Retweets—but Not Just Retweets:

Quantifying and Predicting Influence on Twitter

A thesis presented

by

Evan T.R. Rosenman

to

Applied Mathematics

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 30, 2012

Abstract

There has recently been a sharp uptick in interest among researchers and private firms in determining how to quantify influence on the microblogging site Twitter. We restrict our attention solely to celebrities, and using data collected from Twitter APIs in February and March 2012, we explore four different influence metrics for a group of 60 prominent and well-followed individuals. We find that retweet-based influence is the most significant type of influence, but other effects—like the adoption of hashtags and links—are comparable in terms of generated impressions, and are governed by fundamentally different dynamics. We use the insights from our analysis to develop predictive models of retweets, hashtag and link adoptions, and increases in follower counts. We find that, across different types of influence, the degree to which a celebrity is discussed on Twitter is an extremely useful predictor, while follower counts are comparatively less predictive.

Acknowledgements

This paper could not have been written without the input and support of many individuals. First and foremost, I thank Mike Ruberry for his invaluable assistance in formulating the ideas and writing the Java code that made this project possible. I thank my adviser Yiling Chen for her sage advice throughout this process, and also thank Michael Parzen and Cassandra Pattanayak for their recommendations regarding the statistical methods utilized in this paper.

I also owe a debt of gratitude to my parents and to my friends for their moral support throughout this process. Thanks to Kristen Hunter for her advice on both the analytical methods and content of my paper. Thanks to Matt Chartier for his encouragement and his relentless enthusiasm regarding programming and computer science. Thanks to Daniel Norris for his support. Lastly, thanks to Kevin Fogarty and Danielle Kolin for working alongside me and motivating me at the end of this process.

Contents

1 Introduction **4**

1.1 Motivation 6

1.2 Defining Influence and Our Question 8

1.3 Overview of Results 9

2 Background and Related Work **10**

2.1 Twitter Review and the Twitter Graph 10

2.2 Using the Static Twitter Graph to Determine Influence 12

2.3 Using Activity to Determine Influence 14

2.4 Tweet Virality 16

3 Data Collection and Methods **18**

3.1 Data Collection 18

3.2 Data Processing 20

4 Retweet Influence **24**

4.1 Introduction to the Data 24

4.2 Relationships Among Variables 26

4.3 Retweet Count Variability 29

4.4 Groupings and Ratios 31

5 Non-Retweet Influence **38**

5.1 Local Adoption 39

5.1.1 Definitions and Methods . . . 40
5.1.2 Per-Follower Effect . . . 41
5.2 Co-Mention Adoption . . . 45
5.3 Word Adoption . . . 48
5.4 Emotional Transmission Analysis . . . 54

6 Predictive Models **59**
6.1 Methodology . . . 59
6.2 Retweet Impressions Predictive Model . . . 60
6.2.1 Model Construction . . . 61
6.2.2 Discussion . . . 65
6.3 Non-RT Hashtag and Link Impressions Predictive Model . . . 66
6.3.1 Model Construction . . . 66
6.3.2 Discussion . . . 67
6.4 Follower Uptick Predictive Model . . . 68
6.4.1 Model Construction . . . 69
6.4.2 Discussion . . . 71

7 Conclusions **73**
7.1 Review of Major Contributions . . . 73
7.2 Future Work . . . 75

A List of Celebrities and Designations in the Dataset **77**

B Stop-Word List **79**

C Candidate Predictor Variables **80**

Chapter 1

Introduction

In the summer leading up to the 2012 Republican Presidential primary, former Speaker of the House Newt Gingrich defended his flailing candidacy by telling the *Marietta Daily* that he had "six times as many Twitter followers as all the other candidates combined" [40].

The claim was technically true—with 1.3 million followers, Gingrich had a much larger Twitter presence than any other contender for the Republican nomination. But the day after the Gingrich interview was published, gossip blog Gawker posted a story accusing Gingrich of faking the vast majority of his followers by creating dummy follower accounts and by paying real Twitter users to follow him. Soon thereafter, social networking research firm PeekYou went public with its claim that an analysis of Gingrich's followers suggested that 92% of them were fake [10].

These reports were never confirmed, and some later analyses seemed to vindicate Gingrich [39]. Nonetheless, the controversy—which came to be known as "TwitterGate"—drew attention to the ambiguous nature of quantifying social media influence.

Later that August, pop star Beyonce Knowles announced at the MTV Video Music Awards that she was expecting her first child with her husband, rapper Jay-Z. News of Knowles' pregnancy spread like wildfire through social media. Later reports found that Knowles' announcement caused the single most active tweeting frenzy in Twitter history, with more than 8,800 tweets sent about the topic per second in the minutes following her announcement [35].

Given the attention she garnered, one might naturally think that Knowles is a highly influential user of Twitter. And indeed, the pop star had attracted an impressive audience on the site, reaching one million followers in March 2011 [41]. Yet, despite her large audience and the enor-

mous interest Twitter users expressed about her life, there was one fact to call Knowles' "Twitter influence" into question: she had never sent a single tweet.

Stories like these demonstrate two important consequences of the ascendancy of social networking and micro-blogging services over the past half decade. First, the explosion in popularity of these sites has created a new paradigm in the ways prominent individuals can communicate with and potentially influence others. Celebrities using Tumblr, Instagram, and, most of all, Twitter [24] can communicate with millions of fans many times a day through the messages they send. They can use this secular pulpit to transmit news, promote new products, introduce new topics into the national conversation, and even alter their fans' moods.

Second, understanding who is or isn't influential on these sites is a task fraught with ambiguity. As Cha et al. explain, following relationships "in social media could represent anything from intimate friendships to common interests, or even a passion for breaking news or celebrity gossip" [9]. Thus, using the size of a celebrity's follower count to estimate influence is misleading, since Twitter users may choose to follow a celebrity for many reasons other than an interest in the celebrity's message—and follower accounts may even be faked. The degree to which celebrities are *discussed* on Twitter has also been proposed to quantify influence [9], but this metric is similarly problematic. In talking about a celebrity, Twitter users may be expressing their fandom or approval of the celebrity—but they may also be criticizing or condemning him or her [7]. As a result, it would be fallacious to assume that heavy Twitter discussion about a celebrity necessarily indicates a high level of interest in reading the content of his or her messages.

Interest in the topic of Twitter influence has notably spiked over the past year, owing at least partly to the ambiguity about quantifying influence—and the potential marketing benefit of identifying influencers. New services, such as Klout, Peerindex, and Twitter Grader [34] have gained attention by purporting to offer numerical scores that quantify users' social influence. Tech companies such as Yahoo! [31] and Hewlett-Packard [33] have also conducted substantial research on characterizing influential Twitter users. Yet concerns have been repeatedly raised about the accuracy of the influence-scoring services [18, 34], and the diversity of approaches currently being proposed to quantify Twitter influence [9, 33, 42, 7, 13] demonstrates that no single algorithm has

emerged as the obvious best choice.

This thesis focuses on answering the question: "How do we know who is influential on Twitter?" We define an "influence effect" to mean any type of Twitter activity in which an individual demonstrates that he or she has been influenced by a celebrity. Using a group of 60 celebrities as our potential influencers, we begin by exploring the simplest influence effect seen on Twitter: retweeting. We then dive deeper into an analysis of other types influence effects, using the textual content of individual tweets. This exploration provides us with a number of insights into the factors that determine whether a celebrity is influential. Ultimately, this helps us to generate predictive models of Twitter influence.

1.1 Motivation

Though Twitter provides a medium through which celebrities can speak directly to exceptionally large numbers of fans, it is not necessarily obvious that Twitter influence has meaningful consequences outside the realm of social media. However, Twitter's unique popularity and usage profile, as well as the results of a number of preliminary studies, both provide strong evidence that Twitter influence also matters in the real world.

Twitter was founded in 2006, and today it commands more than 100 million active users worldwide [5]. Despite, or perhaps because of, the extreme limitations the site places on content generation—users can post messages of only 140 or fewer characters in length, and many rich social features of other sites, like commenting on posts or photo tagging, are notably absent on Twitter—the site has also attracted a core base of highly devoted users. Half of all active Twitter users log in every day [5], and as of August 2011, Twitter had surpassed 200 million messages sent daily. Activity continues to increase [1].

In venues as diverse as the traditional news media, academic sociology, and market research, Twitter is widely seen as a useful real-time gauge of public reaction to major developments [30, 6]. This characteristic gives Twitter enormous predictive value, and researchers have found that feeds from the site can predict everything from the stock market [6], to box office revenues [3], to public opinion polls [28]. Yet, given its scale and ubiquity, Twitter does not just reflect public sentiment about products, stories, or individuals—it also has the potential to define sentiment about these

topics. This property means that those who wield influence on Twitter can meaningfully direct public opinion, which leads directly to the question that motivates this thesis: the question of Twitter influence.

Early academic work in this area has found that influencers can play a major role in information propagation on the Twitter network. González-Bailón et al. found that individuals who were central to the Twitter network (i.e., who were well-connected and whose friends were also well-connected) played a crucial role in spreading information about a popular protest movement in Spain in May 2011 [14], encouraging more individuals to be recruited to the protest. Similarly, Cha et al. investigated the role of influential Twitter users in spreading information about three major 2010 news stories, and found that the number of retweets and mentions that users received for their posts about the stories followed a power-law distribution. This means that the most influential individuals are many orders of magnitude more influential than the average user, and that these top influentials have the potential to spread information to an extremely broad audience. Validating the private sector research in this area, Cha et al. concluded that "utilizing top influentials has a great potential payoff in marketing strategy" [9].

The encouraging results of these preliminary studies, along with Twitter's unique structure and the private sector interest in developing algorithms to score Twitter influence, lead us to two essential motivations underlying this thesis:

1. Because Twitter provides a unique medium through which celebrities can communicate with large numbers of individuals *whom they do not personally know*, it allows us to explore how influence is exerted among people who are not personally acquainted. This impersonal influence effect is exceedingly important, because in venues from fashion to politics, we are often impacted by the behavior of prominent individuals, even if we never directly interact with them.

2. Our understanding of influence gives us insight into creating a more accurate way to predict which Twitter users can best propagate a message to others. This has enormous implications for viral marketing, public relations, political campaigns, and the dissemination of breaking news by the media. Using our models as a basis, a group interested in spreading a message could potentially predict which individuals would serve as the best message propagators and

pay high-impact propagators to tweet the message to their followers.

1.2 Defining Influence and Our Question

The term "influence" itself is somewhat imprecise, so we begin our analysis by providing a rigid definition and explanation of influence. In this paper, we define the term to mean "*the ability to, through one's own behavior on Twitter, promote activity and pass information to others.*"

Under this definition, there are many different ways in which users can respond to celebrity-generated content to demonstrate that they have been influenced. It is worth noting which behaviors are and are not examples of influence. A tweet that simply talks about a celebrity—e.g. "I love Justin Bieber's new song!!"—is not an example of Twitter influence, because the celebrity's own Twitter behavior plays no role in changing the sentiments or emotions of the tweeter. Similarly, if Justin Bieber tweets "Merry Christmas" and numerous others tweet Christmas messages in the subsequent hours, this is not an example of Twitter influence, because the topical similarity of Bieber's tweet and others' tweets is due to factors exogenous to Twitter. However, if Bieber tweets a specific message which is retweeted by others; if he introduces a new topic or piece of news that is then discussed by others; or if the positive or negative affect of his tweets is reflected in a change in the sentiment of others' tweets, then these are real examples of influence because Bieber's behavior is driving the behavior of other tweeters.

Our analysis also requires that we precisely define the population we wish to study. For the purposes of this thesis, we restrict our attention solely to widely known celebrities. We posit that the dynamics of how individuals exert influence will be similar among members of this group, but may differ from the dynamics of how lesser known individuals exert influence. We define a "celebrity" to be any individual ranked among the 1,000 most followed individuals on Twitter as of January 17, 2012. Sixty celebrities were selected from this group, according to a methodology described in chapter 3, to be used in our analysis.

Lastly, based on the results of initial work in this area (see Chapter 2), we define three key hypotheses which direct our research:

1. A celebrity's total number of followers (his or her *audience*) is not strongly predictive of his or her influence.

2. The degree to which a celebrity is talked about on Twitter (his or her *buzz*) is also not strongly predictive of his or her influence.

3. There exist meaningful influence effects aside from retweets, and these effects can be quantified by analyzing hashtag, link, and word frequencies in individual tweets. These effects should be considered when determining a celebrity's influence, and can be used to produce a more accurate predictive model of retweets.

1.3 Overview of Results

In our initial analysis of retweet-based influence, we find that both follower count and mention count—metrics of audience size and buzz, respectively—are actually quite strongly correlated with retweet counts. We find, also, that the distribution of retweets for each celebrity tends to be dramatically right-skewed and to have high variance. We categorize our celebrities into six different groups and introduce two ratios, the influence to buzz and influence to audience ratios, which take on distinct values across the celebrity groups. Using these results, we develop an interpretation of how different types of celebrities are able to become influential on Twitter.

We investigate other types of influence—including hashtag and link adoption, word adoption, and emotion adoption—and find that they are fundamentally different from retweet influence. Some of these influence metrics appear to be governed by a celebrity's level of engagement with his or her individual followers, while others are more related to a celebrity's audience size and the type of content that they generate. Lastly, using the insights from our prior analyses, we generate two predictive models of influence and a predictive model of follower upticks. We find that the degree to which a celebrity is being talked about on Twitter is an important predictor in both of the influence models, while follower count is not a meaningful predictor in either model. This indicates that buzz is indeed an important determinant of influence, and that it better captures Twitters users' interest in a celebrity's message than the celebrity's follower count. Furthermore, all three models prove to have impressive predictive ability.

Chapter 2

Background and Related Work

2.1 Twitter Review and the Twitter Graph

A number of simple definitions and behaviors are key to understanding Twitter activity. The central method of communication on Twitter is the *tweet*, a message of up to 140 characters in length. If one person is a *follower* of a celebrity, this means that the celebrity's tweets are immediately visible to the follower from this individual's home page. We say the celebrity is one of the individual's *followees*. Twitter users can also send messages directed specifically at other users, called *mentions*, using the syntax "@[username]" anywhere in the tweet. Mentions can be sent to anyone on Twitter, not merely one's followers or followees.

There also exist several common conventions regarding tweet content. If a user enjoys someone else's tweet and wishes to share it with his own followers, he can *retweet* it, thus sending the same message as one of his own tweets. Retweets often contain an acknowledgment of the original poster, using either "RT @[username]" or "via @[username]" syntax, though this is not universally adopted. Users can also characterize tweets using *hashtags*, denoted by the syntax "#[word]". Hashtags are generally contained at the end of a tweet, and indicate the general topic of the tweet, such as "#Superbowl" or "#SOTU" (State of the Union). Lastly, users can also share links in their tweets using any number of link-shortening services, which create links that redirect to a longer URL, though they can also directly share any URL shorter than 140 characters in length. Researchers at Microsoft [8] investigated the relative frequency of these types of tweets using a random sample of tweets collected in 2009. They found that 36% of tweets were mentions, 5%

of tweets contained a hashtag, 3% of tweets were retweets, and 22% of tweets contained a URL. Furthermore, retweets were substantially more likely to contain a hashtag (18% of retweets) or a URL (52% of retweets).

As a social network, Twitter is unique in several ways. Unlike Facebook and LinkedIn, Twitter employs an *asymmetric* following model. This means that Twitter relationships are directed and not necessarily reciprocal—a user x may follow a user y without user y following user x. This asymmetry is widely employed throughout the network, and non-mutual following is quite common, especially among the most-followed individuals on Twitter. In fact, researchers have found that only about 22% of Twitter relationships are mutual [19].

Furthermore, Twitter does not impose a limit on the number of followers any one user can have or on the number of other tweeters that one user can follow. As a result, the distribution of in-degrees (i.e. number of followers) and outdegrees (number of followees) of Twitter users has a low mean but an extremely long tail. In fact, both quantities are believed to follow a power-law distribution [19], though the distribution of followers is significantly more skewed for in-degrees [4], as heavily followed individuals are significantly more common than aggressive followers. This distribution of followers underscores an important truth about Twitter: since users can choose to follow anyone they like, it is extremely common for active Twitter users to follow celebrities and media personalities who they do not personally know. Defining a "friend" as an individual whom one has mentioned in at least two tweets, researchers have found that the mean friend-to-followees ratio on Twitter is 0.013 and the median is 0.04 [17]. Furthermore, the number of friends generally saturates quickly as the number of followees rises [17]. Thus, it seems that users often connect with individuals with little intention of actively communicating with them; rather, many use Twitter as a way to passively interact by reading others' updates.

Lastly, the long-tailed power-law distribution has been found to describe not only the topology of Twitter, but some types of activity on the network. In particular, the number of retweets that any tweet receives appears to also be power-law [21], indicating that most tweets receive very little attention, but a handful receive a very large amount of attention.

2.2 Using the Static Twitter Graph to Determine Influence

The Twitter "graph"—the collection of nodes (representing users) and edges (representing following relationships) on Twitter—has received substantial attention in prior work as a potential indicator of who is influential. Since the topological relationships within a graph have been shown to have enormous value for predicting the importance of webpages, as demonstrated by the success of Google's PageRank algorithm [29], one might naturally assume that similar properties would hold within a social network. The basic logic underlying this assumption—that Twitter users would tend to form following relationships only with individuals whose tweets they intend to read, internalize, talk about, and be influenced by—seems quite reasonable. Yet existing work has found that the static graph is, at best, a mediocre indicator of who is actually influential on Twitter.

Kwak et al. [19] and Cha et al. [9] both investigated the relationship between the simple in-degree (number of followers) of a Twitter user and his or her influence. Both groups of researchers compiled lists of the most influential Twitter users under a variety of metrics, including in-degree and retweet count. And in both cases, the researchers found that there was substantial difference between the lists of most-followed individuals and most-retweeted individuals. Under the definition of influence provided in Chapter 1, retweet frequency is an extremely meaningful component of a celebrity's overall influence. Therefore, this research strongly indicates that follower count does not accurately capture influence.

Besides merely using in-degree as an indicator of influence, several researchers have also applied graph-ranking algorithms to subgraphs of Twitter, with mixed results. Kwak et al. [19] compiled a list of the top 20 Twitter users with the highest PageRank along with their other rankings. The PageRank-based list was found to align extremely well with the list based on follower counts, but quite poorly with the list based on retweets. And Ghosh et al. [13] tested an alternative graph-ranking algorithm known as Alpha-Centrality on the Twitter graph. The Alpha-Centrality metric is extremely similar to PageRank, but differs in that it is "non-conservative," meaning that it allows one node to donate some of its rank to another node without losing any of its own rank. The researchers hypothesized that since information diffusion is fundamentally non-conservative—i.e., one becomes no less aware of a piece of information by informing others about

it—that Alpha-Centrality might be a better way of quantifying influence on Twitter. However, the researchers actually found that Alpha-Centrality was worse than PageRank at predicting who would drive Twitter activity.

In several other papers, however, both in-degree and other graph-based rankings have been shown to be useful for predicting certain types of influence on the network. In some of these papers, the populations studied appear to exclude the types of big-name celebrities who are posited in this research to be the most influential Twitter users. For instance, Weng et al. [42] successfully utilized an extension of the PageRank algorithm to identify influential individuals among a group of active Singapore-based Twitter users. However, the researchers' sample was relatively small (containing 6,748 individuals), did not appear to contain any major celebrities, and exhibited much higher follower reciprocity (72%) than would be found in a random sample of Twitter. Thus, it is unlikely that any conclusions drawn from this sample would apply to the broader Twitter network. Similarly, Suh et al. [38] analyzed a dataset of 10,000 unique tweets and the retweets they generated. The researchers found that the retweet count of each tweet was almost perfectly linearly correlated with the number of followers of the original tweeter. However, the most-followed individual in the dataset had about 5,000 followers, far fewer than the celebrities we seek to analyze.

Other papers have appeared to find a more generalizable connection between Twitter topology and influence, though the relationship never appears to be strong enough to directly infer influence from the graph alone. Ardon et al. [2] investigated how topics become popular on Twitter, identifying about 6.2 million topics within 52 million tweets sent in 2009. The researchers did find that popular topics are generally initiated by users with very high follower counts (particularly celebrities or web-based news media outlets). However, not all topics started by celebrities became popular; rather, celebrities could influence the spread of topics, but they could not make them popular unless common users picked them up. With a similar focus on seeders of viral topics, Bakshy et al. [4] investigated the "diffusion trees" that occur when a single tweet is retweeted many times across the network, using a dataset of 74 million diffusion events in 2009. They created a predictive model of how many individuals would retweet a given link and found that the number of followers of the original tweeter was an important input into the model, though it was less important than other factors.

Romero et al. [33] looked at a dataset of 22 million tweets referencing 15 million distinct URLs

sent over a period of two weeks in September 2009, and applied a weighted PageRank algorithm to the static graph of tweeters in the dataset. When looking exclusively at the most retweeted 0.1% of links, they found that the PageRank of the original person who tweeted the link was quite a good predictor of the overall traffic the URL received, with an R^2 value of 0.84.

Taken together, this group of papers points to a few meaningful conclusions. First, follower count seems to have a surprisingly weak, though not nonexistent, relationship with a celebrity's ability to drive activity on Twitter. Second, graph-ranking algorithms—which capture more information about the static graph structure of Twitter than in-degree alone—seem to be slightly better, but still inadequate, indicators of influence. Both of these static metrics also seem to be more useful in predicting broad diffusion events across the Twitter network than in predicting everyday activity. Thus, the body of existing research implies that one must analyze more than merely the Twitter topology in order to understand who is truly influential.

2.3 Using Activity to Determine Influence

Researchers appear to have had greater success in investigating influence when they take into account actual *activity* on the Twitter graph.

Cha et al. [9] also compiled a list of the Twitter users who are mentioned the most and correlated it with the list of users who are retweeted the most, finding that "in general, users who get mentioned often also get retweeted often, and vice versa." Because a directed mention toward a celebrity is unlikely to elicit a response, a high number of mentions may be taken as an indication that the celebrity has attracted a lot of *buzz*, while a high number of retweets largely indicates that the celebrity is *influential*. Thus, this work seems to indicate that buzz and influence are indeed fairly well correlated.

Steeg et al. [37] focused even more directly on activity as an indicator of influence. The researchers collected data about the tweet histories of about 800,000 Twitter users and filtered down to those users who tweeted more than 10 URLs over a three-week period in the fall of 2010. They then calculated the "transfer entropy"—the reduction of uncertainty about whether one user will tweet a link given that another user has already tweeted it—among every pair of users in the dataset. In doing so, they created an empirical metric of how much any one user appeared to

influence another. In keeping with previous results, they found that while more-followed individuals had greater total transfer entropy on average, there was still dramatic variation in transfer entropy among individuals with the same number of followers. Though most of the highly influential relationships they identified turned out to be between spammer accounts, the researchers were also able to show that certain individuals with modest followings—like Brazilian politician Marina Silva—were able to have a large effect on the network because their followers were so likely to retweet the links they sent out.

Bakshy et al. [4], who created a predictive model of link retweets, found that follower count was a meaningful indicator of the retweets that one can generate—but that past activity on the graph was most informative. In particular, they found that individuals' past history of getting their immediate friends to adopt links was the strongest predictor of how many retweets they could generate. That past *local adoption*—rather than *global* adoption—was most informative is quite surprising, but appears to be due to the fact that the researchers' dataset consisted of a random sample of links tweeted over a two-month period in 2009. Because most individuals have very little Twitter influence, the median link in the dataset was not retweeted at all, which biased the model toward prioritizing local adoption. For our purposes, most celebrity tweets are likely to attract at least a moderate number of retweets, so these results may not apply to our data.

Lastly, Romero et al. [33] adapted a graph-ranking algorithm to take into account a user's history of activity, and generated a robust predictive model of the traffic received for tweeted URLs. In particular, they weighted edges between nodes based on the proportion of links sent by the follower that the followee had retweeted in the past, and then ran an algorithm somewhat similar to PageRank, calculating a quantity known as the IP Influence Score. They found that this score was the best overall predictor of retweet URL traffic, and that the most influential individuals identified by the algorithm—including Ashton Kutcher, The Onion, and social media blog Mashable—fit well with general intuitions about who should be deemed influential on Twitter.

From these papers, it is clear that analyzing and quantifying Twitter activity is a key step in identifying influential users, as the topology of user following relationships is inadequate. Yet prior research has offered only a patchwork of definitions of influence, with many defining the concept solely using a tweeter's retweets or her ability to generate traffic on a posted link. In this thesis, we strive to capture a broader definition of influence by looking not just at retweets

or a celebrity's ability to propagate a link, but also at a celebrity's ability to propagate hashtags, individual words, and emotional states to others on Twitter. Using this broader approach, we hope to more meaningfully utilize Twitter activity to capture influence.

2.4 Tweet Virality

Because the ability to generate content that "goes viral" across Twitter is an important component of how we define influence, a brief overview of work on Twitter virality is also in order.

Hansen et al. [16] investigated the features of tweets that garner large numbers of retweets, analyzing a dataset of 210,000 "news-y" tweets about the 2009 United Nations Climate Change Conference as well as a random sample of about 350,000 tweets from 2010. The researchers originally hypothesized that negative tweet sentiment would promote more retweets. They found a statistically significant bias toward retweeting negative content among the news-y tweets, but not among the randomly sampled tweets. In fact, when the randomly sampled tweets were filtered down to those with a nonzero "arousal score" (i.e. those tweets which were designated as having either negative or positive sentiment, but not neutral sentiment), there was a significant bias in favor of retweeting *positive* tweets.

Suh et al. [38] also looked at features that promoted retweets in their dataset. They found that tweets with URLs are notably more likely to get retweeted, and that the retweet rate is heavily dependent on the domain of the URL, with sites like twitlonger.com, mashable.com, and nytimes.com engendering the highest retweet rates. The presence of a hashtag was also found to have a positive correlation with retweet rate, with dramatic variation in the retweet rate between different hashtags. Follower and followee count were also correlated with retweet rate. The original tweeter's activity—as measured by the total number of tweets posted from his or her account—did not have a strong relationship with retweet rate, but the age of a Twitter account did have a minor positive relationship with retweet rate.

Lastly, Lehmann et al. [20] analyzed hashtags that were used over six months in 2008 and 2009, and found that the hashtags tended to trend in three ways: continuously (for those hashtags used virtually all the time, like "#music"), periodically (for those hashtags used on specific days like "#followfriday"), or uniquely (for those hashtags referring to a specific event, like "#oscars").

Furthermore, they found that the activity profile surrounding the trending event was very much tied to the content of the hashtag. Hashtags related to *anticipated* events, like "#masters" for the 2009 Golf Masters, had activity concentrated before and during their activity peaks; *unexpected* event hashtags, like "#winnenden" for the Winnenden school shooting, had activity concentrated during and after their activity peaks; *transient* event hashtags, like "#gfail" for a Google service outage, had activity almost totally concentrated on the day of their activity peaks; and *ongoing* event hashtags, like "#watchmen" for the release and performance of the movie *Watchmen*, had activity concentrated symmetrically about their activity peaks.

The latter two categories of hashtags were shown to be significantly more viral in nature—meaning a larger proportion of their usage occurred in retweets—than the former two categories. Furthermore, hashtags with activity concentrated during and after their peaks tended to have a high number of seeders, meaning that information about these hashtags was transmitted largely in a manner exogenous to Twitter. This work thus indicates that individuals can achieve greater influence by tweeting about specific types of topics. It also provides us with a useful framework with which to understand Twitter trends and how they can be impacted by the input of influential users.

Chapter 3

Data Collection and Methods

3.1 Data Collection

We restricted our focus solely to those individuals who are already well-known enough to be characterized as celebrities. For the purposes of this thesis, we define a "celebrity" to mean any individual who was among the 1000 most-followed individuals on Twitter as of January 17, 2012. Due to constraints on the amount of data we could reasonably scrape from Twitter's APIs, we decided to gather data on only 60 individuals.

Knowing that the in-degree distribution for Twitter users follows a power-law [19], we noted that the most-followed individual on Twitter (Lady Gaga) had roughly $10^{7.25}$ followers, while individuals at the bottom end of the top 1000 most-followed list had roughly $10^{5.75}$ followers. We then divided the top 1000 individuals into seven groups based on follower counts, with the follower count ranges defined by quarter powers of 10 below $10^{7.25}$. Thus, the top group included all individuals with $10^{7.0}$ to $10^{7.25}$ followers, the next group included all individuals with $10^{6.75}$ to $10^{7.0}$ followers, etc. Next, we quasi-randomly selected eight English-speaking Twitter users from each group. As a result, we included all of the top eight most-followed Twitter users, but only a sample of those with fewer followers. The follower counts of these 56 individuals, in fact, closely followed a power-law distribution, as can be noted from the linearity the plot in figure 3.1:

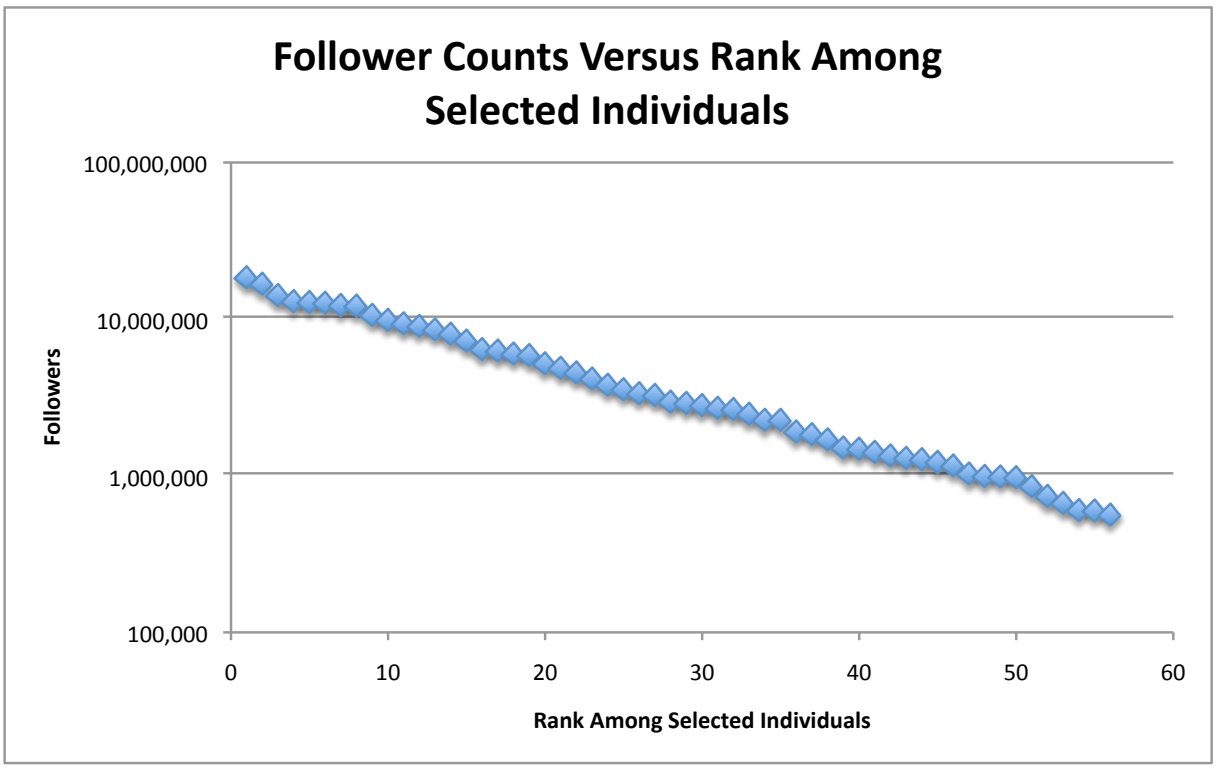

Figure 3.1: Logarithmic plot of follower count versus follower rank for 56 of the 60 celebrities

A handful of strategic selections were also made. Newt Gingrich, who had enough followers to rank in the top 1000 most-followed Twitter users, was intentionally added to the group. We also included four other individuals running for the Republican nomination for President—Mitt Romney, Ron Paul, Rick Perry, and Rick Santorum—who did not have enough followers to fit our definition of a "celebrity," but who were deemed to be sufficiently interesting to warrant inclusion in the dataset. The full list of individuals can be found in chapter 4.

Twitter Queries

Two different Twitter APIs were queried by four different programs over the period from February 15 to March 16, resulting in a full dataset containing about 157 million tweets.

The Streaming API was queried in order to get all tweets sent by the 60 individuals in the dataset, along with any explicit mentions (using the "@[username]" syntax), implicit mentions (using the Twitter user's full name, without an "@" symbol), and retweets of their content. Visual inspection confirmed that the API gave us all the tweets actually sent by the celebrities. The API did not return all of the actual mentions and retweets of the target individuals, though it was

assumed the collected tweets represented a random sample of these tweets. In total, over the time period studied, 47.5 million tweets of this type were collected.

A separate program also continuously queried the Streaming API and downloaded the output of the Twitter "gardenhose"—a random sample of all of the tweets being sent on Twitter at a given time. This data provided a way to compare the tweets being sent by and sent about celebrities with the general Twitter "conversation." According to information publicly available from Twitter, the gardenhose returns roughly 10% of all tweets, though this proportion is heavily dependent on current traffic. In total, 94.2 million tweets were collected from this API.

Lastly, we queried the Twitter Rest API in order to get a sample of the tweets sent by the followers of the target individuals. Because the API was extremely rate limited, two separate programs were set up to query for tweets by the followers of the 60 celebrities in our dataset. In each request, the most recent tweets for 100 followers of each celebrity were returned, though many of these returned values were null due to individuals' privacy settings. Including these null values, a total of 15.7 million tweets were returned by this API over the time period, though due to some redundant sampling, only 6.5 million of these tweets were usable.

3.2 Data Processing

Data was stored in text files containing roughly one million tweets per file. We wrote a custom data store to handle the data, allowing rapid sequential processing over the millions of tweets in our database. Our iterator program returned information about the content, timestamp, and sender for a given tweet, as well as the follower count of the sender and the original tweeter and retweet count for all retweets in our dataset.

English Classifying

Additional processing programs were also written in order to analyze word frequencies and other aspects of tweet content. Because we were solely interested in English-language tweets—and many of our calculations would have been skewed by the inclusion of non-English tweets—we needed a method to detect our tweets' language. However, though several free language-detection APIs exist, we found that the enormous size of our tweet set—each one requiring an HTTP requests on

the order of 10^{-1} seconds—made the use of these sites prohibitively slow.

To circumvent this problem, we used a free language-detection API available at detectlanguage.com to classify 500,000 of our tweets as English or non-English. We then used these tweets as a training set for a Bayesian English-classifying program. Our classifier worked by breaking up the individual words in the tweet. The probability that an individual word w_i is English, denoted $Pr(w_i)$, was calculated according to:

$$Pr(w_i) = \frac{e_i}{n_i + e_i}$$

where e_i = number of appearances in English tweets and n_i = number of appearances in non-English tweets. The following heuristics were also used in cases when one or both of n_i and e_i were zero:

$$Pr(w_i) = \begin{cases} 0.99 & \text{if } n_i = 0; \\ 0.01 & \text{if } e_i = 0; \\ 0.50 & \text{if } n_i = 0 \text{ and } e_i = 0 \end{cases}$$

The overall probability that a tweet T = $\{w_i\}$, where $0 \leq i \leq n$, was an English tweet was then calculated according to the following formula:

$$Pr(T) = \frac{Pr(w_1) * Pr(w_2) * ...Pr(w_n)}{Pr(w_1) * Pr(w_2) * ...Pr(w_n) + (1 - Pr(w_1)) * (1 - Pr(w_2)) * ...(1 - Pr(w_n))}$$

Tweets with a probability $Pr(T) > 0.90$ were designated as English. When training on a subset of 400,000 of the classified tweets and testing against the remaining 100,000, the Bayesian classifier was found to match the results of the language-detection API more than 92% of the time, though visual inspection on a subset of 600 of these results implied that the accuracy was closer to 96%. Furthermore, at a processing rate of roughly 400,000 tweets per minute, the Bayesian classifier was roughly 10^3 times as fast as querying the API repeatedly.

Stop-Listing

When constructing word frequency vectors from different sets of tweets, we used an English stop list derived from the list offered by the Snowball project [32]. These words were removed in order to prevent an illusory correlation between the tweet content of celebrities who use many common words and the tweet content of other Twitter users. The full list of stop-listed words can be found in Appendix B.

Emotional Valence

Emotional valence scores were calculated using AFINN [26], a labeled word list designed for analysis of emotion in text. The list—generated by Danish researcher Finn Arup Nielsen—contains 2,477 words scored between -5 and +5 for their emotional content [27]. Nielsen scored the words by beginning with a set of obscene and positive words, and extended the list by looking at a set of tweets about the United Nations Climate Conference. Internet slang was also added to the list, as well as words from other affective word lists. The list was tested against a set of 1,000 tweets that had been labeled manually using Amazon Mechanical Turk, and the scores generated from the list were found to correlate well (Pearson correlation = 0.564 and Spearman's rank correlation = 0.596) with the AMT scores.

Regression

All regressions were conducted using the Flanagan Java Scientific Library [11], developed by Dr. Michael T. Flanagan at University College London. Several programs were written using the Flanagan library to conduct large-scale regression analyses. The first program, entitled simply *regress.java*, was written to take in a one-dimensional vector of dependent variable data and a two-dimensional vector containing any number of potential predictor variables. The program sequentially dropped predictor variables until all regressors were found, using a t-test, to be statistically significant predictors at the $p \leq 0.05$ level.

A second program, entitled *crossValidationRegression.java*, was built on top of our regression program to conduct the cross-validation analyses used in our modeling in Chapter 6. The program similarly takes in a vector of independent variable data and a two-dimensional vector of dependent variable data. It then sequentially holds out one data point (one horizontal row) from each vector, and conducts a regression on the remaining data, using the functions in *regress.java*. The resultant model is then used to predict the value of the held-out data point. On each iteration, the program collects the error term, and once every data point has been held out in a single regression, the program reports the RMSE of the analysis.

Granger Causality Analysis

A final method utilized in our data analysis is a tool from economics known as Granger Causality Analysis [15], which has been utilized in other papers that have sought to exploit Twitter as a predictor of real-world processes [6, 22].

Granger Causality Analysis is used to discover whether or not one time series, X_t, is useful for predicting the values of another time series Y_t. The analysis is carried out in two steps. First, the dependent time series, Y_t is regressed on its own lagged values Y_{t-1}, Y_{t-2}, etc. All lagged values with a significant t-statistic ($p \leq 0.05$) are included in the regression, leading to a model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + ...$$

Next, the lags of the potential predictor series, X_{t-1}, X_{t-2}, etc. are tested as candidates to be added to the model, using the same significance criteria. This leads to a model:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + ...\gamma_1 X_{t-1} + \gamma_2 X_{t-2} + ...$$

If any of these new terms are statistically significant predictors, then the model's overall R^2 value will go up and its residual sum of squares will go down. However, it is still possible that this expanded model is not significantly more predictive than the original model. Thus, an F-test is conducted, using the following formula for the F statistic, where n represents the number of data points in the analysis, RSS represents the residual sum of squares of a model, and v represents the number of parameters in a model:

$$F = \frac{\frac{RSS_{original} - RSS_{expanded}}{v_{expanded} - v_{original}}}{\frac{RSS_{expanded}}{n - v_{expanded}}}$$

The p value corresponding to this F statistic is calculated using an F distribution with $(v_{expanded} - v_{original}, n - v_{expanded})$ degrees of freedom. If the p value is less than a threshold value of 0.05, then the time series X_t has forecasting value for the time series Y_t, and X_t is said to "Granger cause" Y_t.

In Chapter 5, we develop time-series representations of the emotional valence content of a celebrity's tweets and the tweets of his followers. We then use Granger Causality to see if the celebrity's valence time series is a meaningful forecaster of his followers' valence time series.

Chapter 4

Retweet Influence

In the first step of analysis of our tweet dataset, we looked at the simplest type of influence: the ability to get others to retweet a message. Retweets provide a useful introduction to influence because they are a simple, extremely common way of demonstrating interest in a celebrity's tweet content. Because a retweet requires that an individual read a celebrity's tweet and decide to share it with his or her own followers, there is essentially no ambiguity that a retweet constitutes an influence event. Furthermore, much of the past research on Twitter influence has focused on retweets as the primary metric with which to measure influence [9, 42, 38].

We compare a celebrity's retweet count to a number of other aggregate Twitter metrics—including mention count and follower count—and categorize these metrics into three designations: influence effects, buzz effects, and audience effects. The metrics are then correlated with one another and their relationships are explored in depth.

4.1 Introduction to the Data

Using our listener program, we collected data on the frequency of retweets, mentions, and textual references to a celebrity's name for all the celebrities in our dataset. We also collected data on the follower count and tweeting frequency of these celebrities, positing that both are also potential indicators of influence. The data is collected in the following table:

Name	Total Refs	Followers	Tweets Sent	Retweets	Name	Total Refs	Followers	Tweets Sent	Retweets
Adam Savage	7,272	589,148	84	5,895	Ke$ha	292,509	2,618,337	77	29,637
Alicia Keys	209,451	5,787,491	19	7,417	Keri Hilson	97,736	1,809,084	103	22,256
Arnold Schwarzenegger	354,038	2,233,810	50	2,565	Kim Kardashian	607,090	12,634,587	353	162,476
Ashton Kutcher	100,828	9,231,520	36	9,609	Kirstie Alley	22,433	971,069	640	8,373
Barack Obama	442,112	11,982,070	264	91,760	Kristin Cavallari	6,352	730,580	34	4,462
Beyonce Knowles	54,180	2,775,692	0	0	Lady Gaga	1,929,169	18,171,463	23	155,174
Bill Gates	150,769	5,129,156	24	16,281	Marshall Mathers	68,784	8,441,411	3	6,541
Britney Spears	351,251	12,493,390	8	11,734	Michael Ausiello	5,021	1,142,432	384	7,670
Charlie Sheen	171,243	6,322,335	39	8,026	Mitt Romney	421,558	267,876	33	7,583
Chris Anderson	9,847	1,322,278	98	3,834	Newt Gingrich	165,706	1,395,034	166	23,501
Chris Brown	2,604,831	7,162,927	121	371,310	Nicole Polizzi	146,375	4,102,121	152	44,715
Conan O'Brien	29,649	4,802,595	29	41,883	Oprah Winfrey	183,374	8,881,874	276	32,677
Dane Cook	43,016	2,665,401	144	94,355	Paris Hilton	170,943	5,921,344	190	34,864
Daniel Tosh	125,822	4,484,775	213	22,142	Paula Abdul	24,156	2,234,683	279	8,310
Danny Glover	5,286	1,487,899	6	339	Rick Perry	26,539	122,492	28	1,586
David Guetta	287,233	3,304,534	30	9,363	Rick Santorum	450,335	84,543	135	13,682
Dianna Agron	94,341	842,720	112	27,638	Rihanna	6,041,162	12,083,380	276	929,534
Dita Von Teese	12,534	978,909	64	3,241	Robbie Williams	34,459	658,661	54	1,431
Dr. Phil	42,592	960,346	90	2,122	Ron Paul	309,319	195,434	42	7,504
Felicia Day	12,424	1,882,189	487	4,373	Scooter Braun	281,684	1,277,268	519	274,321
Jim Carrey	36,829	6,182,681	0	747	Selena Gomez	1,111,808	9,711,599	25	182,812
John Cleese	8,042	1,675,133	11	1,821	Shakira	404,879	12,798,273	39	21,977
Jon Favreau	1,671	1,203,384	17	630	Soulja Boy	327,676	3,221,567	1,625	46,240
Jon Stewart	55,586	593,174	54	10,124	Stephen Colbert	66,765	2,920,528	55	61,336
Jonas Brothers	371,974	3,501,782	76	28,126	Stephen Fry	67,718	3,740,361	270	63,100
Jordin Sparks	29,851	1,462,312	261	7,684	Suze Orman	7,357	1,252,205	229	3,617
Joy Behar	4,583	554,405	9	2,104	Taylor Swift	1,161,271	10,456,496	20	89,242
Justin Bieber	14,586,842	16,528,367	572	4,474,298	Tom Cruise	84,626	2,442,627	52	1,393
Justin Timberlake	161,776	7,927,692	28	10,356	Travis Barker	23,706	1,014,596	103	7,547
Katy Perry	1,054,498	13,998,128	58	176,002	Usher	955,578	2,864,352	35	25,361

Table 4.1: List of all celebrities and their reference counts, follower counts, tweets sent, and retweet counts.

4.2 Relationships Among Variables

Among these variables, retweet count, reference count, and follower count have all been proposed in the literature as influence metrics [9]. However, under our stricter definition of influence ("the ability to, through one's own behavior on Twitter, promote activity and pass information to others"), the classification is more ambiguous. Retweets are a canonical example of this type of influence. Non-mention textual references to a celebrity's full name are, largely, a non-example, because the majority of these kinds of tweets are simply about a celebrity, rather than in response to a celebrity's Twitter content. Mentions are somewhere in between, as a mention may be a direct response to a celebrity's Twitter statements or may simply be a message sent to the celebrity, though the vast majority appear to be in the latter category.

Our first two hypotheses are concerned with the relationship between influence, buzz, and audience. To fit into this framework, we treat retweets as an influence metric; we call the sum of mentions and non-mention references "total references" and treat it as a buzz metric; and we treat the number of followers as an audience metric. We also analyze the number of tweets sent, treating it as a measure of activity. We seek to find the relationships among these different variables.

As a preliminary analysis, we calculated the top ten highest ranked individuals under each of these metrics. The results are given in table 4.2:

	Audience (Followers)	Influence (Retweets)	Buzz (Total Refs)	Activity (Tweets Sent)
1	Lady Gaga	Justin Bieber	Justin Bieber	Soulja Boy
2	Justin Bieber	Rihanna	Rihanna	Kirstie Alley
3	Katy Perry	Chris Brown	Chris Brown	Justin Bieber
4	Shakira	Scooter Braun	Lady Gaga	Scooter Braun
5	Kim Kardashian	Selena Gomez	Taylor Swift	Felicia Day
6	Britney Spears	Katy Perry	Selena Gomez	Michael Ausiello
7	Rihanna	Kim Kardashian	Katy Perry	Kim Kardashian
8	Barack Obama	Lady Gaga	Usher	Paula Abdul
9	Taylor Swift	Dane Cook	Kim Kardashian	Rihanna
10	Selena Gomez	Barack Obama	Rick Santorum	Oprah

Table 4.2: Top ten ranked celebrities under each of the four metrics. Note the high degree of overlap among the first three columns.

A few insights are immediately evident from these rankings. There is substantial overlap

among the lists of most-followed, most-retweeted, and most-referenced individuals. Lady Gaga, Justin Bieber, Katy Perry, Selena Gomez, Kim Kardashian, and Rihanna appear on all three lists. However, the list of the ten most active tweeters in the dataset is substantially different, with only Justin Bieber, Kim Kardashian, and Rihanna making all four lists. This seems to indicate that the amount of content that one produces has less to do with the size of one's audience, one's influence, or one's buzz than these three metrics have to do with one another.

We next take an aggregate look at the relationships between these metrics across our 60 celebrities. However, the distributions of these metrics pose a problem. As can be seen in figure 4.1, the metric distributions are roughly linear (with downward slope) on a logarithmic plot against rank; this indicates that variance is extremely high, so a simple correlation analysis could be dramatically skewed by a handful of extremely large values.

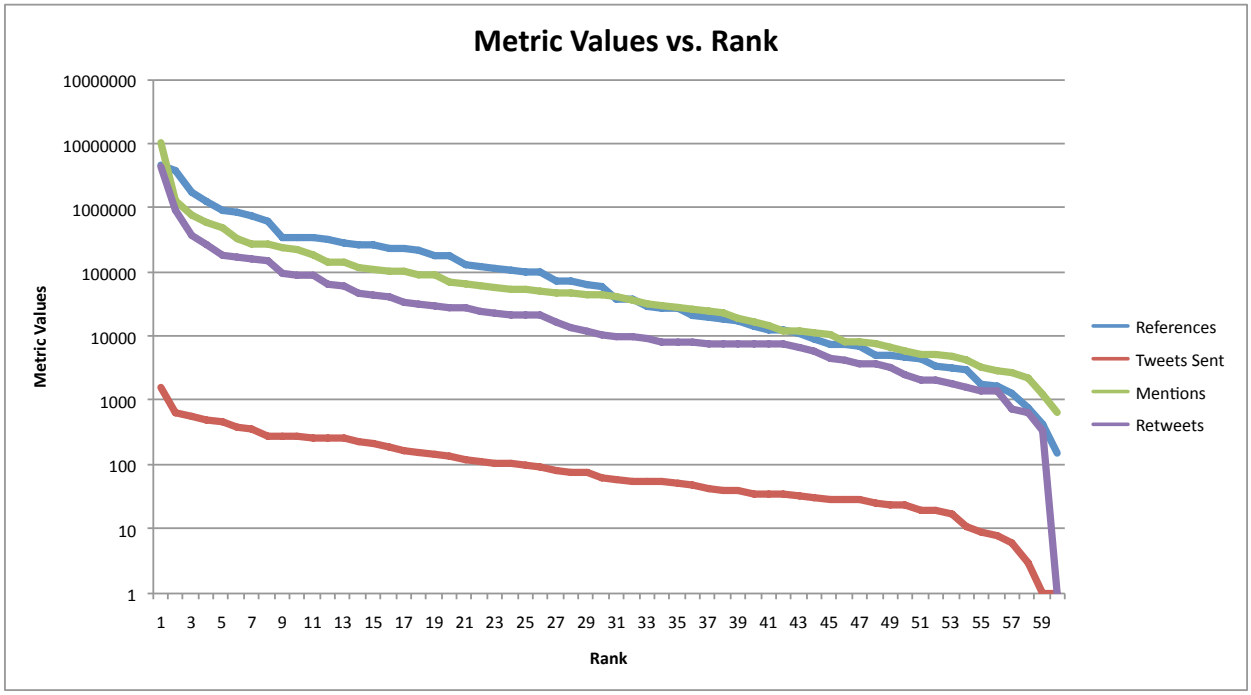

Figure 4.1: Logarithmic plot of each of the metric's values, in order of size. Note that all metrics have consistent, quasi-linear downward slope.

Because of this problem, we use Spearman's rank correlation and log correlation to explore the strength of the relationships among variables. Log correlations are the correlations of variables under a log transformation. Spearman's rank correlation [23] is given by the following formula, where d_i represents the difference in ordinal rank of a single observation when ranked separately

under two distinct metrics:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Log correlation and Spearman's rank correlation are highly insensitive to outliers. The correlations are given in the following table, with the Spearman's rank correlations given as the first entry in each cell and the log correlation given as the second, italicized entry in each cell. The significance of both types of correlations are tested using a two-sided t test. Starred entries are statistically significant at the 0.05 level:

	Followers	**Tweets Sent**	**Retweets**	**Total Refs**
Followers	1.000* *1.000**			
Tweets Sent	-0.056 *-0.045*	1.000* *1.000**		
Retweets	0.579* *0.439**	0.443* *0.566**	1.000* *1.000**	
Total Refs	0.608* *0.524**	0.082 *0.167*	0.700* *0.662**	1.000* *1.000**

Table 4.3: Log correlations and rank correlations among each of the metrics

These correlations tell a slightly different story than the top ten rankings. The number of tweets sent—and, by extension, the overall frequency of Twitter activity—is significantly correlated with the total number of retweets a celebrity receives. This is not surprising, because it is very rare for any user to retweet a single tweet multiple times, but it is much more common to retweet distinct tweets sent by the same celebrity. As a result, every new tweet a celebrity sends is essentially a fresh opportunity to create retweets. Given this fact, we might have reasonably hypothesized that tweeting frequency would have the highest correlation with overall retweet count. However, looking at the above chart, we are surprised to see that the correlation between tweets sent and retweet count is about as high as the correlation between follower count and retweet count, and actually *lower* than the correlations between reference count and retweet count.

Intriguingly, we also see that the correlations between audience size and tweeting frequency, and between reference count and tweeting frequency, are statistically insignificant. The weakness of the former relationship implies that, at least among top Twitter celebrities, followers are attracted

not due to an abundance of content generation on the site, but due to the fame they have attained exogenously to Twitter. The weakness of the latter relationship seems to imply that the amount of buzz a celebrity receives on Twitter is also a factor largely independent of content generation. These facts are helpful in explaining why Beyonce, having never sent a single tweet, was nonetheless able to attract both buzz and followers on Twitter, as explained in the introduction to this thesis.

A number of other insights arise from these correlations. Surprisingly, retweets appear to be significantly and strongly correlated with follower counts, indicating that features of the static Twitter graph may be a more meaningful component of influence than we hypothesized. Nonetheless, the correlation is not perfect; it might be said, then, that when it comes to predicting influence, Twitter topology is not destiny.

Lastly, the relationship between influence and buzz—as measured by the correlation between retweets and total references—is also surprisingly strong. Again, this seems to indicate that the amount of buzz one attracts on the network is a meaningful component of influence, though far from the only meaningful factor. Thus, taken together, this analysis implies that our first two hypotheses were not explicitly correct; *both* buzz and audience size are strong indicators of influence, though they still do not tell the whole story.

4.3 Retweet Count Variability

We now know that, when comparing across celebrities, we learn a lot about a celebrity's aggregate influence by knowing his or her follower count and mention count. Though these relationships appear to be somewhat strong overall, we might wonder if they continue to be true at the resolution of individual tweets. From an application perspective, marketers would want to know that they can generate a certain number of impressions from a tweet by a famous person, irrespective of how that person's other tweets perform.

In order to answer this question, we shift our focus from comparing across celebrities to comparing across tweets sent by the same celebrity. If we find that retweet counts are fairly consistent across tweets for each celebrity—meaning the distribution of retweet counts is peaked, not very skewed, and has low variance—this indicates that the retweet count of an individual tweet could easily be predicted based on the average retweet count. Since we have already found that average retweet count is strongly correlated with audience size and buzz, we could potentially

develop a fairly strong retweet-predictive model using either or both of these variables.

However, somewhat unsurprisingly, we found that the distribution of retweet counts had extremely high variance, *even for a single celebrity.* Assuming a t-distribution of the retweet counts, we calculated the 95% confidence intervals for retweet counts of the 59 celebrities in our dataset who had generated at least one historical retweet (Beyonce being the exception). We found that every single one of the 95% confidence intervals included 0, indicating that we could not be reasonably certain that any celebrity tweet would generate retweets. Plots of the 95% confidence intervals are given in figure 4.2 for the five celebrities with the highest average retweet count. The lower half of the 95% confidence interval is given by the blue bar and the upper half is given by the green bar. The location where the bars meet gives the celebrity's average retweet count.

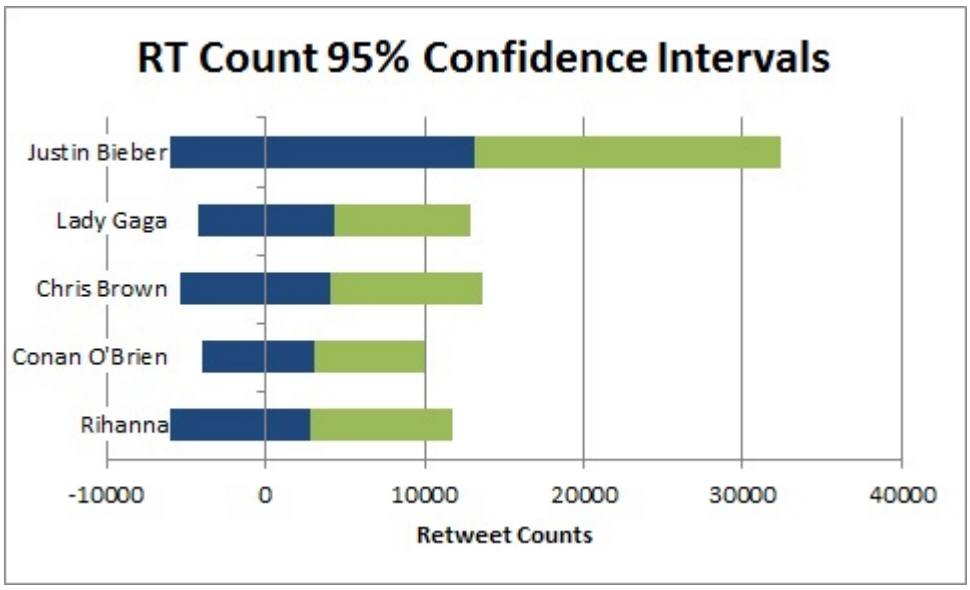

Figure 4.2: 95% confidence intervals for retweet counts for top 5 celebrities

Since retweet counts cannot go below zero, this plot seems to indicate that the distributions of retweet counts for individual celebrities not only have high variance but may also be skewed. Our analysis revealed this to be the case; the average skew of the retweet distributions among our 59 tweeting celebrities was +3.02, and 56 of the distributions demonstrated a skew of at least +1.0. This means that the distributions are, almost uniformly, strongly right skewed. Plots of the distributions for the same celebrities are given in figure 4.3, with the proportion of the celebrity's tweets given on the vertical axis and the number of retweets given on the horizontal axis:

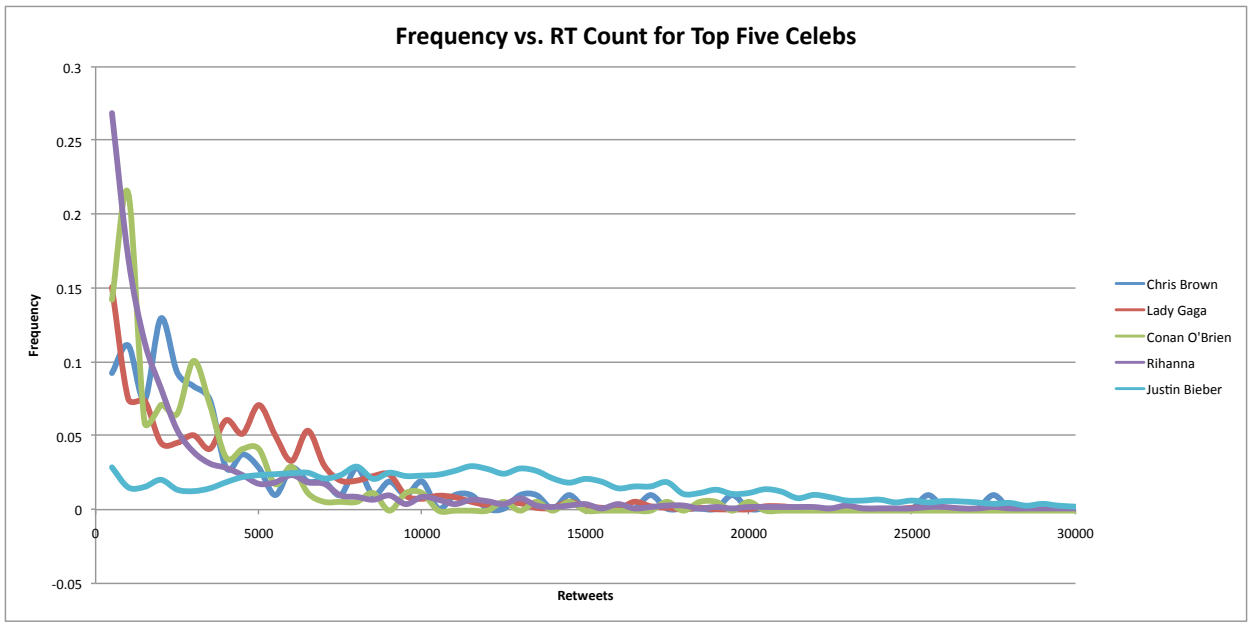

Figure 4.3: Plot of the proportion of a celebrity's tweets which attract a given number of retweets. Note that all celebrities have extremely right-skewed retweet distributions.

The plot confirms that, even for a single celebrity, there is far too much variation in retweet count across tweets to accurately predict the number of retweets that a single post will attract. We have seen that audience size and buzz can serve as moderately strong predictors of aggregate retweet generation, but when looking at a single celebrity (whose retweet generation history is already known), retweet prediction is extremely difficult for individual tweets. This fact provides an important qualification for the model we generate in Chapter 6.2, which predicts average retweet impressions for each celebrity. Though we find that we can predict these average impressions moderately well, this does *not* mean that we can predict the exact impression counts for any individual tweet sent by a celebrity.

4.4 Groupings and Ratios

A final meaningful question we might ask is: "Does the *type* of celebrity impact influence?" This question was explored by grouping the celebrities into six different "fame types" and then calculating the average per-tweet retweet count for each type. Two ratios—the "influence to buzz" and "influence to audience" ratios—were also calculated for each group. These ratios were then compared across the groups to see if celebrity designations actually have a relationship with influence.

The six celebrity designations used were: musician, actor, politician, comedian, TV per-

sonality, and entrepreneur. In instances where celebrities could reasonably receive more than one designation (such as Arnold Schwarzenegger, who is both a famous actor and a former governor), celebrities were given the designation that we deemed to better describe the reason for their fame (actor, in Schwarzenegger's case). The designations were also externally validated by checking that they matched the celebrity descriptions provided on Wikipedia. Overall, the dataset contained 10 actors, 9 comedians, 4 entrepreneurs, 6 politicians, 22 musician, and 9 TV personalities. A list of all the designations can be found in Appendix A.

The influence to buzz ratio and influence to audience ratios are both calculated on a per-celebrity basis and then averaged across celebrities with the same designation. They are calculated using the following formulas:

$$\text{Influence to buzz ratio} = \frac{\text{Avg. retweet count}}{\text{Total references}}$$

$$\text{Influence to audience ratio} = \frac{\text{Avg. retweet count}}{\text{Total followers}}$$

The influence to buzz ratio reflects the degree to which a celebrity is able to drive content on the Twitter network, relative to the degree to which he or she receives attention by being widely discussed. A high ratio indicates that an individual is perceived to be an interesting source of information, but that the person is not very interesting to discuss. A low ratio indicates that an individual is very interesting to talk about, but his or her content is not very interesting. One potential interpretation of this ratio is that it serves as a "conversion metric," indicating the degree to which celebrities are able to convert individuals who talk about them into individuals who retweet their content.

The influence to audience ratio gives a metric of how much a celebrity is able to drive content on the Twitter network, relative to the degree to which he or she is followed. A high ratio indicates that an individual is able to propagate his content widely despite having comparatively few followers; this can be interpreted to mean that the individual's content is particularly interesting or worthy, but that he or she has relatively few individuals who identify as explicit "fans." A low ratio could be interpreted to mean the opposite: that an individual has many self-identified fans but does not produce exciting or intriguing content. The influence to audience ratio can be interpreted to have a meaningful cardinal value (retweets per celebrity tweet per celebrity follower), though we use it only for ordinal comparisons here. The average value of each of these metrics for the six celebrity

groups are given below, in table 4.4:

Designation	Avg. Retweets	Influence to Buzz	Influence to Audience
Actor	347.73	0.0110	0.00012
Comedian	1015.16	0.0323	0.00036
Entrepreneur	767.64	0.0091	0.00044
Politician	291.88	0.0017	0.00091
Musician	1869.69	0.0031	0.00022
TV Personality	309.32	0.0055	0.00007

Table 4.4: Average metric values among the celebrities in each category

We see dramatic variation among these celebrity categories for each metric—and we see different celebrity categories dominating in each metric. Musicians are far ahead in average retweets; comedians are the clear winners in the influence to buzz ratio; and politicians lead in the influence to audience ratio. But what could be causing these discrepancies?

The dominance of musicians in average retweets may be reflective of the general popularity of musicians on Twitter. It is worth noting that all of the top five most followed individuals on Twitter are singers (Lady Gaga, Justin Bieber, Katy Perry, Rihanna, and Shakira), and that, though our group of 60 celebrities was selected irrespective of profession, more than a third of them are musicians. Given that singers seem to be an active and popular presence on Twitter, it might not surprise us that they attract more retweets in general. However, there are two notable reasons to be wary about interpreting a causal link between being a musician and attracting more retweets. First, the singers in our sample tended to be much younger on average (28.2 years) than the overall 60-person sample (41.3 years). Second, the presence of Justin Bieber in the musician set dramatically pulled up the average because, as we saw in figure 4.2, he is uniquely able to engender retweets relative to others in our dataset.

The influence to buzz and influence to audience ratios give us greater insight into the relationship between influence and profession. The above data is visualized in a bubble-chart in figure 4.4. The bubble sizes correspond to the average retweet count attracted by the group of celebrities.

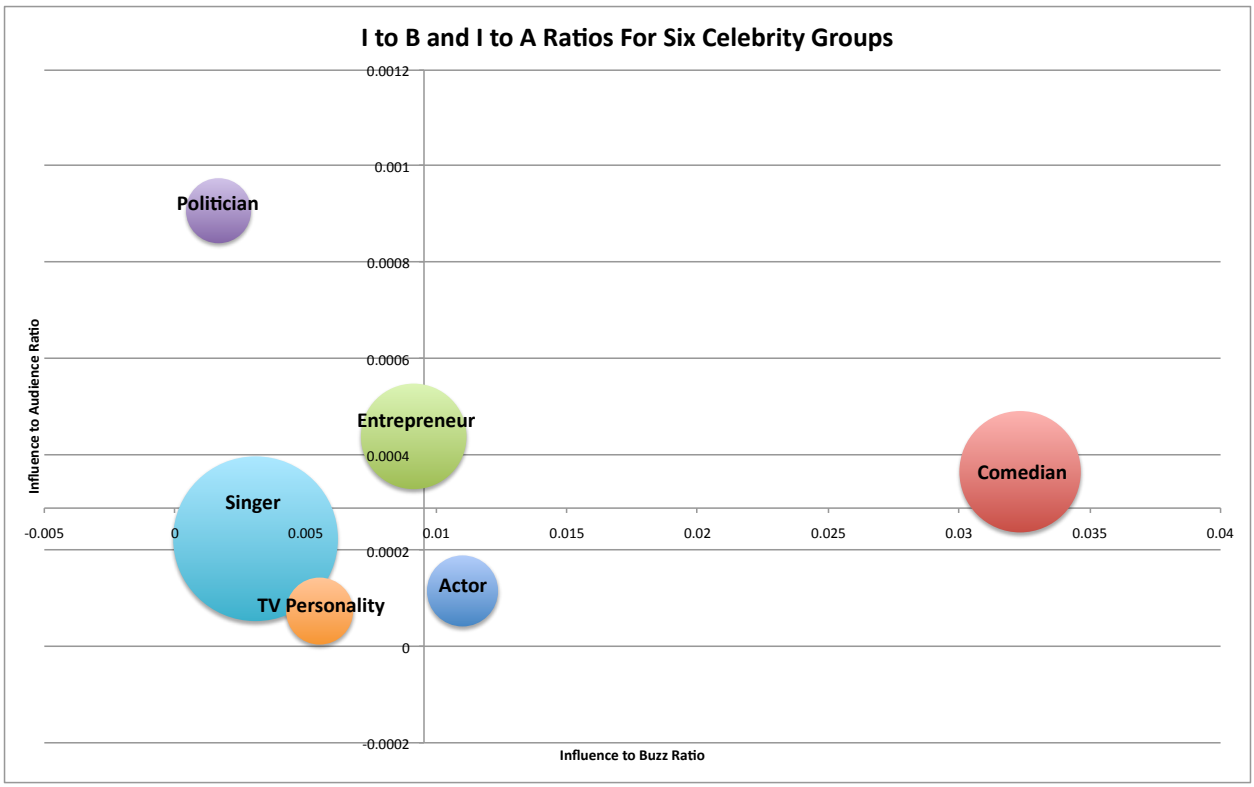

Figure 4.4: Celebrity groups are plotted according to their average influence to buzz and influence to audience ratios. Axes cross at the average value for each ratio.

Comedians clearly lead in the influence to buzz ratio, indicating that they exert exceptionally large influence relative to the degree to which they are discussed on Twitter. Many of the comedians in our dataset tended to tweet jokes and humorous musings which probably attracted retweets because other Twitter users found them funny. Indeed, one of the more heavily retweeted tweets in our dataset was from Conan O'Brien: "Thank God Beyonce had her baby and can go back to work. For the past 6 months that family's had to live entirely on Jay-Z's salary." In all likelihood, the humorous nature of these tweets made Twitter users much more likely to share them than a standard tweet. That the funniness of their tweets increases their "retweetability" means that comedians are less reliant on attracting buzz—which is, basically, attention—in order to garner retweets. At the other end of the spectrum, politicians appear to be highly non-influential relative to the amount they are discussed. This is likely because our data was collected during the period of the Republican presidential primaries, and most of the politicians in our dataset were seeking the Republican nomination. Thus, there was an extremely active conversation *about* these individuals on Twitter, but much of that discussion revolved around the politicians' campaigns and statements

to the media. Thus, many of the people talking about these politicians probably had no interest in reading or responding to the politicians' Twitter posts.

However, politicians came out significantly ahead in the influence to audience ratio, indicating that they exert a large amount of influence relative to their follower counts. We posit that this is due to two effects. First, the choice to a follow a celebrity on Twitter is often a reflection of fandom. Thus, since many Twitter users identify as fans of a singer like Lady Gaga, she is able to attract a large number of followers; but comparatively few people would identify as "fans" of a politician like Mitt Romney or Rick Santorum. This means that politicians may garner fewer followers relative to the public interest in their tweets, driving up their influence to audience ratios. Second, many of the politicians in our dataset only recently rose to national prominence—Rick Santorum and, to a lesser degree, Ron Paul and Mitt Romney were not household names prior to the 2012 Republican primaries. As a result, they may not have had adequate time to acquire followers commensurate with their prominence. This theory is bolstered by the fact that President Barack Obama, who has been widely known since the 2008 Democratic primaries, had a much lower influence to audience ratio than the other politicians in the dataset. If this "newcomer" effect indeed explains the discrepancy between Obama's influence to audience ratio and his Republican rivals, then it leads to an interesting implication: a high influence to audience ratio may be a good predictor of a future rise in follower count. This idea is explored in Chapter 6, when we construct a predictive model of increases in follower counts.

To show these effects at a higher resolution, we visualized the influence to audience and influence to buzz ratios for all 60 celebrities in figure 4.5. Bubble sizes again correspond to average retweet counts. Bubbles for each celebrity group are given the same color.

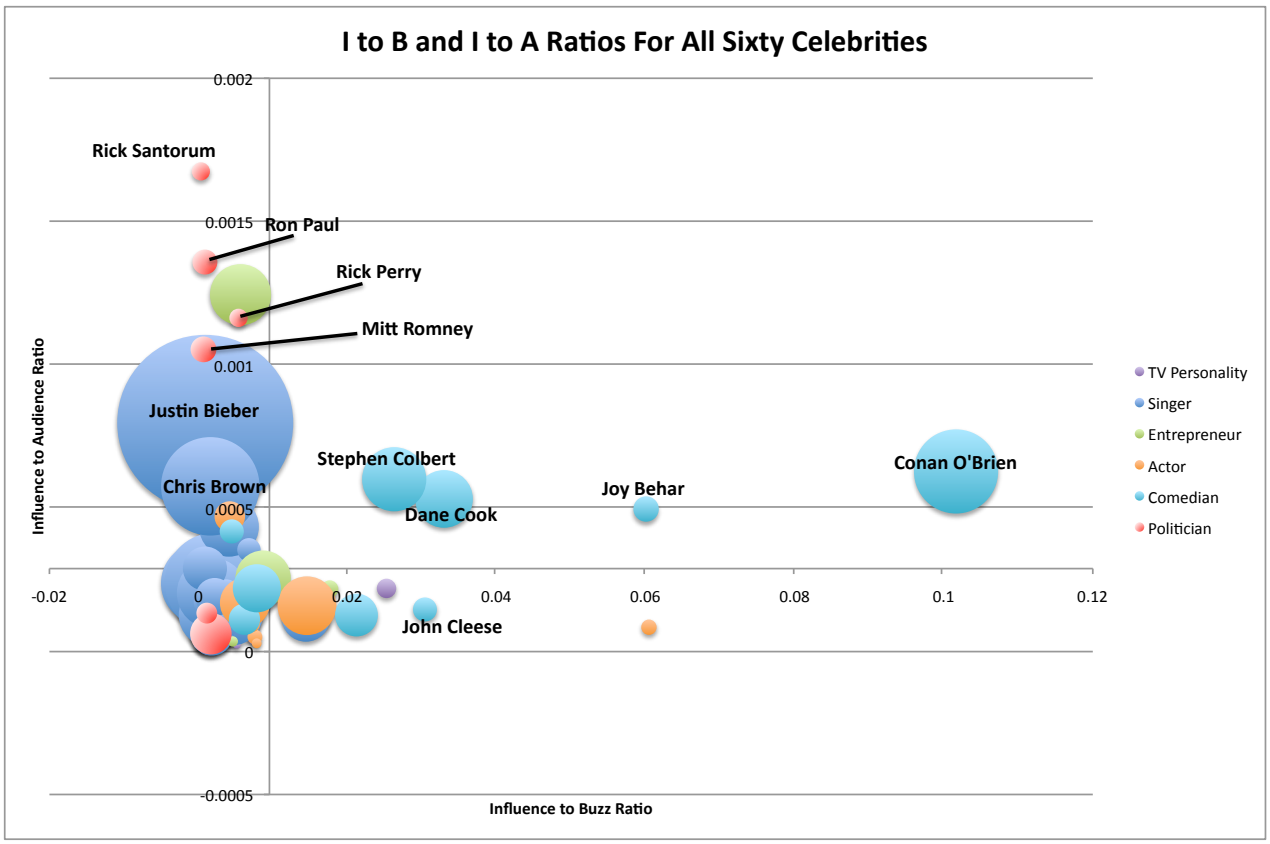

Figure 4.5: Individual celebrities are plotted according to their influence to buzz and influence to audience ratios. Again, axes cross at the average value for each ratio.

The above chart confirms that the effects seen in the averages are not due only to one or two outlier cases. We see that while Conan O'Brien is extremely dominant in influence to buzz ratio, most of his fellow comedians also have influence to buzz ratios that are significantly larger than average. Similarly, while Rick Santorum leads dramatically in influence to audience ratio, almost all of the other Republican contenders for the presidential nomination have above average influence-to-audience ratios. We also note that singers, actors, and TV personalities almost uniformly cluster together in the lower-left area of the graph, implying relatively little diversity in either ratio for these types of celebrities.

The overall interpretation of these charts is that celebrities toward the upper-right quadrant can engender more retweets—and thus exert greater influence—without having a large established fan base and without attracting a significant amount of buzz through being discussed by other tweeters. There are few celebrities who occupy this quadrant, and the ones who do are exclusively comedians, whose tweet content tends to be more enticing to share. Individuals toward the lower-

left quadrant require both large audiences and a lot of buzz in order to propagate their messages, and, unsurprisingly, this quadrant is quite well populated.

We might now naturally ask: do retweets capture the whole picture? Or can celebrities exert influence in ways other than engendering retweets among their fans? In the next section, we address this question.

Chapter 5

Non-Retweet Influence

Having investigated the dynamics of retweeting, we next move on to quantifying other ways that celebrities can exert influence on Twitter. We begin with a specific example from our dataset. In figure 5.1 below, the blue lines represent incidents in which our most-retweeted celebrity, Justin Bieber, sent tweets containing the hashtag "#callmemaybe," a reference to the song "Call Me Maybe" by pop star Carly Rae Jepsen. The green line represents the volume of the resultant retweets of Bieber's message, which peak shortly after each of Bieber's tweets and then quickly decline. In red, we see the volume of tweets which contain the hashtag "#callmemaybe" but are *not* retweets of Bieber's message. These messages follow a similar pattern to retweets.

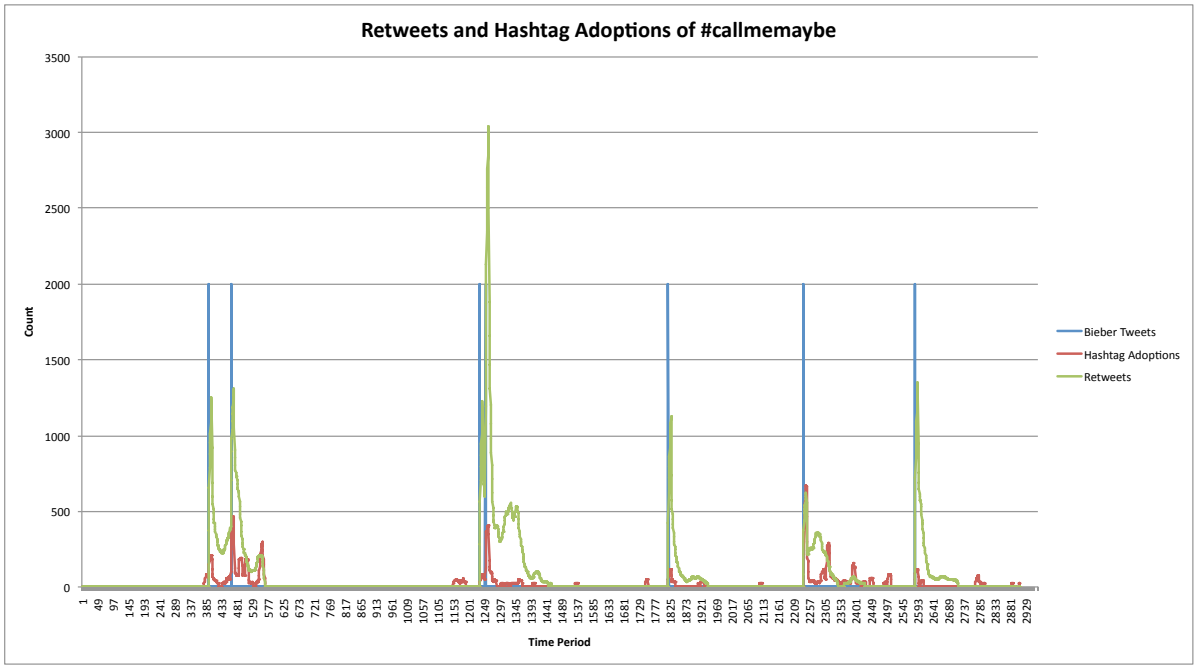

Figure 5.1: The blue line represents tweets by Bieber (vertical height is arbitrary). The red line represents hashtag adoptions and the green line represents retweets. Data smoothing is applied.

Though the volume of retweets is clearly larger in this example, it is important to note that the temporal pattern of non-retweets containing "#callmemaybe" strongly indicates that they, too, occur as a direct result of Bieber's tweets. If we were to solely quantify influence using retweets, we would miss this important effect.

More generally, we note that Twitter users need not explicitly retweet a celebrity to show that they have been influenced. Rather, if a celebrity sends a tweet, and another tweeter sends a message shortly thereafter that bears significant resemblance to the tweet content of the celebrity message (for example, containing an identical hashtag), we may suspect that the celebrity has influenced the other tweeter. If we see many tweets shortly after the celebrity's tweet which resemble the tweet in content, we can be increasingly confident that we are witnessing a real influence event.

In this chapter, we focus on these types of events, where a celebrity's moniker or message is *not* explicitly used, but the celebrity's effect can be inferred based on tweet content similarities. These events represent another type of behavior that fits our definition of influence ("*the ability to, through one's own behavior on Twitter, promote activity and pass information to others*"). Though there are many different types of non-retweet influence events, we focus on three particularly salient examples: adoption of a celebrity's hashtag or link, adoption of terms from a celebrity's tweets, and adoption of the emotional valence of a celebrity's tweet content.

Given the nature of these types of events, inferring the celebrity's impact is much more difficult than with retweets. For example, if a celebrity tweets a more general hashtag—say, "#FF", meaning "Follow Friday," a hashtag commonly used on Fridays—and another user tweets that same hashtag an hour later, can we know for certain that the celebrity has caused the other individual's tweet? Or is it possible that both individuals independently decided to tweet #FF?

Due to this ambiguity, we develop precise methodologies for how to interpret the relationship between a celebrity's tweets and the Twitter trends they may cause.

5.1 Local Adoption

In instances when celebrities tweeted a link or hashtag (the "item"), we make the following assumptions about the celebrity's tweet in relation to tweets of his or her followers:

1. The frequency of follower tweets containing the item over the two hours preceding the celebrity tweet is taken to be the baseline level of activity for the item in the follower population.

2. Any *increase* from the baseline in the frequency of follower tweets containing the item in the two hours after the celebrity tweet can be attributed to the celebrity.

3. Any *decrease* from the baseline in the frequency of follower tweets containing the item in the two hours after the tweet can be attributed to factors other than the celebrity.

These assumptions are obviously overly broad, and the third assumption particularly biases the sample toward demonstrating the desired effect. However, because declines from the baseline after the celebrity tweet were relatively rare (occurring less than 10% of the time in the case of hashtags and less than 3% of the time in the case of links), and because it is highly implausible that a follower would avoid tweeting something because a celebrity had tweeted it, we felt that this assumption was justifiable.

5.1.1 Definitions and Methods

We define the "local adoption effect" to mean the uptick in the frequency of followers tweeting an item after a celebrity has tweeted the item. We exclude all retweets of the celebrity from this analysis, so the local adoption effect is completely separate from retweeting as an influence metric. We calculated the effect for both hashtags and links tweeted by celebrities in our dataset. Our initial data contained 2281 unique celebrity-hashtag-timestamp triples and 2151 unique celebrity-link-timestamp triples tweeted by the celebrities from February 15 to March 16, 2012. Fifty-six of the celebrities in our dataset tweeted at least one hashtag, while 57 tweeted at least one link. However, in order to avoid overweighting outlier cases, we filtered out celebrities from each list who had tweeted less than five unique hashtags or five unique links; our final dataset thus consisted of 2245 celebrity-hashtag-timestamp triples representing 44 celebrities and 2129 celebrity-link-timestamp triples representing 48 celebrities.

The upticks were estimated using tweets from a group of each celebrity's Twitter followers sampled randomly by the Twitter API. Though the sample sizes were not uniform, they averaged about 350 followers, giving us a reasonably large sample to estimate the adoption rates in the celebrity's overall follower population. As defined here, the local adoption effect could loosely be interpreted as the per-follower increase in likelihood of tweeting an item given that the celebrity has tweeted it. One important caveat, however, is that the celebrity followers in our dataset all tweeted during our collection period, meaning they are at least moderately active users. Thus, a more

rigorous interpretation would be that the local adoption effect signifies the increase in likelihood that an *active* follower tweets an item, given that the celebrity has tweeted it.

5.1.2 Per-Follower Effect

The top 10 celebrities with the highest average local adoption effects are given in table 5.1 below:

Link			**Hashtag**		
Rank	Name	Effect Size	Rank	Name	Effect Size
1	Rick Perry	0.0025	1	Stephen Colbert	0.0324
2	Jon Favreau	0.0015	2	Felicia Day	0.0287
3	Felicia Day	0.0011	3	Chris Anderson	0.0275
4	Arnold Schwarzenegger	0.0010	4	Justin Bieber	0.0141
5	Travis Barker	0.0010	5	Newt Gingrich	0.0094
6	Robbie Williams	0.0010	6	Michael Ausiello	0.0079
7	Scooter Braun	0.0009	7	Arnold Schwarzenegger	0.0076
8	David Guetta	0.0008	8	Stephen Fry	0.0069
9	Shakira	0.0008	9	Scooter Braun	0.0065
10	Adam Savage	0.0008	10	Oprah Winfrey	0.0055

Table 5.1: Top ten individuals with highest average local adoption effect for links and hashtags.

A few observations are immediately obvious from these lists. First, the individuals who exert the greatest per-follower adoption effect are largely not the individuals who are the most-followed, most-retweeted, or most-referenced individuals in our dataset. In fact, none of the individuals who rank in the top ten under both of these metrics were consistently top-ten ranked under the more standard influence, audience, and buzz metrics used in the previous chapter. This is not entirely surprising, since the most-retweeted, most-followed, and most-mentioned individuals are not necessarily the ones who have the most loyal followers. But this also confirms that a model of influence that takes into account solely retweets, follower counts, or references would not accurately capture hashtag and link local adoption—and, more generally, follower loyalty.

Second, the hashtag and link local adoption are quite small in magnitude. Even the highest-ranked individuals can, on average, increase the likelihood that their active followers tweet a hashtag by one or two percentage points, and the likelihood that their active followers tweet a link by about a tenth of a percentage point. Altogether, in a dataset of more than 4,000 unique hashtag and link-tweeting instances, there were only 20 times (0.45%) when the percentage of active followers tweeting the hashtag or link increased by more than 10 percentage points in the two hours after a

celebrity tweet. This, too, is somewhat unsurprising. We have already noted that following relationships seem to be dictated by many factors other than interest in a celebrity's content, and it thus seems logical that only the most enticing links and hashtags would be likely to significantly capture the attention of a large swath of a celebrity's followers. Nonetheless, given the enormous follower counts of many of these celebrities, the number of hashtag adoptions could still be substantial even with a relatively modest per-follower effect. For instance, if Justin Bieber can, on average, get 1.41% of his followers to pick up a hashtag after he tweets it—and we assume that, similar to the overall Twitter population, about 50% of his followers are active [1]—this still translates to roughly 125,000 hashtag adoptions in the two hours after Bieber's tweet.

Third, the hashtag local adoption effect appears to be about ten times larger in magnitude than the link local adoption effect. Research conducted by Microsoft [8] found that links were much more commonly used on Twitter than were hashtags, so we might reasonably have expected the link local adoption effect to be substantially larger than the hashtag local adoption effect—not vice versa. However, the Microsoft research is three years old, and it is possible that common practices on Twitter have shifted since 2009. Regardless, it appears that if celebrities seek to propagate content, they are much better at getting their followers to adopt hashtags—perhaps due to hashtags' brevity and oft-humorous nature—than at getting their followers to adopt links.

The hashtag and link local adoption effects are also moderately well-correlated with one another. Analyzing the 41 individuals who tweeted sufficient hashtags and links to be included in both datasets, we find that the two effects have a correlation of about 0.56 under a log transformation. A scatterplot of the relationship is given in figure 5.2 below, where the variables have been multiplied by a constant and log transformed:

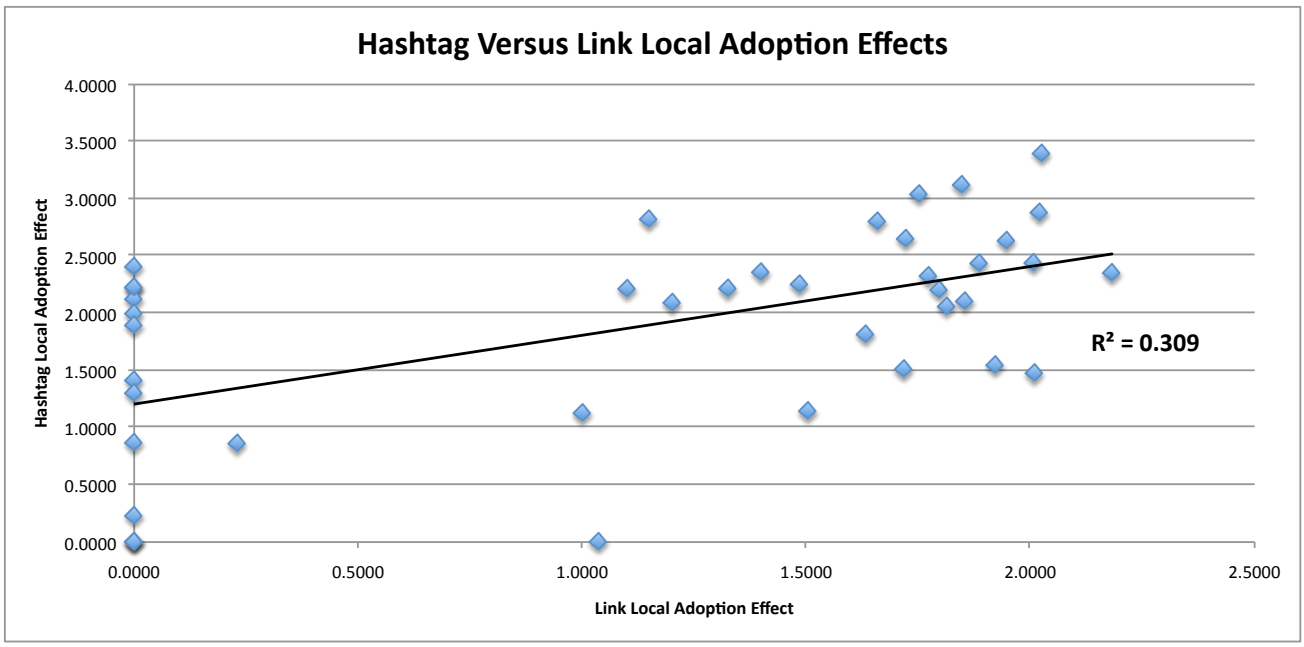

Figure 5.2: Hashtag and link local adoption effects for all celebrities who tweeted sufficient volume of both items. Note that the two effects are moderately correlated.

The moderate strength of this correlation implies that if a celebrity is good at getting his followers to pick up one type of content (e.g., links), the celebrity is generally also good at getting followers to pick up the other types of content (e.g., hashtags). But there do appear to be many individuals who are particularly well-suited to tweet one item or the other. Republican Presidential Candidate Newt Gingrich is one such example. Gingrich tends to send pithy hashtags—like "#250gas", referring to his plan to reduce national gasoline prices—which capture a single idea and are widely picked up by his followers. He also tweets many links to sites that promote his campaign and encourage his supporters to get involved. These links do not have the same shareable quality as Gingrich's hashtags, and he therefore ranks much higher in his hashtag-propagation ability than his link-propagation ability.

We also investigate the relationship between the hashtag and link local adoption effects and other variables we have previously calculated. The correlations of these variables with other variables in our data—as well as the associated significance value under a two-sided t-test—are given in table 5.2 and table 5.3. All variables have also been log-transformed:

Link Local Adoption Effect								
	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets	Tweets w/ Links	I to B	I to A
Correl	0.0909	-0.0641	0.1043	0.1798	-0.1509	0.0424	0.1135	0.2918
Sig	*0.5387*	*0.6652*	*0.4806*	*0.3059*	*0.1329*	*0.7745*	*0.4422*	***0.0442***

Table 5.2: Correlations and significance calculations for the link local adoption effect. The influence to audience ratio is significantly correlated, while all other variables are not.

Hashtag Local Adoption Effect								
	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets	Tweets w/ Hash-tags	I to B	I to A
Correl	0.066	-0.013	0.076	0.1683	-0.084	0.2044	0.0410	0.2020
Sig	*0.670*	*0.932*	*0.624*	*0.281*	*0.3211*	*0.1885*	*0.7940*	*0.1940*

Table 5.3: Correlations and significance calculations for the hashtag local adoption effect. No variables turn out to be significantly correlated.

These correlations lead to a number of intriguing observations. First, neither total retweet count nor mean retweet count is significantly correlated with either local adoption effect. Both retweet count and mean retweet count are aggregate metrics that are not calculated on a per-follower basis. Thus, the lack of a significant correlation here seems to fit with the narrative introduced at the beginning of this chapter: exerting a strong effect across all of Twitter does not seem to meaningfully indicate a high level of engagement with individual followers or fans.

Second, the total volume of tweets with links and tweets with hashtags is not significantly correlated with the local adoption effect. This is interesting, as it would make sense that individuals who tweet hashtags or links more frequently would be regarded as "expert sources" and might develop stronger local adoption effects, but this does not seem to be the case. It was also hypothesized that tweeting both too few and too many hashtags might have reduced the overall local adoption effect—an effect that would not have been captured by the correlation metric—but visual inspection of the scatterplots revealed no particular pattern between the local adoption effect and the frequency of tweeting hashtags or links.

Third, the only variable significantly correlated (at the 0.05 level) with either of the local adoption effects is the influence to audience ratio. However, this effect is significantly correlated with only the link local adoption effect and not the hashtag local adoption effect; the reason for this discrepancy is not immediately obvious. Regardless, it is quite intriguing that the influence to audience ratio is still a significant predictor of link adoption with retweets excluded from the

data, since it is not possible that the two metrics are measuring the same effect. With the caveat that the relationship was not significant for hashtags, this seems to indicate that the influence to audience ratio is a useful indicator of celebrity engagement with followers on an individual basis, and may serve as a useful predictor of certain types of Twitter behavior.

5.2 Co-Mention Adoption

In order to analyze the adoption of hashtags and links in a meaningful way, we would like to have a richer time-series representation of the hashtag and link adoptions that were caused by a celebrity. However, we face a problem in that, as more and more time elapses after a celebrity's tweet, we become less confident that hashtag and link adoptions by the celebrity's followers are a result of the celebrity's influence.

To address this problem, we take a very simple approach: if a celebrity sends a tweet containing a hashtag h or a link l, we assume that any subsequent tweet which contains either h or l, as well as an "@" mention of the celebrity, was caused by the celebrity. We call these events "hashtag co-mentions" and "link co-mentions." As with all heuristics for determining causality, there are clearly some conceivable counterexamples, in which a hashtag or link co-mention is not caused by a celebrity. Nonetheless, this approach has two significant advantages. First, it allows us to extend our conception of which events were caused by a celebrity across the entire data collection period. Second, hashtag or link co-mentions do not necessarily have to come from a celebrity's followers, so it gives us a more global sample of all the individuals on Twitter who are being influenced to adopt a hashtag or link.

As before, we begin by listing the top ten individuals ranked under this metric:

Rank	Name	Average Co-Mentions
1	Justin Bieber	21,853.13
2	Rihanna	1,973.16
3	Ron Paul	1,704.53
4	Chris Brown	1,578.85
5	Usher	1,350.39
6	Scooter Braun	1,084.62
7	Oprah Winfrey	1,037.96
8	Lady Gaga	764.58
9	Katy Perry	722.64
10	Taylor Swift	528.33

Table 5.4: Top ten individuals with highest average co-mentions.

Unsurprisingly, we see that Justin Bieber leads in this metric (as with many of the metrics that we calculate throughout this paper), attracting 11 times more co-mentions per tweet than his nearest competitor, Rihanna. Bieber's undisputed dominance appears to be a result not just of his rabid Twitter following, but also of a cleverly pursued marketing campaign on the part of Bieber's management. Bieber repeatedly tweeted the hashtags "#boyfriend" (the name of his upcoming single) and "#believe" (the name of his upcoming album) during our data collection period, and these hashtags appear to have caught fire not only due to anticipation of Bieber's new music, but also because both terms could be used in other tweet contexts (for example, "@justinbieber, will you be my #boyfriend?"). The simplicity of these hashtags and their potential usage in different contexts appears to have fostered widespread adoption which, in turn, allowed Bieber to broadly publicize his upcoming music. The time series of retweets and hashtag adoptions for "#believe" is given in figure 5.3 below:

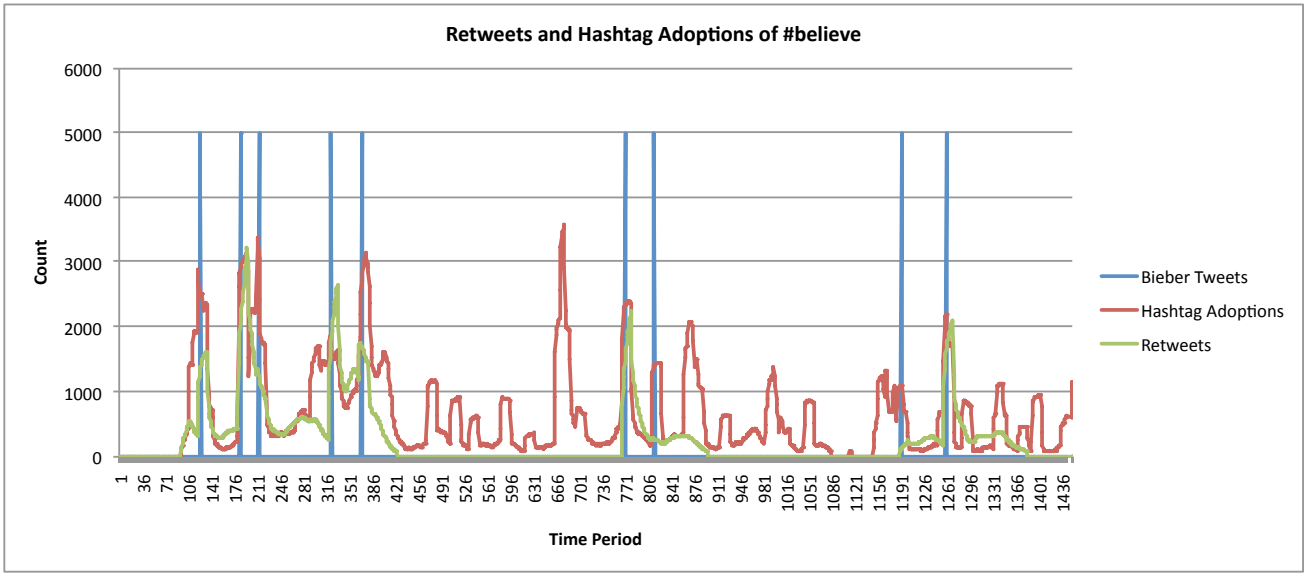

Figure 5.3: As in figure 5.1, the blue line represents tweets by Bieber (vertical height is arbitrary). The red line represents hashtag adoptions and the green line represents retweets. Data smoothing is applied. Note that hashtag adoptions tend to slightly outpace retweets.

It is clear that uses of "#believe" are responsive to Bieber's tweets, but the hashtag is also used heavily in periods when he has not recently tweeted it. Furthermore, non-retweet adoptions of the hashtag noticeably outpace retweets of Bieber's tweets containing the hashtag. This provides strong evidence that hashtag and link co-mentions are a significant component of celebrity influence.

Returning to the list of individuals with the highest average co-mentions, it is also noteworthy

that this list appears to contain many individuals who ranked highly under our influence, buzz, and audience metrics in Chapter 4. There are, however, a number of notable exceptions, including Ron Paul and Oprah. In order to investigate whether average co-mention count is significantly related to these variables, we calculate its correlations and present them in table 5.5:

	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets	Tweets W/ Links or HTs	I to A	I to B
Correl	0.3362	0.2231	0.7735	0.7368	0.2097	0.1983	0.2335	-0.0598
Sig	***0.0086***	*0.0866*	***0.0000***	***0.0000***	*0.1078*	*0.1287*	*0.0725*	*0.6497*

Table 5.5: Correlations and significance calculations for average co-mention count. Three variables turn out to be significantly correlated.

Three metrics are significantly correlated with average co-mention count: total references, total retweets, and mean retweet count. That total references are strongly correlated is not at all surprising. Since a co-mention is, by definition, a Twitter reference to a celebrity, it makes sense that individuals who attract more references overall would tend to attract more co-mentions. It may also be the case that as celebrities attract more buzz on Twitter, more Twitter users begin to read their messages and adopt their links and hashtags.

The strong correlations with total retweets and mean retweet counts are also expected, as individuals who can attract enough interest in their tweets to engender a large number of retweets can probably leverage that same interest in order to get Twitter users to pick up their links and hashtags. Despite the strong correlation, however, it is worth noting that only about 55% of the variation in average co-mention count can be explained using average retweet count. This indicates that if we model influence solely using retweets as our influence metric, we may undervalue individuals who are capable of engendering large numbers of co-mentions but relatively few retweets.

It is also interesting that a number of variables are *not* significantly correlated with average co-mention count. As with the local adoption effect, we see that the count of tweets with links or hashtags is not significantly correlated with average co-mentions. This seems to further underscore the point that tweeting more hashtags and links does not tend to make a celebrity an "expert source," but that other factors dictate the degree to which Twitter users pay attention to these hashtags and links. We see, also, that follower count is not significantly correlated with co-mention count (though, with a p value of about 0.09, it nearly meets the threshold). This is somewhat

surprising, as we could reasonably propose that celebrities with more followers would have more immediate readers of their tweets containing hashtags and links, and should thus see more resultant co-mentions containing these hashtags and links.

The composite story told by these correlations is that hashtag and link co-mention influence is similar in some ways to retweet influence, but there are also important differences between the two effects. This is indicated not just by the imperfect correlation between mean retweet count and mean co-mention count, but also by the fact that certain variables (follower count, reference count) are better correlated with mean retweets than they are with mean co-mentions. Furthermore, the two effects appear to be roughly comparable in size. Averaging first over tweets and then over celebrities, we find that the average retweet count in our dataset was 861, while the average co-mention count was 618. Co-mentions thus seem to be a meaningful factor in overall influence, and one that must be dealt with separately from retweets. This is precisely the approach we take when building predictive models of influence in Chapter 6.

5.3 Word Adoption

We now turn our attention to a more general way that a celebrity can exert influence without engendering retweets—by getting others to adopt her language. If a celebrity is very good at getting either her followers or her mentioners to use the same terms that she uses in her tweets, then she is exerting substantial influence through Twitter. In order to analyze this trend in the aggregate, however, we need a meaningful way to measure term similarity across tweets. The following procedure was used:

1. For each celebrity tweet, we filter out all stopwords (words with little lexical significance, such as "and" and "the"—see Appendix B for the full list) to create a set of unique words, W.

2. For whichever *tweet stream* we seek to analyze (either the follower tweets or the mentioner tweets), we apply the same stopword filtering. We define p_0 as the 15-minute period in which the celebrity tweet was sent. Then for the eight fifteen-minute periods prior to the celebrity's tweet (p_{-t}, $1 \leq t \leq 8$) and the eight fifteen-minute periods during and after the period of the celebrity's tweet (p_t, $0 \leq t \leq 7$), we create vector representations of the frequencies of each of the unique words in the tweet stream.

3. We define $f_t()$ as the function which returns a single word's frequency in the 15-minute period p_t. We calculate the baseline frequency of the terms in W, b_W, in the following way: for each period p_{-t}, $1 \leq t \leq 8$, we find the period-specific frequency $b_{W,t}$ according to:

$$b_{W,t} = \sum_{w_i \in W} f_t(w_i)$$

We then calculate the overall baseline frequency:

$$b_W = \frac{\sum_{i=-1}^{-8} b_{W,t}}{8}$$

4. For each period during and after the period of the celebrity's tweet, p_t, $0 \leq t \leq 7$, we calculate $d_{W,t}$, the difference from the baseline frequency in period t, as:

$$d_{W,t} = (\sum_{w_i \in W} f_t(w_i)) - b_W$$

And, lastly, we calculate the overall difference d_W as:

$$d_W = \sum_{i=0}^{7} d_{W,t}$$

It is worth noting that this mathematical definition of our effect size is one of several possible choices we could have made. We could, for instance, have calculated the L_2 norm distance between the celebrity's term frequency vector and the term frequency vectors from the tweet stream. However, there is a subtle difference between our analytical method and one that uses a distance metric: our analysis prioritizes events in which the use of any term by a celebrity increases the overall usage of that term in the tweet stream, while a distance analysis would instead have prioritized events in which the term frequencies of the tweet stream came more closely into alignment with those used by the celebrity. As a concrete example, suppose that for two hours, no one was tweeting about "Call Me Maybe," and Justin Bieber then sent a tweet about the song in which 20% of the words he used were either "call," "me," or "maybe." Suppose that after his tweet, 100% of the words tweeted by his followers were either "call," "me," or "maybe." Then, using a distance metric, we would find that the distance between Bieber's tweet and his followers' tweets had *increased* because the term frequency vectors would be so dissimilar due to the much more frequent use of "call," "me," and "maybe" among Bieber's followers than in Bieber's tweet. But this would obviously be

a case where Bieber's influence was extremely strong. To avoid such situations, we opted to use a term frequency measure instead. One important caveat, however, is that using our definition of effect size, we do not take into account shifts in the overall volume of tweeting activity over the four-hour analysis period.

To demonstrate this effect graphically, we select one of Conan O'Brien's tweets, sent on March 6th: "I've been practicing for this year's St. Patrick's Day. Every morning, I have my personal trainer punch me in the face." In the graph below, time periods are measured in 15-minute increments since the start of 2012. Conan's tweet was sent in time period 6386, and we can see that every word in the tweet experienced a frequency uptick among tweets from his mentioners, with the blue line representing the total frequency of all the words in the tweet. This is particularly striking given that we are excluding explicit retweets from this analysis. Visual inspection of tweets from this period revealed that many individuals sent what might be called "quasi-retweets"—tweets that contained almost the exact text of O'Brien's joke but were not actually retweets because they contained additional text, like "lol."

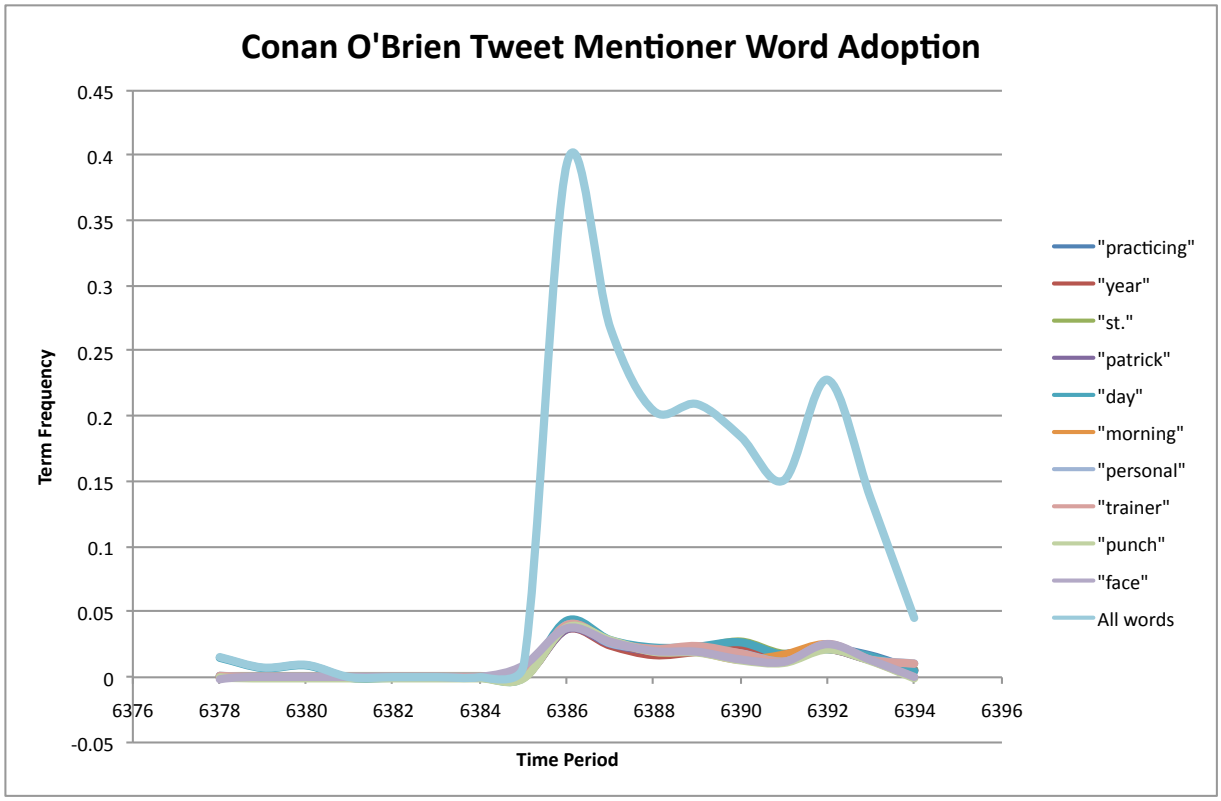

Figure 5.4: The blue line represents the summed frequencies of all the words in O'Brien's tweet. The other lines represent the frequencies of individual words. Note the uptick after O'Brien's tweet is sent in period 6386.

We calculated the d_W value for every celebrity tweet in our dataset which was not itself a retweet (there were just under 4,500 in total). We then filtered out any celebrity who had sent less than five non-retweet tweets during the collection period, and averaged the d_W scores for each celebrity in order to get an aggregate metric of word adoption. The top ranked celebrities under this metric are given below in table 5.6:

Followers			**Mentioners**		
Rank	Name	Effect Size	Rank	Name	Effect Size
1	Justin Bieber	0.0200	1	Conan O'Brien	0.2240
2	Joy Behar	0.0160	2	Chris Anderson	0.1640
3	Bill Gates	0.0121	3	Jon Favreau	0.1328
4	Stephen Colbert	0.0107	4	Joy Behar	0.1052
5	Lady Gaga	0.0104	5	Michael Ausiello	0.1015
6	Taylor Swift	0.0089	6	Dane Cook	0.1008
7	Rick Perry	0.0086	7	John Cleese	0.0928
8	Mitt Romney	0.0086	8	Kristin Cavallari	0.0890
9	Ron Paul	0.0079	9	Adam Savage	0.0886
10	Dane Cook	0.0077	10	Stephen Colbert	0.0873

Table 5.6: Top ten celebrities with highest average word adoption among followers and mentioners

A number of trends are immediately obvious from this data. First, the effect size appears to be about ten times larger among a celebrity's mentioners than a celebrity's followers. We find that this holds true across the entire dataset: the average effect size is 0.0484 among mentioners but just 0.0043 among followers. We posit that this is because an individual who mentions a celebrity is much more likely to be actively engaged with the content of the celebrity's tweets than the celebrity's average follower. Therefore, those mentioning a celebrity would be much more likely to read the celebrity's recent tweets and adopt some of the terms the celebrity used. This would also explain why, in Chapter 4, we found that a celebrity's reference count had a higher log correlation with his total number of retweets than did his follower count; if retweets, too, are a way that a Twitter user demonstrates engagement with celebrity tweet content, then having a large number of highly-engaged mentioners would, on average, engender more retweets than having a large number of passively-engaged followers.

We also see a number of intriguing discrepancies between the rankings for mentioners and followers. The top ten rankings share few individuals in common, with only Joy Behar, Dane Cook, and Stephen Colbert making both lists. Curiously, the follower rankings seem to favor both

politicians and comedians, with three celebrities of each type making the list. But the mentioner rankings favor almost exclusively comedians, with half the list made up of comedians. Lastly, the follower rankings include a number of individuals who ranked highly under our most-followed, most-referenced, and most-retweeted metrics; these include Justin Bieber, Lady Gaga, and Taylor Swift. The mentioner rankings, by contrast, include virtually no individuals who ranked highly under these more traditional metrics.

We investigate the discrepancy between the effect size among followers and mentioners by performing a log correlation. We find that the correlation is, indeed, startlingly low—just 0.16. A scatterplot of the values is given below, where both variables have been multiplied by a constant and log transformed to generate a uniformly positive, normal distribution:

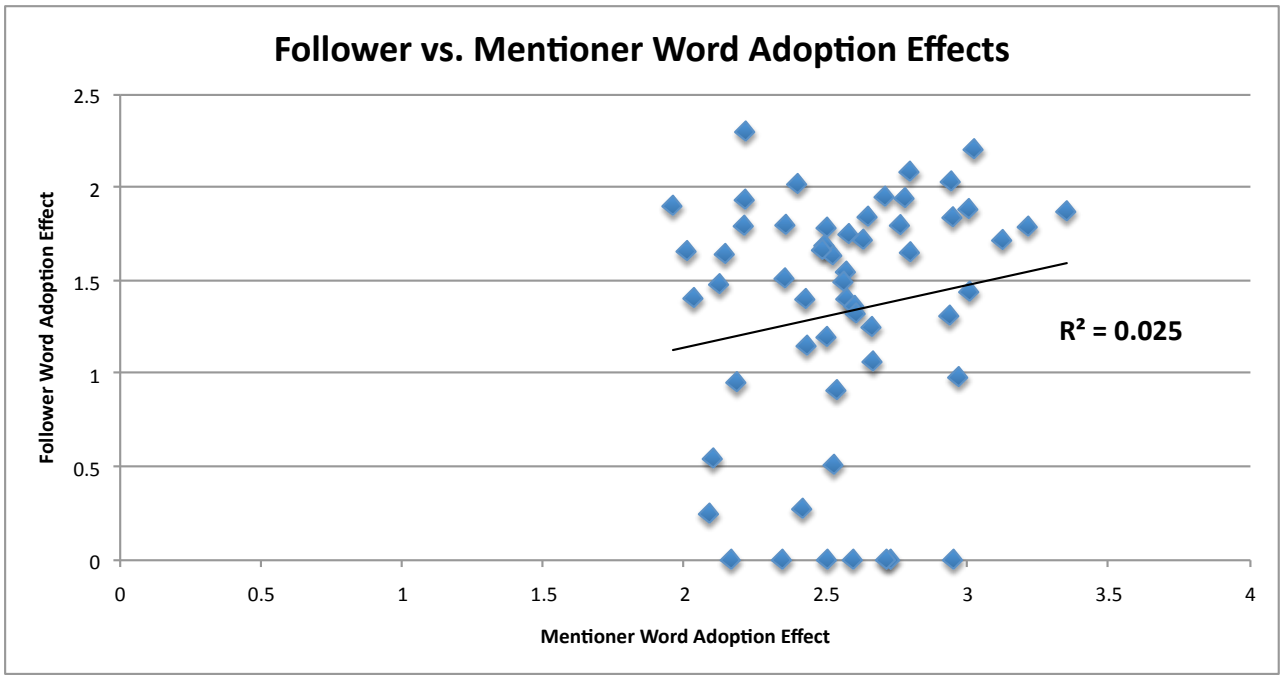

Figure 5.5: Points' positions represent follower and mentioner word adoption effects. Note the nearly random scatter, indicating very low correlation.

As before, we calculate the correlations with a number of our existing metrics to see what kinds of relationships emerge. We make the normal adjustments to ensure uniform positivity and normality. We also include *Comedian*, the binary variable indicating whether or not an individual is a comedian, and *Politician*, the binary variable indicating whether or not an individual is a politician, based on the trends we saw among the top ten rankings. Correlations are given below in tables 5.7 and 5.8:

Follower Word Adoption									
	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets	I to A	I to B	Comedian	Politician
Correl	0.2485	0.0022	0.3385	0.4365	-0.1278	0.5367	0.0918	0.1437	0.1900
Sig	*0.0648*	*0.9869*	***0.0107***	***0.0007***	*0.3481*	***0.0000***	*0.5010*	*0.2908*	*0.1608*

Table 5.7: Correlations and significance calculations for follower word adoption. Three variables turn out to be significantly correlated

Mentioner Word Adoption									
	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets	I to A	I to B	Comedian	Politician
Correl	-0.5566	-0.0177	-0.2129	-0.0127	-0.2961	-0.2838	0.7138	0.4429	-0.3702
Sig	***0.0000***	*0.8970*	*0.1153*	*0.9258*	***0.0267***	***0.0341***	***0.0000***	***0.0006***	***0.0050***

Table 5.8: Correlations and significance calculations for mentioner word adoption. Six variables turn out to be significantly correlated

These correlations show a number of discrepancies between the effect on followers and the effect on mentioners, and seem to indicate that the two processes are governed by different dynamics. Follower word adoption is significantly correlated with both mean and total retweets, as well as the influence to audience ratio. This is quite surprising, since all of these effects are derived from retweet counts, and the analysis of follower word adoption excludes retweets. However, we note that the most significant correlation is with the influence to audience ratio, which we have proposed as an indicator of the level of follower engagement with a celebrity. We therefore posit that a high level of follower word adoption, like a high influence to audience ratio, is the result of follower loyalty to a celebrity, and that this explains the strong correlation between the variables.

Mentioner word adoption, by contrast, is positively correlated with the influence to buzz ratio as well as *Comedian*, and negatively correlated with total references, tweets sent, *Politician*, and the influence to audience ratio. The correlation with the influence to buzz ratio makes intuitive sense. We have previously proposed that the influence to buzz ratio be interpreted as a conversion metric, indicating a celebrity's ability to get those who mention him to retweet his content. However, an alternative interpretation is that, in the same way that the influence to audience ratio reflects engagement with followers, the influence to buzz ratio reflects engagement with a celebrity's *mentioners*. If this interpretation is correct, then the high correlation indicates that well-engaged mentioners are both more likely to retweet a celebrity and more likely to adopt the celebrity's words. We hypothesize that the positive correlation with being a comedian is a result of the fact that indi-

viduals are more likely to adopt the comic terms used by celebrity comedians in their tweets; and the negative correlation with being a politician is a result of the fact that individuals are less likely to adopt the kinds of campaign-promotional terms often used by politicians in their tweets. The strong negative correlations we see with total references and with tweets sent appears to be a result of the fact that these variables happen to be negatively correlated with the influence to buzz ratio. We therefore doubt that any independent relationship is indicated by these correlations.

As a final check on this analysis, we compare the word adoption effect against our other metric of follower engagement: the hashtag and link local adoption effects. We have hypothesized that the follower word adoption effect is driven mainly by follower engagement with a celebrity, and that the mentioner word adoption effect is driven by different dynamics. Under this hypothesis, we would expect to see a positive correlation between the hashtag and link local adoption effects and follower word adoption, but *not* to see a positive correlation between the hashtag and link local adoption effects and mentioner word adoption. This is precisely what we see: the hashtag and link local adoption effects exhibit positive, significant correlations with follower word adoption of 0.2990 and 0.3762, respectively. In contrast, the hashtag and link local adoption effects are not significantly positively correlated with mentioner word adoption.

5.4 Emotional Transmission Analysis

The last non-retweet influence effect we investigate is the transmission of emotion from celebrities to others on Twitter. We hypothesize that individuals who are best able to encourage their followers and mentioners to adopt their emotional states may be particularly well engaged with these groups.

At the core of our analysis is the idea that the emotional intensity of a tweet can be scored, and that transmission occurs when the emotional intensity of a celebrity's tweets is a forecaster of the emotional intensity of followers or mentioners. Emotional scoring was done using AFINN, a labelled word list which assigns emotional intensity scores of -5 to +5 to nearly 2,500 different words (see Chapter 3 for a full explanation). Our analysis methodology is described below:

1. For each celebrity C, for each 15-minute time period in the data t, a set of word-frequency pairs, $(w, f) \in WF_{C,t}$, was constructed, where w represents a unique non-stopword tweeted by the celebrity in t and f represents the number of times the word was used in t. Defining $V()$ as the function that takes in a word and returns its emotional valence score, the overall

valence value for each celebrity for each time period, $v_{C,t}$, was calculated as follows:

$$v_{C,t} = \frac{\sum_{(w,f)\in WF_{C,t}} V(w) * f}{\sum_{(w,f)\in WF_{C,t}} f}$$

2. The same methodology was applied to the tweet stream of interest (the follower stream or mentioner stream) after retweets were filtered from the stream. Thus, for each celebrity C and tweet stream S, we generated values $v_{C,S,t}$ for each period t.

3. A time series v_C was assembled as $[v_{C,1},v_{C,2},v_{C,3},...]$ and a time series $v_{C,S}$ was assembled as $[v_{C,S,1},v_{C,S,2},v_{C,S,3},...]$

4. A Granger Causality analysis, using up to 10 lags for each time series, was performed to see if v_C was a meaningful forecaster of $v_{C,S}$ (see chapter 3 for a full explanation of Granger Causality Analysis). If v_C was found to Granger-cause $v_{C,S}$, then the strength of the relationship was scored according to the positive change in the R^2 value of a regression model of $v_{C,S}$ once the lags of v_C were added to the model. If v_C was *not* found to Granger-cause $v_{C,S}$, then the strength of the relationship was scored as zero.

This technique has a few important caveats. First, if this analysis method returns a nonzero value, this technically only means that the celebrity's valence series can be used to *forecast* changes in the followers' or mentioners' valence series—not that it *caused* changes in the series. In the aggregate, we assume that forecasting ability corresponds to emotional transmission from celebrities to followers or mentioners, but more in-depth analysis would be required to confirm that this was actually the case.

Second, we use an extremely simple method of valence-scoring and, as a result, lose much of the rich textual content of individual tweets. A number of more accurate methods of valence scoring (text-analysis software, the use of human scorers through Amazon Mechanical Turk) are possible, but were prohibitively expensive. We again assume that, in the aggregate, our method provides a useful approximation of the emotional valence content of individual tweets.

We now turn our attention to the top ten ranked individuals for emotional transmission to both followers and mentioners. Again, the effect size is defined as the increase in R^2 for a regression model of the tweet stream's emotional intensity once the lags of the celebrity's emotional intensity

time series are added to the model. The rankings are given in table 5.9:

Followers			**Mentioners**		
Rank	Name	Effect Size	Rank	Name	Effect Size
1	John Cleese	0.0071	1	Keri Hilson	0.0459
2	Snooki	0.0039	2	Kim Kardashian	0.0219
3	Alicia Keys	0.0037	3	Soulja Boy	0.0209
4	Rick Perry	0.0031	4	Paris Hilton	0.0197
5	Dianna Agron	0.0030	5	Michael Ausiello	0.0171
6	Dita Von Teese	0.0024	6	Britney Spears	0.0146
7	Britney Spears	0.0024	7	Chris Brown	0.0134
8	Michael Ausiello	0.0022	8	Jordin Sparks	0.0106
9	Barack Obama	0.0021	9	Rihanna	0.0097
10	Kirstie Alley	0.0019	10	Alicia Keys	0.0078

Table 5.9: Top ten celebrities with the highest average emotional transmission to their followers and to their mentioners

As before, we see that the effect size appears to be substantially larger among mentioners than among followers. We confirm this by comparing the overall data means, finding that the average emotional transmission effect is about eight times bigger among mentioners than among followers.

We see also that very different individuals appear to rank highly on the two lists, with the top ten rankings sharing only three celebrities—Michael Ausiello, Alicia Keys, and Britney Spears. Both lists appear to slightly favor women (six women appear among the top ten in the followers list and seven women in the mentioners list, but only 40% of the 60 celebrities in our dataset are female). It is possible, however, that this is due to random chance. The followers list also appears to show no particular bias toward celebrities of a particular designation, but the mentioners list strongly favors musicians, as seven of the top ten ranked individuals are musicians.

Further supporting the notion that emotional transmission from celebrities to followers is very different from emotional transmission from celebrities to mentioners, we find that the correlation between the two effects is extremely low (0.12). However, we note that the celebrity-to-follower emotional transmission effect was calculated to be zero for all but 13 of the celebrities, while the celebrity-to-mentioner transmission effect had a nonzero value for 46 of our celebrities. We thus posit that, in this case, the apparent discrepancy may be due to a failure to accurately detect

emotional transmission in our follower data, which is relatively sparse compared to our data on mentioners. Because of this sparsity, we restrict the remainder of our analysis to the celebrity-to-mentioner emotional transmission effect.

We now correlate the mentioner emotional transmission effect with several of our existing variables. As before, we create a handful of variables based on our preliminary analysis to see if they are also correlated with the effect. We include $Musician$, the binary variable indicating whether or not someone is a musician; $Female$, the binary variable indicating whether or not someone is female; and $Total\ Valence$, the sum of all the absolute values of the valence scores from a celebrity's tweets throughout the collection period. The correlations are given in table 5.10:

Mentioner Emotional Transmission					
	Total Refs	Followers	Total RTs	Mean RTs	Total Tweets
Correl	0.1692	0.0747	0.3976	0.2039	0.4435
Sig	*0.1963*	*0.5705*	***0.0016***	*0.1180*	***0.0004***
	I to A	I to B	Singer	Female	Total Valence
Correl	0.0718	0.0575	0.2492	0.2724	0.4846
Sig	*0.5855*	*0.6624*	*0.0548*	***0.0352***	***0.0001***

Table 5.10: Correlations and significance calculations for mentioner emotional transmission effect. Four variables turn out to be significantly correlated.

We see that the four variables significantly correlated with the mentioner emotional transmission effect are total retweets, total tweets, $Female$, and $Total\ Valence$. A critical analysis of these correlations indicates that several of them may be illusory. Given that our best-correlated variable is $Total\ Valence$, it appears that those who use more emotionally charged language tend to be better able to transmit their emotions to their mentioners. But this variable, since it is not normalized by tweets sent or words sent, is also correlated strongly with the total volume of content generation—and is thus strongly correlated with both the number of tweets sent and total retweet count. Among these relationships, then, we hypothesize that the "real" relationship is with total valence score and that the other two, weaker correlations are not actually indicative of a meaningful relationship.

The one other statistically significant relationship is with $Female$, the binary variable indicating whether or not an individual is female. This variable does not turn out to be significantly correlated with $Total\ Valence$, so we have more reason to believe that its correlation with the follower emotional transmission effect is actually meaningful. However, it would be wrong to assume

that individuals are better at transmitting emotional states to their followers *by virtue of being female*. A number of explanations are possible, including that the mentioners of female celebrities are more likely to reflect the celebrities' emotional states or that both female celebrities and their mentioners are more likely to adopt emotional states that are trending on Twitter.

Lastly, we compare this effect against the variables we have calculated throughout this chapter to see if it correlates strongly with any other non-retweet influence metrics. The only statistically significant correlation we see is with the mentioner word adoption effect calculated in Chapter 5.3. This is not at all surprising; if the mentioners of a celebrity tend to pick up the celebrity's words in general, they will probably tend to pick up their emotionally charged words as well. If this is case, the time series of emotional valence in their tweet content tend to resemble the celebrity's valence time series and they will score highly for emotional transmission.

Chapter 6

Predictive Models

Using the data we have gathered and the variables we have generated, we now seek to create predictive models of various Twitter influence effects. As we have seen, celebrities can exert influence both by generating large numbers of retweets and thus garnering many views for their messages; or by other, more subtle effects wherein their hashtags, links, words, or emotional states are transmitted to others on Twitter. We have seen, also, that retweets are moderately correlated with some of these subtler transmission effects, but they are not perfectly correlated with the presumed largest effect: hashtag and link adoption. As a result, we model these two influence effects separately. We also create a separate model of celebrities' proportional upticks in follower counts in order to gain some insight into how influence may evolve over time.

In the creation of these models, we place ourselves in the shoes of a marketing firm paying a celebrity to tweet about a particular product or concept. We are particularly interested in the question, "How much is a tweet by a celebrity worth?" Though we do not directly calculate dollar values for tweets, we do drive toward predicting the total number of views ("impressions") that will be garnered by the average tweet sent by a celebrity. In our models, we test characteristics of the celebrities—as well as features of the tweet content that they produce—as potential predictor variables. The full list of candidate predictors can be found in Appendix C.

6.1 Methodology

The modeling tool we use is step-up least-squares multiple linear regression, and we validate our models by using the leave-one-out cross-validation technique [25]. Cross-validation is a method used to prevent over-fitting bias when estimating the residuals of a model [36]. We utilize the

method as follows: defining n to be the number of data points in our sample (in this case, the data points represent celebrities, but not all 60 celebrities are used due to outlier cases), we define distinct, unique partitions p_i, $1 \leq i \leq n$ of the dataset. Each partition p_i consists of a training set T_i, which contains $n-1$ data points, and a test set s_i, which contains a single data point. In each partition, one unique data point is selected for the test set s_i.

For each partition, we train the model on T_i and test its predictive value on s_i. We collect the error from this test (the difference between the model's predicted value and the actual value), then advance to the next partition and repeat the process. Overall, we report the error of our model by calculating the root mean square error (RMSE) using the n distinct error terms, one from each partition.

A number of general ordinary least squares linear regression modeling techniques are adapted for the model construction here. The step-up regression technique requires that we add variables sequentially to the model based on which variable's addition leads to the largest increase in the model's coefficient of determination, R^2. However, since we generate n different models at each step in our modeling process, we instead determine the next addition using the average R^2 uptick over the n models. Furthermore, since cross-validation is used to validate a model, rather than to construct it, it is technically possible that a predictor variable will be significant in one fold and not another. However, because we have removed outliers from the dataset and because the log-transformed data is not excessively variant, we find at each step in our model construction that the *best* subsequent variable addition is uniformly significant in all of our folds. As a result, we report the results of our cross-validation analysis at each step in the model construction.

In order to make our models future-predictive, we divide our dataset into two time periods—February 15 to March 3 (period 1) and March 4 to March 16 (period 2)—and train our model on data from the second period, using predictor variables calculated from the first period.

6.2 Retweet Impressions Predictive Model

We begin by modeling the average number of retweet impressions garnered by each celebrity in period 2. A retweet impression is defined as a single view by a Twitter user of a celebrity tweet that has been retweeted. We calculate this metric under the broad assumption that a Twitter user will eventually look at any tweet posted by someone the user follows. This is technically untrue,

because some Twitter users are completely inactive, but provides a reasonable approximation of the total number of views that any one tweet will receive.

We define $f()$ as the function which, given a Twitter user as input, returns her number of followers. Then, for a tweet t retweeted N times, we define U_j to be the user who retweeted t the j^{th} time with $1 \leq j \leq N$. Then, the total number of retweet impressions for this tweet, I_t, is given by the summation:

$$I_t = \sum_{j=1}^{N} f(U_j)$$

Defining a celebrity's set of tweets as T and the cardinality of T as $|T| = k$, we now define our variable of interest, $Avg(I_{retweet})$, as:

$$Avg(I_{retweet}) = \frac{\sum_{t \in T} I_t}{k}$$

We calculate this value for period 2, denoted $Avg(I_{retweet,2})$, and log transform it in order to achieve normality.

For this analysis, we also throw out four celebrity outliers: Jim Carrey, Beyonce, Eminem, and Justin Timberlake. Each of these celebrities failed to tweet in either one or both of the periods, meaning that we could not assign them a meaningful value for either the predicted variable or most of the predictor variables (or both). As a result, we excluded these celebrities from our analysis.

6.2.1 Model Construction

We consider a wide array of variables (calculated for period 1) as potential predictors in our model. All tested variables can be found in Appendix C.

We find that the largest average R^2 value from a single predictor is given, unsurprisingly, by the period 1 average retweet impression count: $log(Avg(I_{retweet,1}))$. A model containing solely this variable has an average R^2 of 0.572. The model's RMSE is 0.524, for a predicted variable whose mean value is 5.28. However, as we add variables to the model, we find that a more predictive model can actually be created using the combination of the total references to the individual in period 1, $log(tr_1)$ and the individual's influence to buzz ratio from period 1, $log(itob_1)$. Adding these two variables to the model actually "kicks out" $log(Avg(I_{retweet,1}))$ as a significant predictor (meaning its significance as a regressor no longer meets the 0.05 threshold), but the new model has an average R^2 value of 0.697 and a lower RMSE value of 0.445. We also test the addition of the

$log(itob_1)$ variable to a model containing solely $log(tr_1)$ and confirm, via an F-test on the average residual sum of squares values across the 56 program iterations, that the addition of the second predictor variable is statistically significant. As a result, we opt to begin with the two-variable model as our initial model. The predicted vs. actual values for each model are given in figures 6.1 and 6.2:

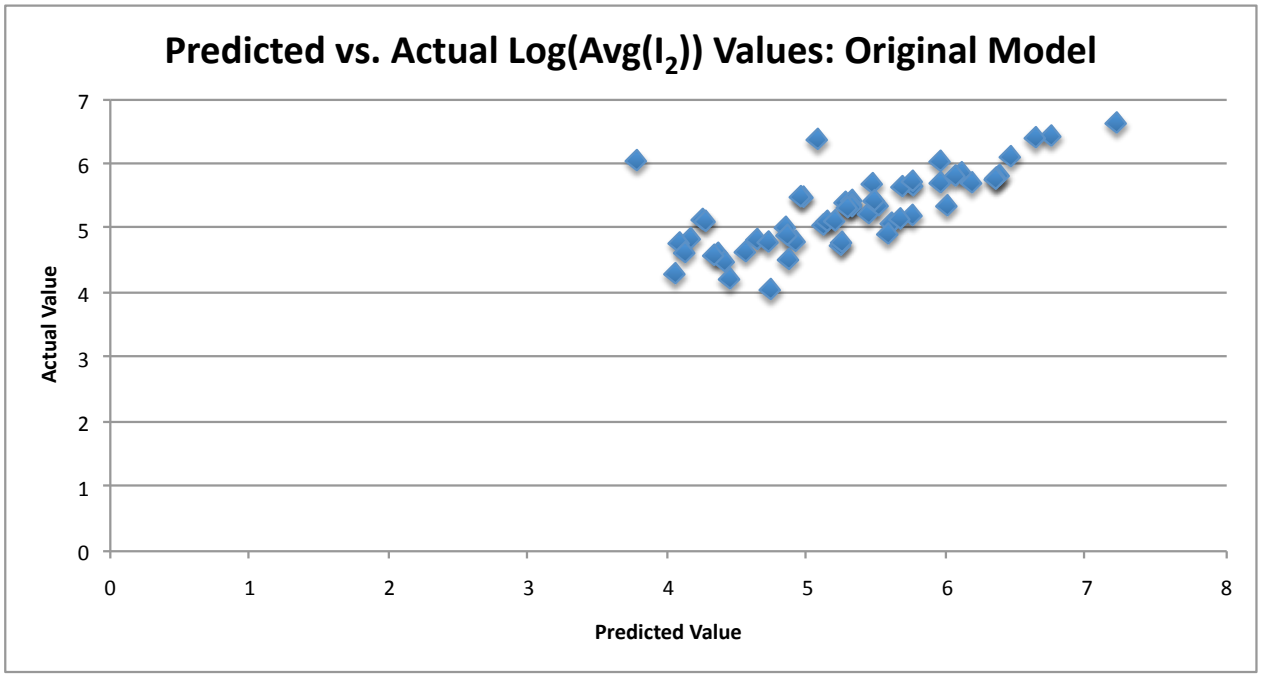

Figure 6.1: Predicted vs. actual values for model of retweet impressions, containing solely the period 1 average retweet impression count as a predictor

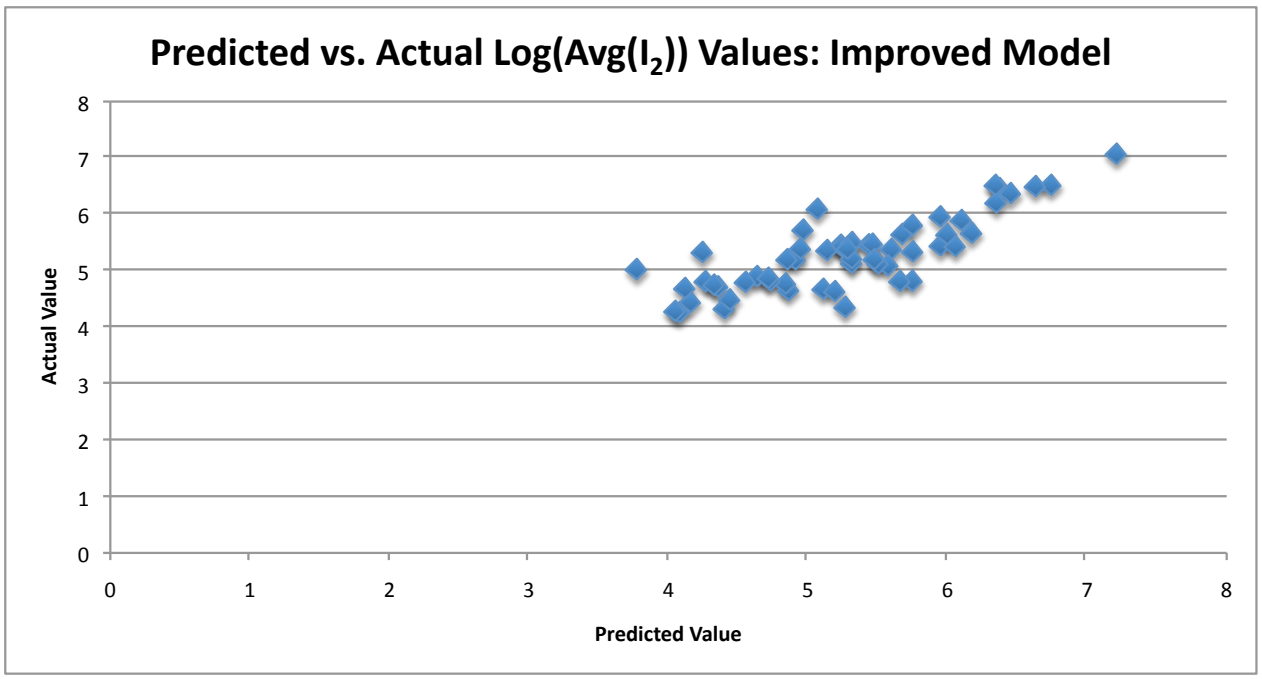

Figure 6.2: Predicted vs. actual values for model of retweet impressions, containing period 1 reference count and influence to buzz ratio as predictor variables

We now seek to further improve the model by the addition of new predictor variables. After testing all of our candidate variables, we find that only one improves the predictivity of our model: the average period 1 word adoption effect among a celebrity's mentioners (wa_1). This metric is a plausible predictor, because it seems to reflect a celebrity's level of engagement with his mentioners. If mentioners tend to pick up a large number of a celebrity's words, that means that they tend to read and be influenced by the celebrity's Twitter content. Other things being equal, we would assume that celebrities better engaged with their mentioners would also be able to generate more retweets.

This addition to the model increases the average R^2 value to 0.730, though an F-test reveals that the p-value associated with this addition is only 0.16, notably above our standard cutoff of 0.05. However, since we are focused primarily on improving our model's predictivity, we opt to include the variable because it lowers the RMSE slightly to 0.437.

The coefficients for each variable—with their 95% confidence intervals across the 56 program

iterations—are given in the equation below:

$$log(Avg(I_{retweet,2})) = (-0.241 \pm 0.101) + (0.966 \pm 0.018) * log(tr_1)$$
$$(0.511 \pm 0.069) * log(itob_1) + (5.050 \pm 1.414) * wa_1$$

The scatterplot of predicted versus actual values for the final model is given in figure 6.3:

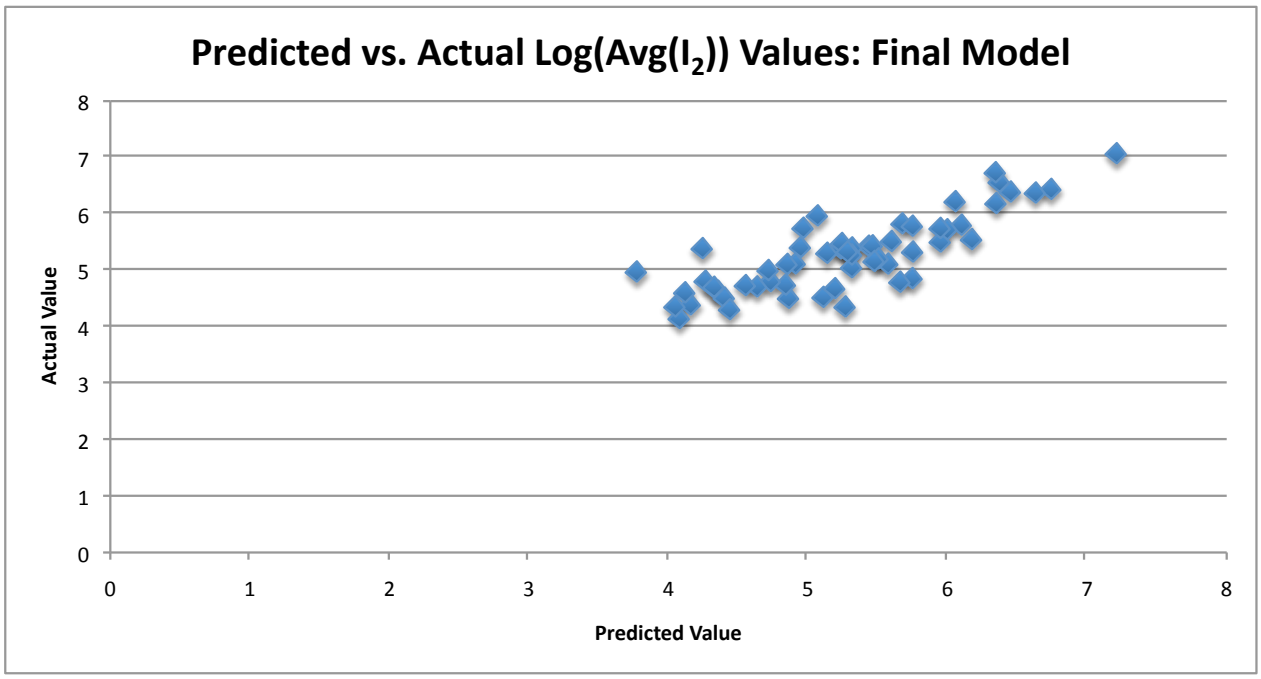

Figure 6.3: Predicted vs. actual values for model of retweet impressions, containing period 1 reference count, influence to buzz ratio, and average mentioner word adoption as predictors

We see that the points adhere quite well to the line $y = x$, with relatively little scatter. This demonstrates that the model is quite predictive.

As a final check of our model's validity, we investigate which individuals' $log(Avg(I_{retweet,2})$ were most poorly predicted by the model. In order, we find that the model was most inaccurate when predicting the value for Danny Glover, Daniel Tosh, Chris Anderson, John Cleese, and Usher. This group appears to be somewhat random—containing individuals representing four different professions who exhibit diversity in the predicted and predictor variables as well as follower count and Twitter activity. This makes us more confident that our model is not somehow systematically biased.

6.2.2 Discussion

This model provides a number of novel observations about the dynamics of Twitter influence, as well as a useful starting point for generating a predictive model to be used in marketing settings.

The significance of total references in period 1 is perhaps the most intriguing and salient aspect of the model. As we recall, an initial hypothesis for this paper was that a celebrity's total reference count was not strongly predictive of influence. But, as we now see, reference counts are an extremely strong predictor of future retweet impressions, which are an enormous component of celebrity influence. It appears that heavy discussion of an individual on Twitter does indicate a strong interest in reading and retweeting that person's messages. Furthermore, the interest in a celebrity's messages indicated by a high reference count is not a transitory or evanescent interest; rather, it appears to last long enough that the reference count over a two-week period can be used to meaningfully predict retweet impressions over the subsequent two-week period.

The other two metrics in the model—the influence to buzz ratio and the average mentioner word adoption effect in period 1—can best be interpreted as "conversion" and "engagement" metrics respectively. The influence to buzz ratio (which, as we recall, is calculated as a celebrity's average retweet count divided by his reference count) gives us a good indicator as to whether a celebrity can efficiently "convert" his buzz into retweets. Individuals with a high ratio may be highly capable of leveraging their buzz in order to increase overall readership of their tweets, and this aspect of a celebrity may be comparatively time-stable. If this is the case, it would provide a meaningful explanation as to why reference count and influence to buzz ratio can together produce a better model of period two retweet impressions than a model based on period 1 retweet impressions. Lastly, the mentioner word adoption metric, we have previously hypothesized, is a measure of how engaged an individual's mentioners tend to be with the individual's content. Since the majority of a celebrity's mentioners are likely also potential retweeters of the celebrity, it makes sense that being well-engaged with these individuals would give a celebrity a boost in his ability to generate retweet impressions.

With greater refinement, this model could be adapted to provide extremely robust predictions of the average number of retweet impressions expected to be generated by celebrity tweets over a given time window. It is important to note that, as discussed in Chapter 4.3, there is high variance in retweet count—and, by extension, retweet impression count—even among a single celebrity's tweets.

Thus, this model would be a less accurate predictor of the exact number of retweet impressions that would be generated by an individual tweet. Nonetheless, by providing an estimate of the average expected count, the model could provide useful information to marketers, campaigners, and the media in the quest to find individuals to propagate an advertisement or message.

6.3 Non-RT Hashtag and Link Impressions Predictive Model

Now that we have developed a predictive model of retweet impressions, we seek to model alternative types of influence. In chapter 5, we found that the average number of impressions generated by the non-retweet propagation of hashtags and links appeared to be comparable in size to the number of impressions generated by retweets. We also found that this process appeared to be governed by somewhat different dynamics than those that govern retweeting, meaning that a predictive model of retweet impressions might not fully capture this type of influence. Thus, we seek to design a second model to predict these types of impressions.

As before, we define $f()$ as the function which, given a Twitter user as input, returns her number of followers. Then, for a link or hashtag x adopted M times, we define U_j to be the user who adopted x the j^{th} time with $1 \leq j \leq M$. Then, the total number of impressions for this link or hashtag, I_x, is given by the summation:

$$I_x = \sum_{j=1}^{M} f(U_j)$$

Defining a celebrity's set of tweets containing links or hashtags as X and the cardinality of X as $|X| = k$, we now define our variable of interest, $Avg(I_{h/l,2})$, as:

$$Avg(I_{h/l,2}) = \frac{\sum_{x \in X} I_x}{k}$$

We calculate this value for period 2, denoted $Avg(I_{h/l,2})$, and log transform it in order to achieve normality. The same four celebrities from the previous analysis are excluded as outliers from this analysis.

6.3.1 Model Construction

We find that the single best predictor variable in our dataset is actually not $Avg(I_{h/l,1})$, but the total references to the individual in period 1, $log(tr_1)$. Using solely this variable in the model, we

get an average R^2 value of 0.606 and an RMSE of 0.570 against a dependent variable whose mean is 5.37. We next set about finding additional predictor variables to improve the model.

The next meaningful predictor variable we find is the total number of *retweet* impressions generated in period 1, $log(I_{retweet,1})$. With this addition, our average R^2 value rises to 0.662 and the RMSE drops to 0.537. An F-test on the average residual sum of squares reveals that this addition is statistically significant at the 0.05 level. After testing our additional candidate predictor variables, we find no other variables that improve the model's predictive ability. Our final model is thus:

$$log(Avg(I_{h/l,2})) = (0.779 \pm 0.132) + (0.536 \pm 0.037) * log(tr_1) + (0.337 \pm 0.030) * log(I_{retweet,1})$$

The scatterplot of predicted vs. actual values is given in figure 6.4:

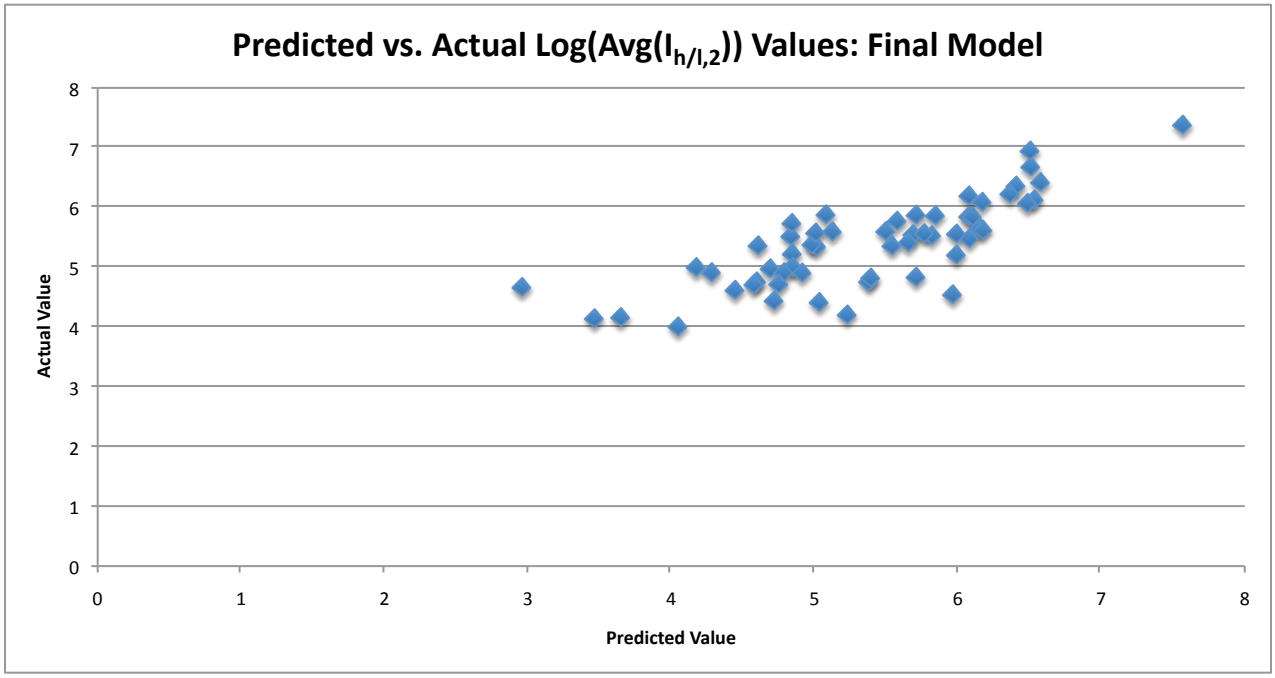

Figure 6.4: Predicted vs. actual values for the final hashtag and link impressions model, containing period 1 reference count and retweet impressions as predictors

6.3.2 Discussion

We note that this model and the model of average retweet impressions share a highly significant regressor—a celebrity's total reference count in period 1. We also note that the second significant predictor in the hashtag and link impressions model is the *total* count of retweet impressions in period 1. These facts imply that there are some indicators of influence that are relatively informative

no matter what type of influence is being exerted. In particular, it seems that a high reference count implies a strong interest in the referenced celebrity's message, which is informative about users' intentions to both retweet the message and to adopt parts of its content.

Similarly, it seems that garnering a large number of retweet impressions indicates a broad interest in a celebrity's message content, explaining why this variable is a useful predictor for hashtag and link impressions. It is plausible that, if a celebrity garners many retweet impressions in period 1, this has the effect of getting her message out there and ensures that many Twitter users are aware of the celebrity's tweets. In period 2, the celebrity can then not only generate many retweet impressions, but many hashtag and link impressions as well.

Even despite these facts, we note that the regressors in this model and the regressors in the average retweet impressions model are not identical. This validates our assumption that retweets do not tell the whole story when it comes to influence, because an effect of comparable size—hashtag and link impressions—can be better predicted when considered independently of retweet impressions.

In terms of applications, this model could similarly serve as the basis for a model to be used in marketing and media. The recent prominence of the "Kony 2012" campaign—which included both hashtags and links that were widely propagated through social media—indicates that this type of viral campaign is a viable way of disseminating a message [12]. If our model were expanded and calibrated separately for link and hashtag impressions, then it could potentially be used to estimate the average number of impressions to be generated by hashtag- and link-containing tweets sent by individual celebrities. Private firms and media groups could then plan to pay prominent individuals at a rate commensurate with the average number of impressions they are expected to generate.

6.4 Follower Uptick Predictive Model

We lastly seek to develop a predictive model of proportional changes in follower counts for the celebrities in our dataset. We define f_{pre} as the vector of follower counts for each celebrity on February 15, 2012, f_{mid} as the vector of follower counts on March 1, and f_{post} as the vector of follower counts on March 16, 2012. Because we want to prioritize large proportional upticks rather

than large absolute upticks, we define our variable of interest, U_2, as:

$$U_2 = \frac{f_{post}}{f_{mid}} - \vec{1}$$

where $\vec{1}$ is the vector of all 1's. This variable is somewhat skewed, so we multiply it by a constant A (10^5) and log transform it to ensure that we have a positive, normally distributed variable. We also define a key predictor variable for our model: the proportional follower uptick from the first period, U_1, calculated as:

$$U_1 = \frac{f_{mid}}{f_{pre}} - \vec{1}$$

and apply the same transformations to U_1.

We remove three outlier celebrities from the dataset prior to developing our model: Beyonce, Michael Ausiello, and Danny Glover. Beyonce was removed because, having never tweeted, she has no meaningful data for many of our predictor variables. The other two celebrities were removed due to the fact that they experienced slight declines in total follower count over the data collection period. Though a more robust model would be able to predict bidirectional changes in follower count, we felt that the rarity of cases in which follower counts declined made it likely that the trends were due to factors exogenous to our data. In removing these data points, we make the implicit assumption that our model is solely predictive of positive changes in follower count over time.

6.4.1 Model Construction

It turns out that proportional follower rises in the first period are an almost perfect predictor of follower rises in the second period. Creating a model using solely this predictor variable, we get an impressive average R^2 value of 0.822 and an RMSE value of 0.171 (against a predicted variable with a mean of 3.28). The coefficients, with 95% confidence intervals, are:

$$log(A * U_2) = 0.1855 \pm 0.047 + (0.935 \pm 0.013) * log(U_1)$$

The scatterplot of predicted versus actual values is given in figure 6.5:

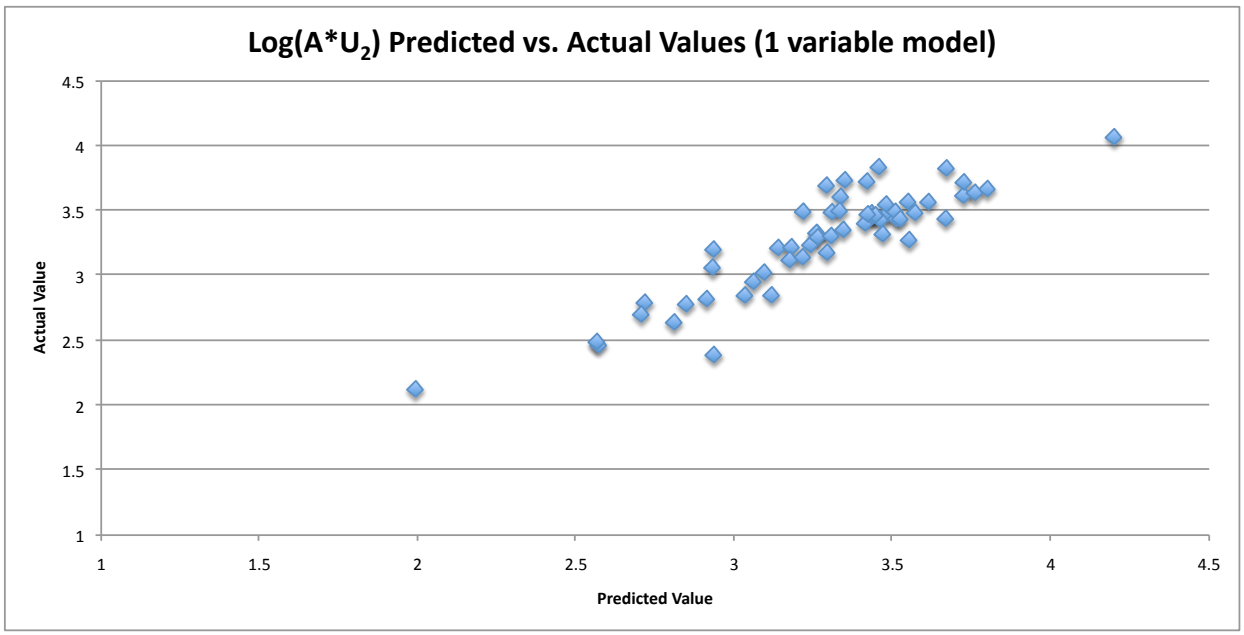

Figure 6.5: Predicted vs. actual values for the original follower uptick model, containing period 1 follower upticks as the predictor variable

Though this model is already extremely predictive, we test out other variables to see if they can improve the model. Due to the fact that the R^2 value is already extremely high—and the residual sum squares value extremely low—we find that no additional variables pass an F-test on the average residual sum squares to warrant inclusion in the model. However, as in section 6.1, we are primarily interested in in absolute increases in predictive ability, and thus include all variables which reduce our model's RMSE. We find one more variable which slightly increases our overall predictivity—total references in period 1 ($log(tr_1)$). The final model is thus:

$$log(A * U_2) = 0.129 \pm 0.046 + (0.840 \pm 0.013) * log(U_1) + (0.810 \pm 0.008) * log(tr_1)$$

The graph of predicted versus actual values for the final model is given in figure 6.6 below:

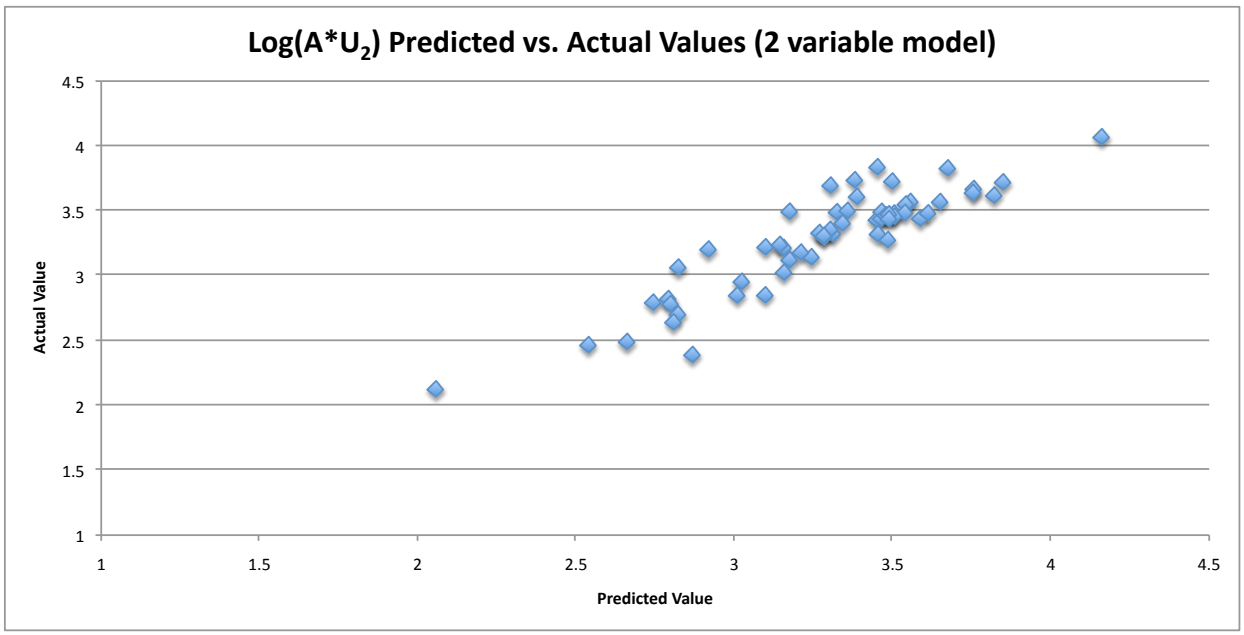

Figure 6.6: Predicted vs. actual values for the final follower uptick model, containing period 1 follower upticks and total reference count as the predictor variables

6.4.2 Discussion

Similar to our prior models, this model not only provides us with a number of insights about Twitter but also has a number of potential applications.

An important question in our analysis has been the degree to which the past can be used to predict the future on Twitter. Given that Twitter activity is dominated by transient events, like retweet cascades or trending topics, there is ample reason to believe that the Twitter of tomorrow might bear no resemblance to the Twitter of today. However, this predictive model gives us reason to question that assumption. As we have seen, the past proportional rise in follower counts is a near-perfect predictor of future proportional rises in follower counts—at least over the relatively short time period that we analyzed. This strongly indicates that, among celebrities who are famous *exogenously* to Twitter, one's star rises and falls over a relatively long time horizon, and one's Twitter following has relatively little to do with day-to-day or week-to-week content generation. It is likely that, looking over a period of months or years, we would find that past follower upticks became a less accurate predictor of future follower upticks. Nonetheless, the sheer strength of the relationship we have found is striking, and makes us reconsider some very basic notions about how and why individuals attract followers on Twitter.

This model also has a number of potential applications, as it could be used to identify individuals who have a high likelihood of rising in follower count. This means that, with a few weeks of data collection, an advertising or media firm could potentially identify celebrities with comparatively low follower counts who are poised to see large increases in their audience over the coming months. Firms could then recruit these individuals to tweet about a particular product or campaign, allowing them to potentially garner many impressions as the celebrity's audience grows. Alternatively, the model could also be used to identify individuals who, based on their anticipated follower upticks, are poised to rapidly ascend to the national stage. In identifying such individuals, firms could potentially gain a new way to identify emerging talent in film, music, or comedy.

Chapter 7

Conclusions

Having developed these predictive models, we now summarize the major findings of this analysis, and set the stage for future work building off of this research.

7.1 Review of Major Contributions

Influence Needs to be Considered Broadly

Our work in Chapter 5 not only demonstrates that several non-retweet influence effects exist—including hashtag and link adoption, word adoption, and emotion adoption—but also confirms that these effects need to be considered alongside retweets when determining celebrity influence.

The justification for considering these effects is four-fold. First, our analysis reveals that the hashtag and link adoption effects appear to be comparable in size to retweet effects in terms of the average number of impressions they generate. Second, while these effects vary in terms of their overall correlation with retweet count, no effect showed a perfect correlation with retweets. And when modeling one of these effects, hashtag and link co-mention adoption, we found that a more predictive model could be constructed using different predictors than those predictors used in a model of retweets. These differences indicate that an influence model which accounts solely for retweets could underweight the influence of individuals who score well on alternative influence metrics. Third, the model of retweet impressions designed in Chapter 6.2—which included the word adoption effect as a predictor variable—demonstrated that some of these alternative influence effects are actually helpful in creating a more predictive model of retweets.

Fourth, several of these effects, including the local adoption effects and the emotional transmission effect, seem to be measures of a celebrity's overall level of *engagement* with his or her

followers. The substantial variation in these metrics indicates that some celebrities have a much higher level of engagement with their mentioners and followers than other celebrities. As a result, impressions generated by these celebrities are likely more meaningful than impressions generated by less well-engaged celebrities. This has meaningful implications for marketers hoping to generate impressions through celebrity tweets.

Past work on the notion of Twitter influence has largely focused on retweets, mentions, and follower counts as potential influence metrics [9, 42, 38], while some other papers have instead focused on link adoption [33, 21] as an influence metric. Our work strongly supports the notion that these metrics cannot be considered in isolation, and that a broad notion of influence is required.

Followers Matter, but References Matter More

At the outset of this paper, we hypothesized that both follower count and reference count would not be significant predictors of influence. In Chapters 4 and 5, we found that both of these variables were correlated with many of our influence metrics, though reference count tended to be better correlated. In Chapter 6, we found that total reference count was a useful predictor variable in both of our influence models (as well as our model of follower upticks); in contrast, neither follower count nor any metric derived from follower count was a meaningful predictor in *any* of our models.

It seems, then, that follower count may be one useful signal of the degree to which individuals on Twitter are "paying attention" to a single celebrity. However, in all of our predictive models, it turned out that reference count was a significantly *better* signal of attention, and that follower counts added no statistically significant predictive power to any model already containing reference counts as a regressor. Somewhat surprisingly, it seems that establishing a following link with a celebrity is only a moderately meaningful indicator of the interest in the celebrity's content; talking about the celebrity is a much more meaningful indicator of this same interest.

For Celebrities, Twitter Celebrity is Enduring

One of the more surprising revelations from our modeling exercise was that past performance on all of our predicted metrics was well correlated with future performance. Particularly in our model of retweet upticks, we found that period 1 follower growth was an astoundingly good predictor of period 2 follower growth. In our two influence models, we also found that modeling period 2 average impression counts solely using period 1 average impression counts produced surprisingly

predictive models—though we ultimately found superior predictors to use in the models instead.

This indicates that Twitter influence and audience growth are *not* only driven by ephemeral trends or by the reception of recent content. Rather, we see that celebrities tend to grow in audience at a fairly consistent rate and tend to generate roughly similar impressions per tweet from week to week. It seems, then, that—at least for those individuals whose fame is developed exogenously to Twitter—Twitter celebrity is a relatively enduring state.

Methodological Contributions

This paper also makes a number of noteworthy methodological contributions.

First, to our knowledge, Granger Causality analysis has to date only been used to determine whether or not processes on Twitter are useful for forecasting processes outside of Twitter, such as the performance of the stock market. If this is true, then this paper is the first to use Granger Causality analysis to explore the relationship between two processes endogenous to Twitter. The Granger Causality technique could potentially be used to relate the time-series representations of many different Twitter activities, not just the tweeting of emotionally charged messages. This method may therefore provide a useful tool for future researchers to determine if a potential influencer's Twitter behavior forecasts the activity of other individuals on the network.

Second, the models provided here demonstrate that relatively simple modeling techniques—ordinary least squares regression and leave-one-out cross-validation—can be used to construct and validate models of Twitter processes. Furthermore, the models generated in Chapter 6 should provide an extremely useful basis for developing future predictive models of Twitter impressions.

7.2 Future Work

We briefly enumerate a number of potential future developments for this research, which could not be included in this paper due time constraints.

First, we hope to develop a composite influence metric in the future, which incorporates a celebrity's ability to engender retweets, hashtag and link adoptions, word adoptions, and adoptions of emotional state. Finding an appropriate weighting of the different effects would require significant empirical research, but would ultimately allow us to assign each celebrity a single numerical influence score and compare across celebrities.

Second, we hope to develop predictive models of the remaining non-retweet influence effects discussed in chapter five (word adoption and emotion adoption). This would allow us to determine whether any of our influence metrics can be predicted using the same regressors, and would provide insight into whether any influence effects are essentially redundant.

Third, we aim to conduct further research into the notion of engagement, and to explore the relationship between influence metrics which appear to be highly reflective of a celebrity's engagement—like local hashtag and link adoption—and aggregate influence metrics, which appear to be less reflective of engagement. Ultimately, we seek to develop a way to rate the impressions generated by a celebrity according to that celebrity's level of engagement with his follower base and mentioners. This would make our impressions more indicative of the degree to which their viewers are actually likely to care about a propagated message.

Lastly, we hope to eventually increase the resolution of our modeling procedure by generating models to predict impressions not for individual celebrities, but rather for individual *tweets*. This would require extensive additional research into how the various features of the textual content of tweets affects their retweet probability. Nonetheless, such a model would be incredibly useful in allowing marketers to project the actual number of impressions to be generated on a tweet-by-tweet basis for prominent celebrity tweeters.

Appendix A

List of Celebrities and Designations in the Dataset

Name	Handle	Designation
Lady Gaga	ladygaga	Musician
Justin Bieber	justinbieber	Musician
Katy Perry	katyperry	Musician
Shakira	shakira	Musician
Kim Kardashian	KimKardashian	TV Personality
Britney Spears	britneyspears	Musician
Rihanna	rihanna	Musician
Barack Obama	BarackObama	Politician
Taylor Swift	taylorswift13	Musician
Selena Gomez	selenagomez	Musician
Ashton Kutcher	aplusk	Actor
Oprah Winfrey	Oprah	TV Personality
Marshall Mathers	Eminem	Musician
Justin Timberlake	jtimberlake	Musician
Chris Brown	chrisbrown	Musician
Charlie Sheen	charliesheen	Actor
Jim Carrey	JimCarrey	Comedian
Paris Hilton	ParisHilton	TV Personality
Alicia Keys	aliciakeys	Musician
Bill Gates	BillGates	Entrepreneur
Conan O'Brien	ConanOBrien	Comedian
Daniel Tosh	danieltosh	Comedian
Nicole Polizzi	snooki	TV Personality
Stephen Fry	stephenfry	Comedian
Jonas Brothers	JonasBrothers	Musician
David Guetta	davidguetta	Musician
Soulja Boy	souljaboy	Musician
Stephen Colbert	StephenAtHome	Comedian
Usher	UsherRaymondIV	Musician
Beyonce Knowles	beyonce	Musician

Selected Celebrities Table – Continued

Name	Handle	Followers
Dane Cook	danecook	Comedian
Ke$ha	keshasuxx	Musician
Tom Cruise	TomCruise	Actor
Paula Abdul	PaulaAbdul	Musician
Arnold Schwarzenegger	Schwarzenegger	Actor
Felicia Day	feliciaday	Actor
Keri Hilson	KeriHilson	Musician
John Cleese	JohnCleese	Comedian
Danny Glover	mrdannyglover	Actor
Jordin Sparks	JordinSparks	Musician
Newt Gingrich	newtgingrich	Politician
Chris Anderson	TEDchris	Entrepreneur
Scooter Braun	scooterbraun	Entrepreneur
Suze Orman	SuzeOrmanShow	TV Personality
Jon Favreau	JonFavreau	Actor
Michael Ausiello	MichaelAusiello	TV Personality
Travis Barker	travisbarker	Musician
Dita Von Teese	DitaVonTeese	Actor
Kirstie Alley	kirstiealley	Actor
Dr. Phil	DrPhil	TV Personality
Dianna Agron	DiannaAgron	Actor
Kristin Cavallari	KristinCav	TV Personality
Robbie Williams	robbiewilliams	Musician
Jon Stewart	TheDailyShow	Comedian
Adam Savage	donttrythis	Entrepreneur
Joy Behar	JoyVBehar	Comedian
Mitt Romney	MittRomney	Politician
Ron Paul	RonPaul	Politician
Rick Perry	GovernorPerry	Politician
Rick Santorum	RickSantorum	Politician

Appendix B

Stop-Word List

i	me	my	myself	we	us	our	ours	ourselves	you
your	yours	yourself	yourselves	he	him	his	himself	she	her
hers	herself	it	its	itself	they	them	their	theirs	themselves
what	which	who	whom	this	that	these	those	am	is
are	was	were	be	been	being	have	has	had	having
do	does	did	doing	would	shall	should	could	must	ought
i'm	you're	he's	she's	it's	we're	they're	i've	you've	we've
they've	i'd	you'd	he'd	she'd	we'd	they'd	i'll	you'll	he'll
she'll	we'll	they'll	isn't	aren't	wasn't	weren't	hasn't	haven't	hadn't
doesn't	don't	didn't	won't	wouldn't	shan't	shouldn't	can't	cannot	couldn't
mustn't	let's	that's	who's	what's	here's	there's	when's	where's	why's
how's	a	an	the	and	but	if	or	because	as
until	while	of	at	by	for	with	about	against	between
into	through	during	before	after	above	below	to	from	up
down	in	out	on	off	over	under	again	further	then
once	here	there	when	where	why	how	all	any	both
each	few	more	most	other	some	such	no	nor	not
only	own	same	so	than	too	very	a	able	about
across	after	all	almost	also	am	among	an	and	any
are	as	at	be	because	been	but	by	can	cannot
could	dear	did	do	does	either	else	ever	every	for
from	get	got	had	has	have	he	her	hers	him
his	how	however	i	if	in	into	is	it	its
just	least	let	like	likely	may	me	might	most	must
my	neither	no	nor	not	of	off	often	on	only
or	other	our	own	rather	said	say	says	she	should
since	so	some	than	that	the	their	them	then	there
these	they	this	tis	to	too	twas	us	wants	was
we	were	what	when	where	which	while	who	whom	why
will	with	would	yet	you	your	ll	ve		

Appendix C

Candidate Predictor Variables

The following metrics were calculated for period 1 and were tested as candidate predictor variables for each of the models developed in Chapter 6:

- Retweet Impression Count
- Retweet Count
- Retweet Impressions/Follower
- Retweet Impressions/Tweet
- Follower Count
- Reference Count
- Tweets Sent
- Hashtag Local Adoption
- Link Local Adoption
- Hashtag and Link Co-Mentions
- Hashtag and Link Co-Mentions Impression Count
- Hashtag and Link Co-Mentions Impressions/Tweet
- Hashtag and Link Co-Mentions Impressions/Follower
- Follower Word Adoption
- Mentioner Word Adoption
- Follower Emotional Transmission
- Mentioner Emotional Transmission
- Followers' Average Follower Count
- Influence to Audience Ratio
- Influence to Buzz Ratio
- Percentage of Tweets Containing Hashtag
- Percentage of Tweets Containing Links
- Percentage of Tweets Containing Mentions
- Average Tweet Length
- Proportional Follower Increase
- Celebrity Designations (represented by 6 binary variables)
- Frequency of terms in overall Twitter feed

Bibliography

[1] Your world, more connected. *Twitter Blog*, 2011.

[2] Sebastien Ardon, Amitabha Bagchi, Anirban Mahanti, Amit Ruhela, Aaditeshwar Seth, Rudra M. Tripathy, and Sipat Triukose. Spatio-temporal analysis of topic popularity in twitter. *CoRR*, abs/1111.2904, 2011.

[3] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *CoRR*, abs/1003.5699, 2010.

[4] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[5] Nick Bilton. Twitter reaches 100 million active users. *New York Times*, 2011.

[6] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[7] Bianca Bosker. To rank influence on twitter, keep quiet and cross your fingers. *Huffington Post*, 2011.

[8] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, volume 0, pages 1–10, Los Alamitos, CA, USA, January 2010. IEEE.

[9] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

[10] John Cook. Update: Only 92% of newt gingrich's twitter followers are fake. *Gawker*, 2011.

[11] M. T. Flanagan. Michael Thomas Flanagan's Java Scientific Library. urlhttp://www.ee.ucl.ac.uk/mflanaga/java, 2008.

[12] Elizabeth Flock. Invisible children responds to criticism about stop kony campaign. *Washington Post*, 2012.

[13] Rumi Ghosh, Kristina Konstantin Voevodski, Lerman Tawan, and Shang hua Teng. Non-conservative diffusion and its application to social network analysis, 1102.

[14] Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Sci. Rep.*, 1, 12 2011.

[15] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969.

[16] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. *CoRR*, abs/1101.0510, 2011.

[17] Bernardo A Huberman, Daniel M Romero, and year=2008 pages=1–9 Wu, Fang. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1).

[18] Mathew Ingram. Should you care how high your klout score is? *Businessweek*, 2011.

[19] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[20] Janette Lehmann, Bruno Gonçalves, Jose J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. *CoRR*, abs/1111.1896, 2011.

[21] Kristina Lerman, Rumi Ghosh, and Tawan Surachawala. Social contagion: An empirical study of information spread on Digg and Twitter follower graphs. February 2012.

[22] Huina Mao, Scott Counts, and Johan Bollen. Computational economic and finance gauges: Polls, search, and twitter. In *Meeting of the National Bureau of Economic Research - Behavioral*

Finance Meeting, Stanford, CT, 11/5/2011 2011. National Bureau of Economic Research, National Bureau of Economic Research.

[23] J.S. Maritz. *Distribution-free statistical methods*. Monographs on statistics and applied probability. Chapman & Hall, 1995.

[24] Marissa McNaughton. Social networking stats: Instagram reaches 27 million users. *The Realtime Report*, 2012.

[25] Andrew W. Moore. Cross-validation for detecting and preventing overfitting. urlhttp://www.autonlab.org/tutorials/overfit10.pdf.

[26] F. Å. Nielsen. Afinn, mar 2011.

[27] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[28] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[29] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[30] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[31] David Pennock. Romney win captivates twitterfor 10 minutes. *Yahoo! News*, 2012.

[32] Martin F. Porter. Snowball: A language for stemming algorithms. Published online, October 2001.

[33] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and Passivity in Social Media. August 2010.

[34] Stephanie Rosenbloom. Got twitter? you've been scored. *The New York Times*, 2011.

[35] Catharine Smith. Beyonce pregnancy: New twitter record set at mtv vmas. *Huffington Post*, 2011.

[36] Jon Starkweather. Cross validation techniques in r: A brief overview of some methods, packages, and functions for assessing prediction models. 2011.

[37] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. *CoRR*, abs/1110.2724, 2011.

[38] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. pages 177–184, August 2010.

[39] Chris Taylor. Exclusive: Twitter analysis vindicates gingrich in followers scandal. *Mashable*, 2011.

[40] Katrina Trinko. Gingrich's twitter followers may be fake. *National Review*, 2011.

[41] Lyndra Vassar. Beyonce has 1 million twitter followers, never tweets. *Essence*, 2011.

[42] Jianshu Weng, Ee P. Lim, Jing Jiang, and Qi He. TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.