Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing

Xiao Liu, Param Vir Singh, Kannan Srinivasan

To cite this article:
Xiao Liu, Param Vir Singh, Kannan Srinivasan (2016) A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing. Marketing Science 35(3):363-388. https://doi.org/10.1287/mksc.2015.0972

MARKETING SCIENCE
Vol. 35, No. 3, May–June 2016, pp. 363–388
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2015.0972

A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing

Xiao Liu
Stern School of Business, New York University, New York, New York 10012, xliu@stern.nyu.edu

Param Vir Singh
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, psidhu@cmu.edu

Kannan Srinivasan
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, kannans@cmu.edu

Accurate forecasting of sales/consumption is particularly important for marketing because this information can be used to adjust marketing budget allocations and overall marketing strategies. Recently, online social platforms have produced an unparalleled amount of data on consumer behavior. However, two challenges have limited the use of these data in obtaining meaningful business marketing insights. First, the data are typically in an unstructured format, such as texts, images, audio, and video. Second, the sheer volume of the data makes standard analysis procedures computationally unworkable. In this study, we combine methods from cloud computing, machine learning, and text mining to illustrate how online platform content, such as Twitter, can be effectively used for forecasting. We conduct our analysis on a significant volume of nearly two billion Tweets and 400 billion Wikipedia pages. Our main findings emphasize that, by contrast to basic surface-level measures such as the volume of or sentiments in Tweets, the information content of Tweets and their timeliness significantly improve forecasting accuracy. Our method endogenously summarizes the information in Tweets. The advantage of our method is that the classification of the Tweets is based on what is in the Tweets rather than preconceived topics that may not be relevant. We also find that, by contrast to Twitter, other online data (e.g., Google Trends, Wikipedia views, IMDB reviews, and Huffington Post news) are very weak predictors of TV show demand because users tweet about TV shows before, during, and after a TV show, whereas Google searches, Wikipedia views, IMDB reviews, and news posts typically lag behind the show.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2015.0972.

Keywords: big data; cloud computing; text mining; user generated content; Twitter; Google Trends
History: Received: December 31, 2013; accepted: September 30, 2015; Pradeep Chintagunta, Dominique Hanssens, and John Hauser served as the special issue editors and Koen Pauwels served as associate editor for this article. Published online in *Articles in Advance* April 28, 2016.

1. Introduction

The average American watches 5.1 hours of television per day, which is more than the 4.6 self-reported daily hours that are spent on all work-related, educational, and housework activities (Nielsen 2011, Bureau of Labor Statistics (BLS) 2011). This statistic explains why TV is the largest ad spending medium in the United States. Still, the ad spending in the medium remains relatively stable despite the significant growth of online advertising.

Accurate forecasts of television ratings are critical for many reasons. First, depending on the forecasts, networks often adjust the number of new shows for each series. Second, the pricing of advertising for TV shows can be made more dynamic and near real time. Current online advertising on content-based sites is realized in as little as 0.35 seconds based on dynamic real-time bidding models. The spillover to traditional television advertising is beginning to occur. Thus, better forecasts of the size of a viewing audience will enhance the ability to conduct price auctions, and advertisers can decide whether to participate and how much to bid. Third, depending on the projected ratings, firms can adjust many (endogenous) actions, such as advertisements for TV shows, paid blogs, and Tweets, to affect the ratings. Therefore, advertisers and broadcasting companies are eager to accurately predict the TV show's ratings.

A recent article in the *New York Times* (2015) emphasizes the industry significance of the issue studied in our paper. "So far, however, Twitter and Nielsen have avoided the most important question posed by

marketers and the TV industry: Exactly how much does chatter on Twitter lift the viewership of a particular show? Although Nielsen published data on the Twitter activity surrounding a show's broadcast as a complement to its more familiar TV ratings, it has said little about the relationship between the two." The issue that we examine in this paper is of paramount importance to the industry.

Currently, consumers use various online platforms to enhance their TV watching experience: They look for show-related information on search engines such as Google and share their viewing experience with friends on social networks such as Twitter and Facebook. These online platform footprints can help advertisers forecast demand. However, the content on these online platforms produces two challenges. First, the data that are produced are in an unstructured form (for example, texts, video, audio, and images). Second, the sheer volume of data makes standard data analysis procedures computationally unworkable or inefficient. Existing literature has attempted to incorporate user-generated content (UGC) in its analyses by including easy-to-calculate measures such as the volume or valence of relevant UGC on online platforms. Recent studies have attempted to examine content information (e.g., Gopinath et al. 2014, Pauwels et al. 2013). These studies depend primarily on manual coders to classify the UGC, which limits the scale of the application, or they follow a supervised learning approach to classify the UGC. However, all of these studies are limited in the extent to which the UGC is typically classified into preconceived labels (such as sentiment, recommendation-oriented or emotion). We show that the information in the textual content can provide greater insight into user behavior and decision making. Moreover, an unsupervised learning approach can provide significant improvement in forecasting performance than measures such as volume and sentiment. We combine the methods from machine learning and text mining to more completely examine the textual content, and we endogenously identify distinct streams of information in the UGC. We use the tools and methods from cloud computing, such as Hadoop MapReduce, to manage millions of text documents. In this way, we confront a problem that is germane to text mining. We find that the number of distinct information streams identified is typically greater by an order of magnitude than the number of observations on the variable of interest. To address this concern, we conduct a massive dimension reduction using the Apache Mahout machine learning library to summarize the significant amount of information into several principal components. We find that these principle components have excellent predictive power in demand forecasting. None of the tasks described above are trivial when the volume of data is significant. The memory space and computing capacity of a single workstation cannot manage data of our scale. Instead, we use Amazon Web Services to perform the cloud computing tasks and pay with only a minimal budget.

We use the unstructured data of consumer behavior on online platforms, including Twitter, Google Trends, Wikipedia, the Internet Movie Database (IMDB), and Huffington Post (TGWIH) to predict consumers' offline TV viewing behavior for 30 TV series and prime time National Football League (NFL) games. We argue that consumers reveal their interests in TV programs through online platforms before actually watching TV. For example, a Twitter user's post, "I am going to watch Breaking Bad tonight," is a direct indication of her future TV watching intent for a specific show. If a user searches Google for an NFL game before the game begins, it is likely that he is going to watch it on TV. Therefore, by collecting publicly available data from TGWIH at a negligible cost, marketers and advertisers can leverage consumer-generated data to accurately forecast future demand rather than relying on the historical information from the Nielsen Rating data.

To achieve our goal, we use a large data set derived from the following five sources of online platforms: (1) Twitter, 1.8 *billion* Tweets for five years from 2008 to 2013; (2) Google Trends,[1] 113.3 *million* Google searches[2] (when combined with the Google AdWords keyword volume service, we can obtain the real search volume); (3) Wikipedia views, 433.6 *billion* Wikipedia page views; (4) IMDB reviews, 4.3 thousand reviews; and (5) Huffington Post news, 5.5 *million* articles. We find that the predictive power of the surface-level measures of the UGC, such as the volume of Google searches (or Wikipedia views) or the volume and valence of Tweets (or IMDB reviews and Huffington Post news), is not as strong as the historical data for forecasting TV ratings. However, the refined information in Tweets exhibits a stronger power to predict TV ratings than historical ratings.

We conduct a rigorous, structured econometrics analysis of the processed unstructured data. Our results show that Tweets, Google searches, Wikipedia views, IMDB reviews, and Huffington Post News volume have a positive impact on TV ratings. The impact of the valance of Tweets on ratings is not statistically significant. Carefully summarized Tweet content that indicates future action has the highest predictive power.

Our paper makes two key contributions. First, from a managerial standpoint, we show that easily accessible

[1] Google Trends is a public Web facility of *Google, Inc.* that is based on *Google Search*. It shows how often a particular search term is entered relative to the total search volume across various regions of the world and in various languages.

[2] Google Trends and Wikipedia views data are structured in a numerical format. The other three sources of data are in a text format.

public information from online platforms can be used to predict TV ratings. Particularly, surface-level information such as volume and valance is not more useful than the historical data; only a sophisticated content analysis can achieve high prediction accuracy. Our proposed method has a distinct advantage because it does not require classification of content into pre-conceived topics that may not be relevant. Instead, our method endogenously summarizes the information in Tweets into topics. That is, our approach classifies the relevant Twitter content into distinct streams of information that consumers are discussing. Second, we introduce state-of-the-art big data processing techniques through a cloud-based distributed computing framework called Hadoop MapReduce, which we demonstrate with Amazon Web Service tools. We hope marketing researchers can use these methods to conduct more structured research on large scale unstructured data.[3]

2. Literature Review

Our paper draws on three lines of literature, i.e., (1) using online platform data for predictions, (2) the effect of online UGC on product sales, and (3) text mining.

2.1. Online Platform Predictions

Research has suggested that Twitter feeds are early indicators of various economic and social phenomena, such as book sales (Gruhl et al. 2005), movie box office sales (Mishne and Glance 2006), opinion polls (O'Connor et al. 2010), elections (Tumasjan et al. 2010), the spread of contagious diseases (Paul and Dredze 2011), stock market performance (Bollen et al. 2011), and NFL game outcomes (Sinha et al. 2013). The predictive power of Twitter is derived from the information that is embedded in consumers' conversations. We follow this literature to predict the ratings of popular TV shows in the U.S. market. The literature concerning the demand prediction of new product introduction has studied products, including movies, books, cellphones, etc. By comparison, our focal product, TV shows, is unique because every week, a new "mini" product (i.e., episode) is released.[4] This continuity makes TV show demand relatively easier to predict than the demand for other products. Beyond Twitter, research has also investigated other online platform data for predictions, such as Google searches (Preis et al. 2010) and Wikipedia views (Mestyán et al. 2013). These studies typically capture information in Tweets by volume, valence or emotion. Furthermore, most papers in this area use machine learning methods with the objective of merely minimizing the prediction error. We instead use an econometrics model that corrects for the Nickell (1981) bias to perform a more structured analysis (about providing economic explanations) of the unstructured data from multiple online platform sources.

2.2. The Effect of Online UGC on Sales

Our paper is also closely related to the literature on the effect of UGC on demand/sales. As listed in Table 1, numerous studies have examined this topic in marketing and computer science for explanatory or prediction purposes. In these papers, researchers have used various forms of UGC, including online reviews, online ratings, and blogs to investigate their impact on the demand for the focal products. We instead use more popular online platforms, including TGWIH, to collect UGC. The advantage of our approach is that these platforms have a much wider user base; therefore, the predicted demand from using information from these platforms is more likely to represent emerging big data information sources.

Additionally, as Table 1 demonstrates, metrics such as the volume, valence, and variance of UGC have been examined. However, the rich content information in text data has been underexploited. In fact, to our knowledge only three of the prior studies (Onishi and Manchanda 2012, Gopinath et al. 2014, and Pauwels et al. 2013) have tried to perform text mining beyond basic sentiment analysis. Our paper extends this line of the literature with two major distinctions. First, we incorporate cloud-based, large scale text mining to extract useful information from a vast amount of data whose size is larger than the data in the previous literature by a magnitude of 1,000. Second, we exploit unsupervised learning techniques to allow the data to determine the issues rather than imposing any label on the features. (For example, Pauwels et al. 2013 selected conversations related to "went there/purchased," and Gopinath et al. 2014 classified online conversations as related to attributes/emotions/recommendations). Instead, we mine the data with an unsupervised learning approach and adopt a dimensionality reduction method (specifically, principal component analysis). By studying the loading of specific content, we can interpret the key principal components. Thus, our approach is consistent with the traditional marketing approach wherein the dimensionality reduction is first undertaken (such as factor analysis and principal component analysis), and then, the factors are interpreted.

[3] Although big data processing techniques are only now entering business academic research, these tools are already being widely used by companies such as Facebook, Yahoo!, and Netflix. Although the industry uses a large volume of data for predictions, regrettably the underlying models are rarely revealed. Academics may focus on sophisticated models while industry may rely on simple but effective methods. Practitioners may also value the scalability of the models more than academic researchers in the business arena.

[4] We thank one of the reviewers for identifying this feature.

Table 1 Overview of Literature on UGC

Author	Year	Product	UGC	Measure	Effect	Outcome measures	Text data	Data size	Text mining tools
Godes and Mayzlin	2004	TV shows	Online review	Volume	∼	Household rating	Yes	21,604	Independent raters
				Valence	∼				
				Variance	+				
Chevalier and Mayzlin	2006	Books	Online rating	Volume	+	Sales rank	No		
				Valence	+				
Liu	2006	Movies	Online review	Volume	+	Box office revenue	Yes	12,136	Independent raters
				Valence	∼				
Mishne and Glance	2006	Movies	Weblog	Volume	+	Sales	Yes	Unknown	Keyword detection
				Valence (sentiment)	+				
Liu et al.	2007	Movies	Weblog	Sentiment		Box office revenue	Yes	45,046	Machine learning
Dhar and Chang	2009	Music	Online review, blog, SNS intensity	Volume	+	Sales rank	No		
				Rating	+				
				Social network intensity	∼				
Sadikov et al.	2009	Movies	Blog	Volume	+	Box office revenue	Yes	Unknown	Machine learning
				Sentiment	∼				
Chakravarty et al.	2010	Movies	Online review	Volume		Box office revenue	Yes		Independent raters
Chintagunta et al.	2010	Movies	Online rating	Valence	+	Opening day revenue	No		
Chen et al.	2011	Digital cameras	Online rating	Valence	+	Sales rank	No		
Karniouchina	2011	Movies	Yahoo! movie site	Internet searches/ review count	+	Box office revenue	No		
Moe and Trusov	2011	Multiple products	Online rating	Valence	+	Ratings and sales	No		
Onishi and Manchanda	2012	Movies, cell phone subscriptions	Blog	Volume	+	Sales volume	Yes	200,000	Wordcount + selected words
				Valence	+				
				(text mining)	+				
Stephen and Galak	2012	Loans	Press and blog	Volume	+	Sales	No		
Dewan and Ramaprasad	2012	Music	Blog	Volume	+	Sampling	No		
Tirunillai and Tellis	2012	Multiple products	Product reviews	Volume	+	Stock prices	Yes	347,628	Machine learning
				Valence	+				
Gopinath et al.	2013	Movies	Blogs	Volume	+	Box office revenue	Yes	Unknown	Independent raters
				Valence	+				
Pauwels et al.	2013	Clothing retailer	Blog, forum, Facebook, and Twitter	Volume	+	Store/Web traffic	Yes	428,450	Machine learning
				Valence	+				
				Content	+				
Gopinath et al.	2014	Cellular phones	Forum	Volume	∼	Sales	Yes	Unknown	Independent coders
				Valence	+				
				Content	+				
This paper	2016	TV shows	Twitter, Google, Wikipedia, IMDB reviews, news	Volume	∼	TV ratings	Yes	6,894,624	Cloud-based machine learning
				Sentiment	∼				
				Content	+				

Table 2 TV Series, Number of Episodes and Rank (2008–2012)

			2008		2009		2010		2011		2012	
	Show	Channel	Ep's	Rank	Ep's	Rank	Ep's	Rank	Ep's	Rank	Ep's	Rank
1	2 Broke Girls	CBS							24	32	24	32
2	30 Rock	NBC	22	69	22	86	23	106	22	130	13	99
3	90210	CW	24	172	22	137	22	133	24	145	22	147
4	Allen Gregory	Fox							7			
5	Blue Bloods	CBS					22	19	22	22	23	14
6	Body of Proof	ABC					9	13	20	44	13	34
7	Breaking Bad	AMC	7	86	13	51	13	39	13	22	16	1
8	Charlie's Angels	ABC							8			
9	Cougar Town	ABC			24	57	22	67	15	107	15	
10	Criminal Minds	CBS	26	11	23	16	24	10	24	15	24	20
11	Desperate Housewives	ABC	24	9	23	20	23	26	23	37		
12	Gary Unmarried	CBS	20	74	17	72						
13	Glee	FOX			22	33	22	43	22	56	22	50
14	Gossip Girl	CW	25	168	22	135	22	139	24	188	10	140
15	Grey's Anatomy	ABC	24	2	24	12	22	9	24	12	24	10
16	Harry's Law	NBC							12	28	22	52
17	Hellcats	CW					22					
18	How I Met Your Mother	CBS	24	49	24	42	24	48	24	45	24	42
19	Lie to Me	FOX	13	29	22	57	13	78				
20	Mike and Molly	CBS					24	35	23	31	23	37
21	NCIS	CBS	25	5	24	4	24	5	24	3	24	1
22	Nikita	CW					22	135	23	182	22	145
23	Parks and Recreation	NBC	6	96	24	108	16	116	22	134	22	111
24	Private Practice	ABC			22	10	23	37	22	48	22	49
25	Rules of Engagement	CBS	13	23	13	50	24	49	15	42	13	52
26	Shark Tank	ABC			14	102	9	113	15	98	26	63
27	Smallville	CW	22	152	21	129	22	131				
28	The Big Bang Theory	CBS	23	44	23	12	24	15	24	8	24	3
29	The Vampire Diaries	CW			22	118	22	193	22	166	23	133
30	Two and a Half Men	CBS	24	10	22	11	16	17	24	11	23	11

Data source. http://tvbythenumbers.zap2it.com.

2.3. Mining Unstructured Text Data

As noted by Archak et al. (2011), textual information embedded in UGC has largely been ignored in business research because of a lack of practical tools to analyze this unstructured data. Netzer et al. (2012) also noted that the overwhelmingly large volume of data has made analysis extremely difficult if not impossible. Recently, with the aid of automated machine learning techniques, research has emerged to exploit the content of text data rather than its simple numeric volume to provide richer insights. These studies include Das and Chen (2007), Ghose et al. (2007), Eliashberg et al. (2007), Decker and Trusov (2010), Ghose and Ipeirotis (2011), Lee and Bradlow (2011), Ghose et al. (2012), and Lee et al. (2013). More applications that use text mining can be found in areas other than marketing, such as computer science (see Pang and Lee 2008 for a review).

Our paper goes beyond text mining by combining it with cloud-computing techniques to quickly and cost efficiently analyze big text data. To our knowledge, the amount of Twitter feed data that we process is much larger than the scale of any previous papers. See §§3 and 4 for details.

3. Data Description

Our primary goal is to use online platform data to predict TV ratings. Below, we explain how we collect the data for TV ratings and five sources of online platforms, including TGWIH.

3.1. TV Series

We study a collection of 30 U.S. TV series during the 2008 to 2012 TV seasons. Table 2 shows the number of episodes and each show's ranking in terms of total viewership over the five TV seasons.[5] Among these TV

[5] Among the shows, some (e.g., Breaking Bad) are cable shows (AMC is a cable network that generates revenue from user subscription fees), whereas other shows (e.g., The Big Bang Theory) are network broadcast shows (CBS is a broadcasting network whose revenue mainly comes from advertisements). Because cable shows generally have fewer viewers, the ranks of cable shows and network broadcast shows are not directly comparable.

Table 3 NFL Primetime Games (2010–2012)

	2010	2011	2012
Sunday Night Football	18	18	19
Thursday Night Football	8	8	13
Monday Night Football	17	17	17

Data source. http://www.nfl.com/schedules.

shows, some are very popular, including The Big Bang Theory, Breaking Bad, Grey's Anatomy, NCIS, and Two and a Half Men. We chose these shows first because advertisers are eager to know their ratings because these shows are costly on a cost per thousand views (CPM) basis (for example, The Big Bang Theory commanded a staggering $326,260 per 30-second spot on CBS[6] in 2013, which ranked behind only Sunday Night Football (SNF)). Second, we chose these shows because their popularity may generate significant talk on online platforms, such as Twitter and Google searches. Other less well known shows are also included, such as Allen Gregory, Charlie's Angels, Hellcats, Harry's Law, and Gary Unmarried. These shows did not last more than two seasons. We include them because their ratings vary dramatically, and they are difficult to predict. We also examine other shows that are neither popular nor unpopular to demonstrate the generalizability of our findings. We focus on shows whose titles are unique enough[7] not to be confused with other common words that may appear in Tweets (e.g., if we search in Tweets for another popular show called Community, many non-related Tweets with the generic word "community" may appear).

3.2. NFL

Football is the most popular sport in the United States, and football games are among the most watched TV programs. We focus only on the regular season of professional football, the National Football League (NFL),[8] prime time games (8:30 P.M.), i.e., SNF on NBC, Thursday Night Football on the NFL Network, and Monday Night Football on ESPN. Data are collected for three regular seasons from 2010 to 2012 for a total of 135 games. As shown in Table 3, there are more games on Sundays than on Mondays or Thursdays.

[6] http://www.adweek.com/news/television/big-bang-theory-gets-highest-ad-rates-outside-nfl-153087.

[7] We also performed an analysis on six shows with common words as titles. We applied machine learning techniques to classify the relevant tweets rather than the method described in §3.4.2. Our results show that the Tweet Content Model also performs significantly better than other models for these shows. We thank the anonymous reviewer for suggesting this generalization test.

[8] The preseason and postseason games' advertising slots are normally not on the scatter market.

3.3. A.C. Nielsen Ratings

We search the A.C. Nielsen Ratings data for the 18–49 year old age range (http://tvbythenumbers.zap2it.com) for each episode of the 30 TV series[9] and each prime time NFL game. By definition, rating means the percentage of all television-equipped households that were tuned in to that program at any given moment.

Figure 1(a) depicts the ratings of the five most popular TV series (Breaking Bad, The Big Bang Theory, Grey's Anatomy, NCIS, and Two and a Half Men, in Figure 1(a)) and five unpopular TV series (Allen Gregory, Charlie's Angels, Gary Unmarried, Harry's Law, and Hellcats, in Figure 1(b)) for the five seasons from 2008 to 2012. Each show has a unique trend throughout the five years. Breaking Bad experienced an upward trend over time and reached a dramatic spike for the final episode of the final season. The Big Bang Theory gradually improved its ratings every year. Grey's Anatomy's ratings decreased. NCIS' ratings remained stable, and Two and a Half Men lost popularity soon after the main character (Charlie Sheen) was replaced in 2011. All of the unpopular shows have a steep, downward-sloping ratings trend.

For the NFL games (in Figure 2), SNF is the most watched, whereas Thursday Night Football is the least watched. Despite the trend across years, when we focus on each season, the ratings time series are relatively stationary.

Based on Figures 1(a) and 2, we know that each TV series and each game day of the week (Sunday, Monday, and Thursday) has a distinct time-series pattern and can be treated as one member of the panel data.

Next, we describe the data derived from five online platform sites, i.e., TGWIH.

3.4. Twitter

Founded in 2006, Twitter is a real-time information service where people around the world can post ideas, comments, news, photos, and videos in 140 characters or fewer. In 2012, there were more than 500 million registered users who posted 340 million Tweets per day. Twitter ranks as the second[10] most visited social

[9] Unfortunately, this website did not collect any ratings data for Breaking Bad for the 2008 and 2009 seasons. Therefore, we use only the 2010–2012 ratings data for that program.

[10] Unfortunately, we could not access Facebook data. Moreover, although Facebook has a wider user base, most people make their Facebook status updates and content viewable only to their friends, which leaves only a small percentage of public Facebook updates involving TV shows that are available to researchers. This restriction significantly constrains the number of Facebook posts that we can analyze. Moreover, Twitter is known as "the place that hosts a real-time, public conversation about TV at scale" (http://www.mediaweek.co.uk/article/1288398/twitter-buys-secondsync-mesagraph). Therefore, we do not use Facebook data in this paper.

Figure 1(a) (Color online) Nielsen Ratings—Five Popular TV Series

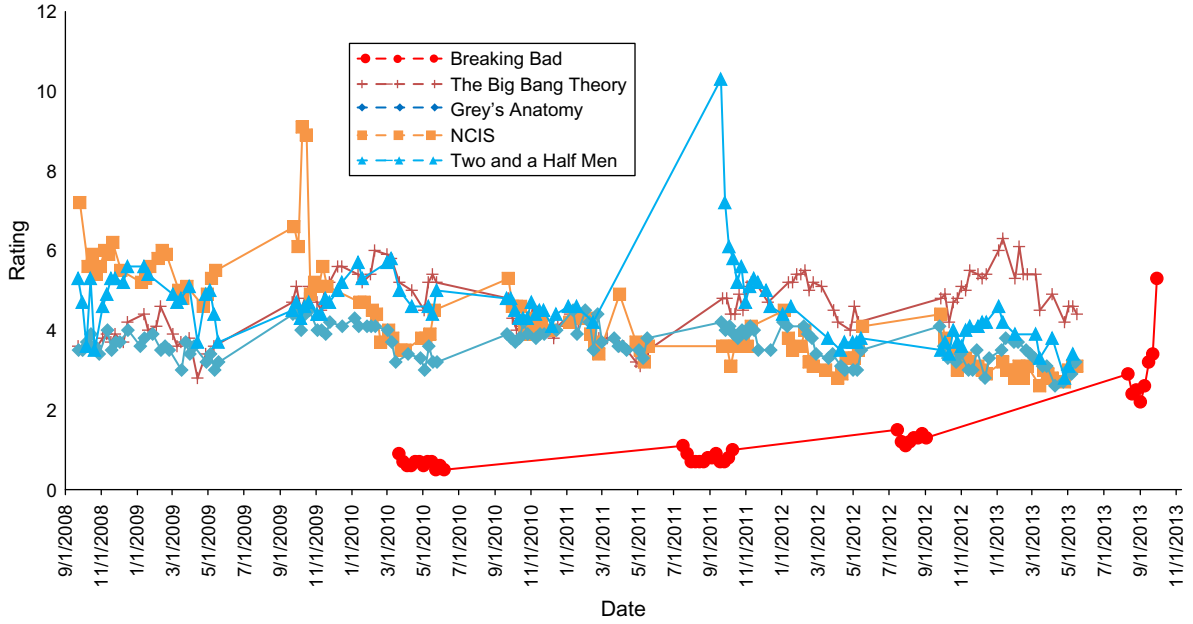
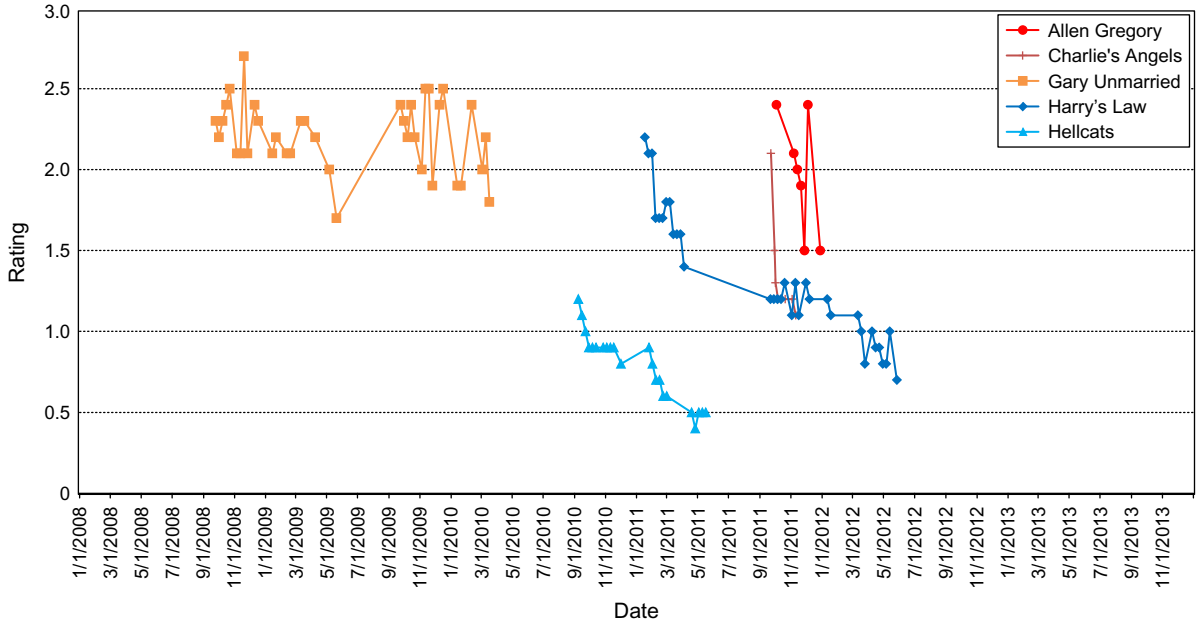


Figure 1(b) (Color online) Nielsen Ratings—Five Unpopular TV Series



networking website according to the Global Alexa Page ranking.[11]

3.4.1. Data Gathering. The data we use are collected from Twitter (www.twitter.com) using the "garden hose" (10%) stream[12] on a daily basis from September 1, 2008 to October 27, 2013.[13] Table 4 shows the size of the data set per month as measured by the number of Tweets (in millions) and the text file storage size.

From this large number of Tweets, we first select the relevant Tweets that discuss the 30 TV series and NFL prime time games.

3.4.2. Selecting Relevant Tweets. We use the following four types of identifiers to search for the relevant

[11] http://www.alexa.com/topsites/category/Computers/Internet/On_the_Web/Online_Communities/Social_Networking.

[12] http://blog.gnip.com/tag/gardenhose/.

[13] We thank Brendan O'Connor and Professor Noah Smith from Carnegie Mellon University for providing us with the data.

Figure 2 (Color online) Nielsen Ratings—NFL

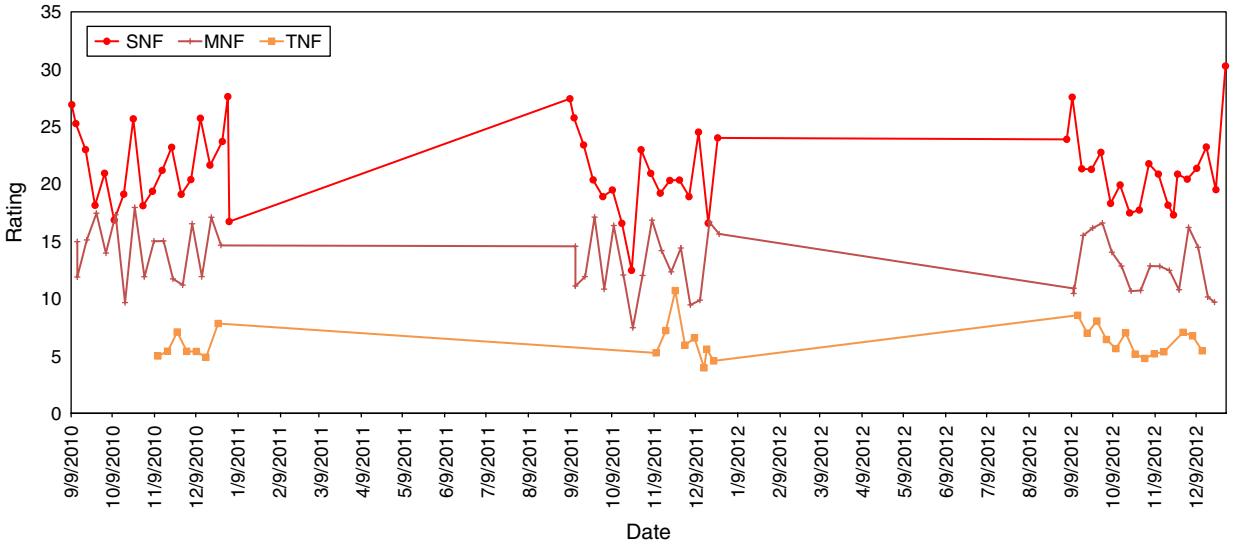


Table 4 Number of Tweets and File Size by Month (2008–2013)

	2008		2009		2010		2011		2012		2013	
	Tweets (Mil)	Size (GB)	Tweets (Mil)	Size (GB)	Tweets (Mil)	Size (GB)	Tweets (Mil)	Size (GB)	Tweets (Mil)	Size (GB)	Tweets (Mil)	Size (GB)
1			22.17	3.67	121.02	32.56	292.99	114.34	762.28	316.53	1,282.95	577.33
2			20.65	3.34	126.70	35.58	300.93	117.83	768.13	319.75	1,170.41	524.69
3			20.89	3.30	163.07	47.48	369.83	147.29	860.95	361.32	1,286.31	600.46
4			23.19	3.39	181.11	54.03	386.54	151.37	893.34	380.43	1,267.80	579.93
5			9.21	1.76	210.18	63.37	429.47	167.58	949.08	407.89	1,342.71	615.73
6			14.26	3.41	244.58	74.56	450.20	177.58	1,004.92	434.91	1,321.88	607.61
7			28.17	6.85	168.61	52.39	508.13	207.34	1,095.50	478.38	1,409.66	650.34
8			50.96	12.61	97.81	31.74	537.77	224.20	1,120.58	493.33	1,418.14	654.46
9	9.45	1.49	65.55	16.34	198.62	69.23	524.76	219.77	1,050.63	464.29	1,287.45	592.95
10	16.17	2.55	73.56	18.49	216.35	80.60	570.28	241.26	1,111.40	506.26	1,299.44	601.49
11	21.06	3.40	82.91	20.99	225.63	88.90	612.91	256.41	1,131.26	519.72		
12	20.65	3.38	93.37	24.27	263.78	101.99	685.82	284.17	1,220.59	555.67		

Tweets for the five TV series: (1) name of the show (e.g., Breaking Bad);[14] (2) official Twitter account of the show (e.g., @TwoHalfMen_CBS); (3) a list of hashtags associated with the show (e.g., #AskGreys); and (4) the characters' names on the show (e.g., Sheldon Cooper). As to the NFL, we use similar identifiers, including (1) a list of the hashtags of the 32 teams[15] (e.g., #gosteelers) and (2) the hashtags of the game (e.g., #SNF). We use Hadoop MapReduce to efficiently select all of the relevant Tweets. See §4 for technical details.

Tables 5 and 6 show some summary statistics of the Tweets for the five TV series and the NFL games.

3.4.3. TV Series. The number of Tweets on Twitter still varies greatly. If we compare the number of Tweets per episode, Breaking Bad created the most buzz, with more than 57,000 tweets, whereas the least popular show, Gary Unmarried, had only 45.

For all 30 TV series, the number of Tweets peaks during the show; consumers Tweet more frequently before the show than afterward. These results reflect that Twitter is a social platform that creates real-time engagement for viewers. It is no surprise that Nielsen is teaming up with Twitter to establish social TV ratings.[16]

3.4.4. NFL. Similarly, the number of Tweets is more intensive during the NFL games than before or after the games (Table 6).

For the 32 NFL teams (Table 7), the most Tweeted team on Twitter was the New York Jets, which had five times more Tweets on average than the least Tweeted team, the Jacksonville Jaguars. The difference in Tweet frequency is largely because of the size of the fan base.

3.5. Google Trends

We also collected data from Google Trends (http://www.Google.com/trends/). Google Trends provides the

[14] We also include some variations of the name, such as Breaking_Bad and BreakingBad.

[15] Hashtags include the names of the teams.

[16] http://www.nielsen.com/us/en/press-room/2013/nielsen-launches-nielsen-twitter-tv-ratings.html.

Table 5 Tweet Frequency, Pre-During-Post

Show	Frequency				Frequency/Episode				Frequency/Hour			
	Pre[a]	During	Post	Total	Pre	During	Post	Total	Pre	During	Post	Total
2 Broke Girls	7,258	3,071	30,861	41,189	100.8	42.6	428.6	572.1	4.2	85.3	3.0	3.4
30 Rock	114,258	31,058	423,803	569,119	1,077.9	293.0	3,998.1	5,369.0	44.9	586.0	27.9	32.0
90210	136,434	38,605	699,171	874,210	1,196.8	338.6	6,133.1	7,668.5	49.9	338.6	42.9	45.6
Allen Gregory	796	1,279	6,106	8,181	113.8	182.7	872.2	1,168.7	4.7	365.4	6.1	7.0
Blue Bloods	21,344	15,445	132,811	169,600	239.8	173.5	1,492.3	1,905.6	10.0	173.5	10.4	11.3
Body of Proof	6,741	4,826	21,066	32,633	160.5	114.9	501.6	777.0	6.7	114.9	3.5	4.6
Breaking Bad	525,234	687,585	1,199,180	2,412,016	12,505.5	16,371.3	28,552.0	57,428.8	521.1	16,371.3	199.7	341.8
Charlie's Angels	2,365	1,014	11,994	15,372	337.9	144.8	1,713.4	2,196.1	14.1	144.8	12.0	13.1
Cougar Town	31,068	7,411	99,986	138,465	349.1	83.3	1,123.4	1,555.8	14.5	166.5	7.8	9.3
Criminal Minds	97,349	56,446	809,187	962,982	804.5	466.5	6,687.5	7,958.5	33.5	466.5	46.8	47.4
Desperate Housewives	58,109	29,285	239,505	326,899	575.3	289.9	2,371.3	3,236.6	24.0	289.9	16.6	19.3
Gary Unmarried	532	67	1,068	1,667	14.4	1.8	28.9	45.0	0.6	3.6	0.2	0.3
Glee	356,269	109,596	1,708,721	2,174,587	3,298.8	1,014.8	15,821.5	20,135.1	137.4	1,014.8	110.6	119.9
Gossip Girl	220,750	70,024	982,357	1,273,131	2,006.8	636.6	8,930.5	11,573.9	83.6	636.6	62.5	68.9
Grey's Anatomy	482,976	870,526	1,067,363	2,420,866	4,093.0	7,377.3	9,045.4	20,515.8	170.5	7,377.3	63.3	122.1
Harry's Law	4,493	2,841	11,994	19,328	132.2	83.6	352.8	568.5	5.5	83.6	2.5	3.4
Hellcats	10,994	4,985	40,707	56,686	499.7	226.6	1,850.3	2,576.6	20.8	226.6	12.9	15.3
How I Met Your Mother	101,727	11,224	580,399	693,350	847.7	93.5	4,836.7	5,777.9	35.3	187.1	33.7	34.4
Lie to Me	2,994	680	14,934	18,608	62.4	14.2	311.1	387.7	2.6	14.2	2.2	2.3
Mike and Molly	1,589	795	4,070	6,454	17.3	8.6	44.2	70.1	0.7	17.3	0.3	0.4
NCIS	223,767	311,939	452,242	987,933	1,849.3	2,577.9	3,737.5	8,164.8	77.1	2,577.9	26.1	48.6
Nikita	49,174	8,495	588,896	646,564	673.6	116.4	8,067.1	8,857.0	28.1	116.4	56.4	52.7
Parks and Recreation	30,637	10,666	97,953	139,256	278.5	97.0	890.5	1,266.0	11.6	193.9	6.2	7.5
Private Practice	41,628	27,003	133,206	201,836	408.1	264.7	1,305.9	1,978.8	17.0	264.7	9.1	11.8
Rules of Engagement	4,457	1,235	55,764	61,456	47.9	13.3	599.6	660.8	2.0	26.6	4.2	3.9
Shark Tank	36,527	14,428	114,773	165,728	392.8	155.1	1,234.1	1,782.0	16.4	155.1	8.6	10.6
Smallville	23,383	7,551	108,993	139,927	365.4	118.0	1,703.0	2,186.4	15.2	118.0	11.9	13.0
The Big Bang Theory	556,636	327,452	1,607,036	2,491,124	4,717.3	2,774.9	13,618.9	21,111.2	196.6	5,550.0	94.9	125.7
The Vampire Diaries	148,009	42,410	562,743	753,162	1,333.4	382.1	5,069.8	6,785.2	55.6	382.1	35.5	40.4
Two and a Half Men	156,894	99,753	471,712	728,359	1,439.4	915.2	4,327.6	6,682.2	59.9	1,830.3	30.2	39.8
Total	3,454,392	2,797,695	12,278,601	18,530,688	39,939.9	35,372.7	135,648.9	210,961.7	1,664.1	39,878.8	948	1,255.8

[a]"Pre" includes all Tweets 24 hours before the show starts. "During" includes Tweets only during the show. "Post" includes Tweets between the end of one episode and the start of the next.

total search volume for a particular search item. The results can be further customized to a certain time range, location, and language. For the TV series data, we use the name of the show (e.g., Two and a Half Men) and character names on the show (e.g., Walden Schmidt) as the keywords. For the NFL data, we use the name of the football team (e.g., Pittsburgh Steelers) as the keyword.

In Figure 3, we present the results for the search item "the big bang theory" for the three months from September to November 2013 in the United States. When a search for a term on Google Trends is performed, a graph similar to Figure 3 is shown. The numbers on the graph reflect how many searches have been conducted relative to the total number of searches that are conducted on Google over time. The numbers on the graph do not represent absolute search volume numbers because the data are normalized and presented on a scale from 0–100. Each point on the graph is divided by the highest point, or 100. For example, on September 21, 2013, the number 35 is relative to the highest point, 100, on September 27.

Table 6 Tweet Frequency—NFL

Time	Frequency	Freq/Game	Freq/Hour
Pre game	1,520,044	3,015.96	125.67
During game	2,532,638	5,045.10	1,261.27
Post game	5,402,517	10,783.47	77.02
Total	9,455,200	18,760.32	111.67

In addition, we use the Google AdWords Keyword planner volume search service[17] to obtain the absolute search volume or, more specifically, a 12-month average of the number of searches for the exact keyword based on the selected location and the Search Network target settings. Combining these data with the Google Trend relative numbers (by multiplying the relative number by a ratio to transform it to an absolute number), we obtain the absolute search volume for each day.

We record the Google Trends data daily (by restricting each search query to three months) for each of the 30 TV series and the 32 NFL teams. When matching the Google search data to a particular NFL game, we

[17] https://support.google.com/adwords/answer/2999770?hl=en.

Table 7 Tweet Summary Statistics—32 NFL Teams

Variable	Mean	Median	Std dev	Maximum	Minimum	*N*
New_York_Jets	487.15	299	1,062.92	21,350	1	1,397
Dallas_Cowboys	412.18	123	1,325.90	20,595	0	1,397
New_England_Patriots	335.86	105	1,449.39	41,909	0	1,397
Pittsburgh_Steelers	323.05	88	1,521.75	40,416	0	1,397
New_Orleans_Saints	304.54	123	1,632.64	51,592	0	1,397
Baltimore_Ravens	299.67	70	1,621.66	47,427	0	1,397
Philadelphia_Eagles	282.29	105	762.12	8,558	0	1,397
Green_Bay_Packers	277.72	70	1,344.21	39,485	0	1,397
Chicago_Bears	273.73	123	779.28	18,011	0	1,397
San_Francisco_49ers	236.90	70	1,181.30	30,312	0	1,397
New_York_Giants	189.97	70	687.34	19,101	0	1,397
Washington_Redskins	171.14	70	454.71	5,412	0	1,397
Indianapolis_Colts	163.26	35	894.73	25,655	0	1,397
Detroit_Lions	161.95	70	440.27	6,467	0	1,397
Oakland_Raiders	151.34	70	308.94	3,251	0	1,397
Denver_Broncos	143.24	53	463.74	9,120	0	1,397
Minnesota_Vikings	131.97	53	547.91	16,043	0	1,397
Atlanta_Falcons	128.68	35	509.82	8,628	0	1,397
Arizona_Cardinals	127.62	53	364.66	9,014	0	1,397
Cleveland_Browns	125.95	70	211.57	2,407	0	1,397
Houston_Texans	123.80	35	435.19	6,414	0	1,397
Buffalo_Bills	116.78	53	228.05	3,497	0	1,397
Kansas_City_Chiefs	101.52	35	250.94	3,901	0	1,397
Miami_Dolphins	93.66	53	179.54	2,179	0	1,397
Seattle_Seahawks	87.84	35	344.30	8,224	0	1,397
Carolina_Panthers	81.94	53	143.11	1,423	0	1,397
Cincinnati_Bengals	81.85	35	240.20	4,393	0	1,397
St__Louis_Rams	71.84	35	115.69	1,494	0	1,397
Tampa_Bay_Buccaneers	70.77	35	104.68	1,125	0	1,397
San_Diego_Chargers	68.10	18	215.54	5,711	0	1,397
Tennessee_Titans	66.09	35	136.66	1,476	0	1,397
Jacksonville_Jaguars	63.89	35	106.47	1,037	0	1,397

add the numbers for the two teams that participated in the game.

3.6. Wikipedia

Wikipedia is a free access, free content Internet encyclopedia. As of February 2014, Wikipedia had 18 billion page views and nearly 500 million unique visitors every month (see Cohen 2014). Many of the Wikipedia editors are committed followers of the TV industry who gather information and edit related articles earlier than the show's release date. Consumers may view the edited information before the show or before the NFL games; therefore, the Wikipedia edits or views may be good predictors of TV ratings.

Figure 3 (Color online) Google Trend Plot

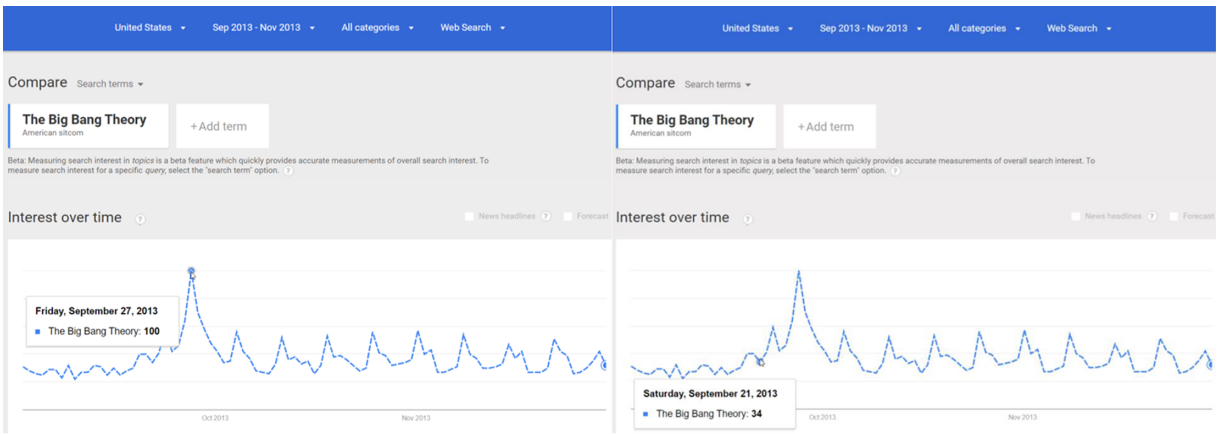


Table 8 Wikipedia Sizes

	2008		2009		2010		2011		2012		2013	
	Pages (Bil)	Size (GB)	Pages (Bil)	Size (GB)	Pages (Bil)	Size (GB)	Pages (Bil)	Size (GB)	Pages (Bil)	Size (GB)	Pages (Bil)	Size (GB)
1	2.5	19.9	4.9	39.0	5.4	42.9	6.2	49.2	7.7	61.5	8.7	69.4
2	2.4	18.8	4.4	35.0	5.3	42.2	6.1	48.3	7.2	57.2	7.7	61.5
3	2.5	19.9	5.1	40.8	5.7	45.7	3.1	24.6	7.5	59.6	8.6	68.4
4	2.6	20.9	4.8	38.7	5.3	42.4	6.2	49.1	7.1	56.8	8.1	64.4
5	3.6	28.9	5.1	40.9	5.4	42.8	6.6	52.3	7.5	59.6	8.6	68.8
6	4.3	34.4	5.0	39.7	4.3	33.9	6.4	51.5	7.3	57.9	8.4	66.9
7	4.3	34.0	5.0	39.9	4.4	35.5	6.6	53.0	7.5	60.0	8.1	64.8
8	4.3	34.1	5.1	40.5	5.1	40.6	6.8	54.0	7.5	60.2	8.2	65.8
9	4.3	34.6	3.5	27.7	5.8	46.6	6.1	48.6	7.8	61.9	7.9	63.0
10	4.5	35.7	5.5	43.7	5.9	47.3	7.3	58.4	8.1	64.6	8.5	67.7
11	4.4	35.0	5.5	43.5	5.7	45.8	7.1	56.7	8.0	64.2	8.0	63.5
12	4.7	37.1	5.4	42.9	6.0	47.6	6.6	52.4	8.2	65.6	8.3	66.2

We extract the edits and page view statistics from the Wikimedia Downloads site (http://dumps.wikimedia.org/other/pagecounts-raw). Table 8 shows the number of webpages and the size of the data files.

Instead of focusing only on the Wikipedia pages designated to the TV programs (for example, http://en.wikipedia.org/wiki/How_I_Met_Your_Mother_(season_8)), we searched through all of the page names that contain the shows' names (for example, A Change of Heart (How I Met Your Mother) or How I Met Your Mother (season 6)) or the keyword NFL (or National Football League). After selecting the relevant pages, we also find the corresponding page edit history using http://tools.wmflabs.org/xtools. Table 9 summarizes the counts of views and edits for all of the shows. As shown, Wikipedia edits are very limited. This is because only 31,000 people around the world are considered to be active editors for Wikipedia; regular consumers do not edit Wikipedia pages. In the following analysis, we drop the "edits" data and keep only "views."[18]

Table 9 Counts for Wikipedia Views, Edits, IMDB Reviews, and Huffington Post Articles

Show	Wikipedia views	Wikipedia edits	IMDB reviews	Huffington Post news
2 Broke Girls	7,072,696	6,282	161	105,000
30 Rock	13,777,360	1,126	80	657,000
90210	17,363,009	4,270	84	33,000
Allen Gregory	912,694	654	48	33,800
Blue Bloods	3,801,506	822	77	44,300
Body of Proof	2,912,555	828	40	187,000
Breaking Bad	41,449,326	4,437	688	229,000
Charlie's Angels	1,880,872	780	35	41,700
Cougar Town	6,586,439	1,622	60	15,100
Criminal Minds	21,460,772	4,752	141	45,400
Desperate Housewives	21,315,771	11,123	155	261,000
Gary Unmarried	978,092	742	20	363
Glee	57,126,905	7,232	158	411,000
Gossip Girl	18,298,602	7,072	183	380,000
Grey's Anatomy	31,073,450	9,463	251	489,000
Harry's Law	1,419,857	654	63	1,360
Hellcats	3,669,048	1,291	20	656
How I Met Your Mother	62,167,139	8,190	321	135,000
Lie to Me	8,458,678	1,631	108	236,000
Mike and Molly	2,477,749	771	39	332,000
NCIS	24,884,220	6,381	153	13,300
Nikita	7,936,616	1,655	77	12,600
Parks and Recreation	9,458,196	1,882	90	263,000
Private Practice	6,213,890	1,893	35	84,700
Rules of Engagement	5,472,316	1,151	37	47,000
Shark Tank	1,025,637	1,023	10	10,600
Smallville	14,857,326	8,737	358	974
The Big Bang Theory	56,785,777	8,864	314	229,000
The Vampire Diaries	26,545,974	4,919	244	23,100
Two and a Half Men	30,355,421	7,674	268	499,000
NFL	10,754,293	10,064		762,000

3.7. IMDB Reviews

Consumers also post reviews on discussion forums such as the IMDB. We choose the IMDB because it has the highest Web traffic ranking (according to Alexa) among all TV-show-related sites. As of January 18, 2015, IMDB had 58 million registered users. The site enables registered users to submit new material and request edits to existing entries.[19] The fourth column in Table 9 describes the number of reviews for each TV series. By contrast to Tweets and Wikipedia views, there are a limited number of IMDB reviews. The show with the most reviews, Breaking Bad, has only 688 posts over more than six years.

3.8. The Huffington Post News

Consumers may also be driven to watch TV series by news articles. Therefore, we collect data from The Huffington Post, a site that offers news, blogs, and

[18] We did not report the results with the edits data in Tables 15–18 and Table 20 because the performance of "edits" is dominated by the performance of "views."

[19] IMDB also features message boards that stimulate regular debates among its authenticated users. However, we cannot access historical discussions before September 2014.

Figure 4 (Color online) Comparing Ratings, Tweets, and Google Searches—Breaking Bad 2011

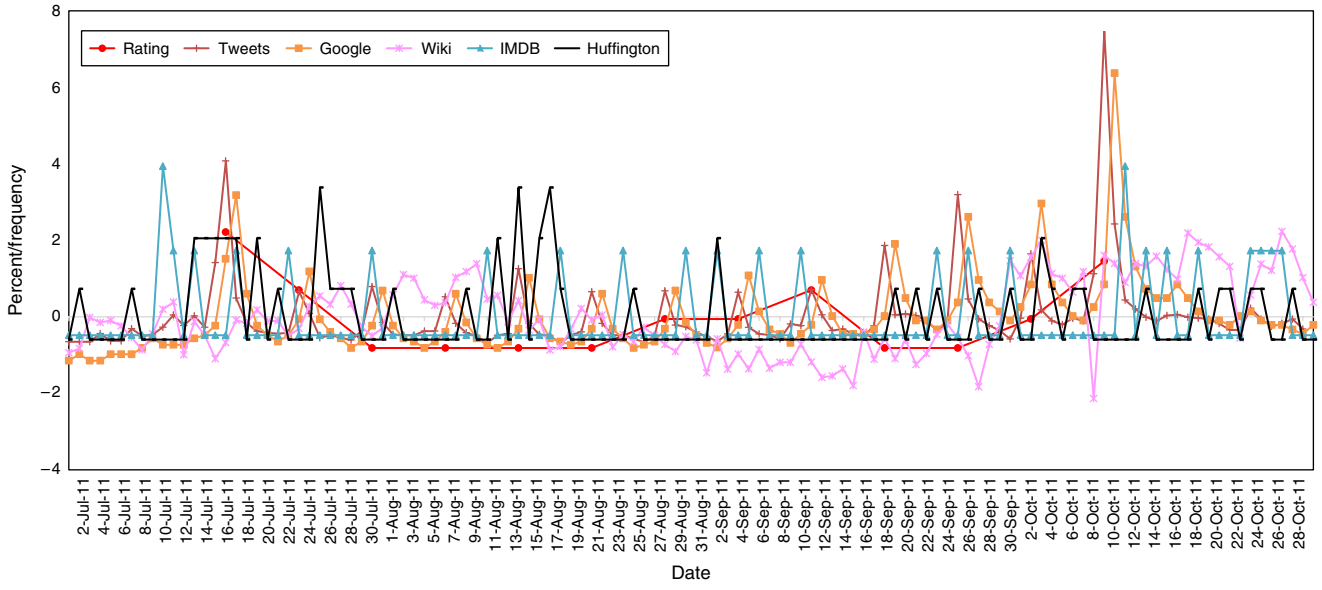
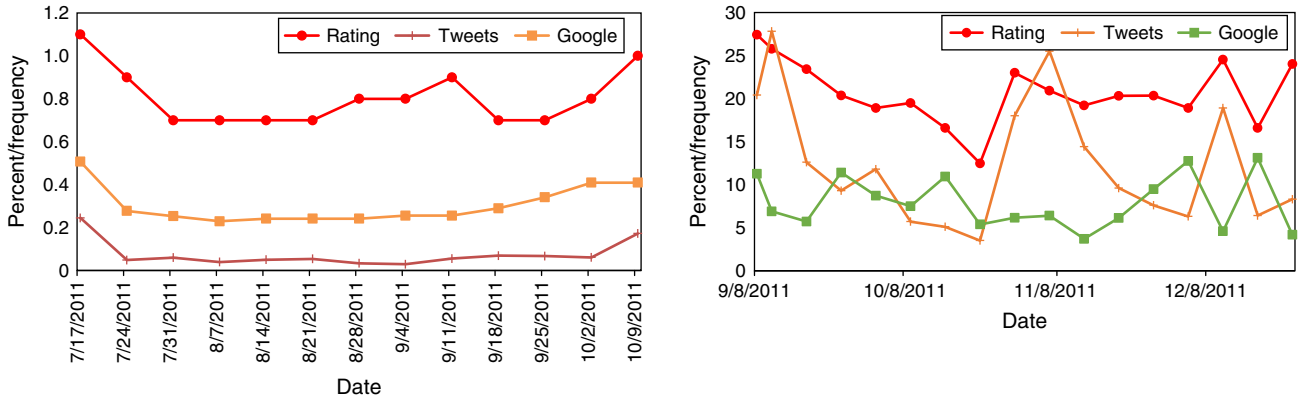


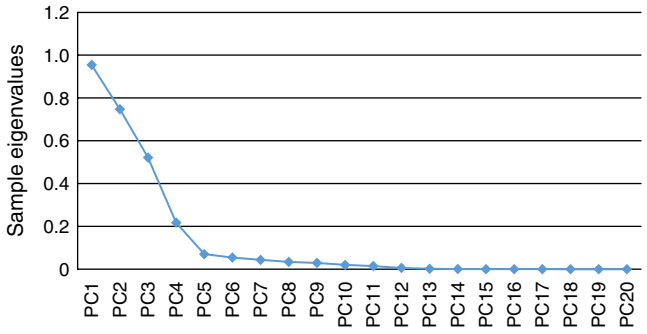
Figure 5 (Color online) Comparing Ratings, Tweets, and Google Searches (1 day)—Breaking Bad 2011[a] and SNF 2011



[a]We adjusted the Tweets and Google searches by a proportion to make the three series more comparable with a similar scale on one plot.

original content, and which covers entertainment, politics, business, etc. The site has extensive traffic and ranks 26th on Alexa as of January 29, 2015. The last column in Table 9 lists the number of news articles related to each TV series and NFL game. Interestingly, some shows, such as 30 Rock, do not have much social buzz (measured by Tweets) but are popular on news sites.

Figure 6 (Color online) Scree Plot



3.9. Comparing Ratings, Twitter Tweets, Google Trends, Wikipedia Views, IMDB Reviews, and Huffington Post News

Figure 4 illustrates the relation among Nielsen Ratings, Tweets, Google Trends, Wikipedia views, IMDB reviews, and Huffington Post news using the example of the fourth season of Breaking Bad in 2011. During that season, there was a new episode every week. Tweets spiked around every show day,[20] and a trail appeared

[20] The viewings of popular shows, such as Breaking Bad, can be time deferred, which may reduce the information relevance of Tweets to predict on-air TV ratings. However, we find that Tweets peak during the show time, which suggests that the majority of consumers still watch the shows on air.

Table 10 MapReduce for Three Tasks

No.	Task	Map		Reduce
		Key	Value	
1	Select relevant tweets/wiki	Keyword	1, text	Summation
2	*N*-gram frequency count	*N*-gram	1	Summation
3.1	SSVD—Matrix multiplication	Matrix row index	Matrix row vector	Null
3.2	SSVD—Orthogonalization	Submatrix	QR matrix	Summation

the day after the show. Google searches peaked after each show day. Wikipedia views and Huffington Post news often gradually increased in the next several days after a show day. However, the timing of IMDB reviews is irregular and somewhat misaligned with the ratings. Over the entire season, the TV ratings first moved down, then gradually increased, and reached a peak at the finale.

Because Tweets and Google searches cluster around the show days, in Figure 4 we determine whether the pattern of Tweets and Google searches on one day (24 hours) before each episode shows a similar pattern to the ratings. In Figure 5, we can see that in the general ratings patterns, Tweets and Google searches are similar, but Tweets and Google searches fail to capture some local variations in the ratings. As to the NFL (right panel of Figure 5), the trend in Tweets and Google searches shows an even larger variation from the ratings.

4. Structured Analysis

4.1.1. Dynamic Panel Data Linear Model. Our model is a dynamic panel data linear model (Bond 2002) that naturally extends Arellano and Bond (1991) (the AB approach). The AB approach is a single equation model with autoregressive dynamics and explanatory variables that are not strictly exogenous. It uses a Generalized Method of Moments (GMM) estimator to consider all of the potential orthogonality conditions. The AB approach solves the Nickell bias problem (Nickell 1981) of a fixed effects model using dynamic panel data. Our model fits the assumptions of the AB approach because we assume that the TV ratings in the current period can be explained by the ratings in previous periods and by the information in TGWIH before the TV program. In Equation (1), $i \in \{1, 2, \ldots, I\}$ denotes the TV series or the day of the week of the NFL prime time game. For example, for a TV series, $1 =$ "2 Broke Girls," with the orders shown in Table 2. For the NFL, $\{1 =$ Sunday Night Football, $2 =$ Thursday Night Football, and $3 =$ Monday Night Football$\}$. The subscript t denotes time. For a TV series, one period of time is one episode of the show, whereas for the NFL, one period is one week in the regular season. The j_{th} lag of the current period's TV show rating is denoted as $Rating_{it-j}$. The variable $F(Tweet_{it})$ denotes the information from Tweets regarding show i at time t. Because Tweets are unstructured text data, we try several functional forms for F that are explained later. The variable $Google_{it}$ measures the number of Google searches. The number of Wikipedia views is denoted by $Wiki$. The variables $F(IMDB_{it})$ and $F(Huffington_{it})$ contain the content information from IMDB reviews and Huffington Post News. The control variables, $Controls_{it}$, include a premier indicator, a finale indicator, the age of the show (which episode in the entire history) and seasonality (a winter dummy variable that indicates whether it is in the winter or not). The show-specific, time-invariant fixed effect that is unobserved to the researcher is captured by u_i, and ϵ_{it} is the unobserved shock. We use TGWIH shortly before the TV program begins so that these measures are exogenous to the current errors $(u_i + \epsilon_{it})$, but not strictly exogenous to the past errors

$$\begin{aligned} Rating_{it} = \alpha + \sum_{j=1}^{J} \beta_j Rating_{it-j} + \gamma F(Tweet_{it}) + \theta\ Google_{it} \\ + \delta\ Wiki_{it} + \eta F(IMDB_{it}) + \phi F(Huffington_{it}) \\ + \lambda\ Controls_{it} + u_i + \epsilon_{it}. \end{aligned} \quad (1)$$

According to Arellano and Bond (1991), we begin by first-differencing the model to eliminate all of the show-specific effects, u_i. Then, we use lags (up to period $t-2$) of the dependent and explanatory variables as instruments to perform a GMM estimation. These lags are valid instruments because they are uncorrelated with the difference in the error term but are correlated with the first differences of the endogenous variables. We are careful to avoid using too many lags as instruments, which Roodman (2009) emphasized as a problem.[21]

4.1.2. Information in Tweets, IMDB Reviews, and Huffington Post News.

Information Measures of Content: Count, Sentiment, and n-grams PCA. We use three measures to extract content information from the unstructured text data, including Tweets, IMDB reviews, and Huffington Post news. Below, we use Tweets as an illustration.

One easy measure of information in Tweets is how many times a TV program is discussed by Twitter users. When more users mention a TV program in Tweets, they are likely to watch the program, and the social network to which they are connected is likely to

[21] To prevent weakening of the Hansen test, we finally chose lags of up to five periods ahead as instruments (Andersen and Sørensen 1996, Bowsher 2002).

Table 11 Most Popular *N*-Grams in Shows

Show	1st	Count	%	2nd	Count	%	3rd	Count	%	Total
2 Broke Girls	2 Broke Girls	21,959	5.0	OfficialKat	19,199	4.3	Watch	7,227	1.6	442,665
30 Rock	30 Rock	568,593	8.0	Watch	127,239	1.8	Love	41,667	0.6	7,148,180
90210	90210	893,597	9.3	Watch	177,318	1.8	Doctor	61,888	0.6	9,601,842
Allen Gregory	Allen Gregory	8,181	9.2	JonahHill	2,100	2.4	Watch	1,713	1.9	89,161
Blue Bloods	Blue Bloods	169,526	8.5	Watch	21,809	1.1	Love	14,294	0.7	1,995,877
Body of Proof	Body of Proof	32,658	8.8	Watch	7,609	2.0	DanaDelany	3,958	1.1	372,839
Breaking Bad	Breaking Bad	113,950	2.7	aaronpaul	28,519	0.7	Tonight	18,407	0.4	4,145,246
Charlie's Angels	Charlie's Angels	15,372	9.4	Watch	3,379	2.1	Cancel	1,182	0.7	162,895
Cougar Town	Cougar Town	138,465	8.8	Watch	28,818	1.8	Show	14,161	0.9	1,575,302
Criminal Minds	Criminal Minds	962,982	9.4	Watch	215,756	2.1	Gublernaton	127,008	1.2	10,284,082
Desperate Housewives	Desperate Housewives	326,899	9.6	Watch	91,934	2.7	Season	26,254	0.8	3,404,224
Gary Unmarried	Gary Unmarried	1,664	9.1	Watch	389	2.1	Funny	143	0.8	18,334
Glee	Glee	2,174,587	8.7	Watch	311,489	1.2	Love	208,418	0.8	25,074,162
Gossip Girl	Gossip Girl	1,273,131	8.4	Watch	257,260	1.7	Now	116,795	0.8	15,177,057
Grey's Anatomy	Grey's Anatomy	111,321	2.6	Tonight	19,369	0.4	Watch	15,302	0.4	4,307,147
Harry's Law	Harry's Law	19,328	8.6	Watch	4,130	1.8	Kathy bates	2,379	1.1	224,088
Hellcats	Hellcats	56,686	9.8	Watch	10,595	1.8	Ashley tisdale	7,811	1.4	577,529
How I Met Your Mother	How I Met Your Mother	693,350	8.0	Watch	139,585	1.6	Neil patrick harris	117,773	1.4	8,625,859
Lie to Me	Lie to Me	18,608	8.7	Tim roth	6,525	3.1	Watch	1,844	0.9	213,067
Mike and Molly	Mike and Molly	6,454	9.9	Watch	2,095	3.2	Checked-in	891	1.4	65,037
NCIS	NCIS	54,054	2.3	Watch	8,014	0.3	Tonight	7,091	0.3	2,325,168
Nikita	Nikita	646,564	9.1	Go	91,309	1.3	Love	47,038	0.7	7,075,657
Parks and Recreation	Parks and Recreation	139,256	8.5	Watch	23,400	1.4	Office	11,899	0.7	1,631,277
Private Practice	Private Practice	201,836	8.6	Grey's anatomy	53,815	2.3	Watch	46,560	2.0	2,354,062
Rules of Engagement	Rules of Engagement	61,456	7.7	David spade	27,061	3.4	Go	8,554	1.1	795,710
Shark Tank	Shark Tank	165,728	7.7	Watch	21,795	1.0	Dragon's den	3,905	0.2	2,141,261
Smallville	Smallville	139,927	8.5	Watch	26,197	1.6	Season	16,833	1.0	1,651,297
The Big Bang Theory	Big bang*	121,590	2.6	Watch	16,103	0.3	Tonight	10,379	0.2	4,639,872
The Vampire Diaries	The Vampire Diaries	753,162	8.3	TV	255,581	2.8	Watch	136,691	1.5	9,025,715
Two and a Half Men	Two and a Half Men	32,337	2.3	Watch	5,043	0.4	Charlie sheen	3,901	0.3	1,390,918
NFL	cowboy	47,251	0.5	Game	36,550	0.4	Tonight	28,432	0.3	8,603,269
Total										135,138,799

Note. The same logic applies to the names of the other shows.
**N*-grams: Big, Bang, and Big Bang have the same frequency.

be influenced to watch the program. We call this the "Tweet volume model."[22]

A second measure of information is sentiment, where Tweets are classified by polarity, i.e., positive, neutral, and negative. Positive (negative) Tweets express favorable (unfavorable) feelings about a show, whereas neutral Tweets are neither positive nor negative. Hypothetically, positive Tweets are likely to generate positive feedback for a show, which increases TV ratings, whereas negative Tweets signal that consumers may stop watching the show because they are dissatisfied, which lowers future ratings. We construct two variables, $t\_pos = \#\ of\ positive\ tweets$ and $t\_neg = \#\ of\ negative\ tweets$, and test their effect on ratings in the "Tweet sentiment model." We constructed a sentiment analysis classifier using the LingPipe[23] linguistic analysis package, which provides a set of open-source Java libraries for natural language processing tasks. We use the DynamicLMClassifier, which is a language model classifier that accepts training events of categorized character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. To obtain labeled training data for the classifier, we hired two independent coders who have expertise in TV shows and NFL games to manually label 4% of the Tweets for each show/NFL team. We further adjusted the classified Tweets using the list of positive and negative opinion words that are provided by Hu and Liu (2004).

The third measure more closely examines the variety of content in Tweets. We discover that some Tweets that are related to a program may express only users' opinions concerning the program rather than indicating

[22] Thanks to one of the reviewers, we also considered a "length model" where the number of words in the documents (Tweets/IMDB reviews and Huffington Post news articles) is used as the explanatory variable. The results are similar to the "volume model." Estimates are available on request.

[23] http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html.

Table 12 Sentiment Distribution

Show	Positive (%)	Neutral (%)	Negative (%)	Total
2 Broke Girls	21.23	78.16	0.6	41,189
30 Rock	30.11	69.85	0.05	569,119
90210	4.85	95.13	0.02	874,210
Allen Gregory	31.56	41.59	26.84	8,181
Blue Bloods	18.93	80.66	0.40	169,600
Body of Proof	17.84	81.38	0.78	32,633
Breaking Bad	12.33	87.11	0.56	2,412,016
Charlie's Angels	41.76	46.31	11.93	15,372
Cougar Town	12.34	87.40	0.26	138,465
Criminal Minds	8.35	91.47	0.18	962,982
Desperate Housewives	20.42	79.55	0.03	326,899
Gary Unmarried	26.17	56.64	17.18	1,667
Glee	23.77	75.37	0.86	2,174,587
Gossip Girl	8.35	91.47	0.18	1,273,131
Grey's Anatomy	33.17	61.22	5.60	2,420,866
Harry's Law	10.26	88.21	1.54	19,328
Hellcats	31.46	67.45	1.08	56,686
How I Met Your Mother	14.55	85.12	0.33	693,350
Lie to Me	18.17	80.55	1.29	18,608
Mike and Molly	18.66	80.60	0.75	6,454
NCIS	36.38	63.56	0.06	987,933
Nikita	5.92	71.39	5.92	646,564
Parks and Recreation	21.60	78.29	0.12	139,256
Private Practice	32.37	67.09	0.54	201,836
Rules of Engagement	4.29	95.61	0.10	61,456
Shark Tank	16.93	78.74	4.33	165,728
Smallville	6.21	93.74	0.04	139,927
The Big Bang Theory	31.91	66.90	1.19	2,491,124
The Vampire Diaries	27.87	72.08	0.05	753,162
Two and a Half Men	11.84	85.30	2.86	728,359
NFL	36.31	57.37	6.32	2,040,139

an intent to watch the upcoming program. For example, consider the following two sample Tweets about the show "Breaking Bad":

"I learnt how to cook meth like the guy in breaking bad"; and "Pumped for the season finale of Breaking Bad tonight. Only 4 hours and 37 minutes to go."

The first Tweet discusses a featured behavior of the character in the show. From this Tweet, we can infer that the Twitter user has watched Breaking Bad and is interested in its story. Nevertheless, the second Tweet directly states the future viewing behavior of the user. If there are many Tweets similar to the second Tweet before the show starts, the show's ratings are likely to be high. By contrast, variations of Tweets similar to the first one may have far less predictive power.

Based on this rationale, we construct a third measure of information to make inferences from the full content of the Tweets. More specifically, we use the counts of each *n*-gram in the Tweet. An *n*-gram is a continuous sequence of *n* words in the text. For example, "big data" is a 2-gram, and "big" is a 1-gram. The sample Tweet "I love Pittsburgh Steelers" contains four 1-gram, three 2-grams, two 3-grams, and one 4-gram. Because phrases provide more interpretable information than a single word, we count *n*-grams rather than counting only words. We label this the "Tweet content model."

Table 13 *N*-Grams with Highest Loadings on First Four PCs

PC1	PC2	PC3	PC4
Tonight	Bed	Season	Excited
Can't wait	Home	Start	Finale
Watch	TV	Premiere	Love

Table 14 Variables in the Regression with Descriptions

Variable	Description
Rating_lag	Lagged rating, i.e., Nielsen rating for the previous episode of this show
Tweets	Number of show-related tweets 24/48 hours before the show starts
Google	Number of show-related Google searches (from Google Trend) 24/48 hours before the show starts
Wiki	Number of show-related Wikipedia views 24/48 hours before the show starts
IMDB	Number of show-related IMDB reviews 24/48 hours before the show starts
Huffington	Number of show-related Huffington Post news articles 24/48 hours before the show starts
Tweet_Pos	Number of show-related positive tweets 24/48 hours before the show starts
Tweet_Neg	Number of show-related negative tweets 24/48 hours before the show starts
Tweet_PC#	#th principal component score for the *n*-gram matrix derived from show-related tweets 24/48 hours before the show starts
Tweet_T#	#th topic model proportion for the *n*-gram matrix derived from show-related tweets 24/48 hours before the show starts
Premier	Indicator of whether this episode is a season premier
Finale	Indicator of whether this episode is a season finale
Age	Total episode number since Season 1 Episode 1
Winter	Indicator of whether the episode is broadcast in December, January or February.

Information Timeliness. Another decision concerns the length of time to collect TGWIH before a show begins. The shows are generally broadcast weekly. Consistent with what we find in §3, Tweets and Google searches also follow a weekly trend where more instances occur around (shortly before and after) the show. Intuitively, on the first or second day after the previous show, consumers are more likely to focus on the old show, whereas one or two days before the new show, consumers are more likely to discuss the new show. Following this intuition, we use TGWIH 24 or 48 hours before the new show starts as the input variable. It is interesting to compare the performance of the 24-hour measure with the 48-hour measure to evaluate the value of information over time.[24]

4.1.3. Challenges in Processing Enormous Unstructured Data—Cloud Computing Techniques. As shown in §3, our analysis involves an enormous amount

[24] We also conducted an analysis for the one-week window. The results are provided in Online Appendix A2 (available as supplemental material at http://dx.doi.org/10.1287/mksc.2015.0972).

of unstructured data. For example, we have approximately 1.8 billion Tweets, 433 billion Wikipedia pages, and 5.5 million Huffington Post news articles. Our data-cleaning process includes the following three major procedures: (1) selecting relevant Tweets/Wikipedia/Huffington Post pages, (2) counting *n*-grams, and (3) using the stochastic singular value decomposition (SSVD). The first two tasks can be performed in a streaming fashion (no out-of-memory problems) but are extremely time consuming on a single machine given the volume of data. The last task cannot be performed on a single machine because the size of the matrix does not fit in memory.

For example, the content information in Tweets is substantial. Even when using the 24-hour measure, we selected 6,894,624 Tweets related to the 30 TV series and 2,004,987 Tweets that are related to the NFL. These Tweets generate 28,044,202 and 9,028,774 *n*-grams that appear at least five times. Moreover, in our regression model for the TV shows, we hope to incorporate all of the content information. One way to do this is to use the frequency of all *n*-grams as features. This approach provides us significant feature space. Therefore, we must rely on dimension reduction techniques such as Principle Component Analysis (PCA) to make the task more tractable. However, performing PCA on a $2{,}339 \times 28{,}044{,}202$[25] matrix cannot be accomplished on a single machine because the matrix is too large to be stored in memory.

Our solution to this challenge is to use the SSVD method developed by Halko (2012). The key idea behind the SSVD is that when a data matrix is too large to be stored in memory, randomized sampling (the stochastic element of SSVD) allows the algorithms to work in a streaming environment to rapidly construct a low-rank approximation of the matrix. The SSVD parallelizes and distributes the randomized sampling and factorization stages using Hadoop MapReduce.

Next, we explain how we solve these challenges using cloud computing services.

Because the computing task cannot be managed by a single machine, programs have been developed to exploit the capacities of massively distributed computational resources. MapReduce is a good example.[26] MapReduce is a programming model for processing large data sets using a parallel, distributed algorithm on a cluster. It is very powerful because it abstracts the complexities of parallel programming down to two operations, i.e., a map and a reduce. Both the map and reduce steps can be parallelized on a cluster of computers. Between the map and reduce, the process involves shuffling and sorting the keys so that all key-value pairs of the same key go to the same reduce for the next step. Thus, data communication occurs only between the map and the reduce. The details, such as distributed computation, file storage, communication, and data transfer, are left for the framework, such as Hadoop,[27] to manage.

We implement MapReduce using Amazon Elastic MapReduce (EMR).[28] Specifically, we use EMR for all three tasks, which are (1) selecting the relevant Tweets/Wikipedia pages, (2) performing the *n*-gram counts, and (3) conducting the SSVD. Table 10 summarizes how we design the map and reduce jobs for each task.

Here we use the Tweet *n*-gram count task as an illustration. In the first procedure, "Map" filters the input data in key value pairs. When reading one Tweet as the input, if we find one *n*-gram in the Tweet, then we set the key as the *n*-gram and the value as 1. The second procedure, "Reduce," then summarizes the key-value pairs generated by the "Map" procedure. "Reduce" adds all of the values of the same key (*n*-gram) as the summary count of the *n*-gram.

When implementing the SSVD, we used Mahout, an open-source machine learning library that uses Hadoop MapReduce to implement scalable distributed algorithms. Essentially, it breaks down the singular value decomposition of a huge matrix into two basic operations, which are matrix multiplication and orthogonalization. Because both operations can rely on MapReduce to be performed in distributed clusters, our computational challenge is resolved.

4.1.4. Alternative Machine Learning Models. In addition to the cloud-based PCA model and the dynamic panel linear model that are explained above, we use alternative content extraction models and machine learning models for prediction comparison. In terms of content extraction, we compare the current cloud-based PCA model with the Latent Dirichlet Allocation (LDA) Topic Modeling approach (Blei et al. 2003). Moreover, we compare the dynamic panel linear model

[25] The number of episodes of the shows is 2,339. The number of *n*-grams that are generated from Tweets 24 hours before a show is 28,044,202. For the larger data set of Tweets 48 hours before a show, the number of *n*-grams is 34,855,764.

[26] MapReduce developers tout MapReduce for its scalability, fault tolerance, and elasticity. Google uses it to process 20 Pb of data per day.

[27] Hadoop is an open-source software framework that allows the distributed processing of large data sets across clusters of computers using simple programming models. It contains (1) the Hadoop Common package, which provides file system and OS level abstraction, (2) Yarn, a MapReduce engine, and (3) the Hadoop Distributed File System. These mechanisms automatically break down jobs into distributed tasks, schedule jobs, and tasks efficiently at participating cluster nodes, and tolerate data and task failures.

[28] EMR was created to enable researchers, businesses, and developers to easily and efficiently process vast amounts of data in a pay-as-you-go fashion. For more detailed information about implementation, see Online Appendix A4.

Table 15 Impact of Previous 24 Hours of TGWIH on TV Ratings: TV Series[a]

24 hr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	LR	T	G	W	I	H	LR + T	LR + G	T + Sen	LR + T + Sen	Topic	LR + Topic	LR + T + G + W + I + H + Topic	PC	LR + PC	LR + T + G + W + I + H + PC
Rating_lag	0.473 (< 0.001)						0.499 (< 0.001)	0.483 (< 0.001)		0.498 (< 0.001)		0.237 (< 0.001)	0.217 (< 0.001)		0.187 (< 0.001)	0.161 (< 0.001)
Tweets		0.001 (< 0.001)					0.001 (< 0.001)		0.001 (0.031)	0.001 (0.145)			2.06E−04 (0.169)			2.93E−04 (0.016)
Google			1.01E−06 (0.037)					4.39E−06 (< 0.001)					1.09E−06 (0.001)			1.27E−06 (< 0.001)
Wiki				1.90E−05 (< 0.001)									4.46E−06 (0.014)			4.55E−06 (0.002)
IMDB					0.020 (0.280)								0.008 (0.642)			0.010 (0.483)
Huffington						0.011 (0.672)							−0.028 (0.092)			−0.014 (0.180)
Tweet_Pos									0.001 (0.252)	0.001 (0.512)						
Tweet_Neg									0.006 (0.438)	0.008 (0.317)						
Tweet_PC1														0.255 (< 0.001)	0.220 (< 0.001)	0.213 (< 0.001)
Tweet_PC2														0.424 (< 0.001)	0.367 (< 0.001)	0.356 (< 0.001)
Tweet_PC3														0.670 (< 0.001)	0.577 (< 0.001)	0.559 (< 0.001)
Tweet_PC4														0.730 (< 0.001)	0.628 (< 0.001)	0.608 (< 0.001)
Tweet_T1											0.250 (< 0.001)	0.243 (< 0.001)	0.233 (< 0.001)			
Tweet_T2											0.737 (< 0.001)	0.701 (< 0.001)	0.670 (< 0.001)			
Tweet_T3											0.324 (< 0.001)	0.309 (< 0.001)	0.296 (< 0.001)			
Tweet_T4											0.848 (< 0.001)	0.810 (< 0.001)	0.777 (< 0.001)			
Tweet_T5											1.061 (< 0.001)	1.012 (< 0.001)	0.097 (< 0.001)			
Premier	0.191 (0.164)	0.267 (< 0.001)	0.352 (< 0.001)	0.273 (< 0.001)	0.346 (< 0.001)	0.348 (< 0.001)	0.333 (< 0.001)	0.303 (< 0.001)	0.266 (< 0.001)	0.333 (< 0.001)	0.360 (< 0.001)	0.391 (< 0.001)	0.324 (0.001)	0.391 (< 0.001)	0.387 (< 0.001)	0.309 (< 0.001)
Finale	−0.209 (0.377)	−0.039 (0.288)	−0.025 (0.497)	−0.076 (0.035)	−0.027 (0.450)	−0.027 (0.453)	0.009 (0.809)	0.004 (0.921)	−0.040 (0.279)	0.008 (0.833)	−0.046 (0.097)	0.013 (0.578)	−0.014 (0.570)	−0.029 (0.270)	6.94E−05 (0.998)	−0.043 (0.122)
Age	−0.019 (0.909)	−0.135 (< 0.001)	−0.133 (< 0.001)	−0.215 (< 0.001)	−0.143 (< 0.001)	−0.143 (< 0.001)	−0.128 (0.001)	−0.016 (0.003)	−0.135 (< 0.001)	−0.128 (0.001)	−0.177 (< 0.001)	−0.136 (< 0.001)	−0.146 (< 0.001)	−0.179 (< 0.001)	−0.139 (< 0.001)	−0.153 (< 0.001)

Table 15 (Continued)

24 hr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	LR	T	G	W	I	H	LR+T	LR+G	T+Sen	LR+T+Sen	Topic	LR+Topic	LR+T+G+W+I+H+Topic	PC	LR+PC	LR+T+G+W+I+H+PC
Winter	0.086	−0.022	−0.021	0.030	−0.034	−0.034	0.092	0.117	−0.024	0.091	0.021	0.051	0.068	0.068	0.065	0.085
	(0.529)	(0.482)	(0.511)	(0.319)	(0.275)	(0.275)	(< 0.001)	(< 0.001)	(0.438)	(< 0.001)	(0.289)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
R2	0.751	0.065	0.049	0.121	0.048	0.048	0.756	0.755	0.065	0.756	0.620	0.890	0.893	0.756	0.881	0.906
Wald Chi2	114.890	150.900	112.800	309.800	113.400	112.150	244.680	181.930	152.600	277.210	3,661.110	2,500.590	8,206.060	6,953.600	1,761.670	3,688.090
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
AR(1)	−2.463						−2.806	−3.022		−2.083		−4.254	−4.237		−4.318	−4.369
	(0.014)						(0.005)	(0.003)		(0.005)		(< 0.001)	(< 0.001)		(< 0.001)	(< 0.001)
AR(2)	1.646						1.935	2.227		1.923		1.829	1.794		0.792	0.331
	(0.100)						(0.053)	(0.026)		(0.054)		(0.067)	(0.073)		(0.428)	(0.740)
Sargan Chi2	22.179						22.477	23.503		18.277		19.957	7.999		15.986	11.153

[a]We also conduct the sentiment analysis and *n*-gram PCA for IMDB reviews and Huffington Post news. The results are similar to columns 5 and 6 in Table 15.

with the machine learning models that are widely used in other papers related to Twitter predictions, including Auto Regression with Exogenous Independent Variables (Autoregression X), Multilayered feedforward neural networks, and the Self-organizing Fuzzy Neural Network (SOFNN) models. For detailed information concerning these alternative models, see Online Appendix A1. We discuss the prediction performance of these competing models in §5.

5. Results

5.1. Specification Tests and Fit

5.1.1. Stationarity Tests. We apply the Augmented Dickey-Fuller (ADF) unit root test on all of the variables (Ratings, Twitter Tweets, Tweet sentiments, Tweet Principal Components, Google searches, Wikipedia views, IMDB reviews, and Huffington Post news). In all cases, the null hypothesis that the series is nonstationary is rejected.

5.1.2. IV Validity and Serial Correlation. We use the Sargan Chi-squared statistic to test the validity of the instruments. Tables 15 to 17 report that all of the over-identifying test statistics are insignificant. Therefore, we cannot reject the joint validity of the instrument sets. In addition, we use the method developed by Arellano and Bond (1991) to determine whether the errors are serially correlated (the AR(1) and AR(2) test scores in Tables 15–17). There is no evidence of first-order serial correlation.[29]

5.1.3. Number of Lags. To determine the number of lags (J in Equation (1)) of the dependent variable in our model, we use the MMSC-BIC statistics developed by Andrews and Lu (2001). Comparisons show that including only the first lag yields the lowest MMSC-BIC. Therefore, in Tables 15 to 17, we report the results with only the first lag included.

5.2. Main Findings

5.2.1. People Tweet About What They Are Going to Do.

The n-grams. In Table 11, we list the *n*-grams with the 1st, 2nd, and 3rd highest frequencies for each of the five TV series and the NFL. Across all five TV series, the most mentioned topic is the name of the program, such as Breaking Bad. Moreover, "watch" and "tonight" appear with very high frequency. In fact, we find many Tweets that discuss the consumer's intention to watch the show, for example, "I can't wait to watch Breaking Bad tonight." The content in this type of Tweet is a useful predictor for the ratings of

[29] There is also no second-order serial correlation. The results are not reported here but are available on request.

Table 16 Impact of Previous 48 Hours of TGWIH on TV Ratings: TV Series

48 hr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	LR	T	G	W	I	H	LR + T	LR + G	T + Sen	LR + T + Sen	Topic	LR + Topic	LR + T + G + W + I + H + Topic	PC	LR + PC	LR + T + G + W + I + H + PC
Rating_lag	0.473 (< 0.001)						0.492 (< 0.001)	0.423 (< 0.001)		0.495 (< 0.001)		0.288 (< 0.001)	0.299 (< 0.001)		0.308 (< 0.001)	0.265 (< 0.001)
Tweets		0.001 (< 0.001)					0.001 (< 0.001)		0.001 (0.037)	0.001 (0.155)			0.001 (0.163)			0.001 (0.023)
Google			1.68E−06 (0.043)					2.87E−06 (< 0.001)					2.09E−06 (< 0.001)			1.43E−06 (< 0.001)
Wiki				9.38E−06 (< 0.001)									1.66E−05 (< 0.001)			3.88E−05 (< 0.001)
IMDB					−0.027 (0.293)								−0.311 (0.397)			0.026 (0.459)
Huffington						0.012 (0.521)							0.049 (0.046)			−0.025 (0.326)
Tweet_Pos									0.001 (0.235)	0.001 (0.418)						
Tweet_Neg									0.005 (0.418)	0.006 (0.500)						
Tweet_PC1														0.504 (< 0.001)	0.569 (< 0.001)	0.566 (< 0.001)
Tweet_PC2														0.834 (< 0.001)	0.859 (< 0.001)	0.839 (< 0.001)
Tweet_PC3														0.960 (< 0.001)	0.946 (< 0.001)	0.915 (< 0.001)
Tweet_PC4														1.394 (< 0.001)	1.273 (< 0.001)	1.062 (< 0.001)
Tweet_T1											0.363 (< 0.001)	0.387 (< 0.001)	0.329 (< 0.001)			
Tweet_T2											1.342 (< 0.001)	1.500 (< 0.001)	1.648 (< 0.001)			
Tweet_T3											0.603 (< 0.001)	0.589 (< 0.001)	0.701 (< 0.001)			
Tweet_T4											1.275 (< 0.001)	1.404 (< 0.001)	1.322 (< 0.001)			
Tweet_T5											1.874 (< 0.001)	1.852 (< 0.001)	1.725 (< 0.001)			
Premier	0.191 (0.164)	0.294 (< 0.001)	0.308 (< 0.001)	0.268 (< 0.001)	0.351 (< 0.001)	0.386 (< 0.001)	0.343 (< 0.001)	0.329 (< 0.001)	0.271 (< 0.001)	0.334 (< 0.001)	0.361 (< 0.001)	0.380 (< 0.001)	0.384 (< 0.001)	0.372 (< 0.001)	0.363 (< 0.001)	0.395 (< 0.001)
Finale	−0.209 (0.377)	−0.036 (0.265)	−0.024 (0.553)	−0.089 (0.025)	−0.081 (0.733)	−0.081 (0.020)	0.003 (0.903)	0.006 (0.748)	−0.053 (0.307)	0.051 (0.699)	−0.072 (0.054)	0.035 (0.367)	−0.053 (0.189)	−0.094 (0.633)	−0.043 (0.216)	−0.066 (0.491)
Age	−0.019 (0.909)	−0.136 (< 0.001)	−0.130 (< 0.001)	−0.233 (< 0.001)	−0.129 (< 0.001)	−0.023 (< 0.001)	−0.132 (< 0.001)	−0.107 (< 0.001)	−0.256 (< 0.001)	−0.183 (0.001)	−0.183 (< 0.001)	−0.169 (< 0.001)	−0.149 (< 0.001)	−0.201 (0.001)	−0.148 (< 0.001)	−0.170 (< 0.001)

Table 16 (Continued)

48 hr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	LR	T	G	W	I	H	LR+T	LR+G	T+Sen	LR+T+Sen	Topic	LR+Topic	LR+T+G+W+I+H+Topic	PC	LR+PC	LR+T+G+W+I+H+PC
Winter	0.086	−0.025	−0.022	0.032	−0.035	−0.031	−0.174	0.162	0.123	0.126	0.096	0.063	0.036	0.045	0.083	0.066
	(0.529)	(0.481)	(0.447)	(0.497)	(0.403)	(0.251)	(< 0.001)	(< 0.001)	(0.432)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
R2	0.751	0.078	0.062	0.133	0.055	0.053	0.764	0.769	0.086	0.770	0.727	0.895	0.899	0.787	0.901	0.910
Wald Chi2	114.890	152.381	116.006	311.587	132.389	133.556	276.158	279.465	155.469	279.370	2,854.333	3,006.375	8,243.639	6,437.452	1,824.338	3,943.277
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
AR(1)	−2.463						−2.835	−3.338		−3.089		−4.020	−4.326		−4.138	−4.391
	(0.014)						(0.005)	(0.004)		(0.002)		(0.001)	(< 0.001)		(< 0.001)	(< 0.001)
AR(2)	1.646						1.828	2.279		1.762		1.244	1.555		0.459	0.293
	(0.100)						(0.063)	(0.073)		(0.059)		(0.089)	(0.058)		(0.668)	(0.773)
Sargan Chi2	22.179						23.412	24.939		23.684		19.568	8.137		15.678	11.477

the program. Not surprisingly, we also find that the celebrities on the TV programs are another salient topic, such as Charlie Sheen in Two and a Half Men. Twitter users express their preference for a celebrity and also re-Tweet what the celebrity says. For example, many people Tweeted that "Two and A Half Men is never the same without Charlie Sheen" after Sheen was replaced on the show in 2011. Fans of The Big Bang Theory often Tweeted phrases such as "Bazinga! Love Sheldon," where "Sheldon" is a main character on the show, and "Bazinga" is a phrase that he frequently utters.

Sentiment. We find several interesting phenomena in the sentiment analysis of the Tweets (Table 12). First, the majority of the Tweets are neutral and document consumers' mundane lives, such as their actions and plans. Second, consistent with the previous findings of Godes and Mayzlin (2004), consumers are much more positive than negative about the TV shows that they watch. On average, there are 7.7 times more positive Tweets than negative. However, this ratio is relatively lower for NFL games than for TV series.

One way to explain this result is self-selection bias. Consumers who Tweet about TV shows are those who enjoy the shows most. Thus, they are more positive than the entire population. However, the theory of cognitive dissonance indicates that people tend to justify their own mistakes. In our context, even if a consumer dislikes a show that she watched, she may be unwilling to admit that she made a mistake in choosing the show by broadcasting it on Twitter. These properties of the sentiment analysis of Tweets prevent this aspect from predicting TV ratings.

PCA. In the PCA, we follow the standard approach and use a "scree plot" (Cattell 1966) to decide how many principal components to retain. If the largest few eigenvalues in the covariance matrix dominate in magnitude, then the scree plot will exhibit an "elbow." We apply this "elbow" rule and select four principal components from the 28,044,202 *n*-gram features based on Figure 6.

It is interesting to examine the matrix of eigenvectors (loadings) to observe which *n*-grams contribute the most to the chosen principal components. Table 13 shows the three *n*-grams with the largest loadings on each of the first four principal components (PC).

Consistent with our findings from the *n*-gram count, words and phrases such as "tonight," "can't wait," and "watch" have the largest projection on the first PC. Location- and device-related words such as "home" and "tv" contribute most to the second PC. The third PC captures consumers' attention to the "premiere" of the "season;" the fourth PC contains positive emotions such as "excited," "love," and another prominent topic, "finale." Overall, the first four PCs cover consumers' intention to watch the shows. Later, in the regression

Table 17 Impact of TGWIH on TV Ratings: NFL

Rating	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	LR	T	G	W	H	D	LR + T + D	LR + G + D	T + Sen	LR + T + Sen	Topic	LR + D + Topic	LR + T + G + W + H + D + Topic	PC	LR + D + PC	LR + T + G + W + H + D + PC	LR + T + G + W + H + PC
Rating_lag	0.053 (0.641)						0.066 (0.336)	0.177 (0.002)		0.160 (0.003)		0.019 (0.691)	−0.054 (0.931)		−0.006 (0.808)	0.092 (0.116)	0.071 (0.459)
Tweets		0.026 (< 0.001)					0.018 (< 0.001)		0.064 (< 0.001)	0.083 (< 0.001)			0.006 (< 0.001)			0.016 (< 0.001)	0.031 (< 0.001)
Google			2.68E−05 (0.995)					0.001 (0.750)					0.001 (0.580)			0.001 (< 0.001)	0.002 (< 0.001)
Wiki				2.13E−04 (< 0.001)									−0.001 (0.060)			0.001 (< 0.001)	4.44E−04 (< 0.001)
Huffington					0.028 (0.269)								0.093 (0.461)			0.021 (0.492)	0.074 (0.474)
Team dummy						~	~	~				~	~		~	~	
Tweet_Pos									0.010 (0.430)	0.006 (0.238)							
Tweet_Neg									0.010 (0.672)	0.005 (0.745)							
Tweet_PC1														1.247 (< 0.001)	1.173 (< 0.001)	0.631 (< 0.001)	0.854 (< 0.001)
Tweet_PC2														0.478 (< 0.001)	1.053 (< 0.001)	1.454 (< 0.001)	1.559 (< 0.001)
Tweet_PC3														1.064 (< 0.001)	1.345 (< 0.001)	0.880 (< 0.001)	0.995 (< 0.001)
Tweet_PC4														1.039 (< 0.001)	1.151 (< 0.001)	1.271 (< 0.001)	1.288 (< 0.001)
Tweet_T1											0.701 (< 0.001)	0.720 (< 0.001)	1.902 (< 0.001)				
Tweet_T2											1.194 (< 0.001)	1.387 (< 0.001)	0.754 (< 0.001)				
Tweet_T3											1.592 (< 0.001)	1.283 (< 0.001)	1.119 (< 0.001)				
Tweet_T4											1.261 (< 0.001)	0.841 (< 0.001)	1.182 (< 0.001)				
Tweet_T5											0.425 (< 0.001)	1.056 (< 0.001)	0.668 (< 0.001)				
Premier	1.671 (< 0.001)	−2.073 (< 0.001)	1.618 (< 0.001)	1.640 (< 0.001)	1.737 (< 0.001)	1.457 (< 0.001)	−1.604 (< 0.001)	1.644 (< 0.001)	1.626 (< 0.001)	1.584 (< 0.001)	1.313 (< 0.001)	0.999 (< 0.001)	1.177 (0.001)	1.469 (< 0.001)	1.594 (< 0.001)	1.972 (< 0.001)	1.426 (< 0.001)
Finale	1.176 (0.371)	0.901 (0.263)	1.246 (0.204)	1.682 (0.096)	1.721 (0.231)	2.168 (0.071)	1.813 (0.236)	1.940 (0.140)	2.012 (0.213)	0.996 (0.391)	1.485 (0.258)	2.161 (0.205)	1.993 (0.216)	1.661 (0.243)	1.832 (0.238)	1.571 (0.242)	1.821 (0.239)
Age	−0.311 (0.025)	−0.261 (< 0.001)	−0.228 (< 0.001)	−0.225 (< 0.001)	−0.302 (< 0.001)	0.018 (< 0.001)	−0.188 (< 0.001)	−0.123 (< 0.001)	−0.308 (< 0.001)	−0.207 (0.002)	−0.297 (< 0.001)	−0.249 (0.019)	−0.257 (0.001)	−0.220 (0.001)	−0.388 (< 0.001)	−0.281 (< 0.001)	−0.323 (< 0.001)

Table 17 (Continued)

Rating	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	LR	T	G	W	H	D	LR+T+D	LR+G+D	T+Sen	LR+T+Sen	Topic	LR+D+Topic	LR+T+G+W+H+D+Topic	PC	LR+D+PC	LR+T+G+W+H+D+PC	LR+T+G+W+H+PC
Winter	0.032	0.023	0.021	0.029	0.032	0.030	0.031	0.025	0.030	0.031	0.037	0.030	0.034	0.037	0.023	0.022	0.032
	(0.489)	(0.489)	(0.503)	(0.485)	(0.488)	(0.479)	(0.495)	(0.486)	(0.487)	(0.475)	(0.479)	(0.485)	(0.490)	(0.497)	(0.494)	(0.496)	(0.487)
R2	0.248	0.342	0.039	0.059	0.055	0.692	0.790	0.778	0.363	0.777	0.815	0.835	0.883	0.845	0.868	0.916	0.900
Wald Chi2	64.759	238.192	226.827	250.048	238.887	69.505	127.236	113.371	252.320	246.614	325.728	304.949	316.085	331.147	337.293	331.981	336.962
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
AR(1)	−1.972						−2.235	−2.051		−1.972		−2.001	−1.963		−2.390	−2.043	−2.514
	(0.049)						(0.026)	(0.041)		(0.048)		(0.046)	(0.050)		(0.017)	(0.040)	(0.012)
AR(2)	0.056						0.754	−0.317		0.441		0.649	0.844		0.636	0.873	0.805
	(0.965)						(0.450)	(0.750)		(0.652)		(0.522)	(0.397)		(0.530)	(0.382)	(0.419)
Sargan Chi2	124.100						57.114	63.062		105.399		52.309	47.959		45.812	46.276	75.884

Table 18 Prediction: TV Series

Model	MAPE	MSE	Model	MAPE	MSE
24 hr			48 hr		
1	0.121	0.082	1	0.126	0.083
2	0.466	0.339	2	0.464	0.333
3	0.476	0.348	3	0.420	0.337
4	0.439	0.320	4	0.404	0.310
5	0.481	0.373	5	0.483	0.360
6	0.503	0.369	6	0.500	0.381
7	0.130	0.091	7	0.123	0.083
8	0.115	0.077	8	0.118	0.085
9	0.468	0.333	9	0.452	0.313
10	0.118	0.079	10	0.111	0.077
11	0.125	0.098	11	0.120	0.083
12	0.088	0.061	12	0.075	0.051
13	0.078	0.052	13	0.073	0.051
14	0.119	0.089	14	0.125	0.088
15	0.056	0.041	15	0.050	0.040
16	0.070	0.043	16	0.066	0.041

results analysis, we confirm that this information summary is indicative of users' upcoming consumption (i.e., watching the show).

5.2.2. Twitter Content Is a Lead Indicator of TV Ratings. In Table 14, we first list the variables included in the regression.

The results show that the lagged rating itself explains 75.1% of the variation in the current ratings (column 1 of Table 15). By contrast, the sheer volume of Tweets 24 hours before the show accounts for only 6.5% of the variation (column 2). Similarly, Google searches, Wikipedia views, IMDB review count, and Huffington Post news articles in the previous 24 hours explain a variation of only 4.9% (column 3), 12.1% (column 4), 4.8% (column 5), and 4.8% (column 6), respectively. Moreover, when we combine the lagged rating and Tweet counts (column 7) or Google searches (column 8), the *R*-squared increases little from the *R*-squared that includes only the lagged ratings (column 1). This result implies that the pure "volume model" has little explanatory power. Sentiments such as the proportion of positive or negative Tweets are not much better than volume in predicting TV ratings. As shown in Models 9 and 10 in Table 15, with Tweet volume, the effects of positive and negative Tweets are not statistically significant. We also experiment with other functional forms of the sentiment variables, such as quadratic terms and exponential terms; the results are

Table 19 Prediction MAPE of Popular vs. Unpopular Shows

Show	M1 (Only lagged rating) MAPE	M14 (Twitter content) MAPE
Unpopular	0.2853	0.0563
Medium	0.1379	0.0578
Very popular	0.1035	0.0583
Total	0.1214	0.0577

Table 20 Prediction: NFL

Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
MAPE	0.18	0.34	0.48	0.47	0.49	0.16	0.11	0.12	0.32	0.11	0.09	0.09	0.07	0.07	0.07	0.06	0.05
MSE	0.12	0.24	0.35	0.34	0.35	0.11	0.07	0.08	0.23	0.08	0.06	0.06	0.04	0.05	0.04	0.04	0.03

qualitatively similar to the linear form. Therefore, the "sentiment model" is also not useful for predicting TV ratings.

However, the first four principal components from the Tweet *n*-gram features produce an *R*-squared of 0.756 (column 14), which is comparable with the R-squared of the model with only the lagged rating included. A better model fit is found when we combine PCs with the lagged rating (column 15) and both the lagged rating and Google searches (column 16). This result shows that the "Tweet content model" outperforms both the "Tweet volume model" and the "Tweet sentiment model" in predicting TV series ratings.

Similarly, when we use the LDA (Blei et al. 2003) technique to extract the content information, we obtain an *R*-squared of 0.620 (column 11), which is close to the result that we obtain in the *n*-gram PCA case. Adding the lagged rating and other online platform information improves the $R2$ (columns 12 and 13). However, *n*-gram PCA performs better overall than LDA (we confirm this result in §5.3).

We also find marginal improvement in the *R*-squared by adding Google searches, Wikipedia views, IMDB reviews, and Huffington Post news. This result suggests that the information derived from Google searches (or Wikipedia views, IMDB reviews, and Huffington Post news) and Twitter posts may be mostly substitutes.

Table 15 considers TGWIH measures 24 hours before the new TV shows start. Will it help to involve more information by extending the data collection period to 48 hours? Comparing the *R*-squareds from Table 16 with the *R*-squareds from Table 15, we find that adding more information in TGWIH increases the model fit, but the increments are very small. For example, if we compare column 8 in Table 16 with column 8 in Table 15, the *R*-squared improves by less than 2%. This result implies that consumer buzz on online platforms about TV shows is highly transient. Information value rapidly deteriorates with time. Most of the other findings from Table 15 are replicated in Table 16.

Table 21 Prediction Comparison with LDA

MAPE (Model/Feature)	Cloud (Hadoop) PCA	LDA
Dynamic panel linear model	0.0577	0.0595
Auto regression *X*	0.0613	0.0622
Multi-layered feedforward neural networks	0.0664	0.0681
Self-organizing fuzzy neural network (SOFNN) model	0.0640	0.0679

As to the NFL (Table 17), we find that the lagged rating is not a good predictor of the current rating ($R2 = 0.248$). The reason for this result is likely because the size of the fan base changes because the teams that play change every week. For example, if last week's SNF game was between the Dallas Cowboys and the Pittsburgh Steelers, the ratings would be much higher than this week's SNF game between the Jacksonville Jaguars and the Tennessee Titans.

Instead, when we use the number of Tweets related to two teams 24 hours before the game begins (Table 17, column 2), the $R2$ becomes 0.34, which is much higher than when we use the lagged rating as the only explanatory variable (column 1). Surprisingly, the number of Google searches, Wikipedia views, and Huffington Post news articles related to the two teams 24 hours before the game starts can explain variations in the rating of only 4%, 6%, and 6%, respectively; the estimated coefficients are not significant.

To resolve the problem of changing teams, we add team dummies (home and away separately) in the sixth specification of the model. The resulting $R2$ of 0.69 supports our conjecture that the size of the fan base is an important determinant of ratings. In columns 7 and 8, we combine the lagged rating with the team dummies and Tweets or Google searches. Together, this combination can explain approximately 78% of the variation in the ratings.

We confirm that the content analysis (using *n*-gram PCs or topics as features) of Tweets is a powerful predictor of TV ratings. Most strikingly, the good model fits remain even after we remove the team dummies. This is shown in columns 11, 14, and 17 of Table 17. Topics and PCs alone can explain variations in the ratings of 82% and 85%, respectively. If we combine all of the information, including the lagged rating, Tweets, Google searches, Wikipedia views, Huffington Post news articles, and PCs, almost 90% of the variation can be explained.

This result indicates that team-specific Tweets that capture consumers' intention to watch the games are the lead indicators of actual future consumption.

Table 22 Computational Time Comparison

	Cloud (Hadoop) PCA	LDA
Time (Minutes)	6.2	36.8

5.3. Forecasting

After the model is calibrated, we test how well it can be used to forecast TV ratings. For this, we use Nielsen ratings data from September to November 2013 as the test sample. These data include 411 episodes[30] of the 30 TV series and 39 NFL games. To test the model's performance, we use two measures, i.e., (1) the mean absolute percentage error (MAPE) and (2) the mean squared error (MSE).

Largely consistent with the results in Tables 15 and 16, Table 18 demonstrates that without content analysis (Models 2 to 6), the mere volume of TGWIH performs much worse in predicting ratings than using content analysis (Models 11 to 16).

Interestingly, we find that the content information from Tweets can still adequately forecast ratings for obscure titles provided that the predictive power of the lagged rating is greatly decreased (Table 19). Specifically, for the five unpopular shows that did not last more than two seasons,[31] including Allen Gregory, Charlie's Angels, Hellcats, Harry's Law, and Gary Unmarried, we found that the prediction MAPE for the model that includes only the lagged ratings is significantly larger than the prediction MAPE for the other popular shows. However, the prediction MAPE for the model that includes Twitter content information is not much different from the prediction MAPE of the popular shows.

Similarly, Models 11 through 17 use content-selected Tweets and have the smallest prediction errors for the NFL sample (Table 20).

Next, we show the comparative advantage of our model. As shown in Table 21, our Dynamic Panel Linear Model with Cloud PCA outperforms the alternative models (including Auto Regression with Exogenous Independent Variables (Autoregression X), Multilayered feedforward neural networks, and the SOFNN models) in terms of out-of-sample prediction accuracy.

Moreover, cloud-based PCA is much faster than LDA. As shown in Table 22, cloud-based PCA takes approximately six minutes for the TV series task, whereas LDA takes approximately 37 minutes. This result shows that cloud computing can make computation more efficient, which can help advertisers predict demand faster, potentially in real time.

6. Conclusions

Our paper shows that easily accessible online information such as Twitter Tweets, Google Trends, Wikipedia views, IMDB reviews, and Huffington Post news can be useful for marketers to accurately predict consumer demand for TV shows. We demonstrate the power of using machine learning, text mining, and cloud computing techniques to process large-scale unstructured data to conduct a structured econometric analysis. We conclude that information from online platforms, if carefully extracted and sorted, can provide a timely representation of consumer intentions. These results have important implications for forecasting purchases/consumption that should be of great interest to all firms. We acknowledge that the industry is leading marketing academia in conducting cloud analytics in the context of TV viewing.[32] For example, Netflix is analyzing petabytes of data to recommend TV shows to consumers, optimize playback quality, and identify poorly translated subtitles.[33] The cloud tools used by Netflix go beyond the simple MapReduce programming model and include Hive, Pig, Parquet, Presto, Spark, etc.[34] Future marketing research that requires large-scale data analytics may consider adopting these tools.

Our paper has certain limitations. First, the real mechanism between online behavior and offline consumption is not revealed. Our study is not based on a well grounded theory that explains the entire path of a consumer's consumption experience. Thus, caution should be exercised in interpreting the results. Second, although Twitter has a wide user base, it is relatively more appealing to the young and urban demographic group, which is different from the general U.S. population. For example, our data show a discrepancy in the rankings of a TV series based on the volume of Tweets and ratings. NCIS has the highest average ratings but the lowest number of Tweets. This result is likely because of a mismatch between the Twitter user population and the fan base for the show. This limitation constrains Twitter's predictive power for consumption that is targeted toward other demographics. Finally, we predict the most popular TV shows, which have relatively stable ratings patterns. Predicting low-rated shows or newly debuted shows may pose significant additional challenges.

Our paper is only a first step in using consumers' online activities to predict offline consumption. Future research may consider gathering information from more consumer access points to predict demand and for other noninformation goods. More heterogeneous or location-specific analyses can also be performed to predict demand for certain demographics or in a specific local market. Methodologically, other text mining methods could be developed to extract the most useful information for predictions. Another promising

[30] The shows that have already ended are not included, such as Breaking Bad and Allen Gregory.

[31] We use the last season of these obscure titles for the prediction.

[32] We thank the anonymous reviewer for emphasizing this point.

[33] https://gigaom.com/2015/01/21/netflix-is-open-sourcing-tools-for-analyzing-data-in-hadoop/.

[34] http://strataconf.com/big-data-conference-ca-2015/public/schedule/detail/38649.

venue to explore is to disrupt the current Nielsen Rating system and replace it with a real-time measure of audience size/composition based on Twitter or Facebook conversations. Researchers may also want to link TV advertising to online platforms to accurately measure consumer responses to TV ads in real time. We hope that future research will address these issues.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2015.0972.

References

Andersen TG, Sørensen BE (1996) GMM estimation of a stochastic volatility model: A Monte Carlo study. *J. Bus. Econom. Statist.* 14(3):328–352.

Andrews DW, Lu B (2001) Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econometrics* 101(1):123–164.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econom. Stud.* 58(2):277–297.

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. *J. Machine Learning Res.* 3:993–1022.

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J. Comput. Sci.* 2(1):1–8.

Bond SR (2002) Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Econom. J.* 1(2):141–162.

Bowsher CG (2002) On testing overidentifying restrictions in dynamic panel data models. *Econom. Lett.* 77(2):211–220.

Bureau of Labor Statistics (BLS) (2011) American time use survey, Table 1. http://www.bls.gov/news.release/atus.t01.htm.

Cattell R (1966) The Scree test for the number of factors. *Multivariate Behavioral Res.* 1(2):245–276.

Chakravarty A, Liu Y, Mazumdar T (2010) The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation. *J. Interactive Marketing* 24(3):185–197.

Chen Y, Wang Q, Xie J (2011) Online social interactions: A natural experiment on word of mouth versus observational learning. *J. Marketing Res.* 48(2):238–254.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Cohen N (2014) Wikipedia vs. the small screen. *New York Times* (February 9).

Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the Web. *Management Sci.* 53(9):1375–1388.

Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.

Dewan S, Ramaprasad J (2012) Research note—Music blogging, online sampling, and the long tail. *Inform. Systems Res.* 23(3-Part-2): 1056–1067.

Dhar V, Chang EA (2009) Does chatter matter? The impact of user-generated content on music sales. *J. Interactive Marketing* 23(4):300–307.

Eliashberg J, Hui SK, Zhang ZJ (2007) From story line to box office: A new approach for green-lighting movie scripts. *Management Sci.* 53(6):881–893.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Ghose A, Ipeirotis PG, Sundararajan A (2007) Opinion mining using econometrics: A case study on reputation systems. *Proc. 45th Annual Meeting-Assoc. Comput. Linguistics* 45:416–423.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Gopinath S, Chintagunta PK, Venkataraman S (2013) Blogs, advertising, and local-market movie box office performance. *Management Sci.* 59(12):2635–2654.

Gopinath S, Thomas JS, Krishnamurthi L (2014) Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Sci.* 33(2):241–258.

Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. *Proc. Eleventh ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 78–87.

Halko NP (2012) Randomized methods for computing low-rank approximations of matrices. Unpublished doctoral dissertation, University of Colorado, Boulder.

Hu M, Liu B (2004) Mining and summarizing customer reviews. *Proc. Tenth ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 168–177.

Karniouchina EV (2011) Impact of star and movie buzz on motion picture distribution and box office revenue. *Internat. J. Res. Marketing* 28(1):62–74.

Lee D, Hosanagar K, Nair HS, Stanford GSB (2013) The effect of advertising content on consumer engagement: Evidence from Facebook. Working paper, Carnegie Mellon University, Pittsburgh.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Liu Y, Huang X, An A, Yu X (2007) ARSA: A sentiment-aware model for predicting sales performance using blogs. *Proc. 30th Ann. Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 607–614.

Mestyán M, Yasseri T, Kertész J (2013) Early prediction of movie box office success based on Wikipedia activity big data. *PLoS One* 8(8):e71226.

Mishne G, Glance NS (2006) Predicting movie sales from blogger sentiment. *AAAI Spring Sympos.: Comput. Approaches Analyzing Weblogs*, 155–158.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.

Nielsen (2011) State of the media 2010: U.S. audiences and devices. http://blog.nielsen.com/nielsenwire/wp-content/uploads/2011/01/nielsen-media-fact-sheet-jan-11.pdf.

New York Times (2015) Study of TV viewers backs Twitter's claims to be barometer of public mood. (March 8), http://bits.blogs.nytimes.com/2015/03/08/twitter-chatter-reveals-what-we-like-to-watch/?_r=0.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Nickell S (1981) Biases in dynamic models with fixed effects. *Econometrica: J. Econometric Soc.* 49(6):1417–1426.

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: Linking text sentiment to public opinion time series. *Proc. Fourth Internat. AAAI Conf. Weblogs Social Media*, 122–129.

Onishi H, Manchanda P (2012) Marketing activity, blogging and sales. *Internat. J. Res. Marketing* 29(3):221–234.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Foundations Trends Inform. Retrieval* 2(1–2):1–135.

Paul MJ, Dredze M (2011) You are what you tweet: Analyzing Twitter for public health. *Proc. Fifth Internat. AAAI Conf. Weblogs Social Media*, 265–272.

Pauwels H, Stacey E, Lackman A (2013) Beyond likes and tweets: Marketing, online platforms content, and store performance. MSI Report.

Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philosophical Trans. Roy. Soc. London A: Math., Physical Engrg. Sci.* 368(1933):5707–5719.

Roodman D (2009) How to do xtabond2: An introduction to difference and system GMM in Stata. *Stata J.* 9(1):86–136.

Sadikov SE, Parameswaran AG, Venetis P (2009) Blogs as predictors of movie success. *Proc. Third Internat. AAAI Conf. Weblogs Social Media*, 304–307.

Sinha S, Dyer C, Gimpel K, Smith NA (2013) Predicting the NFL using Twitter.

Stephen AT, Galak J (2012) The effects of traditional and social earned media on sales: A study of a microlending marketplace. *J. Marketing Res.* 49(5):624–639.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proc. Fourth Internat. AAAI Conf. Weblogs Social Media*, 178–185.