Formality Identification in Social Media Dialogue

Partha Mukherjee[1(✉)] and Bernard J. Jansen[2]

[1] College of Information Science and Technology, Pennsylvania State University, University Park, USA
pom5l09@ist.psu.edu
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
jjansen@acm.org

Abstract. Researching second screen interactions that form a social soundtrack concerning a major broadcast media event, we perform statistical analysis on more than 800 K postings and 50 K blogs of Super Bowl XLIX on Instagram and Tumblr respectively for three categories (commercials, music and game) during three phrases (*Pre*, *During*, and *Post*) identifying the influence of different social soundtrack features of the postings on formality of contents during three phrases (*Pre*, *During*, and *Post*). For Instagram, the positive influence of URL-based postings in relative scale on formality is significant, but other features have significant negative impact in *Pre* and *Post* phases. For Tumblr, undirected broadcast pattern of conversation and number of sentences in relative scale in *Pre* and *Post* phases have a positive influence on formality. The *During* phase does not show any significant influence between any of the social soundtrack feature and formality of the postings for either Instagram or Tumblr. It is important to note that formality is significantly increased on Instagram, but it exhibits significant reduction on Tumblr. We further evaluate the effects of categories on top of the influence of social interaction features on contents of social media platforms for a fixed effects model. For Instagram's formality aspect, the fixed effects estimate of the game category significantly outperforms the other two categories in all three phases, while for Tumblr, the music category fixed effects plays the lead role in *Pre* and *Post* phases. These results assist in identifying the strength of linkage among broadcast categories, social media postings, and inherent formality, providing insights into viewer reactions to the broadcast of In-Real Life events.

Keywords: Social soundtrack · Formality · Fixed effects · Instagram · Tumblr

1 Introduction

The integration of broadcast media events, mobile technologies, and social media sites has facilitated synergic online interactions that impart feelings of togetherness, information sharing, and conversation among people in dispersed locations, leading to the creation of an online conversation for events and associated content, such as advertising. Viewers exchange information related to the event via second screen devices in terms of posting of social media comments [8]. The use of secondary screens affords the creation

K.S. Xu et al. (Eds.): SBP-BRiMS 2016, LNCS 9708, pp. 13–22, 2016.
DOI: 10.1007/978-3-319-39931-7_2

of what we refer to as the social soundtrack, the online interactions with others regarding broadcast programs, particularly for In-Real Life event (IRL) events such as sporting events and award shows. The effect on the role of the viewer is profound, as the nature of viewership is transiting from a passive to an active one, where the viewer can, to some degree, take action while watching and engaging in an event.

In this research, we consider Super Bowl XLIX as an IRL broadcast media event. We investigate the use of second screens in the *Pre*, *During*, and *Post* phases of Super Bowl XLIX, specifically examining if second screen formalities, or use of proper linguistic terms and syntax, from viewer interactions concerning Super Bowl commercials, game, and music categories are related to the social soundtrack features in each of event's phases. We select Instagram and Tumblr as our data collection sites. The temporal change in patterns used in social media discourse and the quantum of sentences and unique words in the social soundtrack conversation intuitively are the factors for temporal shift in observance of rules of social media etiquette. This intuition motivates our research.

The research is important as changes in language style indicate the credibility and rapport building between people in second screen conversations who do not know each other, which influences the impact of the information shared [6]. The formality of language also has significant impact on how messages are received, and it can be used to identify disparate user groups. Additionally, it is essential to understand how users engage and leverage the affordances of their technology devices for information sharing, which can have a profound effective on areas such as online advertising.

2 Related Work

There has been limited research on formality analysis on social media conversations. Understanding the social soundtrack formality of the conversations can shed light on the goals, needs, and desires of the conversation participants while viewing an event.

Sabater [12] examined Facebook messages of native and non-native English speakers to identify the stylistic variations in their online writing. The result showed that non-native speakers exhibited more formal traits in a university context. In another study on postings of two communities on Twitter, it was found that there was marked difference in formality and tone between the two user groups and the underlying differences of communication goal was cited as the reason [11]. In a separate research using WhatsApp in an university context, Alamri revealed that, over time, instructor discourse became informal [1]. Similar characteristics were found for blogs that claimed to be more informal, and it was shown that the personality of the author influenced the formality of text [9, 10]. Lee, Ham, Kim, and Kim [7] used Twitter as the social media platform to assess people's interest in Super Bowl 2012 car commercials.

None of these prior research studies assessed the temporal interaction effects of social networks and second screens from the temporal strength of linkage between social soundtrack features, patterns, and the content of second screen conversations from a formality perspective concerning live broadcast of major IRL events. Understanding the temporal aspects of formality within the social soundtrack has implications

for leveraging social media within a variety of domains, including online marketing and public service communication. Also, much of the previous studies are limited to a single social media platform.

3 Research Question

In our research, we classify second screen interactions appearing in the social soundtrack concerning Super Bowl XLIX into three event categories: (a) commercials, (b) music and (c) game. There is considerable sharing of feelings in the social soundtrack on three aforementioned categories not only during but before and after the event. We label these temporal phases of the social soundtrack as: (a) *Pre* phase, (b) *During* phase and (c) *Post* phase. As we collect data related to Super Bowl XLIX from the 10th of January 2015 and continued till the 24th of February 2015 on Instagram and Tumblr, the *Pre* phase begins on 1/10/2015-00:00:00 and continue till 2/1/2015-18:29:59 (till the start of the kick off). The *During* phase is the period of the live broadcast of the event, i.e., from 2/1/2015-18:30:00 to 2/1/2015-22:30:00. The *Post* phase is the social soundtrack beginning on 2/1/2015-22:30:00 and lasting till 2/24/2015-00:00:00. We amass 811,262 Instagram media posts and 51,569 Tumblr blogs using respective APIs and secret tokens. We chose Instagram and Tumblr as they are major social media platforms with limited investigation in prior work, relative to Twitter.

In this study, we extracted the social soundtrack features in terms of count of posts corresponding to (a) pattern of viewers' conversation, (b) number of sentences in the postings, and (c) number of unique words present in the texts of the posts. The identifiers for categories of social media post patterns are listed in Table 1.

Table 1. Categories of social soundtrack conversation patterns common to the social media platforms

Category	Description
Referral (RF):	Any full length or shortened URL.
Response (RS):	Postings intentionally engaging another user by means of '@' symbol which does not meet the other requirements of containing referrals.
Broadcast (BC):	Undirected statements (i.e., does not contain any addressing) which allow for opinion, statements and random thoughts to be sent to the author's followers. Any undirected statement followed by questions '?' belongs to Broadcast (BC) category.

For Instagram and Tumblr, RF categories may contain the URLs for images and videos in addition to general full length or shortened URLs in Instagram captions and Tumblr blogs. For Instagram captions, and Tumblr blogs, we set the priority order as: RS > RF > BC. The sentences of the posts are parsed based on the punctuations such as ".", "?" and "!". The number of unique words within each posting is determined by excluding the stop words and the hashtags present in the sentences of each post.

We have an intuition that relationship between interaction features present in social media postings and the sentiments extracted from the social soundtrack conversation regarding specific categories changes in phases. The social soundtrack feature-formality linkage will also most likely change in specific phases for different categories. These feature-formality relationship of language shed light on the manner of information processing and dissemination with the social soundtrack.

Based on this intuition, we frame our research question that deals with influence of interaction features on formality of social media conversations in each phase on different categories. The research question is evaluated by linear regression on balanced panel data.

RQ1. Do the features of social soundtrack conversations on different social media platforms affect the formality of social media conversations in each phase?

4 Research Design

We classify the Instagram and Tumblr data into the three categories of second screen interaction. We identify the categories by means of the keywords collected from relevant websites related to Super Bowl commercials [2, 15], Super Bowl music [16], and Super Bowl game [13]. The list of Super Bowl commercial keywords contains the ad titles, titles of the themes / videos for the ads, hashtags associated with the spots (e.g., '#realstrength', '#likeagirl' etc.), and the first and last names of actors participating in Super Bowl commercial videos. The list of Super Bowl music keywords contains the first name and last name of the performers of the halftime and the pre-game show, terms that describes the half time show (e.g., 'shark', 'palm' etc.), and the songs (e.g., 'california girls'). The list of keywords related to Super Bowl game contains the first name and last name of the players, coaches, umpires, referees, commentators, the field positions(e.g., 'quarter-back', 'red zone' etc.), team names and other key terms related to game (e.g., 'punt', 'fumble' etc.). We assign the posts on Instagram and Tumblr to Super Bowl commercials, music, or game categories, depending on the presence of terms from the respective keyword lists. Prior to any analysis, we perform the following pre-processing steps as: (a) remove punctuations from the sentences of the posts, (b) remove the hashtags, as this does not contribute to the frequency of parts of speech (POS) tags, (c) remove the usernames addressed by "@" and "RT" within the messages, (d) remove the special characters such as "@", "RT", "via" and URLs, (e) replace all contraction of verb forms to the corresponding verbs (e.g., "'ll" to "will", "'ve" to "have", "'re" to "are") (f) replace all negations ("neither", "nor", "never", "no", "negative", "not", "n't", "won't" etc.) to "not", (g) replace a sequence of repeated characters by two characters (e.g., "coooool" to "cool", "ooooh" to "ooh" etc.), (f) lowercase the letters and expand the acronyms in the posts to its meaning extracted from relevant resources.

After pre-processing, we use the Stanford POS Tagger to identify the Part of Speech (POS) tags from the tokenized posts. We use the standard tag-set defined by Penn Treebank to identify the tags as element of POS class (i.e., noun, pronoun, verb, etc.) from the tokens by means of Stanford POS Tagger. Formality is expressed as a function of such POS class elements present in a post. We compute the F-score for

formality as defined in [5], and thereafter, we calculate the aggregated F-score in five-minute intervals in phase-category space, used as the unit of analysis for testing the research hypotheses. Higher F-score of formality indicates more careful and less casual social soundtrack conversations.

We further segregate the posts regarding volume of posts, patterns of conversations exist in posts, number of sentences present in the posts and number of unique words in sentences of posts into five minutes intervals for *Pre*, *During* and *Post* phases. We have an intuition that the volume, sentence and unique words will affect formality score, as those attributes are functions of the social media texts. We transform the five minute time count data regarding volume, each of the social features and the derived formality score into a relative scale using equation: $rel\_value_j^i = Score_j^i / Max_i \{Score_j^i\}$, i denotes the index of the five minute time slot within a specific phase and j is the specific attribute of the posting. Score denotes the values of attributes. *Max* function returns the highest value of the count for a specific attribute within a phase. Here, the attributes are social features (i.e., volume of posts, each of the conversation patterns, sentences, and unique words) and quantized formality.

Once the computation of relative scaling of the attributes is done for the social soundtrack conversation, we organize the categorical time count data into a balanced panel [4] data for both the social media platforms, where each of the three categories has relative values of the social soundtrack attributes across total number of five minute slots for data collection in each phase. Panel data, also known as cross-sectional time series data, can control for variables whose behavior cannot be observed (i.e., behavior of Super Bowl categories). In our study, for each phase, the balanced panel dataset can be viewed as a three dimensional space where the dimensions are (1) Super Bowl categories (commercial, music, game), (2) time stamps for each category (number of five minute time slots i.e., 6558, 49, and 6534 for *Pre, During* and *Post* phases respectively), and (3) social media platforms (Instagram and Tumblr). In each social media dataset, we have a total of 19674 (3×6558), 147 (3×49) and 19062 (3×6354) records each with relative values of attributes of posts for *Pre, During* and *Post* phases, respectively. Each record is the unit of analysis in evaluating the research question for each phase in two different social media platforms.

5 Methodology

We use panel data regression with fixed effects [3, 14] on balanced feature-formality panel data to evaluate the relationship between the features of social soundtrack conversations and formality concerning Super Bowl categories. In the regression model, relative formality scores data is the dependent variable, while the relative values of social soundtrack features (i.e., volume, patterns of social soundtrack conversations from Table 1, number of sentences, and number of unique words) are the cofactors. We conduct the fixed effects regression model on the panel data. The fixed effects model assumes that individual specific effect is correlated with the independent variable. We set the Super Bowl commercials as the baseline category for finding relative categorical effect in the fixed effects model. We are estimating the pure effect of social soundtrack features by controlling the unobserved heterogeneity with the addition of dummy variables for each category in the fixed effects model.

6 Result

The estimates of regression coefficients of the social soundtrack features (cofactors) identify how much second screen formality changes over time on average per category when the respective cofactor is increased by one unit. In the fixed effects regression model, we also evaluate the effect of categories (i.e. unobserved variable) on the linkage between social soundtrack features (cofactors) and the quantified formalities (response) in each phase on Instagram and Tumblr.

6.1 Instagram

From Table 2, we find that formality increases by 10.76 times in relative scale with one unit increase of URL-based captions (RF) in the *Pre* phase for Instagram; however, the unit increase of other cofactors reduces the formality significantly. In the *During* phase, the coefficients of the majority of cofactors are large and positive except sentences and unique words; however, they are not significant (p-value > 0.05) in measuring the effect of the cofactors on the formality of the Instagram captions. In the *Post* phase, formality increases 3.73 times with a unit increase of referral (RF) pattern. It is important to note that increased number of postings with captions increases formality significantly in *Pre* and *Post* phases, while in *During* phase the effect is also positive but not significant (i.e., p-value > 0.05). The variance explained (R^2) in *Pre* and *Post* phases in Instagram is lower than that explained in *During* phase.

Table 2. Fixed effects model results for Instagram

Phase	Cofactors	Coeff	p-value	R^2
Pre	**volume**	**2.834**	**0.011***	0.23
	mention	−4.173	0.057	
	referral	**10.758**	**1.4e-05***	
	broadcast	**−5.720**	**0.009***	
	sentences	**−3.349**	**7.6e-13***	
	unique words	**−2.845**	**0.021***	
During	volume	18.262	0.141	0.57
	mention	39.692	0.662	
	referral	24.200	0.785	
	broadcast	27.728	0.758	
	sentences	−13.142	0.083	
	unique words	−25.610	0.175	
Post	volume	1.967	0.091	0.25
	mention	**−8.815**	**1.4e-06***	
	referral	**3.731**	**0.010***	
	broadcast	**−8.519**	**2.8e-06***	
	sentences	**−3.890**	**2.2e-16***	
	unique words	**−8.543**	**9.9e-08***	

Figure 1 depicts the effects of music and game categories on linkage between social soundtrack features and formality on Instagram. It is seen from Fig. 1 that music and game categories have significant increased effect on feature-formality linkage relative to Super bowl commercials in all three phases. So, among three categories, the fixed effects of Super Bowl commercials is least, while the fixed effects of game category is the highest on the relationship between social soundtrack feature and formality of the content in all phases for Instagram.

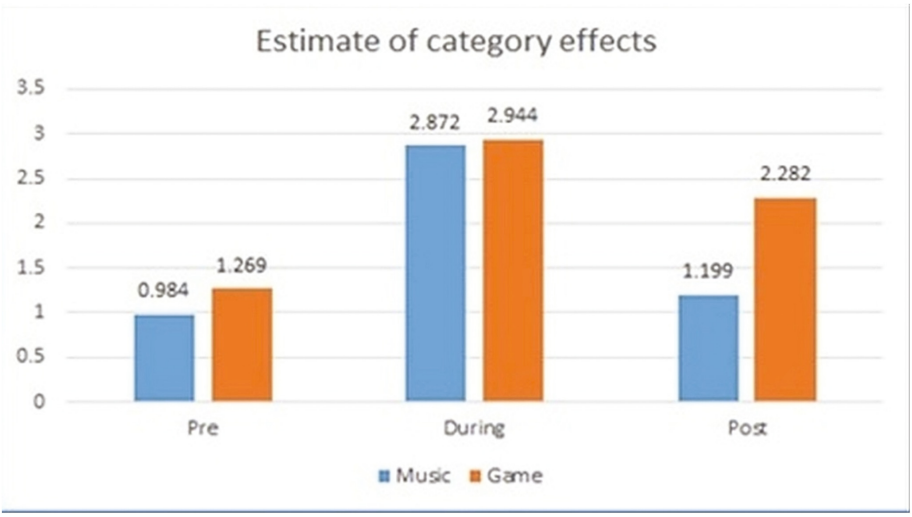

Fig. 1. Effect of music and game in relation to commercials for Instagram by phase

6.2 Tumblr

We present the results of fixed effects regression model done on the panel data for Tumblr in all three phases in Table 3. We find that formality significantly increases by 19 times and 41.19 times in relative scale for a one unit increase of undirected broadcast (BC) pattern and number of sentences respectively in the *Pre* phase for Tumblr, while for unit increase of other cofactors the formality reduces significantly. In the *During* phase, the coefficients of the cofactors are large and positive, but they are not significant (p-value > 0.05) in measuring the effect of the cofactors on the formality of the Tumblr blogs. In *Post* phase, formality increases 25.7 times and 43.21 times with unit increase of undirected broadcast (BC) pattern and number of sentences, respectively. It is important to note that the increased number of Tumblr postings reduces formality significantly in *Pre* and *Post* phases, while in *During* phase the effect is positive but not significant (i.e., p-value > 0.05). Unlike Instagram, variance explained (R^2) in Tumblr by the model in *During* phase is lower than that explained in *Pre* and *Post* phases.

Figure 2 depicts the effects of music and game categories on linkage between social soundtrack features and formality on Tumblr. It is seen from Fig. 2 that music and

Table 3. Fixed effects model results for Tumblr

Phase	Cofactors	Coeff	p-value	R^2
Pre	**volume**	**−44.977**	**2.2e-16***	0.49
	mention	**−14.988**	**1.2e-14***	
	referral	**−7.306**	**0.000***	
	broadcast	**19.034**	**2.2e-16***	
	sentences	**41.195**	**2.2e-16***	
	unique words	**−7.036**	**3.8e-08***	
During	volume	7.848	0.679	0.28
	mention	12.579	0.647	
	referral	26.479	0.352	
	broadcast	17.810	0.539	
	sentences	1.796	0.931	
	unique words	8.161	0.618	
Post	**volume**	**−43.561**	**2.2e-16***	0.57
	mention	**−19.534**	**2.2e-16***	
	referral	**−9.922**	**7.4e-06***	
	broadcast	**25.700**	**2.2e-16***	
	sentences	**43.211**	**2.2e-16***	
	unique words	**−8.462**	**2.8e-10***	

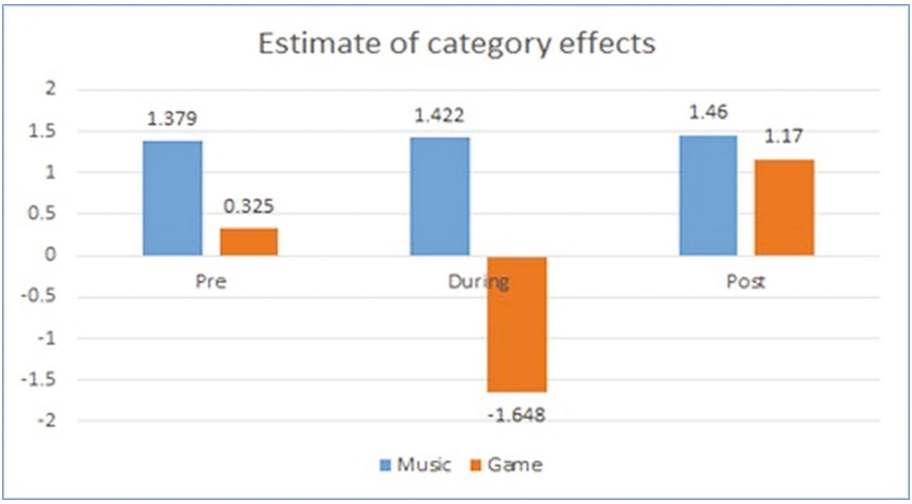

Fig. 2. Effect of music and game in relation to commercials for Tumblr by phase

game categories have significant increased effect on feature-formality linkage relative to Super Bowl commercials in the *Post* phase. In *Pre*, though relative fixed effects of music and game categories are positive, the effects of game in *Pre* phase is not significant (p-value > 0.05). In the *During* phase, there is no significant effect of music

and game categories relative to commercials, though the effect of game is least in the *During* phase. The fixed effects of Super Bowl music category is most pronounced on the relationship between social soundtrack feature and formality of the content in all phases for Tumblr.

7 Discussion and Implication

The variation of conversation pattern based formality on Instagram and Tumblr facilitates to identify the demography of viewers from the change in stylometric variation for categories in different phases. It is also observed that game category has higher estimates of fixed effects on Instagram feature-formality relationship, while the impact of commercials is the lowest. This informality inherent in commercial related posts increase the personalization of the brands and marketing, allowing brands to communicate with viewers in a manner they are comfortable with via social media platforms. In *Pre* and *Post* phases for Instagram, referral or URL based recommendations (RF) pattern has the positive influence on social soundtrack formality, while the other patterns have significant negative influence (see Table 2). For Tumblr, the blogs with more undirected broadcast patterns become more formal in *Pre* and *Post* phases, while blogs containing more of other conversation patterns become less formal (see Table 3). From a feature-formality influence perspective, the relative volume of postings has positive correlation with Instagram formality, while for Tumblr it is negative (see Tables 2 and 3). The feature-formality relationship is insignificant in *During* phase, while the R^2 value is higher for Instagram. This is because of the huge difference in the sample sizes (i.e., 6500 for the *Pre-* and *Post-* phases, only 49 for the *During*). From category effect perspective, for Instagram game category outperforms other two categories in the strength of feature-formality linkage in all three phases. In Tumblr, the music category plays the lead role. This seems to infer that the media based posts that contain URLs in Tumblr are more informal relative to Instagram. It is also observed that game category has higher estimates of fixed effects on Instagram feature-formality relationship, while the impact of commercials is the lowest. This informality inherent in commercial-related posts increases the personalization of the brands, allowing brands to communicate with viewers in a manner they are comfortable with via social media platforms.

8 Conclusion

Our research provides contributions concerning understanding user behavior and interaction in terms of the shift in users' temporal formality concerning effects of categories treated as unobserved variables in the IRL event in a traditional formality computation framework. In future research, we will analyze the relationship between temporal informality of posts and different features of second screen interaction taking care of idiosyncrasies of IRL event related social media texts on more social media platforms with the estimates of fixed and random models, comparing the results with formality settings presented in this research.

References

1. Alamri, J.: (In) formality in social media discourse: The case of instructors and students in Saudi higher education. In: Proceedings of Global Learn (AACE), pp. 101–108 (2015)
2. Anonymous: 2015 Super Bowl commercials (2015). http://www.superbowl-commercials.org/2015
3. Bell, A., Jones, K.: Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. Polit. Sci. Res. Methods **3**(01), 133–153 (2015)
4. Berrington, A., Smith, P., Sturgis, P.: An overview of methods for the analysis of panel data. NCRM Methods Review Papers. NCRM/007, pp. 1–57 (2006)
5. Heylighen, F., Dewaele, J.-M.: Formality of language: definition, measurement and behavioral determinants. Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel (1999)
6. Jansen, B.J., Sobel, K., Cook, G.: Classifying ecommerce information sharing behaviour by youths on social networking sites. J. Inf. Sci. **37**(2), 120–136 (2011)
7. Lee, H., Kim, Y.K., Kim, K.K., Han, Y.: Sports and social media: Twitter usage patterns during the 2013 super bowl broadcast. In: Proceedings of the International Conference on Communication, Media, Technology and Design, (ICCMTD), pp. 250–259 (2014)
8. Mukherjee, P., Jansen, B.J.: Social TV and the social soundtrack: significance of second screen interaction during television viewing. In: Kennedy, W.G., Agarwal, N., Yang, S J. (eds.) SBP 2014. LNCS, vol. 8393, pp. 317–324. Springer, Heidelberg (2014)
9. Nowson, S.: The Language of Weblogs: A study of genre and individual differences. Doctoral Thesis. University of Edinburgh (2006)
10. Nowson, S., Oberlander, J., Gill, A.J.: Weblogs, genres and individual differences. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci), pp. 1666–1671 (2005)
11. Paris, C., Thomas, P., Wan, S.: Differences in language and style between two social media communities. In: Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM) (2012)
12. Pérez-Sabater, C.: The linguistics of social networking: A study of writing conventions on facebook. Linguistik Online **56**(6), pp. 111–130 (2013)
13. Schalter, T.: Super Bowl XLIX: Power ranking the top 25 players in this year's game. http://bleacherreport.com/articles/2343013-super-bowl-xlix-power-ranking-the-top-25-players-in-this-years-game/page/2
14. Schmidheiny, K., Basel, U.: Panel data: fixed and random effects. Short Guides to Microeconometrics, pp. 1–16 (2011)
15. Staff, A.A.: Super Bowl XLIX ad chart: Who bought commercials in Super Bowl (2015). http://adage.com/article/special-report-super-bowl/super-bowl-xlix-ad-chart-buying-big-game-commercials/295841/
16. Wikipedia: Super Bowl XLIX halftime show (2015). http://en.wikipedia.org/wiki/Super_Bowl_XLIX_halftime_show

http://www.springer.com/978-3-319-39930-0

Social, Cultural, and Behavioral Modeling
9th International Conference, SBP-BRiMS 2016,
Washington, DC, USA, June 28 - July 1, 2016,
Proceedings
Xu, K.S.; Reitter, D.; Lee, D.; Osgood, N. (Eds.)
2016, XVIII, 412 p. 131 illus., Softcover
ISBN: 978-3-319-39930-0