Can Econometrics Tell Us Which High Dimensional Text Clusters have Short Signatures?

Abstract

The past decade has witnessed an unprecedented generation of unstructured information. Given this mountain of unstructured information, a paradigm that is of increasing importance to industrial applications is to extract a "small amount" of this information for use by the application.

As a separate, recent development, quantities that were earlier modeled using the classical Pareto "80-20" rule were found to display a distinct phenomenonthe so-called "long-tail." This phenomenon, seen, for example, in the effect of the internet on sales curves, has influenced recent thinking in econometrics.

In this interdisciplinary paper, we demonstrate links between the above two seemingly disparate issues. We show that the problem of extracting a small amount of information in data mining contexts is amenable to analysis using recent ideas from econometrics. In particular, we view the centroids produced during document clustering, which capture important statistical information about the corpus, through the econometric lens of "Pareto vs. long-tail." We ask when such clusters have short signatures.

Our approach is as follows. We propose generic parametric definitions for "Pareto" and "long-tail" that are tailored towards document clustering. We choose three standard document collections of very different provenances. We then investigate whether the centroids produced by document clustering follow a Pareto or long-tail distribution, for suitable choices of parameters in these definitions. Our choices of parameter values are driven by applications. We argue that clusters have short signatures for the purposes of these applications only when their centroids are Pareto. However, the behavior of cluster centroids in this regard shows a pattern: we show that the centroids tend to *switch from long-tail to Pareto as the number k of clusters is increased*. Our work shows that the choice of the "right amount" of information to extract is considerably more intricate than is presently believed. Our framework allows users to decide when high dimensional clusters have short signatures, and when not, using an econometric lens on data mining constructs.

1 Introduction

The past decade has witnessed an unprecedented generation of *unstructured* information in textual format. Such unstructured information is being generated in a wide variety of domains such as the world-wide-web, enterprise intranets, newsfeeds, email, digital libraries, medical records, and many others.

Among the many ideas that are discussed in order to tame this information generation is to extract just the "right amount" of information from this mountain of data. This idea must be judiciously applied, however. In this paper, we approach the question of the "right amount" of information, for some important applications, in a principled manner, drawing upon insights from econometrics. In particular, we explain when a small amount of information suffices for important applications.

One of the primary techniques for the analysis and management of unstructured information is document clustering. Document clustering can be used at various stages in unstructured information management: to understand high-level organization of data [6, 14, 19, 23], to organize search results [11], to extract semantic information such as labels [6, 13], and so on. These technologies have high commercial importance: in enterprise unstructured information management, clustering technologies are a part of enterprise content management (ECM), which was valued at $5.1B in 2012, and is growing at close to 10% annually as per Gartner [2].

There is a large body of research on better and faster algorithms to produce a clustering of a document collection. However, several *industrial* applications of clustering require, in addition to a good and fast algorithm, a deeper understanding of the *structural properties* of the clusters that are produced by these algorithms. In particular, we will show that such an understanding is needed in order to extract just the "right amount" of information in the context of document clustering. We provide below two examples of frequently occurring applications of clustering in unstructured information management where these considerations are natural and important. Both of these examples are generic, and applicable to the gamut of uses of document clustering mentioned in the previous paragraph.

Example 1 (Labeling)**.** *In several important applications of clustering, such as cluster based retrieval*[1]*, a cluster is "labeled" by a digest of terms. These terms are supposed to convey the content of the cluster to the user. A standard way to extract these terms from the documents in the cluster is by frequency: namely, the cluster is labeled by the most frequent terms that occur in it. A natural question that arises now is: how many terms should we use in the label?*

[1] Especially in the interactive setting or as part of scatter-gather workflows [6] that occur during commercially important applications such as eDiscovery (estimated market size $2.1B by 2017 as per Gartner [1]).

Example 2 (Dimensionality Reduction)**.** *A natural means to perform feature selection for clustering is to begin with a clustering, extract the most important features from each cluster, and use this collection of features as part of the reduced feature set (this is the so-called "wrapper" technique [12]). Again, the question arises: how many features should we extract per cluster?*

In both of the examples above, we need to extract just the "right amount" of information from a cluster. However, in order to characterize the amount of information sufficient for these applications, we must understand the distribution of this information. Therefore, our broad problem statement is to understand the distribution of information in the context of the structure of clusters produced by standard clustering algorithms in large corpora of text documents.

Arguably the most important intra-cluster structure produced by a clustering is the *cluster centroid*. The centroid itself supports several applications of clustering. For example, both motivating examples presented earlier are intimately related to the cluster centroid: labels and important features can both be gleaned from cluster centroids. The most frequent terms in the cluster are precisely the coordinates of the centroid having the highest magnitude, and these can be used as labels [6].

So, what would be the "right amount" of information to be extracted from a cluster centroid? In particular, we are interested in cases where a small amount of information would suffice. In light of our motivating examples, we may phrase this question as follows:

> When does a high dimensional text cluster have a "short signature?"

In order to answer this question, we must understand the distribution of information in the centroid. Given the importance of the cluster centroid, both from theoretical and application standpoints, we believe this is a worthwhile study in and of itself. In this paper, we propose a principled approach to this study based upon ideas from econometrics.

Consider the weights of terms that occur in the centroid. We will informally refer to this information as the "distribution of weights:" a formal definition is provided in §5.2. The distribution of weights in a cluster centroid follows a classic shape of a quantity that falls monotonically and rapidly along the X-axis. In other words, the first few terms with the highest weights contribute disproportionately to the total weight of the centroid. This property of the distribution of weights is frequently used for a variety of tasks in text mining and retrieval, including the two tasks of Ex. 1 and 2.

In econometrics, one makes a distinction between two types of such distributions based on their cumulative distribution: they can be either Pareto-style or long-tailed. In Pareto-style distributions, the tail does not account for much of the cumulative distribution, hence the colloquial "80-20 rule" which says that only 20% of the observations account for 80% of the cumulative weight. A quantity that is often modeled using a Pareto-style distribution is the book sales

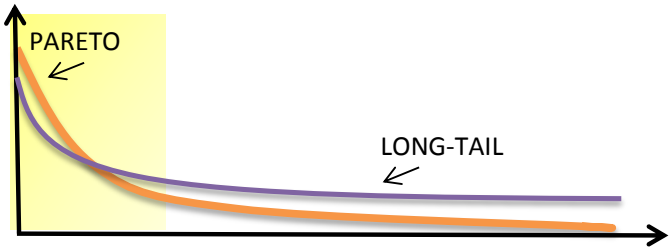


Figure 1: A Pareto-style distribution vs. a long-tailed one. Both power laws superficially look alike: resembling an inverted J-shape of exponential decay. However, they differ in the proportion of the cumulative that lies at the "heavy front" of the curve (shown in shaded). For Pareto-style distributions, the area in the head is the majority of the total area, while for long-tailed distributions, the area under the curve to the right of the head is significant, and cannot be ignored.

from a brick-and-mortar store. Such a store makes most of its sales from a relatively small number of top-selling books. On the other hand, with the growth of the internet, certain distributions that were earlier modeled as Pareto-style are now recognized to be "long-tailed," meaning that the tail now accounts for a significant portion of the cumulative (see Fig. **??**). In econometrics, this effect is used to model diverse phenomena such as internet sales of vendors like Amazon. In other words, it is recognized that vendors such as Amazon are able to monetize the long-tail by selling a large number of esoteric books, each selling very few copies, to a large number of niche buyers.

How do the ideas from the preceding discussion apply to our problem? In light of the above discussion, let us ask whether the weight distribution of the centroid is a Pareto-style distribution or a long-tailed distribution? Why would the answer to such a question be important? Clearly, the answer would affect the selections in both our motivating examples. For suppose that the distribution is Pareto-style, then we can choose a small number of labels/features and put a safe upper bound on the amount of information lost. In this case, the cluster has a short signature. Whereas, if it is a long-tailed distribution, then a choice of just a few highly weighted terms would result in a loss of significant information. In such a case, the cluster does not admit a short signature.

At this time, we introduce two interesting aspects of this problem statement.

1. The high-dimensionality of text corpora. Dimensionalities of tens of thousands are the norm. Therefore, a literal "80-20" Pareto rule would still give us a few thousand features to consider. Clearly, a signature comprising these features would not be "short" for the applications (such as labelling) in our motivating examples. This points to the need to (a) provide a generic parametrized definition that reworks the Pareto rule for the high dimensional setting of text, and (b) let applications drive the actual values of the parameters in this definition.

2. The dependence of these phenomena on the number of clusters k: The number of clusters is a natural parameter to all clustering-related phenomena. Is it possible that for certain values of k, the weight distribution is Pareto-style, but for others, it becomes long-tailed?

Our contribution is to systematically study the distribution of weights, and study its structural properties in the context of ideas of econometrics as applied to information management problems. We show that the observations already made about centroid weights ([13] and [7]) are important, but do not suffice to answer the meta-question of "how much information must be extracted from centroids?" for various applications. Our broad study aims to piece together the entire picture and answer this question in a principled manner.

2 Research Questions

Motivated by the discussion of the preceding section, we formulate the following specific research questions.

1. What are appropriate versions of Pareto-style and long-tail properties to be applied to text clusters?
2. What is the average distribution of coordinate weights of the centroid?
   (a) Where is most of its weight concentrated?
   (b) Is the distribution Pareto-style or long-tailed?
3. What is the deviation between distributions of weights between different clusters in a clustering? In particular, is the deviation low enough that we may meaningfully speak about a representative average?
4. What is the shape of the cumulative distribution? What value does it take at the beginning of the tail?
5. What is the effect of the number of clusters k on these behaviors? Is there a shift from Pareto-style to long-tail (or vice versa) as we vary k?

We provide the scope and limitations of our work in the supplementary notes.

3 Related work

3.1 Power Law Distributions and their Prevalence When the probability of occurrence of a particular value of a variable varies inversely as a power of that value, the resulting distribution is said to follow a power law. Power law distributions occur in a diverse range of domains, from economics and finance, planetary and earth sciences, physics and computer science. Canonical power laws include the discrete Zipf's law of word occurrences [25], and the continuous Pareto law, from which the Pareto "80-20" principle gets its name. There is a vast amount of literature on power laws. We refer the reader to [17] for an overview of the empirical evidence for the existence of such laws, as well as theories that explain them. [20] provide a simple plausible explanation for several domain examples of power laws (including Zipf and Pareto) as follows. They claim that when the time of observation is itself regarded as an exponentially distributed random variable, the resulting observed distributions tend to follow a power law. [21] suggest that when processes whose growth is exponential are randomly observed, their state follows a power law. [16] offers a comparative review of power law distributions with a close relative—log-normal distributions—across many domains of modeling. Their stress is on understanding the underlying generative models for both distributions.

3.2 Observations on Centroid Weights The fact that the first few terms of the cluster centroid have significantly higher weight than the rest has been observed by both [13] and [7]. However, their valuable observation is not sufficient to distinguish between the Pareto and long-tail cases, since their observation holds true in both cases. We argue, based on ideas from econometrics, that one must carefully differentiate between these two broad cases in order to make correct inferences on the amount of information to be extracted from the centroids.

3.3 Long-Tailed Distributions, Internet Sales The term "long-tail" was first used in [3] (later expanded into a book). [4] investigate the switch from Pareto to long-tail caused by internet sales. They provide empirical evidence that the internet sales channel exhibits a significantly less concentrated sales distribution when compared with traditional channels such as brick-and-mortar stores. [9] study the problem of predicting the power law exponent for long-tailed distributions such as internet book sales.

3.4 Domain Applications of Long-Tailed Distributions The long-tail phenomenon has been applied to various domains. [8] use long-tails to analyze network performance. [10] model internet traffic queuing delays using long-tails. [18] use long-tails for the analysis of recommender systems. [22] model the long-tail of short articles on the wikipedia for the task of information extraction.

4 Preliminaries

4.1 Data Mining Preliminaries Since this work is interdisciplinary, we describe the data mining setting of our study in the supplementary notes for the benefit of readers from econometrics.

4.2 Econometrics Preliminaries For the benefit of our data mining readers, we provide a brief introduction to the standard Pareto distribution and the Pareto principle.

4.2.1 Standard Pareto distributions The standard two-parameter continuous Pareto distribution of a random variable X is a canonical power-law distribution with cdf

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^\alpha & x \geq x_m, \\ 0 & x < x_m. \end{cases}$$

The parameter α is the power in the power-law governing the distribution. The parameter x_m denotes the minimum value above which the distribution is considered a power law: this is necessary because any pdf of the form $Cx^{-\alpha}$ will diverge as $x \to 0$. On a standard scale, the distribution takes the shape of an inverted J-curve, On a log-log scale, it takes the form of a straight line. It has a finite mean of $\alpha x_m/(\alpha-1)$ when $\alpha > 1$, and a finite variance of $\alpha {x_m}^2/[(\alpha-1)^2(\alpha-2)]$ when $\alpha > 2$.

4.2.2 Pareto principle The Pareto distribution is a basis for the so-called "Pareto principle:" an observation on the distribution of various quantities of social and economic interest (for example, incomes). It says that the top 20% of the observation (for example, incomes) account for 80% of the cumulative distribution. Applied to the distribution of incomes, the Pareto principle would say "the top 20% richest people have 80% of the total wealth." Similar statements hold about other quantities of interest, such as populations of cities in a given area, etc.

The Pareto principle holds precisely for a Pareto-distributed variable when $\alpha = \log_4(5)$.

The Pareto principle is often used to distinguish between distributions that are "heavy-front" from those that are "long-tail." We have included a preliminary discussion of these phenomena in §6.3.

5 Data Mining Framework

In this section, our goal is to construct a method to study the k weight distributions of the k centroids that result from a k-way clustering $\{C_1, C_2, \ldots, C_k\}$ of the corpus $\mathcal{D}$. In order to do this, we define certain functions that combine these k weight distributions into. Analyzing these "combined functions" will give us the structural properties that we seek.

5.1 Norm and Ordering on Terms in Clusters First, we need a norm for terms in clusters that takes into account the clustering structure. Let t be a term that occurs in C. Normalize the centroid of C so that the sum of its coordinates is one. Then we denote the weight of the term t in the centroid by $\pi(t, C)$.

Next, we impose an ordering on the terms that occur in a cluster C by means of $\pi(\cdot, C)$.

Definition 1. *We denote by $L_C(n)$ the n-th term in $\mathcal{V}$ ranked in descending order of $\pi(\cdot, C)$.*

It follows that $L_C(n)$ is the n-th most frequent term in C. For example, $L_C(1)$ and $L_C(2)$ are the two most frequent terms in C, and therefore the two highest weighted terms in the centroid of C.

5.2 Functions to Study Cluster Centroids Recall that our setting is a k-way clustering $\{C_1, \ldots, C_k\}$ of $\mathcal{D}$. Each centroid has a different distribution of norms. It will help us to incorporate all these distributions at once into a single function. Doing so allows us to study both intra and inter-cluster structural properties. We define this function next.

Definition 2. *The* combined ranked weight function *$f(n, j, k)$, where $1 \le n \le |\mathcal{V}|$ and $1 \le j \le k$, is defined as $\pi(L_{C_j}(n), C_j)$ for cluster C_j. Namely, it is the weight of the term $L_{C_j}(n)$.*

To be clear, $f(n, j, k)$ is the weight of the n-th term in cluster C_j, ranked by the norm $\pi(\cdot, C_j)$. Therefore, for a fixed k-way clustering, and fixed value of j, $f(1, j, k)$, $f(2, j, k)$, are the first, second, and so on most frequent terms in the cluster centroid of C_j.

Before proceeding, we note that the first index in $f(n, j, k)$ refers to the terms within a single cluster, the second index refers to a cluster in the clustering, and the third index captures the number of clusters in the clustering. Therefore $f(n, \cdot, k)$ encodes the distributions of all the k clusters generated by the k-way clustering of $\mathcal{D}$. Specifically, $f(n, 1, k), \ldots, f(n, k, k)$ are the k distribution functions that correspond to the k clusters.

Note that $f(n, j, k)$ is indeed a distribution function: it is non-negative, upper bounded by one, and its values sum up to one.

$$\sum_{n>0} f(n, j, k) = 1.$$

We may now define the *cumulative* distribution of $f(n, j, k)$. The cumulative has the same indexing scheme as $f(n, j, k)$.

Definition 3. *The* combined ranked cumulative function *$F(n, j, k)$ is defined as*

$$F(n, j, k) = \sum_{i=1}^{n} f(i, j, k).$$

Similar to the case with $f(n, \cdot, k)$, the cumulative function $F(n, \cdot, k)$ encodes the cumulative norm distribution functions for all k clusters.

In light of the above, we may define the *average* of the distribution functions taken over the k clusters in the clustering.

Definition 4. *The functions $\overline{f}(n, k)$ and $\overline{F}(n, k)$ are defined as the average of $f(n, j, k)$ and $F(n, j, k)$, respectively, over the k clusters in the k-way clustering. Namely,*

$$\overline{f}(n, k) = \frac{\sum_{j=1}^{k} f(n, j, k)}{k}, \quad \overline{F}(n, k) = \frac{\sum_{j=1}^{k} F(n, j, k)}{k}.$$

Definition 5. *The functions $\sigma_f(n, k)$ and $\sigma_F(n, k)$ are defined as the standard deviation of $f(n, k)$ and $F(n, k)$ around their means $\overline{f}(n, k)$ and $\overline{F}(n, k)$, respectively.*

6 Econometric Framework

We wish to transfer and tailor econometric ideas to the setting of users interacting with a clustering system. Our stress will not be on mathematical models, but rather on extracting the ideas underlying these models, and investigating whether they can be meaningfully tailored to provide insightful approaches to real-world problems in information management. We also stress that the values of parameters suggested below can be changed based upon the application and dataset; we have provided some nominal values along with a justification.

6.1 Form of Generic Pareto for Interaction with Cluster Centroids We first formulate an analog of the Pareto principle for cluster centroids. Let us write a generic candidate Pareto principle in terms of both the distribution $f(n, j, k)$ and the cumulative $F(n, j, k)$. Recall that our setting is a k-way clustering $(C_1, \ldots, C_k)$ of $\mathcal{D}$.

Definition 6. *The cluster C_j is said to be* Pareto-style with parameters X and Y *if the top X terms in the distribution $f(n, j, k)$ account for Y proportion of the weight of the centroid of C_j.*

Since the cumulative function is asymptotically unity, it allows for a concise formulation of the principle above.

Definition 7. *The cluster C_j is said to be* Pareto-style with parameters X and Y *if*

$$F(X, j, k) = Y.$$

Finally, we note that the standard Pareto principle corresponds to the values of $X = 20\%$ and $Y = 80\%$.

6.2 Nominal Values of Parameters for Text Clustering As noted, the standard Pareto principle in econometrics suggests values of $X = 20\%$ and $Y = 80\%$. We will, instead, propose the following nominal form and parameters for a Pareto-style principle applied to text clustering. Justification follows the proposed definitions.

Definition 8 (Pareto for Centroids)**.** *A cluster C_j is said to be* Pareto-style for text *if the top 20 terms in the distribution $f(n, j, k)$ account for at least 50% of the weight of the centroid of C_j. Namely, the median of $f(n, j, k)$ occurs at $n \leq 20$.*

Once again, the cumulative function allows for a concise formulation of the proposed principle.

Definition 9. *A cluster C_j is said to be* Pareto-style for text *if*

$$F(20, j, k) \geq 0.5.$$

Now, we justify why this is an appropriate form for a Pareto-style principle. Let us begin by applying the standard Pareto principle to the setting of text clustering. Text is characterized by extremely high dimensionality. For example, our datasets have dimensionality of the order of 10,000 terms. A 20% proportion *of the entire dimensionality* would give us of the order of 2,000 terms. This is a few orders of magnitude more than any user would be willing to inspect in an interaction with a clustering system: for example for scatter-gather or cluster-based retrieval.

On the other hand, extensive experience with cluster-based systems in enterprise information management has shown us that 20 terms per cluster is around the number that users expect to inspect when making decisions on the relevance and importance of a cluster. This is why we set the value of X to 20 terms, rather than 20% of the total terms.

Next, we explain why we set the value of Y to 50% of the centroid's weight, rather than 80% as would be the case in the standard Pareto principle. There are three reasons for this.

1. A user trying to make a decision on relevance or importance of a cluster would likely be able to make their decision once 50% of the total information were available to them. This can be further justified since terms are inter-related in their usage within documents. Therefore, having the top 50% of the weight does also shed some light on the remaining 50%.

2. Since the weight distribution falls rapidly after the first few terms, if the 50% cumulative weight is not reached therein, a user would have to inspect a large number of additional terms in order to add significant cumulative weight, and this also would push the user towards making a compromise on the amount of information they wish to have.

3. The 50% cumulative weight point—the median—is a natural object of study. In particular, in skewed distributions, the median is often used as a measure of centrality, in preference over the mean. This adds more importance to the median.

4. The value of 50% is sometimes used in econometrics to characterize long-tails (see §6.3).

In light of the above factors, we feel that 50% of the cumulative was a good choice for Y.

We stress that in our approach, the values of the parameters are application driven. Furthermore, their range may be empirically determined for a given dataset. The proposed values are suitable for our motivating applications, and for our datasets. For other applications and datasets, other values of parameters can be tried. Whatever be the values chosen, it is the *phenomena* of Pareto vs. long-tail that is important, and should be studied.

6.3 Long-Tails in Cluster Centroids The Pareto principle holds for distributions that have a "heavy front" where 20% of the top observations account for 80% of the cumulative. A long-tail is, in some sense, the opposite of this phenomenon. A distribution is said to have a long-tail when the tail has more cumulative than the front of the distribution. In particular, the bottom 80% of the observations account for over $B\%$ of the cumulative. The choice of B may be domain, problem, and context specific. A value of $B = 50$ is sometimes used in econometrics. For example, studies of internet sales show that niche products can account for up to 50% of the total sales of a vendor. This is contrast to the typical brick-and-mortar showroom sale where the Pareto principle often applies: namely, most of the sales are generated by a small number of top-selling items. In particular, in econometrics, the internet has been identified as an engine that tips sales curves from Pareto to long-tail.

Let us continue our effort to tying some of these ideas to the setting of a user interacting with a clustering system. A "long-tail" for such a user would mean a cluster whose infrequent terms carry "too much" weight. In other words, there is insufficient concentration of the distribution around the most frequent terms. In particular, in accordance with

the proposed form and parameter values for a Pareto-style principle for cluster centroids, we propose the following nominal definition of a long-tail, expressed in terms of both the weight function and the cumulative function.

Definition 10. *A cluster C_j is said to* have a long-tail *if the most frequent 20 terms in it account for less than 50% of the weight of its centroid. Namely, the median of $f(n, j, k)$ occurs at $n > 20$.*

Definition 11. *A cluster C_j is said to* have a long-tail *if*

$$F(20, j, k) < 0.5.$$

Once more, we recall the comments about the values of parameters being application-driven.

6.4 Effect of Varying the Number of Clusters Our discussion thus far on Pareto and long-tail ideas applied to the setting of clusters has ignored arguably the most important parameter in clustering systems—namely, the number of clusters k. The distribution $\overline{f}(n, k)$ depends upon k. In this section, we frame a natural question that arises by applying some of the ideas in the previous sections to the variation in $\overline{f}(n, k)$ brought about by k.

Switch between Pareto and long-tail. How does k affect the behavior of the pdf $\overline{f}(n, k)$? In particular, does a change in k turn a Pareto cluster into a long-tailed one? At what value of k does this happen? If this were so, it would have an impact on the design of clustering systems through multiple channels. For example, if it were known that users will be interacting with the system through 20 (or fewer) labels, then it may be advantageous to choose k such that the clustering follows a Pareto-like distribution.

7 Experiments

7.1 Datasets We experimented with three datasets that are widely used as benchmarks in document clustering. The first is the 20 Newsgroups dataset, denoted by `N20`, which contains roughly 20,000 articles posted to 20 usenet newsgroups. The articles are more or less evenly divided between the newsgroups; however some newsgroups are highly related, while others are not. The second is a subset of `REU`, the Reuters-21758 dataset, collected and labeled by the Carnegie group and Reuters while developing the CONSTRUE system. There are 82 primary topics in the documents in this dataset. The R52 subset is obtained by considering only documents with a single topic, and retaining only those topics that have at least one test and one training document. This leaves only 52 of the original 90 classes, and has 9,100 documents. Following several previous studies, we used the 'R8' subset obtained by taking the eight most frequent topics of R52. This is done in order to reduce the skewness in the class distribution of the original `REU` dataset. The resulting R8 dataset has 7,674 documents. The third dataset is the `WebKB` data set.

Table 1: Properties of the three datasets used in this study.

Dataset	Documents	Terms	Classes
`N20`	18821	68911	20
`REU` R8	7674	16984	8
`WebKB`	4199	7537	4

It consists of webpages gathered from computer science departments at various universities. Each page falls into one of seven categories: student, faculty, staff, course, project, department, and other. Previous studies have used only the four most populous categories—student, faculty, course, and project—and we do the same. These four categories consist of 4,199 web pages total.

There are multiple versions of these datasets. For replicability, we used the prepared versions available at `http://web.ist.utl.pt/acardoso/datasets/`. A summary of our datasets is provided in Table 1.

7.2 Protocol We performed a K-way clustering of each dataset, which we denote by $\mathcal{D}$. For replicability, we used the repeat-bisect algorithm from the CLUTO toolkit[2], rather than those which require random seeding. We varied K over a wide range:

$$K \in [5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100].$$

7.3 Experiment 1: Characteristics of $\overline{f}(n, k)$

Objective We wish to examine the average $\overline{f}(n, k)$, and characterize its shape.

Methodology For each k-way clustering, we computed the average $\overline{f}(n, k)$ over the k clusters, as well as the deviations $\sigma_f(n, k)$ from the average.

Results $\overline{f}(n, k)$ is shown in Fig. 2, on the following page, for $k = 5, 15, 30$ for our three datasets. The plots for higher values of k are shown in §7.6. The plots for deviations from the mean are shown in Fig. 3. Numerical values for sample deviations from the mean for $k \in [5, 15, 30, 100]$ and $n \in [1, 2, 3, 5, 10, 15, 20]$ for `N20` are shown in Table 2.

Discussion The following salient features of the function $\overline{f}(n, k)$ may be observed from the plots. Note that $\overline{f}(n, k)$ is, by definition, a monotonically decreasing function.

1. $\overline{f}(n, k)$ is significantly higher for $1 \leq n < 5$ than it is for higher values of n. This initial short segment represents the "heavy front" of $\overline{f}(n, k)$, shown shaded in the graphs.

2. The value of $\overline{f}(1, k)$ is always significant: we rarely see values lower than five, and the averages are almost 10, 20, and 30 for $k = 5, 15, 30$, respectively for `N20`. `WebKB` is similar, with the exception of $k = 15$, where the average is close to 25. For `REU`, the curves are all somewhat higher than both other datasets.

[2] Available at `http://glaros.dtc.umn.edu/gkhome/views/cluto`.

Table 2: Standard deviations $\sigma_f(n,k)$ around the mean $\overline{f}(n,k)$ of the individual clusters, for representative values of k and n for the N20 dataset. The qualitative description remains the same for REU and WebKB.

k	n						
	1	2	3	5	10	15	20
5	3.61	1.89	0.64	0.44	0.26	0.12	0.10
15	4.75	3.56	1.36	0.63	0.39	0.32	0.23
30	7.72	2.91	1.68	0.98	0.55	0.38	0.28
100	6.93	3.34	2.13	1.09	0.49	0.29	0.21

3. We find that by $n = 20$, the function $\overline{f}(n,k)$ is, in a great majority of cases, less than one. From here on, it follows a slow drop to zero.

4. Fig. 3 shows that the standard deviation is, roughly, half of the mean. This indicates that the distribution is well-behaved, and we may regard the mean $\overline{f}(n,k)$ as conveying a reasonable characterization of the distributions $f(n,k)$.

7.4 Experiment 2: Cumulative Distribution $\overline{F}(n,k)$

Objective We wish to quantify the cumulative weight of the first n coordinates of the centroid, captured by $\overline{F}(n,k)$.

Methodology We compute $\overline{F}(n,k)$ for the range of values of k.

Results Shown in Fig. 2: the upper curves in each of the plots.

Discussion The three datasets behave markedly differently with respect to $\overline{F}(n,k)$. For N20, about 30% of the total weight happens by the time $n = 20$ at $k = 5$. This number roughly doubles for $k = 30$, and then mostly saturates from then on. For REU, on the other hand, between 75% and 85% of the total weight occurs by the time $n = 20$, for all values of k we tested. WebKB is in between the N20 and REU in its cumulative distribution.

7.5 Experiment 3: Distribution of Medians of $f(n,k)$

Objective We wish to understand the distribution of the median for $f(n,k)$.

Methodology We compute each of the k functions $f(n,k)$ for the range of values of k. For each such function, we compute the median point. This gives us k medians for each value of k. We plot these as a boxplot against k.

Results Shown in Fig. 4.

Discussion The values of the median are skewed for every k. This skew itself varies with the dataset. For reference, the values of the median point for $k = 5, 10$ are given for each of the three datasets in Table 3. We use a boxplot in order to illustrate this spread of median values. The dotted line on the boxplot represents the number of clusters whose median was not reached by the time $n = 30$.

We can see that the three datasets behave differently: N20 has a fall in the median line of the box plot until we arrive at stability. REU has a less dramatic fall before stability. WebKB does not have a falling pattern before, nor does it stabilize as much as the other two datasets.

Table 3: Sample values of the median of $f(n,k)$.

	k	
dataset	5	10
N20	[26]	[5, 17, 26, 12, 25, 30]
REU	[1, 5, 3, 25]	[1, 4, 2, 3, 4, 5, 6, 10, 17, 27]
WebKB	[3, 20]	[2, 2, 3, 5, 5, 5, 16]

7.6 Experiment 4: The Effect of k: Switch between Pareto and Long-Tail

Objective We want to understand the impact of k on the Pareto vs. long-tail nature of $\overline{f}(n,k)$.

Methodology We plotted the functions $\overline{f}(n,k)$ and $\overline{F}(n,k)$ for each value of k in our range for each of our datasets and inspected the behavior at $n = 20$.

Results Shown in Fig. 2.

Discussion The first observation is that the distributions are *not always Pareto-style*. For lower values of k, they may display long-tailed behavior. We observe a marked difference between the three datasets in this regard. In REU, through the range of k, the curves are all Pareto-style. In contrast, for both N20 and WebKB, the curves *are long-tail at lower* k. As k is increased, the curves gradually become Pareto-style. This change happens at around $k = 15$ for N20, but at lower k for WebKB.

This can be tied in to Experiment 3. We see that for REU, the median is always reached early, and this corresponds to the fact that REU always remains in a Pareto-style regime throughout the range of k. In contrast, N20 has a long-tail for lower k, and we see that the boxplots for the distribution of medians are significantly higher than for high k.

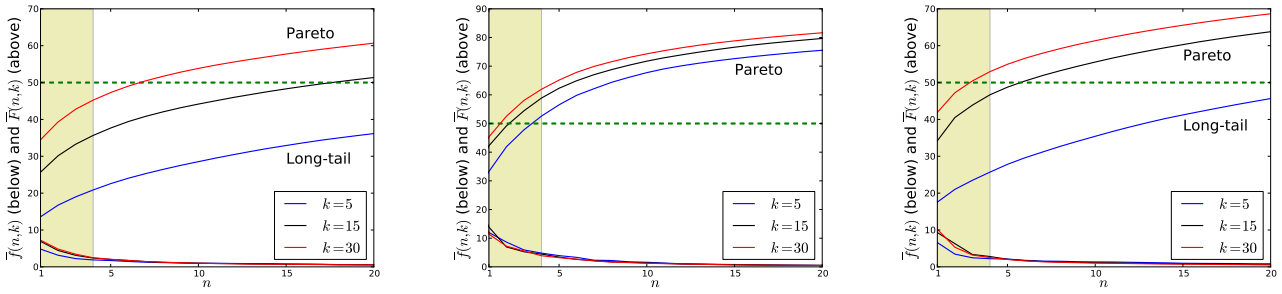
In summary, there is no single regime that characterizes the distribution of centroid weights, and the determination of which regime the centroid falls into must be done carefully in order to extract sufficient information from it.

8 Conclusions and Future Work

We have conducted a broad structural study on the distribution of weights in document cluster centroids. Our study characterizes this weight distribution into either a Pareto, or a long-tail style distribution. We show that as k is increased, the distribution tends towards a Pareto style.

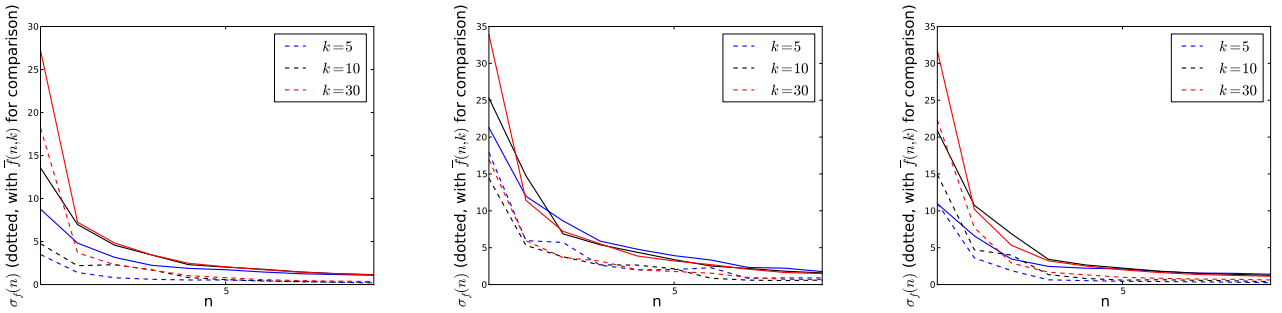
Specifically, the distribution $f(n,j,k)$ follows a power-law like distribution, with a heavy front, and a significant tail. By $n = 20$, the value of $f(n,j,k)$ has usually fallen to below one. This may mislead practitioners into thinking that a label or signature comprising the first 20 coordinates is adequate. This type of cluster labelling is frequently used in several applications.

However, even though the value of $f(n,j,k)$ falls to below one, the *cumulative remaining weight* of the function $F(n,j,k)$ for $n > 20$ varies considerably based on dataset



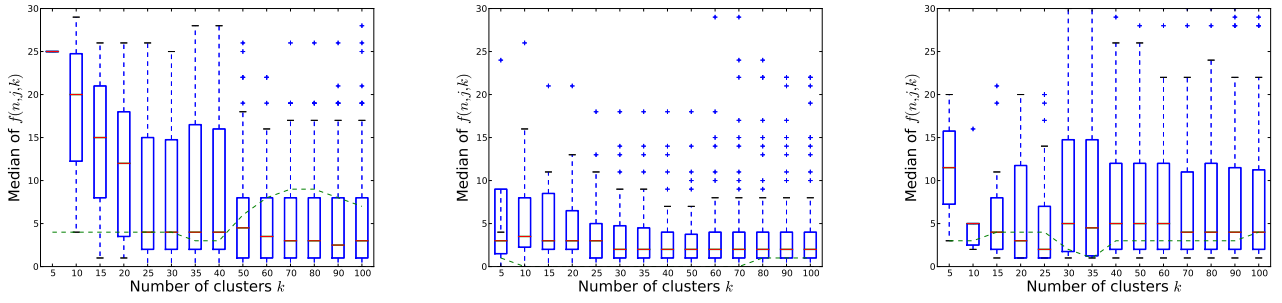
(a) N20 (b) REU (c) WebKB

Figure 2: The $\overline{f}(n,k)$ and $\overline{F}(n,k)$. The shaded region shows the "heavy front" of the $\overline{f}(n,k)$ where the values of $\overline{f}(n,k)$ are high, and are falling rapidly. To the right of this shaded portion is the tail. The dotted horizontal line represents the cumulative distribution reaching 50% of its total. Curves that cross the dotted line represent Pareto-style distributions, while those that remain below the dotted line represent long-tailed distributions. Notice how the datasets behave differently with regards to this characterization: for REU, all the curves are Pareto-style, while for the other two datasets, there is a switch from long-tail to Pareto-style as we increase k.



(a) N20 (b) REU (c) WebKB

Figure 3: The standard deviations (shown in dashed line) of the individual $f(n,j,k)$ around their mean $\overline{f}(n,k)$. We consistently see a standard deviation of around half the mean, which indicates that the mean is a good representative of the family of individual distributions.



(a) N20 (b) REU (c) WebKB

Figure 4: The distribution of the medians of $f(n,j,k)$ as a function of k. For each cluster C_j, we obtain a median; therefore there is a distribution of k such medians for each value of k. This distribution is shown as a boxplot. The dotted line shows the number of clusters whose median was not reached by the time $n = 30$.

and k. In cases where the distribution $f(n,j,k)$ follows a Pareto-style principle, the cumulative is low (nominally, less than half of the total weight). However, in many cases, the distribution has a long-tail, which means that in spite of the value at $n = 20$ being low, the cumulative of the values for $n > 20$ is highly significant, and can not be ignored. In particular, this effect has ramifications on problems such as labelling and feature selection (the two motivating examples we provided in §1). When the distribution has a long-tail, can we safely keep only the first few terms and ignore the rest? Not if we consider the cumulative weight distribution. In other words, we would expect heuristics that keep only a

few of the highest weight terms (for various applications) to have poorer accuracy in the long-tailed regimes of the distribution. This brings us back to the question of which document clusters have short signatures, and the answer in light of our framework is "those whose centroids follow a Pareto style distribution."

8.1 Future Work The following offer interesting avenues for future work.

1. From this work, it emerges that the principled way to perform information extraction from structures such as centroids is to adaptively perform the extraction based on the regime (Pareto or long-tail) the cluster centroid exhibits. Investigating such adaptive approaches provides a fruitful avenue for future work. Consider the motivating application of feature selection in a wrapper for clustering. How can we adapt the wrapper for Pareto vs. long-tail clusters? Likewise, how do we adaptively label clusters based on econometric considerations.

2. The values of the parameters for the generic definitions of Pareto and long-tail are application and dataset driven. What are reasonable heuristics and algorithms for this choice?

3. Can we classify the centroid distribution, in certain ranges of k, into power-law, exponential, or log-normal? What is the effect of k on this classification? How could such models for the centroid weight distribution be gainfully used for information extraction applications?

References

[1] Magic quadrant for e-discovery. *Gartner Report*, 2013.

[2] Market share analysis: Enterprise content management, worldwide. *Gartner Report*, 2013.

[3] Chris Anderson. The long tail. *Wired Magazine*, 2004.

[4] Erik Brynjolfsson, Yu (Jeffrey) Hu, and Duncan Simester. Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. *Management Science*, 57(8):1373–1386, 2011.

[5] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[6] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[7] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1/2):143–175, 2001.

[8] Anja Feldmann and Ward Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3):245–279, 1998.

[9] Trevor Fenner, Mark Levene, and George Loizou. Predicting the long tail of book sales: Unearthing the power-law exponent. *Physica A: Statistical Mechanics and its Applications*, 389(12):2416–2421, 2010.

[10] Michele Garetto and Don Towsley. Modeling, simulation and measurements of queuing delay under long-tail internet traffic. In *ACM SIGMETRICS Performance Evaluation Review*, volume 31, pages 47–57. ACM, 2003.

[11] Marti A. Hearst and Jan O. Pedersen. Re-examining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual ACM International SIGIR Conference on Research and Development in Information Retrieval*, pages 76–84, New York, 1996. ACM Press.

[12] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning (ICML '94)*, pages 121–129. Morgan Kaufmann, 1994.

[13] George Karypis and Eui-Hong Han. Fast supervised dimensionality reduction algorithm with applications to document categorization and retrieval. In *Proceedings of 9th ACM International Conference on Information and Knowledge Management, CIKM-00*, pages 12–19, New York, US, 2000. ACM Press.

[14] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *KDD*, pages 238–243, 1996.

[15] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, page 14. California, USA, 1967.

[16] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226251, 2004.

[17] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, May 2005.

[18] Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 11–18. ACM, 2008.

[19] Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, volume 1 of *PAPERS: Interactive Information Retrieval*, pages 213–220, 1996.

[20] William J. Reed. The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19, December 2001.

[21] William J. Reed and Barry D. Hughes. From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Physical Review E*, 66(6):067103, 2002.

[22] Fei Wu, Raphael Hoffmann, and Daniel S. Weld. Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 731–739. ACM, 2008.

[23] Yiming Yang, Tom Pierce, and Jaime Carbonell. A study on retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36, Melbourne, AU, 1998.

[24] Ying Zhao and George Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.

[25] George K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

A Supplementary Notes

A.1 Scope and Limitations Our stress is not on mathematical modeling of the centroid's weight distribution. In particular, given any empirical data, one can rarely be certain what distribution underlies it (usually, one only rules out competing hypothesis). We do not assert that the centroid weight distribution follows a power law, or related laws such as log-normal. Such an investigation would be valuable, but is beyond the scope of the present work. See [5] for the intricacies involved in such an investigation; they also show that several distributions that had previously been modeled as power laws, in fact, cannot be considered power laws. In summary, determining whether, and over what range of values, a distribution follows a power law is a separate issue, and beyond the scope of our work.

Our focus is only on phenomena of distributions that have been uncovered during the study of power laws. Our contribution is that we generalize, rework, and tie these phenomena to important aspects of unstructured textual information management, and comprehensively study them for the centroids of clusters. These properties are at a higher level than the distribution itself: namely, they can be defined and investigated for a centroid without ascertaining precisely what distribution is most likely to have given rise to it.

A.2 Data Mining Preliminaries Since this work is interdisciplinary, we describe the data mining setting of our study in some detail for the benefit of readers from econometrics.

A.2.1 Corpus and vocabulary Our data $\mathcal{D}$ will comprise a document collection (or corpus) $\mathcal{D} = \{D_1, \ldots, D_n\}$. We assume a standard preprocessing of documents for clustering: tokenization of words, stemming, removal of stoplists, and removal of infrequently occurring words (less than thrice) in the corpus. The resulting set of tokens are called *terms*. The *vocabulary* of $\mathcal{D}$, denoted by $\mathcal{V}$, is the set of all terms in the documents of $\mathcal{D}$. Typical text corpora have vocabularies of size $|\mathcal{V}|$ of the order of tens of thousands of terms. We assume a lexicographic ordering of the terms in the vocabulary, so that it makes sense to talk of term t_i: the i^{th} term in this ordering.

A.2.2 The vector space model There are various approaches to clustering a collection of text documents. However, most of them rely on some form of a vector space representation. A commonly used vector space representation is TF-IDF. A document D is represented by the vector

$$\mathbf{v}_D := \left(\mathrm{tf}_1 \log \frac{|\mathcal{D}|}{\mathrm{df}_1}, \mathrm{tf}_2 \log \frac{|\mathcal{D}|}{\mathrm{df}_2}, \ldots, \mathrm{tf}_{|\mathcal{V}|} \log \frac{|\mathcal{D}|}{\mathrm{df}_{|\mathcal{V}|}}\right),$$

where the "term-frequency" tf_i is the frequency of occurrence of term t_i in D, and the "document-frequency" df_i is the number of documents in $\mathcal{D}$ that contain t_i. The TF-IDF vectors so obtained are of differing lengths. Therefore, a length-normalization step is included where all the vectors are normalized to having unit length.

A.2.3 Similarity metrics All clustering algorithms rely on some underlying notion of similarity on the document universe. Formally, similarity is a function $s(\_, \_) : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^+$ on pairs of documents that satisfies our intuitive idea of what it means for two documents to be "similar." It will, in general, not be a metric. Various notions of similarity have been proposed and studied in literature. By far, the most commonly used for the purposes of clustering of textual documents is cosine similarity given by

$$s(D_i, D_j) := \cos(\mathbf{v}_{D_i}, \mathbf{v}_{D_j}) = \frac{\mathbf{v}_{D_i}.\mathbf{v}_{D_j}}{\|\mathbf{v}_{D_i}\|\|\mathbf{v}_{D_j}\|}.$$

We will use this notion of similarity for our clustering algorithms.

A.2.4 Clustering algorithms The general problem of clustering is defined as follows. Given a universe $\mathcal{D}$ of documents, we would like to partition them into a pre-determined number of k subsets (known as clusters) $\{C_1, C_2, \ldots, C_k\}$ such that the documents within a cluster are more similar to each other than they are to documents that lie in other clusters.

There are various approaches to clustering documents, and the clusters produced by different approaches produce different clusters, based on the notion of cohesiveness of document classes used by the approach. Clustering algorithms can be categorized based either on the underlying methodology of the algorithm, or on the structure of the clusters that are output by the algorithm. The first approach results in a division into agglomerative or partitional approaches, while the second leads to a division into hierarchical or non-hierarchical solutions.

Agglomerative algorithms work "bottom-up." They find clusters by engaging in a while loop that initially assigns each object to its own cluster and then repeatedly merges pairs of clusters until a certain stopping criterion is met. On the other hand, partitional algorithms work "top-down". They either find the k clusters directly, or through a sequence of repeated bisections where they create finer clusters at each step. Classical partitional algorithms include k-means [15] and k-medoids, among many others.

It is generally accepted [6] that when clustering large document collections, partitional clustering algorithms are preferable due to their relatively low computational requirements. Following a comprehensive comparative study of various clustering algorithms by [24], we use the repeat-bisect algorithm to generate our clusters. This has the added benefit of producing identical clusters at each run, thereby making the functions that we define in later sections purely functions of the dataset and K. Our choice of repeat-bisect was also guided by its good scale-up to enterprise class problems in information management.