Twitter and the Development of an Audience: Those Who Stay on Topic Thrive!

Yi-Chia Wang
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
yichiaw@cs.cmu.edu

Robert Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
robert.kraut@cs.cmu.edu

ABSTRACT
Although economists have long recognized the importance of a critical mass in growing a community, we know little about how it is achieved. This paper examines how initial topical focus influences communities' ability to attract a critical mass. When starting an online community, organizers need to define its initial scope. Topically narrow communities will probably attract a homogeneous group of interested in its content and compatible with each other. However, they are likely to attract fewer members than a diverse one because they offer only a subset of the topics. This paper reports an empirical analysis of longitudinal data collected from Twitter, where each new Twitter poster is considered the seed of a potential social collection. Users who focus the topics of their early tweets more narrowly ultimately attract more followers with more ties among them. Our results shed light on the development of online social networking structures.

Author Keywords
Social networking sites, online communities, community startup, critical mass, topical focus, natural language analysis

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

INTRODUCTION
The vast majority of Internet users participate in some online groups or social networking sites [4]. Yet for every successful Facebook, Wikipedia or Linux project, there are dozens of failed attempts to start an online group. For example, 50% of IRC chat groups die within 24 hours after they are formed [11]. Understanding why some online groups succeed and many other fail is a practical and scientifically important research problem.

Online communities take many forms - from well-defined open source projects with defined boundaries, stable membership and clear production goals to the more diffuse social networks on Facebook, to short-lived IRC channels, to high turnover Usenet groups. These social collections differ on many dimensions – e.g., definition of boundaries, stability and homogeneity of membership, degree of interaction among members and the degree that members and outsiders think of the collection as a group. Twitter collections share some attributes of groups: they have a defined membership, some homogeneity of interests and some interaction among members. McMillan and Chavis's definition of community suggests that an online social collection will be a more successful group if it has enough members to provide resources and ties among members to facilitate social interaction and information exchange [8]. Therefore, in this study we consider larger membership size and higher social tie density as signs of a more successful online social group.

This paper examines whether the initial topical focus of content in new online groups predicts their subsequent success – eventually developing a larger and denser network of members. We investigate this question with longitudinal data from Twitter, a popular micro blogging service used by 8% of American Internet users [10].

Critical Mass and Initial Topical Focus
Interactive media must develop a critical mass of users if they are to be self-sustaining [7]. Theories of network externalities partially explain the process [6]. For goods and services characterized by network externalities, such as the telephone network, the value users receive grows with the number of users.

The content offered by someone starting an online group – its quality and scope – may determine the number and type of people who initially join it, thus triggering the upward or downward spirals associated with network externalities. The group founder has many choices about its scope. Some researchers believe that topical focus (or its inverse, content diversity) will influence an online group's ultimate success [12]. Network externalities suggest that a group should start with diverse content, because the diversity of topics can potentially appeal to more people, each of whom is interested in a subset of topics, and thus help the group more quickly reach critical mass. In contrast, starting with a

CHI 2012, May 5-10, 2012, Austin, TX, USA.

focused topic may restrict membership to the subset of the larger population that is interested in that topic.

H1a: The topical diversity of initial content in online groups will be positively correlated with their ultimate size.

On the other hand, when a community starts with a focused topic, people with similar interests can identify themselves more easily with it and can better anticipate what they can get from it. Moreover, groups organized around a narrow topic are likely to attract members who are similar to each other, and are likely to form relationships that maintain each others' commitment to the community because of homophily [9]. If this is the case, then initial content focus, not diversity, should lead to community success.

H1b: The topical focus of initial content in online groups will be positively correlated with their ultimate size.

H2: The topical focus of initial content in online groups will be positively correlated with the density of ties among members.

Little empirical research has examined the role of topical focus in online groups. Using *listserv* data, Butler showed that topic variation had both positive and negative effects on the sustainability of online groups, but did not look at the impact of topic variation on achieving a critical mass of users [3]. Adamic et al. analyzed knowledge sharing activity in an online question-answer forum [1]. They found that answerers who specialized in a narrow range of topics produced better quality answers. Since potential members are likely to cluster around groups that provide better content, we can infer that they will be more likely to join online groups with a focused topic.

We chose Twitter as the site for this research because of its popularity and ability to support needed data collection. We assume each new person who posts on Twitter is a seed of a potential social collection that can become more or less successful as a group. They can start by applying for a Twitter account and posting status updates. Others can join this new collection by becoming followers. In turn, these followers can connect with each other, creating a more interactive group. We examined the effect of the topical focus of the initial tweets on measures of group success – membership size and the density of ties among followers – a year later. We show that compared to Twitter posters who tweet about a broad range of topics, those who focus their initial tweets on a smaller range attract more followers and furthermore, these followers form a denser network.

TWITTER AND DATA SET

Twitter is a conversational micro-blogging website. It allows users to post short messages called "tweets" of at most 140 characters. Twitter provides several functions to facilitate social interaction and conversation among users. Users can give attention to others by "following" them. They can mention other users in their tweets. If Twitter users find someone else's tweet interesting, they can "retweet" it to share it with their own followers.

The data collection consisted of the three steps below.

Twitter Directories. Twitter directories organize Twitter users by genre and allow users to find other Twitter users to connect with. We downloaded the liberal and conservative political lists of Twitter users from three online Twitter directories – *Twellow.com*, *wefollow.com*, and *MyTwitterDirectory.com*. A total of 34,827 Twitter user identities were extracted.

User Selection. Since the formation of a new social group takes time, we sampled users who joined Twitter between August 2009 and December 2009 and examined their success a year later. Measuring topical focus requires a sufficient corpus of textual data, given that each tweet contains fewer than 140 characters. Therefore, we limited the sample to users who tweeted at least two times per day on average. 480 users fit these criteria.[1] On average, user in the sample had 1,574 tweets. This sample represents very active users.

User Information Extraction. We then retrieve the Twitter data from these 480 heavy Twitter users. We gathered all the tweets they posted, as well as the meta-information for each tweet, such as its timestamp, whether it was a reply, whether it was a retweet, etc. We also took a snapshot of these users' and their followers' social graphs, including all of the followers' followers, on November 20, 2010, about a year after the date they created their accounts.

METHODS

We examined the relationships between the initial topical focus of new users' online social groups and their success a year later. We measured topical focus using text processing techniques to assess the similarity among the tweets posted by a single user. We evaluated group success in terms of membership size and social tie density.

Membership Size and Tie Density

Specifically, we examined two dependent measures relating to the success of online social groups:

Membership Size is the total number of followers a user had on November 20, 2010.

Social Tie Density is the number of links among a user's followers. We normalized this measure by the number of followers a user has, so that it corresponds to the average number of links a follower has to other followers. The social tie density measure of a Twitter user is computed according to the following formula:

$$\text{SocialTieDensity}(\text{user}) = \frac{\text{num of links between followers}}{\text{num of followers}}$$

[1] Due to the limitation of Twitter API, this study only considered users who tweeted 3,200 or fewer times.

	Mean	Median	Std. Dev.	Min	Max
Membership Size	435.41	217.50	587.60	12	4916
Social Tie Density	23.98	10.26	32.82	0	204.45
AvgCosSim	0.07	0.06	0.05	0.01	0.59
NumTweet	250.60	121.50	339.23	0	2318
NumReply	34.44	27	31.51	0	141
NumRetweet	19.05	10	24.50	0	150
NumDay	408.77	415.00	45.25	324	476
Celebrity (X 1000)	405.2	1.6	52100	0	1000000
Politics	0.50	0	0.50	0	1

Table 1. Descriptive Statistics for Twitter Dataset

Users' social tie density increases when their followers are more connected to and potentially have more interactions with each other. As a consequence, the social group centered on a particular user is less likely to break down.

Topical Focus

Communication on Twitter is conducted through text messages. We measured users' initial topical focus by applying simple language processing techniques – average pairwise cosine similarity – to their first 150 tweets.

Average pairwise cosine similarity (AvgCosSim) is the similarity of vocabulary in a user's first 150 tweets. When people talk about one topic, they tend to use a common vocabulary. For example, words like "Obama", "Obamacare", "socialism" and "repeal" are frequently seen in tweets where conservatives discuss healthcare.

Cosine similarity measures similarity between two vectors of words by calculating the cosine angle between them in a high dimensional space. This study represents each tweet as a term vector consisting of the presence or absence of each word in the sample. High cosine similarity between two tweets indicates high text similarity. We first calculated the cosine similarity between all pairs of tweets produced by a single user and then computed the mean of all these pairwise cosine similarities. A high average cosine similarity indicates that a user's tweets were on similar topics and more topically focused. AvgCosSim can theoretically range from 0 to 1 and in the current sample ranged from 0 to .59 (mean .07 and median .06; Table 1).

Two examples of Twitter users with high AvgCosSim are *TwitterUser-A* (AvgCosSim=.08) and *TwitterUser-B* (AvgCosSim=.07). *TwitterUser-A* frequently wrote about Obama's health care policies, whereas *TwitterUser-B* was a former marine writing about military affairs and terrorism. Although neither was a celebrity, they attracted, 479 and 671 followers respectively. In contrast, *TwitterUser-C* has a low AvgCosSim (AvgCosSim=.04). He wrote about a wide variety of topics, such as 3C devices, politics, music, and so on. By November 20, 2010, he had only 69 followers.

We also measured the following control variables:

Dependent Var.	Membership Size		Social Tie Density	
Independent Var.	**IRR**	**Std. Err.**	**IRR**	**Std. Err.**
AvgCosSim	26.4016***	24.65	35.8315**	42.44
NumTweet (Log)	1.0187	0.02	1.0331	0.03
NumReply (Log)	0.7912***	0.03	0.9151*	0.04
NumRetweet (Log)	1.1271***	0.04	1.1671***	0.04
NumDay	1.0035***	0.00	1.0044***	0.00
Celebrity (Log)	1.0098	0.01	1.0107	0.01
Politics	1.4941***	0.14	2.4909***	0.27

*: p<0.05, **: p<0.01, ***: p<0.001

Table 2. The Effect of Topical Focus (H1 and H2)

Num of tweets (NumTweet) is the number of tweets a user produced in the first 60 days after joining Twitter. The restriction of the sample to heavy tweeters leads to underestimates of the effects of number of tweets on group success.

Number of reply tweets (NumReply) is a measure of how many tweets among the first 150 tweets posted by a user were reply tweets. A reply tweet is a tweet which is written to respond to someone else's tweet.

Number of retweets (NumRetweet) is a measure of how many tweets among the first 150 tweets posted by a user were retweets.

Days on Twitter (NumDay) is the number of days between the date a user joined Twitter and November 20, 2010.

Degree of celebrity (Celebrity): A social group started by a celebrity outside of Twitter can attract more followers, so we controlled for users' external popularity. This is estimated by the number of pages returned by Google when the screen name of the Twitter user is searched.

Politics is a binary variable describing a user's political view. 1 is conservative view; 0 is liberal view.

Table 1 describes descriptive statistics for the variables entered into regression models. For example, users posted 251 tweets in the first 2 months after joining Twitter.

ANALYSIS

The analysis seeks to identify the effects of topical focus on the number of followers a user eventually attracts and the density of ties among them. We used negative binomial regression models to predict membership size and social tie density. Negative binomial regression models are appropriate for count data, which are truncated at zero and highly non-normally distributed. Table 2 displays the results of the regression analysis with the dependent, independent, and control variables used to test the hypotheses. The effect of an independent variable on a dependent variable is reported using Incidence Rate Ratios (IRR). IRR is the change in a dependent variable expressed as a ratio when an independent variable increases by one

unit. An IRR of 0 means no change, 0.5 means halving the count, and 2 means doubling it.

RESULTS

In general, the regression results suggest that the initial topical focus of online social groups had a large impact on their success. The significant IRR for AvgCosSim in the two models predicting Membership Size and Social Tie Density suggests that topical focus has a strong positive impact on a group's ability to attract members and breed social connections among them. When AvgCosSim increases by 0.01 (from the 25th to the 50th percentile), users attracted on average 111 more followers. Moreover, each of their followers had on average 8 more connections within the community. The result supports H1b and H2 suggesting that initial topical focus may lead to more success turning a collection into a true online social groups; it disconfirms H1a that diversity would lead to success.

DISCUSSION AND FUTURE RESEARCH

This paper studied the relationship between topical focus and the formation of online social groups. The results demonstrated that at least in the domain of political discussion, more topically focused start-up groups are more likely to be ultimately successful. This finding suggests that people are more likely to be drawn to and join online groups with a focused topic, because they anticipate they will acquire content of interest to them and to meet people whose interests match their own.

If these results generalize, the implications are clear: Founders should start an online group with a well-defined topic initially in order to develop more audience with more connections among them.

This research examined the effects of topical focus within users who had self-identified as providers of political tweets. These users were already more likely to be topically focused than the typical Twitter user, who is less likely to limit posts to a single topic, like politics. Statistically this restriction of range can lead to underestimates of the power of topical focus on shaping the success of online groups. However, the effect may not be substantive, because one might expect that diversity injected into homogeneous discussions can enliven it [5]. Subsequent research will need to test the effects of topical focus on a more heterogeneous sample.

One might argue that cosine similarity can only measure degree of content overlap but not degree of focus in terms of content. In the future we plan to apply topic modeling techniques [2] to analyze text and infer the topic distribution of each group.

Finally, our current findings are based on a snapshot of users' social graphs. We can only make correlation claims about the relationship between starting conditions of a group and its later structure, not causal ones. Future research should monitor the growth of online groups. For example, founders of online groups would like to know if the effects of topical focus influence the number of new members they attract or the time they stay once they join.

ACKNOWLEDGMENTS

This research was supported by NSF grant IIS-0968485. We want to thank Shay Cohen who helped provide the machine for the data collection for this study.

REFERENCES

1. Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. WWW 2008*, 665-674.
2. Blei, D., Ng, A., & Jordan, M. Latent Dirichlet allocation. (2003). *Journal of Machine Learning Research, 3*(Jan), 993-1022.
3. Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research, 12*(4), 346-362.
4. Hampton, K. N., Sessions, L., & Her, E. J. (2011). Core networks, social isolation, and new media: How Internet and mobile phone use is related to network size and diversity. *Information, Communication & Society, 14*(1).
5. Harper, F. M., Frankowski, D., Drenner, S., Ren, Y., Kiesler, S., Terveen, L., Kraut, R., & Riedl, J. (2007). Talk amongst yourselves: Inviting users to participate in online conversations. In *Proc. ACM IUI07*, 62-71.
6. Katz, M. L., & Shapiro, C. (1985). Network Externalities, competition, and compatibility. *American Economic Review, 75*(3), 424-440.
7. Markus, L. (1987). Towards a "critical mass" theory of interactive media: Universal access, interdependence, and diffusion. *Communication Research, 14*(5), 491-511.
8. McMillan, D., & Chavis, D. (1986). Sense of community: A definition and theory. *Journal of Community Psychology, 14*(1), 6-23.
9. McPherson, J. M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415-444.
10. Pew Internet. 8% of online Americans use Twitter. (2011). *http://pewinternet.org/Reports/2010/Twitter-update-2010.aspx*, retrieved on Feb 4, 2011.
11. Raban, D., Moldovan, M., & Jones, Q. (2010). An empirical study of critical mass and online community survival. In *Proc. CHI 2010*, 71-80.
12. Resnick, P., Chen, Y., Konstan, J., & Kraut, R. E. (In press). Starting new online communities. In R. E. Kraut & P. Resnick (Eds.), *Building successful online communities: Evidence-based social design* (pp. 231-289). Cambridge, MA: MIT Press.