Journal of Economic Perspectives—Volume 26, Number 2—Spring 2012—Pages 189–206

Using Internet Data for Economic Research

Benjamin Edelman

The data used by economists can be broadly divided into two categories. First, structured datasets arise when a government agency, trade association, or company can justify the expense of assembling records. The Internet has transformed how economists interact with these datasets by lowering the cost of storing, updating, distributing, finding, and retrieving this information. Second, some economic researchers affirmatively collect data of interest. Historically, assembling a dataset might involve delving through annual reports or archives that had not previously been organized into a format ready for research. In some cases researchers would survey stores, factories, consumers, or workers; or they could carry out an experiment. For researcher-collected data, the Internet opens exceptional possibilities both by increasing the amount of information available for researchers to gather and by lowering researchers' costs of collecting information. In this paper, I will explore the Internet's new datasets, present methods for harnessing their wealth, and survey a sampling of the research questions these data help to answer.

The Internet had 346 million sites as of June 2011, growing at a rate of 30 percent per year (Netcraft 2011). Many websites that are household names offer data of interest to economists. For example, Monster.com organizes available jobs; Amazon.com offers data on prices and sales of all sorts of items; eBay posts records of the bidding process for every listing; and Facebook and Twitter organize information about social connections, consumption choices, and preferences for privacy. The first section of this paper discusses "scraping" the Internet for data—that is, collecting data on prices, quantities, and key characteristics that are already available on websites but not yet organized in a form useful for economic research.

■ *Benjamin Edelman is Assistant Professor of Business Administration, Harvard Business School, Boston, Massachusetts. His website is ⟨www.benedelman.org⟩.*

http://dx.doi.org/10.1257/jep.26.2.189. doi=10.1257/jep.26.2.189

The second main section of the paper then considers online experiments. This category includes experiments that the economic researcher observes but does not control—for example, when Amazon or eBay alters site design or bidding rules. It also includes experiments in which a researcher participates in design, including researcher-designed experiments; experiments conducted in partnership with a company or website; and online versions of laboratory experiments. Finally, I discuss certain limits to this type of data collection, including both "terms of use" restrictions on websites and concerns about privacy and confidentiality.

A wealth of data exists online mainly because the general public wants access to it. With so much data readily accessible, researchers often need not obtain any special permission to obtain the data they seek. Furthermore, easy data availability helps avoid selection bias: data providers tend to decline researcher requests for data that (they believe) reflect unfavorably on them, but comprehensive online postings sometimes reveal surprisingly detailed information. The panorama of available online data may especially benefit researchers early in their careers. Online data is often available without delay and thus allows the preparation of original empirical research under tight time constraints—particularly helpful for students writing papers within the constraint of an academic term. Unrestricted access also assists anyone whose credentials might fail to satisfy the gatekeepers who evaluate requests for internal data from companies and organizations. Meanwhile, the costs and difficulty of collecting such data are modest. Most undergraduate computer science students can design a basic system to collect data from a website, using tools as basic as macros or scripts, so lack of advanced programming skills need not stand in the way.

Online data can speak to almost every field of economics, including subjects well beyond software, networks, and information technology. The next section identifies representative examples, focusing on novel data sources while flagging useful tools and techniques as well as recurring challenges.

Scraping the Internet to Collect Data

Consumers and competitors push websites to post remarkable amounts of information online. For example, most retail booksellers would hesitate to share information about which items they sold. Yet eBay posts the full bid history for every item offered for sale, and Amazon updates its rankings of top-selling items every hour. Surveying job seekers can be time consuming, but sites like Monster.com organize available jobs, making it easier and quicker to track job search among narrower groups. Researchers in many fields of economics have discovered the benefits of online data collection. Table 1 presents papers, each using online data, drawn from every top-level category in the *Journal of Economic Literature* classification system. Below, I turn to specific methods of online data collection, with details on their respective applications as well as representative research using each method.

Table 1
Diverse Papers Grounded in Online Data

History of Economic Thought

Azar, Ofer H. 2007. "The Slowdown in First-Response Times of Economics Journals: Can it Be Beneficial?" *Economic Inquiry* 45(1): 179–87.

Examines trends in the timing of journals' responses to submitted manuscript, collecting response time data from journals websites.

Microeconomics

Bajari, Patrick, and Ali Hortacsu. 2003. "The Winner's Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions." *RAND Journal of Economics* 34(2): 329–55.

Bid data from coin sales on eBay reveal bidder behavior in auctions, including the magnitude of the winner's curse.

Macroeconomics and Monetary Economics

Cavallo, Alberto. 2011. "Scraped Data and Sticky Prices." January 27. http://www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.

Daily price data from online supermarkets reveal price adjustment and price stickiness.

International Economics

Philipp Maier. 2005. "A 'Global Village' without Borders? International Price Differentials at eBay." DNB Working Paper No. 044, De Nederlandsche Bank.

Purchases at eBay reveal differences in real prices between countries, including differences between countries without currency friction.

Financial Economics

Antweiler, Werner, and Murray Z. Frank. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59(3): 1259–94.

Finds that online discussions help predict market volatility; effects on stock returns are statistically significant but economically small.

Public Economics

Ellison, Glen, and Sara Fisher Ellison. 2009. "Tax Sensitivity and Home State Preferences in Internet Purchasing." *American Economic Journal: Economic Policy* 1(2): 53–71.

At a comparison shopping service, click patterns reveal users' efforts to avoid sales taxes.

Health, Education, and Welfare

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature,* February 19, Volume 457(7232): 1012–14.

Trends in users' web searches identify influenza epidemics.

Labor and Demographic Economics

Edelman, Benjamin. Forthcoming. "Earnings and Ratings at Google Answers." *Economic Inquiry.*

Measures labor market outcomes in an online research service, including higher earnings for experience, flexibility, and disfavored work schedules.

Law and Economics

Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, and James R. Marsden. 2006. "Impact of Legal Threats on Online Music Sharing Activity: An Analysis of Music Industry Legal Actions." *Journal of Law and Economics* 49(1): 91–114.

Observes the prevalence of users sharing music files via peer-to-peer networks, and analyzes users' response to an increased likelihood of litigation.

Industrial Organization

Chevalier Judith, and Austan Goolsbee. 2003. "Measuring Prices and Price Competition Online: Amazon.com vs. BarnesandNoble.com." *Quantitative Marketing and Economics* 1(2): 203–222.

Uses publicly available price and rank data to estimate demand elasticities at two leading sellers of online books, finding greater price sensitivity at Barnes & Noble than at Amazon.

Table 1—continued

Business Administration and Business Economics; Marketing; Accounting

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Duncan Simester. 2011. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales." *Management Science* 57(8): 1373–86.

Data from an online clothing store show that the Internet reduces product search times and enables successful niche products.

Economic History

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, January 14, 331(6014): 176–82.

Analyzing approximately 4 percent of all books ever printed, as scanned by Google Books, the authors profile changes in word choice and grammar, the duration of celebrity, and the problems of censorship and suppression.

Economic Development, Technological Change, and Growth

Seamans, Robert, and Feng Zhu. 2010. "Technology Shocks in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers." NET Institute Working Paper 10-11.

Explores the response of local newspapers to entry by Craigslist, including increase in subscription price, decreasing advertising price, and decreasing classified price.

Economic Systems

Hsieh, Chang-Tai, Edward Miguel, Daniel Ortega, and Francisco Rodríguez. 2011. "The Price of Political Opposition: Evidence from Venezuela's Maisanta." *American Economic Journal: Applied Economics* 3(2): 196–214.

Drawing on a "punishment list" published by Hugo Chavez, containing the names of people who signed a petition against him, this paper estimates the "price of political opposition"—the lost earnings of disfavored persons.

Agricultural and Natural Resource Economics

Camacho, Adriana, and Emily Conover. 2011. "The Impact of Receiving Price and Climate Information in the Agricultural Sector." IDB Working Paper IDB-WP-220.

Finds that farmers who received SMS (text message) information about price and weather had a narrower dispersion in expected price of their crops and a significant reduction in crop loss.

Urban, Rural, and Regional Economics

Kroft, Kory, and Devin G. Pope. 2008. "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist." http://faculty.chicagobooth.edu/devin.pope/research/pdf/JPE_Final_with_figures.pdf.

Measures the number of Craigslist posts in metropolitan areas to observe Craigslist's effect on classified ads, apartment and housing rental vacancy, and unemployment.

The Basics of Online Data Collection

For a researcher seeking to collect data from the Internet, the simplest approach is typically a one-time collection of structured information from a single site. For example, Roth and Ockenfels (2002) retrieve data from about 480 eBay and Amazon auctions, finding that Amazon's automatic extension of auction closing time obviated the incentive for late bidding (which was observed much more often at eBay). Carlton and Chevalier (2001) retrieve data about prices of fragrances, DVD players, and refrigerators from an array of retail sites to explore the relationship between retail promotion, online availability, and price dispersion. They find

that manufacturer websites tend to charge higher prices than retail websites, and that manufacturers who limit distribution of their product in the physical world also tend to avoid having their product appear on websites that offer especially deep discounts. Freedman and Jin (2011) retrieve data about lenders and borrowers at the microfinance person-to-person lending site Prosper.com, assessing the magnitude of asymmetric information, efforts to mitigate information asymmetry, and pricing of risk. They find that early lenders apparently did not understand the risks involved, and as the lenders learned to evaluate these risks, high-risk borrowers found themselves unable to receive loans through this mechanism.

Online data collection typically calls for a three-step procedure: A first component retrieves web pages from a web server, often checking an input file for a list of pages to retrieve, terms to search for, or similar instructions. Next, a parser receives web server replies and extracts the data elements of interest to the researcher, storing this data in an output file. Finally, a researcher can use standard tools to import and analyze the output file. Of course numerous refinements on this approach are possible, and subsequent sections present some of the variations.

Researchers can implement data collection systems in a number of programming languages. When demonstrating data collection in a classroom or assignment, I often offer Visual Basic for Applications (VBA) code that stores data directly in Excel—letting Excel handle storage of both inputs and outputs. Figure 1 presents sample code from this approach. For larger projects, researchers often choose Perl, PHP, or Python.

Long-Term Data Retrieval

It is sometimes helpful to collect data over an extended period, which yields a larger sample as well as insight into changes over time. When data collection is repeated periodically—a step that can be programmed to occur at preset time intervals—a one-time script can be converted into a system for continued data retrieval.

For example, Ellison and Ellison (2009) collect data on consumers making online purchases of computer memory modules. They examine hourly data over a year, using the Pricewatch.com search engine, which displays lists of products being sold by participating online retailers at their own website. To measure users' incentive to buy from out-of-state online sellers (often thereby avoiding state sales tax), Ellison and Ellison need variation in users' choice sets to draw conclusions about user preferences. Because choices change as online sellers adjust their prices, more-frequent data improves the power of the analysis. They find that states with higher sales taxes for off-line purchases tend to have more online purchases—with buyers thereby often avoiding the state sales tax.

Long-term data collection can be particularly important for analysis of price levels and inflation. Cavallo (2011) collected daily prices of 80,000 supermarket products taken from the public web pages of online retailers in four countries over a period of three years. This price data is then used to explore patterns in price stickiness. Cavallo finds a bimodal pattern of price changes—that is, price changes are significantly positive or negative but few changes are close to zero, confirming

Figure 1
Sample Worksheet and Script for Basic Online Data Collection

	A	B
1	ISBN	Rank
2	0300151241	
3	0691143285	
4	0195340671	

```
Sub GetAmazonDataDemo()
    'retrieves sales ranks from Amazon
    'INPUT:        sheet1 column A – gives a list of ISBN-10 numbers
    'OUTPUT:       sheet 1 column B – reports Amazon sales ranks

    Dim curcell As Range, htmlresponse As String    'required variables

    'iterate through rows of the first column
    For Each curcell In ActiveWorkbook.Sheets("Sheet1").Range("A:A")
        'leave the loop when encounter a blank value in column A
        If curcell.Value = "" Then Exit For

        'row 1 is the header row – don't do anything there
        If curcell.Row > 1 Then
            'get Amazon product detail page with this ISBN
            htmlresponse = GetURL("http://www.amazon.com/o/ASIN/" & Trim(curcell.Value))

            parse1 = GetBetween(htmlresponse, "Sales Rank:", " in")
            rankval = GetAfter(parse1, "#")

            'store rank in column B
            curcell.Offset(0, 1).Value = rankval
        End If
    Next curcell    'proceed to next row
End Sub

Function GetURL(url)
    'INPUT:    a URL to be retrieved
    'OUTPUT:the HTTP body at the specified URL
    'REQUIREMENTS:      a HTTP object: Winhttp, Xmlhttp, Winhttprequest, or Serverxmlhttp
    'ERROR HANDLING:    none – should be added for production use!

    On Error Resume Next
    Set objHTTP = CreateObject("WinHttp.WinHttpRequest.5.1")
    If objHTTP Is Nothing Then Set o = CreateObject("Microsoft.xmlhttp")
    If objHTTP Is Nothing Then Set o = CreateObject("winhttp.winhttprequest")
    If objHTTP Is Nothing Then Set o = CreateObject("MSXML2.ServerXMLHTTP")
    objHTTP.Open "GET", url, False
    objHTTP.send
    GetURL = objHTTP.responsetext
End Function

Function GetBetween(s, s1, s2)
    'INPUT:    s - string to search;
              s1 - string that marks start of retrieval
              s2 - string that marks end of retrieval
    'OUTPUT: the portion of s that comes strictly between s1 and s2
    'ERROR HANDLING: if s1 or s2 is not found, returns a blank string

    p1 = InStr(s, s1)
    If p1 = 0 Then Exit Function        's1 was not found
    p2 = InStr(p1 + Len(s1), s, s2)
    If p2 = 0 Then Exit Function        's2 was not found
    GetBetween = Mid(s, p1 + Len(s1), p2 - p1 - Len(s1))
End Function

Function GetAfter(ByVal s As String, ByVal s1 As String)
    'INPUT: s - string to search; s1 - string that marks start of retrieval
    'OUTPUT: the portion of s that comes strictly after s1; if s1 is not found, all of s
    p1 = InStr(s, s1)
    If p1 = 0 Then GetAfter = s: Exit Function
    GetAfter = Mid(s, p1 + Len(s1), Len(s) - p1 - Len(s1) + 1)
End Function
```

the predictions of menu cost models. He also finds synchronization of prices for goods that are close competitors within product categories.

Multistep Data Retrieval

More complex data retrieval systems can monitor multiple sources and adjust their configurations based on what occurs. For example, in Edelman (2002), I use a two-step process to measure the effects of recommendations by Amazon's editorial staff. A first system identifies which books Amazon's editorial staff recommend, running repeatedly to uncover new books soon after the recommendations begin. Then a second system tracks the sales rank of each such book—allowing measurement of the sales increase attributable to Amazon's recommendation. This approach allows examination of interaction between multiple economic actors, and if one set of events is at least locally exogenous, this approach can identify causal effects.

To measure users' sharing on peer-to-peer networks, Bhattacharjee, Gopal, Lertwachara, and Marsden (2006) also rely on a multistep collection process. A first system searches for randomly selected genres to retrieve a list of users sharing music. For selected users, a second system then activates a "Find More From Same User" function to retrieve information about the total number of songs that user is sharing. The authors collect weekly data for a year, yielding a measurement of users' response to threatened litigation by the recording industry. Notably, Bhattacharjee et al. collect data not from sites presented to users in browsers such as Internet Explorer and Firefox, but rather from Kazaa and WinMx, two file-sharing programs users can install on their computers. Data collection in this context calls for scripting to operate programs' menus and buttons and to capture results from program windows.

Collecting Data from Secondary Sources

For some purposes, researchers may find it preferable to collect data from aggregators that assemble data from multiple underlying sources. For example, Baye, Morgan, and Scholten (2004) examine retailers' prices at a comparison shopping service (a website that lists prices of selected items at multiple retailers) to measure the breadth of price dispersion. Baye, Morgan, and Scholten find substantial price dispersion and little evidence of "the law of one price."

For any researcher needing information about the history of a website—whether on a one-off basis or for large-sample analysis—the Internet Archive is the natural choice. With copies of 150 billion pages dating back to 1996 (Internet Archive 2011), the Internet Archive provides no-charge access to prior versions of most online materials, facilitating all manner of historic analysis. Seamans and Zhu (2010) use the Internet Archive to gather historic data on Craigslist postings to explore relationships between the entry of Craigslist into a market and newspaper circulation and pricing.

While Internet Archive preserves historic materials, certain secondary sources analyze and tabulate *current* user behavior. For example, Google Trends reports the frequency of particular searches at Google. Using Google Trends data, Choi

and Varian (2009) predict future filings for unemployment benefits, finding an improvement over official government forecasts. Wu and Brynjolfsson (2009) use Google searches to predict housing prices and sales, while Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant (2009) use data from Google searches to detect influenza epidemics.

Web 2.0 Data Sources

The term "Web 2.0" denotes the Internet's change from simple screens of unchanging information to interactive communications in which users contribute ever more content—often making sharply more information available to the general public and to researchers. Facebook and Twitter are prominent examples in this vein.

Researchers seeking to study activity at Facebook benefit from default privacy settings that let the general public view each user's name, friends, networks, wall posts, photos, likes, and more. The resulting data can facilitate research on myriad topics. For example, Baker, Mayer, and Puller (2011) use Facebook to assess the diversity effects of randomized dormitory assignment. They find that students randomly exposed to persons of a different race have more friends of that race within the dormitory, but no greater diversity in social networks outside that environment.

Iyengar, Han, and Gupta (2009) look at data from Cyworld, a social networking site in Korea, in which users often decorate their mini-homepages with items like wallpaper or music purchased from Cyworld. Using 10 weeks of purchase and nonpurchase data from 208 users, they identify a low-status group that is not affected by the purchases of others; a medium-status group that has a positive correlation with the purchases of others; and a high-status group that has a negative correlation with the purchases of others. Acquisti and Gross (2006) examine demographic and behavioral differences in users' views of privacy. They find that privacy concerns expressed in survey results did not seem to limit which people joined Facebook nor how much information they revealed—in part, because those who joined did not fully understand what data was public. Default privacy settings at the short-message service Twitter also facilitate research: with few exceptions, Twitter messages are public, and Twitter also publishes the list of all authors each Twitter user is "following." While few researchers have embraced Twitter data, Vincent and Armstrong (2010) assess high-frequency trading strategies grounded in messages on Twitter, finding a profit opportunity in fast-breaking Twitter discussions.

The Internet's newest services also facilitate research on users' views of companies and organizations. Every company and organization page on Facebook includes a "like" button, and Facebook, Google, and others now let sites present "like," "+1," and similar buttons to garner user endorsements. The number and/or identity of users clicking these buttons is often available to researchers and the interested public—facilitating research about trends and trendsetters, reaction to news, and more.

Monitoring Network Activity

The computer science literature features papers that analyze users' network traffic to draw conclusions about online activity. For example, Saroiu, Gribble, and Levy (2004) monitor the online activity of students at a university to identify computers infected with "spyware"—software like Gator, Cydoor, SaveNow, and eZula that gathers information about computer use and relays it to a third party, often without the consent (or informed consent) of the computer user. The authors monitor infections to identify types of users and behaviors particularly likely to suffer from spyware. Karagiannis, Broido, Brownless, claffy, and Faloutsos (2004) observe network activity at two Internet service providers to measure the prevalence of peer-to-peer file-sharing software, finding that contrary to a common belief at that time, peer-to-peer file sharing was not declining in response to concerns over its legality, but was in fact continuing to increase. In principle, economists could use similar methods. As Saroiu, Gribble, and Levy (2004) and Karagiannis et al. (2004) demonstrate, network monitoring systems are notable for the breadth of data they can observe—in principle, every online activity of users on the corresponding networks. They are also notable for their ability to collect data no individual site can, or cares to, assemble.

On a larger scale, commercial services analyze network traffic to identify trends in user behavior and site popularity. Best known is comScore (2011), which offers a two-million member panel of users recruited to install tracking software that monitors their browsing, purchasing, and other online activities. Compete offers a similar service, also via a panel of participating users, while Hitwise collects data from both users and Internet service providers.

What Types of Data are Abundant or Scarce?

Online data collection projects tend to reveal certain data far more readily than others. For example, researchers seeking prices are easily satisfied; myriad sites post product price information as they offer items for purchase. But a researcher needing sales quantities faces greater difficulty: while a few sites, most notably eBay, post substantial information about each sale, most sites have no business need to distribute sales information systematically and publicly.

Work-arounds can yield quantity data. Some sites report the quantity of in-stock inventory available for each listed product. Slow decreases in this quantity are typically interpreted as indicating items sold. Large increases are understood to reflect arrival of additional inventory. By checking sufficiently frequently, a researcher can infer purchases.

Other sites report sales ranks, which offer insight into sales quantities. Best known in this area is Amazon, where many an author diligently tracks sales rank. Chevalier and Goolsbee (2003) pioneered procedures for converting Amazon sales ranks to sales quantities using multiple methods including cross-checks with publishers (who report that a given sales rank in a given week matches a given quantity), controlled experiments (purchasing and/or returning a given quantity of books and monitoring changes in rank), and fitting a portion of the distribution

using publicly known sales quantities (for example, for bestsellers whose sales are occasionally revealed to the public). Depending on available data, these approaches can be used to extract quantities from other ranking systems.

In certain contexts, researchers may be able to infer useful information about sales from other materials that are posted at sites, such as "top sellers," "recommended items," or "people who bought this also bought . . ." suggestions. However, such methods are not yet well developed.

Data Requiring Company or Site Cooperation

While the preceding sections identify data that researchers can collect directly from websites, some research questions require data that sites decline to publish to the public. In this context, it has proven fruitful to request additional data from the companies and organizations that run such sites. For example, Hitsch, Hortaçsu, and Ariely (2010) obtained records from an online dating site revealing all aspects of users' activities, including browsing, viewing photos, and sending and receiving messages. This data is necessary for Hitsch, Hortaçsu, and Ariely's evaluation of the quality of users' matches; they find that at the dating site, "the actual matches are approximately efficient." Data providers are often concerned about distribution of their internal records. Below, I discuss methods to address such concerns and protect privacy.

Online Experiments

Scraping the web for existing data can raise concerns about whether observed changes can be treated as exogenous or endogenous. After all, when users and sites make decisions based on changing external circumstances to advance their respective objectives, it can be difficult to draw inferences about what factor caused what outcome. An experimental method can yield better insight into causation. Online experiments can also observe long-run behavior changes, whereas most short-run experiments risk overemphasizing short-run substitution effects.

In the taxonomy of List (2011) in this journal, online experiments can take four forms: "Natural experiments" arise from exogenous changes created by third parties (or nature) that mimic the conditions of an experiment. "Laboratory experiments" place the agents in an artificial game-like setting to see how they react. "Field experiments" present agents with randomized variation in conditions—changes agents experience while carrying out their usual online activities in their homes or at work, though they are aware that an experiment is being conducted. Finally, "natural field experiments" present agents with randomized variation in natural settings without informing agents that they are involved in an experiment.

Natural Experiments

Sometimes, a site or service changes design parameters arbitrarily or in a time or manner unlikely to be correlated with other outcomes. Such circumstances can create a natural experiment yielding insight into lines of causation.

For example, Roth and Ockenfels (2002) point out that system designers set the ending rules for eBay and Amazon auctions without consideration of implications for bidder behavior. Thus, these sites provide a reasonable context to assess the effect of such rules. At the time of their study, eBay auctions closed at a fixed time, while Amazon auctions continued until ten minutes had gone by without a bid. Bidders reacted to such rules, submitting a higher fraction of late bids on eBay.

Other exogenous variation comes from unexpected changes. For example, Miller (2011) relies on a large increase in the amount of information available about borrowers on Prosper.com, a change which made lenders more selective among high-risk borrowers. Chiou and Tucker (2011) note a 2009–2010 dispute between Google and Associated Press that led to the temporary removal of AP stories from Google News (which aggregates news content from many sources). During that period, Google News referred fewer users to *all* traditional news sites, compared to other news aggregators that continued to host AP articles.

Researcher-Designed Experiments

Some researchers participate in online markets in order to build what are, in essence, online field experiments. For example, Hossain and Morgan (2006) list items on eBay with varying listing prices and shipping prices, showing that users undervalue shipping cost relative to item price. By varying reserve policies in listings at eBay, Katkar and Reiley (2006) find that secret reserve prices deter bidder entry and reduce the likelihood of a listing resulting in a sale. Resnick, Zeckhauser, Swanson, and Lockwood (2006) auction matched pairs of items on eBay, some using a seller's well-established identity and others using new identities, thereby identifying the willingness of buyers to pay for seller reputation.

Many companies have recognized the benefit of experiments in improving their own operations. For example, online marketers test dozens of alternative advertisements. Tools like Optimizely and Visual Website Optimizer let a designer evaluate user behavior in multiple variants of a site—testing alternative layout, color, text, and more. Varian (2010) discusses the benefits Google has achieved through comprehensive experiments to evaluate possible changes to its services.

In lieu of a researcher running experiments, Einav, Kuchler, Levin, and Sundaresan (2011) flag the possibility of a researcher identifying experiments others are already running. If an eBay seller is testing variations in item listing (perhaps which format, description, or pricing achieves the highest price), a researcher can find these variations, retrieve data about both the experimenter's changes and the public's response, and thereby draw conclusions about the effect of the changes at issue. Einav et al. find that of the 100 million listings on eBay each day, more than half will reappear on the site, often with differing parameters for the sale. Assembling a dataset with hundreds of thousands of such matching pairs during a single year, these authors examine questions about price dispersion, bidding under different sets of rules, and customers' reaction to shipping fees. Practitioners' experiments can offer a vastly larger sample than researcher-implemented experiments—in turn, yielding more precise estimates. Practitioners' experiments also often occur across

product categories, whereas practical concerns often limit researcher-implemented experiments to narrow categories, impeding inferences about other areas.

Experiments in Partnership with a Company or Site

Some kinds of online field experiments tend to require cooperation from a company or site operator. For example, Chen, Harper, Konstan, and Li (2010) look at MovieLens, an online site that offers recommendations for movies. Submitting movie recommendations is a public good—benefiting all other users of the site, at some cost to the specific user who makes time to contribute. Many movies had too few recommendations for the software to match them with potential users, but Chen et al. find that when MovieLens subscribers are informed of their standing in terms of how many recommendations they make relative to other users, they tend to contribute more recommendations. Chen et al. partnered with Movielens in order to provide such notifications to a random set of users. Online advertising has proven particularly well suited to experiments with company cooperation. Reiley, Li, and Lewis (2010) change the number of advertisements presented at the top of the page at an Internet search engine, finding that when more such advertisements are presented, users click more often on the top-most advertisement. Ostrovsky and Schwarz (2011) adjust reserve prices in Yahoo! auctions for online advertisements, finding large revenue increases when reserve prices are set optimally.

With additional technical complexity, researchers may be able to conduct experiments entailing modification of a website even without participation by or cooperation from that site. In Edelman and Gilchrist (2010), my coauthor and I build a proxy that presents some users with modified search result pages showing hypothetical alternative advertisement labels: in place of the usual "sponsored link" or "ad" labels, some users instead saw labels reading "paid advertisement." Users with low education or little online experience benefit most from the "paid advertisement" label, which the Federal Trade Commission has sought in other media. Similarly, Schechter, Dhamija, Ozment, and Fischer (2007) present varying security warnings as users attempt to access online banking applications, finding that few users recognize the warnings intended to flag possible attacks. They also flag the importance of realistic experimental conditions: users who participated in the experiments as role players were far less concerned with security than those who used their own actual passwords.

Online Lab Experiments

Online experiments can address many of the questions historically explored in real-world economics laboratories. For example, Horton, Rand, and Zeckhauser (2011) replicate three classic lab experiments in an online lab: the extent of cooperation in a one-shot prisoners' dilemma game; playing the prisoners' dilemma game after being "primed" by reading various religious or nonreligious texts (which tends to reduce rates of defection); and testing the "framing" result of Kahneman and Tversky(1979) that choices will differ depending on how questions are framed, because people are risk averse in the domain of gains but risk seeking as to losses.

Horton, Rand, and Zeckhauser also conducted a natural field experiment in which subjects were offered the opportunity to be paid to transcribe a simple passage of text in return for a compensation that had been randomly determined—demonstrating an upward-sloping supply of labor. Mason and Suri (2011) argue that online lab experiments using Mechanical Turk (which allows hiring people anywhere in the world to carry out tasks that can be performed online) can offer important benefits over traditional laboratory experiments, including easier access to a large and diverse subject pool, low cost, and faster deployment of new experiments.

Online lab experiments also present potential downfalls. As in physical economics laboratories, participants who sign up to participate are unlikely to be representative of the population as a whole, and their differences might be correlated with some treatments. Online subjects can exit a study more readily than subjects in a lab, which may be a concern if certain treatments disproportionately prompt early exit. Communication between online subjects may be possible, both during an experiment and between experiments, depending on a researcher's method of recruiting subjects. To deter communication among subjects, Horton, Rand, and Zeckhauser (2011) suggest running online experiments quickly and avoiding notoriety.

If researchers so choose, online experiments can blur the boundaries between the lab and the field. For example, Centola (2010) builds an online social network where participants can see the health behaviors of selected other participants assigned to be their "health buddies." From one perspective, this appears to be a natural field experiment: users participate from home or work, for an extended period, not knowing that they are subject to randomized variation in an academic research project. Yet participants are interacting in an environment constructed from scratch specifically for research purposes. Just as designers of a lab experiment design the rules of their system, Centola controlled most aspects of what participants could see, say, and do. With the right design, online experiments may be able to combine positive aspects of lab and field experiments.

Limits to Internet-Based Data Collection

Terms of Use and Similar Restrictions

Most websites present a "Terms of Use" or similar document that purports to restrict the methods and purposes of data access. Such statements are widespread; for example, Amazon, eBay, and Google all include provisions stating that users must not copy data from their sites. A series of court cases hold that such agreements are enforceable against competitors seeking to copy data for reasons courts view as improper. For example, in *eBay vs. Bidders' Edge* (100 F.Supp.2d 1058 [N.D. Cal. 2000]), Bidders' Edge sought to copy eBay data to build an auction aggregation service—a service which, if successful, would have undercut eBay's competitive advantage. eBay could therefore offer a cogent notion of harm—not just a few extra requests for its web server to answer, but a genuine business loss.

In contrast, researchers are far less disruptive to the site whose data is being copied. For example, researchers' activities are usually limited to analyzing data but not republishing or redistributing—and certainly not reselling—the information they collect. Furthermore, most researchers access online data that sites produce and distribute incidental to other activity, and researchers' activities do not interfere with sites' core business models. In this context, researchers typically perceive that they have strong defenses to any claims that data providers might bring. Finally, for lack of an urgent business harm, a target site is less likely to press the point.

In practice, sites most often respond to researchers' activities not by filing lawsuits, but by blocking access from computers that send too many requests. Seeing many requests from a single IP address (roughly, a single computer), it is usually straightforward for a site to configure its web server to deny further requests from that computer. Such a blockage often suffices to prompt a researcher to scale back data collection. That said, many researchers nonetheless continue requesting data even after a blockage—for example, using a different computer or a different IP address. Excessive requests could slow access by others and invite further bans, but most researchers' data requirements can be adequately addressed using a data collection system that operates at a rate similar to an ordinary user, sending perhaps one request every few seconds.

To date, to the best of my knowledge, no data provider has filed suit against a researcher collecting data that is available, in smaller quantities, to the general public without charge. Of course a researcher facing notable problems—perhaps accessing data that is otherwise made available only under a paid license—might do well to seek guidance from a qualified attorney.

Privacy and Confidentiality

A researcher collecting online data must also consider privacy concerns. If collecting data about individuals, a researcher should consult the appropriate human subjects committee. That said, many human subjects committees will readily give the researcher permission for an online data collection project to proceed quickly and without conditions. In particular, human subject committees often give blanket permission or even waivers for studies that involve observation of public behavior, analysis of existing data, and gathering information in a way that that presents minimal risk to subjects.

When collecting data from a secure or semisecure site or when analyzing company data or other internal data, researchers may want to shield themselves from user-specific data. Sometimes, researchers need not even receive sensitive information. For example, when Hitsch, Hortaçsu, and Ariely (2010) analyzed user behavior at a dating website, they analyzed user data that contained no names, contact information, or images. But other research requires at least limited analysis of sensitive information or data derived from sensitive information. For example, Ian Larkin and I analyzed working paper downloads at the SSRN (Social Science Research Network) website to investigate whether the numbers were being "gamed" by authors downloading their own papers repeatedly to increase reported download

counts. In doing this analysis, we did not want to see the names of the authors whose papers were most downloaded in circumstances suggesting gaming, but we did seek to analyze relationships between gaming and authors' professional standing, coauthors, and peers. To limit data in this way, we kept author names in a restricted table with limited access rights. When we needed analysis of authors' resumes and biographical information, we provided our research assistants with access to authors' names, but our assistants could not view information about paper downloads, and they did not know the purpose of our study. These procedures prevented anyone, including us, from connecting particular download data to a particular author. We found limited evidence of gaming due to demographic factors and career concerns, but strong evidence of gaming driven by social comparisons with various peer groups (Edelman and Larkin 2009).

When requesting data from companies, additional protections can help address concerns about confidentiality and about the possibility of readers uncovering company identity. It is routine to describe a corporate data source in general terms (like sector and approximate size) but to decline to name the specific company. But creative researchers can do more to protect data details. For example, when analyzing effectiveness of an online advertising campaign performed by Yahoo! and a major retailer, Lewis and Reiley (2011) created a database of over one million customers matched in the databases of the two companies. However, they then hired an outside vendor to render the data anonymous by merging together all of the personally-identifying information about users' online and offline activities. In addition, the vendor multiplied actual sales amounts by an undisclosed number between 0.1 and 10—preventing readers, or even the researchers, from learning the true amount of the company's advertising costs, incremental revenue, or other dollar figures.

Even data not intended to identify individuals may prove easily linked to specific persons. For example, in 2006 AOL's research division posted search data from 650,000 users—a dataset AOL intended to offer for academic research by anyone interested. AOL believed users' privacy was adequately protected because AOL published only users' search requests, not their names, usernames, or e-mail addresses. But some users could be identified from their unusual searches—including searches for their own names, value of their homes, and the like (Barbaro and Zeller 2006). Indeed, even a narrow range of possibilities for users' Social Security numbers can be inferred based on birthplace and date of birth, which are often publicly available (Acquisti and Gross 2009). With re-identification of individuals possible in unexpected circumstances, protecting privacy requires careful planning and ongoing vigilance.

Opportunities

Opportunities for research using the Internet expand every year. New sites and services collect and retain ever more data, while mobile devices collect data

even more widely. Sites and users have been remarkably willing to share much of their data with anyone interested, and the opportunities for economic research are limited primarily by researcher time and creativity. Furthermore, with certain kinds of data increasingly widely available, future researchers may be able to replicate results with similar methods and different datasets rather than using different methods on the same sections.

Meanwhile, in the realm of experiments, online data offers advances on questions of exogeneity and identification. A website design change is often plausibly exogenous, whereas real-world events like government policies are typically correlated with other events. Online systems also make it particularly easy—and, in some contexts, increasingly routine—for different users to receive different treatments on a widespread and ongoing basis. These circumstances combine the identification offered by experiments with the realism of naturally occurring data—potentially giving researchers the best of both methodologies.

■ *I thank Paul Kominers and Xiaoxiao Wu for excellent research assistance.*

References

Acquisti, Alessandro, and Ralph Gross. 2006. "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook." Presented at the Sixth Workshop on Privacy Enhancing Technologies, Cambridge, United Kingdom, June 28 –30.

Acquisti, Alessandro, and Ralph Gross. 2009. "Predicting Social Security Numbers from Public Data." PNAS 106(27): 10975–80.

Baker, Sara, Adalbert Mayer, and Steven L. Puller. 2011. "Do More Diverse Environments Increase the Diversity of Subsequent Interaction? Evidence from Random Dorm Assignment." *Economics Letters* 110(2): 110–112.

Barbaro, Michael, and Tom Zeller, Jr. 2006. "A Face is Exposed for AOL Searcher No. 4417749." *New York Times,* August 9.

Baye, Michael R., John Morgan, and Patrick Scholten. 2004. "Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site." *Journal of Industrial Economics* 52(4): 463–96.

Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, and James R. Marsden. 2006. "Impact of Legal Threats on Online Music Sharing Activity: An Analysis of Music Industry Legal Actions." *Journal of Law and Economics* 49(1): 91–114.

Carlton, Dennis W., and Judith A. Chevalier. 2001. "Free Riding and Sales Strategies for the Internet." *Journal of Industrial Economics* 49(4): 441–61.

Cavallo, Alberto. 2011. "Scraped Data and Sticky Prices." January 27. http://www.mit.edu/~afc/papers/Cavallo-Scraped.pdf.

Centola, Damon. 2010. "The Spread of Behavior in an Online Social Network Experiment." *Science,* September 3, 329(5996): 1194–97.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review* 100(4):1358–98.

Chevalier Judith, and Austan Goolsbee. 2003. "Measuring Prices and Price Competition Online: Amazon.com vs. BarnesandNoble.com." *Quantitative Marketing and Economics* 1(2): 203–222.

Chiou, Lesley, and Catherine Tucker. 2011. "Copyright, Digitization, and Aggregation." NET Institute Working Paper No. 11-18.

Choi, Hyunyoung, and Hal Varian. 2009. "Predicting Initial Claims for Unemployment Benefits." July 5. http://research.google.com/archive/papers/initialclaimsUS.pdf.

comScore. 2011. "Methodology." Information on a webpage. http://www.comscore.com/About_comScore/Methodology.

Daniel Kahneman, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47(2): 263–92.

Edelman, Benjamin. 2002. "The Effect of Editorial Discretion Book Promotion on Sales at Amazon.com." http://www.benedelman.org/publications/thesis-intro.pdf.

Edelman, Benjamin, and Duncan S. Gilchrist. 2010. "'Sponsored Links' or 'Advertisements'?: Measuring Labeling Alternatives in Internet Search Engines." Harvard Business School Working Paper 11-048.

Edelman, Benjamin, and Ian Larkin. 2009. "Demographics, Career Concerns or Social Comparison: Who Games SSRN Download Counts?" Harvard Business School Working Paper 09-096.

Einav, Liran, Theresa Kuchler, Jonathan D. Levin, and Neel Sundaresan. 2011. "Learning from Seller Experiments in Online Markets." NBER Working Paper 17385.

Ellison, Glen, and Sara Fisher Ellison. 2009. "Tax Sensitivity and Home State Preferences in Internet Purchasing." *American Economic Journal: Economic Policy* 1(2): 53–71.

Freedman, Seth, and Ginger Zhe Jin. 2011. "Learning by Doing with Asymmetric Information: Evidence from Prosper.com." NBER Working Paper 16855.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature,* February 19, Volume 457(7232): 1012–14.

Hitsch, Gunther J., Ali Hortaçsu, and Dan Ariely. 2010. "Matching and Sorting in Online Dating." *American Economic Review* 100(1): 130–63.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3): 399–425.

Hossain, Tanjim, and John Morgan. 2006. ". . . Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay." *Advances in Economic Analysis & Policy* 6(2): Article 3.

Internet Archive. 2011. "About the Wayback Machine." Information on a webpage. http://www.archive.org/web/web.php.

Iyengar, Raghuram, Sangman Han, and Sunil Gupta. 2009. "Do Friends Influence Purchases in a Social Network?" HBS Marketing Unit Working Paper 09-123.

Karagiannis, Thomas, Andre Broido, Nevil Brownlee, kc claffy, and Michalis Faloutsos. 2004. "Is P2P Dying or Just Hiding?" Presented at the 47th Global Telecommunications Conference, Globcom 2004, Dallas TX, Nov. 29–Dec. 3.

Katkar, Rama, and David H. Reiley. 2006. "Public versus Secret Reserve Prices in eBay Auctions: Results from a Pokémon Field Experiment." *Advances in Economic Analysis and Policy* 6(2): Article 7.

Lewis, Randall A., and David H. Reiley. 2011. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via Controlled Experiment on Yahoo!" http://www.davidreiley.com/papers/DoesRetailAdvertisingWork.pdf.

List, John A. 2011. "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off." *Journal of Economic Perspectives* 25(3): 3–16.

Mason, Winter, and Siddharth Suri. 2011. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods.* Online publication, June 30. doi: 10.3758/s13428-011-0124-6.

Miller, Sarah. 2011. "Information and Default in Consumer Credit Markets: Evidence from a Natural Experiment." https://netfiles.uiuc.edu/smille36/www/InformationDefault062011.pdf.

Netcraft. 2011. "June 2011 Web Server Survey." http://news.netcraft.com/archives/2011/06/07/june-2011-web-server-survey.html.

Ostrovsky, Michael, and Michael Schwarz. 2011. "Reserve Prices in Internet Advertising Auctions: A Field Experiment." Presented at the 12th ACM Conference on Electronic Commerce, San Jose, California, June.

Reiley, David H., Sai-Ming Li, and Randall A. Lewis. 2010. "Northern Exposure: A Field Experiment Measuring Externalities between Search Advertisements." *Proceedings of the 11th ACM Conference on Electronic Commerce* (EC-2010), pp. 297–304. ACM Digital Library.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. "The Value of Reputation on eBay: A Controlled Experiment." *Experimental Economics* 9 (2): 79–101.

Roth, Alvin E., and Axel Ockenfels. 2002. "Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet." *American Economic Review* 92(4): 1093–1103.

Saroiu, Stefan, Steven Gribble, and Henry Levy. 2004. "Measurement and Analysis of Spyware in a University Environment." Presented at the First Symposium on Networked Systems Design and Implementation (NSDI '04), San Francisco, CA, March 29–31.

Schechter, Stuart, Rachna Dhamija, Andy Ozment, and Ian Fischer. 2007. "The Emperor's New Security Indicators." Presented at the IEEE Symposium on Security and Privacy, Oakland, CA, May.

Seamans, Robert, and Feng Zhu. 2010. "Technology Shocks in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers." NET Institute Working Paper 10-11.

Varian, Hal R. 2010. "Computer Mediated Transactions." *American Economic Review* 100(2): 1–10.

Vincent, Arnaud, and Margaret Armstrong. 2010. "Predicting Break-Points in Trading Strategies with Twitter." SSRN Working Paper 1685150.

Wu, Lynn, and Erik Brynjolfsson. 2009. "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales." Presented at the NBER meeting on Technological Progress & Productivity Measurement, December 4.

This article has been cited by:

1. Manuel Leonard F. Albis, Sabrina O. Romasoc, Shushimita G. Pelayo, Bea Andrea C. Gavira, Jazzen Paul J. Asombrado. 2023. Web scraping for price statistics in the Philippines. *Statistical Journal of the IAOS* **39**:4, 933-945. [Crossref]
2. Joao F. Bigotte, Filipa Ferrao. 2023. The Future Role of Shared E-Scooters in Urban Mobility: Preliminary Findings from Portugal. *Sustainability* **15**:23, 16467. [Crossref]
3. Jasper Tjaden. 2023. Web Scraping for Migration, Mobility, and Migrant Integration Studies: Introduction, Application, and Potential Use Cases. *International Migration Review* **172**. . [Crossref]
4. Dehua Shen, Zezheng Tong, John W. Goodell. 2023. Do online message boards convey cryptocurrency-specific information?. *International Review of Financial Analysis* **75**, 102950. [Crossref]
5. Karen Evelyn Hauge, Andreas Kotsadam, Anine Riege. 2023. Culture and Gender Differences in Willingness to Compete. *The Economic Journal* **133**:654, 2403-2426. [Crossref]
6. Catherine Brown, Sharon Christensen, Andrea Blake, Karlina Indraswari, Clevo Wilson, Kevin Desouza. 2023. Is mandatory seller disclosure of flood risk necessary? A Brisbane, Australia, case study. *Journal of Property, Planning and Environmental Law* **15**:2, 83-105. [Crossref]
7. Jitka Poměnková, Petr Koráb, David Štrba. Text Data Pre-Processing for Time-series Modelling 1-6. [Crossref]
8. Qing Zeng, Jiawei Cao, Yangli Guo, Dayong Dong. 2023. The macroeconomic attention index: Evidence from China. *Finance Research Letters* **52**, 103567. [Crossref]
9. Sebastiano Manzan. Big Data and Computational Social Science for Economic Analysis and Policy 231-242. [Crossref]
10. Johannes Boegershausen, Hannes Datta, Abhishek Borah, Andrew T. Stephen. 2022. Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing* **86**:5, 1-20. [Crossref]
11. Marc Hasselwander, Joao F. Bigotte, Miguel Fonseca. 2022. Understanding platform internationalisation to predict the diffusion of new mobility services. *Research in Transportation Business & Management* **43**, 100765. [Crossref]
12. Bernd Süssmuth. 2022. The mutual predictability of Bitcoin and web search dynamics. *Journal of Forecasting* **41**:3, 435-454. [Crossref]
13. Brian Ratchford, Gonca Soysal, Alejandro Zentner, Dinesh K. Gauri. 2022. Online and offline retailing: What we know and directions for future research. *Journal of Retailing* **98**:1, 152-177. [Crossref]
14. Ti-Ching Peng, Chun-Chieh Wang. 2022. The Application of Machine Learning Approaches on Real-Time Apartment Prices in the Tokyo Metropolitan Area. *Social Science Japan Journal* **25**:1, 3-28. [Crossref]
15. Jens Kolbe, Rainer Schulz, Martin Wersing, Axel Werwatz. 2021. Real estate listings and their usefulness for hedonic regressions. *Empirical Economics* **61**:6, 3239-3269. [Crossref]
16. Konstantinos N. Konstantakis, Despoina Paraskeuopoulou, Panayotis G. Michaelides, Efthymios G. Tsionas. 2021. Bank deposits and Google searches in a crisis economy: Bayesian non-linear evidence for Greece (2009–2015). *International Journal of Finance & Economics* **26**:4, 5408-5424. [Crossref]
17. Yucheng Zhang, Shan Xu, Long Zhang, Mengxi Yang. 2021. Big data and human resource management research: An integrative review and new directions for future research. *Journal of Business Research* **133**, 34-50. [Crossref]
18. John P. Berns, Abu Zafar M. Shahriar, Luisa A. Unda. 2021. Delegated monitoring in crowdfunded microfinance: Evidence from Kiva. *Journal of Corporate Finance* **66**, 101864. [Crossref]

19. Chien-Chiang Lee, Mei-Ping Chen. 2020. Happiness sentiments and the prediction of cross-border country exchange-traded fund returns. *The North American Journal of Economics and Finance* **54**, 101254. [Crossref]
20. Egidio Farina, Colin Green, Duncan McVicar. 2020. Zero Hours Contracts and Their Growth. *British Journal of Industrial Relations* **58**:3, 507-531. [Crossref]
21. Andrés Vallone, Coro Chasco, Beatriz Sánchez. 2020. Strategies to access web-enabled urban spatial data for socioeconomic research using R functions. *Journal of Geographical Systems* **22**:2, 217-239. [Crossref]
22. Judith Hillen. 2019. Web scraping for food price research. *British Food Journal* **121**:12, 3350-3361. [Crossref]
23. Brian T. Ratchford. The Impact of Digital Innovations on Marketing and Consumers 35-61. [Crossref]
24. Motilal Bicchal, S. Raja Sethu Durai. 2019. Rationality of inflation expectations: an interpretation of Google Trends data. *Macroeconomics and Finance in Emerging Market Economies* **12**:3, 229-239. [Crossref]
25. Desamparados Blazquez, Josep Domenech, Jose A. Gil, Ana Pont. 2019. Monitoring e-commerce adoption from online data. *Knowledge and Information Systems* **60**:1, 227-245. [Crossref]
26. Marco Castellani. 2019. Does culture matter for the economic performance of countries? An overview of the literature. *Journal of Policy Modeling* **41**:4, 700-717. [Crossref]
27. Pascal Courty, Sinan Ozel. 2019. The value of online scarcity signals. *Information Economics and Policy* **46**, 23-40. [Crossref]
28. Julian Hackinger. 2018. DataGorri: a tool for automated data collection of tabular web content. *NETNOMICS: Economic Research and Electronic Networking* **19**:1-2, 31-41. [Crossref]
29. Xiao Li, Dehua Shen, Wei Zhang. 2018. Do Chinese internet stock message boards convey firm-specific information?. *Pacific-Basin Finance Journal* **49**, 1-14. [Crossref]
30. Desamparados Blazquez, Josep Domenech. 2018. Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change* **130**, 99-113. [Crossref]
31. Arne Feddersen, Brad R. Humphreys, Brian P. Soebbing. 2018. Sentiment Bias in National Basketball Association Betting. *Journal of Sports Economics* **19**:4, 455-472. [Crossref]
32. Xulia González, Daniel Miles-Touya. 2018. Price dispersion, chain heterogeneity, and search in online grocery markets. *SERIEs* **9**:1, 115-139. [Crossref]
33. Desamparados BLAZQUEZ, Josep DOMENECH. 2018. WEB DATA MINING FOR MONITORING BUSINESS EXPORT ORIENTATION. *Technological and Economic Development of Economy* **24**:2, 406-428. [Crossref]
34. Dehua Shen, Xiao Li, Wei Zhang. 2018. Baidu news information flow and return volatility: Evidence for the Sequential Information Arrival Hypothesis. *Economic Modelling* **69**, 127-133. [Crossref]
35. Dehua Shen, Lanbiao Liu, Yongjie Zhang. 2018. Quantifying the cross-sectional relationship between online sentiment and the skewness of stock returns. *Physica A: Statistical Mechanics and its Applications* **490**, 928-934. [Crossref]
36. Stefan Bechtold. Law and Economics of Copyright and Trademarks on the Internet 7649-7659. [Crossref]
37. Byung–Cheol Kim, Jeongsik "Jay" Lee, Hyunwoo Park. 2017. Two-sided platform competition with multihoming agents: An empirical study on the daily deals market. *Information Economics and Policy* **41**, 36-53. [Crossref]

38. Christoph M. Flath, Sascha Friesike, Marco Wirth, Frédéric Thiesse. 2017. Copy, Transform, Combine: Exploring the Remix as a Form of Innovation. *Journal of Information Technology* **32**:4, 306-325. [Crossref]
39. Francesco D'Amuri, Juri Marcucci. 2017. The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting* **33**:4, 801-816. [Crossref]
40. Xiao Li, Dehua Shen, Mei Xue, Wei Zhang. 2017. Daily happiness and stock returns: The case of Chinese company listed in the United States. *Economic Modelling* **64**, 496-501. [Crossref]
41. Georgios Nalbantis, Tim Pawlowski, Dennis Coates. 2017. The Fans' Perception of Competitive Balance and Its Impact on Willingness-to-Pay for a Single Game. *Journal of Sports Economics* **18**:5, 479-505. [Crossref]
42. Arne Feddersen, Brad R. Humphreys, Brian P. Soebbing. 2017. SENTIMENT BIAS AND ASSET PRICES: EVIDENCE FROM SPORTS BETTING MARKETS AND SOCIAL MEDIA. *Economic Inquiry* **55**:2, 1119-1129. [Crossref]
43. Alberto Cavallo. 2017. Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers. *American Economic Review* **107**:1, 283-303. [Abstract] [View PDF article] [PDF with links]
44. Carlianne Patrick, Amanda Ross, Heather Stephens. Designing Policies to Spur Economic Growth: How Regional Scientists Can Contribute to Future Policy Development and Evaluation 119-133. [Crossref]
45. Yongjie Zhang, Yuzhao Zhang, Dehua Shen, Wei Zhang. 2017. Investor sentiment and stock returns: Evidence from provincial TV audience rating in China. *Physica A: Statistical Mechanics and its Applications* **466**, 288-294. [Crossref]
46. Norma Burow, Miriam Beblo. 2017. Why Do Women Favor Same-Gender Competition? Evidence from a Choice Experiment. *SSRN Electronic Journal* . [Crossref]
47. Xiong Xiong, Jin Zhang, Xi Jin, Xu Feng. 2016. Review on Financial Innovations in Big Data Era. *Journal of Systems Science and Information* **4**:6, 489-504. [Crossref]
48. Luc Anselin, Sarah Williams. 2016. Digital neighborhoods. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* **9**:4, 305-328. [Crossref]
49. Dehua Shen, Wei Zhang, Xiong Xiong, Xiao Li, Yongjie Zhang. 2016. Trading and non-trading period Internet information flow and intraday return volatility. *Physica A: Statistical Mechanics and its Applications* **451**, 519-524. [Crossref]
50. Alberto Cavallo, Roberto Rigobon. 2016. The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives* **30**:2, 151-178. [Abstract] [View PDF article] [PDF with links]
51. Miroslav Beblavý, Lucia Mýtna Kureková, Corina Haita. 2016. The surprisingly exclusive nature of medium- and low-skilled jobs. *Personnel Review* **45**:2, 255-273. [Crossref]
52. Jordana Viotto da Cruz. 2016. Beyond Financing: Crowdfunding as an Informational Mechanism. *SSRN Electronic Journal* . [Crossref]
53. Lucia Mýtna Kureková, Miroslav Beblavý, Anna Thum-Thysen. 2015. Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics* **4**:1. . [Crossref]
54. Derrick M. Anderson, Barry C. Edwards. 2015. Unfulfilled Promise: Laboratory experiments in public management research. *Public Management Review* **17**:10, 1518-1542. [Crossref]
55. Sergio Rey, Luc Anselin, Xun Li, Robert Pahle, Jason Laura, Wenwen Li, Julia Koschinsky. 2015. Open Geospatial Analytics with PySAL. *ISPRS International Journal of Geo-Information* **4**:2, 815-836. [Crossref]

56. Liz Browne, Steve Rayner. 2015. Managing leadership in university reform. *Educational Management Administration & Leadership* **43**:2, 290-307. [Crossref]
57. Xulia Gonzzlez, Daniel Miles. 2015. Price Dispersion and Supermarket Heterogeneity in Spanish Food Retailing. *SSRN Electronic Journal* . [Crossref]
58. Juan Carlos Guataquí R, Jeisson Cárdenas R, Jaime Montaña. 2014. La problemática del análisis laboral de demanda en Colombia. *Perfil de Coyuntura Económica* :24. . [Crossref]
59. Dean Fantazzini. 2014. Nowcasting and Forecasting the Monthly Food Stamps Data in the US Using Online Search Data. *PLoS ONE* **9**:11, e111894. [Crossref]
60. Daniel Arribas-Bel. 2014. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography* **49**, 45-53. [Crossref]
61. Yongjie Zhang, Lina Feng, Xi Jin, Dehua Shen, Xiong Xiong, Wei Zhang. 2014. Internet information arrival and volatility of SME PRICE INDEX. *Physica A: Statistical Mechanics and its Applications* **399**, 70-74. [Crossref]
62. Sascha Friesike, Hendrik Send, Robin P. G. Tech. 2014. What Do Consumers Use 3D Printers For?. *SSRN Electronic Journal* . [Crossref]
63. Stefan Bechtold. Law and Economics of Copyright and Trademarks on the Internet 1-12. [Crossref]
64. Wayne Simpson, J.C. Herbert Emery. 2012. Canadian Economics in Decline: Implications for Canada's Economics Journals. *Canadian Public Policy* **38**:4, 445-470. [Crossref]
65. Byung-Cheol Kim, Jeongsik Lee, Hyunwoo Park. 2012. Two-Sided Platform Competition in the Online Daily Deals Promotion Market. *SSRN Electronic Journal* . [Crossref]