The Spoiler Effect?:
Designing Social TV Content That Promotes Ongoing WOM

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

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

ABSTRACT

Television shows are now instigating online social interactions between viewers by requesting viewers, as part of the first broadcasts, to participate in simultaneous discussions about shows. In this work, we analyze and quantify the effect of strategies used in an American reality singing show to engage viewers. We show three main results: 1) when specific tweet messages are posted on the TV screen during the show, the tweets 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; 2) there is a difference between social media engagement response to specific tweet messages and more general social media content-- specific tweets with new information result in more engagement on the first show airing; and 3) there is a difference in overall response to specific social media tweet messages on the West Coast of the US since results by the time of airing have been spoiled.

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 look 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]. It has been reported that people watch more live TV to both avoid spoilers and to communicate with other viewers. This is in contrast to the trend only a few years ago when people started using DVR machines to watch shows at their own pace.

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. For example, they might ask how much engagement specific types of programming can elicit from viewers. If an announcer tells a consumer to go to a certain Web site, to send a tweet, or to purchase a particular item, does that influence the consumer to actually comply? The response on sites like Twitter 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 ripe 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 prevalent 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 correlate with viewership or sales, let alone drive 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 effect 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 this work, we analyze several aspects of such tweets and our work to analyze Facebook posts is underway.

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 provide 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 correlation with content on the online community's response and sentiment. In addition, we show that the engagement response varies across different airings of the show (East Coast versus West Coast). To our knowledge, this is the first academic study 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 communal 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 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 at the individual 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.

In this work, we track buzz and sentiment 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. The second is the placement of the #thevoice hashtag, 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. In addition, we compare how the response varies across different airings of the show in particular focusing on the fact that the results of the show may have been spoiled for west coast viewers.

Spoiler effect refers to the phenomenon that the disclosure of information that reveals any plot or specific of experiential goods (e.g., movies, fictions, TV shows) can significantly reduce people's interest in the goods. While spoiler effect is commonplace in daily life, studies on this problem are rare in the literature. In an attempt to explain the spoiler effect, Wilson et al. [9] found that the uncertainty following a positive event can prolong the pleasure it causes, whereas the cognitive efforts to make sense of the positive events reduce the pleasure people can obtain from them. Further, Tsang et al. [10] argued that the spoiler effect was caused by a cognitive bias called focusing illusion [11]. In addition, they presented a possible way to neutralize the spoiler effect – highlighting non-disclosable attributes of the experiential goods [10]. On the technical side, Eliashberg et al. [12] presented how to screen movies which are worth investment based on the spoilers of movies. Guo et al. [13] proposed to detect comments and reviews containing spoilers automatically by incorporating linguistic dependency information into topical models. Our paper focuses, in large part, on the level of engagement in response to two straightforward social TV strategies and how the response changes as a result of spoilers.

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 celebrity 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 on the show during the broadcast to keep viewers engaged. These displays are usually just presentations of specific Tweets or hashtags but they also ask the viewing audience a question or request that audience members post something that happened on social media. The audience then engages with the show and other viewers using social media outlets such as Twitter and Facebook to respond. The audience also participates in the show by voting for contestants through text messages, iTunes song purchases, and social media messages. In this study, we will combine two types of data: time/geo-stamped social media buzz on Twitter and time-stamped TV dialogue and events.

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 spread and 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 pre-specified terms and tags related to the show as they are posted. All data are publicly available.

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 are 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 airing of a 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 specific 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 gives an example of the tweets over the course of a day for April 3, 2012. Over time, the number of tweets go down, but there are certain peaks, during the show. In this paper, we address whether those peaks happen when something interesting happens during the show and/or whether it is a result of social media stimulus.

In addition to the timestamped tweets, we collected the exact times specific tweets and hashtags appeared on the TV as part of the viewing. Recordings of the shows were viewed and research assistants noted the exact times messages were displayed. We use these data to link the real time display of specific social media content to social media buzz. A list of features constructed from the TV content can be found in Table 1. It is important to note that, in this work, we are able to identify the source of exact social media buzz because new news is happening for the first time on the TV show.

4. HYPOTHESES

In this work, we aim to address 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, judges, or the Master of Ceremonies. The second strategy is the placement of the show hashtag, #thevoice, on the show that serves as a reminder to viewers to continue to tweet and get involved with the conversation about the voice. We first evaluate the strategies individually, then we compare them, and finally we compare their effectiveness across time zones in the United States. For each of our hypotheses, we use ordinary least squares models to link factors of interest to viewers engagement outcomes.

4.1 Placement of Tweets and Hashtags on Screen

While placing related and specific tweets and general hashtags on screen has become 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 five hypotheses for the tweets that were made throughout the show by participants in the show and hashtags that were displayed on the screen:

H1: Displaying atweet on the show will result in 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 2 provides a list of these variables.

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.

We compare the change in buzz during a commercial break when #thevoice is shown prior to it to commercial break buzz when #thevoice is not shown, with respect to the level of buzz before the commercial break occurred. 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.

4.2 A Comparison of the two Strategies: Specific Tweets to general Hashtags

In previous work [14] we have focused on increase in buzz or participation, and do not make a distinction between posting a tweet on screen or a hashtag/handle. In this work we examine the effect of these two different social media strategies more closely, and assess their efficacy using additional metrics (e.g., sentiment on social media and diversity of response). In addition, we analyze how these two different strategies affect social media response during the east and west coast showings. Our sixth hypothesis is that response to specific tweets will be greater than response to general tweets on a number of social media engagement dimensions.

H6: Response to specific tweets will be greater than response to general tweets

In order to test hypothesis 6, we coded each episode for hashtags/handles displayed (general) mentions/tweets displayed (specific) calling these "social media mentions" (SMMs). Displays of

SMMs on west coast were inferred by pushing timestamps of these events forward by 3 hours (except for the first show, which was shown simultaneously on both east and west coasts). For this result, we assumed that the program content was the same between east and west coast showings (i.e., that commercials occurred at the same points in the episode, that the commercials had the same content, and that the program content in the west coast airing was the same as the content in the east coast airing). The fact that the times of west coast events were inferred may result in the west coast timestamps becoming slightly off. However, we believe that the window size we used allows for a small amount of perturbation in the timestamps.

For each event, we were able to observe the proportion/difference of change in total Voice-related buzz 3 minutes before each SMM to 3 minutes after each SMM. Along with this metric, we included other dependent variables such as the difference of proportion tweets from in-network users (capturing change in the proportion of buzz coming from Voice fans), difference of proportion original tweets (tweets that were not replies or retweets), and difference of average tweet sentiment after the SMM to before the SMM.

We built linear models to predict these dependent variables, controlling for overall Voice-related buzz on Twitter (within a two hour window around the SMM timestamp), viewership of episode, and program content (e.g., time since last performance, popularity of last performance, time during course of show, time since last commercial). We also included the variable IS_SECOND_HALF, and its interaction with whether or not the SMM was specific or general, since after episode 12 (April 9), all specific tweets displayed during the show were constrained to the "social media room", so should be treated differently. After removing seven outliers from our data, we were left with 967 SMMs, 483 during the east coast airing and 484 during the west coast.

4.3 The Spoiler Effect: A Comparison of East Coast West Coast Response

We hypothesize that West Coasters will not be as sensitive to social media stimulus, especially for content specific to the time and first airing of the show. We use the same methodology as above and test Hypothesis 7 below.

H7:Response to specific tweets will not be as strong on the West Coast after the show has aired the first time and has had the opportunity to be spoiled.

5. RESULTS

5.1 Placement of Tweets and Hashtags on the Screen

In Table 2, we show that displaying the hashtag on the screen and the content of the tweet are correlated with the number 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 in general [15]. Our contribution arises from 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).

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, 2012. 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. However, this does not control for the fact that the #thevoice hashtag was displayed at points which were deemed tweet-worthy, and increased user engagement could be a result of the program content, not the display of the hashtag. The model in section 5.2 attempts to show a relationship between SMMs and user engagement while controlling for program content.

5.2 A Comparison of the two Strategies: Specific Tweets to General Hashtags

In Table 4 we show that being a specific tweet has an effect on the proportion of total buzz, even with controlling for other variables, although it has no effect on the other dependent variables. We also find that the response is correlated with episode number. Even though number of viewers/tweeters decreases/stays the same over subsequent episodes, their reaction to SMMs decreases over time. Figure 4 depicts the reduction in average response to SMMs as a function of episode number. Note that the outliers that were removed were high response posts early on, from episodes two through five. We believe that the downward slope is partially due to higher variance in the response during earlier episodes.

Although having a specific or general SMM seems to have no effect on the proportion of original user-generated content, we did notice a strong effect of the show itself on the proportion of original content generated. Figure 5 shows that there seems to be a moderately strong correlation between the seconds until a commercial is displayed with change in original content for the SMMs in our data. As the time till the next commercial break increases, there tends to be a higher change in the proportion of original content generated the time the SMM is displayed. This effect can best be explained by Figure 6. Here the distribution of original content proportion is displayed for different time segments related to the show airing. This suggests that users are generating original content as the events are occurring and use the "downtime" during commercials and after the show is over to converse and propagate the information. This may be expected, but explains the relationship between "seconds till commercial" and change in the proportion of original content.

5.3 The Spoiler Effect: A Comparison of East Coast West Coast Response

Whether an SMM is general or not is non-significant in the west coast model, while it is significant in the east coast model. We hypothesize that west coast users are de-sensitized/spoiled by knowing that these mentions are first displayed on the east coast. There is no point in reacting to these messages, since they are old news. Total buzz is also much less than on the east coast (for example, during the finale we captured approximately 20,000 Voice-related tweets for the west coast airing, whereas the east coast airing generated approximately 136,000), and they respond less to specific tweets. In general, that is not surprising since Nielsen reports over 55 million east coast viewers and 18 million west coast during 2011. Figure 7 displays the distribution of response to SMMs as a function of whether it is a specific tweet or a general hashtag/Twitter handle, is during the second half of the program season, and whether it was shown during the east or west coast airing.

In this results section we show three things: 1) Displaying a hashtag or tweet leads to more engagement 2) in general, specific tweets lead to more engagement and 3) west coast viewers respond at a lower rate to specific social media content during shows, perhaps due to a spoiler effect.

In addition, we observed that 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 during the broadcast.

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 level 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 for 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 capability, 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. We attempt to control for this in the model in section 5.2 via the “seconds since last performance” and “sum of followers of last performers” features.

Other limitations include that we cannot be certain that people are from east/west coast just because they happen to tweet about The Voice when it is airing in that particular time zone. This could disturb the data if people are tweeting after or before the show airs. By conference we will do analysis with actual location data, either geo-tagged or inferred from a free-text location description field, when possible. We performed a sanity check during the first airing of the show, which was broadcast on the east and west coasts simultaneously and found the same response rates on both coasts and therefore don’t think our location identification strategy is a problem. Another issue could be that west coasters are just different somehow and don't respond to social media stimulus. The same sanity check discussed above gives us evidence that that is not the case.

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.

7. DISCUSSION

From these preliminary results, it seems as though displaying these specific tweets and general hashtags causes a tangible increase in viewers' Twitter activity overall, as well as increases their engagement with the program during commercial breaks. We attempted to control for the effect of program content with the models in Table 4, yet, there may be other factors which we did not take into account. For example, perhaps although the last performer was not very popular, their performance was stellar, generating a high level of buzz. There are many other questions we would like to address with this dataset. For example, how do Nielsen ratings relate to characteristics of the The Voice tweets? Can characteristics of the last episode's tweets (e.g., sentiment) predict the ratings of the subsequent show? How does this Twitter activity vary over different regions in the U.S.?

We have begun preliminary analysis using a sentiment model, from which the sentiment change dependent variable was generated. 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 [16]. This training set is undoubtedly. However, we believe that the trained model is still useful and requires no hand-coding of tweets. We also use this model to generate the "change in average sentiment" dependent variable in Table 5.

Using this model, we discovered that the average sentiment per episode increases over time, as depicted in Figure 8. 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 9).

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 10).

Table 4 suggests that there is no difference in the type of social media strategy on the proportion of original content, proportion of tweets from in-network users, or sentiment. However, this does not mean that there is no effect at all, and we intend to compare the effect of SMMs on these outcomes to randomly sampled times that do not coincide with SMMs. In spite of no effect of the type of social media strategy, we discovered an interesting fact about viewers' activity on Twitter with respect to the show duration.

This work is very important as it allows for the real time identification of effective social media strategies and user engagement.

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 the type of SMM displayed on the program has an effect on social media user engagement. Namely, specific tweets relating to what is currently happening on screen tend to generate more buzz than general hashtags or handles. We also show that response to these SMMs during the west coast airing is reduced, perhaps due to a "spoiler effect".

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 over time with a 15 minute sample rate, coded recordings of all episodes to date, and sales data pertaining to the show collected every 10 minutes. 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 origin of the tweet.

Finally we are interested in how the network of users changes over time as a function of what is being said on TV. This can be thought of as yet another social media metric and would . We chose not to include this dependent variable in our analyses since the network sampling rate was too slow for the time frames we were examining. We could however evaluate the effect of particular types of episodes or appearances of contestants on the change in their number of Twitter followers.

10. ACKNOWLEDGEMENTS

Data collection was a tremendous undertaking for this project. We would like to thank Anthony Crawford, Stan Liu, John Paul Lacovora, and Jing Peng for their help.

11. REFERENCES

Dumenco, S. 2011. "7 Things You Need to Know About 'Social Tv' Right Now the State of the Art -- and Science -- of Monitoring Social-Media Chatter About Tv." from http://adage.com/article/the-media-guy/social-tv/229845/

Talbot, D. 2011. "A Social Media Decoder." from http://static.bluefinlabs.com/website/bluefin_mit-tech-review.pdf

Proulx, M., and Shepatin, S. 2012. *Social Tv: How Marketers Can Reach and Engage Audiences by Connecting Television to the Web, Social Media, and Mobile*. Wiley.

Nathan, M., Harrison, C., Yarosh, S., Terveen, L., Stead, L., and Amento, B. 2008. "Collaboratv: Making Television Viewing Social Again," in: *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*. Silicon Valley, California, USA: ACM.

Wohn, D., and Na, E.-K. 2011. "Tweeting About Tv: Sharing Television Viewing Experiences Via Social Media Message Streams," *First Monday [Online]* (16:3).

Harboe, G. 2009. "In Search of Social Television," in *Social Interactive Television: Immersive Shared Experiences and Perspectives*, D.G. P. Cesar, and K. Chorianopoulos (ed.). IGI Global, pp. 1-13.

Torrez-Riley, J. 2011. "The Social Tv Phenomenon: New Technologies Look to Enhance Television's Role as an Enabler of Social Interaction ". Elon, North Carolina: Elon University, p. 23.

Helweh, A. 2011. "Twitter, Time Shifting, Technology & Television. Social Media Explorer." from http://www.socialmediaexplorer.com/social-mediamarketing/twitter-time-shifting-technology-television/.

Wilson, T. D., Centerbar, D. B., Kermer, D. A., and Gilbert, D. T. "The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate," *Journal of personality and social psychology* (88:1) 2005, p 5.

Tsang, A. S. L., and Yan, D. "Reducing the Spoiler Effect in Experiential Consumption," in: *Association for Consumer Research Annual Conference*, San Francisco, CA, 2008.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., and Stone, A. A. "Would you be happier if you were richer? A focusing illusion," *Science* (312:5782) 2006, pp 1908-1910.

Eliashberg, J., Hui, S. K., and Zhang, Z. J. "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science* (53:6), June 1, 2007 2007, pp 881-893.

Guo, S., and Ramakrishnan, N. "Finding the storyteller: automatic spoiler tagging using linguistic cues," Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Beijing, China, 2010, pp. 412-420.

Hill, S., Benton, A. 2012. "Social TV: Analyzing the Impact of Social Media Strategies on Viewer Engagement." In Proc. of Workshop on Social Media Analytics, KDD.

Suh, B., Hong, L., Pirolli, P., Chi, E. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In Proceedings of the 2010 IEEE Second International Conference on Social Computing. pp. 177-184.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of LREC.

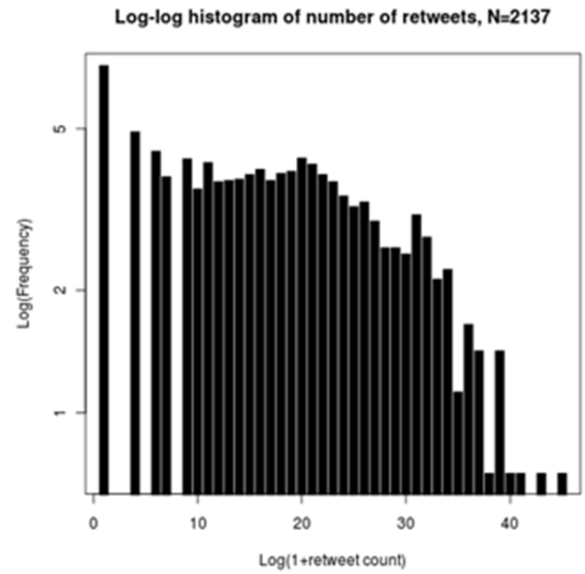

Figure 1. Distribution of log(retweets) for all tweets in sample.

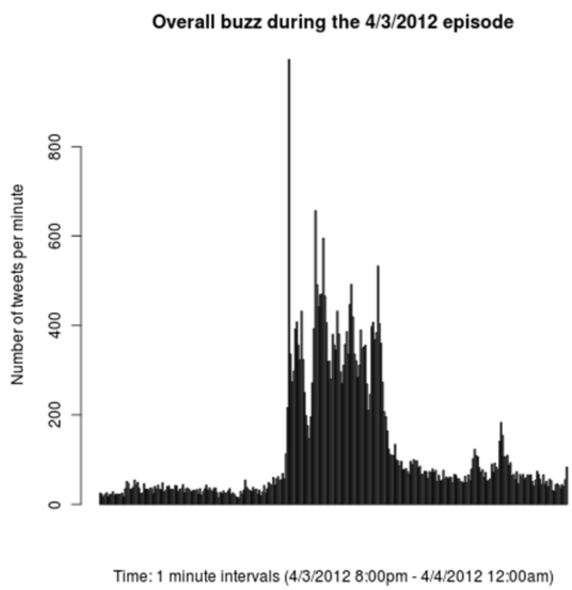

Figure 2. Tweets over time for April 3, 2012.

Table 1. Variables associated with the general or specific social media mention, in regard to the TV show.

Factor type	Factor	Description
Episode feature	IS_SECONDHALF	Is in the second half of the season, after April 9 showing. All "specific" tweets occurred during the social media room after this date.
	EPISODE_NUM	The episode number, from 0 to 20.
Episode context	SECS_SINCE_PERFORMANCE	Seconds since the last performance.
	LAST_PERFORMERS_POPULARITY	The sum of number of followers for each of the previous performers.
	SECS_SINCE_COMMERCIAL	Seconds since the last commercial break.
	SECS_SINCE_START	Seconds since the start of the show.

Table 2. Variable description for predicting retweet count (hypotheses 1-5).

	Description
Control variables	
log(1 + num followers)	Number of followers of message poster
log(general buzz)	Total number of tweets occurring within hour
time from start	Number of seconds since show start
Independent variables	
was displayed	Whether or not the tweet was displayed on TV
log(1 + num mention followers)	Number of followers of judge/contestant mentioned in tweet
num hashtags	Number of hashtags in tweet
num exclamation points	Number of exclamation points in the tweet, attempts to capture arousal
Dependent variable: log(1 + num retweets)	Number of retweets elicited by tweet

Table 3. Summary of weights learned for predicting number of retweets. after controlling for popularity of tweeter, overall activity, and position during show.

	Model 1 (control) R^2=0.6162	Model 2 R^2=0.6409	Model 3 R^2=0.6425	Model 4 R^2=0.6453	Model 5 R^2=0.6444
Intercept	-0.8 ***	-0.8 ***	-0.8 ***	-0.8 ***	-0.9 ***
Controls					
log(1+number of followers)	0.4 ***	0.4 ***	0.4 ***	0.4 ***	0.4 ***
number of tweets within 2 hour window	4.8*10^-8	2.2*10^-9	0.2*10^-8	4.1*10^-8	3.9*10^-08
time from start of show (seconds)	5.4*10^-6	6.8*10^-7	0.7*10^-6	7.1*10^-6	7.1*10^-06
Hypotheses					
was displayed on TV		0.6 ***	0.7 ***	0.6 ***	0.7 ***
Log(1 + num followers of judge mention)			-6.1*10^-3 **	-5.8*10^-3 **	-6.0*10^-3 ***
number of hashtags in tweet				4.9*10^-2 ***	5.0*10^-2 ***
number of exclamation points in tweet					8.3*10^-3 *

Table 4. Change in tweet volume for commercials that were preceded by a #thevoice display.

Condition	N	Total tweets in previous window	Total tweets during commercial	Proportion change in tweet volume
Preceded	6	10454	12788	0.22
not preceded	9	17860	16868	-0.056

Table 5. Weights and significance levels of social TV factors affecting different social media engagement outcomes on east and west coast. . < 0.1, * < 0.05, ** < 0.01, * < 0.001**

		Buzz		Diversity/Dispersion		Sentiment
Factor type	**Factor**	**Total Tweet Proportion Change**	**Total Tweet Difference**	**Proportion in Network Difference**	**Proportion Original Content Difference**	**Average Sentiment Difference**
	Intercept	1.683***	928.1***	-5.312*10^-2***	0.1252***	-1.492*10^-2***
Episode feature	Is season second half	1.418*10^-2	12.91	4.362*10^-3	9.978*10^-3	-1.431*10^-2***
	Episode number	-1.971*10^-2***	-23.46***	1.141*10^-3*	-3.275*10^-3***	1.249*10^-3***
Episode context	Seconds since performance	3.15*10^4***	0.2355**	-2.453*10^-5***	5.313*10^-5***	-7.879*10^-6**
	Sum of last performers network sizes	-8.648*10^-9**	-2.964*10^-5***	6*10^-10	-3.206*10^-9***	4.899*10^-10**
	Seconds since performance X Sum of last performers network sizes	-8.765*10^-11***	-7.132*10^-8*	-5125*10^-12	-8.030*10^-12	7.756*10^-13
	Seconds since commercial	-2.887*10^-4***	-0.2706**	8.109*10^-5***	-2.294*10^-4***	3.253*10^-5***
	Seconds since start	2.088*10^-5***	1.906*10^-2*	9.385*10^-7	3.636*10^-7	-7.585*10^-8
Online context	Total number of tweets around (60 min/10 min)	-2.81*10^-6***	-3.391*10^-3***	1.186*10^-7.	-2.747*10^-7*	1.881*10^-8
Hypothesis	Is general social media mention	-7.515*10^-2* (p=0.049)	-124.8 (p=0.031)	-2.396*10^-3 (p=0.675)	-2.107*10^-2* (p=0.023)	-1.431*10^-2 (p=0.23)
	Is west coast showing	-3.432*10^-1*** (p<2*10^-16	-579.5 (p<2*10^-16)	-5.319*10^-3	-1.642*10^-2* (p=0.04113)	-4.962*10^-4
	Is general social media mention X Is season second half	-3.5*10^-2	-29.41	2.359*10^-3	2.939*10^-3	6.981*10^-3**
	Is general social media mention X Is west coast showing	8.849*10^-2 (p=0.015)	123.8* (p=0.024)	6.54*10^-3	1.366*10^-2	-2.563*10^-3
	Adjusted R^2	0.339	0.3334	0.1906	0.3869	0.2528

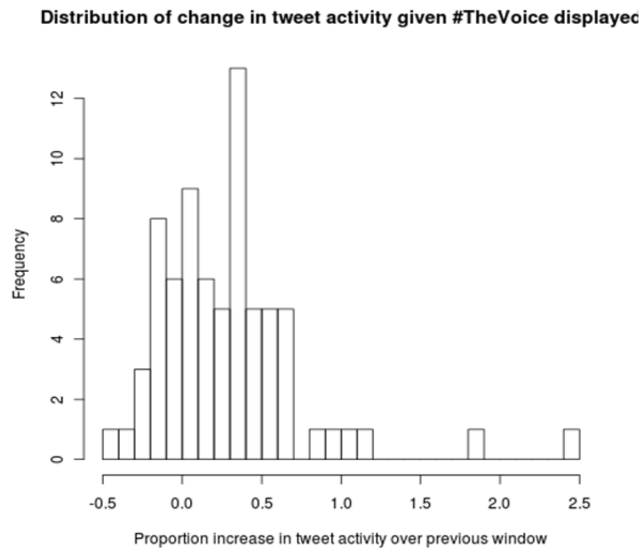

Figure 3. Distribution of proportion change in tweet activity given #thevoice displayed.

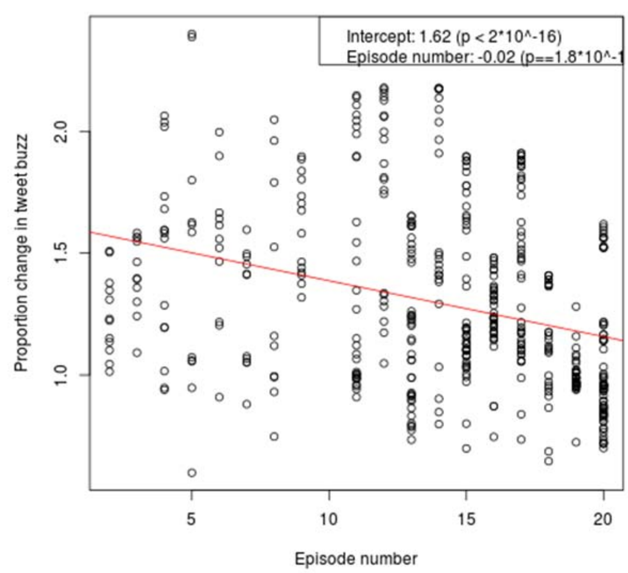

Figure 4. Buzz response to SMMs over time, by episode number. Only east coast SMMs shown.

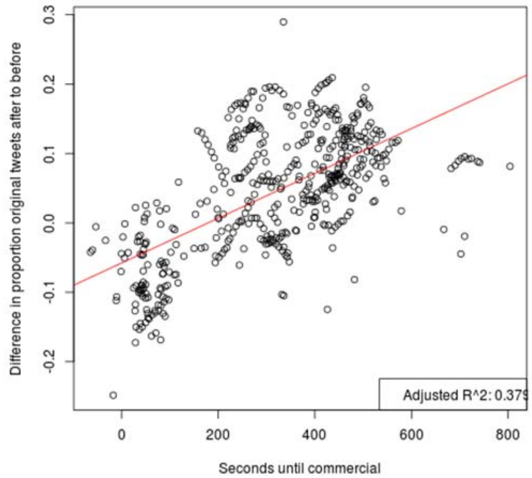

Figure 5. Difference in proportion of original (non-retweet/non-reply) tweets after a SMM compared to before as a function of seconds until a commercial. This linear relationship is due to difference in social media content generated during commercial breaks and during show.

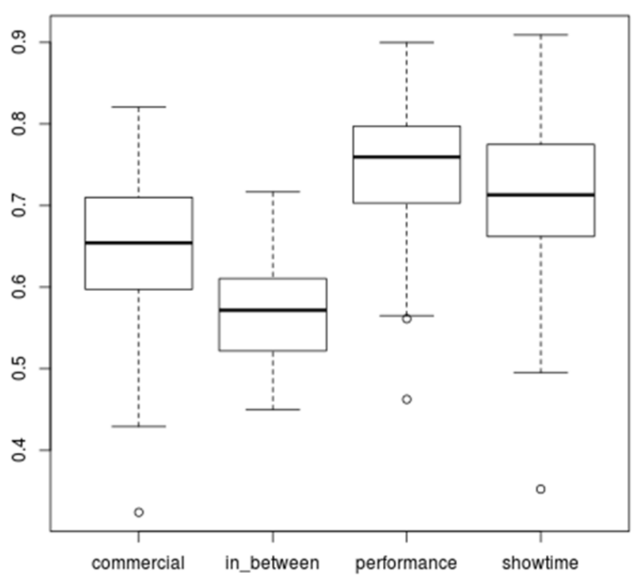

Figure 6. Distribution of proportion original content generated about The Voice during different time periods. "commercial" is during a commercial break, "in_between" is during the time between subsequent episodes, "performance" is during an artist performance, and "showtime" are during times when the episode is playing but there are no performers on. More original content tends to be generated when the episode is on than when it is off.

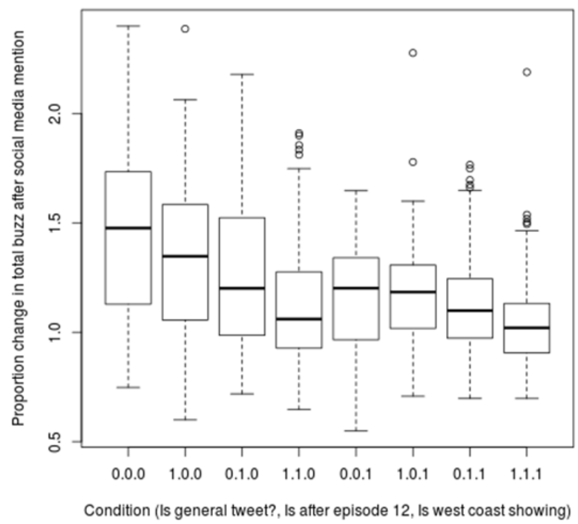

Figure 7. Distribution of total change in buzz as a function of the SMM being general, was during the second half of the season, or was during the west coast showing. Reaction seems to decrease for the west coast showing compared to the east coast. Specific tweets also seem to generate more reaction than general hashtags/handles. User engagement seems to be higher earlier in the season as well.

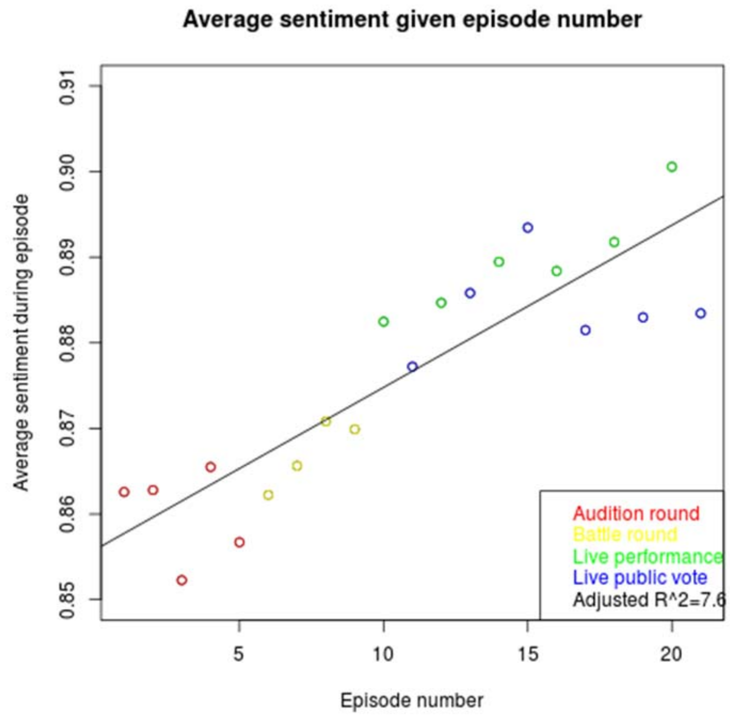

Figure 8. Average sentiment of tweets during an episode as a function of the episode number. There seems to be a strong linear relationship (adjusted R^2: 0.76).

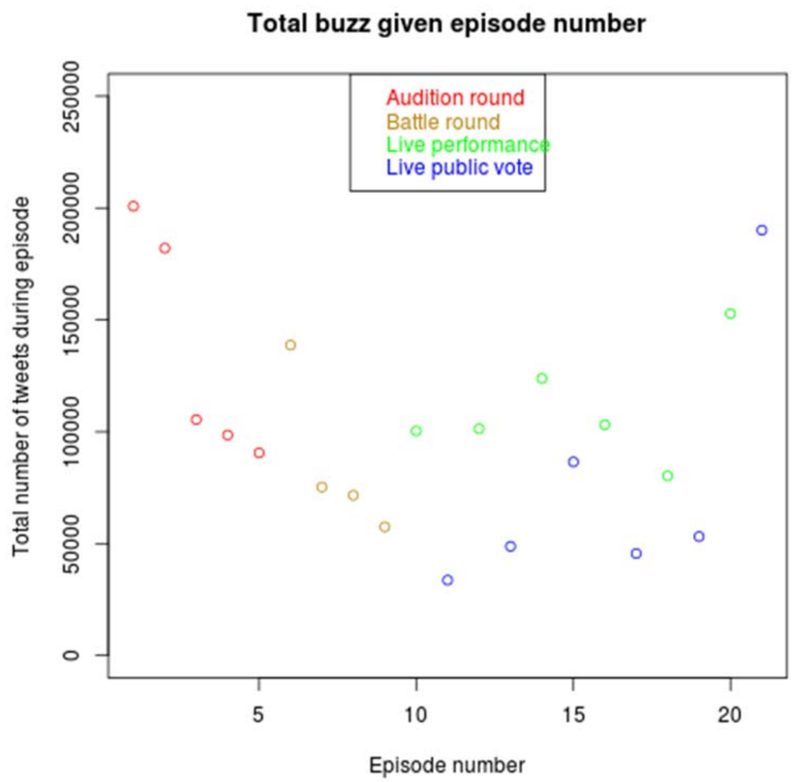

Figure 9. Number of The Voice related tweets occurring per episode. The colors refer to different types of episodes.

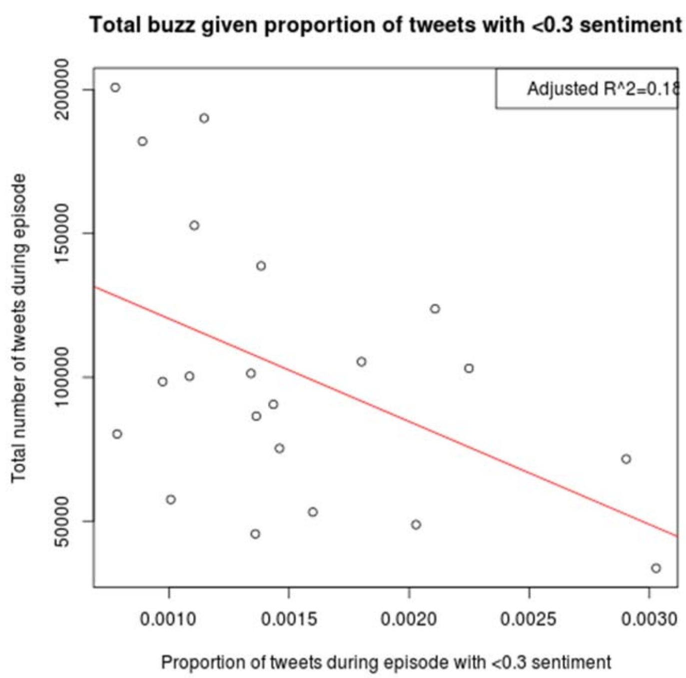

Figure 10. Plot of total Voice-related buzz per show given the proportion of tweets below the sentiment threshold of 0.3. This metric of sentiment seems to predict the total buzz much more closely than just using average sentiment alone, which increases linearly over time.