A Longitudinal Study of Follow Predictors on Twitter

C.J. Hutto
School of Interactive Computing
Georgia Institute of Technology
cjhutto@gatech.edu

Sarita Yardi
School of Information
University of Michigan
yardi@umich.edu

Eric Gilbert
School of Interactive Computing
Georgia Institute of Technology
gilbert@cc.gatech.edu

ABSTRACT
Follower count is important to Twitter users: it can indicate popularity and prestige. Yet, holistically, little is understood about what factors – like social behavior, message content, and network structure – lead to more followers. Such information could help technologists design and build tools that help users grow their audiences. In this paper, we study 507 Twitter users and a half-million of their tweets over 15 months. Marrying a longitudinal approach with a negative binomial auto-regression model, we find that variables for message content, social behavior, and network structure should be given equal consideration when predicting link formations on Twitter. To our knowledge, this is the first longitudinal study of follow predictors, and the first to show that the relative contributions of social behavior and message content are just as impactful as factors related to social network structure for predicting growth of online social networks. We conclude with practical and theoretical implications for designing social media technologies.

Author Keywords
Social networks; social media; computer-mediated communication

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Asynchronous interaction - Web-based interaction.

General Terms
Human Factors; Design; Measurement.

INTRODUCTION
Followers are Twitter's most basic currency. Building an audience of followers can create access to a network of social ties, resources, and influence. Yet, little is understood about how to grow such an audience. This paper examines multiple factors that affect tie formation and dissolution over time on the social media service Twitter. We collected behavioral, content, and network data approximately every three months for fifteen months. We examine specific user *social behavior choices*, such as: proportions of directed communications versus broadcast communications [6]; the total number of tweets produced; communication burstiness; and profile completeness [30]. We also assessed numerous attributes specific to the *content* of users' tweets, such as: propensity to express positive versus negative sentiment [26,37]; topical focus [40]; proportions of tweets with "meformer" content versus informational content [33]; frequency of others "retweeting" a user's content [5]; linguistic sophistication (reading difficulty) of tweets; and hashtag usage. Finally, we evaluated the impact of users' evolving *social network structure*, collecting snapshots of their friends and followers every three months for fifteen months. With this, we can evaluate the effects of network status, reciprocity [18], and common network neighbors.

Our variables were selected from prominent theoretical constructs bridging social science, linguistics, computer mediated communications, and network theory. This paper compares the relative contributions of factors from each perspective for predicting link formations in online social networks. We take a temporal perspective and develop a model that accounts for social behavior, message content, and network elements at several intervals for over a year. We use an auto-regressive, negative binomial regression model to explore the changes in users' follower counts over time. We find that *message content significantly impacts follower growth*. For example, in contrast to [26], we find that expressing negative sentiment has an adverse effect on follower gain, whereas expressing positive sentiment helps to facilitate it. Similarly, we show that informational content attracts new followers with a relative impact that is roughly *thirty times higher* than the impact of "meformer" content, which deters growth. We also find that *behavioral choices can also dramatically affect follower growth*. For example, choosing to complete one's profile and choosing directed communication strategies over broadcast strategies significantly stimulates follower growth over time. Finally, we show that *even simple measures of topology and structure are useful predictors of evolutionary network growth*.

Comparing across multiple variables related to message content, social behavior, and network structure allows us to interpret their relative effect on follower growth from different theoretical perspectives. We believe this is the first paper of its kind to compare the impact of all these factors together within a single longitudinal study. The temporal nature of the longitudinal method is crucial because it allows us to suggest causal relationships between these factors and network growth on Twitter.

CHI 2013, April 27 – May 2, 2013, Paris, France.

BACKGROUND

Next, we consider related work showing how social behavior, message content, and network structure relate to follower growth. Our study draws from this prior work in deciding which variables to include in our analysis, and contributes new results to this body of literature by considering these variables temporally, and in conjunction with one another. For convenience and organizational purposes, we group our variables into three categories: social behaviors (e.g., interactional communication choices that a user makes), message content (e.g., linguistic cues), and social network structure. These categories are intended to be neither mutually exclusive nor exhaustive. However, we specifically call attention to content variables because they seem to be underrepresented in much of the related literature on follower growth dynamics [18,19,26,28,31].

Social Behavior and Follower Growth

Social Capital and Communication Behavior

Social capital refers to "the actual or potential resources which are linked to a durable network of more or less institutionalized relationships of mutual acquaintance or recognition" [4]. It is your relative social "worth," resulting from your position in a social network: i.e., the number and kind of the ties you maintain, your relative access to resources desired by those in your network, as well as your level of access to the resources your network ties possess [41].

In prior work, researchers distinguished between three kinds of social behavior that affect social capital in the social networking site, Facebook: (1) directed communications with specific, target individuals; (2) broadcast communications, which are not targeted at anyone in particular; and (3) passive consumption of content [6]. Because personalized messages are more likely to contain content that strengthens social relationships (such as self-disclosure and general supportiveness), it has been suggested that directed communications are useful for maintaining existing ties and for encouraging the growth of new ones. Indeed, previous research found that, when compared to broadcast communications and passive consumption, personalized one-on-one communication strategies have a measurably greater impact on self-reported social capital of Facebook users [6]. Other research suggests that informal personal conversation is a major reason for using a social media like Twitter [22,23], even for work and enterprise purposes [42,43]. However, the volume of messages and the rate at which they are transmitted (i.e., their "burstiness") are both correlated with unfollowing on Twitter [28]. Here, we test whether these behaviors help to grow followers on Twitter.

Profile Elements as Social Signals

Because there is some cost incurred with producing it, user-generated profile content is an important signal for conveying a trustworthy identity [11,12,30]. The shared context of social networking sites such as Facebook helps facilitate explicit and implicit verification of identity claims, and users are motivated to present their "ideal self" [17] in order to attract new connections. In [30], the authors explore the relationship between profile structure (namely, which fields are completed) and number of friends on Facebook. Based on a static snapshot of the social network at a large university, the authors found that the act of populating profile fields was strongly correlated with the number of friendship links. Compared to users without profile elements, users who had entered profile content had about two to three times as many friends. Based on this prior literature as well as our own intuition, we anticipate similar effects in our longitudinal data regarding network growth on Twitter. Assuming that people will be more likely to follow those who include identity cues in their profile (such as description, location, and personalized URL), we expect that the more these elements are included, the more successful one will be in growing an audience. Our study tests these assumptions.

Message Content and Follower Growth

Sentiment and Emotional Language

Sentiment analysis refers to the computational treatment of opinion, sentiment, and subjectivity in text [35]. Previous research found significant correlations between the number of followers of a Twitter user and that user's tendency to express emotions like joy and sadness [26] or positive versus negative sentiments [37] in their tweets. However, the authors in [26] acknowledge that an important limitation of the study was the static nature of the correlation analysis. In particular, we note the following passage from the paper:

> With the current analysis we cannot deduce causality; e.g., whether the emotional richness of interactions draws more followers or whether people tend to share more emotional content when they have larger audiences. (p. 382)

Although not explicitly stated, this same limitation also applies to [37]. We build on their prior work and extend it by studying changes in audiences *over time*. By relying on time-dependent regression analysis of longitudinal data to identify the relative effects of sentiment expression on follower gain, we are able to address the limitation noted above. This is conceptually similar to the approach used by [20] to characterize the relative effects of various factors on predicting Twitter adoption among young adults. Exploring dynamics over time gives us a stronger case for causality.

We also build on the approach in both [26] and [37] by extending our analysis beyond the LIWC2007 text analysis package to automatically classify positive and negative sentiment. LIWC [36] is a widely used and validated dictionary-based coding system often used to characterize texts by counting the frequency of more than 4,400 words in over 60 categories. However, LIWC does not include many features that are important for sentiment analysis of tweets. For example, our study also includes the 408 words in LIWC categories for P*ositive Emotion* and *Negative Emotion*, plus an additional ~2,200 words with positive or negative senti-

ment[1], as well as considerations for sentiment-laden acronyms/initialisms, emoticons, negations, and slang. These additional characteristics are known to be important features of sentiment analysis for microblogs like Twitter [10]. Also, some words are bound to connote more extreme sentiment than others (e.g., "good" versus "exceptional"). Thus, in addition to simply counting occurrences of positive or negative words (i.e., the LIWC method), we also assess the directional magnitude (i.e., *intensity*) of the sentiment for each word, associating human coded valence scores ranging from -5 to +5 for each word in our dictionary.

Topical Focus

The principle of *homophily* asserts that similarity engenders stronger potential for interpersonal connections. In the selection of social relationships, people tend to form ties to others who are like them – a finding that has been one of the most pervasive empirical regularities of modern social science [32]. Sharing interests with another person is one form of similarity [14]. A Twitter user who discusses a wide range of topics may appeal to a broader audience, therefore attracting more followers – a notion that, according to [40], is supported by the economic theory of network externalities [24,38]. In [40], the authors describe how initial topical focus affected users' ability to attract followers. However, the users in [40] self-identified as providers of *politically oriented* tweets, and it is unknown whether the findings from [40] will hold for a more heterogeneous sample of Twitter users. Our study addresses this question.

Informativeness: Information Brokering and "Meformers"

In [29], the authors highlight the dual nature of Twitter as both a social network and as a news/information medium. Also, [33] suggests two basic categorizations of Twitter users as Informers (those who share informational content) versus "Meformers" (those who share content about themselves). Meformers were reported to have almost three times fewer followers than Informers. The authors note that "the direction of the causal relationship between information sharing behavior and extended social activity is not clear" [33:192]. We explore whether this type of message content affects growing a social media audience over time.

Network Structure and Follower Growth

Network Size, Reciprocity and Mutuality

Preferential attachment, or the phenomenon whereby new network members prefer to make a connection to popular existing members, is a common property of real life social networks [3] and is useful for predicting the formation of new connections [31]. The number of followers a person maintains has been shown to reduce the likelihood that the person will be unfollowed in the future [25], meaning popular people often remain popular. Additionally, we can calculate the "attention status" of an individual within their own Twitter network by taking the ratio of followers (those who pay attention to the user) to following (those among whom the user divides their attention). Such measures reflect ego-level network attributes that affect the decision of others to follow the user. On the other hand, [18] shows that follower counts alone do not fully explain interest in following. In other words, popularity, in and of itself, does not beget popularity. Dyadic properties such as *reciprocity* and *mutuality* also play key roles in the process of tie formation and dissolution [18,25].

Common Neighbors: Structural Balance and Triadic Closure

In addition to dyadic structural properties, we also consider *triads* (structures of three individuals). Specifically, we are interested in the concepts of *structural balance* and *triadic closure*. For example, consider the case where three people form an undirected network. If A is friends with X, and X is friends with B, then according to Heider's theory of cognitive balance, the triad is "balanced" when A is friends with B, but "unbalanced" when A is not friends with B [21]. As the number of common neighbors (occurrences of "X") between A and B increases, the likelihood of the A-B tie being formed also increases [8]. This principle of structural proximity is known as *triadic closure* [13]. Measuring the occurrences of common network neighbors is useful for link predictions in real life social networks [31] as well as online social networks [18,19,25]. We explore the extent to which these network structures impact follower gain as compared to message content and social behavior.

Limitations (and Benefits) of Longitudinal Observations

Making causal claims with observational data can be problematic. It is impossible to absolutely rule out every possible "third factor" that might account for some portion of an association between an independent variable and its effect on the dependent variable. We try to mitigate this problem by accounting for as many "third factors" as is feasible. Longitudinal studies are still correlational research, but these correlations have greater power because we have time-dependent, repeat observations. In other words, when input A is consistently and reliably observed preceding outcome B for the exact same group of individuals time after time, then we have greater confidence in suggesting a causal relationship between A and B.

METHODS

Data Collection and Reduction

We collected data from 507 active Twitter users who collectively provided us with a corpus of 522,368 tweets spanning the 15 months between August 2010 and October 2011. In addition to the tweets, we also have snapshots of friends and followers taken at periodic intervals (a total of five periods, each approximately three months in duration). We were interested in discovering the relationship between the factors discussed above within each three-month period and the subsequent changes in follower counts at the end of that period. To build our dataset, Twitter accounts were obtained by recording unique account IDs that appeared on

[1] http://fnielsen.posterous.com/afinn-a-new-word-list-for-sentiment-analysis

the public timeline during a two-week period in August 2010, and then screened for certain attributes. The subset selected for inclusion in this study consisted of those accounts that met the following four criteria when sampled approximately every three months:

1. Tweet in English, as determined by inspecting the users' profiles for the designated language via Tweepy[2], a Twitter API library for Python, as well as Python's Natural Language Tool Kit[3] (NLTK) for language detection on the users' 20 most recent tweets. This filter is necessary for our linguistic predictors (described later), although it may restrict the generalizability of our results.
2. Have Twitter accounts that are at least 30 days old at the time of the first collection period, and are therefore not new to the service. This was done to avoid the potential confounding effects of users who have just joined and are likely building up their followership based on existing friends and acquaintances (rather than attracting followers based on the variables we track).
3. Follow at least fifteen other "friends" and have at least five followers. This removes a large portion of unengaged or novice users, and is close to Twitter's own definition of an "active user"[4,5].
4. Tweet at least twenty times within each time period (a *time period* is the approximately three-month interval between snapshots of users' social networks). This removes the confounding effects of inactive accounts.

Response Variable (dependent measure)

Follower growth: change in follower counts for users at the end of a given three-month time period, as compared to the follower counts at the end of the previous period.

Predictor Variables

Behavioral and Social Interaction Variables

Tweets in period: the total number of tweets produced by a user in a three-month time period.

Peak tweets per hour ("burstiness"): for a given three-month time period, the maximum rate of tweets per hour.

Directed communications index: captures replies and mentions, as well as consideration for the social signal sent when the person "favorites" someone else's tweet, calculated as "@" count plus favorites count divided by the total number of tweets in a period.

Broadcast communication index: the ratio of tweets with no "@" at all in the tweet to total number of tweets in a period.

Profile cues of "trustworthiness" of Twitter identity: (1) the length, in characters, of the user's self-defined profile description, (2) whether the user has indicated a personal URL in their profile, and (3) whether the user has indicated their location. We collected data about whether the user had a personal profile image or the default egg image, but there was insufficient variation in the data to use this variable (all users in our sample had non-default images).

Message Content Variables

Positive (Negative) sentiment intensity rate: ratio of the sum of the valence intensity of positive (negative) language used in tweets to the total number of tweets in a period. In a separate formative evaluation involving a small subset of tweets from the corpus (n=300), our custom sentiment analysis engine performed quite well. The correlation coefficient between our sentiment analysis engine and ratings from three human judges was high (r = 0.702); better than the Pattern.en sentiment analysis engine[6] (r = 0.568). The correlation among human judges was r = 0.851.

Informational content index: the ratio of tweets containing either a URL, "RT", "MT", "HT" or "via" to total number of tweets in the period.

Meformer content index: the ratio of tweets containing any of the 24 self-referencing pronouns identified in LIWC (e.g., words like "I", "me", "my", "we", "us") to total number of tweets in the period.

Topic focus: following [40], this is the average cosine similarity (ranging between 0 and 1) for every unique paired combination of a user's tweets in a given time period.

User tweets retweeted ratio: the total number of times a user's tweets were retweeted, relative to the total number of tweets produced by the user in the period.

Hashtag usage ratio: the total number of hashtags used in a period relative to the total number of tweets in the period.

TReDIX: the "Tweet Reading Difficulty Index" is a measure developed by the authors to capture the linguistic sophistication of a set of tweets. It is inspired by the Readability Index (RIX, c.f. [1]) and is based on the frequency of real English words with 7 or more letters. TReDIX is a ratio of the total count of long words appearing in tweets within a time period relative to the number of tweets in the period.

Network Topology/Structural Variables

In-link reciprocity rate: the number of followers that the user is also following relative to the total number of followers in the user's social network at the end of each period.

Attention-status ratio: ratio of followers (those who pay attention to the user) to following (those among whom the user divides their attention), calculated based on the user's existing social network at the end of each period.

---

[2] http://code.google.com/p/tweepy

[3] http://www.nltk.org

[4] http://www.businessinsider.com/chart-of-the-day-how-many-users-does-twitter-really-have-2011-3

[5] http://techland.time.com/2011/09/09/twitter-reveals-active-user-number-how-many-actually-say-something

[6] http://www.clips.ua.ac.be/pages/pattern-en#sentiment

Network overlap: where A is the user of interest and B is either a follower or a friend of A, this is the raw network overlap (count of common neighbors) between A and B. The final measure is the sum for user A's entire network.

Other (Control) Variables

Age of account: the age of a user's Twitter account (in days) at the end of a time period, to control for the likely differences between older, more established accounts and newer, developing accounts.

No. of followers: The total number of followers at the end of a given period, a plausible criterion used by other potential followers when evaluating whether or not to follow the user. We include the number of followers as a control to account for popularity-based preferential attachments.

No. of friends ("followees"): The number of accounts the user is following at the end of a given period, also a plausible criterion used by potential followers when deciding whether to follow a user.

Change in followers (previous period): change in follower count at the end of time period t_{-1} (the previous time period), is a lagged variable used to control for second order follower growth dynamics for the dependent variable in the time-dependent auto-regressive model. This addresses the issue of possible preferential (de)attachment for rising or falling "stars" [3], and helps mitigate concerns related to lack of independence among repeated observations.

We test the predictive power of these variables by incorporating auto-regression into a negative binomial regression model. Negative binomial regression is used for modeling count variables, and is well-suited to modeling dependent variables of count data which are ill-dispersed (either under- or over- dispersed) and do not have an excessive number of zeros [7], as is the case with our data set. Auto-regressive models attempt to predict an output of a system based on previous observations [34], which allows us to mitigate concerns associated with lack of independence for repeated measures by incorporating a lagged variable into our model. In the present study, we use auto-regression to account for the overall slope of follower gain heading into a given time period. Change in follower growth at the end of time period t_0 is therefore conditioned upon the change in follower growth at the end of t_{-1} (the previous time period).

After removing tweets from the first time period interval (it only provides the initial baseline of counts from which we derive changes in follower growth for subsequent periods) and the second time period (in order to incorporate dependency on change in growth for the auto-regressive model), we have 507 unique active Twitter users who collectively provided 1,836 instances of follower growth across the remaining four time periods of our analysis.

RESULTS

We first present descriptive statistics for the dependent measure (follower growth) and the twenty-two predictor and control variables. These variables are organized into three convenience categories: behavioral/social interaction, message content, and network topology/structure.

Descriptive Statistics

Table 1 shows descriptive statistics (mean, standard deviation, minimum, first quartile, median, third quartile, maximum, and density plots) for the response variable (follower growth) as well as seventeen of the twenty-two predictor and control variables. For space reasons, we omit user profile data from the table, and instead provide the following summary: the majority of users had URLs listed in their profile (mean=86%, SD=35%), most listed their location (mean=97%, SD=16%), and the average profile description was 85 characters long. We also omit the lagged variable *change in followers (previous period)* (mean=106.96, SD=551.84, median=25). The density plots in Table 1 show some skewness (lack of symmetry) and generally high kurtosis (peaked, rather than flat, distributions) for many of the variables. This makes the median a better measure of central tendency than the mean for many of the variables, and the density plots reveal the distributions for each variable.

Behavioral and Social Interaction Variables

Most users tweetd between 131-364 times in three months (median=222), usually with bursts of no more than eight tweets within a single hour. The Broadcast Communication Index shows the proportion of tweets that are not directed to any specific person. Most people use broadcast communication strategies for about 30%-60% of their messages (median=45%).

Message Content Variables

Proportionally, most people tweet about twice as much positive and neutral content as negative content, with an average of 106 tweets identified as positive (the same average were neutral tweets), and 51 tweets labeled as negative. (Note: this data did not fit in Table 1). In terms of *intensity* of positive or negative language, most people are generally about three times more positive than they are negative in their tweets (see Table 1). The proportion of users' tweets identified as "meformer" content was nearly normally distributed – users talk about themselves in 41% of their messages, on average. Informational content accounted for 24% of messages. This closely resembles the results from [33]. The mean and median of topical focus (average cosine similarity of one's own tweets) indicate that in general, people post a fairly diverse range of content. The ratios of retweets (0.02-0.12, median=0.05) and hashtag usage (0.06-0.26, median=0.13) to total number of tweets in a period are moderate for the majority of users – retweets comprised about 12% of users' messages, and hashtags were used in about 26% of tweets. The Tweet Reading Difficulty Index (TReDIX) is evenly distributed, with most people using moderately sophisticated language – about 2.36 long words per tweet, on average. On the original RIX scale, an index of 2.4 is equivalent to a seventh grade reading level [1].

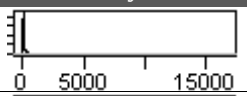
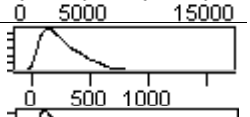
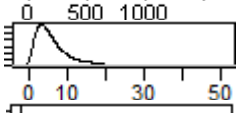
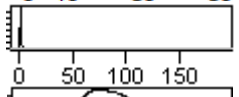
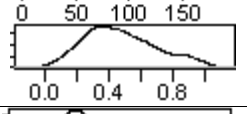
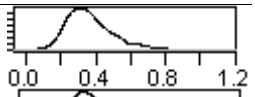
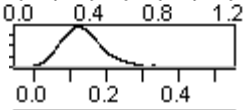
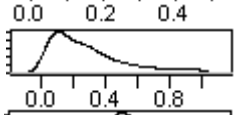
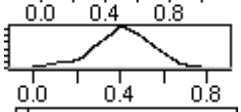
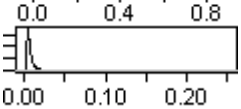
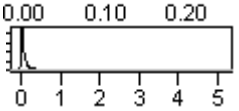
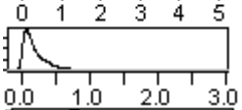
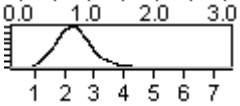
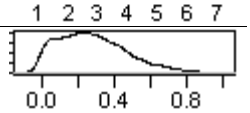
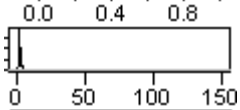
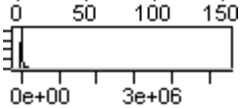
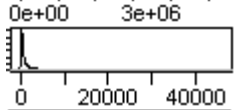
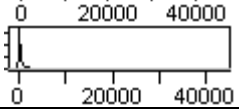
	Variable	Mean	Std Dev	Min	1st Q	Median	3rd Q	Max	Density Plot
D.V.	Follower Growth	194.2	832.7	0	12	36	106	16,623	
Behavioral and Social Interaction Variables	Number of Tweets in period (a control)	262.6	176.3	21	131	222	364	1,552	
	Peak tweets per hour ("Burstiness")	6.39	5.78	0.15	2.79	4.79	7.9	48.9	
	Directed communications	1.91	7.4	0	0.58	0.83	1.22	190.25	
	Broadcast communications	0.48	0.22	0	0.31	0.45	0.62	1	
Message Content Variables	Positive Sentiment Intensity Rate	0.37	0.14	0.05	0.27	0.35	0.44	1.08	
	Negative Sentiment Intensity Rate	0.14	0.06	0	0.095	0.13	0.17	0.5	
	Informational content index	0.3	0.23	0	0.12	0.24	0.41	1	
	"Meformer" content index	0.41	0.14	0	0.33	0.41	0.50	0.79	
	Topic focus	0.008	0.01	0.002	0.005	0.008	0.01	0.25	
	User RT ratio	0.15	0.4	0	0.02	0.05	0.12	5.1	
	Hashtag usage ratio	0.2	0.24	0	0.057	0.13	0.26	2.82	
	TReDIX	2.36	0.64	0.84	1.94	2.31	2.696	6.95	
Network Topology / Structural Variables	Reciprocity rate	0.28	0.19	0	0.125	0.25	0.4	0.9	
	Attention-status ratio	2.18	7.06	0	0.895	1.19	1.90	149.25	
	Network overlap	94,730	351,388	0	2,070	10,472	50,263	5,308,200	
	No. of followers at end of period (a control)	1,145.42	3391.93	15	175.8	391.5	948.8	45,932	
	No. of friends at end of period (a control)	830.63	2879.43	18	135	289.5	661.2	42,797	

Table 1: Descriptive statistics for the dependent variable (follower growth) and seventeen of the twenty-two predictor and control variables. The x-axes of the density plots represent the measured value of the variable, and the y-axis indicates the density of users observed at a particular value. For example, one can interpret the table to indicate that most users grew their Twitter audience at a rate of about 12 to 106 new followers (median=36) every 3 months. The density plot indicates that most users fell within this range.

Network Topology / Structural Variables

The majority of users have 176-949 followers, and 135-661 friends (medians are 391.5 and 289.5, respectively). The density plots indicate that few users fell outside these ranges, but those that exceeded the range did so by a large margin. In general, users reciprocally follow-back about a quarter of their followers (mean=28%, median=25%). The density plot for attention-status ratio (that is, followers to following) shows a very tight distribution around the range 0.895 to 1.9, indicating that many people have similar numbers of in-degree connections (followers) as out-degree connections (friends). About 2K-50K overlapping network neighbors are typical, though some users with very large networks have over two orders of magnitude more.

Comparing Predictors of Follower Content

We now turn to the core of our results: how well do these variables predict follower growth over time and by how much? The overall significance of the negative binomial auto-regressive model is very high (p < 2e-16), meaning the model is very well-suited to characterizing the effects of the described variables on follower growth over time. Significance was determined by testing for the reduction in deviance from a null model, $\chi^2(22, N=1{,}836) = 5943.9 - 2111.9 = 3832.0$, p < 2e-16. This is important in order to have confidence when interpreting the regression coefficients of the model components (b and β), which are depicted in Table 2.

The unstandardized b coefficients in Table 2 are useful in that they can be directly interpreted according to the native units of each predictor: for each one unit change in the predictor variable, the log count of the response variable is expected to change by the respective b coefficient (all else being equal). While this is valuable for a broad range of prediction and forecasting purposes, we are also interested in comparing the *relative* impact of each predictor; we therefore report the standardize beta (β) coefficients (see also Figure 1 – not pictured are three of the control variables used in this study: extant friends and followers, age of account, and the lagged variable). As expected, these controls absorb comparatively large portions of the variance (see Table 2). We are interested in how much our other variables contribute above and beyond these controls.

Among the behavioral and social interaction variables, the Broadcast Communications Index (BroadcastComms), the burstiness measure (PeakTPH), and all three of the profile elements (length of description, URL, and location) each emerge as significant predictors of follower growth. The moderately strong negative effect of BroadcastComms (b = -1.02, β = -2.67e-04) suggests that having too many undirected messages will hinder audience growth. Interestingly, the Directed Communications Index (DirectedComms) was not significant in the model. Apparently, in the presence of all the other variables, the significance of social interactions using @replies and @mentions is muted, at least in terms of its effect on attracting *new* followers.

	b	Std. Err.	Std. *β*	*p*-value
NumTweetsPd	2.63e-04	1.62e-04	5.57e-05	0.104
PeakTPH	2.35e-02	4.94e-03	1.63e-04	1.96e-06***
DirectedComms	4.24e-03	3.37e-03	3.77e-05	0.208
BroadcastComms	–1.02	1.28e-01	-2.67e-04	1.89e-15***
ProfDescLen	3.09e-03	5.57e-04	1.72e-04	2.94e-08***
ProfHasURL	3.91e-01	7.14e-02	1.65e-04	4.27e-08***
ProfHasLocation	3.29e-01	1.52e-01	6.30e-05	0.03995 *
PosSentiRate	8.19e-01	1.96e-01	1.37e-04	2.87e-05***
NegSentiRate	–2.38	4.82e-01	-1.75e-04	7.53e-07***
InformContent	1.18	1.41e-01	3.31e-04	< 2e-16 ***
MeformerContent	–6.72e-02	1.99e-01	-1.12e-05	0.736
TopicFocus	3.75e-01	2.32	5.13e-06	0.872
UserTweetRT'd	9.53e-01	7.23e-02	4.60e-04	< 2e-16 ***
HashtagUseRate	–4.28e-01	1.12e-01	-1.23e-04	1.33e-04***
TReDIX	1.28e-01	4.22e-02	9.85e-05	2.43e-03 **
Reciprocity	3.52e-01	1.46e-01	7.95e-05	0.01597 *
Attn-Status	1.63e-02	4.48e-03	1.38e-04	2.79e-04***
NetworkOverlap	1.20e-06	1.26e-07	5.06e-04	< 2e-16 ***
NumFriends	–1.73e-04	2.88e-05	-5.98e-04	1.96e-09***
NumFollowers	2.70e-04	2.4e-05	1.10e-03	< 2e-16 ***
ChngFollPrevPd	–2.71e-04	8.82e-05	-1.79e-04	2.17e-03 **
AgeOfAccount	4.10e-03	2.26e-04	5.50e-04	< 2e-16 ***

Table 2: Neg. Binomial Auto-Regressive Model Coefficients.

Message content variables are evenly distributed along the rank ordered list of predictors (see Figure 1). Of the 17 (non-control) variables depicted, expressing negative sentiments in tweets is the second most harmful factor to growing a Twitter audience. Interestingly, overuse of hashtags in message content ("hashtag abuse") will also significantly reduce follower gain. On the other hand, producing or passing along informational content is among the top predictors, having a significant positive effect on follower growth rates (β = 3.31e-04). Also, having content that is "retweet worthy" is a very good indicator that a user will gain followers (β = 4.60e-04). Using more sophisticated language in messages also has a moderately strong relative effect on attracting and retaining followers (β = 9.85e-05).

Network oriented variables are also evenly distributed along the ranked list in Figure 1. Reciprocity, status, and network overlap were each significant in the model, even in the presence of the variables controlling for network size and user popularity.

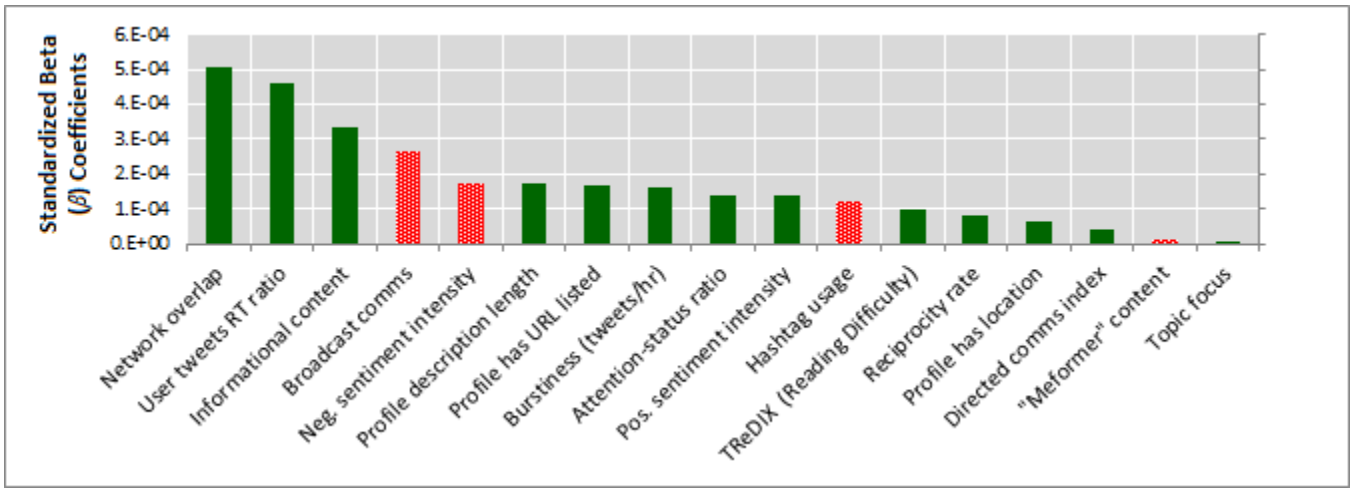


Figure 1: Standardized beta coefficients (β) show the relative effect sizes that each input variable has on follower growth. Green bars indicate positive effects on follower gain, and red bars indicate negative effects (i.e., suppression of follower growth).

DISCUSSION

It Matters What You Say, And How You Say It

Our first major finding is that *message content significantly impacts audience growth*. Six of our eight content variables (negative and positive sentiment, informational and "retweetable" content, hashtag usage, and linguistic sophistication) were found to be significant predictors of audience growth. We find that expressing negative sentiment has an adverse effect on follower gain. This is a contrast to [26], where social sharing of negative emotions correlates to higher numbers of followers. However, [26] studied a static snapshot of existing network ties. Our longitudinal data suggest that sentiment expression may have different (indeed opposite) effects on the formation of new ties over time. This might be because Twitter is a medium dominated by very weak social ties [16], and negative sentiment from strangers may be unpleasant or uncomfortable for a potential new follower to see. For existing ties, on the other hand, negative expressions such as the sharing of a death, poor health, bad news, or a state of unhappiness, can trigger opportunities to build bonding social capital between stronger ties who want to seek and provide emotional support [41]. Or, as [26] put it, "gift giving where users directly exchange digital 'gifts' in terms of emotional messages".

We also found that informational content attracts followers with an effect that is roughly *thirty times higher* than the effect of "meformer" content, which deters growth. We think this is due to the prevalence of weak ties on Twitter [16], and that informativeness [19,28] is a more palatable alternative to meforming among such networks. Kollock [27] describes information as a public good that anyone can consume and share. Retweeted content is another such digital public good that provides both attribution—and thus, motivation—to the original author as well as informational content for the community. Retweeted content also provides *social proof* [9] that a user may be worth following, enabling the process of triadic closure [13] to unfold, whereby followers of a user's followers complete the triad with the user [18].

The mean and median of topical focus (average cosine similarity of tweets) for our heterogeneous group is roughly an order of magnitude less than those same measures from a more homogenous group of politically-oriented tweeters described in [40], but like [40], we also find that topically focused users tend to attract more followers. Twitter users are likely driven by homophily [32], where they seek out content and users who are similar to themselves.

Finally, we found that the Tweet Reading Difficulty Index (TReDIX) has a positive impact on audience growth. Walther's Social Information Processing (SIP) theory suggests that people rely on linguistic cues like spelling and vocabulary to compensate for the lack of traditional contextual cues available in face-to-face settings [39]. Twitter users apparently seek out well-written content over poorly written content when deciding whether to follow another user.

Behavioral Choices Also Matter

Our second major finding is that *social behavioral choices can dramatically affect network growth*. Similar to previous research that showed positive effects of profile completeness for static Facebook networks [30], we find similar results for evolving Twitter networks. Signaling theory suggests that choosing to complete user profile elements helps persuade other users one's authenticity and trustworthiness, making them more likely to become followers [12]. Profile content provides at minimum *conventional signals* of identity (which are easy to fake), but the nature of profiles on social network sites makes these signals somewhat more reliable due to social accountability [12]. Regardless, users who do take the time to give profile information have the opportunity to emphasize the characteristics that they think will present them in the best light without necessarily being deceptive [17]. Others can use this profile information to form impressions prior to deciding whether to pursue or continue a connection [30]. Likewise, choices about inter-

actions and communication techniques, such as sending directed versus broadcast messages, will also impact the rate at which a user will grow their audience. Consistent with previous research studying existing Facebook networks [6], we find that directed communications have a positive effect on follower growth for Twitter. Unlike [6] however, we note a very strong negative effect of broadcast communication techniques during the process of network formation. Such undirected messages are a relatively novel feature of social media; our results suggest that relying on such communication techniques will significantly suppress growth.

Even Simple Measures of Network Structure Are Useful

Our third finding is that *variables related to network structure are useful predictors of audience growth*. This finding is not necessarily surprising, given the emphasis on such factors in much of the related literature [3,18,19,25,31]. Indeed, while our results indicate that even simplistic calculations of network structure can prove to be quite powerful, we stress that such factors should not necessarily be privileged over message content or social behavior measures.

Practical Implications

A vital prerequisite to building social capital of any kind (bonding or bridging) is that a connecting tie must exist between individuals. The practical implication of this fundamental antecedent to social capital motivates the selection of our dependent variable. The number of followers you have is arguably the most important status symbol on Twitter. Rapid follower growth may be an early indication of a rising star, or an emerging leader, within the network. A rapid gain in followers intuitively implies that people like what you're posting and want more of the same. Thus, social capital is a necessary (though not sufficient) precursor to the notion of *interpersonal influence* in social networks [2] – an attribute of interest to strategic communicators, marketers, advertisers, job seekers, activist groups and any entity or organization wishing to disseminate specific messages in a timely manner. Additionally, many users are simply interested in knowing their own relative degrees of popularity or social networking "clout". Sites like HootSuite.com and SocialFlow.com offer web services oriented towards helping its users capture and retain the attention of social media audiences. Companies like these can directly leverage our results to build tools that that make recruiting and retaining network members easier and more effective. For example, in conjunction with a validated tie-strength model (c.f., [15] or [16]), the results of our study suggest that social media technology developers can help users retain existing followers by actively promoting negative sentiment content to strong ties, and experiment with demoting it with weak ties. Similarly, to attract the attention of new audience members, developers can consider implementing user interface components which a) facilitate the sharing of informational content through positive reinforcement, b) encourage directed communications and group discussions, c) provide feedback regarding behavioral patterns (e.g., burstiness), and so on.

Theoretical and Methodological Implications

Our findings also have theoretical (and, by extension, methodological) implications. Our variables were selected from prominent theoretical perspectives bridging social science theory (social capital, signaling theory, presentation of self, homophily, status/power), and network theory (size and preferential attachment, tie strength, reciprocity, balance and closure). We also consider behavioral aspects of computer mediated communications (profile completeness, directed versus broadcast communication strategies) and message content (sentiment, informational versus meformer content, topical focus, linguistic sophistication). Few social media studies have attempted to report on relative impacts of such diverse variables. Compared to how much is known about each theory, very little is known about how they relate to one another. Our research compares their relative contributions to predicting link formations in online social networks. This was a significant undertaking, but more work should be done to understand the relative effects of different variables—as well as different theoretical perspectives and methodological approaches—on study outcomes.

Study Limitations

We have been as thorough as we can within the page limit. However, other variables could explain some of our results. For example, a person's real-world celebrity status, or other exogenous factors like being publicly mentioned in mass communications (news media, printed press, commercials and advertisements, etc.) may contribute to audience growth. Secondly, we do not segment our Twitter sample into types of users or types of uses, although [6], [33], and [37] suggest ways in which categories for specific user and uses may illuminate the processes of attracting network members. Thirdly, this is a quantitative study based on observations with calculated latent measures from those observations. Our approach is useful for describing *what* happens, but without a corresponding qualitative approach, we can only speculate on *why*. Future work could explore why certain variables predict follower growth more than others. Finally, Twitter is one site. We don't know if the results presented here translate into other sites.

CONCLUSION

We believe this is the first longitudinal study of audience growth on Twitter to combine such a diverse set of theory inspired variables. For the first time, we explore the relative effects of social behavior, message content, and network structure on follow behavior and show which of these has more power than the others. Though these results are specific to Twitter and a particular dataset, we think they are important for the following reasons. First, multiple snapshots can help us begin to offer casual explanations for audience growth. Second, comparisons across many variables inspired by different theoretical perspectives allow us to interpret relative effects of each. Third, the impact of message content and social behavior are comparative to network structure, which suggests future work should take caution in privileging any one perspective over another.

REFERENCES

1. Anderson, J. Lix and Rix: Variations on a Little-Known Readability Index. *J. of Reading 26*, 6 (1983), 490–496.
2. Bakshy, E., Hofman, J.M., Mason, W., and Watts, D. Everyone's an Influencer: Quantifying Influence on Twitter. *ICWSM* (2011).
3. Barabási, A.L. and Albert, R. Emergence of Scaling in Random Networks. *Science 286*, (1999), 509–512.
4. Bourdieu, P. The forms of capital. In J.C. Richardson, ed., *Handbook of Theory and Research for the Sociology of Education*. Greenwood, New York, 1985, 241–258.
5. boyd, danah, Golder, S.A., and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS* (2010).
6. Burke, M., Kraut, R., and Marlow, C. Social Capital on Facebook: Differentiating Uses and Users. *CHI* (2011).
7. Cameron, A.C. and Trivedi, P.K. *Regression Analysis of Count Data*. Cambridge University Press, NY, 1998.
8. Cartwright, D. and Harary, F. Structural balance: A generalization of Heider's theory. *Psychological Review 63*, 5 (1956), 277–293.
9. Cialdini, R.B. *Influence: the psychology of persuasion*. Collins, (2007).
10. Davidov, D., Tsur, O., and Rappoport, A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *COLING* (2010).
11. Donath, J.S. *Signals, Truth, and Design*. MIT Press, Cambridge, MA, forthcoming.
12. Donath, J.S. Social Signals in Supernets. *J. of Computer-Mediated Communication 13*, 1 (2007), article 12.
13. Easley, D. and Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, Cambridge, MA, 2010.
14. Feld, S. The Focused Organization of Social Ties. *The American J. of Sociology 86*, 5 (1981), 1015–1035.
15. Gilbert, E. and Karahalios, K. Predicting tie strength with social media. *CHI* (2009), 211–220.
16. Gilbert, E. Predicting Tie Strength in a New Medium. *CSCW* (2012).
17. Goffman, E. *The Presentation of Self in Everyday Life*. Anchor, 1959.
18. Golder, S.A. and Yardi, S. Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality. *IEEE SocialCom*, (2010).
19. Haewoon Kwak, Sue Moon, and Wonjae Lee. More of a Receiver Than a Giver: Why Do People Unfollow in Twitter? *ICWSM* (2012).
20. Hargittai, E. and Litt, E. The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption among a Diverse Group of Young Adults. *New Media & Society 13*, 5 (2011), 824–842.
21. Heider, F. Attitudes and cognitive organization. *The Journal of Psychology 21*, 1 (1946), 107–112.
22. Honeycutt, C. and Herring, S.C. Beyond Microblogging: Conversation and Collaboration via Twitter. *HICSS* (2009).
23. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. *Web-KDD* (2007), 56–65.
24. Katz, M.L. and Shapiro, C. Network Externalities, Competition, and Compatibility. *The American Economic Review 75*, 3 (1985), 424–440.
25. Kivran-Swaine, F., Govindan, P., and Naaman, M. The impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. *CHI* (2011).
26. Kivran-Swaine, F. and Naaman, M. Network Properties and Social Sharing of Emotions in Social Awareness Streams. *CSCW* (2011).
27. Kollock, P. The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. In M. Smith and P. Kollock, eds., *Communities in Cyberspace*. Routledge, New York, 1999, 220–239.
28. Kwak, H., Chun, H., and Moon, S. Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. *CHI* (2011).
29. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? *WWW* (2010).
30. Lampe, C.A.C., Ellison, N., and Steinfield, C. A familiar face(book): profile elements as signals in an online social network. *CHI* (2007), 435–444.
31. Liben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. *CIKM* (2003), 556–559.
32. McPherson, M., Smith-Lovin, L., and Cook, J.M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology 27*, 1 (2001), 415–438.
33. Naaman, M., Boase, J., and Lai, C.-H. Is it Really About Me? Message Content in Social Awareness Streams. *CSCW* (2010), 189–192.
34. Pandit, S.M. and Wu, S.-M. *Time Series and System Analysis With Applications*. Krieger, Malabar, FL, 2001.
35. Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*, 1-2 (2008), 1–135.
36. Pennebaker, J.W., Francis, M., and Booth, R. *Linguistic Inquiry and Word Count: LIWC 2001*. Erlbaum Publishers, Mahwah, NJ, 2001.
37. Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. In the Mood for Being Influential on Twitter. *IEEE SocialCom*, (2011), 307–314.
38. Resnick, P., Konstan, J., Chen, Y., and Kraut, R. Starting New Online Communities. In *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press, Cambridge, MA, 2012.
39. Tidwell, L.C. and Walther, J.B. Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations: Getting to Know One Another a Bit at a Time. *Human Communication Research 28*, 3 (2002).
40. Wang, Y.-C. and Kraut, R. Twitter and the development of an audience: those who stay on topic thrive! *CHI* (2012).
41. Wellman, B. and Wortley, S. Different strokes from different folks: Community ties and social support. *American J. of Sociology 96*, (1990), 558–588.
42. Zhang, J., Qu, Y., Cody, J., and Wu, Y. A case study of micro-blogging in the enterprise: use, value, and related issues. *CHI* (2010), 123–132.
43. Zhao, D. and Rosson, M.B. How and why people Twitter: the role that micro-blogging plays in informal communication at work. *GROUP* (2009), 243–252.