Analyzing the Impact of Social TV Strategies on Viewer Engagement

Shawndra Hill  
University of Pennsylvania  
The Wharton School  
3730 Walnut Street  
Philadelphia, PA 19104  
shawndra@wharton.upenn.edu

Adrian Benton  
University of Pennsylvania  
The Wharton School  
3730 Walnut Street  
Philadelphia, PA 19104  
adrianb@sas.upenn.edu

ABSTRACT

“Social TV” is a term that broadly describes the online social interactions that occur between viewers while watching television. Television shows are now instigating online social interactions between viewers by requesting viewers, as part of the first broadcasts, to engage in simultaneous discussions about the shows. Examples include explicitly asking users to follow and comment on topics on the popular social media site, Twitter. A more passive approach used by TV shows displays messages from social media users on the TV screen during broadcasts. While many shows now include social media in their broadcasts, very little is known about the effectiveness of these strategies on engagement and viewership. In this work, we analyze features of the social media strategy of an American reality singing show. We show that when messages are posted on the TV screen during the show, they are much more likely to be discussed by viewers than messages created by the same online social media users during the show that are not posted on the screen, indicating that this strategy works. We also show that when a prompt is given to use a specific term in discussions online, viewers are more likely to remain engaged with the show on social media. We believe this is the first scholarly work to link real-time TV posts to real-time TV engagement response. Our contribution is a set of results, which show that prompting viewers during the show with social media posts on screen results in increased and prolonged engagement during the show.

1. INTRODUCTION

"Social TV" is the term used to describe the current integration of social media interaction with television programming. Social television has sought to recapture the early days of television, when families gathered in their homes to share the experience of watching television together [1]. Over the past several years, online social media communities such as message boards, Twitter, and Facebook have become the new virtual water cooler for today's tech-savvy television viewers. With the proliferation of social media applications and Smartphone technology, social interaction around television programming can now be shared amongst millions of viewers simultaneously. It is estimated that on average, 10 million public online comments are made each day worldwide related to television content. [2]. Twitter and other social media platforms have "become an integral outlet for TV viewers who are looking to express themselves while watching broadcasts of their favorite television programs." This "backchannel" of communication during TV shows has also led to the resurgence of people's interest in watching live shows [3].

For the first time in history, advertisers and TV programmers are able to receive real-time feedback in the form of not only viewership numbers, but also sentiment from large audiences about their products and ads. In addition, networks are able to capture detailed comments from viewers throughout a television show and therefore can ask and try to answer important questions, such as how much engagement specific types of programming can elicit from viewers. If an announcer tells a consumer to go to a particular Web site, to send a tweet, or to purchase a particular item, does that influence the consumer to actually comply? The response is both immediate and measurable in terms of the level of buzz. However, deriving answers to the important questions above is not always straightforward.

While the amount of data generated by users in the context of TV is enormous and convenient for data mining and business analytics, the problem is that the raw data are a noisy stream of consciousness. For researchers and firms alike, the prevailing question is, how can one make sense of and derive value from it all? Companies such as Bluefin Labs, Networked Insights, Social Guide, and Trendrr specialize in analyzing these data and generating metrics for clients interested in their "online image." However, it is unclear whether the current metrics used to quantify these online social communities' engagement in a program actually correlates with viewership or sales, let alone drives it. Sophisticated tools are needed to extract meaningful text-based features to be used in business decision making. The challenge for researchers and industry is to be able to move beyond merely tracing the amount of buzz to providing precise content analysis of what is being said by potential consumers in response to what is being shown on TV.

In this paper, we present work aimed at quantifying the impact of social TV exposure on viewership and engagement in the program, as evidenced by the amount of discussion about it. We focus on a popular American reality singing talent show, which made headlines last season (2011) due to its use of social media content within the show. Tweets and Facebook messages are frequently displayed on the show during the broadcast to keep viewers engaged. These displays often ask the audience a question. The audience then engages using social media outlets such as Twitter and Facebook. In our work, we analyze several aspects of such tweets.

In order to extract information from these varied and rich data, we draw on techniques from natural language processing, network analysis, and time series analysis. We present preliminary answers to the following question using rigorous content and statistical analysis: When are social TV strategies correlated with increased social network buzz? We show that the specific content of the television program has a strong relationship with content on the online community's response and...
sentiment. To our knowledge, this is the first academic research linking social TV strategy to explicit online outcomes.

2. LITERATURE REVIEW
Since the 1950s, television technologies have become an increasingly integral part of the home. As seen in a Rockwellian portrait, families and neighbors would gather around the television to share the viewing experience together. As television grew more popular, increasing numbers of people tended to share the experience with others. The so-called “water cooler effect” had co-workers across all industries collectively discussing what had been on television last night [4]. As televisions became smaller and cheaper, households were able to purchase more than one. Television viewing shifted from being a shared experience to more of an individual one [5]. “Time-shifting” technologies such as DVR, streaming online video, and video on demand also directed television toward individualism, taking away any social sense of the experience [4].

Past attempts to develop social television ranged from the very cumbersome to the innovative. In the early 1980s, Zenith’s Spacephone incorporated a speakerphone system into the television set, allowing the viewer to make a phone call through the television via the remote control, all while still watching the television [6]. AOL offered AOLTV, which uses a thin client computer adapted to display on the television, allowing the viewer to surf the Internet, check email, chat, and watch television all on the same screen [6]. These and prior attempts were hindered largely in part by inadequate technology. However, the development and proliferation of laptops, smartphones, and tablets, along with new innovations in social media such as Twitter and Facebook, have drawn people back to real-time viewing of television programs and sparked a virtual social community around television content.

Currently, social television operates with the aim of encouraging people to watch television live so that they can participate in that specific social experience [7]. Viewers watch television programs and make comments about the shows using social media platforms such as Twitter and Facebook. Researchers and industry have attempted to analyze the buzz generated from viewers by examining a variety of metrics, including who is tweeting, what they are tweeting about, whether they are interacting with one another, and overall, how frequently people are tweeting. What is occurring are large-scale, real-time conversations amongst geo-displaced audiences that are mimicking the nostalgic past of the family gathering around the television [8].

With this real-time interactivity, television programs can be tailored on a micro level to respond to viewer input. However, there is also a greater potential for the interactivity that social television has now provided. Television programmers now have the capability to try and influence viewer behavior using social media by encouraging them to perform specific tasks, such as respond to questions asked during the show, visit a particular Web site, or buy a particular product. Never before has there been the ability to measure, in real-time, the effectiveness of such behavioral influence on television. Similar techniques have been applied linking activity on social media to real world events. One salient example of this is work demonstrating a relationship between sentiment of Twitter posts and changes in season and time of day [9]. However, to our knowledge, there has been little work on demonstrating a relationship between television content and activity on online social media.

In this work, we track buzz in real-time response to what is said on TV. We evaluate two social TV strategies regularly used by the show The Voice. The first strategy is the placement of specific tweets on the screen during the show. We will test four hypotheses related to the content of the tweets that are displayed. The second is the placement of the #thevoice, which is less obtrusive, on the screen during the show. For both social TV strategies, we compare user engagement when the strategy is being used on the show to when it isn’t. This paper focuses primarily on the level of engagement in response to two straightforward social TV strategies. In future work, we will also explore more deeply the content of tweets, or what is being said as well as many other questions linking what is said to social media response and engagement.

3. DATASET
We have decided to focus on the popular TV show, The Voice. The Voice, an American reality singing talent show, made headlines last season due to its use of social media content during the show. The singing contestants on The Voice are mentored by one of four coaches: Adam Levine, Cee Lo Green, Christina Aguilera, or Blake Shelton. The show, the contestants, and coaches are widely discussed throughout the season by viewers. The Voice frequently displays Tweets and Facebook messages on the show during the broadcast to keep viewers engaged. These displays often ask the viewing audience a question or request that audience members post something that happened on social media. The audience then engages using social media outlets such as Twitter and Facebook to respond. The audience also participates in the show by voting for contestants. The voting happens in part by purchasing songs from the online music store, iTunes, data from which can be tracked over time. In this study, we will combine two types of data: timestamped TV dialogue and geo-stamped social media buzz.

We primarily use Twitter as our test bed. Reportedly, Twitter has over 200 million registered users sharing information about a wide range of topics daily. Twitter is a social networking microblogging solution where users of the service follow other users who answer the question, “What are you doing?” in 140 characters or less. These short messages are called tweets. Using tweets, people can reference other users, as well as links to Web pages and Web stories. Of particular importance to our work is the fact that users can “retweet” messages that other users have posted; these are indicated by the tag “RT.” Both links between users and tweets about particular topics spreading on the networks can be observed, including retweets on the network, which can be studied to map information contagion directly.

Twitter makes a subset of its data available to researchers through a portal-supplied application programming interface (API). The Twitter Search API will be queried to collect data relating to users, such as their friends and followers. The Twitter Streaming API was used to collect real-time Twitter statuses that contain prespecified terms and tags related to the show as they are posted. These terms included the Twitter handles of the contestants and coaches and the hashtags associated with the coach teams. All data are publicly available. In addition to the timestamped tweets, we collected the exact
times specific tweets appeared on the TV as part of the viewing. The data are anonymized for research purposes by mapping all users in our database to our own set of anonymous IDs. Metadata referring to usernames or Twitter IDs and all "@" and reply-to mentions of other usernames within the body of the statuses will be replaced by their corresponding anonymous ID. Twitter status updates are particularly amenable to anonymizing because sensitive fields such as usernames and personal names are encoded in separate fields in the JSON object returned by the Twitter API, and other users’ tweets are prefaced by an "@" character within the text body of the status.

In total, we collected over 5.6 million tweets from February 5th to April 28th. Of those, 3.3 million were contributed during the show. We restrict the set of tweets to those made by accounts that are made public and are associated with the show. Of those tweets, 3.0% were displayed on the TV during the show and 97.0% were not. Each tweet is a candidate for being retweeted. Some are retweeted in great numbers while others are not. Likewise each tweet can be replied to. In Figure 1, we show the distribution of the number of retweets for tweets that were shown on the show.

We also study the aggregate level tweets over time and how that level changes as a result of stimulus. Figure 2 is an example of the tweets over time for April 3 2012. You see that over time, the tweets go down, but there are certain peaks, even during the show. In this paper, we ask do those peaks happen when something interesting happens during the show or there is a social media stimulus.

4. HYPOTHESES

In this work, we aim at addressing the question, when are social TV strategies correlated with increased social network buzz during the popular American reality singing show, The Voice? We attempt to answer this question by evaluating two strategies the show uses as part of their viewing. The first strategy is the placement of specific tweets on the screen. These tweets often come from contestants on the show or the judges or the MC. The second strategy is the placement of the show hashtag on the show that serves as a reminder to viewers to continue to tweet and get involved with the conversation about the voice.

4.1 Placement of Tweets on Screen

While placing related tweets on screen has been becoming more and more popular in TV programs, the effectiveness of this strategy in promoting users’ online activities has yet to be empirically verified. As a first attempt toward this end, we test four hypotheses for the tweets that were made throughout the show by participants in the show:

H1: A tweet displayed on the show will have a much higher engagement level with viewers measured by number of retweets.

H2: Given the tweet is on the show, a tweet that mentions a high profile user (one that has a high number of followers) on the show will have a greater likelihood of being retweeted.

H3: Tweets that include hashtags will be more likely to be retweeted.

H4: Tweets that include exclamation marks indicating expression will be more likely to be retweeted.

The first hypothesis focuses on the binary attribute, does the tweet show up on the show or not. The subsequent hypotheses focus on the content of the tweet. We use linear regression to test these hypotheses with our data. We include a number of control variables, the attributes of interest to test our hypothesis, and the dependent variable, number of retweets. Table 1 provides a list of these variables.
A particularly interesting way to interact with viewers is to ask them questions through tweets on the screen. While there were only 3 questions that were in Tweets displayed on the TV show, we found that questions generated 14 times more replies than retweets (average of 257.3 replies per tweet), compared to the overall ratio of 0.5 (average of 15.9 replies per tweet). This indicates that if you want to have a conversation with viewers, you should ask a question.

4.2 Placement of Hashtags on Screen

The advantage of displaying a specific tweet on the screen is that it provides a very concrete topic for viewers to talk about, enabling in-depth discussion along a given thread of retweets or replies. However, it might fail to draw viewers’ general attention to the TV program. A more direct way to achieve this goal is to display a hashtag on the screen that uniquely identifies the TV program. The hashtag can pull people interested in the same TV program together, facilitate communication between them, and allow the program marketing to more easily track buzz relating to their program. Given that viewers’ attention might be largely attracted by the show, we posit that

$H5$: The buzz during commercial breaks should be greater when they are preceded by a hashtag than when they are not.

In this paper, we compare the change in buzz during a commercial break when #thevoice is shown prior to it to the change in buzz when #thevoice does not precede it. The difference in buzz is calculated by counting the number of program-related tweets that occurred during the commercial, then observing the total activity that occurred in the time period prior to the commercial. We also note the change in activity over all hashtags displayed during the show. This is done by comparing the activity in the three minute window before the hashtag was displayed, and the window directly after the tag was shown.

5. RESULTS

We test four hypotheses using a linear regression model. In Table 2, we show that displaying the hashtag on the screen and the content of the tweet are predictive of retweets. This is after controlling for the number of followers (popularity) of the tweet poster, the general buzz at the time, and the time the tweet occurred during the episode. Similar work has been performed on predicting retweetability of tweets in general [10]. Our contribution arises from the use of the “was displayed on TV” attribute, linking exposure on television to a higher rate of retweeting.

We test one hypothesis, that engagement stays at higher levels during commercials when the hashtag for the show is displayed on the TV. In Table 3, we compare the number of tweets during commercials that were preceded by #thevoice displayed on TV compared to those that were not preceded by a similar hashtag. The number of tweets occurring in the time previous to the commercial were counted (the length of the window same size as the commercial length).

Table 2. Summary of weights learned after controlling for popularity of tweeter, overall activity, and position during show.

<table>
<thead>
<tr>
<th>Model 1 (control)</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2=$0.6162</td>
<td>$R^2=$0.6409</td>
</tr>
</tbody>
</table>

- Intercept -0.8 ***

- Controls
  - log(1 + number of followers) 0.4 ***
  - number of tweets within 2 hour window 4.8*10^-8
  - time from start of show (seconds) 5.4*10^-6

- Hypotheses
  - was displayed on TV 0.6 ***
  - Log(1 + num followers of judge mention) 0.7 ***
  - number of hashtags in tweet 0.7 ***
  - number of exclamation points in tweet 0.7 ***

Table 3. Model comparison

<table>
<thead>
<tr>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2=$0.6425</td>
<td>$R^2=$0.6453</td>
<td>$R^2=$0.6444</td>
</tr>
</tbody>
</table>

- -0.8 ***
- 0.4 ***
- 0.2*10^-8
- 0.7*10^-6
- 0.7 ***
- -6.1*10^-3 **
- 4.9*10^-2 ***

- 0.6 ***
- 4.1*10^-8
- 7.1*10^-6
- 0.6 ***
- -5.8*10^-3 **
- 4.9*10^-2 ***
- 8.3*10^-3 *
In addition, we explore whether displaying a hashtag at any point during the program tends to increase social media activity. We collected all #thevoice posts from the initial episode on February 5 to April 17. In this time period, the hashtag was displayed 73 times. Summing over all occurrences of #thevoice, the previous three minute windows contained 170,283 tweets, and the following three minute windows contained 202,362 tweets; an increase of 18.8% in buzz. Figure 3 depicts the distribution of proportion activity changes over all hashtag displays. From this figure, it is clear that this overall positive change is not likely to be due to outliers, but is a consistent tendency of hashtag displays.

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Total tweets in previous window</th>
<th>Total tweets during commercial</th>
<th>Proportion change in tweet volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>preceded</td>
<td>6</td>
<td>10454</td>
<td>12788</td>
<td>0.22</td>
</tr>
<tr>
<td>not preceded</td>
<td>9</td>
<td>17860</td>
<td>16868</td>
<td>-0.056</td>
</tr>
</tbody>
</table>

Despite these limitations, we are very excited about this work because we believe it is the first evidence that social TV strategies do actually work at creating significant engagement. In the future, we will control for what was happening on the show at the time of presenting the Tweets and/or hashtag. Examples include the distance from a performance, from voting on the show, from commercial breaks, from a positive comment by the judges, etc. We have already begun this analysis and our preliminary results suggest that the increase in buzz is still significant after controlling for a wide range of features about the show.

**6. LIMITATIONS**

While The Voice might be one of the most salient TV programs using the social TV strategy, we still feel the need to test our result on other programs. A possible concern is that some TV programs could be more engaging than others. We hope to control for this potential factor in our future studies. Users might have different levels of activities on Twitter at different time of a day/week. So far, we haven't considered the impact of the overall level of activities on Twitter to our result. While we do believe that a large proportion of the bumps on Twitter result from the social TV strategies, controlling the effect of the overall temporal dynamics on Twitter can make our estimate more accurate. We still only use very shallow metrics of tweet content, by simply counting the occurrence of popular users, hashtags, and exclamation marks. This can be a good start for understanding the effect on content on its popularity, but we still need to analyze the content of the tweets more comprehensively to capture the underlying reasons that result in a high volume of buzz. Some other useful features describing tweet content are the sentiment of tweets and its emotional arousal, generated by trained classifiers. Given that the dataset we analyze is observational, there are many events that are difficult to control for. For example, it could be the case that #thevoice was only displayed after a very popular contestant performed. The fact that they performed could thus drive the increase in tweets seen after posting it. We did not observe such an effect, but it could nevertheless be the case.

We have begun preliminary analysis using a sentiment model. This model is a maximum entropy model that was trained on a subset of 10% of the Twitter stream during 2009. All messages in this set that contained a form of the word “love” were considered positive instances, and all messages that contained a form of the word “hate” were considered to be negative instances. This approach to generating a sentiment classifier training set was inspired by Pak and Paroubek, 2010 [11]. The resulting training set contained over 2.7 million positive instances and 770 thousand negative instances. This training set is undoubtedly noisy since users may also use the terms “love” and “hate” in negative and positive tweets respectively. However, we believe that these cases are negligible and this method allows us to train a Twitter-specific sentiment model with no hand-coding of tweets.

An interesting effect that this model illuminates is that the average sentiment per episode increases over time. Figure 4 depicts this. There seems to be a linear relationship between the episode number and the average sentiment during it, with a relatively high R-squared value of 0.74. This differs from the total buzz per episode, which seems to show that viewers are very engaged in the program both at the beginning and end of its run.

**7. DISCUSSION**

From these preliminary results, it seems as though displaying these hashtags correlates with a tangible increase in viewers’ Twitter activity overall, as well as increases their engagement with the program during commercial breaks. However, although we did not notice systematic placement of hashtags during the show, there could be a correlation between noteworthy content during the show and the hashtag display. In this case, much of the additional buzz that we observe could be attributed to the content of the episode. There are many other questions we would like to address to this dataset. For example, how does the sentiment on Twitter relate to the Nielsen ratings of a television program? How does this Twitter activity vary over different regions in the U.S.? How does the network of users grow over the course of the season and what prompts specific types of growth?

We have begun preliminary analysis using a sentiment model. This model is a maximum entropy model that was trained on a subset of 10% of the Twitter stream during 2009. All messages in this set that contained a form of the word “love” were considered positive instances, and all messages that contained a form of the word “hate” were considered to be negative instances. This approach to generating a sentiment classifier training set was inspired by Pak and Paroubek, 2010 [11]. The resulting training set contained over 2.7 million positive instances and 770 thousand negative instances. This training set is undoubtedly noisy since users may also use the terms “love” and “hate” in negative and positive tweets respectively. However, we believe that these cases are negligible and this method allows us to train a Twitter-specific sentiment model with no hand-coding of tweets.

An interesting effect that this model illuminates is that the average sentiment per episode increases over time. Figure 4 depicts this. There seems to be a linear relationship between the episode number and the average sentiment during it, with a relatively high R-squared value of 0.74. This differs from the total buzz per episode, which seems to show that viewers are very engaged in the program both at the beginning and end of its run.
but not during the intervening episodes (see Figure 5). The colors in the plots differentiate the different types of episodes that aired. However, if we modify our sentiment metric and instead consider the proportion of tweets with sentiment below a threshold of 0.3, it seems to predict the buzz regarding the program well (see Figure 6).

Figure 4. Average sentiment of tweets during an episode as a function of the episode number. There seems to be a strong linear relationship, \( R^2 = 7.6 \)

Figure 5. Number of “The Voice” related tweets occurring per episode. The colors refer to different types of episodes.

Figure 6. Plot of the total Voice-related buzz per show given the proportion of tweets below the sentiment threshold of 0.3. This metric of sentiment seems to predict the total buzz much more closely than just using average sentiment alone, which increases linearly over time.

In addition to examining tweet sentiment and general buzz during episodes, there also seems to be a relationship between Voice-related buzz on Twitter and the average rank of The Voice songs that are listed in the top 100 in sales on ITunes (see Figure 7). Note that episodes during which no The Voice songs were listed in the top 100 are omitted from this plot. This relationship is partially due to increased buzz during the final two episodes, but further establishing the relationship between social media and sales will be considered in further work.

Figure 7. Plot of average ITunes sales rank during an episode against the total number of Voice-related tweets during the episode. In this case, a lower rank corresponds to higher sales.
8. CONCLUSIONS

In this research, we show two main effects. The first is that displaying a tweet during a program will increase its retweet rate. Even controlling for the popularity of the tweeter, the content of the tweet will also affect its expected number of retweets. Second, we show that displaying hashtags during a program seems to increase the number of program-related tweets, in this case by a relatively high proportion, 18.8%. This may be surprising since this hashtag is displayed very frequently, and one might expect that the viewers would become immune. We also show that there may be interesting relationships between the sentiment of messages tweeted during an episode and the total buzz around that episode.

9. NEXT STEPS

We believe that the data set we have collected is very rich. There has been little research, to our knowledge, on the direct impact of television social media content on the online social network activity and sales. We have collected over 5.6 million messages pertaining to this show, the networks of all show-related handles with a 15 minute sample rate, coded recordings of all episodes to date, and sales data pertaining to the show collected every 10 minutes. In addition, we have collected a set of Tweets related to major TV finales, American Idol, Dancing with the Stars and the Grammys and Superbowl. We have already begun to link features of advertisements during Super Bowl 2012 to social media buzz [12]. We intend to use all of these data to further understand the relationship between program’s use of social media content and viewership, engagement, and sales. We plan on releasing this dataset to the general public once it has been properly compiled and anonymized.

Some of the independent variables which we used to quantify tweet content are relatively shallow (e.g., number of exclamation points, hashtags). In order to get a more assessment of the tweet content, we intend to use LDA topic modeling to capture general topic of tweets. The text associated on web pages referenced in the tweets can also be used to determine the topic of a tweet. We also intend to apply our sentiment model to particular contestants over time to gain a better understanding of public sentiment regarding contestants for different episodes.

Another interesting question which we would like to address is how the buzz regarding contestants varies by location. In our collected corpus, approximately 1% of tweets are geocoded with latitude and longitude. However, tweets also contain a free text “location” field which we can also scrape in an effort to infer the tweet origin without knowledge of the exact latitude and longitude.

10. Acknowledgements

The data collection for this project was labor intensive and the final dataset including timed TV events, tweets and sales data was the product of the work of many research assistants in our lab (see www.thesocialtvlab.com). We would like to thank Anthony Crawford, John-Paul Lacovara and Haftom Khasai, in particular, for their help with data collection and Jing Peng for both data collection and discussions on an earlier draft of this work.

11. REFERENCES