Marketing the Mountain State:

A large N study of user engagement on Twitter

Abstract

Much of the evolving research on the use of social media in destination marketing emphasizes how information diffusion influences the reputational image of place. The present study uses Twitter data to focus on the relative differences in user engagement across discrete account types. Specifically, this is done to examine how the official destination marketing organization of Montana—the Montana Office of Tourism (MTOT)—performs relative to other account types. Several regression analyses conducted on Twitter data associated with an ongoing MTOT place branding campaign reveal that tweets sent from 'official' accounts are more likely to be retweeted, and are estimated to receive more total retweets. The inclusion of a URL or mention, and the number of followers an account has, are also predicted to positively impact retweets. These results will be useful for economic development professionals working in state and local governments, tourism and marketing companies and nonprofits, university researchers, and community members who seek to understand how destination marketing is being conducted. Those interested in methodology and data collection techniques for Twitter-based research, and data manipulation in Python may also benefit from the study.

Keywords: Twitter, Destination Marketing, Social Media, Account Type, Retweets, User Engagement, eWOM, Python, Stevenson Center

Kirk Douglas Richardson

https://github.com/kirkdrichardson/twitter-analysis

This research fulfills capstone project requirements for the Politics and Government Applied Community and Economic Development M.S. program in partnership with the Stevenson Center for Community and Economic Development at Illinois State University. *Published June 2017.*

Table of Contents

Scripts used for the collection and manipulation of data used in the study are available at https://github.com/kirkdrichardson. The datasets and executable Stata files used are available upon request: kirk.douglas.richardson@gmail.com

Tables

Charts

Introduction

While place marketing as practiced has been fast to adopt digital branding and content delivery practices, industry best practices for the management of social media tools have been slow to emerge. Increasingly fundamental to any digital communications strategy is the effective use of the microblogging platform Twitter. Its status as one of the most successful social platforms in existence, its public by default nature, and its ability to drive global conversations among diverse audiences make it ripe for analysis and a reasonable proxy for social media strategies generally (Guo and Saxton 2014; Rossi and Magnani 2012; Shannon, Perrin, and Duggan 2016). The addition of a public application programming interface (API) and clear metrics for measuring user engagement add to the attractiveness of using Twitter data for a broad range of analyses. For organizations, the use of Twitter and other social media sites provides a low-cost means to engage a target audience to convey information on a particular program, opportunity, or product, or to collect information on user engagement, opinion, or relevant habits (Chao and Saxton 2014).

While destination marketing organizations (DMOs) slowly develop Twitter best practices, much of the evolving research emphasizes how digital platforms are changing the nature of communication between marketer and targeted audience. Indeed, the degree to which these remain discrete groups has shifted considerably, and will likely continue to do so as branding organizations pursue more creative ways to distribute content. Twitter and the ways in which it provides for the linking and sharing of content is ripe for an exploration of how destination branding agencies are utilizing or failing to utilize unique and personal communication strategies to reach simultaneously broad and targeted swaths of their potential audience.

The present study will focus on the relative difference in user engagement across discrete categories of users to examine how the Montana Office of Tourism (MTOT) is using Twitter to reach an audience of potential tourists. Specifically, the study will measure user engagement through the number of retweets received by each original tweet containing a hashtag associated with an ongoing campaign initiated by MTOT. These results will be useful for economic development professionals working in state and local governments, tourism and marketing companies and nonprofits, university researchers, and community members who seek to understand how destination marketing is being conducted.

Place, eWOM, and social media

The utility of social media phenomena to organizations is created through the shift in how organizations are able to interact with target audiences, and the abundance and relative ease of gathering data associated with those interactions. Social media facilitate highly visible user commentary on a brand, often in the form of various platform-specific methods of approval or disapproval. This phenomenon, often referred to as electronic word-of-mouth (Barbagallo et al. 2012), represents a substantial shift in the potential for how organizations promote a brand or product. Since peer-based recommendations impact an individual's behavior more than exposure to traditional marketing materials, organizations have in social media the potential to dramatically improve their marketing efforts (Govers and Go 2009).

Gartner (1993) labeled information sources influencing the way people interpreted the world as 'image formation agents.' The four most important agents as identified by Gartner (1993) are listed below in order of the most impactful in terms of influence upon an individual.

Organic Agents	personal experiences
Social Agents	word of mouth based on the experiences of peers
Autonomous Agents	news media
Induced Agents	commercially-biased marketing communications

Traditionally, commercial brands tend to rely on organic and induced agents (Govers 2015). In the case of destination marketing, the organic agent would be tangible experience with the location of interest (such as a previous trip), whereas the induced agent might take the form of exposure to a state-promoted advertising campaign.

Naturally, organic agents produce a deeper impression. This is particularly the case with hedonic consumption experiences such as travel, movies, or similarly immersive experiences (Govers 2015). It has also been shown that people are more likely to share stories based on such experiences over those related to the use of more utilitarian goods such as household products (Dhar and Wertenbroch 2000). As Govers (2015) posits, hedonic consumption experiences are often related to place due to the increased engagement and interactivity enabled by the complexity of place.

In this context, the notable impact of social media upon destination marketing is unsurprising. Citing their findings as evidence of the challenge to traditional travel marketing, Xiang and Gretzel (2010), for example, demonstrated that social media sites constitute a substantial portion of the top search engine results related to US tourist destinations. Of course, this phenomenon is not isolated to destination marketing, but part of a broader shift in how information is consumed.

The link between personal experience and social sharing, particularly manifest in online social media, is captured by the term electronic word-of-mouth (eWOM). In marketing literature, eWOM is often used to describe the process and product of user-shared opinions on organizational brands, products, and services (Barbagallo et al. 2012). The clear business value of understanding the eWOM regarding a brand or product has made the mining of the internet, and social media in particular, a strong focus of many organizations. Increasingly, organizations have integrated processes and tools for analyzing social media data into their traditional suite of business intelligence tools (Barbagallo et al. 2012).

However, as a comparatively new area of marketing, established approaches to analyzing and influencing eWOM are slow to form, and slower to diffuse to subdomains of branding. Research on the effective management of digital place branding is notably sparse, which leaves Destination Marketing Organizations (DMOs) with little to direct their efforts (Govers 2015; Hanna and Rowley 2015). Nevertheless, Govers (2015) predicts that information technologies and social media will overtake advertising-driven branding in the future due to the connection between personal experience and eWOM. The present lag in this area may be explained by the considerable difficulty in connecting old metrics of success such as reach and frequency to the ROI of social media efforts. However, marketing management literature is increasingly shifting away from a focus on the short-term ROI of digital campaigns to emphasize long-term relationship building and brand management (Hoffman and Fodor 2010).

Although the processes affecting how information is spread have always been difficult to identify and control, the current availability of data linked to information propagation contributes significantly to a clearer understanding of these processes (Barbagallo et al. 2012). Websites and social media accounts tend to provide the digital access points for most organizations. Hannah

and Rowley (2015) classify these as channels, the first of their "7 C's of Digital Strategic Place Brand Management." They make the point that multi-channel marketing—strategically choosing which message to send through which channel to which audience—poses a significant challenge to place marketers. However, though the complexity of digital marketing is high, it is relatively easy to collect data to inform decision making.

User-engagement is a common metric that companies use to evaluate the success of a digital campaign, and although there are many ways it can be evaluated, social media provide particularly relevant and accessible data that is nearly standardized across various sites (channels). For example, Facebook has metrics such as the number of 'likes', 'shares', and 'friends', whereas Twitter has the number of 'favorites', 'retweets', and 'followers'. Both sites also provide application programming interfaces (APIs) to facilitate the collection and analysis of this data. As one of the most popular social media sites, Twitter has frequently been used to study user engagement with brands (Barbagallo et al. 2012; Sevin 2013), hashtags (Ross and Magnani 2012), organizations (Bhattacharya, Srinivasan, and Polgreen 2014; Guo and Saxton 2014; Yasugi et al. 2013), and related phenomena.

Similar to how an organization might seek to put information in front of a target audience via traditional marketing channels, Twitter functions as a channel with a global audience contained within digital space. Twitter also has the great advantage of being free and relatively simple to use, significantly reducing the barriers many organizations would face utilizing traditional marketing channels. The difficulties in Twitter-based marketing compared to traditional approaches stems from the culture of the Twitter community. Naaman, Boase, and Lai (2010) found that the most popular use of Twitter was as a space for users to post messages related to themselves or their thoughts, followed by general information sharing and

opinions/complaints. Even among a random sample of 350 individual Twitter users whose purpose on the platform was not to sell a product, self-promotion was a notably less popular category. Nevertheless, there are many successful examples of organizations using Twitter to market in ways that are non-intrusive and add value to other users. The potential reach of these creative campaigns is enough to continue driving efforts in this area.

Although competing microblogging services such as Tumblr, Jaiku, Posterous, and Google Buzz exist, Twitter retains a clearly dominant position (Boyd, Golder, and Lotan 2010; Suh et al. 2010). Twitter use has also grown substantially over time. As early as 2010 (four years after its 2006 creation), Twitter had about 105 million users producing more than 50 million tweets per day (Suh et al. 2010). By March 2013, more than 200 million users were producing over 400 million tweets per day (Kim et al. 2013). This is not particularly surprising, as 15 percent of adult internet users in the US used Twitter in 2012, and over half of Twitter users utilized Twitter on a daily basis (Kim et al. 2013).

The Pew Research Center's Social Media Update 2016 demonstrates further growth in usage habits (Greenwood, Perrin, and Duggan 2016). While the number of adult internet users in the US grew by 7.5 percent from 2012 to 2016, adult Twitter users grew by 60 percent in the same time period to 21 percent of all US adults (Greenwood, Perrin, and Duggan 2016). In other words, the odds an adult internet user in the US will use Twitter are 1.4 times greater today than in 2012. In the US, Twitter is also more likely to be visited by someone who is young and educated. Approximately 36 percent of adult internet users between the ages of 18 and 29 use Twitter, compared to only 10 percent of those who are 65 or older. Twitter users are also 1.5 times more likely to have a college degree than no college education (Greenwood, Perrin, and Duggan 2016).

While the non-random demographics of Twitter introduce sampling concerns, the volume and ease of collection make it an attractive choice for studying the ways in which information diffuses through a network based on the engagement of its participants. With the proper controls, user engagement on Twitter has the potential to function as a sort of quantitative focus group built from a global participant pool. The most common metric used to study user engagement, eWOM, and general information propagation on Twitter is a retweet.

Retweeting is a practice of information diffusion in which an original tweet travels via retweet to new audiences comprised of the followers of each of the retweeters (Suh et al. 2010). Retweeting is a core component of how Twitter works. This function is also complementary to the primary motivations of most Twitter users whose purpose for using the platform is to access information (Java et al. 2007; Recuero, Araujo, and Zago 2011). There are many rationales for a user to retweet a tweet— to entertain a particular audience, promote a personal interest, comment on a tweet, et cetera—but in each case retweeting suggests that an original tweet contains valued data (Boyd, Golder, and Lotan 2010).

Recuero, Araujo, and Zago (2011) elaborate on this notion by placing it within a social capital framework. They argue that users make decisions on Twitter regarding whether to share information based on the potential benefits sharing may confer upon them. Specifically, they attempt to determine the benefits a retweet will bring to a user's social circle and argue that the act of retweeting, which shares not only the information of the original tweet but of the original sender as well, is a means by which users accrue and distribute social capital.

For organizations, the accrual of social capital by aggregating retweets is beneficial in two complementary ways. First, a higher number of retweets directly contributes to brand recognition through the expansion of an organization's distribution channels. Second, as a Social

Agent as defined by Gartner (1993), peer-forwarded messages ostensibly lend greater credence to an organization's original message. Given the social media imperative to create value and the great benefit to organizations successful in the task, the adaptation required of marketing departments and organizations has become quite clear. Thus, although principles for the effective use of social media remain underdeveloped, Twitter and other platforms have increasingly become an important component of DMOs (Govers 2015).

Research questions

The present study seeks to contribute to an understanding of how DMOs can use Twitter more effectively. It does so by exploring how the type of account from which a tweet is sent influences retweets in the destination marketing realm. Specifically, the study focuses on the use of Twitter related to an ongoing campaign initiated by the Montana Office of Tourism (MTOT). Since the study will be limited to activities associated with the DMO of a single state, it will function primarily as an applied analysis with the potential to support more representative studies of how account type influences retweets.

The effect of account type upon information diffusion has not been widely studied. This is largely because the analysis of Twitter data is a new field characterized by rapid growth in the available data and technological change of the platform itself. For this reason, standardized metrics and methodological approaches have yet to materialize (Kim et al. 2013), and relatively little is known about why some content becomes highly diffused and other information does not (Suh et al. 2010).

Attempts to better understand information diffusion by estimating predictors of retweets

comprise a large portion of Twitter-based research (Hong, Dan, and Davison 2011; Naaman, Boase, and Lai 2010; Petrovic, Osborne, and Lavrenko 2011; Suh et al. 2010). Other frequent approaches to analysis include the study of a phenomenon through hashtag selection (Rossi and Magnani 2012), or the study of organizations or governments based on account-level data or content analysis of selected tweets (Guo and Saxton 2014; Saxton and Waters 2014; Sevin 2013; Yasugi et al. 2013).

Representative of and frequently cited among studies focused on information diffusion, Suh et al. (2010) sought to determine the factors impacting retweets by dividing content features such as whether a tweet contained a URL, hashtag, or mention from contextual features such as the number of followers an account has, the number of people an account follows, account age, and the total number and frequency of tweets. Using a Generalized Linear Model to regress these predictor variables against retweet count, they found that the number of followers an account had and the number of accounts a user followed strongly predicted retweet probability, while total statuses did so only marginally. In the contextual category, they found that tweets containing a URL or hashtag were more likely to be retweeted.

Similarly, Stiegliz and Dang-Xuan (2012) used a Poisson model to estimate factors associated with retweet count within a sample of 64,431 tweets associated with a 2011 parliamentary election in Germany. They found content features such as the inclusion of a hashtag or URL to be strong predictors of retweets relative to the sentiment-related variables at the focus of their study. Also positively related to the quantity of retweets in their study were the account features, number of followers, and age of the account. While the number of followers of an account has consistently been shown to positively impact retweets, other account features such as the number of public lists to which an account is subscribed has been estimated to both

positively (Petrovic, Osborne, and Lavrenko 2011) and negatively (Yasugi et al. 2013) impact retweets.

This past research informs the first of two research questions of the present study: How and why are #MontanaMoment tweets retweeted? (**RQ1**). The hashtag 'MontanaMoment' has been the focus of an ongoing crowdsourced campaign that encourages Twitter users to tweet their 'Montana Moment', which often takes the form of nature photography aligned with the outdoor recreation emphasis of the Montana tourism industry (AP 2017; MTDC 2017). By isolating tweets in this manner, the present study hopes to gain insight into which factors drive the success of MTOT as measured in user engagement with its Twitter initiatives. While retweets will remain the primary metric of success, overall usage of the hashtag by non-MTOT accounts and related summary statistics will also serve as metrics of engagement.

Noting the dearth of research to guide DMOs, the Montana Tourism and Recreation Strategic Plan 2013-2017—co-authored by the Montana Department of Commerce and MTOT—calls for increasing interaction with visitors through content submission via social media channels, expanding its Twitter presence, and a deeper evaluation of social media efforts (MTDC 2012, p. 68-69). Specifically referencing the potential of positive word-of-mouth expressed on social media to contribute to their destination marketing efforts, the strategic plan identifies the establishment of product/service quality as the first step in solidifying a social media presence. By using retweets as a proxy for word-of-mouth and analyzing aspects of tweets containing the 'MontanaMoment' hashtag, RQ1 will attempt to estimate the contribution of Twitter to MTOT's destination marketing efforts.

The selection of tweets by hashtag is a common strategy for data collection. Using this selection method, Rossi and Magnani (2012) highlight the complementarity of dual networks.

The first network they identified was the more stable, localized networks comprised of followers and friends. The second was the 'topical' network capturing the rhythms of global phenomena. This topical network, they realized, was constructed primarily around hashtags associated with events or other phenomena. Rossi and Magnani (2012) analyzed the directionality of replies and the number of retweets of tweets using the hashtag 'XF5', associated with the real-time comments on the TV show Xfactor Italia. Visualizing communicative patterns in a density graph with account-specific nodes of traffic, Rossi and Magnani (2012) found that the Twitter accounts that received the most replies from users were those officially associated with the show: either the official account of the show, the account of the show host, or the account of one of the judges.

However, in analyzing the retweet patterns, distinctions between official and unofficial accounts began to fade. While the official account of the show was among the top three accounts in terms of retweets, the remaining two accounts were from individual users with no official connection to the show. From this, Rossi and Magnani (2012) concluded that social practices on Twitter vary according to the specific communicative tool. While user identity played a large role in reply-based conversations, content of the tweet played a larger role in whether or not it was retweeted.

This predicted relationship between retweets and account type forms the second research question of the present study: what effect does type of account have upon retweets? (**RQ2**). Insight into this relationship can contribute to the social media strategy of MTOT, and may provide insight to similar DMOs. If, for example, Twitter accounts associated with local businesses on average performed much better than a given DMO, then leveraging the support of the businesses associated with those accounts might serve as a goal of a future campaign.

To an extent, the relative competitiveness of MTOT against other accounts as measured in retweets might serve as a litmus test for its efficacy within its domain. This is because, when isolating tweets by topic, the success of an account is expected to be a function of its centrality to the topic and the overall quality of its content. In estimating the degree of influence various users held over topics on Twitter, Cha et al. (2010) found that consistent levels of involvement focused on a specific topic led to greater levels of topical influence. While their sample was limited to users who had increased their influence over a short period of time, the sample included both 'regular' individual users, celebrities, and larger organizations. Similar to Rossi and Magnani (2012), they concluded that while account type had an effect, the primary driver of retweets was quality of content. Thus, while MTOT may begin with the advantage of being the 'official' account associated with the hashtag, performance as measured in retweets is expected to be explained primarily by the content value of its tweets.

Together, RQ1 and RQ2 are expected to contribute to an understanding of the success of MTOT on Twitter, as well as the broader diffusion of Montana's reputational image as a tourist destination. Metrics such as overall user engagement measured in retweets, ratios of original tweets to retweets, and disaggregation of retweets by type of account will aid in depicting the practical realization of MTOT's Twitter initiatives (RQ1). The relative success of the hashtag campaign, and the comparative success of user engagement by type of account, will also be examined through a series of regression analyses (RQ2). If MTOT consistently generates content of high value—accruing the social capital standing that was the focus of Recuero, Araujo, and Zago (2011)—high retweet performance is expected, similar to Rossi and Magnani (2012) and Cha et al. (2010). If Montana performs better relative to other accounts, it would indicate success as defined in the social media section of the Montana Tourism and Recreation Strategic Plan

2013-2017 objective to implement relationship-building activities that produce identifiable promotional benefits (MTDC 2012, p. 69). Relative success would also suggest that, despite its marketing interest, MTOT succeeded in providing content worthy of being spread by social agents (in this case, other Twitter users) in a way that would contribute to the broader reputational image of Montana.

Research Design

Data used in the study

The collection and cleaning process

To collect the data for each of the datasets employed in this study, open source Python scripts modified from the Github repository gdsaxton/Twitter and Weiai Wayne Xu's (2016) personal site were used to access Twitter's REST application programming interfaces (API) (Saxton 2015). The REST APIs allow for reading and writing Twitter data via a program that provides authentication credentials associated with a Twitter Developer's account and outputs the requested data in JSON[1] (Rest 2017).[2] The Twitter APIs allow the automation of data collection and have the advantage of being free. However, the number of tweets are capped at approximately one percent of all tweets, or 3,200 per handle for the REST API, which can make random sampling difficult, and is one of the methodological concerns in the analysis of Twitter data (Bhattacharya, Srinivasan, and Polgreen 2014; Kim et al. 2013). However, this issue is avoided by creating the datasets employed in the study over a period of time and by targeting either a particular account or hashtag (i.e., the datasets are not intended to be a random sample of

[1] JavaScript Object Notation, a data-interchange format consisting of name/value pairs
[2] Twitter provides other APIs that are more suitable to real-time analysis, such as their Streaming APIs. REST APIs are typically used for reading the profile information of users, conducting singular searches, and posting Tweets.

all tweets).

After using #MontanaMoment to access the Twitter API, the scripts employed in this study were used to create the variables of interest by parsing the raw JSON output before inserting the variables into a SQLite database. Several SQLite scripts were then used to perform some additional manipulation to ensure a full sample of the selected date range before the final variables were constructed using Python's *pandas* package for data analysis.

Python, and the *pandas* package in particular, is a particularly powerful tool for data manipulation and preparation. In the case of the Twitter data employed in this study, the *pandas* package allows for quick and efficient manipulation of variables that would become cumbersome in other programs. For example, the data-type formats of crucial variables such as the UTC timestamp conveying the exact moment in time each tweet was sent is provided through the Twitter API as a string (text) rather than an integer to preserve its precise value. Python's *pandas* package allowed this variable to be converted to a datetime format from which sorting, indexing, and aggregation based upon isolated values within the variable (month, day, hour, minute, second, etc.) became possible.

The present study primarily used the 2-dimensional DataFrame structure of the *pandas* package, similar to R's data.frame, to generate the graphs, variables, and datasets employed in the analysis.[3] Where Python falls short is in modeling the more complex statistical procedures employed in this analysis. Although Python's *statsmodels* package[4] provides strong support for many regression models, the GLM family of models is less developed. For these reasons, analysis for the study was conducted in Stata and R.

---

[3] For more information on Python's *pandas* package, see http://pandas.pydata.org/pandas-docs/stable/
[4] For more information on Python's *statsmodels* package, see http://www.statsmodels.org/stable/index.html

Datasets used in the study

The complete hashtag dataset (HT-full) contains variables associated with any tweet containing the hashtag 'MontanaMoment' (not case sensitive) sent between December 24, 2016, and April 17, 2017. It is a tweet-level dataset with 7,669 rows, and 3,100 unique accounts. In other words, 3,100 independent twitter users sent a cumulative 7,669 tweets containing the hashtag 'MontanaMoment' over a period just short of four months. This data was collected by running a Python script once every seven days over the period of the dataset to collect tweets containing the hashtag 'MontanaMoment'. The seven-day interval period was chosen due to the limitation to the past seven days of tweets (or 1,500 tweets) of the Twitter Search function (Kim et al. 2013). Duplicate tweets were automatically omitted through a comparison with the unique numeric value of each tweet ID.

The HT-full dataset contains all unique tweets (by numeric ID) sent over the period, of which 5,482 (71%) cases are retweets. For this reason, the HT-full dataset will be used primarily to provide summary statistics of the overall activity surrounding the hashtag in question. Such an examination of trends in user activity over time and associated with a given topic is the most common form of Twitter analysis (Kim et al. 2013). For the purposes of the present study, a side-by-side comparison of retweets to original tweets is expected to reveal useful information concerning the dynamics of retweeting (RQ1).

The hashtag dataset containing only original tweets (HT-orig) was constructed from the HT-full dataset. It covers the same period of time, and each tweet similarly contains the hashtag 'MontanaMoment'. In order to build the HT-orig dataset, a binary variable *retweet_dummy* was created from variable *retweeted_status* collected through the Twitter API. The *retweet_dummy* variable was used to select only those tweets in the HT-full dataset that were not retweets. The

tweet-level HT-orig dataset contains 2,187 original tweets sent by 546 unique users. In other words, the 2,187 original tweets of the HT-orig dataset comprise only 29 percent of all tweets, and only the 546 users (18 percent of users in HT-full) that originated a tweet containing the hashtag 'MontanaMoment' over the period.

Since the goal of the study is to isolate how and why these tweets are retweeted, the HT-orig dataset will be the focus of the study, and used for deeper analysis. The variables below are those of the HT-orig dataset. Although several variables are common to both HT-full and HT-orig, most of the variables are unique to HT-orig and serve the purpose of capturing the effect size of theorized predictors of retweets.

Variables used in the study

The HT-orig dataset contains two dependent variables and 23 independent variables organized by their role in the study. The independent variable categories are account-type (7), user-objects (3), and tweet-objects (13). The account-type variables are those of particular interest to the study (RQ2), while the remaining variables function as controls.

The only non-numeric variable in the dataset is *from_user_screen_name*, which represents the unique account name of each user. This variable will be used for indexing and aggregating data in order to determine various aspects of unique users within the dataset, determining outlier cases and constructing variables to handle these, and for clustering standard errors by user to reflect the non-independence of tweets within the dataset.

Dependent:

The first dependent variable of the study, *retweet_dummy*, is a binary variable coded as 1 if a tweet was retweeted, and 0 otherwise. With only 667 (30.5%) positive cases for *retweet_dummy*, retweets are moderately rare events. The second dependent variable of the study, *retweet_count*, is a count variable that measures the number of times an original tweet was retweeted during the period of the study. It ranges from 0 (69.5% of cases) to 136, with a mean of 2.38 retweets per tweet.[5] Table 2a below displays the top 10 *retweet_count* values aggregated by user account. Although the *retweet_count* variable is a tweet-level variable, the totals per account will have bearing upon several independent variables.

5 The discrepancy between the mean of retweet_count and total retweets in the HT-full database (i.e. $2.83 \times 2187 \neq 5482$) is due to the variable's basis on original tweets. The difference is created by 287 retweets that appended the 'MontanaMoment' hashtag to original tweets that did not include it. Since the retweet_count variable is associated with original tweets, the 287 retweets without an original tweet counterpart in the HT-orig dataset are not reflected in the retweet_count variable.

Table 2a: Aggregate retweets by account (HT-full)

Unique Account	Total Retweets
visitmontana	2,433
LeonKauffman	728
GlacierNPS	487
DancingAspens	163
KLeaguePhoto	118
BlueMountainBB	96
mislaphotoguy	91
RadleyIce	44
earthXplorer	44
WildReflections	40
All other Accounts	962
Total	**5,206**

Independent:

The independent variables consist of a series of binary variables constructed by parsing the *from_user_screen_name* variable and manually assigning values. Each is intended to contribute in some way to answering RQ2: what effect does type of account have upon retweets?

The primary independent variables consist of five categories determined by the type of account from which an original tweet is sent: the official Montana Office of Tourism account, other official state accounts, accounts of individuals with a clear business interest, accounts of other individuals, and accounts of businesses and organizations not associated with the state. Combined with an additional two outlier account variables described below, the account-type

variables are mutually exclusive and jointly exhaustive of the *from_user_screen_name* value associated with each tweet. Table 2b displays the relationships between these seven variables.

Although the nature of the account categories selected for analysis is unique to the study, the process of measuring the relationship between type of account and number of retweets has been conducted previously. Rossi and Magnani (2012) found that, while account type was important in determining the number of replies to a tweet, account type mattered less in determining retweets. However, the nature of their sample—tweets associated with an episode of a popular show, captured in real-time—was considerably different, and does not provide the predictive portability to inform the estimated effects upon retweets of the account variables in the present study.

The nature of Twitter is such that most studies measuring retweets with consideration for accounts do so by collecting all of the tweets sent from particular accounts (often competing or complementary) and analyze them in comparative perspective (e.g., Bhattacharya, Srinivasan, and Polgreen 2014; Guo and Saxton 2012). The process by which one must collect tweets by a particular hashtag precludes nuanced analysis by type of account through means other than manually coding variables. This similarly makes detailed content analysis onerous, as the natural language processing employed in sentiment analyses of tweets typically provides only categorical series of positive, negative, or neutral sentiment variables (Barbagallo et al. 2012; Bhattacharya, Srinivasan, and Polgreen 2014; Bruni and Francalanci 2012).

Since the HT-orig dataset contains only 546 unique accounts ('users' will be employed interchangeably), it was possible to manually sort accounts into one of the five buckets identified above. Each variable in the five-variable series is coded as 1 if it meets the condition specific to the variable, and 0 otherwise. The first binary variable in the series, *mtot*, is coded as 1 if the

from_user_screen_name variable is equal to 'visitmontana', the official handle of the Montana Office of Tourism (MTOT). As the account name was known beforehand, it was possible to code this variable by using Python to loop through the *from_user_screen_name* column of the dataset.

With 108 tweets in the dataset (4.9 percent), MTOT was the second highest consumer of the hashtag among individual accounts, but fifth for most tweets when compared to the outliers and other account-type variables (see Table 2b). Despite this, *mtot* received the highest percentage of its tweets retweeted (100%), and the most total retweets at 2,433 (46.7% of all retweets).

The second binary variable in the account-type series is *state_account.* It is coded as 1 if the Twitter account is associated with an official state office that is not MTOT. Similar to the remaining three in this five-bucket series, *state_account* was determined by researching the handle provided by the *from_user_screen_name* variable. Examples include the state-sponsored workforce attraction agency Choose Montana (@chooseMontana), Glacier National Park (@glacierNPS), and several city-based destination marketing and economic development organizations. As Table 2b below reveals, *state_account* has the fewest number of tweets in the database (66), although the second highest percentage of tweets retweeted among the account-type variables (third highest including outliers).

The third binary variable in the series is *bus_int_ind,* and is coded as 1 if the account is associated with an individual with a clear business interest such as a photographer, peddler in wares, or established blogger. This variable contains the single largest number of cases, and represents a relatively diverse group of people. The fourth binary variable in the series is *other_ind,* and is coded as 1 if the account is associated with an individual without any apparent business interests. Though it was the third largest group in terms of cases, the *other_ind*,

category received the lowest percentages of tweets retweeted (17.6%), as well as the lowest number of total retweets (131, only 2.5% of all retweets) of the account-type variables.

The fifth binary variable in the series is *bus_orgs_unofficial.* It is coded as 1 if the associated Twitter handle is one of a business or organization that is not associated with the state. The overwhelming majority of these organizations are based in Montana, although several out-of-state organizations are also included. The range of businesses is broad, but some representative examples include Billings Gazette (@billingsGazette), a local news organization; Dude Ranch Vacations (@dudeRanchers), an all-inclusive vacation company; and Wild Reflections (@wildReflections), a fine art gallery.

The two remaining account variables are outliers, characterized as those accounts that have more than 100 tweets (cases) in the dataset. One such account, @visitmontana, is the *mtot* variable previously selected for independent analysis, and thus categorized as one of the account-type variables above. However, the remaining two accounts with greater than 100 tweets were not selected for analysis in isolation. Rather, @TheExceptionMag with 318 tweets would have fallen into the unofficial organizations and businesses (*bus_orgs_unoffical*) category, and @LeonKauffman with 101 tweets would have fallen into the business-interested individuals (*bus_int_ind*) category.

The *exception_mag* variable is coded as 1 if the tweet was sent from the account 'TheExceptionMag' associated with the online food and travel industry publication *The Exception Magazine* based in Portland, Maine. The account is the top consumer of the 'MontanaMoment' hashtag, with 318 positive cases (14.5 percent of all tweets). However, as Table 2b below reveals, none of its tweets were retweeted, which it makes it an outlier for several crucial reasons.

The account is a highly active tweeter (76,900 tweets, 569% more than MTOT), but has relatively few followers (2,269 followers, only 3.8% of MTOT's 60,500), and a much higher followed-accounts to following-accounts ratio than MTOT (0.48 to 0.006). Examining the account's Twitter page reveals very few retweets for any of the account's original tweets. Furthermore, tweets sent from the account contain no identifiable human contribution. Rather, each tweet contains a link to a news article on the organization's website, with the tweet content a replication of the article's title. Considering these details, and following a classification system proposed by Chu et al. (2012), the *exception_mag* account is suspected to be a bot.[6]

Following this assumption, the *exception_mag* variable will be used to identify and exclude the 318 positive cases from the regression models, reducing the cases analyzed in each model to 1,869. This omission will also help to ensure the significant deviation of *exception_mag* (none of its tweets were retweeted) does not influence the covariate estimates of the other variables.

Unlike *exception_mag*, the second outlier is suspected to be human, and will be included within the regression models. The *kauffman* variable is coded as 1 if the tweet was sent from the account @LeonKauffman associated with the Montana-based photographer of the same name. The @LeonKauffman account commands both the third highest number of tweets in the dataset (101) and the second-highest total number of retweets at 728 (14% of all retweets). With 92.1 percent of tweets from the account retweeted, *kauffman* is the only other categorization of user

[6] Bots are relatively common on Twitter, due to lax automation controls. Only during the initial registration must a user fulfill a CAPTCHA image request, after which a bot could use the gained login information to access Twitter APIs and perform routine human tasks. Chu et al. (2012) created an automated classification system to determine whether a Twitter account could be classified as human, bot, or cyborg (a human-assisted bot, or bot-assisted human). In order to develop their proposed model, they created a ground-truth set of data containing accounts known to be human, bot, or cyborg. Their process for attempting to discern these initial accounts is what is employed here.

accounts that comes close to the 100 percent retweet rate of the *mtot* variable. By separating this outlier from its natural category, the study seeks to prevent the impact of the *bus_int_ind* variable from becoming an impact of the single account @LeonKauffman.

Table 2b: Relationships between the dependent and account-type variables

Variables	Positive cases (number of tweets)	% total	Cases retweeted sum(retweet_dummy)	% total	Total retweets sum(retweet_count)	% total
mtot	108	4.9%	108	100.0%	2,433	46.7%
state_account	66	3.0%	37	56.1%	619	11.9%
bus_int_ind	709	32.4%	215	30.3%	853	16.4%
other_ind	391	17.9%	69	17.6%	131	2.5%
bus_org_unofficial	494	22.6%	145	29.4%	442	8.5%
kauffman	101	4.6%	93	92.1%	728	14.0%
exception_mag	318	14.5%	0	0.0%	0	0.0%
Total	2,187	100.0%	667	30.5%	5,206	100.0%

Control:

The right-side variables classified as control fall into either a user-object or tweet-object category. This follows a classification scheme similarly employed by other studies analyzing factors impacting retweets (e.g. Bhattacharya, Srinivasan, and Polgreen 2014; Suh et al. 2010). The purpose of this distinction is to differentiate between those variables that are constructed based upon characteristics associated with a specific user from those associated with a specific tweet. In both cases, the tweet remains the unit of analysis. Data associated with the 546 unique users will simply be included as a non-independent column of each tweet (n=2,187).

User-object

The user-objects series of variables are those variables that are specific to a particular account. Thus, while the HT-orig dataset contains 2,187 rows, the uniqueness of the user-object variables is a function of how metrics associated with 546 unique accounts change over the period of collection.[7] For example, if an account has 100 tweets in the dataset, a variable capturing the number of followers of an account would change only if the user lost or gained followers between sending a #MontanaMoment tweet.

In their unabridged form, each of the user-object variables are prefaced with the identifier 'from_user_', which is how they are denoted by the Twitter API. Each of these continuous predictor variables is standardized to have a mean of zero and a standard deviation of

7 As the exception_mag variable will be used to exclude the 318 tweets sent by @TheExceptionMag, this value will change to 545 unique users in the regression model.

one to facilitate interpreting the effect size of the predictor variables in relation to one another[8]. Appendix A provides summary statistics for the original values of these variables.

The user-object variables account for those factors independent of message crafting and instead tend toward the social interconnectivity and distribution reach of a particular account. These variables are included largely due to their expected impact upon retweets, although findings regarding the effect-size of this series could also inform organizational strategy. If, for example, the number of followers an account has outweighs any other category in terms of impact upon retweets, an organization may wish to direct more of their social-media efforts toward recruiting followers themselves rather than partnering with 'influencers' in a given area.

The standardized variable *followers_count* measures the number of followers an account had at the time an associated tweet was inserted into the dataset. Assuming people follow accounts that publish content of interest, *followers_count* is expected to capture an interesting-content aspect of user accounts. Insofar as number of followers broadens the distribution channel of an account, *followers_count* is expected to have a strong, positive impact upon retweets, similar to other studies (e.g. Suh et al. 2010; Yasugi et al. 2013). Simply put, the larger the distribution channel of a given user, the more likely a tweet is to be seen and chosen to be retweeted by a follower.

The variable *listed_count* is a standardized measure of the number of public lists on which a user appears. Within Twitter, a list is a group of curated accounts from which a user can

8 As the exception_mag cases excluded in the regression model are included in the standardization, the final values for these variables vary slightly. See Table 2d below for the summary statistics of all variables as they are included in the regression models discussed below.

choose to receive tweets if she subscribes to a list of which an account is a part (Petrovic, Osborne, and Lavrenko 2011). Users are able to add followed accounts to lists in order to organize them, and in this sense *listed_count* captures the degree to which an account is actively followed. While Yasugi et al. (2013) identified a weak correlation between number of retweets and membership in public lists and a strong correlation between membership in lists and number of followers, Petrovic, Osborne, and Lavrenko (2011) found membership in lists to be a moderate predictor of retweets. Since the HT-orig dataset does not contain many of the high-status accounts (federal offices, celebrities, etc.) included within other samples, *listed_count* is expected to have only a minimal impact upon retweets when controlling for *followers_count*. This is based upon the assumption (supported by Yasugi et al. 2013) that in order to be added to a large number of lists, an account must have a large number of followers relative to the global and geographically independent Twitter community.

The variable *statuses_count* is a standardized measure of the total number of tweets sent by a user since creation of the account.[9] It is expected to have only a minimally positive relationship with retweets when controlling for factors such as *followers_count* (similar to Suh et al. (2010) and Yasugi et al. (2013)). While Bhattacharya, Srinivasan, and Polgreen (2014) found number of statuses to be negatively associated with retweets, their sample was isolated to official organizations within the medical industry. For this reason, the overall number of tweets associated with accounts within their sample was much higher than in the HT-orig dataset. Since the organizations with the highest number of statuses tend to be those that post regularly and

[9] It is also worth pointing out that the *statuses_count* variable is unique to each tweet, since every new tweet adds a value of one in the unstandardized form of the variable. So, while the other user-object variables may or may not change on a case-to-case basis, *statuses_count* always will.

receive considerable user-engagement as opposed to occasional posters whose primary interests may be in information consumption (rather than information distribution), it is expected that findings will be contrary to those of Bhattacharya, Srinivasan, and Polgreen (2014).

Tweet object

In contrast to the user-object variables, the tweet-object variables are those that are unique to a particular tweet. They concern either the content of a particular message (2 variables), or the day of the week or time of day it was sent (11 variables). The tweet-object variables are intended to control for past findings shown to impact retweets as well as for phenomena not captured elsewhere. The first, *url_dummy*, was coded as 1 for all cases in which the url_count variable provided by the Twitter API was not equal to 0. Of the 1,869 tweets included in the regressions below, 1,010 (54 percent) contain a URL.[10] Similar to other studies (Bhattacharya, Srinivasan, and Polgreen 2014; Suh et al. 2010), *url_count* is expected to be positively associated with retweets, although this may be due in part to the established use of URLs in tweets.

The *mention_dummy* variable was similarly coded as 1 if a tweet contains an @user mention (i.e. a direct address of sorts by one user to another) by identifying all positive cases of the mentions_count variable provided by the Twitter API. A zero value represents the absence of a user mention. The *mention_dummy* variable is expected to be positively associated with retweets similar to the study by Bhattacharya, Srinivasan, and Polgreen (2014), which found that

[10] Each of the 318 excluded *exception_mag* tweets include a URL.

tweets containing a mention were both more likely to be retweeted and more likely to receive multiple retweets.

The remaining control variables in the tweet-object category pertain to time, which has been shown to impact whether a tweet is retweeted (Petrovic, Osborne, and Lavrenko 2011). For example, Semiz and Berger (2017) found that tweets are most likely to be retweeted in the evening[11], while Luo et al. (2013) found night to be the least active time, with only 12.4 percent of tweets retweeted. Semiz and Berger (2017) also estimated tweets sent on a Saturday or Sunday were more likely to be retweeted relative to the remaining five days of the week. To control for such time factors that might impact retweets, several time variables are included in the regression models. The time variables were created in Python by indexing the dataset to various aspects of the *created_at* variable provided by the Twitter API and creating variables unique to the index of each tweet. Seven binary variables (*monday, tuesday, wednesday, thursday, friday, saturday, sunday*) are included to control for the day of week a tweet was sent, each coded as 1 if a tweet was sent on that day.

Given the nature of how the dataset was collected, the day of the week variables are also expected to control for an additional aspect of the tweets. One of the methodological concerns with the present data involves the age of a tweet within the dataset. Once inserted into the SQLite database, the tweets accessed each time a Python script was run remained immutable. That is, each variable associated with a tweet did not change after it was downloaded, thus any retweets an original tweet received after the running of the script are not captured by the dependent

[11] Defined as within the interval 6:00 PM to 11:59 PM.

retweet count variable. This makes the dependent variables of this study (*retweet_count* and *retweet_dummy*) relevant to the unmeasured time-to-be-retweeted aspect of each tweet. Since the collection script was run once every seven days, the time-to-be-retweeted interval ranges from seven days up until the moment the script was run.

However, the rapid consumption and general volatility of tweets significantly reduce the extent to which this is a problem with the data. Kwak et al. (2010), for example, show that more than half of the retweets a given tweet receives will occur within one hour, and 75 percent within a day (also see Luo et al. 2013). In a temporal sentiment analysis, Barbagallo et al. (2012) found that 92 percent of tweets classified as carrying sentiment (positive or negative) were retweeted within five hours, and that 17 percent of all tweets peaked within the first minute. By distributing retweet counts for each tweet over a period of four hours, Barbagallo et al. (2012) also display a graph resembling a power-law distribution, suggesting that retweets for a given tweet decline rapidly the older a tweet becomes. Specifically, a third of peaks (in retweet counts) occurred within two minutes, over half in five minutes, and 80 percent within 31 minutes.

For these reasons, the uncaptured age of the tweets within the HT-full and HT-orig datasets is perceived as a notable though minor concern with the data used for the study. Nevertheless, the day-of-week controls are instituted as a proxy for a time-to-be-retweeted variable. Given past research and the shape of the present data, these stand-in variables are expected to function as hypothesized.

Assuming a standard aggregate retweet-to-tweet ratio, the consistency of the retweet-to-tweet ratio would seem to suggest that most retweets occur on a single day, which is consistent with past research. Chart 2a and 2b show, respectively, the total number of tweets and total

number of retweets occurring on each day of the week. Table 2c reveals that the average number of retweets per tweet by day of the week is 2.9, with the largest deviation being only 0.6 retweets. This is reflected by comparing the Tuesday values for each of the graphs. However, every other day is within 0.2 retweets of the mean. The combination of these comparisons and the light of past research suggests that the large majority of retweets will occur on the same day a tweet is posted, and thus mitigate the potential issue of a missing time-to-be-retweeted variable. Nevertheless, the day-of-week variables will not be interpreted as capturing the impact of day-of-week upon retweets, but as the combined influence of day-of-week and time-to-be-retweeted.

Chart 2a: Number of original tweets by the day of week posted (n = 1,869)

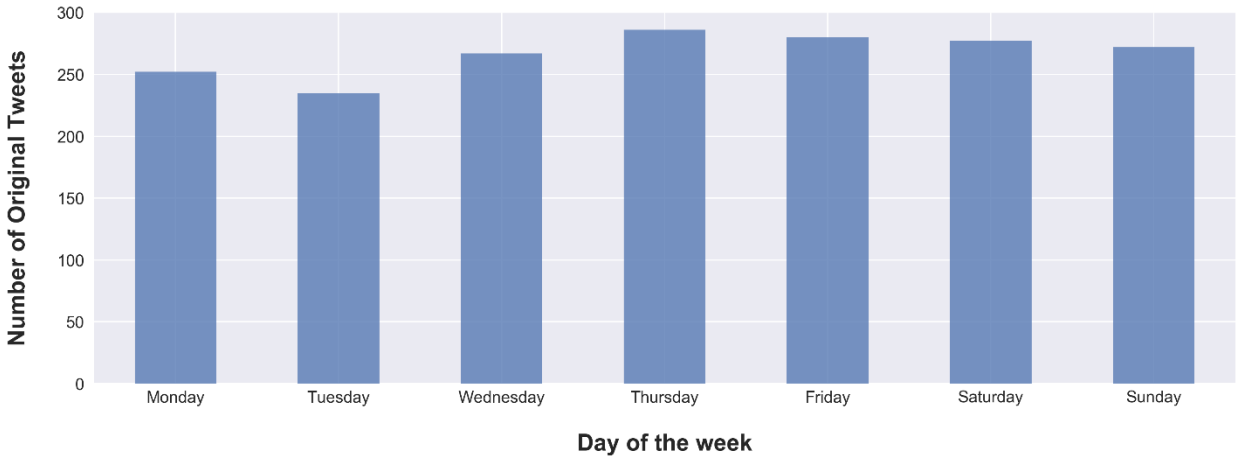


Chart 2b: Number of retweets by the day of week retweeted (n = 5,482)

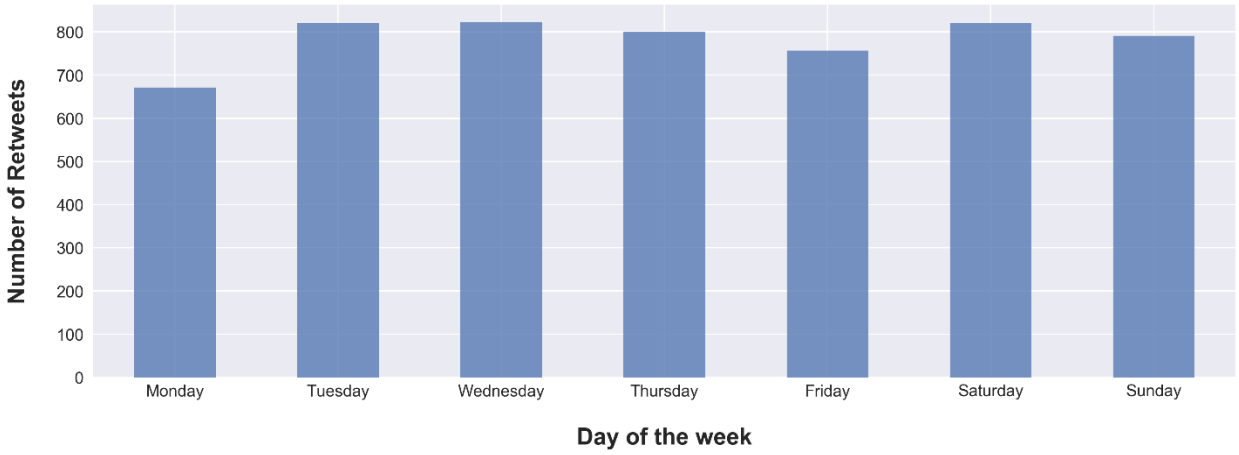


Table 2c: Average retweets per tweet by day of the week

Day of Week	Original Tweets	Retweets	Retweets Per Original
Monday	252	671	2.7
Tuesday	235	821	3.5
Wednesday	267	822	3.1
Thursday	286	800	2.8
Friday	280	757	2.7
Saturday	277	820	3.0
Sunday	272	791	2.9
N	**1,869**	**5,482**	
Average			2.9

The remaining time variables of the tweet-object category include measures for the time of day a tweet was sent. Four time-of-day variables were created by indexing each tweet by the hour of the day it was sent in local time[12] and separating tweets into four buckets based on the hour in which it was sent. The *night* variable is coded as 1 if a tweet was sent between 1:00 AM and 6:00 AM and 0 otherwise. The *morning* variable is coded as 1 if a tweet was sent between 6:00 AM and 12:00 PM, and 0 otherwise. The *day* variable is coded as 1 if a tweet was sent

[12] The Twitter API's *created_at* variable provides a UTC timestamp, from which local times were derived. Since MDT is six hours behind UTC and the time-of-day variables are constructed at six-hour intervals, what would have been *day* in the original dataset was categorized as *morning* here. The underlying assumption for doing so is that most of the tweets and retweets are sent from Montana-based accounts. As the most represented accounts in terms of either tweets or retweets (*mtot*, *kauffman*, *state_account*) are Montana-based there is substantial support for this assumption. However, since accurate geographic information is difficult to obtain it would be a significant challenge to ensure that times are local to where a tweet was sent. A pure time accounting would involve capturing the local times of sending users, retweeting users, and the followers of each unique user in the database. However, the difficulty, or possibility, of doing so is such that the theoretical and empirical suggestion that most tweets are tweeted and retweeted in local or near-local time is deemed sufficient.

between 12:00 PM and 6:00 PM, and 0 otherwise. The final time-of-day variable, *evening*, is coded as 1 if a tweet was sent between 6:00 PM and 12:00 AM, and 0 otherwise.

Table 2d below displays summary statistics for the complete set of variables as they are included in the regression models, organized by their respective category.

Table 2d: Summary statistics of all variables by category

Variables	Count	Mean	Std. Dev	Min	Max
dependent					
retweet_dummy	1,869	0.36	0.48	0	1
retweet_count	1,869	2.79	8.63	0	136
account-type					
mtot	1,869	0.06	0.23	0	1
state_account	1,869	0.04	0.18	0	1
bus_int_ind	1,869	0.38	0.49	0	1
other_ind	1,869	0.21	0.41	0	1
bus_orgs_unofficial	1,869	0.26	0.44	0	1
kauffman	1,869	0.05	0.23	0	1
user-objects					
followers_count	1,869	0.04	1.08	-0.31	8.17
listed_count	1,869	0.02	1.08	-0.29	10.23
statuses_count	1,869	-0.28	0.78	-0.60	8.08
tweet-objects					
mention_dummy	1,869	0.18	0.38	0	1
url_dummy	1,869	0.54	0.50	0	1
monday	1,869	0.13	0.34	0	1
tuesday	1,869	0.13	0.33	0	1
wednesday	1,869	0.14	0.35	0	1
thursday	1,869	0.15	0.36	0	1
friday	1,869	0.15	0.36	0	1
saturday	1,869	0.15	0.36	0	1
sunday	1,869	0.15	0.35	0	1
morning	1,869	0.07	0.26	0	1
day	1,869	0.30	0.46	0	1
evening	1,869	0.31	0.46	0	1
night	1,869	0.32	0.47	0	1

Model

The present study employs binary logistic regression and negative binomial regression to estimate the effect sizes of a series of predictor variables upon whether or not a tweet is retweeted (*logit*), and how many times a tweet is retweeted (*nbreg*). While the *logit* model was a clear choice in terms of appropriateness for the data, the count portion of the analysis was carefully selected through a process discussed in greater detail below.

Considerations in model selection

Ultimately, negative binomial regression was chosen over Poisson regression for its ability to handle the overdispersion of the *retweet_count* variable by parameterizing the overdispersion parameter (α) as $\ln(\alpha)$ (Cameron and Trivedi 2013, 80-89). However, several competing models were considered before concluding negative binomial regression was the best fit for the data. As the standard count model, Poisson regression was the first choice. However, the non-normal distribution of the *retweet_count* variable and the general spreadness of the data favor the derivative negative binomial regression, which better handles overdispersion. However, it was deemed appropriate to evaluate this assumption more thoroughly, especially considering the variety of statistical methods that have been employed by other studies to analyze retweets. Further exploration of model fit was guided primarily by the desire to handle the overdispersion of the *retweet_count* variable while maintaining theoretical credibility in regard to the process by which zeros are generated in the dataset.

Even with *exception_mag* omitted (n = 1,869), the *retweet_count* variable has a variance of 74.4 and a skewness of 6.7. In other words, the distance between these measures and the mean

of 2.8 is far from meeting the underlying assumption of Poisson regression that the variance is equal to the mean (Rodriguez 2007). The hypothesis that negative binomial regression would better handle such a situation was measured formally by conducting a model fit test in Stata using the user-defined SPost13 Package (Long and Freese 2014). As expected, the difference in BIC scores between the models strongly favored the negative binomial model over the Poisson.[13]

Zero-inflated models (Poisson and negative binomial) were also considered for their goodness-of-fit. Of particular interest was the potential of a zero-inflated model to handle the large number of tweets never retweeted. Only 35.7 percent (667) tweets of the 1,869 included for analysis in the models were retweeted. In the interest of measuring what a Bayesian model selection approach (see Raftery 1995) would suggest regarding the utility of a zero-inflated model to handle the large number of zeros in the dataset, a series of model fit tests were performed.

Using the fitstat model estimation procedure to compare zero-inflated Poisson regression (zip) to the negative binomial model (nbreg), the zero-inflated model failed to converge when including the full variable series for the inflate portion of the model. Fitting the inflate portion to the constant, the zip model converged. However, the BIC metric still provided strong support for the nbreg model.[14] Given the relatively weak (though improved) ability of zero-inflated Poisson regression to handle overdispersion, these results aligned with expectations.[15]

A Bayesian comparison of zero-inflated negative binomial regression (zinb) with

---

[13] At 5,551.4, the BIC of the negative binomial regression model is smaller than the Poisson model by a difference of 3,570.2. Raftery (1995) explicates in detail the benefits of using a Bayesian approach to model selection.
[14] At 5,551.4, the BIC of the negative binomial regression model is smaller than the zero-inflated Poisson model by a difference of 2,209.8.
[15] The failure of the zip model to converge in the non-constant-only model is estimated to be the result of some collinearity among some of the predictor variables (account-type and user-object). Theoretically, this is not expected to influence the results in the other non-linear models, although it does provide some limitations in model selection.

negative binomial regression (nbreg) revealed support for the zinb model. However, the difference in BIC scores between zinb and nbreg was much lower than the difference between zip and nbreg and between Poisson and nbreg. In other words, the 'recommendation' of the Bayesian model selection approach is much weaker than in the former cases.[16] Fundamentally, however, the decision to choose negative binomial regression over its zero-inflated variant to model the count variable was one of theory.

Using a zero-inflated model presumes that zeros in a dataset are generated by two independent processes, whereas the choice to use a standard count model assumes that both zeros and positive cases originate through the same process. A useful example based on Allison (2012) is a dependent count variable measuring childbirth among women over 18 living in the US. Regressing on this measure variables for age, sexual activity, and marriage status in a standard count model assumes that if a woman did not give birth, that result can be measured as a function of the independent variables. Conversely, a zero-inflated model would control for the impossibility of a sterile woman to give birth, irrespective of any right-side variables. In other words, the zero-inflated model presumes that zeros in a dataset are generated through two independent processes. Statistically, this is represented by a two-component mixture model in which a binary model is used to model the unobserved state based upon a zero-inflated density that compares a zero-based point mass with a count distribution (Zeileis, Kleiber, and Jackman 2008; for more detail, see Cameron and Trivedi 2013).

Given the theoretical constraints of employing a zero-inflated model, Allison (2012)

---

[16] At 5,551.4, the BIC of the negative binomial regression model is greater than the zero-inflated negative binomial model by a difference of only 150.7. While this still provides strong support for the zero-inflated model, it is not as conclusive as the former comparisons. Fundamentally, the decision is one of theory.

recommends against selecting a zero-inflated model to handle overdispersion even in cases where a BIC comparison would suggest otherwise. Applying this consideration to the present data, it appears the use of a zero-inflated model would bias the covariate estimates since zeros are not structurally determined. In other words, there is no circumstance that would render it impossible for a tweet in the dataset to have been retweeted. The near exception to this is a single account that had no followers at the time a tweet was imported. With no followers, a tweet from this user would only be seen and potentially retweeted if other users searched by #MontanaMoment or otherwise stumbled upon the sending user's page. If a greater number of accounts had no followers, a zero-inflated model might be more appropriate. However, even in this case, zero values would remain only nearly structural given that a tweet sent from an account without followers could still be retweeted. For these reasons, the *followers_count* and time variables are expected to control for possible outcomes in which tweets were never seen. Following this assumption, a zero-inflated model is not supported, even though some studies are more flexible in its application.

Right-side variables

Three variables, one from each of the mutually exclusive and jointly exhaustive dichotomous variables series represented in the model, are omitted from the regression equations in order to avoid their perfect prediction by the linear combination of the remaining variables in the series. As the account of primary interest, the *mtot* variable is omitted from the account-type series so that the coefficients of the remaining variables may be interpreted in relation to it. For those variable series serving as statistical controls, the decision of which variable to omit carries

less weight. From the time-of-day variables in the tweet-object category, the *morning* variable was omitted, and from the day-of-week variables in the tweet-object category, the *sunday* variable was omitted. Any discussion of the remaining variables in these series will thus be interpreted in relation to these omitted variables.

With these exceptions, each of the predictor variables discussed above was regressed on both the binary variable *retweet_dummy* and the count variable *retweet_count.* As such, both the *logit* and the *nbreg* models include 19 right-side variables representing the above explicated categories account-type (5), user-object (3), and tweet-object (11). Following a comparative analysis of retweets in relation to original tweets, the results for the *logit* and *nbreg* models are discussed in the findings section below.

Findings

The findings section is broken into two primary parts. The first will focus on summary statistics and data visualization of the HT-full dataset in comparative perspective with the HT-orig dataset. The purpose is to gain a deeper sense of the overall activity and usage habits surrounding the 'MontanaMoment' hashtag. The nature of Twitter data is such that a significant amount of useful information can be gleaned by isolating various elements of tweets categorized, in this case, by whether or not the tweet is a retweet or an original. Some of the manipulations below involve summary statistics and deeper analyses of variables of interest, and others involve aggregation of the number of tweets in the dataset associated with measures of time. By 'collapsing' the data into a time variable, it becomes possible to identify patterns shared by tweets and retweets.

For a deeper analysis of specific factors associated with retweets, the second section will include the binary logistic and negative binomial regression results. An interpretation of the signs and significance of the right-side variables in the models will be followed by a comparison of the relative strengths of the modified regression coefficients for each of the significant predictor variables within the two models.

Hashtags

There are 1,272 unique hashtags in the HT-orig dataset, compared to 1,309 unique hashtags in the HT-full dataset. In other words, the 5,485 retweets of HT-full excluded from HT-orig contain only 37 additional hashtags inserted by users as a comment on the original tweet. While this suggests, perhaps unsurprisingly, that original tweets drive messaging, what is not evident by aggregate counts of hashtags is how rare each hashtag is within the datasets. In both the HT-full and HT-orig datasets, the graph of hashtag frequency resembles a power law distribution, with over half of the hashtags in each dataset used only once. This is made clearer by Table 3a, which provides the percentages for the number of times a hashtag is used. That 80.8 percent of hashtags are used four times or less provides a sense of the skewness of the distribution. Chart 3a reveals data associated with the peak of the hashtag frequency distribution by graphing the top ten most used hashtags in the HT-orig dataset.

Table 3a: Percentage of hashtags by number of times used in the HT-orig dataset

Number of Times Used	HT-full	HT-orig
1	52.6%	62.7%
2	15.1%	15.0%
3	8.3%	72.3%
4	4.8%	4.1%

Chart 3a: The top 10 hashtags in the HT-orig dataset

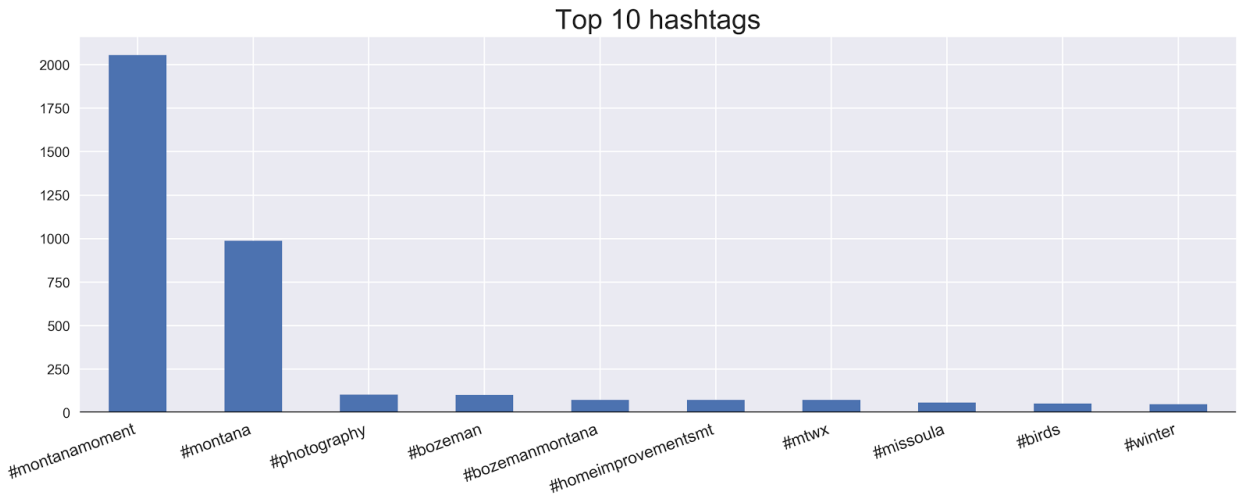
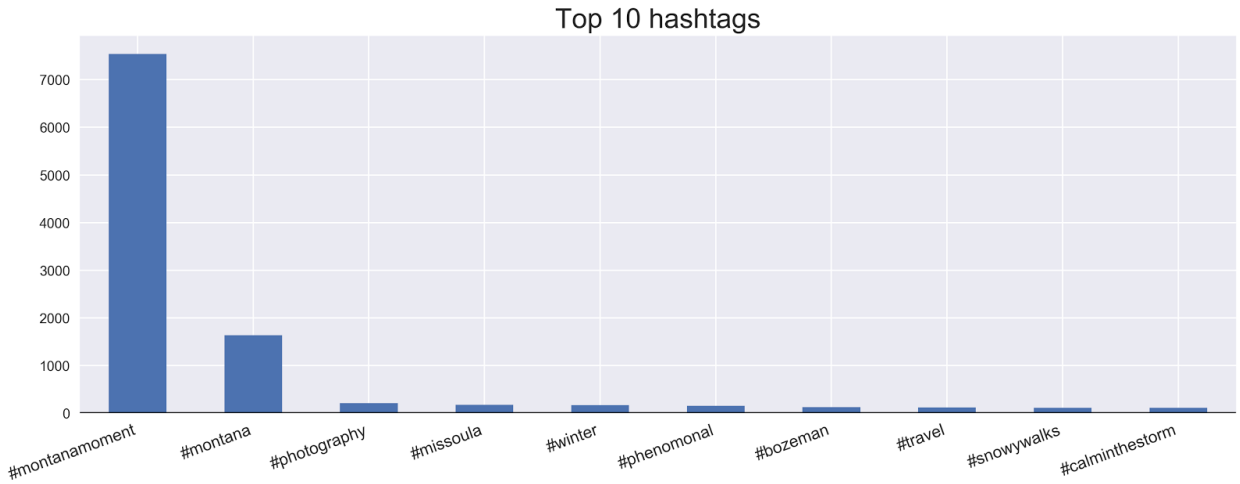


Chart 3a makes it apparent that #MontanaMoment is used in conjunction with #Montana about half of the time, but does not have any other common pairings. The fact that #photography is the third most popular hashtag (and second most popular pairing, as all hashtags in the dataset contain #MontanaMoment) may hint at the frequency with which the many photographers within the *bus_int_ind* category use #MontanaMoment. However, this cannot be certainly determined given the popularity of visual media usage on Twitter and in conjunction with the hashtag of

interest in particular. Chart 3b below displays the top ten hashtags used in the HT-full dataset. While #montana and #photography remain the top two pairs to #MontanaMoment, the symmetric difference of the two sets has an index of eight. In other words, eight total hashtags are unique to either Chart 3a or Chart 3b, suggesting that, while original tweets may determine hashtag usage to an extent, retweets introduce a notable variance.

Chart 3b: Top ten hashtags in the HT-full dataset (tweets & retweets)



Tweets & retweets over time

For each time-based variable used as an axis in the following graphs, the HT datasets were collapsed into the variable of interest. Chart 3c below, for example, displays the number of original tweets that occurred on each day of the collection period. Chart 3d displays the total number of retweets occurring on each of these days. Unlike Chart 2a and 2b used to contextualize the day-of-week variables, the symmetry between original tweets and retweets begins to break down in comparing tweets by day. From this, it is clear that the number of retweets occurring on a given day is impacted by factors beyond the number of original tweets.

A perfect association between original tweets and number of retweets would be suggested if the shapes of the graphs were identical. However, as this study hypothesizes and others suggest, the number of retweets an original tweet receives is a function of more factors than accounted for here. The predictors hypothesized to account for this differentiation are those included as right-side variables in the *logit* and *nbreg* models discussed in the following section. However, it is interesting to note that many of the peaks and troughs between the two graphs are shared, even though many of the peaks in Chart 3d are significantly larger than those of Chart 3c. This would appear to provide some additional support for the assumption that most retweets occur on the same day an original tweet is posted. What is not clear, is why peaks in retweets vary with a high degree of independence from peaks in original tweets.

Chart 3c: Number of original tweets by day posted (n = 1,869)

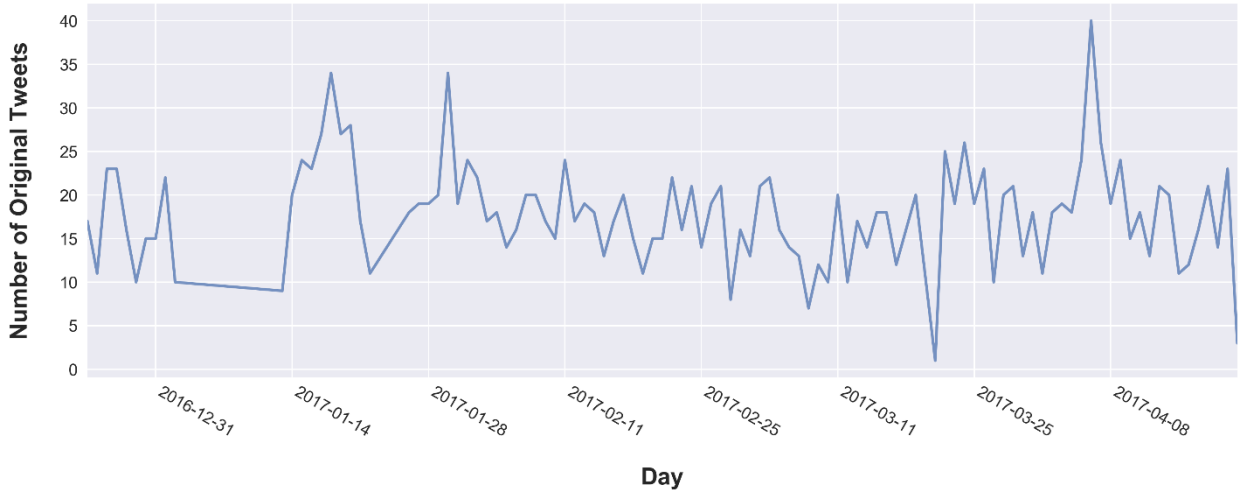
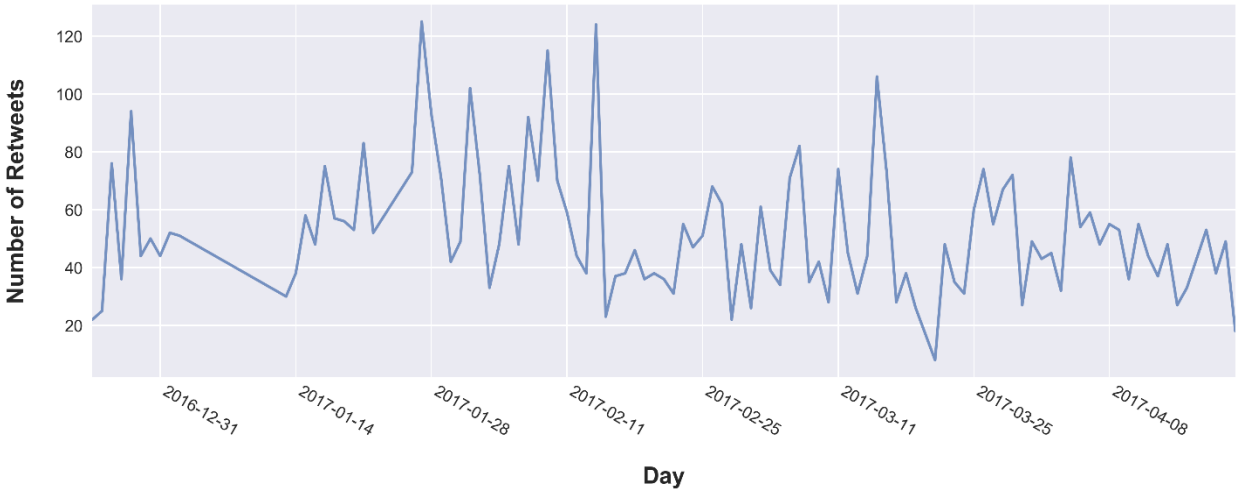


Chart 3d: Number of retweets by day posted (n = 5,482)



Respectively, Charts 3e and 3f provide the total number of original tweets and retweets aggregated by the hour of the day in which they occurred, beginning at 6:00 PM. Considering that previous research (see Barbagallo et al. 2012) has determined that the vast majority of retweets occur within the first hour of a tweet, it seems reasonable to assume that most of the original-tweet-to-retweet relationships that constitute the graphs are non-spurious. That is, within a granted margin of error, most retweets in the 12:00 PM spike, for example, are of tweets occurring in the same day. In other words, any hour-to-hour change between Chart 3e and Chart 3f is expected to be a function of a tweet-to-retweet time interval.

Given this, it seems the largest divergence in the shape of the graphs (i.e., a break in the tweet-to-retweet ratio) occurs between 6:00 AM and 11:00 AM (captured by the *morning* variable). During this period, original tweets balloon with a rapid increase in volume. However, a closer examination reveals that what appears to be a divergence may be more akin to a lag. Over the entire period, both tweets and retweets roughly triple (from approximately 40 to 120 in the

case of original tweets, and 100 to 300 in the case of retweets).

Regardless of the accuracy of the tweet-to-retweet time assumption, the two charts provide useful information regarding user engagement with the hashtag. Night is clearly a period of low activity, whereas 12:00 PM—a typical lunch hour—is a highly active period. User engagement then declines until 6:00PM when retweeters are perhaps settling at home after a day of work. This activity peaks at 10:00 PM, after which users are likely beginning to go to bed, and doesn't pick up again until the morning.

What cannot be discerned from Charts 3e and 3f is whether a tweet is more likely to be retweeted during a given period. Since the ratio of tweet-to-retweet appears relatively proportional, it would seem not. However, this could be the result of a number of phenomena that cannot be further determined. For example, a handful of organizations may ramp up Twitter activity during peak times because they have determined that they receive more retweets and thus drive the trends reflected in the graphs. Alternatively, general activity from all users may increase with tweet-to-retweet proportions remaining unchanged as tweets are equally distributed among users relative to predictor variables. These questions will be further explored in the regression tests discussed below.

Chart 3e: Number of original tweets by hour sent (n = 1,869)

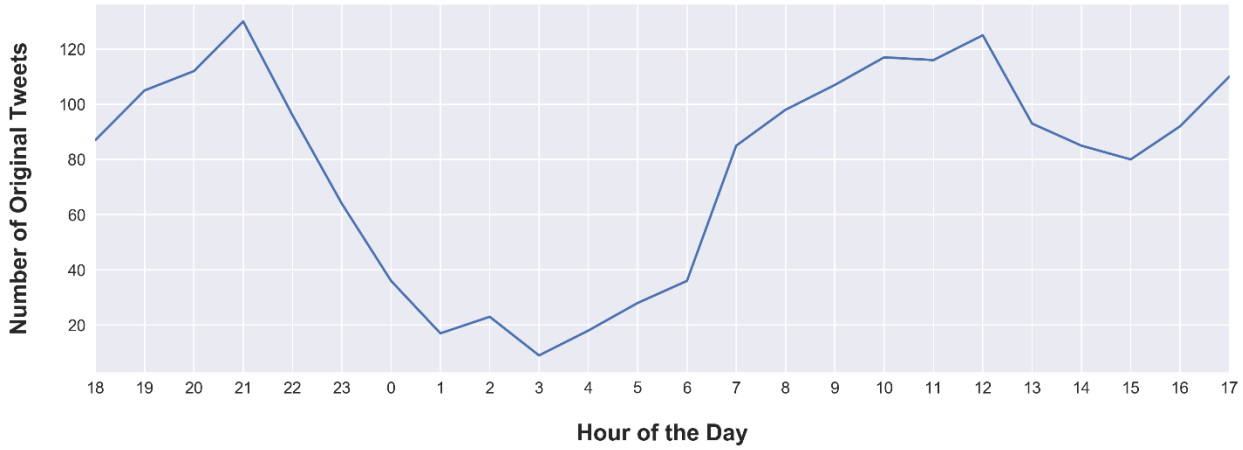


Chart 3f: Number of retweets by hour sent (n = 5,482)

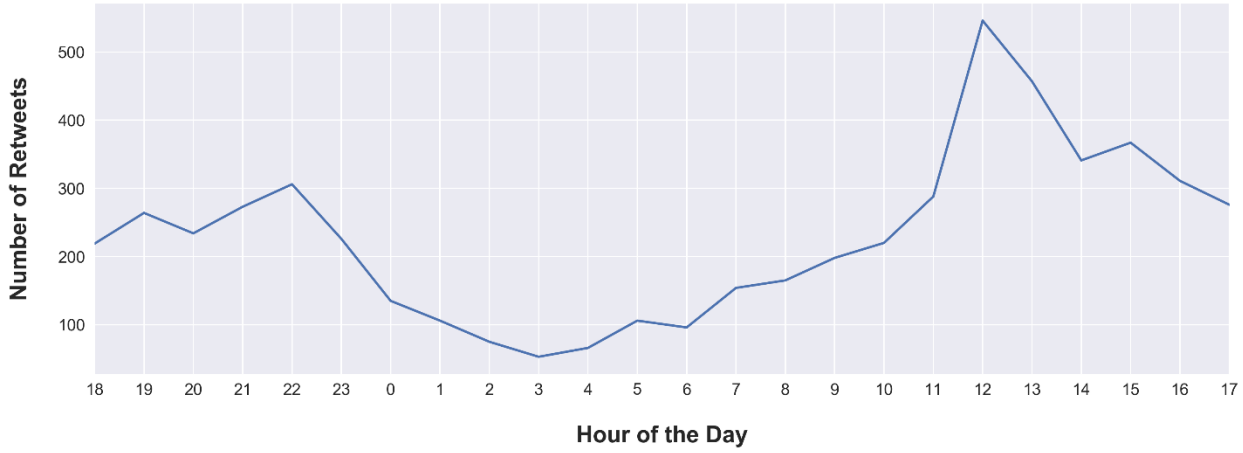


Chart 3g displays the total number of original tweets posted during each month in the dataset, whereas Chart 3h includes only retweets occurring in each month. It is important to recall that only January, February, and March include all the tweets that used #MontanaMoment during the period. Thus, the most obvious difference in month-to-month volume is that both fewer original tweets and retweets occurred in the month of January relative to other months.

The present data does not support an inference as to why January was a slower month for retweets. However, it is possible that harsher weather and reduced outdoor activity played a role, since photographs of the outdoors are a primary use case for the hashtag.

The more interesting aspect of Charts 3g and 3h is the similarity in shape, which suggests that aggregate original tweets influences aggregate retweets to a limited extent. It appears a change in the ratio of original tweets to retweets occurs in the difference between February and March across the two graphs. However, Table 3b reveals that this difference is slight, and that the outlying month is February by a margin of less than 0.2 retweets per tweet from the mean for the three months. The minimal variance in the number of retweets generated per tweet across the three months is perhaps a function of their chronological and seasonal similarities, but it may also suggest the potential to predict total retweets given a number of original tweets with a particular set of shared characteristics.

This is particularly interesting given the ability of the Twitter API to supply data samples approximating the population for research questions narrowed to a group of user accounts, the usage of a particular hashtag, and similarly selective identification strategies. Under such circumstances, summary statistics may be sufficient to provide considerable predictive power. For example, given a large enough timeline, it may be possible to estimate retweet counts based upon the past performance of an organization's original tweets. This could be of potential use in guiding the Twitter activity of organizations seeking to hit certain metrics in terms of user engagement. However, a deeper understanding of this relationship is limited by the current data and is beyond the scope of the present study.

Chart 3g: Original tweets by month (n = 1,869)

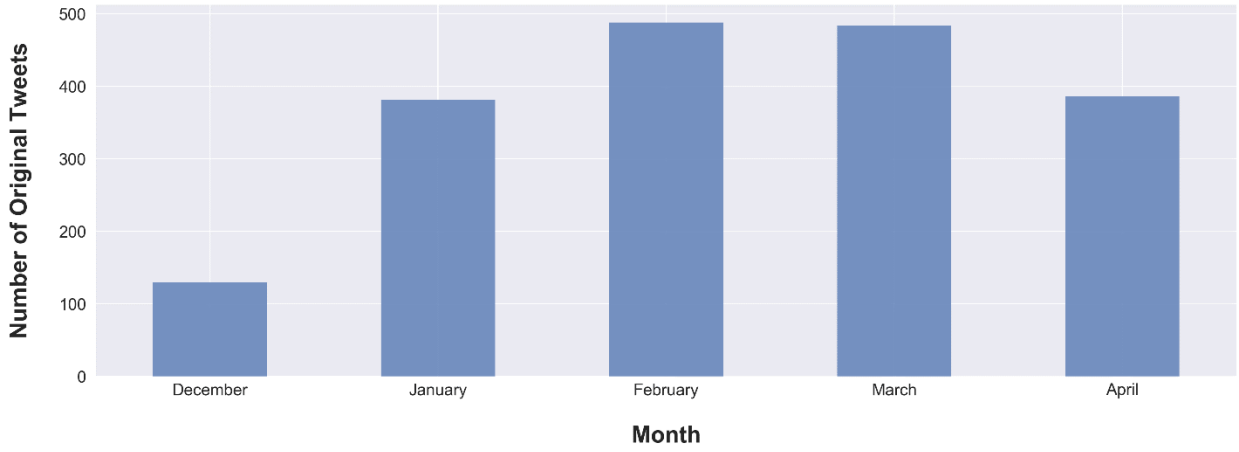


Chart 3h: Retweets by month (n = 5,482)

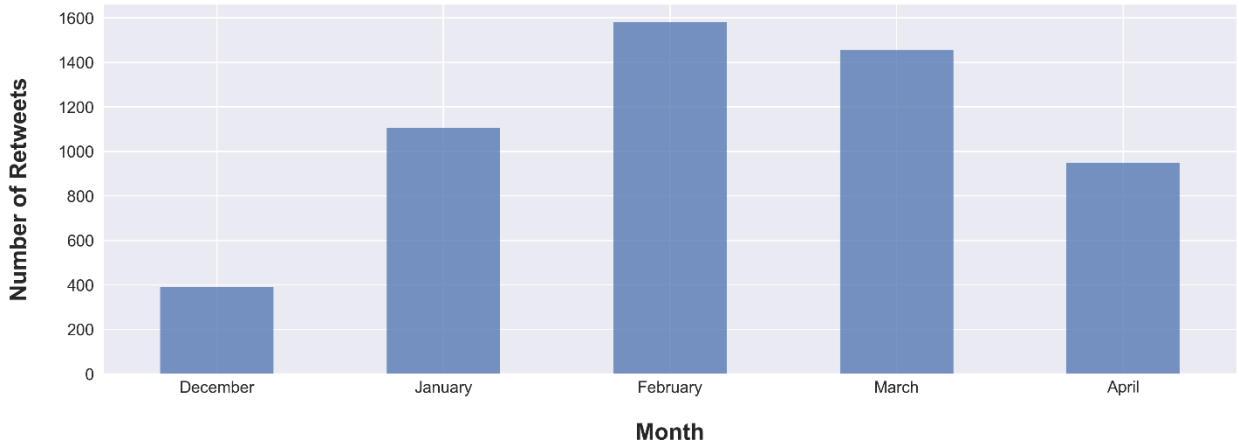


Table 3b: Retweets per tweet by month

Month	Original Tweets	% Total	Retweets	% Total	Total	Retweets per Original
January	381	25.6%	1,106	74.4%	1,487	**2.9**
February	488	23.6%	1,580	76.4%	2,068	**3.2**
March	484	24.9%	1,457	75.1%	1,941	**3.0**

Regression results

Table 3c below displays the results of regressing the outcomes of the *retweet_dummy* variable (column I: logit) and the *retweet_count* variable (column II: nbreg) on each of the 19 independent variables. The models are intended to estimate the effects of account type upon whether or not a tweet is retweeted (logit) and upon how many times it is retweeted (nbreg). The control variables shared between the models include measures for number of followers, membership in public lists, number of statuses, whether a tweet contains a URL or direct mention (@user), on which day of week a variable was sent, and the period of the day in which a tweet was sent

Table 3c: Reg. results for *retweet_dummy* (I) and *retweet_count* (II) outcome variables

Variable	**I. logit** Coefficient (Robust St. Error)	**II. nbreg** Coefficient (Robust St. Error)
state_account	-15.18*** (1.18)	-1.31*** (0.24)
bus_int_ind	-16.09*** (1.17)	-2.08*** (0.27)
other_ind	-16.83*** (1.18)	-3.48*** (0.27)
bus_orgs_unofficial	-16.19*** (1.16)	-2.55*** (0.26)
kauffman	-12.99*** (1.18)	-0.30 (0.21)
followers_count	0.85* (0.35)	0.55*** (0.08)
listed_count	-0.43 (0.27)	-0.52** (0.17)
statuses_count	0.02 (0.14)	0.48 (0.29)
url_dummy	-0.47* (0.20)	-0.26 (0.18)
mention_dummy	0.51 (0.26)	0.62* (0.31)
monday	0.00 (0.24)	-0.08 (0.21)
tuesday	0.31 (0.23)	0.20 (0.26)
wednesday	0.22 (0.23)	0.03 (0.21)
thursday	0.12 (0.21)	-0.14 (0.20)
friday	0.45 (0.30)	0.03 (0.21)
saturday	-0.11 (0.21)	-0.04 (0.19)
night	0.37 (0.45)	0.01 (0.28)
day	0.07 (0.20)	0.31 (0.17)
evening	0.32 (0.21)	0.42 (0.22)
constant	15.23*** (1.17)	2.22*** (0.27)
N	**1,869**	**1,869**

$\frac{Coeff.}{(S.E.)}$ *p < .05, ** p < .01, *** p < .001, for a two-tailed test [+]account-type reference category is *mtot*

Panel I of Table 3c displays the regression coefficients for the logit model. These can be interpreted as the expected change in the natural log of the odds for a one unit change in the respective predictor variable, holding all other variables constant. Panel II of Table 3c displays the results of the negative binomial regression, which models the natural log of expected retweets as a function of the predictor variables. The coefficients in Panel II can be interpreted as the expected change in the difference of the natural log of retweets for a one unit change in each of the predictor variables, holding all else constant. In both panels, standard errors are displayed in parentheses below the regression coefficients.

Looking at the account-type variables, it is clear that account type matters both in terms of how many times a tweet is retweeted, and whether it is retweeted at all. In both the logit and nbreg models, each of the account-type variables is highly significant and negative in relation the *mtot* reference category. That is, an account classified as any other than *mtot* is expected to receive fewer retweets, and is less likely to be retweeted at all, holding all other variables constant. These findings are understandable considering the shape of the data associated with the *mtot* variable (see Table 2b in the variables section). That is, in terms of total retweets, *mtot* was a clear winner with 46.7 percent of all retweets, a full 30.3 percent above the next highest category *bus_int_ind.*

The *kauffman* variable, though classified as an outlier for purposes of clarity, is also compared against the *mtot* reference variable and is the only variable of this series to break the form. Although a highly significant negative association with likelihood to be retweeted is shared between the *kauffman* variable and the other account-type variables, the *kauffman* variable is not

significant in the nbreg model. That is, while tweets sent by the *kauffman* account are less likely to be retweeted (logit) than those sent from the *mtot* account, there is not enough evidence to conclude that Kauffman's tweets were retweeted more or less than MTOT's tweets, all other model variables constant.

Within the user-object category of variables, the number of followers an account has is significant in both the logit and nbreg models. This is not surprising, as the number of followers an account has is among the retweet predictor variables that remain relatively constant across studies (e.g. Suh et al. 2010; Yasugi et al. 2013). The more followers an account has, the more likely a tweet will be seen and retweeted.

What is more interesting is that the *listed_count* variable is negatively associated with the number of times a tweet is retweeted. In other words, the more public lists to which a Twitter account has been registered by followers of the account, the fewer retweets a tweet is likely to receive when holding all other variables constant. Although it was expected the impact of *listed_count* would be minimal when controlling for the number of followers, the hypothesized relationship was positive as *listed_count* was expected to capture a measure of popularity of an account. While the present findings partially contradict those of Petrovic, Osborne, and Lavrenko (2011)[17], there are studies with findings more aligned with those of the present study (e.g. Yasugi et al. 2013). The negative association between membership in lists and retweets in the present

[17] Petrovic, Osborne, and Lavrenko (2011) found *listed_count* to be a moderate predictor of whether or not a tweet was retweeted (binary), and not significant for the number of times a tweet was retweeted (count). As *listed_count* is insignificant in the logit model, it is possible there is some additional nuance involved in the relationship between retweets and membership in public lists. Currently, this remains one of the less examined areas of Twitter-based research. Referencing this, Petrovic, Osborne, and Lavrenko (2011) note that more research is necessary to understand the role upon retweets of membership in public lists.

study is expected to be capturing an interested-but-not-that-interested measure of the followers of the listed accounts. This is because when a Twitter user assigns an account to a list, typically curated by category of account, he or she receives tweets from the list as a whole on a regulated basis. This reduces the potential retweeter's exposure to any one account in the list. Thus, the negative association of *listed_count* with number of retweets is hypothesized to be the result of a more passive consumption of information distributed by listed accounts than those whose tweets are received directly by followers.

This is similar to conclusions made by Yasugi et al. (2013) based on findings that accounts assigned to lists were those that also sent many tweets. They reasoned that assignment to a list was a way for users to continue following accounts of interest without being inundated by their tweets. In this sense, public lists might function as a sort of sanction imposed by a user upon an account, although this inquiry is best left to future research. The peculiarities of the *listed_count* variable also broach a much larger issue concerning the difficulties of accounting for predictors of retweets yet to be examined in detail due to problems of measurement. This is discussed in more depth in the limitations of present findings subsection below.

Another interesting, though not particularly surprising result is that the number of statuses an account has (total tweets) has no significant impact on retweets in either model. Hasty logic might reason that the more tweets an account sends, the more likely some of them are to be retweeted. However, it is more likely that a variety of users are highly active and that an active account with no followers cannot compete with an active account with many followers. Thus, when controlling for additional explanatory factors, the level of activity of an account is rendered insignificant.

Future research, however, might test this assumption by running tests on an account-level rather than tweet-level dataset to examine whether or not aggregate tweets influence aggregate retweets when controlling for some of the user-objects included in the present models. Similar to the suggestion for the public lists variable, such could provide insights for broader organizational strategies over time. It might also be useful to test whether the number of statuses an account has relative to its age is associated with membership in public lists. An alternative hypothesis would be that accounts with many statuses are more likely to be sanctioned by users by being added to a public list in the manner discussed above.

In the tweet-object category, *url_dummy* is not significant in the nbreg model, but has a significant negative effect on whether or not a tweet is retweeted. This finding diverges from other studies that have found the inclusion of a URL to positively impact retweets (Bhattacharya, Srinivasan, and Polgreen 2014; Suh et al. 2010). However, this is expected to be a difference in sampling. The very active and highly successful account @visitMontana (*mtot*), for example, includes a URL in only 8.3 percent of its tweets. The rarity of URLs within retweeted tweets is not isolated to one account either. For all tweets containing a URL, only 19.9 percent are retweeted. In contrast, out of all tweets that do not contain a URL, 46.9 percent of tweets are retweeted. These figures are expected to be a result of the dominance of visual media usage with #montanaMoment. Many of the samples used in other studies focus on topics that are more likely to reference external links to research, organizational posts, or other news events. With the hashtag of interest to this study, the focus on the outdoor beauty of Montana is particularly well-suited to the inclusion of a photograph rather than an external link.

Although not significant in the logit model, tweets that contained a mention were more

likely to be retweeted, similar to past findings (Bhattacharya, Srinivasan, and Polgreen 2014). By directly addressing one or more users through a mention, it is expected that the greater probability of receiving a retweet is a function of the targeted attention the tweet is likely to receive by the mentioned user or users. In this case, it is not surprising that the variable is insignificant in the nbreg model as the targeted effect would be isolated to mentioned users.

None of the remaining tweet-objects—the day-of-week controls for time-to-be-retweeted and time-of-day variables—were significant in either model. However, in the nbreg model, while the day-of-week variables are far from statistically significant[18], the *day* and *evening* variables come close with respective p-values of 0.078 and 0.056. In other words, there is less than an eight percent chance of obtaining the present figures or larger figures for these variables if they did not impact number of retweets when compared against the *morning* reference category.

Percentage of tweets retweeted appears to provide some additional support for the notion that tweets may be more successful in the day and evening compared against the morning. Of all tweets sent during the day and night, 32.8 and 35 percent are retweeted, respectively. This compares to only 26 percent of all morning tweets and 22.8 percent of night tweets that are retweeted. Thus, while it would be an error to infer causality under such circumstances, it is clear that time-of-day is a relevant variable for such studies, as might be expected after reviewing the tweet and retweet volume displayed in Charts 3g and 3h.

[18] Measured against the Sunday reference category, p-values for the day-of-week variables range from 0.436 to 0.900.

Predicted probabilities

Table 3d below displays the predicted probabilities for changes in each of the account-type coefficients of the logit model and the changes in the incident rate ratios for significant variables in the nbreg model. The transformation of the logit coefficients was achieved through manual calculation of the predicted probabilities while holding all other model variables at 0. Since the *mtot* reference category predicts retweets perfectly, the probability that a tweet will be retweeted when sent from each remaining account type is subtracted from the probability an MTOT tweet will be retweeted. Accordingly, values in the logit panel of Table 3d can be interpreted as the degree to which the predicted probability of a tweet being retweeted is lower than a tweet sent from MTOT.

The incident rate ratio (IRR) coefficients of the nbreg model may be interpreted as the estimated rate ratio for either a one unit increase in the variable of interest (continuous variables), or between a value of 1 or 0 (dichotomous variables), holding all other variables at their means. Since the response variable (*retweet_count*) is technically a rate defined as the number of retweets per collection period[19], the original regression coefficients can be interpreted as the log of the rate ratio (UCLA 2017). From this, incident-rate ratios[20] are derived from the regression coefficients. IRR coefficients below 1 represent the factor by which overall retweets are estimated to decrease, and those above 1 represent the factor by which a tweet's total retweets would be expected to increase, holding all other variables at their means.

[19] Due to difficulties in ensuring a collection period of the same duration for each tweet in the dataset and the lack of a precise exposure variable, it cannot be argued with certainty that these results are unbiased by variation in the time-to-be-retweeted between cases. However, as presented at length in the variables section, existing data provide strong support for the assumption that most retweets occur shortly after a tweet is retweeted. As such, any biases in the IRR estimates for the nbreg model are expected to be minimal.

[20] That is e^{β_i} rather than β_i (Stata 2017)

Table 3d: Change in predicted probability (logit) and incident rate ratio (nbreg) for significant variables in each model

Variable	logit[+]	nbreg
state_account	0.49	0.27
bus_int_ind	0.70	0.12
other_ind	0.83	0.08
bus_orgs_unofficial	0.72	0.95
kauffman	0.10	-
url_dummy	-	-
mention_dummy[++]	-	1.86
followers_count*	-	1.73
listed_count*	-	0.59

[+] Logit is predicted pr. relative to MTOT, holding all other variables at 0

[++] Missing values indicate a p-value > 0.05

* Based on a 1-unit change, which represents a standard deviation

Since values in the logit panel are relative to *mtot*, lower numbers represent a higher probability of being retweeted. Thus, the *kauffman* outlier is the most successful account relative to *mtot,* with only 10 percent less chance of being retweeted. Other state accounts are a distant second with only 49 percent less chance to be retweeted. Both individuals with a business interest and business/organization accounts have about 70 percent less chance, ahead of all other individuals, who are 83 percent less likely to be retweeted than MTOT. These findings make sense as 92 percent of *kauffman* tweets, 56 percent of *state_account*, 30 percent of *bus_int_ind,* 29 percent of *bus_orgs_unofficial*, and 18 percent of *other_ind* tweets were retweeted. However, as the primary interest is to estimate factors associated with higher levels of engagement rather than the threshold between engagement and non-engagement, the logit model is intended only to supplement the nbreg model.

Within the nbreg model, it is interesting to note the relative strength a followers count one standard deviation above the mean and the inclusion of a mention have on how many retweets a tweet receives. Holding all other variables at their means, tweets with a mention are expected to receive 86 percent more retweets than those that do not, while more followers result in an estimated 73 percent more retweets. Conversely, tweets sent from accounts associated with a public list are expected to receive 41 percent fewer tweets than those that are not. As discussed above, this is likely due to the suppressed presence of a tweet (sent from a listed account) within a user's newsfeed. Similarly, all tweets sent from a non-*mtot* account are estimated to receive fewer retweets than one from *mtot*.

As none of the account-type variables within the nbreg model are greater than 1, it is clear that the omitted category of MTOT tweets is the most successful in terms of number of retweets. On the far end of the spectrum, a tweet from an average, individual Twitter user (*other_ind*) is expected to receive a retweet count that is only 8 percent of what a tweet from MTOT would be expected to receive when holding all other model variables constant. Contrasted with individuals, businesses and organization not associated with the state (*bus_orgs_unofficial*) do well in relation to MTOT. A tweet sent from one of these accounts is expected to receive only 5 percent fewer retweets than a tweet from MTOT, all else equal.

In terms of retweet counts the remaining two account types (*state_account* and *bus_int_ind*) align more closely with average individuals than the businesses and organizations category. However, non-MTOT state accounts are expected to perform much better than individuals with a business interest. While these state accounts are expected to receive 27 percent of the retweets an MTOT tweet is expected to receive, a tweet sent from an individual with a

business interest is expected to receive only 12 percent, all other model variables constant. However, individual Twitter users with a clear business interest are expected to perform better than average individuals, whose tweets are expected to receive only 12 percent of the retweets MTOT tweets receive, holding all other model variables constant.

In other words, retweet counts appear to chart upward on a scale of official hierarchy. An average user may be considered the least official, followed closely by individuals representing themselves as business owners. The entrepreneurs are succeeded by official state accounts, which sit a fair degree below businesses and organizations not associated with the state. At the top of the official hierarchy stands MTOT, the creator and core promoter of the hashtag.

Though the results seem to suggest account status matters in determining retweet counts, a core difficulty lies in determining the degree to which this is driven by quality of content, and not a metric omitted or accounted for only partially. Given the theoretical prominence of number of followers as a predictor of retweets and its account-level association in the present dataset, *followers_count* is one such variable that requires closer examination. Although the user-object variables such as *followers_count* may change between cases in the dataset, they are unlikely to vary to a high degree.[21] Thus, it is possible a particularly high number of followers could contribute to the strength of the *mtot* category. To address this, Table 3e below displays summary statistics for the ten accounts with the most followers.

[21] As explained in the variables section, this is because each tweet in the dataset is relevant to the user-object values at the time it was downloaded. For example, a change in the *followers_count* variable between a tweet downloaded in December versus one sent from the same account in January represents the change in the number of followers over the one-month period.

Table 3e: Retweets per 100k followers for the ten accounts with the most followers

Account	Total Retweets	Total Tweets	Total Retweets per Tweet	Total Followers[+]	Retweets per 100k Followers[++]
earthXplorer	44	16	3	**183,103**	2
GlacierNPS	487	6	**81**	181,840	**45**
MalloryOnTravel	1	1	1	101,157	1
StephanieQuayle	23	1	23	81,327	28
LuxuryTravel77	5	2	3	71,776	3
visitmontana	**2,433**	**108**	23	57,882	39
ManTripping	1	4	0	57,166	0
robertserian	0	2	0	54,846	0
ECAatState	2	1	2	54,770	4
StormHour	18	1	18	47,673	38

[+] included in the models as a standardized variable

[++] ((total retweets / total tweets) / total followers) * 100,000

Since MTOT's ('visitmontana') retweets per 100,000 followers is either higher or within a similar range of accounts with more followers, it is clear that the success of MTOT is due to more factors than a high number of followers[22]. For its 16 cases in the dataset, the account with the most followers (@earthXplorer) averages only two retweets for every 100,000 followers. On the other hand, @StormHour—an account with followers only 26 percent of @earthXplorer—has a much higher average retweet count. While it is important to note the wide variance in total

[22] It is important to note that these results are based on a linear interpretation of followers count. That is, the addition of one additional follower is expected to mean the same for an account with 10 followers as it does for an account with 10,000 followers. While this is often the norm for studies measuring retweets, the comparative interpretation of the account-type variables is dependent upon the linear coding of the *followers_count* variable.

tweets (i.e. cases in the dataset), it does seem apparent that followers are only one aspect of retweets.

Much of the variance in retweet-defined account success is likely determined by the quality of content, which aligns with the findings of Rossi and Magnani (2012). Recalling the hierarchy of officiality heuristic used to interpret the relative success of account categories, it is reasonable to expect certain account types to output higher quality tweets more consistently. MTOT is the steward of the hashtag, and businesses and organizations depend upon successful marketing for much of their success. Other state accounts likely have processes for ensuring consistent quality of content, but perhaps have less motivation than the two former categories. On average, individuals with a business interest are likely to have less sophisticated campaigns than larger businesses, but more motivation to drive engagement than average individuals.

For small to medium state tourist organizations formulating a campaign to drive user engagement, the present findings suggest that partnership with private sector businesses and organizations may be more successful than encouraging tourists to create and share their own content. This is not to say that the frequently employed attempts to generate user content are not worthwhile. It is entirely possible that an *other_ind* category within a sample capturing information flows related to more popular destinations might be a top performing account. This would likely be due to a threshold drawn at the point where the influence of aggregate users (*other_ind*) overtakes the reach of account groupings more limited in number. Such would have to be determined on grounds of both summary statistics (total retweets) and the relative strength of account-type categories. This is because the relative celebrity status of official accounts (i.e. those associated with a business or organization of any kind) will likely outperform a single

individual on average. However, when the total reach of individuals overtakes those of the state accounts, an organization might consider shifting its strategy to focus more on user-generated content.

However, given the circumstances of the present findings, the data suggest that an increase in tweets sent from the MTOT account is the most promising way to increase user engagement with the hashtag. One caveat to this is that tweets may reach a saturation point at which additional tweets either do not increase user engagement, or potentially reduce it. Reduction in engagement might result from incurring a user-initiated sanction such as inclusion in a public list as discussed above, or a loss of followers. Another important note is that by only tweeting more frequently, an organization limits itself to its own network and the networks of the followers who retweet. Thus, significant user engagement may occur within a relative echo chamber if an organization does not encourage the broader use of a hashtag. The verification or refutation of such assumptions as these is a promising direction for future research.

Conclusions

Regarding RQ1—how and why are #MontanaMoment tweets retweeted?—the present study first examined overall activity of the hashtag through the relationship between original tweets and retweets. This analysis suggested that original tweets largely determine message visibility in terms of hashtag characteristics. On average, there were 2.5 retweets to every tweet, and only 37 new hashtags introduced by the 5,485 retweets analyzed in the study. Moreover, the majority (80.8%) of hashtags are used fewer than four times. This line of inquiry also determined

that the volume of retweets (as well as original tweets) peaked at noon and 10:00 PM, with considerable activity during typical work hours, and very little activity between midnight and 6:00 AM.

In the logit and count models, respectively, the inclusion of a URL and the use of a mention were estimated to be positively associated with retweets, reflecting the findings of Bhattacharya, Srinivasan, and Polgreen (2014) and Suh et al. 2010. The negative association between membership in public lists and retweets was also significant in the count model. This negative association contradicts some of the limited research estimating the relationship between membership in public lists and retweet count (e.g. Petrovic, Osborne, and Lavrenko 2011; Yasugi et al. 2013), and is discussed in more detail below. Lastly, the number of followers, sometimes referred to as in-degree, was positively associated with retweets in both models, similar to the findings of nearly all past studies (Hong, Dan, and Davison 2011; Petrovic, Osborne, and Lavrenko 2011; Suh et al. 2010; Yasugi et al. 2013).

Regarding RQ2—what effect does type of account have upon retweets?—the results of both regression models demonstrate that MTOT performed significantly better than other accounts in terms of retweets. While a partial explanation for this likely stems from the shape of the data—each tweet from MTOT (46.7 percent of all tweets) was retweeted—it also aligns with expectations. As an authority on the topic #MontanaMoment, MTOT was consistent and focused in its use of the hashtag. This partial explanation of MTOT's success aligns with the findings of Cha et al. (2010) as discussed previously. Also relevant are the findings of Rossi and Magnani (2012) that, while official accounts ranked among the top in number of retweets, 'regular' users ranked near or above the official accounts in the study in terms of retweets. From this they

concluded, as did Cha et al. (2010), that the quality of content is primary in determining retweets. The fact that a relatively unestablished photographer in the present dataset captured the third highest number of total retweets among account types (and second only to MTOT in terms of discrete accounts) lends additional credence to the notion that content quality is the primary driver of retweets.

These indications suggest that MTOT creates better content relative to other accounts using the hashtag, and is therefore best poised to influence the eWOM regarding Montana's reputational image as disseminated via Twitter. Although it may have an additional advantage as the official representative of the hashtag, as Rossi and Magnani's (2012) findings would suggest, the sum of its content quality, top quartile followers count, and account status determine its efficacy in engaging users through retweeting.

Relative to MTOT, businesses and organizations not associated with the state performed better than all other accounts. Non-MTOT state accounts were a distant second to businesses, relative to MTOT. Individual users performed the worst in terms of estimated retweets, with individuals with a business interest performing only slightly better than 'regular' individuals.

This suggests that if MTOT were to attempt to expand its influence by partnering with other Twitter users, it would gain the most traction in targeting businesses and organizations. This might appeal to MTOT and other DMOs since the control over image formation agents would likely be easier than in a 'regular' user-generated campaign as Govers (2015) has suggested. Considering that original tweets appear to drive retweets, with little modification of the original tweet, message control becomes easier when tweets are sent from more centralized accounts such as partnering businesses. However, as the brand authority of its own destination

marketing campaigns, MTOT is estimated to be the most capable of driving further engagement. Thus, contributing the most effort to consistent representation of a topic and creating quality content is likely the most effective way for MTOT to maintain a hashtag campaign.

What cannot be stated from the analysis of the present study is the degree to which the influence of MTOT is strengthened by a hashtag campaign. It is likely the case that the increased publicity, production of quality content, and capture of new followers from retweet exposure contribute significantly to the strength of MTOT's social media presence. Simply by generating a higher aggregate retweet count, MTOT may enhance Montana's reputational image through the greater distribution and retweet-based validation. Place-based image formation in this context would be a co-creation of the DMO and the social agents who lend it credence through retweets.

Limitations of present findings

Among the primary limitations of the present study involve those that arise from a limited sample. Most of these have been discussed as they have become relevant to the study, although it is worth briefly summarizing them here. Among the largest sampling concerns is the possibility of a seasonal bias. Because it was not possible to gather more than a few months of data for the present study, fluctuations in tourism and in general outdoor activity might have artificially suppressed some user-generated content. Another time-based concern involves the time-to-be-retweeted phenomena discussed at some length in the Research Design section. While considerable care has been taken to alleviate the potential of resultant biases, future studies accessing the Twitter API in a similar manner should take care to create a complete timestamp for both the moment a tweet was created and the moment it is inserted into the dataset.

Additional sampling concerns raised in the above section involve the overall generalizability of the present study. The number of ways to analyze Twitter data is vast and the field is new. This results in a diverse range of studies, samples, and methodological approaches. With the theoretical landscape still tenuous, it is difficult to determine how broadly the findings of any one study should be applied. In many cases, the sample has a large bearing on results. For the present study, this issue presents itself in the decision to generalize to small- to medium-sized destination marketing organizations. It is not argued that findings related to account type will hold when examining highly popular tourism destinations. In such cases, for example, it is expected that individual Twitter users will have more impact as their relative numbers will be larger. However, for state tourism agencies similar to MTOT, the present findings are expected to make a reasonable contribution to organizational strategy.

Other limitations relate more closely to larger methodological concerns with Twitter analyses generally. For example, the questions raised by the negative association between *listed_count* and number of retweets pointed to larger concerns with the way the uniqueness of a user's newsfeed impacts whether or not a tweet is retweeted. That is, variation in user-exposure is a phenomenon difficult both to identify and to control.

Similar to Facebook, Twitter uses an algorithm-based newsfeed, which it adopted in 2015. The stated purpose for doing so was to provide for a better presentation of the activity that had occurred while a user was away by making prominent tweets selected on grounds such as user-engagement (Twitter 2015). Thus, the unknown parameters guiding the newsfeed algorithm determine much of the variation in how a user experiences information flow on Twitter. Two identical users, for example, could view entirely different information given the single deviation

of variation in frequency of Twitter usage. However, the difficulty in controlling for environmental factors is one of the defining characteristics of the social sciences. The concern with Twitter data is that such environmental factors are not random events but complex formulas designed to make a user's experience unique.

The most immediate concern to the present study is whether tweets that receive a certain number of retweets are more likely to be seen than those that do not reach a certain threshold. It is likely that this would be influenced to a significant extent by the relative number of retweets (and likes, and favorites) received by the tweets sent by all other followed accounts since a user's last login. However, any attempt to control for such factors would be an effort of reverse engineering.

Another issue related to user exposure includes the variation in methods of receiving information on Twitter. Notification preference is perhaps the most significant area of divergence for users, and subsumes similar disparities in regard to how users access Twitter. Smart phones, for example, introduce the possibility of push notifications or instant updates through a connected email account. The increasing complexity of web applications and similar phenomena shaping internet usage habits will likely increase the difficulty of accounting for elements of user-experience. Potential approaches for dealing with a predestination bias of certain tweets are in the Suggestions for Future Research subsection below.

Finally, the single largest limitation of the present study is the lack of controlling for content of tweets. The issue introduces a cognitive dissonance of sorts. For many, the strength of Twitter as a platform for study emanates from the sheer volume of data. However, without further advancement in natural language processing and similar A.I. tools, content analysis

becomes a stopgap that either reduces sample size to allow for manual coding or introduces uncertainty into the studies that omit it. While a significant branch of Twitter research focuses on automated content analysis, this has not evolved far beyond positive, negative, or neutral sentiment analyses.

While natural language processing poses a difficult though partially addressable problem, the processing of visual content is far more challenging. Given the usage of #MontanaMoment, a visual quality rating is likely the largest gap in the right-side regression equations of the present study. It may be the case that retweets in the present study are determined to a high degree by the 'quality' of a photo tweeted with the hashtag. This is likely why certain photographers (such as the account associated with the *kauffman* variable) have attained such an elevated status within the dataset. However, despite the technical nature of the current iteration, the fundamental challenge of categorizing taste is familiar and by definition resistant to classification. Some suggestions for how future research might approach the issue are offered below.

Suggestions for future research

Although not a variable of great concern the present study, the finding that membership in public lists is negatively associated with retweets raises interesting questions for future research. Hypothesizing that assignment to a list may be a user-enforced sanctioning of accounts that send too many tweets, an interesting inquiry might attempt to discern this saturation threshold. Such an inquiry might take the form of an organization-level analysis examining when an additional tweet begins to negatively impact aggregate retweet count for an organization over a period of time. This might provide organizations with information to guide their chosen level

of aggressiveness on Twitter at a given point in time.

As discussed in the limitations of present findings subsection of the findings section, the difficulty of controlling for user-exposure is among some of the methodological challenges of analyzing Twitter data. Under question was how important promotion of popular tweets by the newsfeed algorithm employed by Twitter is to the number of retweets a tweet receives. One interesting approach might be to include a series of binary threshold variables for the number of retweets a tweet receives (see Hong, Dan, and Davison (2011) to view a similar approach). By comparing the relative strength of the coefficients, it might be possible to identify some non-regular patterns in the relative strengths of coefficients. However, the largest challenge in this area involves isolating the popular-tweet-promotion effect from characteristics such as content or sending account that make a tweet popular before a viral effect carries it further.

While many public research barriers have yet to be crossed in user-engagement studies of social media, significant resources are being allocated to a greater understanding. As the theoretical and methodological sophistication of Twitter analyses grows, it will become easier to approach such studies with greater confidence. Motivation for doing so is also enhanced by the volume of data, ease of collection, and entwinement with commercial interests. Although destination marketing lags the curve of understanding, the actionable and increasingly measurable connection between place-based image formation and social media enhanced word-of-mouth is sure to drive further efforts.

References

Akehurst, Gary. 2008. "User Generated Content: The Use of Blogs for Tourism Organisations and Tourism Consumers." *Service Business* 3 (1): 51.

Allison, Paul. 2012. "Do we really Need Zero-Inflated Models?" *Statistical Horizons*.ast modified August 7, 2012. https://statisticalhorizons.com/zero-inflated-models. (Accessed May 15, 2017).

Anttiroiko, Ari-Veikko. 2014. *The Political Economy of City Branding*. Routledge Advances in Regional Economics, Science and Policy (Book 2). New York: Routledge.

AP. 2017. "Montana Tourism Campaign Targets Visitors from Nearby States." *U.S. News*, April 21, 2017. https://www.usnews.com/news/best-states/montana/articles/2017-04-21/montana-summer-tourism-campaign-focused-on-nearby-states (Accessed July 16, 2017).

Barbagallo Donato, Leonardo Bruni, Chiara Francalanci, and Paulo Giacomazzi. 2012. "An Empirical Study on the Relationship between Twitter Sentiment and Influence in the Tourism Domain". In: Fuchs M., Ricci F., Cantoni L. (eds) *Information and Communication Technologies in Tourism*. 2012. Springer, Vienna. DOI: 10.1007/978-3-7091-1142-0_44.

Batrinca, Bogdan and Philip C. Treleaven. 2015. "Social Media Analytics: A Survey of Techniques, Tools and Platforms." *Ai & Society* 30 (1): 89-116.

Bekk, Magdalena, Matthias Spörrle, and Joachim Kruse. 2016. "The Benefits of Similarity between Tourist and Destination Personality." *Journal of Travel Research* 55 (8): 1008-1021.

Bhattacharya, Sanmitra, Padmini Srinivasan, and Phil Polgreen. 2014. "Engagement with Health Agencies on Twitter." *PLoS ONE*, 9(11), e112235. http://doi.org/10.1371/journal.pone.0112235

Boyd, Danah, Scott Golder, and Gilad Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." HICSS-43. IEEE: Kauai, HI, January 6.

Bruni, Leonardo and Chiara Francalanci. 2012. "An Empirical Study on the Relationship between Twitter Sentiment and Influence in the Tourism Domain." DOI:10.1007/978-3-7091-1142-0_44. https://link.springer.com/chapter/10.1007%2F978-3-7091-1142-0_44 (Accessed May 1, 2017).

Cameron, A. Colin and Pravin K. Trivedi. 2013. *Regression Analysis of Count Data*. 2nd edition, Econometric Society Monograph No.53, Cambridge University Press, 1998 (566 pages.)

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy". *Fourth International Conference on Weblogs and Social Media*. ICWSM 2010. Washington, DC, USA, May 23-26, 2010.

Deskins, John and Matthew T. Seevers. 2011. "Are State Expenditures to Promote Tourism Effective?" *Journal of Travel Research* 50 (2): 154-170.

Dhar Ravi and Klaus Wertenbroch K. 2000. "Consumer Choice between Hedonic and Utilitarian Goods". *Journal of Market Research*. 37 (1):60–71.

Ekinci, Yuksel and Sameer Hosany. 2006. "Destination Personality: An Application of Brand Personality to Tourism Destinations." *Journal of Travel Research* 45 (2): 127-139.

Gartner, William. 1993. "Image Formation Process". *Journal of Travel & Tourism Marketing*. 2(2/3):191–215. doi: 10.1300/J073v02n02_12.

Govers, Robert. 2015. "Rethinking Virtual and Online Place Branding." In *Rethinking Place Branding: Comprehensive Brand Development for Cities and Regions*, edited by Mihalis Kavaratzis, Gary Warnaby and Gregory Ashworth, 73-84: Springer.

Govers, Robert and Frank Go. 2009. *Place Branding: Global, Virtual and Physical, Identities Constructed, Imagined and Experienced*. Palgrave Macmillan, Basingstoke.

Greenwood, Shannon, Andrew Perrin, and Maeve Duggan. 2016. *Social Media Update 2016*. Washington, DC: Pew Research Center. http://www.pewinternet.org/2016/11/11/social-media-update-2016/ (Accessed May 1, 2017).

Guo, Chao and Gregory D. Saxton. 2014. "Tweeting Social Change: How Social Media are Changing Nonprofit Advocacy." *Nonprofit and Voluntary Sector Quarterly* 43 (I): 57-79.

Hanna, Sonya and Jennifer Rowley. 2015. "Rethinking Strategic Place Branding in the Digital Age." In *Rethinking Place Branding Comprehensive Brand Development for Cities and Regions*, edited by Mihalis Kavaratzis, Gary Warnaby and Gregory Ashworth, 84-100: Springer.

Hanna, Sonya and Jennifer Rowley. 2008. "An Analysis of Terminology use in Place Branding." *Place Branding and Public Diplomacy* 4 (1): 61-75.

Hoffman, Donna L. and Marek Fodor. 2010. "Can You Measure the ROI of Your Social Media Marketing?" *MIT Sloan Management Review* 52 (1): 41-49.

Hong, Liangjie, Ovidiu Dan, and Brian D. Davison. 2011. "Predicting Popular Messages in Twitter." *Proceedings of the 20th International Conference Companion on World Wide Web*: 57-58.

Java, Akshay, Xiaodan Song, Tim Finn, and Belle Tseng. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities." Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop, ACM Press.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. Finding Influentials Based on the Temporal Order of Information Adoption in Twitter. In Proceedings of the 19th international World Wide Web conference. ACM, Raleigh, North Carolina, USA, pp. 1137-1138. http://dl.acm.org/citation.cfm?id=1772842 (Accessed May 5, 2017).

Long, Scott J. and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables using Stata*. 3rd ed. Stata Press. (589 pages).

Luo, Zhunchen, Miles Osborne, Jintao Tang, and Ting Wang. 2013. "Who will retweet me?: finding retweeters in twitter." *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 869-872. DOI=http://dx.doi.org/10.1145/2484028.2484158

Kim, Annice, Heather Hansen, Joe Murpy, Ashley Richards, Jennifer Duke, and Jane Allen. 2013. "Methodological Considerations in Analyzing Twitter Data." *Journal of the National Cancer Institute Monographs*47: 140-146. DOI:10.1093/jncimonographs/lgt026.

MTDC. 2017. "Montana Launches New Warm Season Campaign." *Montana.gov*. Montana Department of Commerce. accessed June 14, 2017. http://commerce.mt.gov/News/PressReleases/montana-launches-new-warm-season-marketing-campaign.

———. 2012. *Montana Tourism & Recreation Strategic Plan 2013-2017*. 2012. Helena, MT: Montana Department of Commerce. http://marketmt.com/Resources/StrategicPlan

Naaman, Mor, Jeffrey Boase, and Chih-Hui Lai. 2010. "Is it Really About Me? Message Content in Social Awareness Streams." Proc CSCW'10, 189-192.

Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. 2011. "RT to Win! Predicting Message Propagation in Twitter" in International AAAI Conference on Web and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2754

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25: 111-163.

Recuero, Raquel, Ricardo Araujo, and Gabriela Zago. 2011. "How Does Social Capital Affect Retweets?". In Fifth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2807 (Accessed May 4, 2017).

Rodríguez, German. 2007. Lecture Notes on Generalized Linear Models. http://data.princeton.edu/wws509/notes/ (Accessed May 7, 2017).

Rossi, Luca and Matteo Magnani. 2012. "Conversation Practices and Network Structure in Twitter." In Association for the Advancement of Artificial Intelligence, May 20. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4634 (Accessed June 16, 2017).

Saxton, Gregory D. 2015. "Analyzing Big Data with Python." *Social Metrics.* Buffalo, NY: http://social-metrics.org (Accessed December 20, 2016).

Saxton, Gregory D. and Richard D. Waters. 2014. "What Do Stakeholders Like on Facebook? Examining Public Reactions to Nonprofit Organizations' Informational, Promotional, and Community-Building Messages." *Journal of Public Relations Research* 26: 280–299.

Semiz, Gulsah and Paul D. Berger. 2017. "Determining the Factors that Drive Twitter Engagement". Archives of Business Research 5(2): 38-47. DOI: http://dx.doi.org/10.14738/abr.52.2700

Sevin, Efe. 2013. "Places Going Viral: Twitter Usage Patterns in Destination Marketing and Place Branding." *Journal of Place Management and Development* 6 (3): 227-239.

Stata. "Nbreg." Negative Binomial Regression. Stata.com. http://www.stata.com/manuals13/rnbreg.pdf. (Accessed March 20, 2017).

Stieglitz, Stefan and Linh Dang-Xuan. 2012. "Political Communication and Influence through Microblogging--An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior," *2012 45th Hawaii International Conference on System Sciences*. Maui, HI. pp. 3500-3509. DOI: 10.1109/HICSS.2012.476

Suh, Bongwon, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network." Minneapolis, MN, USA, IEEE, 20-22 Aug. 2010. http://dl.acm.org/citation.cfm?id=1907388 (Accessed March 14, 2017).

Twitter. 2017. "Rest APIs." Twitter Developer Documentation. Twitter. https://dev.twitter.com/rest/public (Accessed March 14, 2017).

———. "While You were Away..." Twitter Blog, last modified January 2017. https://blog.twitter.com/official/en_us/a/2015/while-you-were-away-0.html (Accessed May 20, 2017).

UCLA. 2017. "Negative Binomial Regression | Stata Annotated Output." IDRE. UCLA Institute for Digital Research and Education. https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/ (Accessed May 20, 2017).

Williams, Richard. "Scalar Measures of Fit: Pseudo R2 and Information Measures (AIC & BIC)." University of Notre Dame, last modified January 14, 2016. https://www3.nd.edu/~rwilliam/stats3/L05.pdf (Accessed May 20, 2017).

Xiang Zheng and Ulrike Gretzel. 2010. "Role of Social Media in Online Travel Information search". *Tourism Managent*. 31 (2):179–188. http://www.sciencedirect.com/science/article/pii/S0261517709000387 (Accessed June 16, 2017).

Xu, Weiai W. 2016. "Five Steps to Search and Store Tweets by Keywords." Curiositybits. http://www.curiositybits.com/new-page-2/. (Accessed January 2, 2017).

Yasugi, Naoya, Yasuyuki Nishigaki, Wong Meng Seng, Liu Chang Yu, and Hideki Nishimoto. 2013. "Use of Twitter as an Instrument for Disseminating Public Information in Providing Public Goods and Roles of e-Government: Evidence from Japanese Prefectures." *International Journal of Engineering and Innovative Technology* 3 (4): 128-133.

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. 2012. "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing* 9 (6): 811-824. http://ieeexplore.ieee.org/document/6280553/ (Accessed June 16, 2017).

Appendix A

Summary statistics for the original user-object variables in the HT-orig dataset

Variables	Count	Mean	Std. Dev	Min	Max
followers_count	1,869	7,929	24,351	0	191,658
listed_count	1,869	194	668	0	6,514
statuses_count	1,869	9,239	22,356	2	247,758

[+] *exception_mag* cases excluded