Navigating Economic Analysis in a World of Big Data

Allan Shampine, Loren Poulsen, and Michael Sabor

The availability of more and larger data sets has enabled consultants to develop more complex analyses that may offer greater precision in the estimation of the antitrust effects of alleged mergers or anticompetitive practices (e.g., price fixing). New types of projects are being undertaken, as are more complex versions of the type of analyses offered in prior matters.

This is an evolving area, and even counsel who have often worked with consultants on empirical projects may be surprised that those which had become seemingly routine in the past are now proving slower and harder to adjust to changing circumstances—due to both the increase in big data availability and the more complex analytic tools developed to process those data. While there is no bright line for what constitutes “big data,” if you are dealing with multiple terabytes of data then the analysis is likely to have moved past the common practice of 20 years ago.

This article discusses some practical considerations practitioners may encounter as a result of larger and more diverse data sets and provides an example of taking an analysis from initial stages to final work product, focusing on how changes in big data may raise new concerns at various stages. Practitioners should be aware of both the advantages and potential pitfalls of relying on big data.

At a big picture level, the questions being asked are not new. But there are now new ways to try to answer those questions, which can be seen in three types of commonly used analyses: (1) cross-sectional analysis, (2) longitudinal studies, and (3) demand modeling.

Using Common Analyses and Modeling with Big Data Sets

In the context of antitrust, cross-sectional analysis refers to studying differences in prices and demand patterns between different geographic areas, taking advantage of differences in competitors’ presence or relative size in different regions, or the presence or absence of particular concerns (e.g., discrimination across different customer groupings). These sorts of analyses are common in antitrust practice and may be relevant for examining price fixing, information sharing, mergers, or market definition. Here big data does not qualitatively change the sorts of questions that might be asked, or even at a general level how one might ask them, but it does allow levels of granularity that were unimaginable a few decades ago. That is, big data can help us, for example, by looking at very specific cross-sections of customers to determine if there are different behaviors across different classes of customers. Rather than looking at the “representative consumer” in a demand analysis, we can ask if these same patterns persist in households with different levels of income, education, native language spoken, or ethnic background. This allows for a more flexible functional form of demand (in technical terms, essentially an unspecified, non-parametric demand function) that can then be used to ask the counterfactual questions that are typical of those used in antitrust.

Big data can also be used in a cross-sectional analysis to estimate possible competitive effects directly. For example, how does the presence or absence of a competitor or particular policy
change impact consumers? Can differences in consumer prices or behaviors be observed between different areas with different competitive conditions? For example, it is becoming more common for retailers to engage in controlled experiments with pricing or policy changes, applying those changes to only some customers, or only for a short period of time. Such experiments are not only more common than in the past, but are performed at much larger scales, possibly affecting millions of consumers, and thus yielding far more data than in the past. These experiments may provide indirect evidence of competitive effects, or, at the least, more evidence on the form of consumer demand.

Longitudinal studies look at differences across time addressing the same kinds of questions as above but adding a time dimension to the analysis. Again, these are commonly used in antitrust analysis, particularly when assessing the existence of price fixing and evaluating the competitive risk from mergers. With the cost of data storage decreasing, it is more often the case that the client may now have much more historical data than previously available. But in addition to the standard questions, longitudinal big data also allows us to ask more interesting questions related to changes in market structure over time. For instance, perhaps all markets today have the same competitors but those competitors historically have expanded from different regions. Time series analysis can allow us to understand how competition has evolved over time region by region, state by state, or even city by city in some cases. Here, we can now do analyses that were not previously possible because the data did not exist. Big data has been around long enough that it is becoming more likely that there will be rich data sets available far enough into the past that they can address these sorts of questions.

Demand modeling is an example of a specific type of analysis, mentioned above, that can be approached either with cross-sectional or longitudinal data. In practice, demand modeling has changed considerably in recent decades. Indeed, large, sophisticated companies with many consumers may perform a series of experiments holding constant demographics and changing prices or product attributes to understand how it changes demand for their products and where consumers go (in terms of locations or a different product space) in the face of a price increase or quality decrease. Historically, firms like these have conducted surveys or small-scale experiments and then built a formal model of demand using the insights of those initial efforts with the goal of predicting how consumers will react to various product and price changes or sales approaches.

With the availability of big data, firms may approach the problem of estimating consumer price sensitivity rather differently with less theoretical demand modeling, allowing for more flexible functional forms and “letting the data do the talking.” This difference has important implications as the investigation proceeds while using big data. Specifically, there can be conceptual differences between client firms and economists in how to approach these problems. As noted above, under the traditional approach, a firm might have a theory on what drives consumer demand and build a model based on that theory. The firm might then use data to try and calibrate the model and make predictions. With big data now available, a firm might work in the other direction—start by measuring the relationships between all of the available variables and consumer behavior and then write down a predictive model based on those observed relationships without any attempt to motivate the underlying economic relationships between the variable of interest and its covariates. Given enough data and computing power, such predictions may be very useful to a business. However, an economist may find it challenging to fit a pre-existing ad hoc predictive model of this sort into a traditional antitrust analysis. Here, the economist may either construct a new demand model or try to observe competitive impacts directly.1

1 Of course, the economist may have to explain to the court or the antitrust authorities how and why the models differ.
The key to all three of these analyses (cross-sectional, longitudinal, and demand modeling) is having access to a rich set of covariates, such as demographics, competitive pricing, and changes in structure of competition or product attributes. Big data can be helpful here but can also present challenges. For example, there are some peculiar tensions with big data when it comes to statistical significance. As an illustration, big data may contain an enormous amount of granularity with respect to all sorts of consumer characteristics, so much so that any particular slice of the data may end up being quite small. That is, an analyst might have access to years’ worth of transactions data for millions of consumers, but there might be so much detail about the consumers that the analysis is being done based on small income slices at the census block level, which may result in any given slice having only a handful of consumers in it. Thus, data that is big in the sense of having many, many fields for a particular observation can lead to extremely small relevant groups, while data that is big in the sense of having many different people or firms represented can lead to extremely big groups, which can pose a different set of problems, such as improper aggregation (which could mask the effects of relevant characteristics).

The lesson is that big data can give us the freedom to ask more detailed questions and tease out specific nuances of consumer behavior but can also result in losing sight of the forest because of the trees. Just because we can look at extremely granular data does not mean we necessarily should. The granularity may or may not shed additional light on the question of interest, and may obscure it, but it may also reveal subtleties that aggregation obscured. Another challenge of big data is that conventional tools may not have been designed to handle such large data sets, which can pose practical problems (e.g., will a particular software package be able to process the millions of observations at all). But there also are significant theoretical problems: if the package does run, should the interpretation of the results change? A great deal of econometric work is focused on drawing inferences and making predictions about a population based on studying a sample. However, with big data we may not have to settle for a sample—we may literally be able to study the entire relevant population, or a very large portion of it. That is, traditionally, regression analysis (and statistics in general) have been used to infer central tendencies of various populations from samples of that population. Because they were only small samples, questions about the reliability of the estimate (e.g., confidence intervals and statistical significance) played a central role in the analysis. The legal framework recognizes the analytic importance of those questions, and counsel will often expect any estimates to be accompanied by statistical measures of the reliability of those estimates.

The challenge with a small data set is that any model may be imprecisely estimated because, by definition, there are not enough data to generate very precise estimates. With big data, that particular limitation may disappear, but that does not mean that a model with very small standard errors is necessarily reliable. Rather, the focus shifts to other sources of error. That is, with billions of observations, any model may be very precisely estimated in the sense of the coefficients having very small standard errors, but the same may be true for any number of possible model spec-

---


3 It may also be the case that the interpretation of some particular statistical tests may change. For example, some tests may be motivated based on small sample properties, and those properties may be different in a big data world. In the extreme, the analyst might not have a “sample” at all, but might have data on the entire population. However, the more relevant issue from the practitioners’ perspective in the legal context is likely to be a shift in how much attention is paid to different potential sources of error.
ifications. Debates over model specification may therefore become more critical than in the pre-big data universe, while debates over statistical significance become less so. Similarly, other sources of error (e.g., measurement error, bias, model robustness, etc.) are also likely to become more important topics. The presentation and framing of results in terms of statistical significance and the choice of estimated models may therefore be different than in the past. This is very much an evolving discussion amongst practitioners and one that needs to be closely watched.4

Anatomy of a Big Data Engagement

The discussion above is general. But it is also important to consider the specific ways in which a big data engagement, from data acquisition to data processing and then to final output, differ from a more traditional engagement (i.e., one with more limited data). The remainder of this article discusses a hypothetical big data engagement, and describes how differences from a traditional engagement can affect the practicality of using big data to assess antitrust questions.

A first step is to determine whether the engagement involves big data. Here, the analyst will discuss with the client the scope of the data available: number of tables, number of observations per table, number of columns per table, number of string versus numeric columns, and, of course, the total size of the data. As noted earlier, there is no bright line for what constitutes “big data,” but if the answer to the “data size” question is “multiple terabytes,” then you are well beyond the data sizes of even five years ago.

The data collection process, requiring as it does close coordination between the analysts and clients, has always posed challenges, but those challenges can be substantially different with big data. The difference can be thought of as between a trickle of water and a fire hose. Historically, the analyst may have had to make due with a trickle of data, making the best use possible of limited data. Every drop was precious. With big data, the analyst may be confronted with a fire hose of data, much of which may ultimately be irrelevant. Determining and extracting the relevant parts (which are still likely to be extremely large) can be a substantial challenge. Thus, finding the “correct” database or data files amongst the client’s collection is always an issue, but the particular challenges may change from trying to find any relevant data at all to paring down the many possible data analyses to a manageable subset of the data.

Once the analyst and client have reached agreement on the data to be transferred, the data export process itself will depend greatly on whether big data are involved. There are often trade-offs between ease of export and ease of use by the analyst. For example, an Oracle database can be easily exported using Oracle’s proprietary backup facility, but this is typically a costly and inconvenient method for the analyst. Alternatively, the client can create tables that include the relevant data by writing to “flat text files,” which are easy to read and use in many other software packages but require more effort on the part of the client to assemble the relevant data set.

Many export methods allow users to break up and compress a data set, e.g., to export the data set to many files versus a single large file, and some allow for specific data size limitations imposed by the analyst. For example, the client may choose to break the output up into many files where the maximum size is specified, so that the client can be assured the files will fit over some transport medium, such as an e-mail attachment limit or a high-capacity thumb drive. This consideration can impact the difficulty of transporting and the ingesting of data and will be a function of both the size of data being sent and the facilities of the analyst.

---

4 See Athey et al., supra note 2.
Practitioners are generally familiar with compression algorithms such as “zip” files used for data “export” to the consultant, and there are many compression algorithms, many of which will also allow for the compressed data to be broken up into multiple pieces for transport. For large data files, however, this process of compression and decompression, dividing and recombining, can be time consuming and may provide less assistance in transport than might be expected. That is, compressing a few hundred megabytes of data into three ten-megabyte files that can be e-mailed may be very useful. Compressing a few hundred terabytes of data into thousands of ten megabyte files will be time consuming and will still not make e-mail practical for transmission. We discuss the transport issues further below.

Privacy issues are also common with big data and can complicate production and transport because encryption is more challenging and time consuming than in the past, simply because of the volumes involved. The encryption may occur on different “data size” levels, such as encrypting individual files or full disk encryption. Furthermore, both compression and encryption are likely to be used together. In conjunction, and for large amounts of data, compression/encryption and decompression/decryption can be time consuming, particularly if the data are also being broken up into smaller files. Errors and incompatibilities can also arise, which will be time consuming to sort out. We have all experienced problems with zip file productions, for example, and know how the back-and-forth to resolve those problems can take days. Now imagine doing that with multi-terabyte files that are being transferred back and forth by hand-delivered hard drives and that take a day to decompress/decrypt even when everything works correctly. The time required to resolve these types of transfer issues could be substantial, even before any actual antitrust analysis begins.

The physical transport of big data poses challenges as well. E-mail is typically not a realistic option for big data. But there are other common methods. First, an electronic transfer. A secure FTP site with significant bandwidth capabilities might be used. The “significant bandwidth” condition is critical. Both the FTP host and recipient need to have significant bandwidth available, given the large size of the data. As a result, most client and analyst FTPs in everyday use will not be appropriate.

Second, hard drives or thumb drives can be used. There are practical issues involving making multiple copies and arranging for physical delivery, but for large data sets, that is not necessarily slower than the use of a FTP site. However, if the data sets are very large, then the client will likely have to consider a third option—a stand-alone server. There are services that will securely transport and deliver servers and racks of hard drives. Indeed, major cloud services vendors will come pick up a server and take it to their facilities for data transfer. If a stand-alone server or rack of hard drives is used, the client and analyst will also likely wish to discuss the physical security of the server during transit and at the final facilities where it will be stored, whether that is at the analyst’s facilities or the facilities of a vendor (or both).

Getting the data physically in the hands of the analyst does not mean analysis can immediately begin. The analyst must first “ingest” the data. This includes decompressing it, decrypting it, checking for corruption of files, recombining any files that were broken up for transit, and transferring all of the data into whatever hardware and software environment the analyst wishes to use. This can take weeks for big data, particularly if the analyst is relatively unfamiliar with working with such large data sets.

Basic analysis can begin once the transfer issues are resolved. The first step here is usually a cleaning process: run basic summary statistics, review outliers, identify and clean errors, etc. As part of this, the client will likely have provided anecdotes and stylized facts about what the data
show. Do the data received appear to be consistent with those anecdotes? If not, is it because there is a problem with the data or with the anecdote? For example, an analyst might have received geospatial data on the locations of stores and delivery centers. The analyst might then confirm its location accuracy, e.g., a simple test might be whether some locations are on water rather than on land; a more specific test might be to look at a sample and confirm the coordinates are in the general expected location of stores and delivery centers. Again, this process can be time consuming, generally, and is even more so with big data. For example, with ten times as many variables and ten thousand times more observations as a more conventional data set, there are likely more oddities to be explored. That does not mean big data are any less reliable, simply that the data may cover much more ground and so take more time to understand. This will likely involve considerable back-and-forth between the analyst and the client.

Next is to determine the logic for creating a “record” of each observation—the inclusion of all of the relevant data for each observation—for the analysis itself. Most data come in a form that is not immediately relevant to answer the questions that are typically asked for the purposes of an antitrust analysis. For instance, sales data often needs to be rolled up to a daily, weekly, or monthly price net of discounts and rebates. Such data are often produced in an invoice file and a rebate accrual file.

The analyst may also be considering a sampling strategy. Are the resources available to employ the full data set or will the analyst wish to extract a valid statistical sample? This is an area where even with cloud computing, using the full data set may not be practical. It will depend in large part on the question being asked. If the analysis is amenable to being done in parallel (running lots of processes simultaneously), it is more likely that cloud computing will allow even big data to be used in full. Alternatively, if the analysis requires processes to be run sequentially, one after another, the analysis of the entire data base may be too time consuming and so data aggregation or sampling techniques may be called for.

Big data can have a significant impact on processing time. In some cases, computing power has grown quickly enough that processing time can be shorter than in the past despite the size of the data set being much larger, but that depends greatly on the specific type of problem being analyzed. Consider the following example. A merger litigation several decades ago involved shop keeping unit (SKU) level sales data for two firms. The data were beyond the processing capabilities of most analysts at the time and the authors’ firm was engaged because we had a supercomputer. Even so, running a regression on the data could take upwards of 12 hours, which in the middle of litigation is an eternity.

Recently, the authors have worked with large (multiple terabyte) data sets where statistical work on local computing facilities in SAS, MS-SQL, or Stata still takes over eight hours. However, using an optimized cloud computing framework we have developed, that particular statistical work that in a local SAS, MS-SQL, or Stata environment took over eight hours now runs in only 15 seconds.\(^5\) So while it is generally the case that big data means everything takes longer and costs more, there are some exceptions.

Having resolved these various issues, the output of this process at this point will likely look familiar to all those involved. Basic charts, tables, and summary statistics are produced to get a more detailed understanding of the data and to “gut check” the data build with the client (e.g., are aver-

age prices what would be expected, are regional differences in prices reasonable, etc.). As noted earlier, big data is likely to produce models with small standard errors, so practitioners may want to inquire more closely about other measures of error, such as robustness of the results to small changes in the specification of the model, or the possibility of measurement bias. Moreover, just because a statistic is precisely measured does not mean that it is necessarily economically relevant (e.g., accounting margins and economic margins are not the same thing, and precisely measuring the accounting margin does not itself make the accounting margin a better measure of the economic margin). Similarly, even if a variable has a statistically significant impact on price, for example, it does not mean that the effect of that variable on price is quantitatively important. The effect could be trivial.

**Conclusion**

While the basic analyses today are the same whether or not big data is involved, the mechanics of engagements can differ in important ways. There are logistical challenges unique to big data, and the methodologies used in the analyses may be influenced by hardware and software limitations. Practitioners may also wish to pay less attention to standard errors and more attention to other forms of error, such as misspecification of the model.●