Modeling Spread of Word of Mouth on Twitter

by

Xiaoyu Zhang

B.A. Psychology, Franklin & Marshall College, 2011

M.A. Psychology, Harvard University, 2013

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN TRANSPORTATION

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2015

Signature of Author .... Signature redacted

Department of Civil and Environmental Engineering

May 21, 2015

Certified by .... Signature redacted

David Simchi-Levi

Professor of Civil and Environmental Engineering and Engineering Systems

Thesis Supervisor

Accepted by .... Signature redacted

Heidi Nepf

Donald and Martha Harleman Professor of Civil and Environmental Engineering

Chair, Graduate Program Committee

Modeling Spread of Word of Mouth on Twitter

by

Xiaoyu Zhang

Submitted to the Department of Civil and Environmental Engineering
on May 21, 2015, in partial fulfillment of
the requirements for the degree of
Master of Science in Transportation

Abstract

Twitter is a popular word-of-mouth microblogging and online social networking service. Our study investigates the diffusion pattern of the number of mentions, or the number of times a topic is mentioned on Twitter, in order to provide a better understanding of its social impacts, including how it may be used in marketing and public relations.

After an extensive literature review on diffusion models and theories, we chose the Bass diffusion model, because it allows us to achieve a relatively good estimation for the diffusion pattern of a trending topic. Furthermore, we extend the Bass model in two ways: (1) incorporating the number of mentions from influential users on Twitter; (2) aggregating the hourly data observations into daily data observations. Both extensions significantly improve the model's ability to predict the total number of mentions and the time of highest mentions.

In the future, we hope to extend the applications of our study by incorporating external data from the news and other sources, to provide more comprehensive information about what people are saying and thinking. We also hope to analyze the data in terms of demographics and user networks, to potentially predict everything from new product introduction to conversations about defective products.

Thesis Supervisor: David Simchi-Levi

Title: Professor of Civil and Environmental Engineering and Engineering Systems

Acknowledgement

First, I would like to thank my advisor, Professor David Simchi-Levi. Throughout my thesis, he has given me a great amount of guidance and encouragements. He is a celebrated researcher, supportive advisor, and exceptional role model. He opened the world of revenue management to me and has provided me an example of working hard toward what he loves doing.

I would also like to thank my friends at MIT and Harvard. Cambridge became a lovely place and a second home to me because of them. My special thanks to Yingzhen Shen, my friend and project partner. She has given me a great amount of help throughout my research, and made my years at MIT enjoyable and fruitful.

Last but not the least, I would like to thank my parents, for their unconditional support and understanding. I am also indebted to my boyfriend, for his love and inspiration. Words cannot express how grateful I am to them.

Contents

1. Introduction

Twitter is a microblogging and an online social networking service. Users post short 140-character messages called "tweets" or "root tweets" and re-post someone else's tweet called "retweets". They can also follow other users, who can see their root tweets and reply to them in retweets.

This tweeting and reply speaks to its role as a microblog, while following, retweeting and mention behavior speaks to its role as a social network (Thelwall, Buckley, and Patoglou, 2011). As one of the most popular platforms in the world, Twitter has 302 million monthly active users, with 500 million Tweets sent per day, according to the statistics listed on the Twitter official website in May, 2015.

Due to Twitter's popularity and its real-time nature, certain trending topics could become viral on Twitter very quickly (Ma, Sun, and Cong, 2013). It is of particular interest to marketing and public relations professionals and researchers to understand how these topics go viral, and such an understanding has the potential to significantly help firms predict the likelihood of their marketing campaign's success (Schultz, Utz, and Gritz, 2011).

In this project, we predict the spread of specific topics on Twitter. This is challenging, because of an overwhelming amount of information that is short and noisy (Li, Sun, and Datta, 2012). One helpful feature is the # symbol, called a hashtag, used to mark a keyword or topic in tweets. This symbol was created originally by Twitter users as a way to categorize messages. Most topics or events that we will model and discuss in this paper are based on hashtag events.

This thesis is organized in the following way. In the second section, we offer a literature review on relevant work, including 1) research on social media, 2) diffusion theory and models, and 3) social influence through online word of mouth. In the third section, we summarize our study and compare it with existing work on similar topics. We also discuss the fundamental differences between these studies. In the fourth section, we focus on the data we collected. We describe the data collection process and report results of our preliminary analysis.

In the fifth section, we present empirical analysis of the data, describing several model specifications, and estimating and evaluating these model performances.

In the last section, we discuss the contribution and limitations of our study. In addition, we suggest potential directions to explore in future research.

2. Literature Review

In this section, we will review existing works that are relevant to our study. First, we will focus on studies on Twitter and other social media in marketing and computer science areas. Second, in order to understand the diffusion of the word of mouth on Twitter, we discuss the existing diffusion theories and models, including the information cascading theory, as well as the Bass diffusion theory and its extended model. Finally, as word of mouth on Twitter is a specific case of online word of mouth, there might be some similarities or inspiring insights from other types of online word of mouth. Therefore, we review the mainstream literature on other types of online word of mouth in marketing science, especially on the social influence of customer reviews.

2.1 Study on Twitter and other Social Media

There has been a growing trend of social media research in the marketing area in the past few years. These studies cover many areas. Some research studies the posting behaviors of consumers (Berger and Milkman 2012), others focus on the impact of social influence (Aral and Walker, 2012) and the influence of social networks (Katona, Zubcsek and Sarvary, 2011). There are also studies that examine the influence of social media on various performance measures, such as product sales and stock prices (Schweidel and Moe, 2014).

In general, consumers have many social media venues on which they can express their opinions about products or firms. Schweidel and Moe (2014) reveal that the differences in venues formats (e.g., blog, forum, and microblog) will result in different posting behaviors on the social media. Additionally, Smith, Fischer and Chen (2012) find that brand-related sentiments are different across platforms including YouTube, Twitter and Facebook.

There are social media platforms that limit the length of posts (e.g., Twitter) and those that do not (e.g., discussion forums). As a result, people are more likely to post richer opinions on discussion forums and blogs. In contrast, people have to express themselves in a constrained number of words in microblogs and Twitter, where more extreme opinions can be found (Schweidel and Moe, 2014).

There is also research that studies how users' motivations vary across venue formats (Toubia and Stephen, 2013; Hsu and Lin, 2008; Yang et al., 2007). These studies suggest that one's motivation to participate in the social media activity varies across venue formats.

Toubia and Stephen (2013) examine users' different motivation when posting on Twitter. Their study focuses on noncommercial users who do not have financial incentives. There are two main types of utility motivations from previous literature: 1) intrinsic utility, which suggests that users post for their intrinsic satisfaction (Ryan and Deci 2000), and 2) image-related utility, which suggests that users post to influence others' perception and to seek status (Glazer and Konrad 1996). Because these two kinds of utility result in different predictions, Toubia and Stephen (2013) use that to understand whether users should increase their contributions when their number of followers increases. An experiment was conducted where researchers added followers to a group of users, and then compared their posting behaviors between the experimental group and the control group. A dynamic discrete choice model was applied to estimate each treated users' utility function. The results show that the majority of users have larger image-related utility.

In contrast, Yang et al. (2007) show that in discussion forums users are mostly driven by intrinsic motivations. Finally, Hsu and Lin (2008) show that on blogs, both intrinsic and image-related motivations can be found in the posting behavior.

Many computer scientists have also contributed to this area. Bauckhage and colleagues (2014) examine how the adoption of several social media services grows. They obtain the aggregate search frequencies for a specific social media service (e.g. Facebook, Twitter) on Google trend. For each time series, they apply economic diffusion models, including Gompertz, Bass, and Weibull models, and their model fitting is reported to be accurate.

At smaller granularity level, Ma et al. (2013) apply classic classification models to predict which topics on Twitter will turn popular in the next day. As the authors point out, because viral information are influential, it is very important to predict which Twitter topics will turn into the popular ones. Before predicting, Ma et al. categorize popularity into a few strength areas, including not popular, marginally popular, popular, very popular, and extremely popular. Then they change the original problem into an easier classification task. That is, they are predicting which level of popularity the event is in. Ma et al. applied several common classification models, including logistic regression, Naïve Bayes, k-nearest neighbors, support vector machines, and decision trees.

Ma et al. identified 7 content features from the tweets that have a specific hashtag and 11 contextual features from the social graph of the users that have adopted the hashtag. For example, the content features can be ContainingDigits (a variable that indicate whether a hashtag has digits) and SentimentVector (a variable that indicates the neutrality, positivity, or negativity associated with the tweet), whereas the contextual features can be UserCount (number of Twitter users in the social graph), TweetsNum (number of tweets in the social graph), and ReplyFrac (percentage of tweets that are replies to the root tweet).

Their experiments are conducted on a very large data set including millions of tweets and users. The classifiers using Ma et al.'s chosen features perform significantly better as compared to the baseline models not using these features. In addition, the authors demonstrate that the contextual features have more predictive power than the content ones, and that the logistic regression model shows the best performance among the five classification models in the study.

2.2 Diffusion Theory and Models

At the individual level, the diffusion of event information is close to information cascading. In the next few paragarphs we describe the concept of information cascade. An information cascade happens when people observe others' behavior and make the same choice as the others, abandoning their own information or private signal (Bikhchandani et al., 1996). In the Twitter event diffusion process, people who observe others tweeting about the event will also retweet the tweet, which contributes to the total number of mentions and lets their friends observe their behavior.

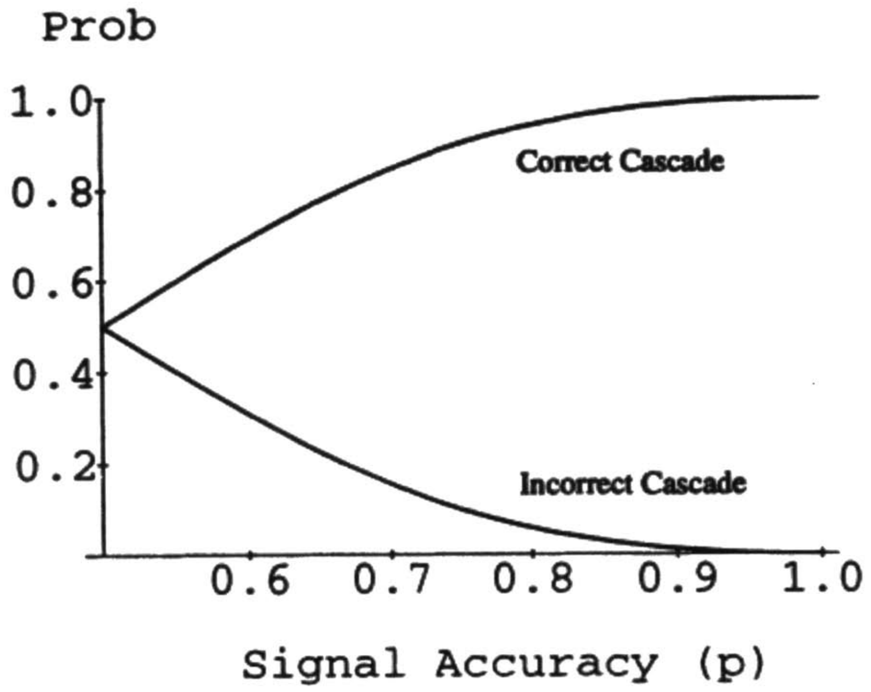
The research on information cascade originates from Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1992). These two groups of researchers demonstrate separately that the information cascades will inevitably occur. Here we briefly review the paper by Bikhchandani, Hirshleifer and Welch (1992).

Bikhchandani et al. consider the mass behavior to be delicate, and that small shocks can result in salient changes in mass behavior. To capture this, the authors present a model, in which there are consecutive individual arrivals. Everyone needs to decide between adopting and rejecting some behavior, and they could see decisions by the people who are in front of them. The sequence of individuals is public information and exogenous. For adopting, the cost is the same, which for now set to 1/2. The gain is the same as well, and is set to be either 0 or 1, both with half probability. Each individual has a conditional private signal about the value, which is either high (H) or low (L). The high signal is observed more often than low if true value is 1 (p>1/2) and otherwise 1-p. For the first person, if his signal happens to be H, he will choose to adopt and vice versa. For the second person, he could speculate about the signal of the first individual using the decision of the first person: if the first person adopted, then it is H and otherwise it is L. If the first person adopted and the second person has signal L, the second person will have equal probability of each decision. If the first person rejected (therefore inferred value would be L), the second person will reject if he has L. Otherwise, the second person will adopt half of the time. The process continues, and according to Bikhchandani et al., the fourth person will end up in the same situation as the second person, fifth following the same situation with the third person, etc.

Therefore, such cascading behavior prohibits the possibility that multiple individual collectively gather information. Once a cascade starts, the following individual's private signal does not matter and therefore, his information will not enhance the later decision. In the ideal situation, if we can aggregate information from all previous people, the following people will converge to behave flawlessly. The following figure in the original paper show the idea of information cascading. Observing from the data, even if we have a large correct signal (p), we could still have a considerable probability that ends up in incorrect cascade. For example, with p equaling 0.8, we still have around 10% probability ending up in wrong cascading.

This information cascading theory could explain the spread of epidemics and information, and explain why sometimes the wrong information such as rumor could spread through online or offline network.

Figure 2.1 Information Cascading



While the diffusion of event can be modeled by information cascading theory at the individual level, we think that at the aggregate level, the diffusion process of event or topic on Twitter is close to the new product adoption process. In our opinion, those who retweet are similar to the imitators who observe other people's adoption. Those who initiate the tweet independently or under outside influence are similar to the innovators from the new product adoption perspective.

Many studies have shown that natural growth of many events follows an S-shaped pattern (Meade and Islam, 1998). The examples include infectious disease diffusion, innovation adoption, and number of future sales of durable products (Radas, 2006). The diffusion theory reveals why there is an S-shaped pattern.

As Radas (2006) explains, diffusion model is a traditional tool in modeling a brand new durable product's lifecycle fluctuation. In addition, it is been used in predicting new products' demands. The Diffusion model describes the change of a new product's sales over time. Among numerous external influences, the product's price change and the advertising level of the product are two of the most important ones that influence the sales of the product.

To give a clear picture of the diffusion model, we here review the original Bass model (Bass, 1969) and a few other extensions of the original Bass model. In 1969, Bass suggested that the probability of current purchase is a function of cumulative number of previous purchase. If we use $f(t)$ as the probability density function of adoption at time t, and $F(t)$ as the cumulative distribution function (probability of adoption by time t), we have

$$\frac{f(t)}{1 - F(t)} = p + qF(t) \qquad (2.1)$$

In equation (2.1), parameters p serves as the coefficient of innovation and q serves as imitation. According to Bass, we could think of coefficient of innovation as the external influence of product adoption, and coefficient of imitation as the interpersonal communication.

Because $F(t)$ is differentiable, we have

$$f(t) = \frac{dF(t)}{dt} \qquad (2.2)$$

Substitute equation (2.2) into equation (2.1), we could rewrite equation (2.1) as

$$\frac{dF(t)}{dt} = p + (q - p)F(t) - qF(t) \qquad (2.3)$$

Let m denote total number of potential buyers of the product, $n(t), N(t)$ represent sales and cumulative sales of the new product respectively at time t. We assume that

$$n(t) = mf(t) \qquad (2.4)$$

$$N(t) = mF(t) \qquad (2.5)$$

Substituting above two equations into equation (2.3), we have

$$n(t) = pm + (q - p)N(t) - \frac{q}{m}N(t)^2 \qquad (2.6)$$

As Radas (2006) points out, there are two basic shapes for the Bass model. When the coefficient of innovation is smaller as compared to the coefficient of imitation, the curve is a bell shape. When the coefficient of innovation is larger than the coefficient of imitation, the shape is downward sloping.

However, for real sales data of durable products, sometimes we observe jumps and sharp curves that do not conform to the classic Bass model. These jumps and sharp curves come from the external influences. Therefore, it is necessary to incorporate external variables such as level of advertising and price into the model.

As stated above, one major limitation of the original Bass model is that it does not incorporate external influences into the model such as marketing mix variables. In response, many researchers have extended the original Bass model by adding marketing mix variables into the model. For a good summary of these types of models, see Table 1 in Radas (2006).

In 1975, Robinson and Lakhani (1975) first incorporate new decision variables with the original Bass model. Their model introduces the price of product into the original Bass expression:

$$\frac{dF(t)}{dt} = [p + (q - p)mF(t) - qmF(t)^2]e^{-kPr(t)} \qquad (2.7)$$

In the above model, $Pr(t)$ represents the price at time t and k is the coefficient. Other variables are the same as in the original Bass model. Subsequently, Bass (1980) did a similar thing by introducing price as a new term, with a slightly easier estimation method.

Afterwards, Bass, Jain and Krishnan (1994) developed the generalized Bass model. The model is

$$\frac{dF(t)}{dt} = [p + (q - p)mF(t) - qmF(t)^2]x(t) \qquad (2.8)$$

The variable in the right-most captures the external influence from price and advertising.

$$x(t) = 1 + \beta_1 \frac{Pr'(t)}{\mathrm{Pr}(t)} + \beta_2 \frac{A'(t)}{\mathrm{A}(t)} \qquad (2.9)$$

$\mathrm{Pr}(t)$ represents the price at time t, and $Pr'(t)$ represents the rate of change in price at time t. $\mathrm{A}(t)$ means the advertising at time t, and $A'(t)$ means the rate of change in advertising at time t. β_1 and β_2 are coefficients.

2.3 Social Influence through Online Word of Mouth

Besides understanding studies on social media and word of mouth diffusion, we are also interested on how word of mouth can make social influence. The objective of our study is to understand the diffusion of word of mouth on Twitter, and it is an initial and important step of understanding how it will make social impact and be of real-world use. In marketing area where the impact of online word of mouth has been recognized very early, there have been many studies on social influence through all kinds of online word of mouth. Notably, online customer review is one of most well studied types of online word of mouth, as it provides the consumers with information about product quality (Chevalier and Mayzlin, 2006). In this subsection, we will review some work on the topic of social influence through online word of mouth.

There are a great number of studies done on social influence of online word of mouth. Some of them focus on impact of online word of mouth on firm performance (Chevalier and Mayzlin, 2006; Anderson and Simester, 2014). Others study how the firm will engage or respond to these influence (Mayzlin 2006; Dellarocas 2006).

Chevalier and Mayzlin (2006) use scrapped data from two largest online book retailers to study how the book reviews impact on sales of books. The author analyzed customer reviews on Amazon and on Barnes and Noble, and concluded that Amazon has more reviews and their reviews are longer, and that the majority of reviews on both websites are positive. Their results also suggest that the sales of a book is positively correlated with an improvement of a book's review (e.g., number of reviews, average star ranking of the reviews).

On the other hand, Anderson and Simester (2014) work on deceptive reviews. They reveal that a considerable number of product reviews are from individuals who have not even bought the product, and these reviews are found to be more negative. The study of Anderson and Simester suggest that the reason for people to post such deceptive reviews can be either seeking social status, or behave like self-appointed product managers.

Besides empirical work in online word of mouth, there are also theoretical papers such as Mayzlin (2006) and Dellarocas (2006). Mayzlin (2006) modeled conditions where reviews, or online word of mouth, influence consumer choice. Ideally, reviews are all from consumers who have used the product. However, the company of the product as well as its competing companies also post reviews, which is hard to identify from the reviews from real customers. Mayzlin demonstrate that the companies with low quality products tend to spend more expense on promotional reviews, and she derive the welfare loss in the system because of such behaviors.

3. Our Study and Related Works

In this section, we first describe what we are going to do in the study. Then we will discuss the difference between our study and the existing studies.

The objective of our study is to understand how mentions of certain trending topic or event accumulate on Twitter. We define the mentions to be either root tweet, retweet or reply on Twitter. Our data also includes mentions from influential Twitter users. More detailed discussion of the data will be shown in the following section. More specifically, we want to predict when the peak of mentions will occur and what the final cumulative mentions would be. We found similarities between our data pattern and data patterns in the studies of new product adoption, and thus we decided to use the Bass model to model the spread of word of mouth on Twitter by event.

To distinguish from other similar studies, we will emphasize features of our study in two aspects. First, our identification of the event is through identifying a certain fixed string, such as “Boston Marathon”. The string could be a hashtag such as #mynypd, although a hashtag is not required. In addition, each time, we are studying the diffusion of one event. Thus, it is different from the work that predicts diffusion or adoption of hashtag (Chang, 2010).

Chang (2010) applies various diffusion models on Twitter hashtag adoption. In their study, the hashtag is treated as a new product, which is different from what we are doing. Chang also points out the importance of hashtag, as it has become a unique tagging tool to help link Twitter messages with certain topics or events.

Second, our study does not use individual behavior data. It is not our objective to study the micro-behavior of retweeting (Suh, Hong, Pirolli, and Chi, 2010) or to study what degree of popularity the event will be (Ma et al, 2013; Zaman, 2014). Those works usually have detailed information of each mention, including content and features of social network users who post these mentions, as well as the network structure information. With those data, the prediction of the next stage information diffusion could be formulated as a standard Bayesian model (Zaman, 2014).

In contrast, our aim of the study is to predict the diffusion at the aggregate level. Importantly, the next stage value such as cumulative mention does not have a clear Bayesian relationship with previous stage value. Thus, our task is more difficult. In addition, because our data is at the aggregate level, our diffusion pattern or data curve is likely to be influenced or mainly driven by external force that we do not know about.

Because in in the Twitter network, retweeting is an essential component of information diffusion, Suh, Hong, Pirolli, and Chi (2010) study retweeting behavior. In particular, they examine why some tweets spread more extensively than other information. In their research, a number of features are examined about how they might affect retweetability of tweets. A total of 74M tweets are gathered to extract content and contextual features, which are used to identify significant factors for retweet rate. The authors develop a predictive model, and find that retweetability is strongly dependent on both URLs and hashtags, which are among content feature. For contextual features, the factors that have an impact on retweetability include the number of followers, the number of followees, and the age of the account. However, the number of past tweets is not a good predictor for the tweet uretweetability of a user.

Ma and colleagues (2013) suggest methods to predict the popularity of a new hashtag. They categorize popularity into several strength areas on Twitter, and then change the prediction problem into a classification one. In other words, they are predicting whether the event falls into the categories named not popular, marginally popular, popular, very popular, or extremely popular.

Zaman, Fox, and Bradlow (2014) build a theoretical model to predict the spread of an individual Tweet on Twitter. Using a probabilistic model, Zaman et al. forecast the final total of retweets with 52 root Tweets and achieve more accurate results as compared to existing models such as the dynamic poison model and regression model. Their model follows a Bayesian approach and can predict the spread of an individual Tweet with only the retweet times and the network structure of the Twitter users involved. Their study has potential implication for understanding the spread of broader topics and trends, and is thus relevant to our research of predicting the spread of a topic in Twitter.

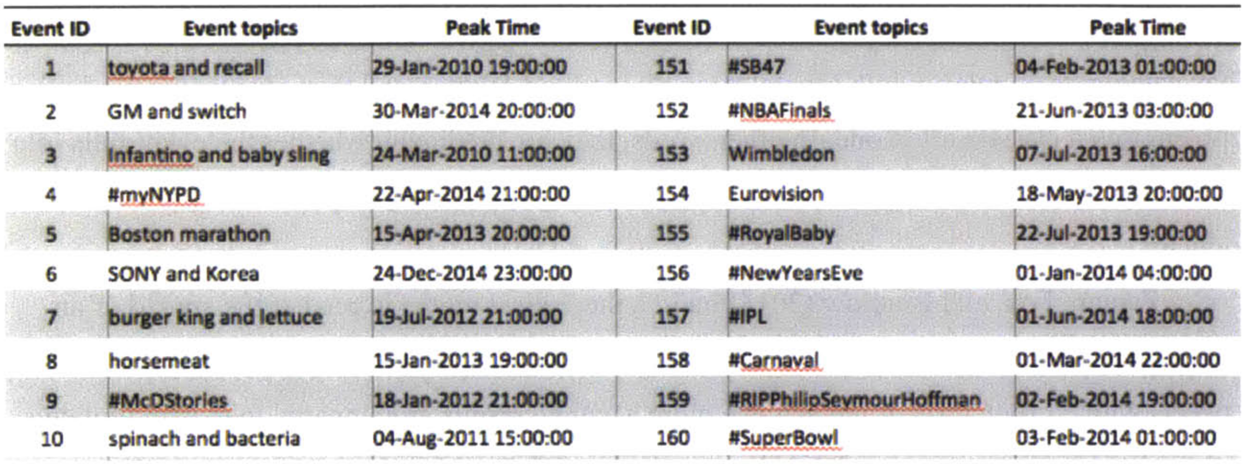
These previous two factors distinguish our study from existing studies, but at the same time these issues limit our model performances and provide us with new research directions. We will have detailed discussion about limitation and future research direction in the last section.

4. Data and Preliminary Analysis

4.1 Data Collection

We collected information of 220 events on Twitter by scraping Topsy Application Program Interface (API) in March 2015. Topsy is a Twitter partner and a certified data reseller. We wrote MATLAB code (See Appendix) to obtain data in JSON format and decode it using json parser (obtained from http://www.json.org/). Most topics keywords of the 220 events are Twitter's self-selected most popular moments of 2010, 2012, 2013, and 2014. Additional topics are selected by researchers, such as "“Toyota + recall” and "horsemeat". The topics include defect in product, terrorist attack, food poison, new product introduction, and miscellaneous important or popular events. A sample of the event keywords can be found in the table below.

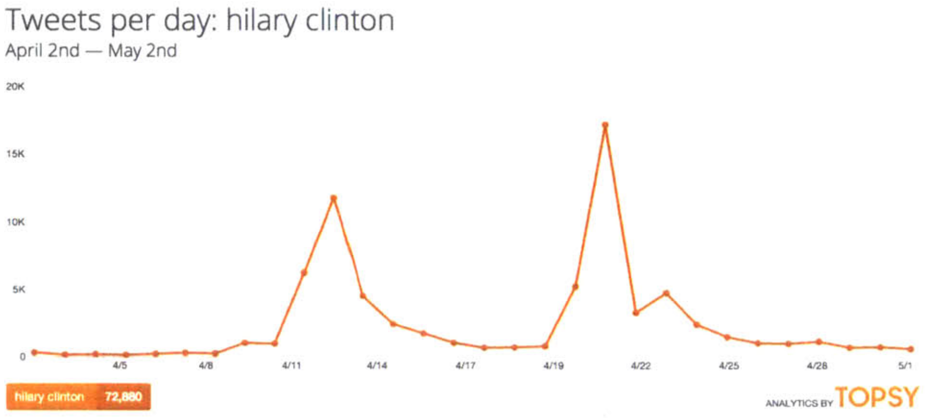
Figure 4.1 Selected Event Descriptions



Event ID	Event topics	Peak Time	Event ID	Event topics	Peak Time
1	toyota and recall	29-Jan-2010 19:00:00	151	#SB47	04-Feb-2013 01:00:00
2	GM and switch	30-Mar-2014 20:00:00	152	#NBAFinals	21-Jun-2013 03:00:00
3	Infantino and baby sling	24-Mar-2010 11:00:00	153	Wimbledon	07-Jul-2013 16:00:00
4	#myNYPD	22-Apr-2014 21:00:00	154	Eurovision	18-May-2013 20:00:00
5	Boston marathon	15-Apr-2013 20:00:00	155	#RoyalBaby	22-Jul-2013 19:00:00
6	SONY and Korea	24-Dec-2014 23:00:00	156	#NewYearsEve	01-Jan-2014 04:00:00
7	burger king and lettuce	19-Jul-2012 21:00:00	157	#IPL	01-Jun-2014 18:00:00
8	horsemeat	15-Jan-2013 19:00:00	158	#Carnaval	01-Mar-2014 22:00:00
9	#McDStories	18-Jan-2012 21:00:00	159	#RIPPhilipSeymourHoffman	02-Feb-2014 19:00:00
10	spinach and bacteria	04-Aug-2011 15:00:00	160	#SuperBowl	03-Feb-2014 01:00:00

Topsy (http://topsy.com/) is a platform to search information about a topic or existing tweet via keywords. For example, if we wanted to know the number of tweets that mention “Hilary Clinton” per day over the past month, we could search on Topsy, and below is a figure that we obtained. We could see from the figure that the total number of mentions from April 2, 2015 to May 2, 2015 is 72880, and the first peak in the figure occurred at April 12, 2015, when Hilary Clinton announced that she would seek the presidential nomination for the 2016 election.

Figure 4.2 Tweets Per Day



The advantage of using Topsy API (http://api.topsy.com/) is that Topsy API provides multiple metrics (statistics derived from analyzing the full Twitter Firehose) available in minute, hour, or day granularity. We also looked at other data sources including Keyhole (http://keyhole.co) and Twitonomy (https://www.twitonomy.com/), and we decided to use Topsy API because unlike other sources, there is no cost incurred during data collection. Examples of metrics offered by Topsy API are:

1. *mentions*: number of tweet mentions by time slice for any topic
2. citations: number of total citations (*root tweets, retweets,* and *replies*) for a particular topic
3. impressions: number of potential impressions by time slice for any topic
4. *sentiment*: Topsy Sentiment Score (0-100) by time slice for any topic. A score of 50 means neutral, 0 means highly negative, and 100 means highly positive.
5. *mentions by influential user only*: number of tweet mentions by time slice for any topic from only influential Twitter users. Influence is a score from 0 to 10 computed by Topsy. An influence score of 8 and greater are defined as influential users.

For each of the 220 events, we first found the peak time for the number of total mentions. For our data collection purpose, we defined the start time of the event to be 4 months prior to the peak time and the end time of the event to be 2 months after the peak time. For each event, we collected 6 variables: *unix_timestamp*, *mentions*, *sentiment_score*, *root tweet, retweet, reply*, and *mentions by influential users only*. We collect one data point of each variable per hour, and for each event, we have 4320 data points for each variable because the duration of these events are all 6 months. For clarity, we list the definitions of the 6 variables again:

1. *unix_timestamp:* Unix time can be conveniently convert to human readable date and time, and is easier to track in programming script.
2. *mentions:* number of tweet mentions by time slice for any topic
3. *sentiment_score:* Topsy Sentiment Score (0-100) by time slice for any topic
4. *root tweet:* number of tweet mentions by time slice for any topic for only the root tweet
5. *retweet:* number of tweet mentions by time slice for any topic for only the retweet
6. *reply:* number of tweet mentions by time slice for any topic for only the reply
7. *influential users only:* number of tweet mentions by time slice for any topic from only influential Twitter users

Table 4.1 Selected Event Data

unix_timestamp	mentions	sentiment_score	root tweet	retweet	reply	influential users only
1358762400	14	59	11	2	1	3
1358766000	40	62	28	11	1	2
1358769600	29	92	22	6	1	2
1358773200	40	89	28	10	2	3
1358776800	80	92	72	8	0	5
1358780400	557	92	100	454	3	12
1358784000	728	94	370	353	5	24
1358787600	629	94	376	237	16	27
1358791200	180	92	128	46	6	8

The above table shows a slice of the data for one event. The data for each event is output into one sheet of an Excel document.

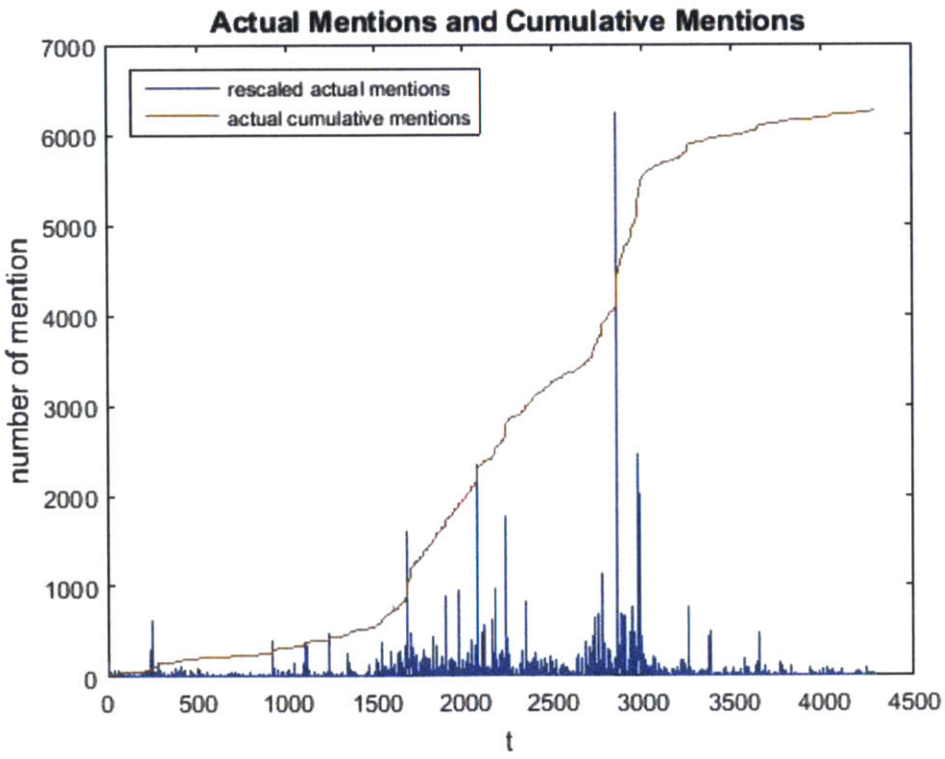
4.2 Preliminary Analysis

Although we have collected data of many events, we will only show our Bass model analysis on one typical event. The event name and keyword we used is IPL. The majority of other events we collected follow the same pattern with the one we show here.

The Indian Premier League (IPL) is an annual tournament for Twenty20 (T20) cricket since 2008. T20 cricket is a shorten form of cricket. The IPL is the most watched and highest paid T20 league in the world. The media says that IPL is massively popular in India, where cricket is like a religion to its people. IPL starts from April and ends in June, and because of its massive popularity, it became the first sports event that has a live broadcast on Youtube.

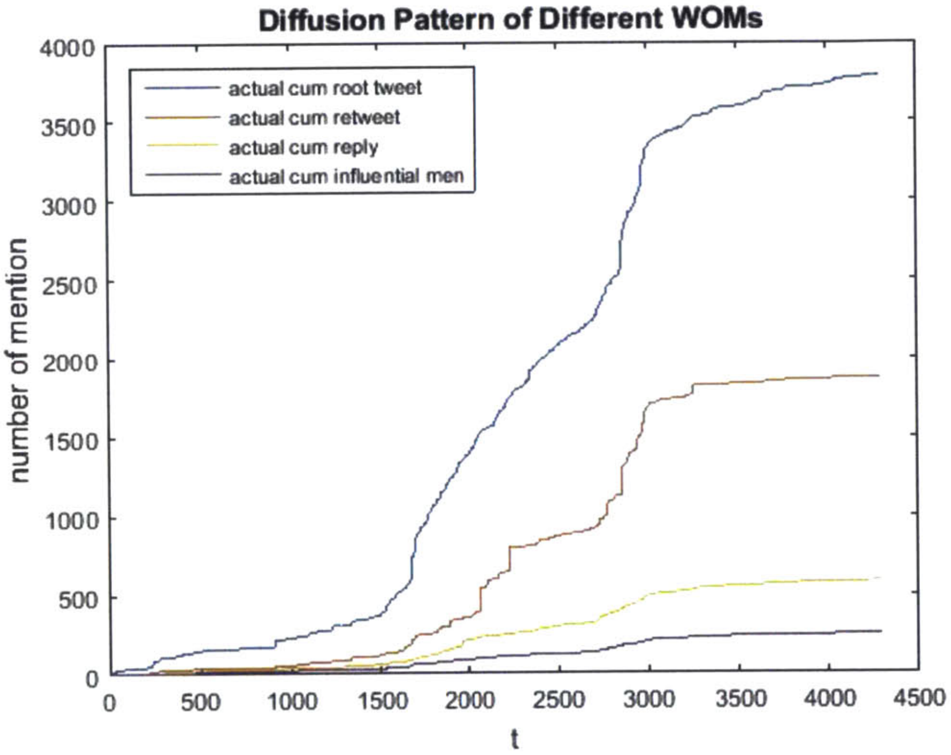
The following figure shows the cumulative mentions (in red) and individual/hourly mentions (in blue). Since the individual mention is much fewer than cumulative mentions, we rescale the mention by multiplying the ratio of maximum of cumulative mentions and maximum of individual mentions. The x axis means the time. In total, we have 4320 time points (hours). The y axis means the number of mentions.

Figure 4.3 Actual Mentions and Cumulative Mentions



The following figure shows the cumulative numbers of root tweet (blue), retweet (red), replies (yellow) and cumulative mentions from influential users (purple).

Figure 4.4 Diffusion Pattern of Different WOMs



5. Model and Analysis

5.1 Basic Bass Model

First, we used basic Bass model to capture the diffusion process. As illustrated in the literature review section, we will have three parameters p, q, m which are respectively coefficient of innovation, coefficient of imitation and ultimate market potential. Let $n(t)$ denote mentions of the event at time t (hour), and $N(t)$ denote cumulative mentions of the event till time t (hour). We will have our benchmark model as follows.

$$n(t) = \left(p + \frac{q}{m}N(t)\right)\left(m - N(t)\right) \quad (5.1)$$

In order to estimate the coefficient, we could rewrite the above equation as

$$n(t) = pm + (q - p)N(t) - \frac{q}{m}N(t)^2 \quad (5.2)$$

If we consider $n(t)$ as dependent variable, $N(t)$ and $N(t)^2$ as independent variables, we could estimate the coefficients by fitting a linear regression.

5.2 Extended Bass Model

To improve the performance of the Bass model, we extended the original Bass model formulation to an extended Bass model as follows.

$$n(t) = \left(p + \frac{q}{m}N(t)\right)\left(m - N(t)\right)g\left(I(t)\right) \quad (5.3)$$

Everything else is same except for $g\left(I(t)\right)$ at right hand side. $I(t)$ denotes the number of cumulative mentions by influential users till time t. We believe it is an important indicator of the stage of diffusion in Twitter network. $g(\cdot)$ denotes a functional form which will transform the data $I(t)$. Here we use natural log function as $g(\cdot)$, which turns the final extended Bass model into

$$n(t) = \left(p + \frac{q}{m}N(t)\right)\left(m - N(t)\right)ln\left(I(t)\right) \quad (5.4)$$

The transformation is applied because the number of cumulative mentions by influential users is right-skewed. For simplicity, we dropped those observations that number of cumulative influential user mentions is zero in the original dataset. The truncation will conform to the log transformation of the data (since we cannot have log of zero) and also focus our analysis on those important time range. To keep consistency, we did same thing for the data used in basic Bass model. The estimation procedure is similar to the basic Bass model.

5.3 Basic Bass Model Analysis

In this and the next section, we will apply the basic and extended Bass model described in the previous two sections on the IPL data. The total data has 4320 observations, and we truncated first 27 observations since there is no influential user mention. Thus, 4293 observations were used in analysis. The same data will be used in extended Bass model analysis.

We used data from first 100% to 10% (10% each step) from the beginning of the time line in the analysis. In other words, when we say 40%, 60% of the total data from the end are dropped before the analysis. By analyzing different proportions of original data, we will have a nuanced understanding of model performance.

We evaluate the performance of our model estimation by two criterions. The first one is the total number of mentions, which is similar with the total potential market in the product adoption case. The second one is the location of the maximum mentions, which tells us when the mention will reach the highest point. Those two criterions are of practical importance. In the following sections, we show the estimation results by a table summarizing the two criterions and ten pictures with different proportion of data.

Table 5.1 Estimation Results of Basic Bass Model

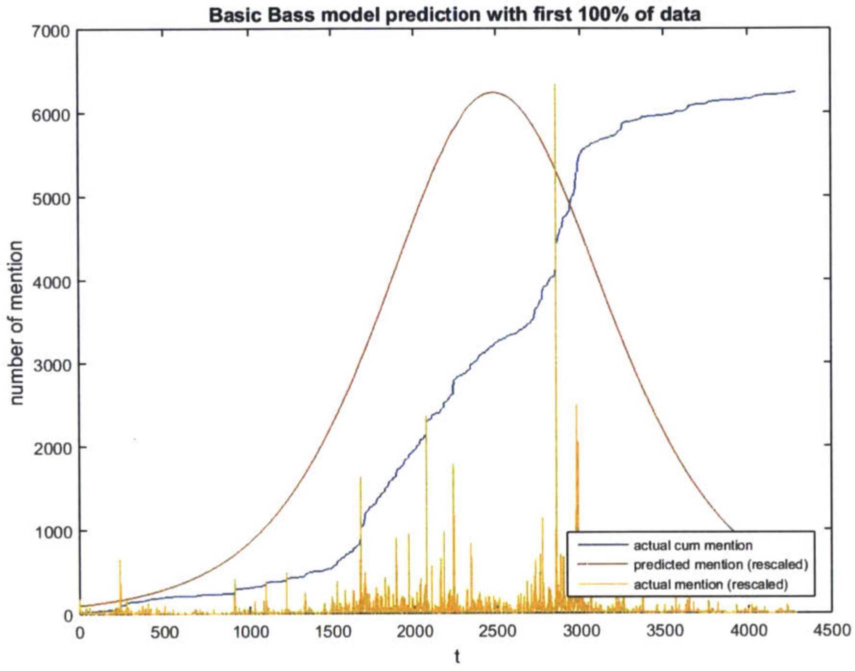
Proportion Data	Est. Tot. Men. (m)	Actual Tot. Men. (cum mentions)	Percentage Error	Est. Loc. Max (time of maximum mention)	Actual Loc. Max	Percentage Error
1	6518	6235	4.5%	2489	2854	12.8%
0.9	6620	6235	6.2%	2868	2854	0.5%
0.8	7128	6235	14.3%	2979	2854	4.4%
0.7	2074	6235	66.7%	590	2854	79.3%
0.6	3938	6235	36.8%	968	2854	66.1%
0.5	3717	6235	40.4%	911	2854	68.1%
0.4	510	6235	91.8%	306	2854	89.3%
0.3	582	6235	90.7%	379	2854	86.7%
0.2	231	6235	96.3%	303	2854	89.4%
0.1	206	6235	96.7%	290	2854	89.8%

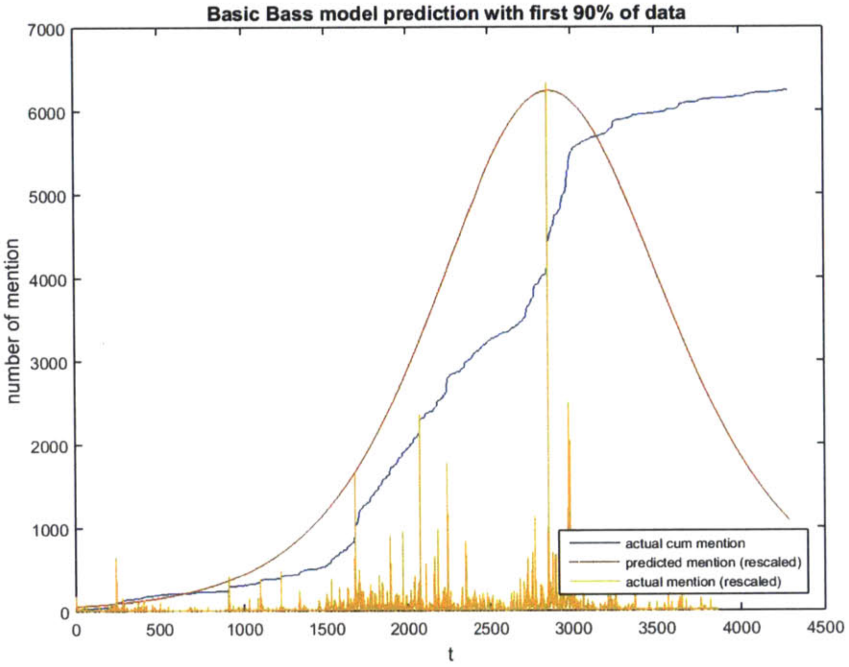
From Table 5.1, we could notice that the model performance is quite good from first 100% to first 80% from both criterions. The percentages of errors are all below 15%. When we use 70% data, the model performance goes very bad (percentage error > 65%). The model did well from first 60% to 50% in terms of predicting the total mentions, where the error rates are around 40%. The rest estimation results are bad as all of their error rates are above 65%.

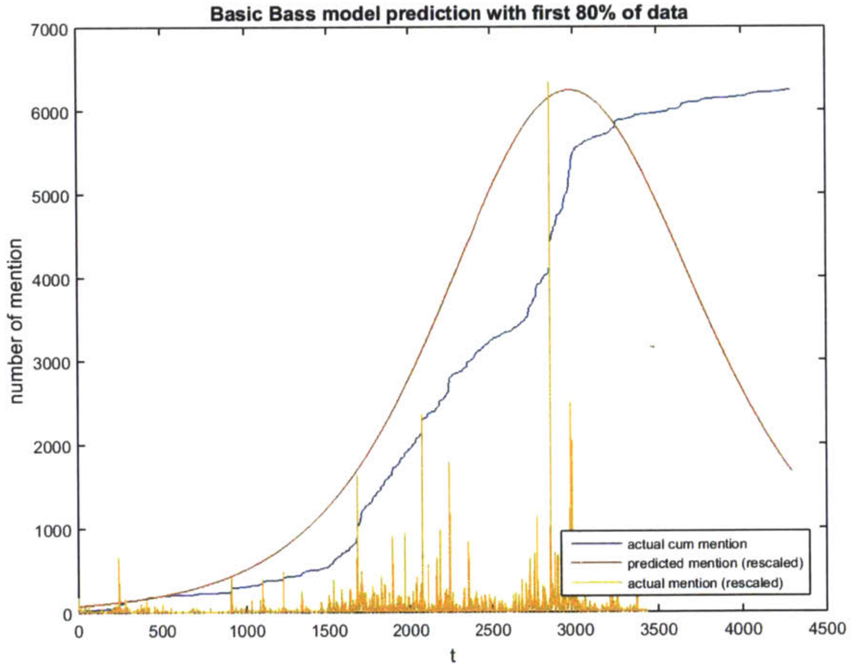
The following ten graphs give more direct ideas of the model performances, especially in terms of predicting the peak time of the mention. The following graph describes the basic Bass model prediction with 100% of data. The blue line shows the actual cumulative mentions. The red and yellow lines show the predicted and actual mentions of first proportion of data. Because the time range is large, the magnitude of mention is much lower than that of cumulative mention. Therefore, we rescaled both mention data by multiplying the ratio of maximum of cumulative mention to maximum of actual (or predicted) mention.

Noticing from first 100% to first 80%, the peaks of red curves are not exactly consistent with the yellow lines (the actual ones) but they are very close with each other.

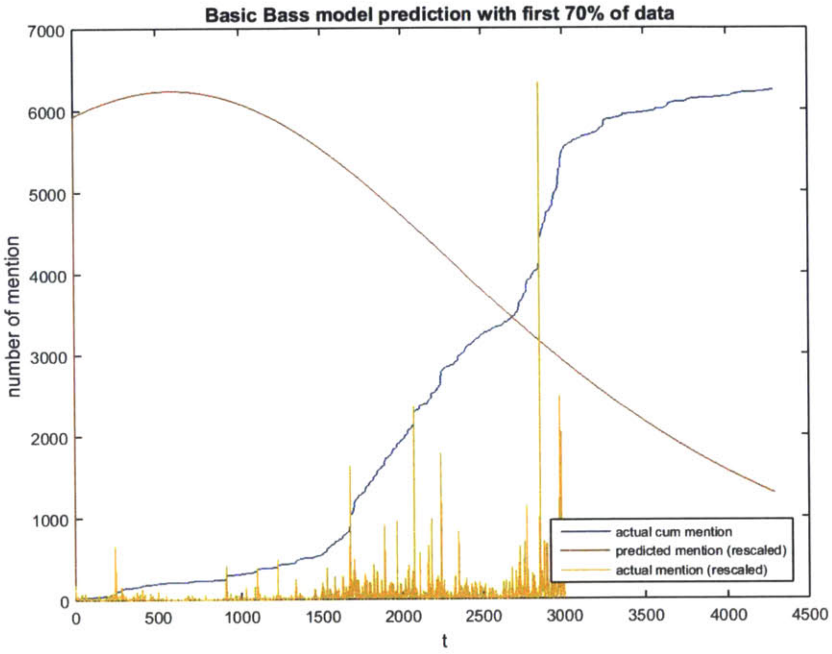
Figure 5.1 Basic Bass Model Prediction

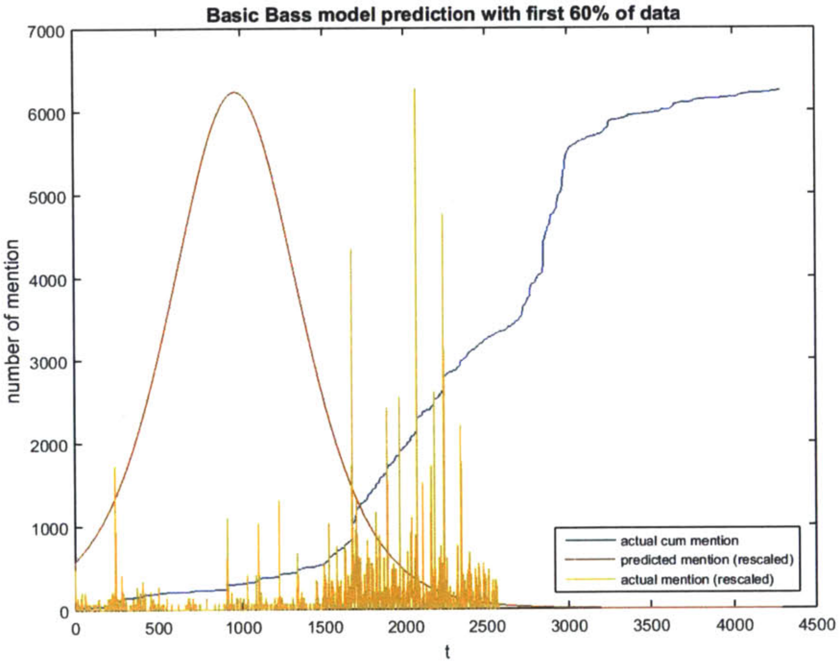


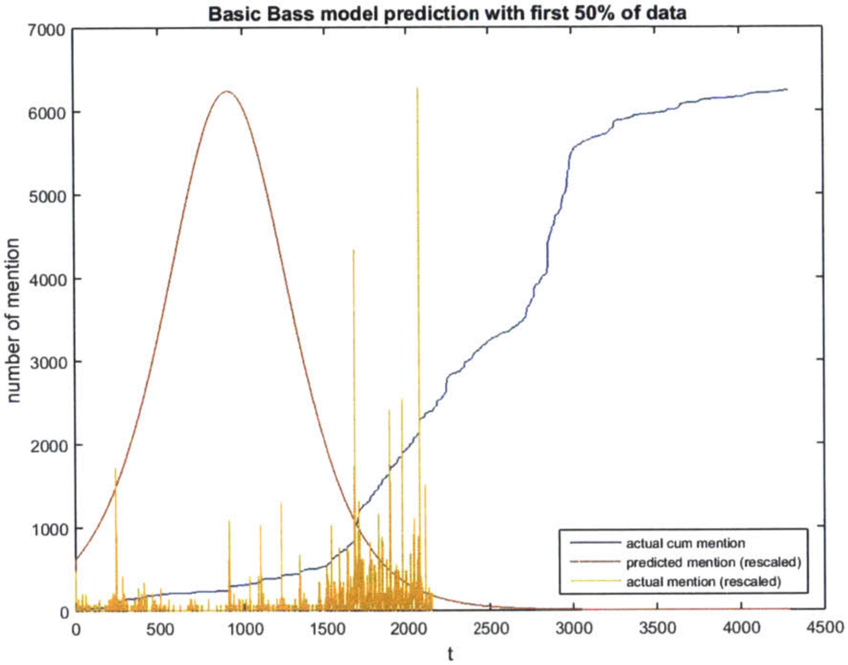

Basic Bass model prediction with first 90% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

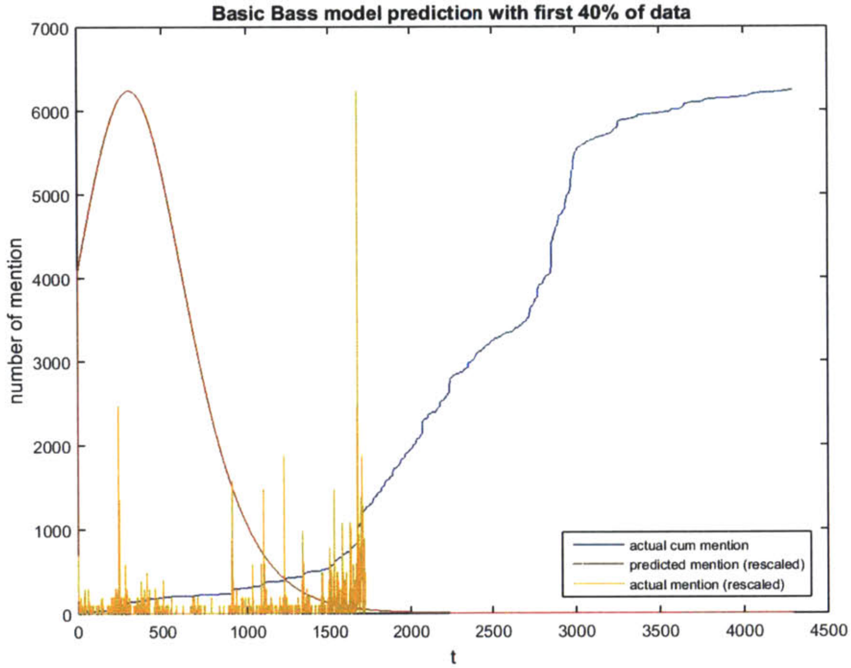

Basic Bass model prediction with first 80% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

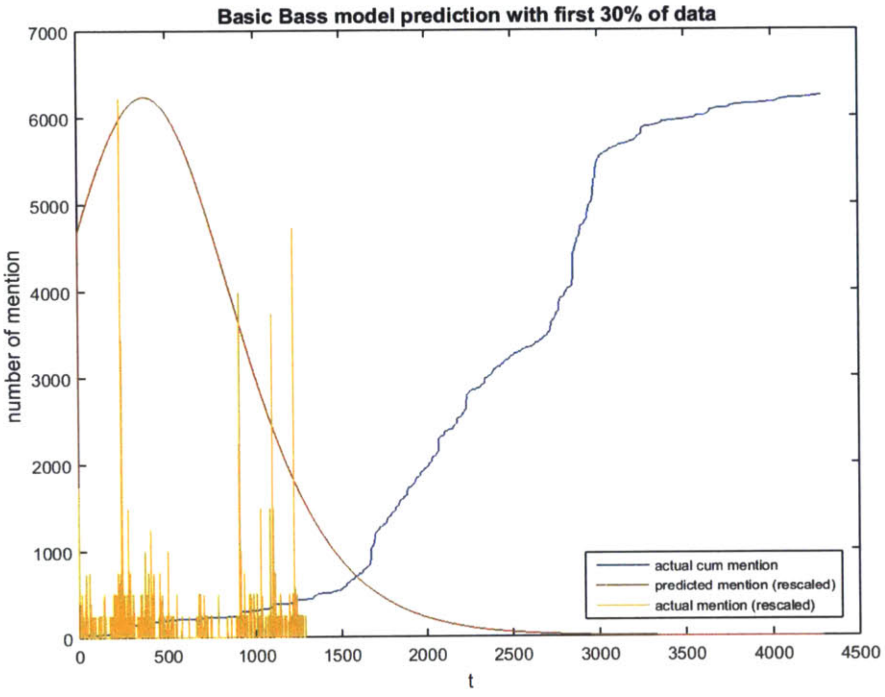
We noticed that for first 70% of data, the predicted curve does not capture the actual peak at all (the estimation result is very poor). We notice the similar things from the graphs with first 60% to fist 10% data. The model could not capture the peak time of the actual diffusion.

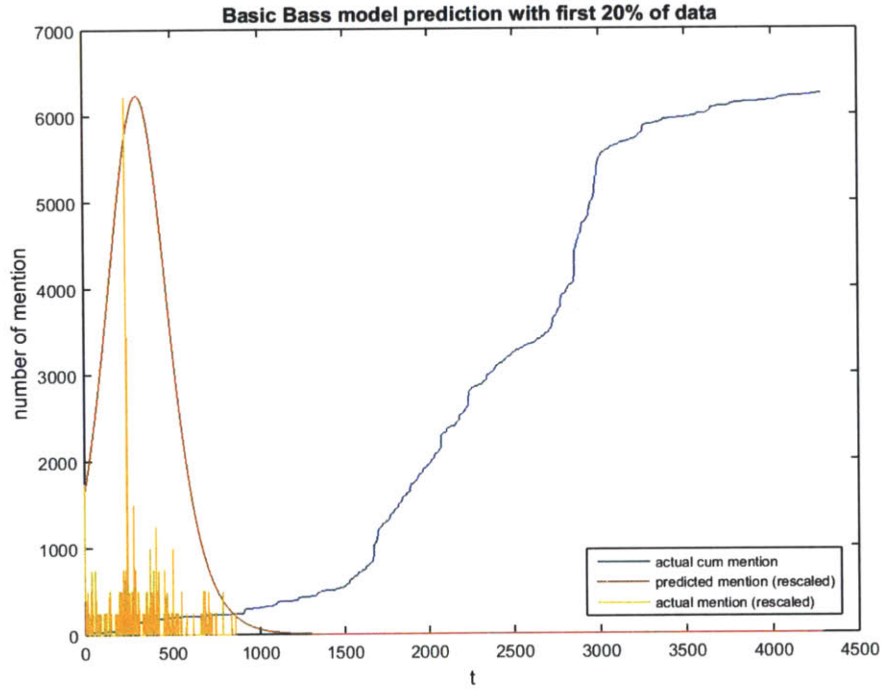


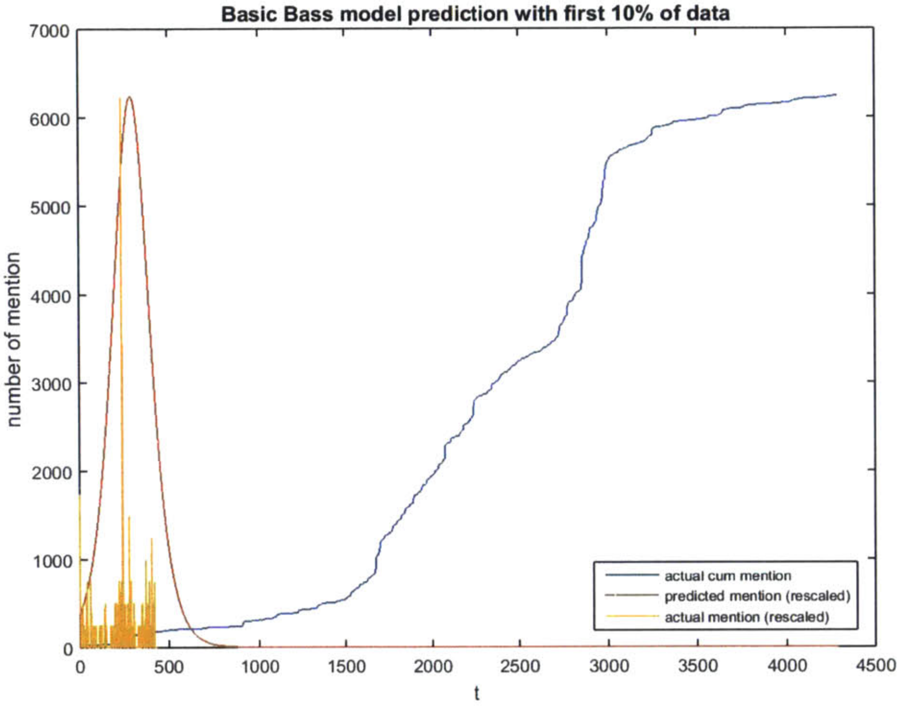



Basic Bass model prediction with first 50% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Basic Bass model prediction with first 40% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Basic Bass model prediction with first 30% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Basic Bass model prediction with first 20% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Basic Bass model prediction with first 10% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

5.4 Extended Bass Model Analysis

Since the prediction results of basic Bass model are not good for those from with the first 70% to first 10% of data, we consider incorporating more information to capture the actual diffusion pattern. The extended Bass model described in section 5.2 gives us a direction of improvement. We thus add the number of mentions from the influential users into the model.

Table 5.2 Estimation Results of Extended Bass Model

Proportion Data	Est. Tot. Men. (m)	Actual Tot. Men. (cum mentions)	Percentage Error	Est. Loc. Max (time of maximum mention)	Actual Loc. Max	Percentage Error
1	6394	6235	2.6%	2779	2854	2.6%
0.9	6481	6235	3.9%	2779	2854	2.6%
0.8	6900	6235	10.7%	2775	2854	2.8%
0.7	3060	6235	50.9%	1701	2854	40.4%
0.6	3780	6235	39.4%	1701	2854	40.4%
0.5	3229	6235	48.2%	1126	2854	60.5%
0.4	460	6235	92.6%	744	2854	73.9%
0.3	711	6235	88.6%	374	2854	86.9%
0.2	538	6235	91.4%	280	2854	90.2%
0.1	561	6235	91.0%	280	2854	90.2%

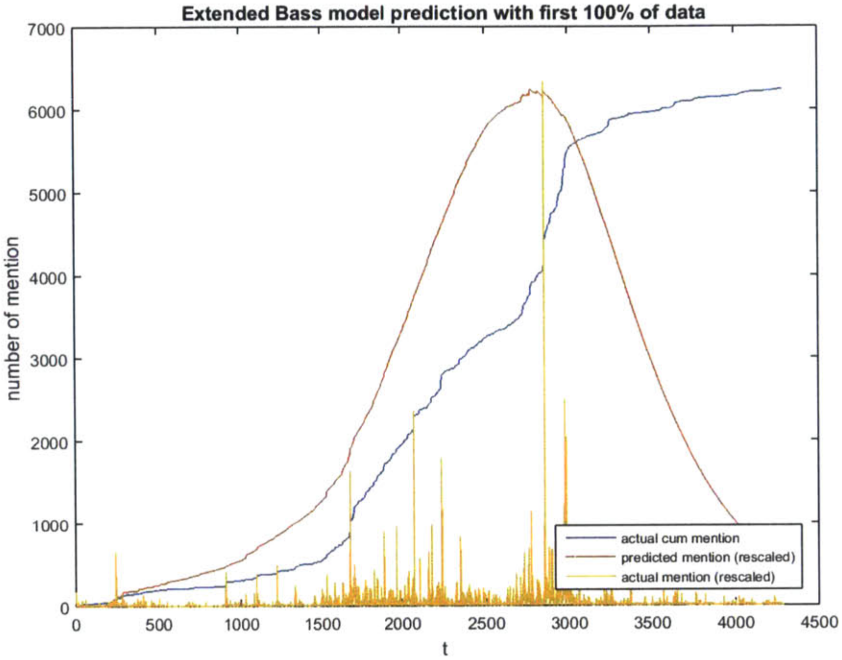
Comparing the estimation results from Table 5.1 and Table 5.2, we can observe significant improvements, in particular from first 100% to first 70% in terms of predicting total mentions. The percentages of errors decrease (from 4.5% to 2.6%, from 6.2% to 3.9%, from 14.3% to 10.7%, etc). The error rate of first 70% decrease dramatically from 66.7% to 50.9%.

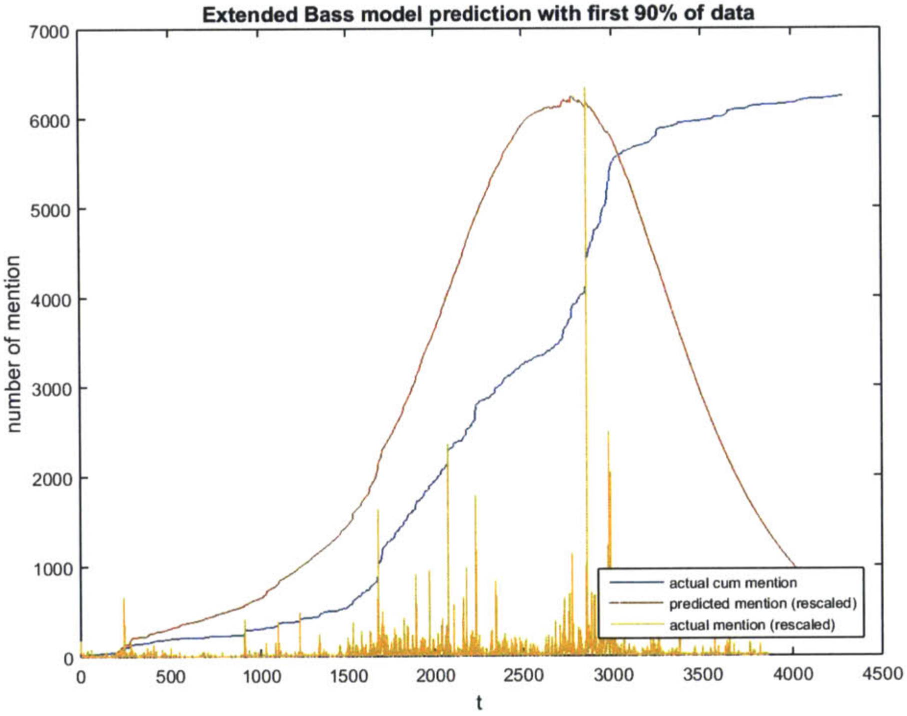
In addition, we observe significant improvements in terms of predicting the peak time. The percentages of errors decrease dramatically for those with 70% to 50% (from 79.3% to 40.4%, from 66.1% to 40.4% and from 68.1% to 60.5%) of data. Thus, we conclude that the extended Bass model does improve the estimation performance.

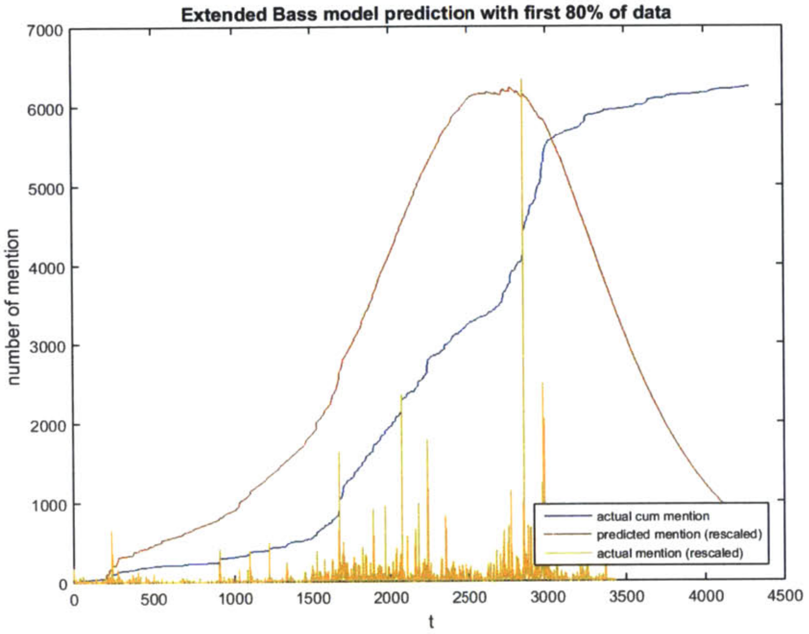
The following ten graphs give a good grasp of the improvements in terms of predicting the peak time.

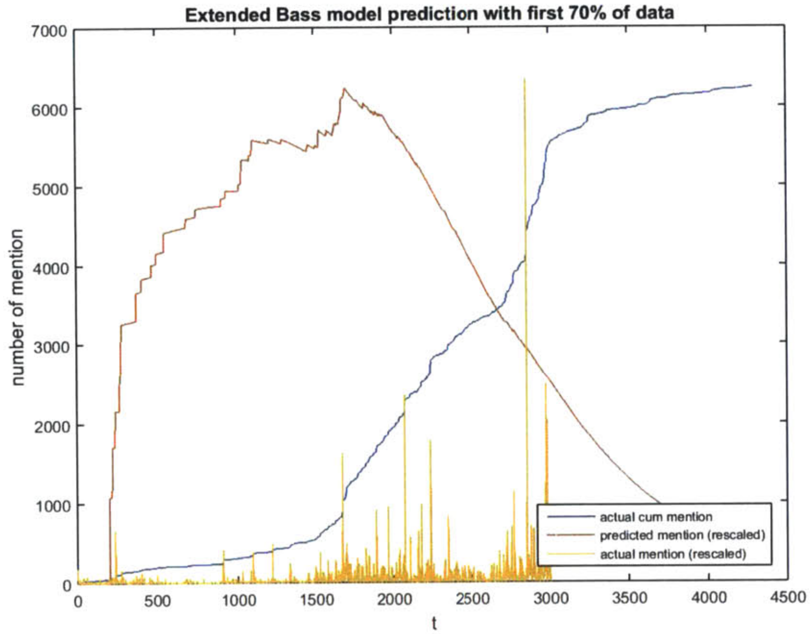
In the first three graphs (first 100%, 90% and 80%), we observe that the red curves always capture the peak of the yellow lines.

Figure 5.2 Extended Bass Model Prediction

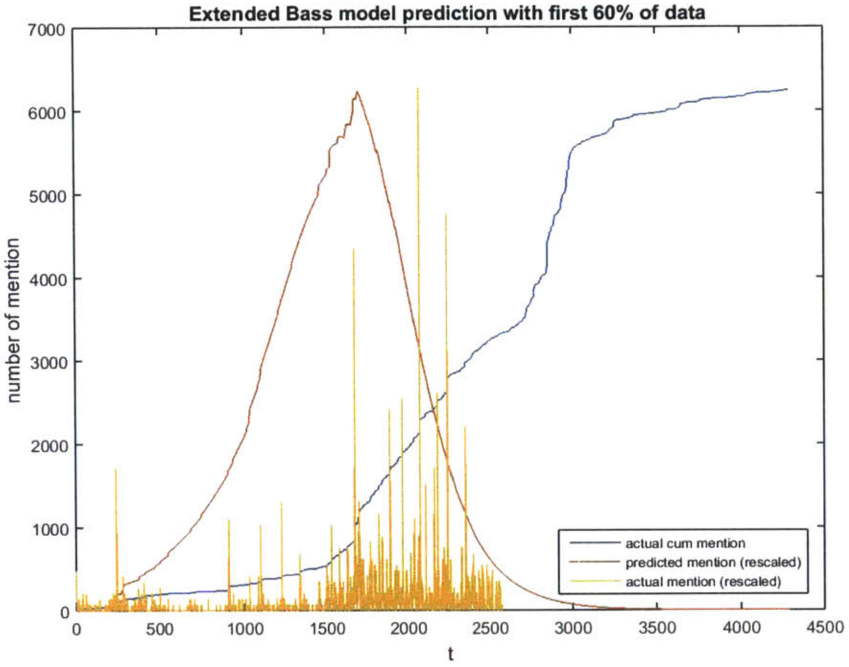


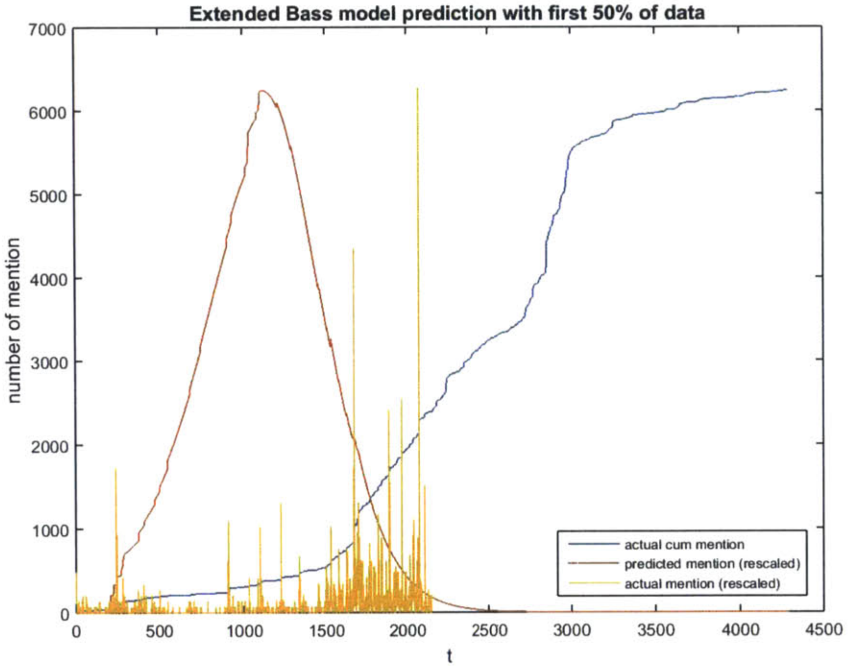


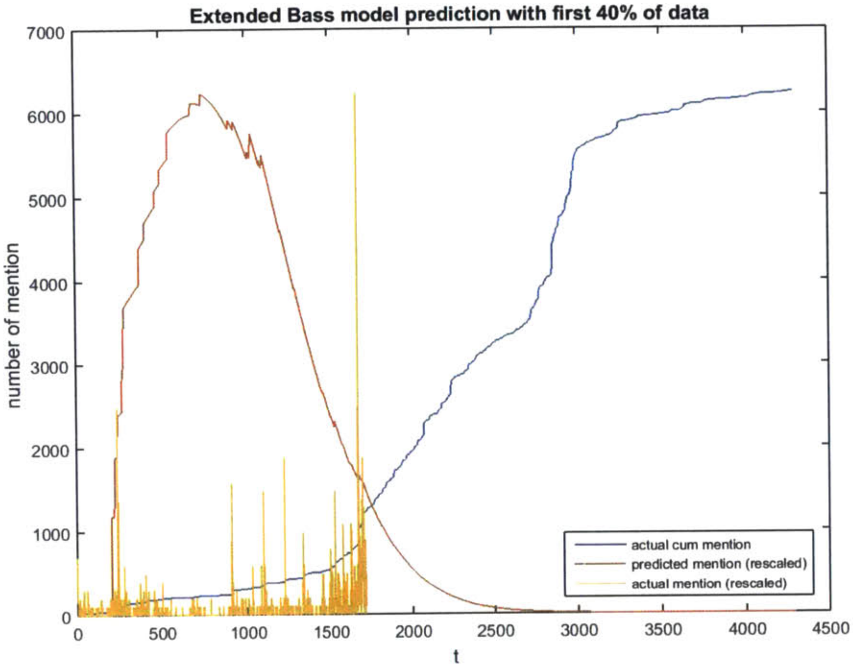


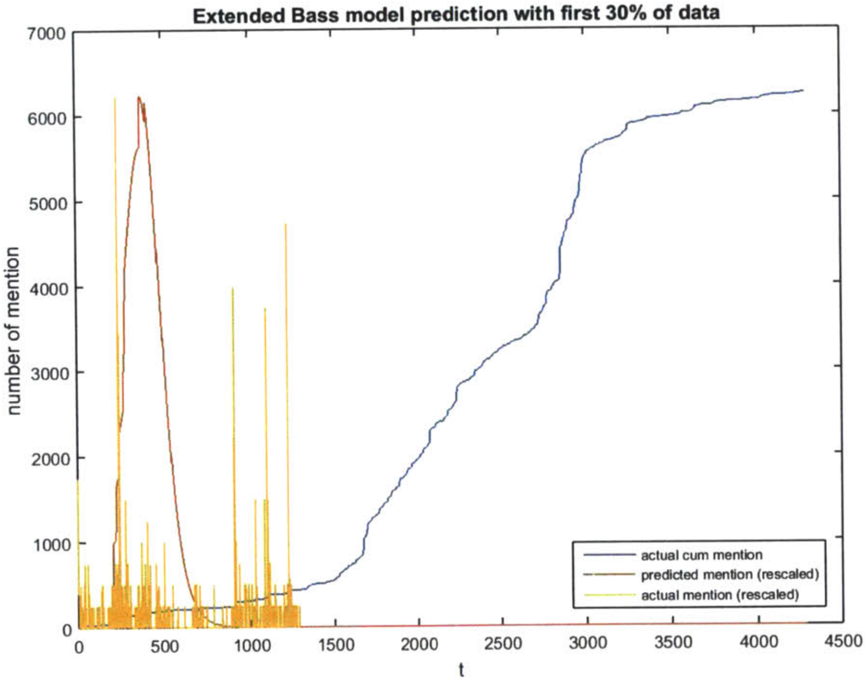


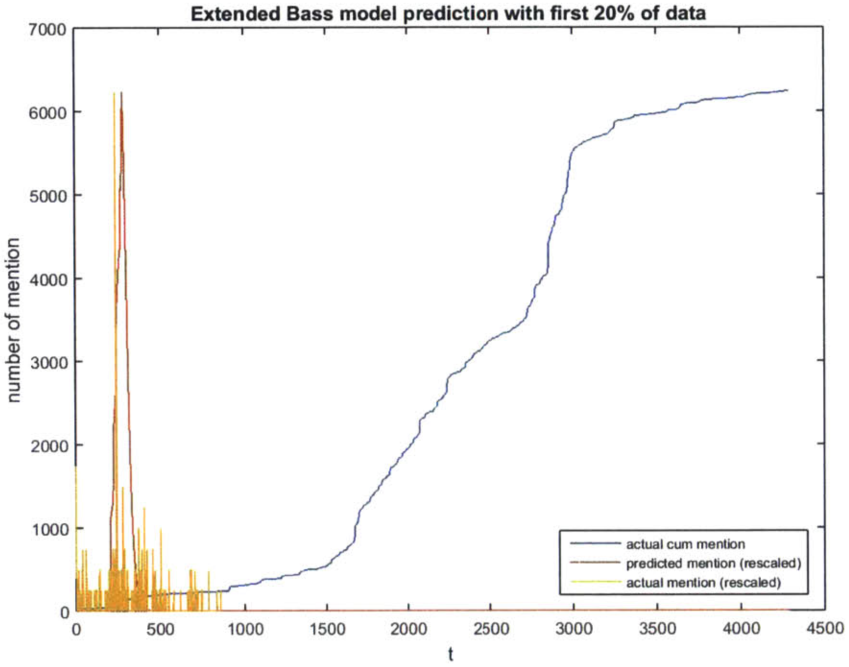
The performance of first 70% is still not good but we can see that the peak of red curve is not far away from the yellow lines. It is likely to capture the early peaks. The similar results could be observed from the graphs of first 60% to first 10%.


Extended Bass model prediction with first 60% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

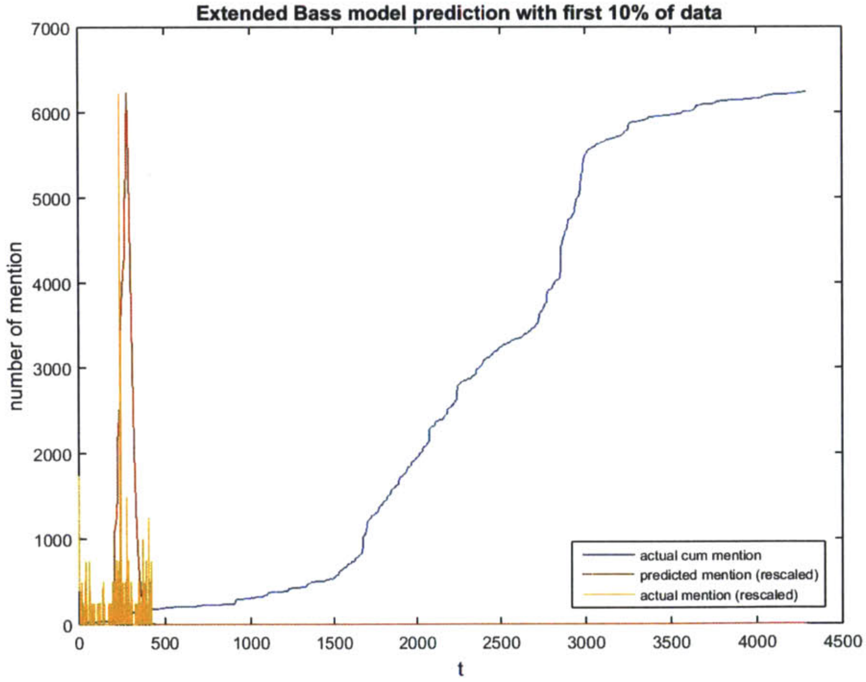

Extended Bass model prediction with first 50% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Extended Bass model prediction with first 40% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Extended Bass model prediction with first 30% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)



Extended Bass model prediction with first 20% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)



Extended Bass model prediction with first 10% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
500
1000
1500
2000
2500
3000
3500
4000
4500
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

5.5 Basic Bass Model Analysis with Aggregate Data

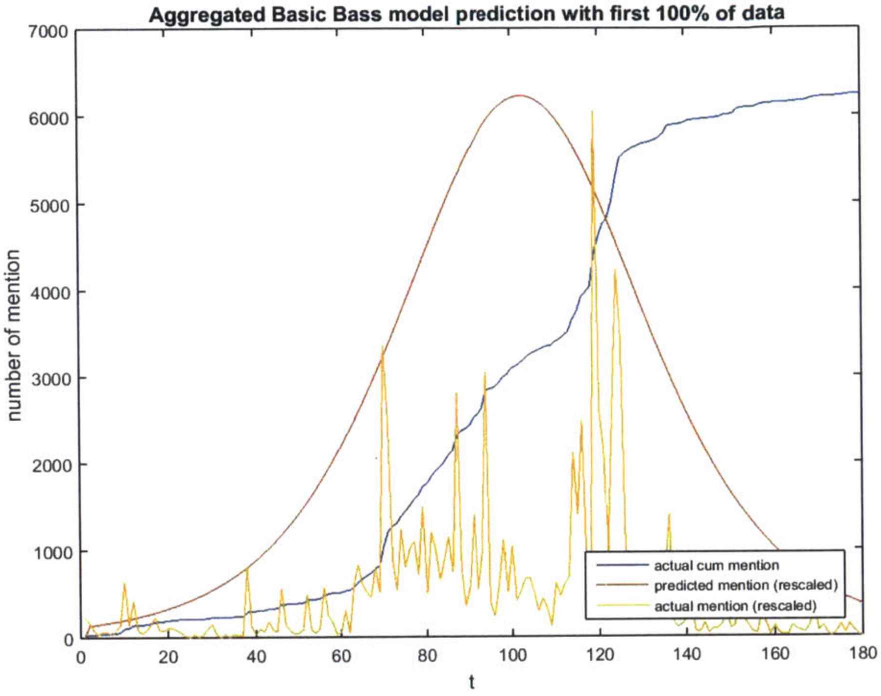
We also aggregated data in time dimension to reduce fluctuation and to ensure an increasing trend. We used 5 days, 1 day (24 hours), 12 hours, 6 hours as our time dimension for data aggregation, and found the best model performance when we aggregate data per day (24 hours). Thus, for each data observation in the model, we aggregated mention count per day instead of per hour, and used aggregated data in the model prediction. In the IPL tournament dataset, the total data size is 180 after data aggregation, meaning that there are 180 days in total. The following table shows the summary of the analysis.

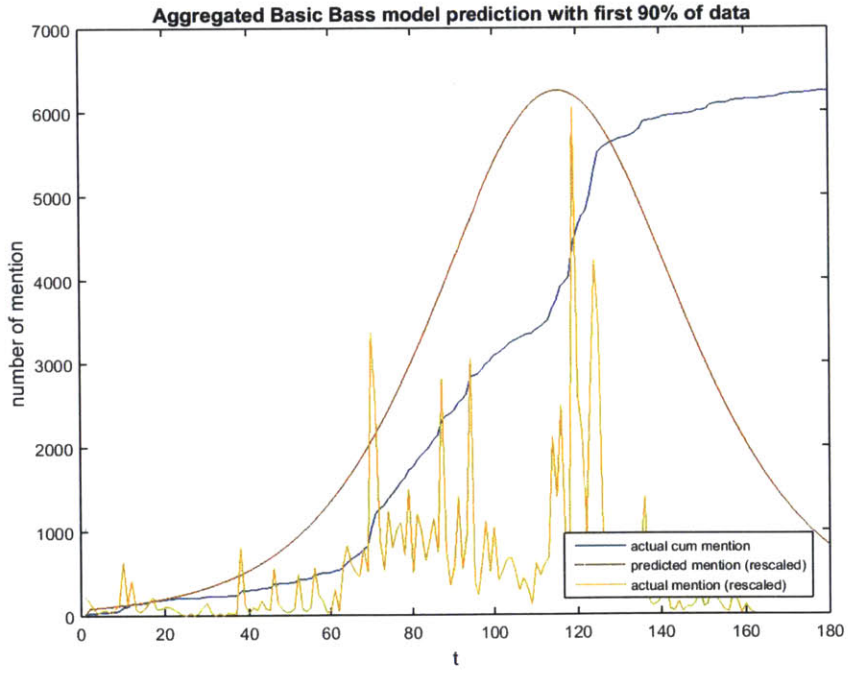
Table 5.3 Estimation Results of Aggregated Basic Bass Model

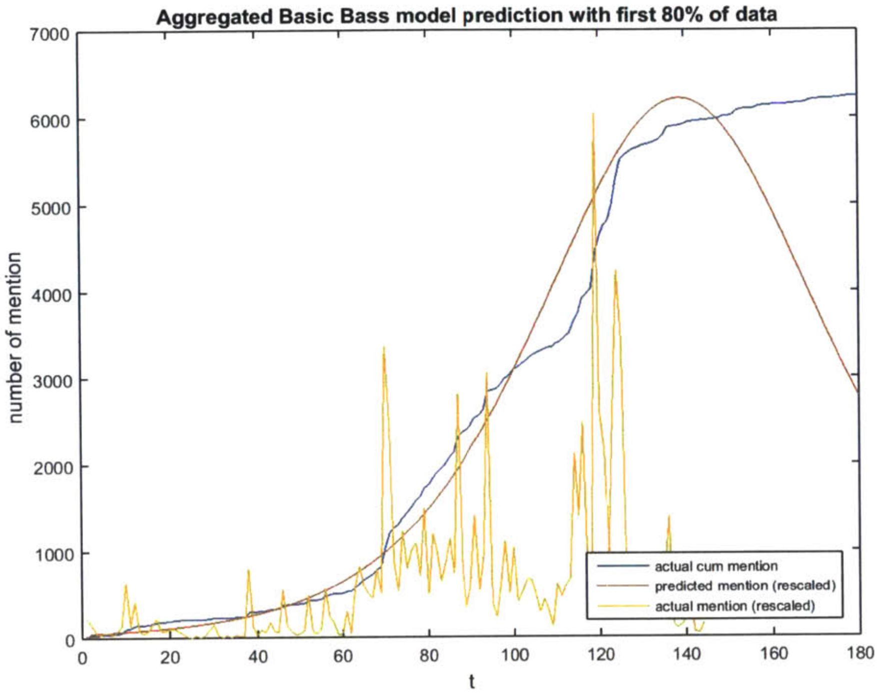
Proportion Data	Est. Tot. Men. (m)	Actual Tot. Men. (cum mentions)	Percentage Error	Est. Loc. Max (time of maximum mention)	Actual Loc. Max	Percentage Error
1	6555	6235	5.1%	103	119	13.4%
0.9	6664	6235	6.9%	115	119	3.4%
0.8	7162	6235	14.9%	139	119	16.8%
0.7	4368	6235	29.9%	98	119	17.6%
0.6	3987	6235	36.1%	43	119	63.9%
0.5	3779	6235	39.4%	41	119	65.5%
0.4	449	6235	92.8%	17	119	85.7%
0.3	567	6235	90.9%	18	119	84.9%
0.2	233	6235	96.3%	17	119	85.7%
0.1	217	6235	96.5%	14	119	88.2%

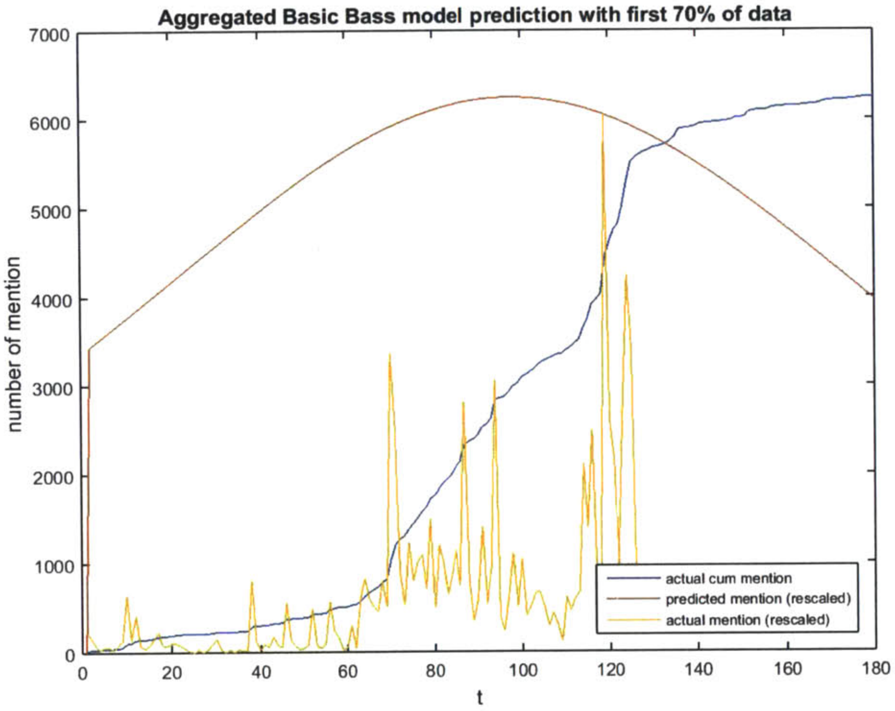
Comparing Table 5.1 and Table 5.3, we notice that the majority of estimation results are similar. However, we could observe huge improvements in terms of the estimation results for first 70%. The percentages of errors in previous model are 66.7% and 79.3% for total mention and peak, but now the error rates become 29.9% and 17.6%. The following ten graphs illustrate the model prediction performance using different proportion of data.

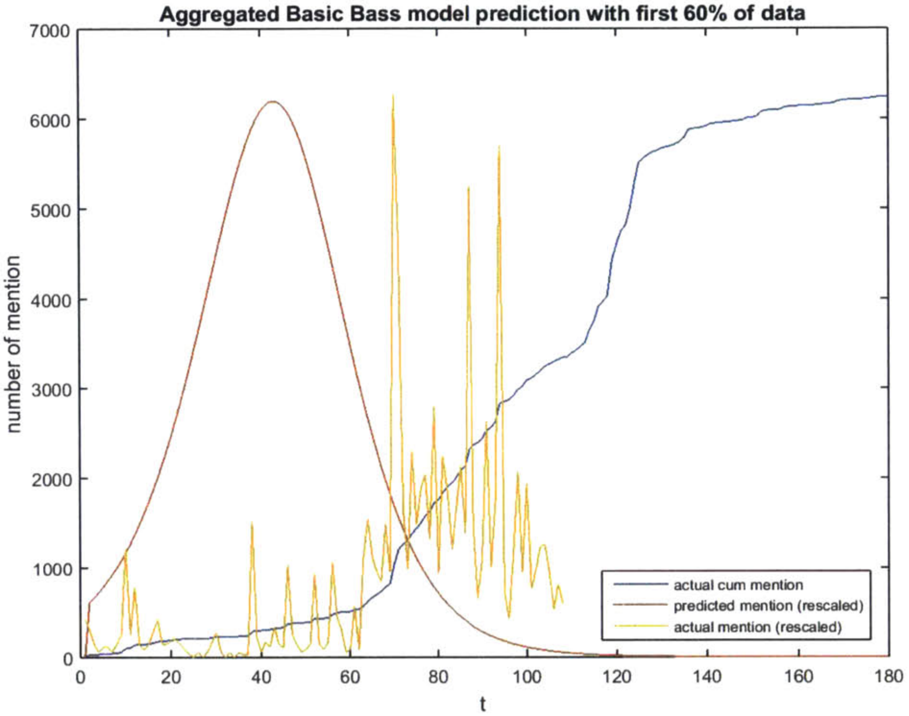
Figure 5.3 Aggregated Basic Bass Model Prediction

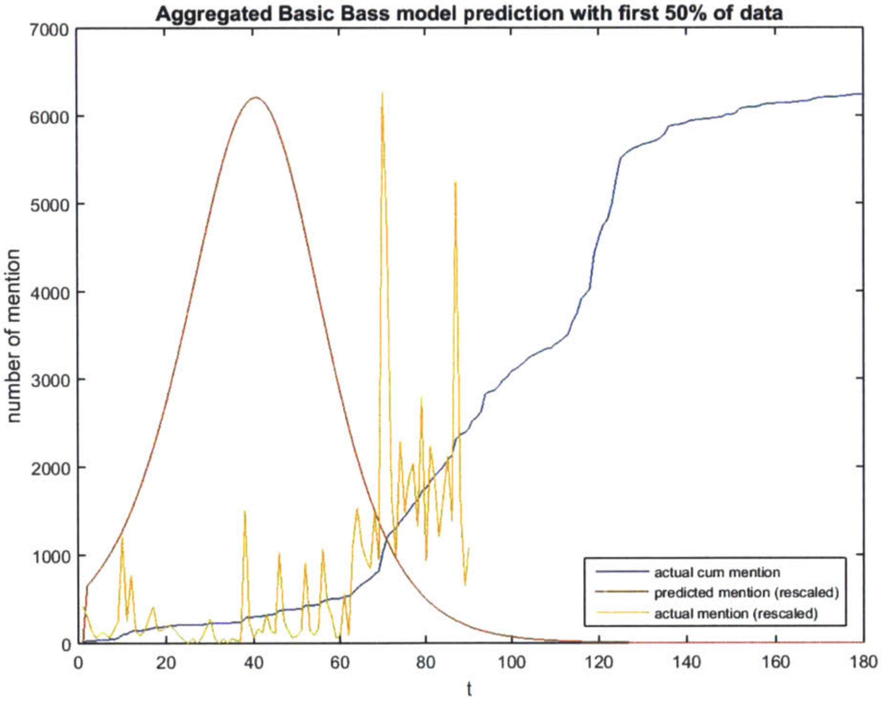


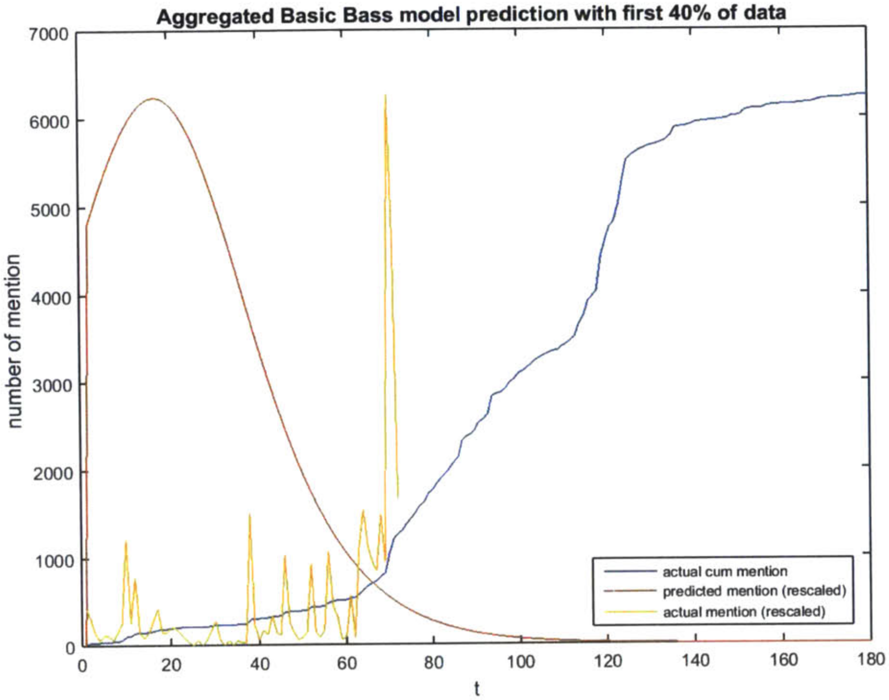


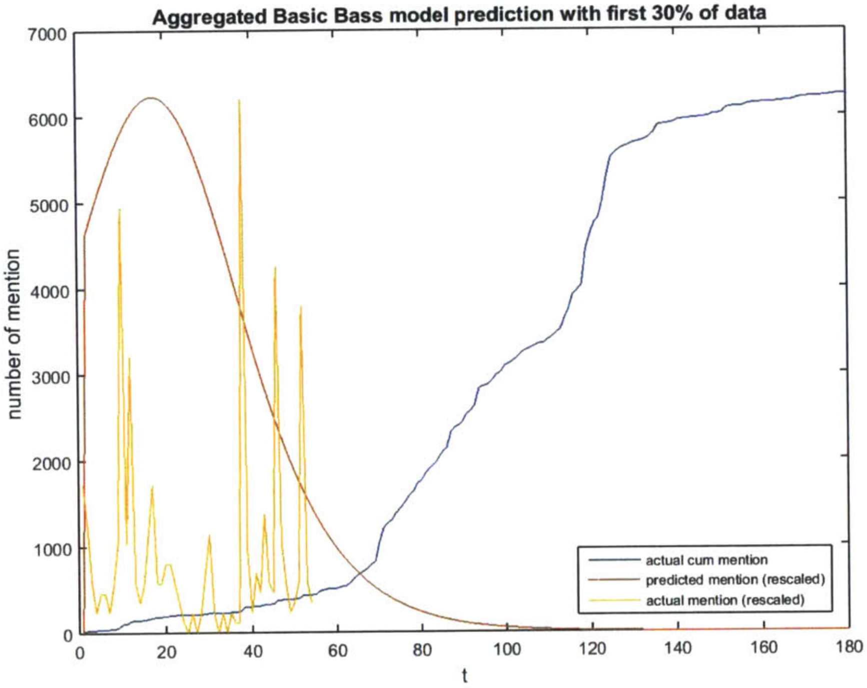

Aggregated Basic Bass model prediction with first 80% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

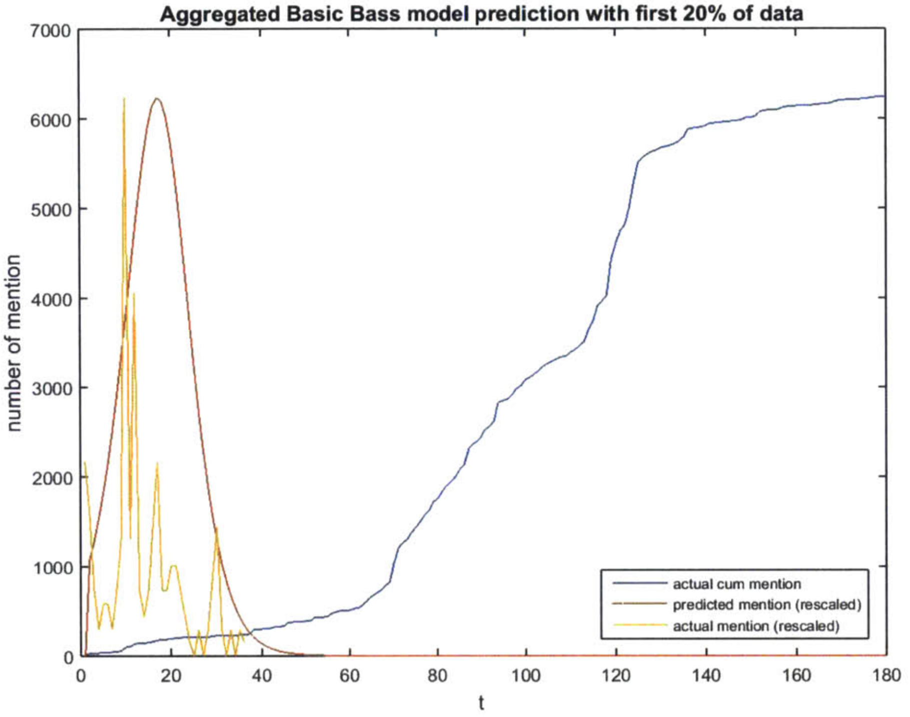

Aggregated Basic Bass model prediction with first 70% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

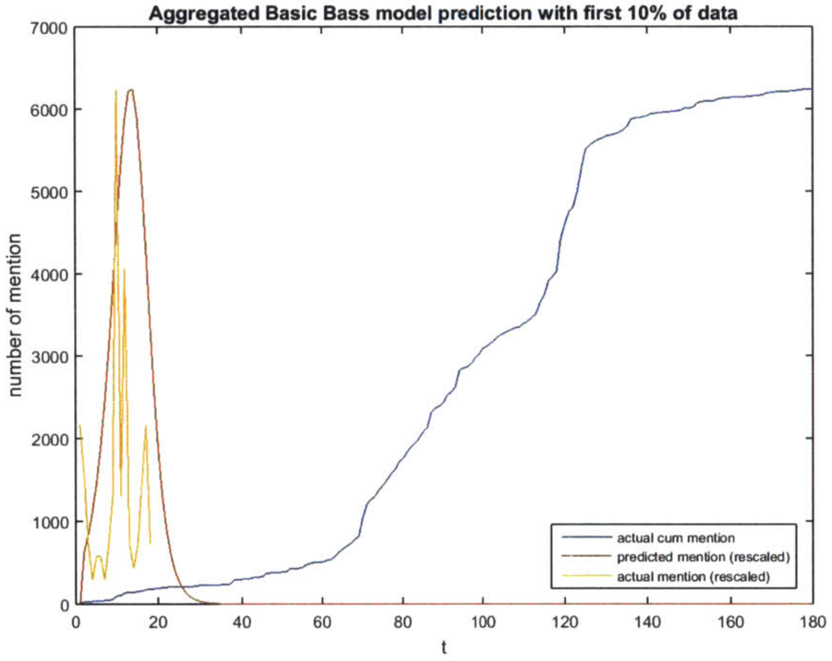

Aggregated Basic Bass model prediction with first 60% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Basic Bass model prediction with first 50% of data
number of mention
0
1000
2000
3000
4000
5000
6000
7000
t
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Basic Bass model prediction with first 40% of data
7000
6000
5000
4000
3000
2000
1000
0
number of mention
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Basic Bass model prediction with first 30% of data
7000
6000
5000
4000
3000
2000
1000
0
number of mention
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Basic Bass model prediction with first 20% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Basic Bass model prediction with first 10% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

5.6 Extended Bass Model Analysis with Aggregate Data

Here, we again used data from 100% to 10% (10% each step) from the beginning of the time line in the analysis. Following the data aggregation step performed for the basic Bass model, we also aggregated mention count per 24 hours (day) instead of per hour, and used aggregated data in the extended Bass model prediction. The following table shows the summary of the analysis.

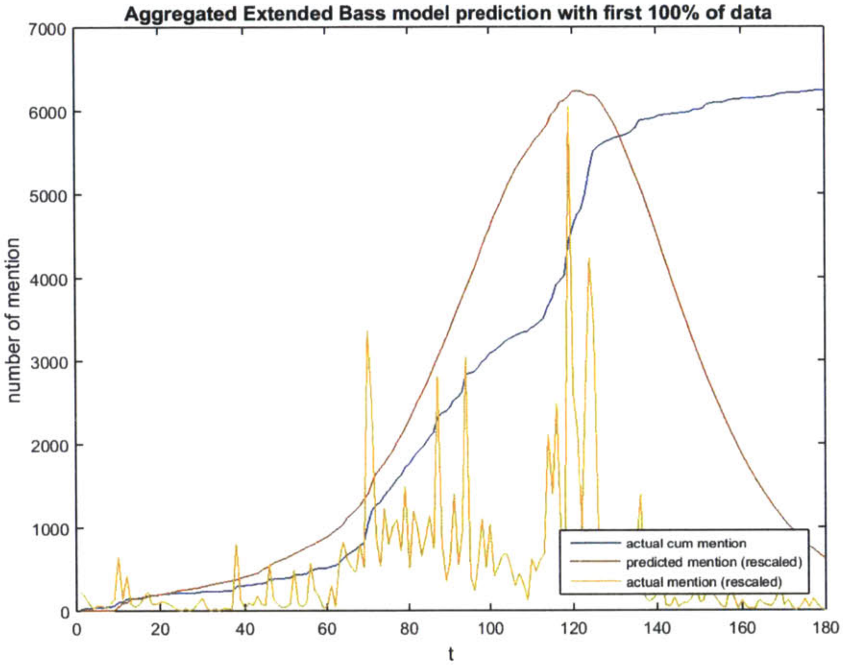
Table 5.4 Estimation Results of Aggregated Extended Bass Model

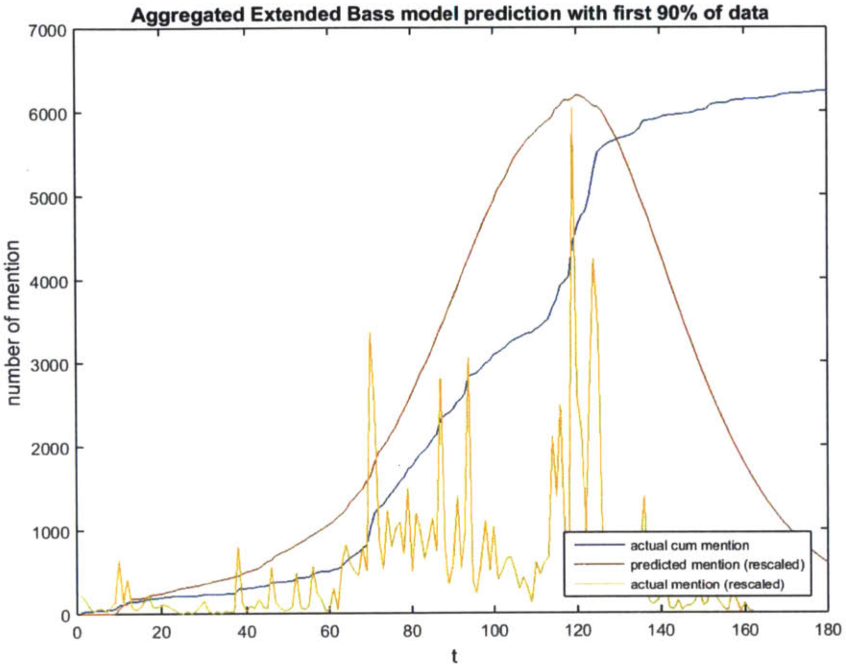
Proportion Data	Est. Tot. Men.	Actual Tot. Men.	Percentage Error	Est. Loc. Max	Actual Loc. Max	Percentage Error
1	6412	6235	0.03	121	119	0.02
0.9	6505	6235	0.04	120	119	0.01
0.8	6930	6235	0.11	120	119	0.01
0.7	2954	6235	0.53	72	119	0.39
0.6	3834	6235	0.39	72	119	0.39
0.5	3305	6235	0.47	50	119	0.58
0.4	211	6235	0.97	47	119	0.61
0.3	711	6235	0.89	18	119	0.85
0.2	541	6235	0.91	31	119	0.74
0.1	490	6235	0.92	29	119	0.76

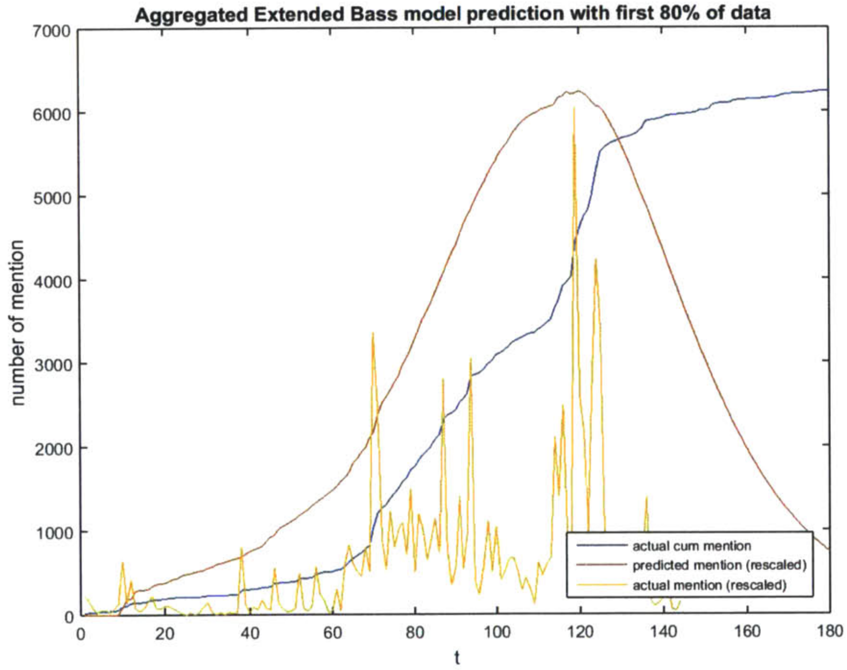
Comparing the estimation results in Table 5.3 and Table 5.4, we can observe significant improvements in both criterions. For example, in terms of predicting total mentions, the percentages of errors from first 100% to first 80% were 5.1%, 6.9% and 14.9% in previous model, but now they become 3.0%, 4.0% and 11.0%. We could observe even more significant improvements in terms of predicting the peak times, the previous error rates are 13.4%, 3.4% and 16.8%, and now the numbers are 2.0%, 1.0% and 1.0%. Furthermore, we observe significant improvements in first 40%, 30% and 10%, which go from 85.7%, 84.9% and 88.2% to 58.0%, 61.0% and 74.0%. The last model has not only shown improvements in using large proportion of the data, but also shown improvements when we are using small proportion of data.

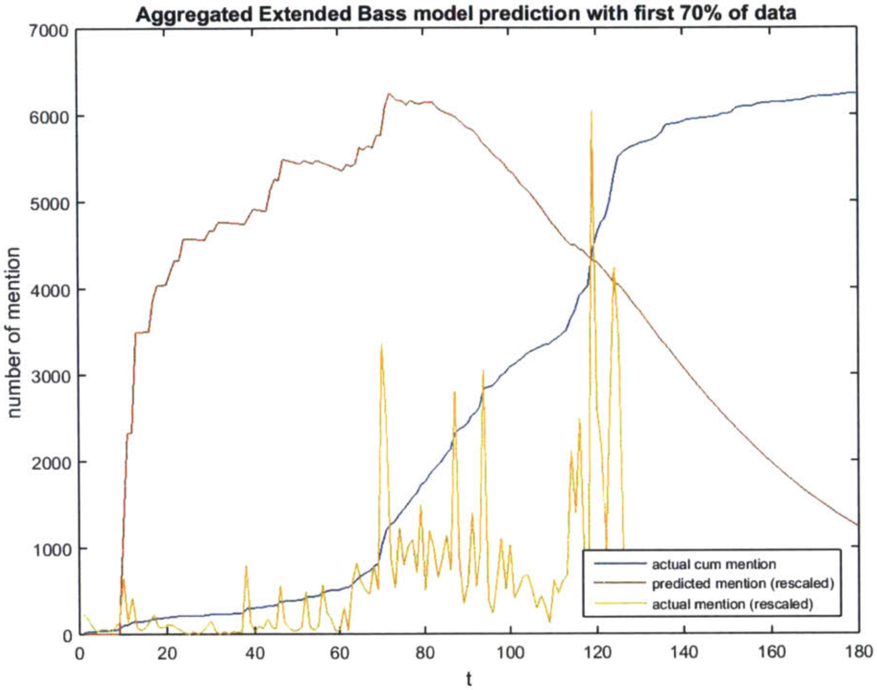
The following ten graphs illustrate the model prediction performance using different proportion of data. If compared with the ten graphs in the previous section, the current model is better at capturing the peak times of the actual trending topic.

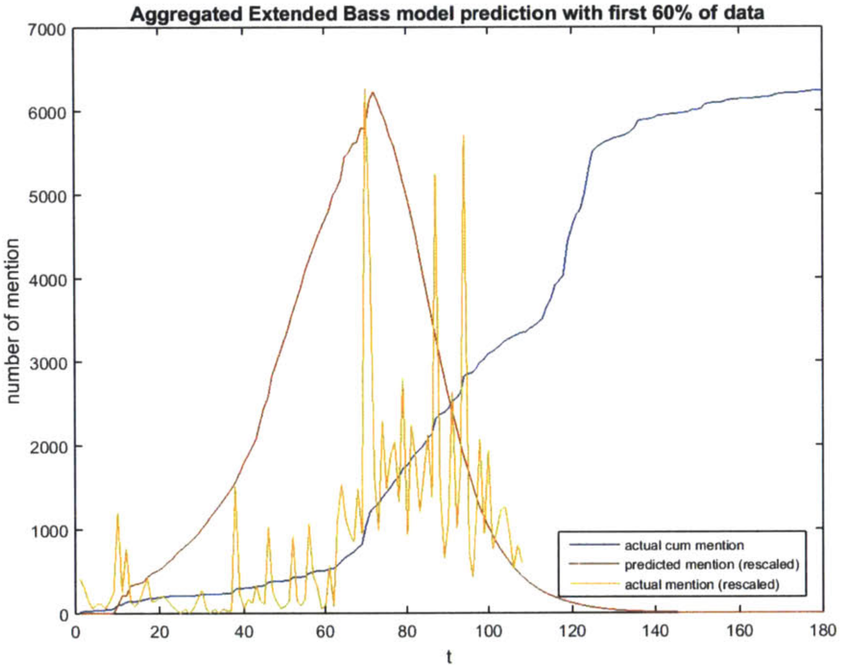
Figure 5.4 Aggregated Extended Bass Model Prediction

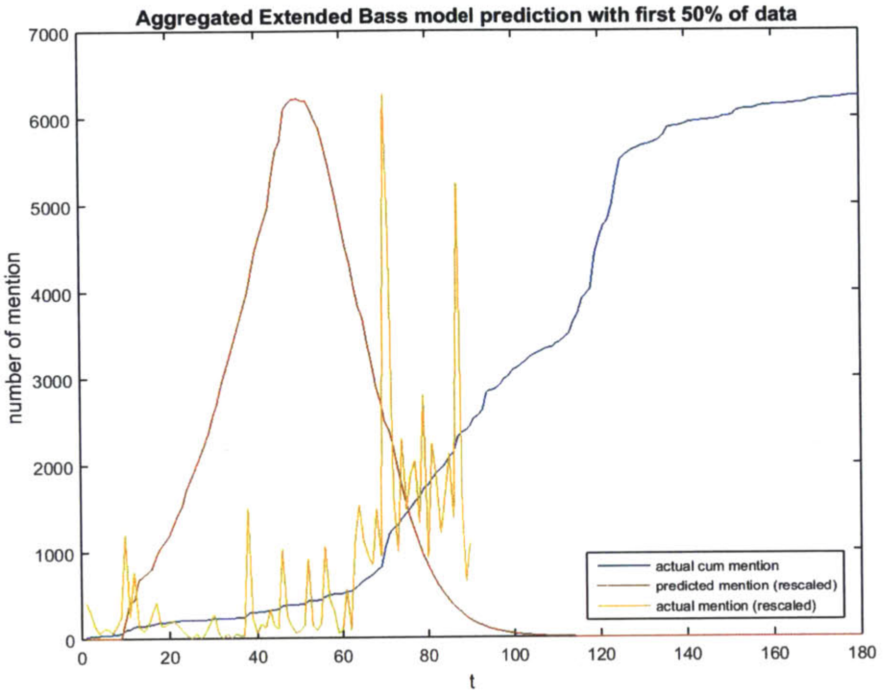


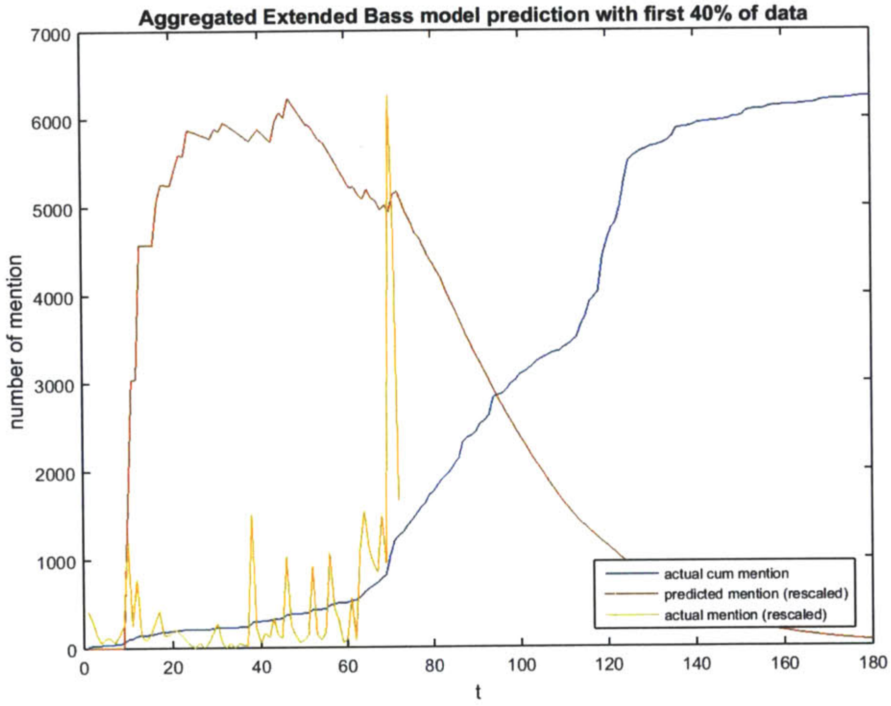

Aggregated Extended Bass model prediction with first 90% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

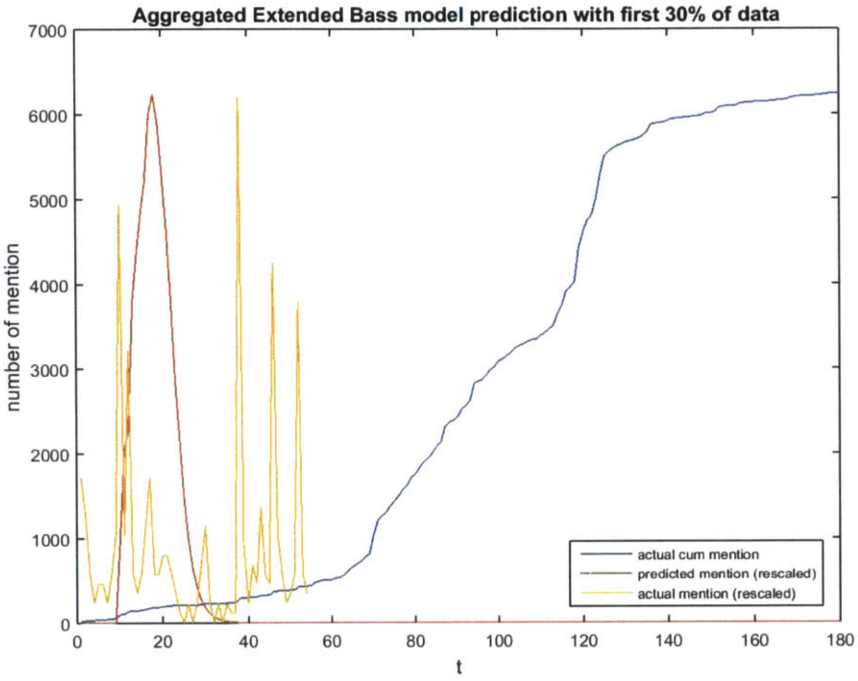

Aggregated Extended Bass model prediction with first 80% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

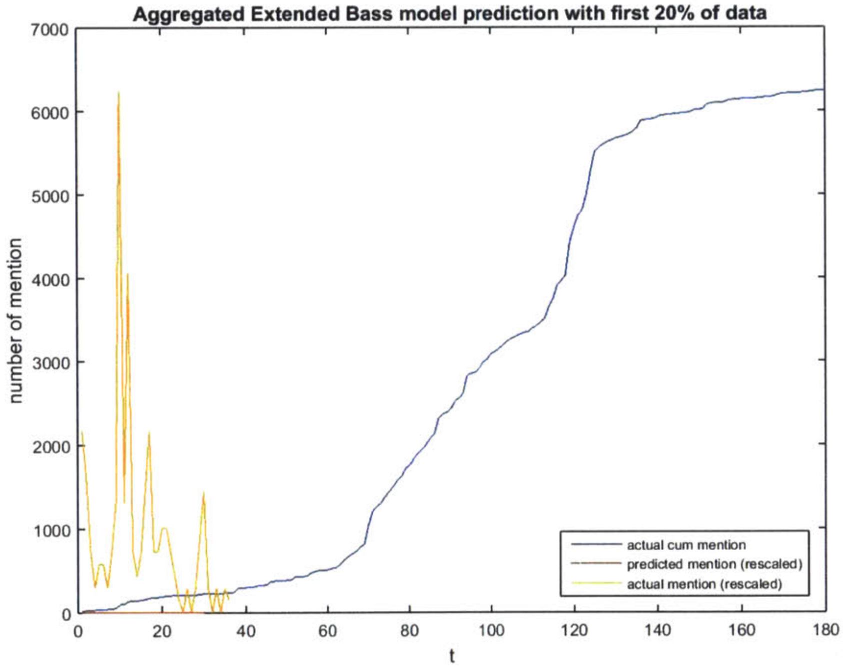

Aggregated Extended Bass model prediction with first 70% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)

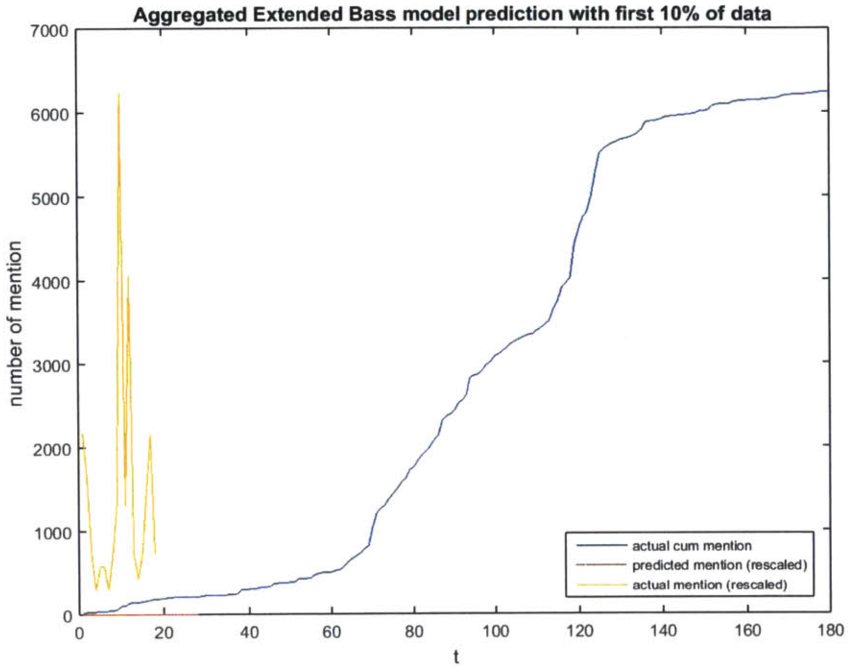

Aggregated Extended Bass model prediction with first 60% of data
number of mention
t
0
1000
2000
3000
4000
5000
6000
7000
0
20
40
60
80
100
120
140
160
180
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Extended Bass model prediction with first 50% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Extended Bass model prediction with first 40% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Extended Bass model prediction with first 30% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)


Aggregated Extended Bass model prediction with first 20% of data
number of mention
7000
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
120
140
160
180
t
actual cum mention
predicted mention (rescaled)
actual mention (rescaled)



In sum, we have shown four different models and their estimation results. By comparing their performances, we observed significant improvements from basic Bass model to extended Bass model. Furthermore, we observed that aggregating the observation from hourly to daily improve the model prediction, especially when we are using first 70% of the data. While the original prediction is poor, aggregation solves this issue and it achieves good estimation performance. Lastly, we tried extended model with aggregated data and it significantly improved model performance as compared to the basic bass model with aggregation. More importantly, we not only observed improvements when we use large proportion of data (say 80% or 70%), but also observed improvements when we use small proportion of data (say 30% or 10%).

6. Discussion

In this paper, we modeled the spread of word-of-mouth topics about specific events on Twitter. By reviewing existing work in diffusion, we conclude that the Bass model fit our objective: to model the diffusion of specific topics to provide important predictions.

Notably, our model achieved reasonable accuracy in predicting the total number of mentions and the times when the maximum mentions occurred. We extended the classic Bass model to incorporate cumulative mentions by influential users, thus successfully improving the model's prediction performance. Furthermore, we aggregated original hourly data observations into daily data observations, and saw significant improvements in prediction, especially when predicting with the first 70% of data. We found that prediction accuracy improved using the extended Bass model with aggregation, as compared to the basic Bass model with aggregation. More importantly, we observed prediction improvement using both a small and large proportion of data.

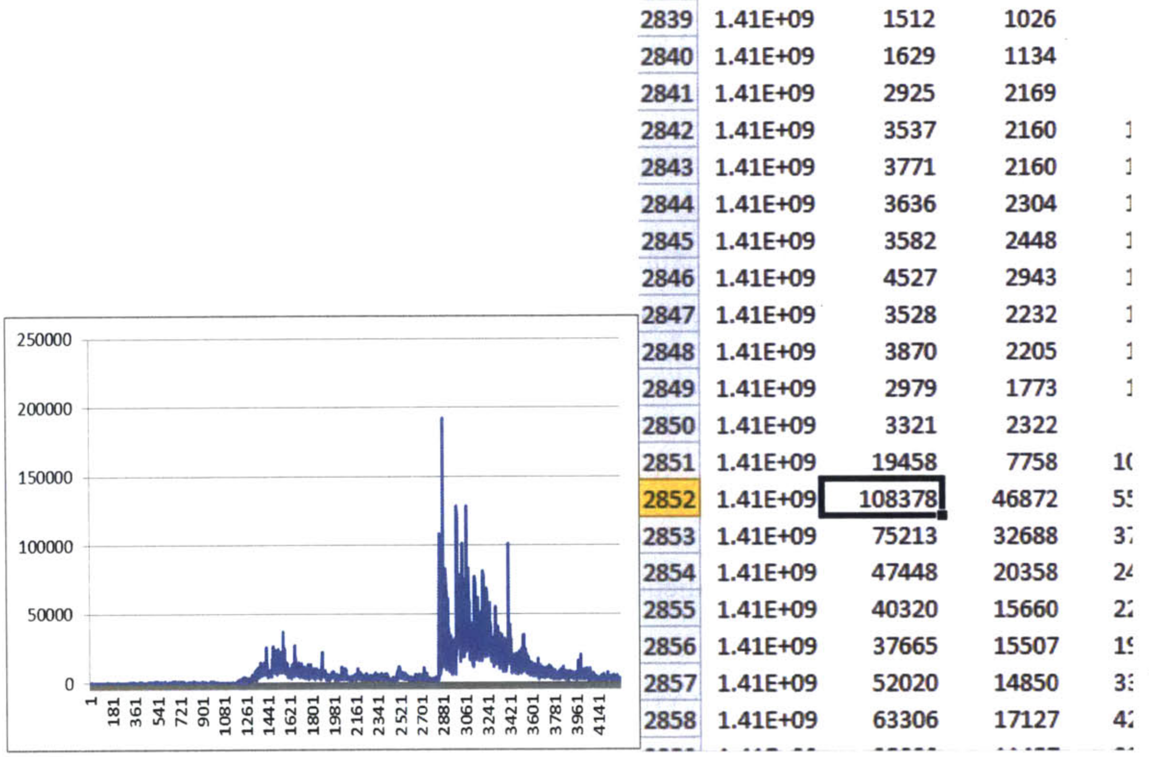
The reason for the first improvement may lie in that the cumulative number of influential users could capture the large external shock of the topic or event. The second improvement may lie in that by smoothing the peaks in the data points, we can fit the smooth bell shape curve better.

However, despite our contributions, we see limitation in this project. Here we will discuss them and future directions.

First, our model fitting accuracy is not high enough, especially with small proportions of data. The main reason of the low accuracy is that the peaks of the data are driven by external influences, which is not captured in our data. We think that without an external data source, it is not possible to predict the development of the mention count well. Our idea of incorporating the influential user term to extend the basic Bass model is one approach to improve model performance given the limitation of our data.

To illustrate the idea of external influence, let us look at one of our collected events. The following figure shows the mention-time graph of event "Ebola." The x axis is the time and the y axis is the mentions. We see that the highest peak of mentions occurs suddenly. We suspect that the sudden increase in mentions is purely driven by an external influence, so we checked the data. In Figure 6.1, we show some data for the event "Ebola." Let us take a closer look at the cell in the black box "108378," which is the highest number of mentions.

Figure 6.1 Ebola Diffusion Pattern



2839	1.41E+09	1512	1026	
2840	1.41E+09	1629	1134	
2841	1.41E+09	2925	2169	
2842	1.41E+09	3537	2160	1
2843	1.41E+09	3771	2160	1
2844	1.41E+09	3636	2304	1
2845	1.41E+09	3582	2448	1
2846	1.41E+09	4527	2943	1
2847	1.41E+09	3528	2232	1
2848	1.41E+09	3870	2205	1
2849	1.41E+09	2979	1773	1
2850	1.41E+09	3321	2322	
2851	1.41E+09	19458	7758	1(
2852	1.41E+09	108378	46872	5!
2853	1.41E+09	75213	32688	3;
2854	1.41E+09	47448	20358	2/
2855	1.41E+09	40320	15660	2:
2856	1.41E+09	37665	15507	1!
2857	1.41E+09	52020	14850	3:
2858	1.41E+09	63306	17127	4:

Before row 2851, the number of mentions varies but is always small. Then suddenly the number of mentions goes 30 times larger in the following two hours and stays at the 40,000 – 70,000 level afterwards. Naturally, we would expect that there was a big news break in the outside world at that time, and so we searched for news around that time. Unsurprisingly, we found that around that time of September 28, 2014, the first confirmed case of Ebola in the United States was admitted into a hospital in Dallas. In other words, the sudden increase in the number of mentions is a result of the media news, and not a result of word of mouth diffusion on Twitter.

Therefore, in order to improve the model, we need to collect the data of external driving forces. Section 2 mentions that researchers incorporated advertising and price into the original Bass model. Our potential future extension can refer to these existing works.

In addition to the lack of data that captures external influences, we also do not have the individual level data, such as user characteristics and network information (e.g., friend adoption, tie strength, etc.). Therefore, we do not have important insights regarding the underlying mechanism of how a topic spreads out in the network in the micro-level. Furthermore, we have not captured the heterogeneity across the Twitter users.

If we can find solutions to these issues, we will be able to strategically apply our model into real-world use. For example, we can influence particular users to make a marketing campaign viral. There are some existing studies that work with individual data, but they lack the understanding of diffusion at the aggregate level, which is undoubtedly important. Therefore, it would be of great interest to explore the combination of an aggregate diffusion model and individual level data.

Lastly, we may try other types of methods, such as other diffusion models or machine learning models, and compare them with our current method. Moreover, it may be of great practical importance to model the decline of the diffusion process, or what happens after the highest mention. For example, the company making a defective product may want to know when the attention will begin to fade and thereby determine the best timing for public relations strategies.

Bibliography

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research*, *51*(3), 249-269.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544-21549.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, *337*(6092), 337-341.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485-1509.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 797-817.

Bass, F. (1969). A New Product Growth Model for Consumer Durables.*Management Sci.*

Bass, F. M. (1980). The relationship between diffusion rates, experience curves, and demand elasticities for consumer durable technological innovations. *Journal of Business*, S51-S67.

Bass, F. M., Krishnan, T. V., & Jain, D. C. (1994). Why the Bass model fits without decision variables. Marketing science, 13(3), 203-223.

Bauckhage, C., Kersting, K., & Rastegarpanah, B. (2014, April). Collective attention to social media evolves according to diffusion models. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (pp. 223-224). International World Wide Web Conferences Steering Committee.

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. Journal of Marketing Research, 49(2), 192-205.

Bewley, R., & Fiebig, D. G. (1988). A flexible logistic growth model with applications in telecommunications. International Journal of Forecasting, 4(2), 177-192.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market.Journal of Computational Science, 2(1), 1-8.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy, 992-1026.

Chang, H. C. (2010). A new perspective on Twitter hashtag use: Diffusion of innovation theory. Proceedings of the American Society for Information Science and Technology, 47(1), 1-4.

Çelen, B., Kariv, S., & Schotter, A. (2010). An experimental test of advice and social learning. Management Science, 56(10), 1687-1701.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. Management Science, 54(3), 477-491.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), 345-354.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. Management Science, 52(10), 1577-1593.

Dockner, E., & Jørgensen, S. (1988). Optimal advertising policies for diffusion models of new product innovation in monopolistic situations. Management Science, 34(1), 119-130.

Easingwood, C. J., Mahajan, V., & Muller, E. (1983). A nonuniform influence innovation diffusion model of new product acceptance. Marketing Science, 2(3), 273-295.

Easingwood, C., Mahajan, V., & Muller, E. (1981). A nonsymmetric responding logistic model for forecasting technological substitution. Technological forecasting and Social change, 20(3), 199-213.

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. European economic review, 46(4), 687-724.

Fershtman, C., & Gandal, N. (2007). Open source software: Motivation and restrictive licensing. International Economics and Economic Policy, 4(2), 209-225.

Fisher, J. C., & Pry, R. H. (1972). A simple substitution model of technological change. Technological forecasting and social change, 3, 75-88.

Glazer, A., & Konrad, K. A. (1996). A signaling explanation for charity. The American Economic Review, 1019-1028.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. Marketing science, 23(4), 545-560.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., ... & Verlegh, P. (2005). The firm's management of social interactions. Marketing Letters, 16(3-4), 415-428.

Goldenberg, J., Oestreicher-Singer, G., & Reichman, S. (2012). The quest for content: How user-generated links can facilitate online exploration. Journal of Marketing Research, 49(4), 452-468.

Harbaugh, W. T. (1998). The prestige motive for making charitable transfers.American Economic Review, 277-282.

Harbaugh, W. T. (1998). What do donations buy?: A model of philanthropy based on prestige and warm glow. Journal of Public Economics, 67(2), 269-284.

Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing?. Sloan Management Review, 52(1).

Horsky, D. (1990). The effects of income, price and information on the diffusion of new consumer durables. Marketing Science, 9(4), 342-365.

Horsky, D., & Simon, L. S. (1983). Advertising and the diffusion of new products. Marketing Science, 2(1), 1-17.

Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D. & Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. Marketing letters, 19(3-4), 287-304.

Hsu, C. L., & Lin, J. C. C. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation.Information & Management, 45(1), 65-74.

Kamakura, W. A., & Balasubramanian, S. K. (1988). Long-term view of the diffusion of durables A study of the role of price and adoption influence processes via tests of nested models. International Journal of Research in Marketing, 5(1), 1-13.

Kalish, S. (1985). A new product adoption model with price, advertising, and uncertainty. Management science, 31(12), 1569-1585.

Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. Journal of Marketing Research, 48(3), 425-443.

Kempe, D., Kleinberg, J., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 137-146). ACM.

Lampel, J., & Bhalla, A. (2007). The role of status seeking in online communities: Giving the gift of experience. Journal of Computer-Mediated Communication, 12(2), 434-455.

Lampel, J., & Bhalla, A. (2007). The role of status seeking in online communities: Giving the gift of experience. Journal of Computer-Mediated Communication, 12(2), 434-455.

Li, C., Sun, A., & Datta, A. (2012, October). Twevent: segment-based event detection from tweets. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 155-164). ACM.

Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in twitter. Journal of the American Society for Information Science and Technology, 64(7), 1399-1410.

Mahajan, V., Muller, E., & Kerin, R. A. (1984). Introduction strategy for new products with positive and negative word-of-mouth. Management Science,30(12), 1389-1404.

Manski, C. F. (2000). Economic analysis of social interactions (No. w7580). National bureau of economic research.

Mayzlin, D. (2006). Promotional chat on the Internet. Marketing Science, 25(2), 155-163.

Mayzlin, D., Dover, Y., & Chevalier, J. A. (2012). Promotional reviews: An empirical investigation of online review manipulation (No. w18340). National Bureau of Economic Research.

Mayzlin, D., & Yoganarasimhan, H. (2012). Link to success: How blogs build an audience by promoting rivals. Management Science, 58(9), 1651-1668.

Meade, N., & Islam, T. (1998). Technological forecasting—Model selection, model stability, and combining models. Management Science, 44(8), 1115-1130.

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. Journal of Marketing Research, 48(3), 444-456.

Miklos-Thal, J., & Zhang, J. (2013). (De) marketing to Manage Consumer Quality Inferences. Journal of Marketing Research, 50(1), 55-69.

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

Putsis, W. P. (1998). Parameter variation and new product diffusion. Journal of Forecasting, 17(3-4), 231-257.

Radas, S. (2006). Diffusion Models in Marketing: How to Incorporate the Effect of External Influence?. Privredna kretanja i ekonomska politika, 15(105), 30-51.

Robinson, B., & Lakhani, C. (1975). Dynamic price models for new-product planning. Management science, 21(10), 1113-1122.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist, 55(1), 68.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. science,311(5762), 854-856.

Schelling, T. C. (2006). Micromotives and macrobehavior. WW Norton & Company.

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. Journal of Consumer Research, 32(2), 260-265.

Schweidel, D. A., & Moe, W. W. (2014). Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice. Journal of Marketing Research,51(4), 387-402.

Schultz, F., Utz, S., & Göritz, A. (2011). Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media. Public relations review, 37(1), 20-27.

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?. Journal of Interactive Marketing, 26(2), 102-113.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In Social computing (socialcom), 2010 ieee second international conference on (pp. 177-184). IEEE.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events.Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. Marketing Science, 31(2), 198-215.

Toubia, O., & Stephen, A. T. (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter?. Marketing Science, 32(3), 368-392.

Tucker, C., & Zhang, J. (2011). How does popularity information affect choices? A field experiment. Management Science, 57(5), 828-842.

Tucker, C., & Zhang, J. (2010). Growing two-sided networks by advertising the user base: A field experiment. Marketing Science, 29(5), 805-814.

Yang, X., Li, Y., Tan, C. H., & Teo, H. H. (2007). Students' participation intention in an online discussion forum: Why is computer-mediated interaction attractive?. Information & Management, 44(5), 456-466.

Zaman, T., Fox, E. B., & Bradlow, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. The Annals of Applied Statistics, 8(3), 1583-1611.

Zhang, J. (2010). The sound of silence: Observational learning in the US kidney market. Marketing Science, 29(2), 315-335.

Zhang, J. (2012). Observational Learning: The Sound of Silence. InEncyclopedia of the Sciences of Learning (pp. 2493-2496). Springer US.

Zhang, J., & Liu, P. (2012). Rational herding in microloan markets.Management science, 58(5), 892-912.

Appendix

A. Codes for Data Collection

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%main function
clear; clc;
infile = 'eventlist_peak.xlsx';
% [num,txt,raw] = xlsread(infile,'all');
% clear num; clear raw;
outfile = 'all_features2.xlsx';
slice_str = 'hour';


[num,evtlist,raw] = xlsread(infile,'for_collect'); % evtlist c1-event_name,c2-start_time,c3-
end_time
no_evt = size(evtlist,1);
clear num; clear raw;


for idx_evt = 1:no_evt
   idx_evt
   query=evtlist{idx_evt,1};
   mintime = evtlist{idx_evt,2};
   maxtime = evtlist{idx_evt,3};

   metrics = all_features(query,mintime,maxtime,slice_str);

   xls_delete_sheets(outfile,num2str(idx_evt));
   xlswrite(outfile, metrics, num2str(idx_evt));

end



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%tweet_count function
% return time and tweet count ;
% collect daily metrics every 170 days, collect minute metrics every 3 days
% input time in the format of [year, month, day]

function metrics = tweet_count(query, mintime, maxtime, slice_str)

SECONDS_PER_DAY = 86400;
urlroot_senti = 'http://api.topsy.com/v2/metrics/sentiment.json?';
urlroot_menion =  'http://api.topsy.com/v2/metrics/mentions.json?';
key = '&apikey=09C43A9B270A470B8EB8F2946A9369F3';
```

```
query_encode = ['q=' urlencode(query)];

mintime_unix =  double(floor(86400 * (datenum(mintime(1),mintime(2),mintime(3)) -
datenum('01-Jan-1970'))));
maxtime_unix =  double(floor(86400 * (datenum(maxtime(1),maxtime(2),maxtime(3)) -
datenum('01-Jan-1970'))));
if strcmp(slice_str, 'min')
  slice = 60;
  timespan = SECONDS_PER_DAY * 1;
else
  slice = 3600;
  timespan = SECONDS_PER_DAY * 31;
end

metrics = zeros(7,ceil((maxtime_unix - mintime_unix)/slice));
start_idx = 1;

for piece_idx = 1:ceil((maxtime_unix - mintime_unix)/timespan)

  piece_idx

  %% read mentions and sentiment
  % scrape data
  mintime_str = num2str(mintime_unix + timespan * (piece_idx-1));
  maxtime_str = num2str(min(mintime_unix + timespan * piece_idx - 1, maxtime_unix));
  time = ['&mintime=' mintime_str '&maxtime=' maxtime_str '&slice=' num2str(slice)];
  url = [urlroot_senti query_encode time key '&include_mentions=1'];
  resp_str = urlread(url);
  resp_obj = parse_json(resp_str);

  % output data
  mentions = resp_obj{1,1}.response.results{1,1}.data;
  for resp_idx = 1:length(mentions)
    metrics(1,start_idx + resp_idx - 1) = mentions{1,resp_idx}.timestamp; % unix timestamp
    metrics(2,start_idx + resp_idx - 1) = mentions{1,resp_idx}.mentions; % mentions
    metrics(3,start_idx + resp_idx - 1) = mentions{1,resp_idx}.sentiment_score; % sentiment
  end

  if sum(metrics(2, start_idx:start_idx + length(mentions) - 1)) > 0
    %% root tweets
    url = [urlroot_menion query_encode time key '&tweet_types=tweet'];
    resp_str = urlread(url);
    resp_obj = parse_json(resp_str);

    tweet = resp_obj{1,1}.response.results{1,1}.data;
    for resp_idx = 1:length(tweet)
```

```
      metrics(4,start_idx + resp_idx - 1) = tweet{1,resp_idx}.mentions; % root tweet
    end

    %% retweet and reply
    url = [urlroot_menion query_encode time key '&tweet_types=retweet'];
    resp_str = urlread(url);
    resp_obj = parse_json(resp_str);

    retweet = resp_obj{1,1}.response.results{1,1}.data;
    for resp_idx = 1:length(retweet)
      metrics(5,start_idx + resp_idx - 1) = retweet{1,resp_idx}.mentions ; % retweet
      metrics(6,start_idx + resp_idx - 1) = metrics(2,start_idx + resp_idx - 1) -
metrics(4,start_idx + resp_idx - 1) - metrics(5,start_idx + resp_idx - 1); % reply
    end



    %% influential users only
    url = [urlroot_menion query_encode time key  '&infonly=1'];
    resp_str = urlread(url);
    resp_obj = parse_json(resp_str);

    infonly = resp_obj{1,1}.response.results{1,1}.data;
    for resp_idx = 1:length(tweet)
      metrics(7,start_idx + resp_idx - 1) = infonly{1,resp_idx}.mentions  ; % influential users
only
    end

  end

  start_idx = start_idx + length(mentions)
  L = length(mentions)
  piece_idx
end


%% delete zeros at the start
cols = size(metrics,2);
for start_pos = 1:cols
  if metrics(2,start_pos) ~= 0
    break;
  end
end
start_pos
metrics = transpose(metrics(:,start_pos:cols));
```

```
end


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%cum metrics function
function cum = cum_metrics(in_mtr)
% col 1 - unix timestamp
% col 2 - mentions
% col 3 - sentiment
% col 4 - root tweet
% col 5 - retweet
% col 6 - reply
% col 7 - influential users only

rows = size(in_mtr,1);
cum = in_mtr;
for i = 2:rows
   cum(i,2) = cum(i-1,2) + in_mtr(i,2);
   cum(i,4) = cum(i-1,4) + in_mtr(i,4);
   cum(i,5) = cum(i-1,5) + in_mtr(i,5);
   cum(i,6) = cum(i-1,6) + in_mtr(i,6);
end

end
```

B. Codes for Bass Model

```
%%Basic Bass Model Analysis
clear;
%read in sheets one by one
data1=csvread('ipl.csv');

%% data
size=size(data1,1);
%exclude those zero influential users
start=1;
for i=1:size
if data1(i,11)~=0
   start=i;
   break;
end
end
e=size;
```

```
time=data1(start:e,1);
men=data1(start:e,2);
cmen=data1(start:e,3);
rt=data1(start:e,4);
crt=data1(start:e,5);
re=data1(start:e,6);
cre=data1(start:e,7);
rp=data1(start:e,8);
crp=data1(start:e,9);
inf=data1(start:e,10);
cinf=data1(start:e,11);
sent=data1(start:e,12);
size=e-start+1;
disp('subset of data');
disp([start, e, size]);

% aggregate data by agg times, men, cmen, cinf
agg=24;
k=1;
kn=1;
size1=round(size/agg)+1;
menn=zeros(size1,1);
infn=zeros(size1,1);
cmenn=zeros(size1,1);
cinfn=zeros(size1,1);
while (k<=size)
    menn(kn)=menn(kn)+men(k);
    infn(kn)=infn(kn)+inf(k);
  if mod(k,agg)==0
    kn=kn+1;
  end
    k=k+1;
end
cmenn(1)=menn(1);
cinfn(1)=infn(1);
for i=2:size1
  cmenn(i)=cmenn(i-1)+menn(i);
  cinfn(i)=cinfn(i-1)+infn(i);
end


%% Basic Model
% Use different percent of data
testper=10;
step=0.1;
per=ones(testper,1);
```

```
for i=2:length(per)
   per(i)=per(i-1)-step;
end
disp(per);
display('start');

for t=1:length(per)
   % estimate p, q, m
   sizen=round(size1*per(t));

   n=menn(1:sizen);
   N=cmenn(1:sizen);
   N2=N.*N;
   X=[ones(sizen,1),N,N2];
   b=regress(n,X);
   disp('percent')
   disp(per(t));
   disp('size, max cum')
   display([sizen,max(N)]);
   disp('b');
   disp([b(1),b(2)]);
   disp(b(3));
   b=abs(b);
   syms p q m
   S=solve([p*m==b(1),q-p==b(2),q/m==b(3)],[p,q,m]);
   if (S.p(2)>S.p(1))
      p=double(S.p(2));
   else
      p=double(S.p(1));
   end
   m=b(1)/p;
   q=b(3)*m;

   disp('parameters')
   display([p,q]);
   disp(m);
   N1=zeros(size1,1);
   n1=zeros(size1,1);
   for i=2:size1
     n1(i)=abs((p+q/m*N1(i-1))*(m-N1(i-1)));
     N1(i)=n1(i)+N1(i-1);
   end
   scn1=round(max(cmenn)/max(n1));
   scn=round(max(cmenn)/max(n));
   disp('scale');
   disp([scn1,scn]);
   [peak,peakpos]=max(n);
   [peak1,peakpos1]=max(n1);
   disp('peak');
   disp([peakpos,peakpos1]);


   h=figure;
   plot(1:size1,cmenn,1:size1,n1*scn1,1:sizen,n*scn);
   title(['Aggregated Basic Bass model prediction with first ' num2str(per(t)*100) '% of data'],
'FontSize', 12)
   xlabel('t', 'FontSize', 12) % x-axis label
   ylabel('number of mention', 'FontSize', 12) % y-axis label
   h_legend = legend('actual cum mention','predicted mention (rescaled)',...
   'actual mention (rescaled)','Location','SouthEast');
   set(h_legend,'FontSize',8);
   saveas(h,num2str(t),'jpg')
   h=figure('visible','off');
end
```

C. Codes for Extended Bass Model

```
%%Extended Bass Model Analysis
clear;
data1=csvread('ipl.csv');

% data
size=size(data1,1);
%exclude those zero influential users
start=1;
for i=1:size
if data1(i,11)~=0
   start=i;
   break;
end
end
e=size;
time=data1(start:e,1);
men=data1(start:e,2);
cmen=data1(start:e,3);
rt=data1(start:e,4);
crt=data1(start:e,5);
re=data1(start:e,6);
cre=data1(start:e,7);
```

```
rp=data1(start:e,8);
crp=data1(start:e,9);
inf=data1(start:e,10);
cinf=data1(start:e,11);
sent=data1(start:e,12);
size=e-start+1;
disp('subset of data');
disp([start, e, size]);

% aggregate data by agg times, men, cmen, cinf
agg=24;
k=1;
kn=1;
size1=round(size/agg)+1;
menn=zeros(size1,1);
infn=zeros(size1,1);
cmenn=zeros(size1,1);
cinfn=zeros(size1,1);
while (k<=size)
    menn(kn)=menn(kn)+men(k);
    infn(kn)=infn(kn)+inf(k);
  if mod(k,agg)==0
    kn=kn+1;
  end
    k=k+1;
end

cmenn(1)=menn(1);
cinfn(1)=infn(1);
for i=2:size1
  cmenn(i)=cmenn(i-1)+menn(i);
  cinfn(i)=cinfn(i-1)+infn(i);
end


%% Basic Model
% Use different percent of data
testper=10;
step=0.1;
per=ones(testper,1);
for i=2:length(per)
   per(i)=per(i-1)-step;
end
disp(per);
display('start');
```

```
for t=1:length(per)
  % estimate p, q, m
  sizen=round(size1*per(t));

  n=menn(1:sizen);
  N=cmenn(1:sizen);
  I=log(cinfn(1:sizen));
  N2=N.*N;
  NI=N.*I;
  N2I=N2.*I;
  X=[I,NI,N2I];
  b=regress(n,X);
  disp('percent')
  disp(per(t));
  disp('size, max cum')
  display([sizen,max(N)]);
  disp('b');
  disp([b(1),b(2)]);
  disp(b(3));
  b=abs(b);
  syms p q m
  S=solve([p*m==b(1),q-p==b(2),q/m==b(3)],[p,q,m]);
  if (S.p(2)>S.p(1))
    p=double(S.p(2));
  else
    p=double(S.p(1));
  end
  m=b(1)/p;
  q=b(3)*m;

  disp('parameters')
  display([p,q]);
  disp(m);
  N1=zeros(size1,1);
  n1=zeros(size1,1);
  for i=2:size1
   n1(i)=abs((p+q/m*N1(i-1))*(m-N1(i-1))*log(cinfn(i-1)));
   N1(i)=n1(i)+N1(i-1);
  end
  scn1=round(max(cmenn)/max(n1));
  scn=round(max(cmenn)/max(n));
  disp('scale');
  disp([scn1,scn]);
  [peak,peakpos]=max(n);
  [peak1,peakpos1]=max(n1);
  disp('peak');
```

```
   disp([peakpos,peakpos1]);

   h=figure;
   plot(1:size1,cmenn,1:size1,n1*scn1,1:sizen,n*scn);
   title(['Aggregated Extended Bass model prediction with first ' num2str(per(t)*100) '% of data'],
'FontSize', 12)
   xlabel('t', 'FontSize', 12) % x-axis label
   ylabel('number of mention', 'FontSize', 12) % y-axis label
   h_legend = legend('actual cum mention','predicted mention (rescaled)',...
   'actual mention (rescaled)','Location','SouthEast');
   set(h_legend,'FontSize',8);
   saveas(h,num2str(t),'jpg')
   h=figure('visible','off');

end
```

D. List of Topics Used for Data Collection

Topic ID	Topic Name (denoted by a hashtag or keywords)	Peak Time	Total Mention Count
1	2Chainz	07-May-2012	1250460
2	#AintNobodyGotTimeForThat	10-Jul-2012	294508
3	#ratchet	07-Nov-2012	483155
4	#ChiefKeefMakesMusicFor	17-Sep-2012	7873
5	#coolstorybro	18-Mar-2012	66080
6	#Struggle	07-Nov-2012	120256
7	#TurntUp	27-Oct-2012	331318
8	#yolo	26-Mar-2012	5190369
9	#ThatShitIDontLike	05-Sep-2012	66505
10	#hirihanna	12-Oct-2011	11601
11	#ModernSeinfeld	12-Dec-2012	771
12	#DrakesMusicWillHaveYou	17-Dec-2012	55363
13	#endoftheworldconfessions	21-Dec-2012	581195
14	#2012regrets	27-Dec-2012	140775
15	#MSL	06-Aug-2012	383626
16	#EDL	01-Sep-2012	117272
17	#Sandy	30-Oct-2012	4600864
18	#ISS	30-Oct-2012	104387
19	#Synchro	05-Aug-2012	4324

20	#Swimming	29-Jul-2012	265618
21	#Endeavour	21-Sep-2012	53826
22	#spottheshuttle	21-Sep-2012	73274
23	#austin	18-Nov-2012	220137
24	#Bengals	07-Jan-2012	141095
25	#WhoDey	07-Jan-2012	62526
26	#pandaAI	30-Apr-2012	1167
27	#askneil	24-Oct-2012	5726
28	#bondtweets	23-Oct-2012	654
29	#london2012	27-Jul-2012	5212299
30	#oneweb	27-Jul-2012	9153
31	#openingceremony	27-Jul-2012	678721
32	#blur	02-Jul-2012	39524
33	#RoyalBaby	03-Dec-2012	140523
34	#KONY2012	07-Mar-2012	2305934
35	#StopKony	07-Mar-2012	2094112
36	#TwitternWie1989	09-Nov-2012	3358
37	#JournalistBerlin	09-Nov-2012	286
38	#Houla	27-May-2012	30123
39	#Damascus	18-Jul-2012	316768
40	#Syria	18-Jul-2012	7625671
41	#AskPele	27-Jun-2012	1379
42	#WorldCup	11-Sep-2012	63736
43	#PrayForMuamba	17-Mar-2012	441350
44	#CFC	28-Oct-2012	2846990
45	#VMARedCarpet	06-Sep-2012	67648
46	#nfltotalaccess	05-Feb-2012	6917
47	#summerwars	20-Jul-2012	97225
48	#ntv	20-Jul-2012	705685
49	Obama	07-Nov-2012	51927099
50	Gulf Oil Spill	30-Apr-2010	450756
51	Haiti Earthquake	13-Jan-2010	429935
52	Pakistan Floods	27-Aug-2010	44576
53	Koreas Conflict	23-Nov-2010	394
54	Chilean Miners Rescue	13-Oct-2010	38100
55	Chavez Tas Ponchao	25-Jan-2010	9222
56	Wikileaks Cablegate	10-Dec-2010	14081
57	Hurricane Earl	31-Aug-2010	149641

58	Prince Williams Engagement	16-Nov-2010	4290
59	World Aids Day	01-Dec-2010	153741
60	Apple iPad	27-Jan-2010	852505
61	Google Android	20-May-2010	366282
62	Apple iOS	22-Nov-2010	227047
63	Apple iPhone	07-Jun-2010	1645107
64	Call of Duty Black Ops	09-Nov-2010	430875
65	New Twitter	28-Sep-2010	2485215
66	HTC	15-Sep-2010	1250908
67	RockMelt	08-Nov-2010	120515
68	MacBook Air	20-Oct-2010	601737
69	Google Instant	08-Sep-2010	242225
70	#rememberwhen	22-Nov-2010	346771
71	#slapyourself	17-Nov-2010	274294
72	#confessiontime	07-Nov-2010	312491
73	#thingsimiss	05-Dec-2010	293510
74	#ohjustlikeme	20-Nov-2010	49351
75	#wheniwaslittle	10-Aug-2010	739665
76	#haveuever	17-Nov-2010	111262
77	#icantlivewithout	04-Nov-2010	157055
78	#thankful	25-Nov-2010	340865
79	#2010disappointments	02-Dec-2010	195233
80	toyota and recall	29-Jan-2010	9173
81	GM and switch	30-Mar-2014	2007
82	Infantino and baby sling	24-Mar-2010	46
83	#myNYPD	22-Apr-2014	142702
84	Boston marathon	15-Apr-2013	2470570
85	SONY and Korea	24-Dec-2014	38577
86	burger king and lettuce	19-Jul-2012	391
87	horsemeat	15-Jan-2013	339336
88	#McDStories	18-Jan-2012	25178
89	iwatch	09-Sep-2014	872266
90	Amazon Fire Phone	18-Jun-2014	283652
91	XBox One	02-Dec-2014	4877345
92	Kraft belVita	11-Nov-2011	189
93	Ebola	02-Oct-2014	37062214
94	Senator proposal	08-Mar-2013	2598
95	Japan Secrecy Bill Law	06-Dec-2013	480

96	Virginia AG recount	26-Nov-2013	1313
97	New Pontifex	13-Mar-2013	1291
98	IPL tournament	21-May-2013	6246
99	Castle in the Sky	25-Aug-2013	13460
100	Northern India flooding	21-Jun-2013	1688
101	Brazilian protests	22-Jun-2013	15396
102	Vine resume	21-Feb-2013	2869
103	Primetime Emmys	23-Sep-2013	8679
104	DOMA Prop8	26-Jun-2013	4843
105	Sochi Olympics	18-Dec-2013	686932
106	Salvage Costa Concordia	16-Sep-2013	20988
107	Italy election	26-Feb-2013	43998
108	World Youth Day	28-Jul-2013	35591
109	#hochwasser	03-Jun-2013	154423
110	Australian election	07-Sep-2013	40858
111	Kobe Cuban	24-Feb-2013	11516
112	German election	22-Sep-2013	32456
113	OneDirection	26-Aug-2013	544327
114	#aufschrei	25-Jan-2013	108686
115	Fashion Week	15-Feb-2013	1097929
116	Asiana 214	06-Jul-2013	55804
117	50th anniversary March Washington	28-Aug-2013	59646
118	#IranTalks	24-Nov-2013	47603
119	#NobelPeacePrize	11-Oct-2013	26398
120	France World Cup	19-Nov-2013	71267
121	Dilma	06-Sep-2013	3217901
122	typhoon Philippines	09-Nov-2013	679418
123	Red panda	24-Jun-2013	88255
124	#Troon	17-Apr-2013	169696
125	Australian Open	27-Jan-2013	367373
126	Tour de France	21-Jul-2013	687379
127	Academy Awards	25-Feb-2013	293091
128	#RockInRio	16-Sep-2013	460624
129	MTV VMAs	26-Aug-2013	254083
130	#Ashes	17-Dec-2013	1153050
131	#MalalaDay	12-Jul-2013	92522
132	#MariagePourTous	23-Apr-2013	898580
133	March Madness	21-Mar-2013	1163560

134	Thatcher death	08-Apr-2013	157218
135	jason collins gay	29-Apr-2013	259639
136	#stanleycup	25-Jun-2013	546988
137	#Inauguration	21-Jan-2013	170746
138	#DoctorWho	23-Nov-2013	1828363
139	#ThankYouSachin	15-Nov-2013	1116119
140	#SFBatkid	15-Nov-2013	285846
141	#RIPMandela	05-Dec-2013	159670
142	World Cup Draw	06-Dec-2013	255758
143	#ThankYouSirAlex	08-May-2013	678244
144	#StandWithRand	07-Mar-2013	489851
145	#2013MAMA	22-Nov-2013	545670
146	#UCLFinal	25-May-2013	437898
147	#PLL	28-Aug-2013	2495315
148	#Sharknado	12-Jul-2013	547256
149	Government shutdown	01-Oct-2013	1705491
150	#StandWithWendy	26-Jun-2013	495244
151	#SB47	04-Feb-2013	568084
152	#NBAFinals	21-Jun-2013	1523961
153	Wimbledon	07-Jul-2013	2015368
154	Eurovision	18-May-2013	1693285
155	#RoyalBaby	22-Jul-2013	1592166
156	#NewYearsEve	01-Jan-2014	224751
157	#IPL	01-Jun-2014	374769
158	#Carnaval	01-Mar-2014	477170
159	#RIPPhilipSeymourHoffman	02-Feb-2014	50326
160	#SuperBowl	03-Feb-2014	2755775
161	#Oscars	03-Mar-2014	4251433
162	#UmbrellaRevolution	03-Oct-2014	277261
163	#iVoted	04-Nov-2014	28120
164	#Abdicates	05-Jun-2014	274
165	#IndiaVotes	05-Mar-2014	8172
166	#wt20	06-Apr-2014	695129
167	#Alia	06-Jan-2015	5528
168	#Wimbledon	06-Jul-2014	907759
169	#USOpen	06-Sep-2014	819176
170	#Sochi2014	07-Feb-2014	5069375
171	#BCSChampionship	07-Jan-2014	273487

172	#WorldCup	08-Jul-2014	20084808
173	#FrenchOpen	08-Jun-2014	115527
174	#Coachella	09-Jan-2014	76383
175	#TDF14	09-Jul-2014	18734
176	#BerlinWall	09-Nov-2014	77477
177	#NYFW	09-Sep-2014	653920
178	#BringBackOurGirls	10-May-2014	3868601
179	#MalalaYousafzai	10-Oct-2014	152659
180	#ThanksLD	10-Oct-2014	103118
181	#Spain2014	10-Sep-2014	953698
182	#RIPRobinWilliams	11-Aug-2014	2757478
183	#ComingHome	11-Jul-2014	29180
184	#CometLanding	12-Nov-2014	570908
185	#GoldenGlobes	13-Jan-2014	1618349
186	#GermanyWins	13-Jul-2014	1468
187	#Ferguson	14-Aug-2014	9921883
188	#TheVoiceAU	14-Jul-2014	221374
189	#StanleyCup	14-Jun-2014	490531
190	#NBAFinals	16-Jun-2014	1333518
191	#Formula1	16-Mar-2014	273901
192	#NBAAllStar	17-Feb-2014	375858
193	#MH17	17-Jul-2014	4044279
194	#OnlyOnTwitter	18-Feb-2015	2707
195	#IceBucket Challenge	19-Aug-2014	22373
196	#BRITAwards	19-Feb-2014	45761
197	#LoveTheatre	19-Nov-2014	46692
198	#IndyRef	19-Sep-2014	5440172
199	#NFLPlayoffs	20-Jan-2014	586196
200	#FirstTweet	20-Mar-2014	357941
201	#MarchMadness	21-Mar-2014	1574794
202	#Glasgow2014	23-Jul-2014	811096
203	#ISS	23-Nov-2014	492621
204	#MuseumWeek	24-Mar-2014	203938
205	#MH370	24-Mar-2014	4837319
206	#Cannes2014	24-May-2014	720491
207	#MarsOrbiter	24-Sep-2014	16032
208	#VMAs	25-Aug-2014	3082229
209	#HeForShe	25-Sep-2014	812374

210	#PhotoshopRF	25-Sep-2014	11479
211	#Emmys	26-Aug-2014	863563
212	#AusOpen	26-Jan-2014	804617
213	#Eleições2014	26-Oct-2014	411975
214	#DerekJeter	26-Sep-2014	149596
215	#Grammys	27-Jan-2014	3454718
216	#RIPMaya Angelou	28-May-2014	2705
217	#PutOutYourBats	28-Nov-2014	147452
218	#ModiInAmerica	28-Sep-2014	93461
219	#SOTU	29-Jan-2014	1391601
220	#WorldSeries	30-Oct-2014	1137844