network. (ii) A vector of betweenness centralities [6], denoting how many shortest paths in the network go through a user node. This signifies the importance of a user as being a broker or middleman in the file sharing network. (iii) A vector of node degrees, i.e., how many other users a node is connected to. (iv) A vector of individual contributions to productivity based on equation (8). (v) A flag for internal versus external user.

We also calculate aggregate weekly measures for the entire company, i.e., the following scalar values: (i) The productivity score P in each week, and the breakdown of this score into individual contribution P_I and collaborative contribution P_C. (ii) The amount of productivity contributed by internal versus external users. (iii) The number of links in the network, and the average degree. (iv) Density of the network, i.e., the number of links in the network divided by the total possible links, $n(n-1)/2$. (v) The average size of communities in the network. Communities are detected using a standard community detection algorithm, the greedy algorithm of [2]. (vi) The percentage productivity contributed by the top 5% of users. (vii) Fragility (or virality), which is a measure of how fast information can spread on the network. This is a function of the concentration in links in a few nodes, and is measured as $E(d^2)/E(d)$, where d is the degree of each node. The numerator of this measure is analogous to the standard measure of concentration used by economists, the Herfindahl index.[4] (viii) We also gather data on weekly

[4] https://en.wikipedia.org/wiki/Herfindahl_index

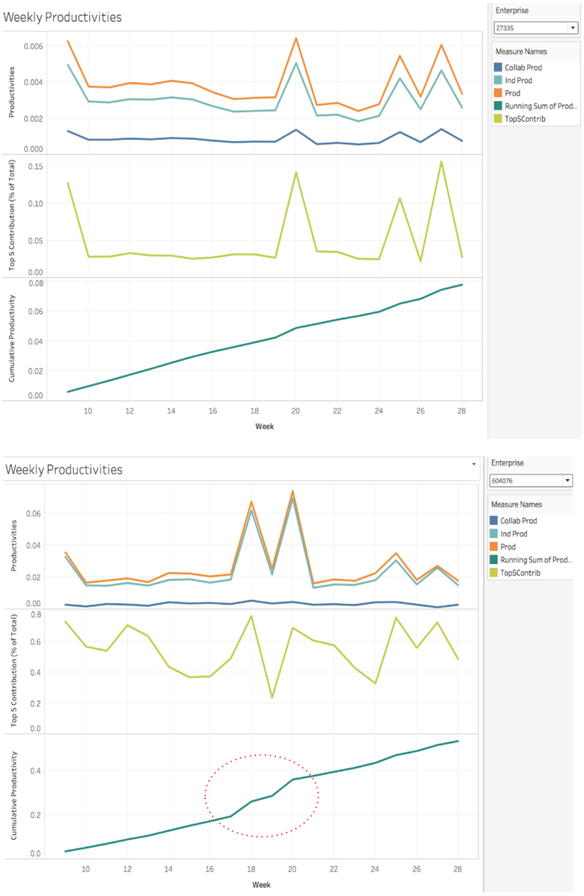


Figure 5: Productivity plots for two sample companies. In each plot the top plot shows daily productivity, total, individual, and collaborative. The middle plot shows the percentage of productivity contributed by the top 5 contributors. The bottom plot shows cumulative productivity. The top company has an almost linear growth in cumulative productivity, whereas the lower one has differing slope as it evidences growth spurts.

active users (WAU) in terms of the percentage of total users. "Active" users are the users who logged into the file-sharing platform at any point in time during the week. For graphing purposes, our data is smoothed to report the rolling weekly average each day.

6.3 Empirical Examples

Figure 5 shows the productivity per employee for two sample companies over time. company productivity, average collaborative productivity, and individual productivity all follow each other closely. Collaborative productivity is much lower than individual productivity across all weeks. Cumulative productivity is almost linear for the first company, i.e., productivity remains steady week over week, though in the bottom plot we see a slight kink where productivity ramps up. In the first company, productivity tracks those of the top 5 contributors, meaning that they are key players. But, this relationship is less marked in the case of the second company. Overall, we see that company productivity does correlate with that of the top few users.

We may also examine metrics within a week, for example, in Figure 6 we plot the user's productivity contribution versus degree of the user, i.e., how many connections they have. Users on the lower right are those with many connections but not too much individual file production, whereas those at the top left have few connections, but generate many files. The user on the extreme right has many connections and also a reasonable level of contribution and may be a good example of a connector, yet, the user at the topmost point in the plot is contributing a lot to the total. The former is more of a connector, but the latter is more of a producer.

Similar analyses of users are shown in Figures 7 and 8. In Figure 7 we note that top productive users may not necessarily be the most central in the network. In Figure 8 we see that the most connected users (in terms of degree) may not be the highest contributors to productivity.

6.4 Analyzing key users

We define key contributors to be the top 5% of total productivity based on the sorted productivity decomposition (D) vector generated using equation (7). We call the vector D_{ct} for company c in week t.

Do the top producers change a lot or remain stable? We

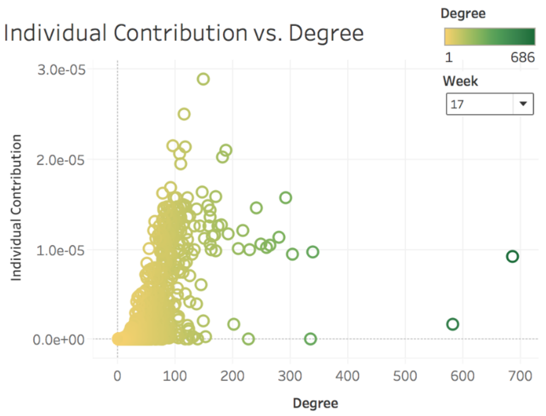


Figure 6: Productivity contribution by user as a function of degree of a node, for a single week.

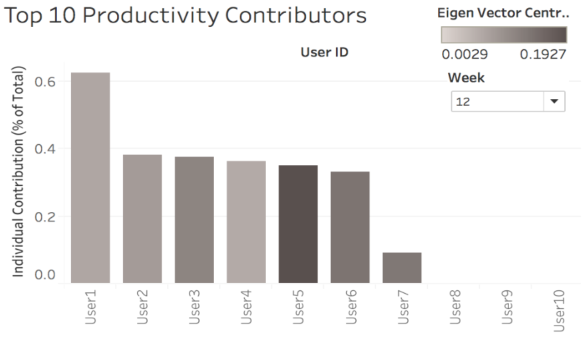


Figure 7: Top 10 productivity contributors. We see that higher individual contribution to total productivity does not necessarily mean a higher EV centrality.

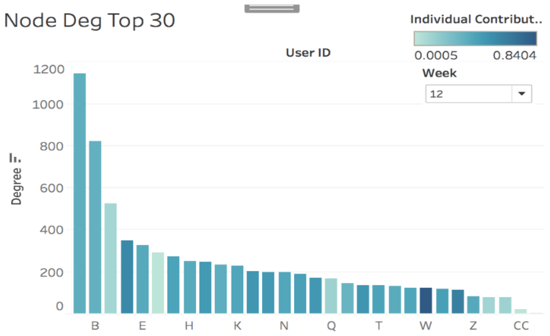


Figure 8: Top 30 users by degree. Higher degrees (ie, having more connections) doesn't mean higher individual contribution. User IDs have been anonymized.

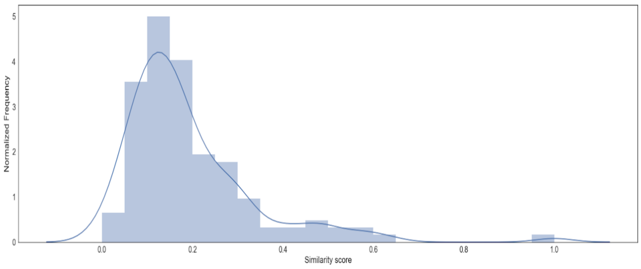


Figure 9: Distribution of Jaccard similarity across consecutive weeks averaged for each company. The histogram shows the distribution for 122 sample companies. Frequency is normalized such that area under the curve is 1.

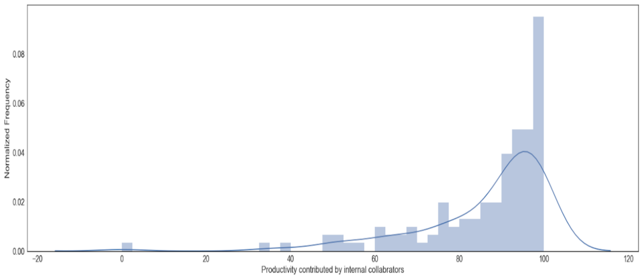


Figure 10: Percentage of productivity attributed to internal users who worked only with internal users. The histogram shows the distribution of this for 124 sample companies. Frequency is normalized such that area under the curve is 1.

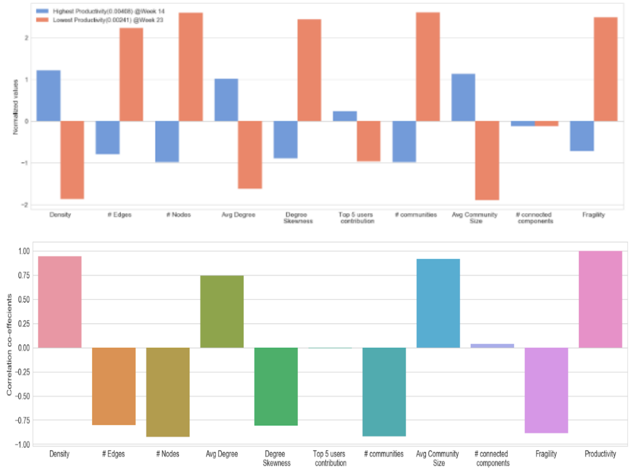


Figure 11: Comparison of per user productivity across low and high productivity weeks (top panel). These measures are described in Section 6.2. The metrics are normalized to make comparisons easier. Correlations are shown in the lower panel, and are consistent with the comparison across high and low productivity weeks.

compute Jaccard similarity between sets of key users in two consecutive weeks to answer this question. For every pair of consecutive weeks in a company, we calculate the similarity between the top contributors as follows.

$$\text{Jaccard similarity}_{ct} = \frac{|D_{c,t-1} \cap D_{c,t}|}{|D_{c,t-1} \cup D_{c,t}|} \in (0, 1) \qquad (16)$$

We calculate for each company the average Jaccard similarity across all weeks to determine how much consistency there is in top contributors. We show the histogram of sample companies' productivity similarity to examine the distribution of stability in top contributors. See Figure 9. Mean and modal similarity is around 0.2, suggesting that for many companies, top producers change from week to week. There is a long right tail, which indicates that for a small fraction of companies, there is stability across time in the top productivity employees.

6.5 External versus internal users

Each company on the file-sharing platform has users who are employees of the company (internal) and also users who are not employees or contractors (external). We examine productivity decomposition by user type: Are users who collaborate with external users more productive? We consider three types of users: (i) internal users who only connect to internal users; (ii) internal users who connect to external users; (iii) external users. For each company, each week, we compute and store the percentage of productivity contributed by each of these three groups. We compute the average share of productivity in each type as well for each company, and display the histogram. We see that most of the productivity comes from internal users who connect with other internal users, see Figure 10. (The share of productivity from external users turns out to be minimal, which is understandable because external users comprise a small fraction of the total user base in a company.)

6.6 Comparing low and high productivity weeks

We also try to get an understanding of the drivers of productivity from a comparison of metrics between the minimum and maximum productivity weeks. This is shown in Figure 11.

We see that network density, average degree, concentration of productivity in the top 5% users, and average community size are correlated with higher productivity per user. Therefore, denser networks with key users drives productivity. On the other hand, too many nodes, skewed degree of nodes, a large number of communities, and higher fragility

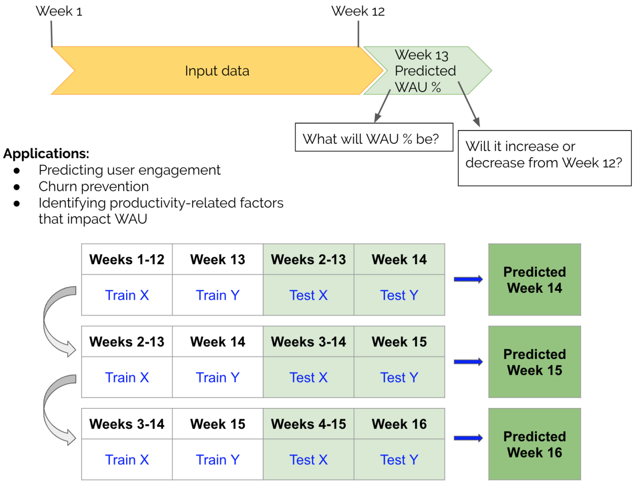


Figure 12: Schematic showing how the model is trained (top) and tested (bottom), with some applications for models trained on this schema. The graphic depicts the three rolling experiments to predict the WAU across all companies in weeks 14, 15, and 16.

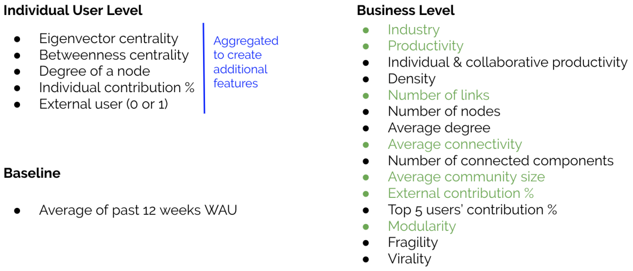


Figure 13: Feature set used for predicting clients' user activity. These features are at the client level. Features that are user level are aggregated by averaging up to the client level across all users.

Metric	Week 14	Week 15	Week 16
MSE (out-of-sample)	0.0026	0.0038	0.0032
MSE (naive)	0.0033	0.0044	0.0045
% Error reduction	21.4%	15.3%	28.6%
Coefficient	0.9306	0.9031	0.9296
Adjusted R^2	0.92	0.89	0.91

Table 1: Metrics of the classification model. In the top half of the table we present the mean squared prediction error (MSE) for both the naive model (only using past 12 week average WAU) versus the full model that also uses network variables. In the bottom half we present the coefficient of a regression of the actual WAU on predicted WAU from the full model and we see that these coefficients are statistically very close to 1 suggesting a good predictive model. The R^2 is also reported to see how much variation is captured, and these are above 90% meaning that most of the variation in the actual WAU is captured by the predicted WAU. All coefficients are highly significant at the 99.99% level (t-stats not reported).

are associated with lower productivity per user. This is because too many users are segmented into communities, leading to lower transmission of productivity across the company. In our cross-sectional analysis, these insights will be useful in determining which network metrics explain (i) productivity across companies, and (ii) predict weekly active users. These statistics are corroborated by the correlation of productivity with the network measures, also shown in Figure 11.

6.7 Predicting client activity

In this section we assess whether our Productivity measure P has predictive power to determine WAU (the ratio of weekly active users to all users). We fit a model to predict WAU in week t using a feature set constructed from data in weeks $t-12$ through $t-1$, i.e., we use the past 12 weeks of data for prediction. We fit one model to the cross-section of client companies and use rolling experiments. Our dataset is short and supports three rolling of out of sample prediction periods.

For example, we will use a feature set constructed from weeks 1–12 to fit WAU in week 13. This trains the prediction model in-sample. We then take this fitted model and use data from weeks 2-13 to predict week 14 out-of-sample as a test of the model. Since we have 16 weeks of data, we are able to run three experiments, predicting out-of-sample WAU for weeks 14, 15, and 16. The schematic is shown in Figure 12.

We have two analytics objectives for the model. One, predict WAU for a client in the following week (this is a "regression" exercise). Two, predict the direction of WAU next week, i.e., the sign of the change in WAU (a "classification" experiment). The naive model for prediction would be to assume that the prediction of WAU for any week is based on the average WAU for the past 12 weeks, which is one variable in the feature set. To this naive variable, we also add the various network measures that were described in Section 6.2. If any variables are highly correlated with each other (collinear) we eliminate one of these variables. This leaves us with an abridged but clean dataset which we then use for prediction. The features used fall into both categories described in Section 6.2, and are shown in Figure 13.

First, we detail the results of the WAU prediction model. Different machine learning models were tried for prediction, but the most successful ones were random forest, gradient boosting, xgboost, and a MLP regression model. Of these, we report the results from xgboost, which gave the best results. The results are shown in Table 1. We see that the prediction model does very well in matching actual WAU in level terms. This can be seen from the fact that the slope coefficient in the regressions lies between 0.9 and 1.0, and the slope is statistically close to 1, though it is always less than 1, suggesting that the model marginally overestimates the next week's WAU. It also improves on the prediction error over the naive model by approximately 15-30%.

Second, we assess how well the model is able to predict the direction of change in WAU for each client. This is of interest because attention may be directed to clients whose WAU is predicted to drop. The confusion matrices for the prediction of the three experiments are shown in Table 2. The diagonals are heavy indicating that the models perform

Week 14	Actual		Metric	Value
Predicted	0	1	Accuracy	0.72
0	132	58	Precision	0.74
1	89	248	Recall	0.81
			F1	0.77
Week 15	Actual		Metric	Value
Predicted	0	1	Accuracy	0.68
0	144	82	Precision	0.71
1	89	214	Recall	0.72
			F1	0.71
Week 16	Actual		Metric	Value
Predicted	0	1	Accuracy	0.69
0	176	96	Precision	0.73
1	68	188	Recall	0.66
			F1	0.70

Table 2: Confusion matrices for the three experiments. In these cases 0 stands for the case where the WAU declined and 1 for when it increased versus the average WAU of the past 12 weeks. We also report accuracy, precision, recall, and F1 score.

well. Accuracy is about 70%, as are precision, recall, and the F1 score. Model performance is stable across time. In sum, a model supported by graph-theoretic features is able to measure productivity and use it to forecast the usage of a file-sharing platform.

7. CONCLUDING DISCUSSION

This paper presents a new productivity metric P based on a novel network model of file-sharing amongst users within a client company. The metric P may be decomposed into productivity coming from individual effort and from collaborative effort. A decomposition of total productivity is also possible by user so as to identify the most productive employees. This network approach generates several metrics at both, user and company level, enabling the creation of a rich feature set for predicting platform metrics.

Using a sample of $\sim$ 525 client companies, over 16 weeks, comprising about 30 million file-sharing records, we find that predicting client usage of the file-sharing platform is improved over a model where the past 12-week average is used as the prediction. Accuracy levels are high when predicting the level of the percentage of active users on the platform, and also when predicting the sign of change in percentage of active users.

The feature set supports many other analyses as well. With longer time-series of data, predicting churn, i.e., client dropout, becomes feasible. Clustering and classification of companies and users by productivity is supported. User engagement can be predicted. And of course, analyses may be provided to clients to enable them to make their companies more productive, while also offering a weekly measure of productivity to track improvements in collaboration.

8. REFERENCES

[1] D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi. The Network Origins of Aggregate Fluctuations. *Econometrica*, 80(5):1977–2016, Sept. 2012.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct. 2008.

[3] P. Bonacich. Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182, Mar. 1987.

[4] T. J. Chemmanur, E. Loutskina, and X. Tian. Corporate Venture Capital, Value Creation, and Innovation. *Review of Financial Studies*, 27(8):2434–2473, Aug. 2014.

[5] R. L. Cross, R. D. Martin, and L. M. Weiss. Mapping the value of employee collaboration | McKinsey. *McKinsey Quarterly*, (August), 2006.

[6] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[7] B. B. Frey, J. H. Lohmeier, S. W. Lee, and N. Tollefson. Measuring Collaboration Among Grant Partners. *American Journal of Evaluation*, 27(3):383–392, Sept. 2006.

[8] R. Gajda. Utilizing Collaboration Theory to Evaluate Strategic Alliances. *American Journal of Evaluation*, 25(1):65–77, Mar. 2004.

[9] E. Hess and J. Liedtka. *The Physics of Business Growth: Mindsets, System, and Processes.* Stanford University Press, Stanford, 2012.

[10] W. W. Leontief. *Input-Output Economics.* Oxford University Press, Oxford, New York, second edition edition, Mar. 1986.

[11] J. Long, B. and C. Plosser. Real Business Cycles. *Journal of Political Economy*, 91(1):39–69, 1983.

[12] K. Olmstead, C. Lampe, and N. B. Ellison. Social Media and the Workplace, June 2016.