Gesundheit! Modeling Contagion through Facebook News Feed

Eric Sun[1], Itamar Rosenn[2], Cameron A. Marlow[3], Thomas M. Lento[4]

[1]Department of Statistics, Stanford University
[2,3,4] Facebook
[1]esun@cs.stanford.edu; [2,3,4] {itamar, cameron, lento}@facebook.com

Abstract

Whether they are modeling bookmarking behavior in Flickr or cascades of failure in large networks, models of diffusion often start with the assumption that a few nodes start long chain reactions, resulting in large-scale cascades. While reasonable under some conditions, this assumption may not hold for social media networks, where user engagement is high and information may enter a system from multiple disconnected sources.

Using a dataset of 262,985 Facebook *Pages* and their associated fans, this paper provides an empirical investigation of diffusion through a large social media network. Although Facebook diffusion chains are often extremely long (chains of up to 82 levels have been observed), they are not usually the result of a single chain-reaction event. Rather, these diffusion chains are typically started by a substantial number of users. Large clusters emerge when hundreds or even thousands of short diffusion chains merge together.

This paper presents an analysis of these diffusion chains using zero-inflated negative binomial regressions. We show that after controlling for distribution effects, there is no meaningful evidence that a start node's maximum diffusion chain length can be predicted with the user's demographics or Facebook usage characteristics (including the user's number of Facebook friends). This may provide insight into future research on public opinion formation.

Introduction

Diffusion models have been used to explain phenomena ranging from social movement participation to the spread of contagious diseases. Some of these models (e.g. Gruhl et al. 2004; Leskovec et al. 2007; Newman 2002) are an extension of epidemiological models of contagion, such as SIR or SIRS (Anderson and May 1991), while others introduce network-based features, such as thresholds for adoption (Centola and Macy 2007). While these models have contributed to our understanding of how diffusions spread across a population, most of them start with an isolated event and explore the conditions under which this event will trigger a global cascade. Although this is a reasonable approach to understanding certain diffusion problems, it may not be the best method for modeling the spread of information through a social media network. Furthermore, with a few notable exceptions, these models are developed without directly relevant empirical data to inform the assumptions and assess the validity of the models. In this paper, we use data from Facebook, a social networking service with over 175 million active users,[1] to empirically assess the conditions under which large-scale cascades occur within a social networking service. We also highlight key differences between cascades in social media and cascades that result from a single isolated event.

How is diffusion currently modeled? A basic contagion model starts with an event, which propagates through a network by spreading along the ties between infected and susceptible nodes. Theoretical models without an empirical basis often focus on isolating the importance of specific characteristics relevant to the diffusion process. Percolation models can isolate the relationship between transmission probability and the spread of an infection through a network (Moore and Newman 2000), while models focusing on structural and individual characteristics assess the impact of network topology or various thresholds for adoption (Granovetter 1978) on the likelihood of a cascade.

One of the most prominent models of the effect of network topology on diffusion shows that rapid global cascades are possible on highly clustered networks even with few ties connecting otherwise disconnected clusters (Watts and Strogatz 1998). More general models suggest that the likelihood of a global cascade – defined as a cascade that eventually reaches a sufficiently large proportion of the network – varies with network connectivity, degree distribution, and threshold distribution (Centola and Macy 2007; Watts 2002).

In addition to adoption thresholds and network topology, diffusion models also assess the importance of influence and connectedness at the level of the individual node. Watts and Dodds (2007) recently published a model specifically designed to determine whether or not "influential," or well-connected, nodes are more likely to trigger a global cascade. Their findings, which suggest that influential nodes are no more likely to trigger cascades than average nodes, run counter to popular suggestions from Gladwell (2000) and others that the key to mass popularity is to identify and reach a small number of highly influential actors (after which everyone else will be reached, essentially for free). These models have significant practical implications for marketers, particularly those who are interested in

[1] http://www.facebook.com/press/info.php?statistics

advertising through social media. However, practitioners and researchers in this area can also benefit from empirical models of how information spreads in social networks.

Models starting from an empirical base typically involve an algorithm that predicts the observed behavior of information propagation. These are commonly developed using blog or web data due to the availability of accurate information regarding links and the time at which data was posted or transmitted. Analysis of link cascades (Leskovec et al. 2007) and information diffusion (Gruhl et al. 2004) in blogs, as well as photo bookmarking in Flickr (Cha et al. 2008), can all be modeled by variations on standard epidemiological methods. Models of social movement participation and contribution to collective action (Gould 1993; Oliver, Marwell, and Teixeira 1985) are often tied to an empirical foundation, but with some exceptions (Macy 1991; Oliver, Marwell, and Prahl 1998) these do not focus as strongly on network properties and cascades.

In most network models of diffusion, the contagion is triggered by a fairly small number of sources. In some cases (e.g. Centola and Macy 2007; Watts and Strogatz 1998), the models are explicitly designed to assess the impact of an isolated event on the network as a whole. In other cases, such as in blog networks, the way information is introduced lends itself to scenarios where one or a few sources may trigger a cascade. This start condition therefore makes sense both in models focused on the endogenous effects of diffusion and in models based on empirical scenarios in which a few nodes initiate cascades. However, it does create a particular conceptualization of how diffusion processes work. At the basis of these models is the assumption that a small number of nodes triggers a large chain-reaction, which is observed as a large cascade. This assumption may not hold in social media systems, where diffusion events are often related to publicly visible pieces of content that are introduced into a particular network from many otherwise disconnected sources. Information will not necessarily spread through these networks via long, branching chains of adoption, but may instead exhibit diffusion patterns characterized by large-scale collisions of shorter chains. Some evidence of this effect has already been observed in blog networks (Leskovec et al. 2007).

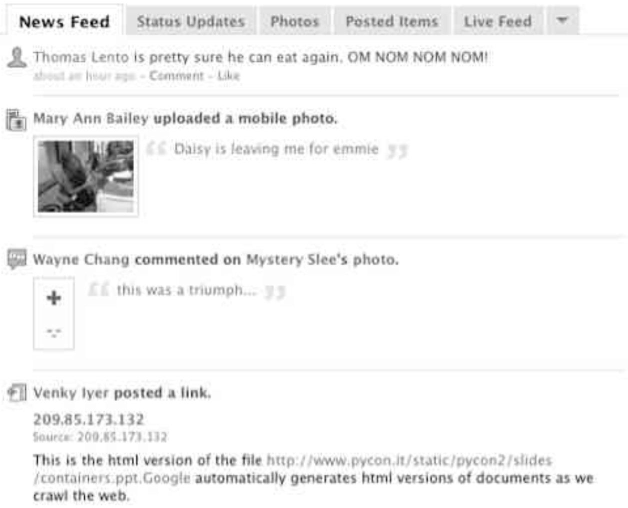
Even if this initial assumption is valid and cascades in social media are started by a tiny fraction of the user population, these models typically lack external empirical validation. Accurate data on social network structures and adoption events are difficult to collect and has not been readily available until fairly recently. In the cases where diffusion data have been tracked, it typically does not include the fine-grained exposure data necessary to fully document diffusion. Given the earlier data limitations facing empirical studies of diffusion, the primary goal of this study is to provide a detailed empirical description of large cascades over social networks.

Using data on 262,985 Facebook *Pages*, this paper presents an empirical examination of large cascades through the Facebook network. We first describe the process of *Page* diffusion via Facebook's News Feed, after which we provide introductory summary statistics that describe the chains of diffusion that result. Unlike previous empirical work, the data we present include every user exposure and cover millions of individual diffusion events. We assess the conditions under which large cascades occur, with a particular focus on analyzing the number of nodes responsible for triggering chains of adoption and the typical length of these chains. We then provide a detailed analysis of 179,010 "chain starters" over a six-month period starting on February 19, 2008 and investigate how their demographics and Facebook usage patterns may predict the length of the diffusion chains that they initiate.

Mechanics of Facebook Page Diffusion

Diffusion on Facebook is principally made possible by News Feed (Figure 1), which appears on every user's homepage and surfaces recent friend activity such as profile changes, shared links, comments, and posted notes. An important feature of News Feed is that it allows for *passive* information sharing, where users can broadcast an action to their entire network of friends through News Feed (instead of *active* sharing methods such as a private message, where a user picks a specific recipient or recipients). These stories are aggregated and filtered through an algorithm that ranks stories based on social and content-based features, then displayed to the friends along with stories from other users in their networks.

To analyze how ideas diffuse on Facebook, we concentrate on the News Feed propagation of one particularly viral feature of Facebook, the *Pages* product. Pages were originally envisioned as distinct, customized profiles designed for businesses, bands, celebrities, etc. to represent themselves on Facebook.



Thomas Lento is pretty sure he can eat again. OM NOM NOM NOM!
about an hour ago - Comment - Like

Mary Ann Bailey uploaded a mobile photo.

" Daisy is leaving me for emmie "

Wayne Chang commented on Mystery Slee's photo.

" this was a triumph... "

Venky Iyer posted a link.

209.85.173.132
Source: 209.85.173.132

This is the html version of the file http://www.pycon.it/static/pycon2/slides/containers.ppt.Google automatically generates html versions of documents as we crawl the web.

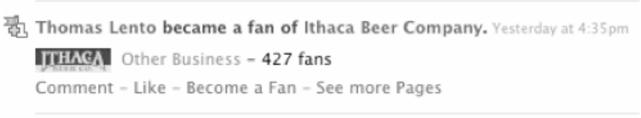
Figure 1. Screenshot of News Feed, Facebook's principal method of passive information sharing.

Since the original rollout in late 2007, hundreds of thousands of Pages have been created for almost every conceivable idea. Pages are made by corporations who wish to establish an advertising presence on Facebook, artists and celebrities who seek a place to interact with their fans, and regular users who simply want to create a gathering place for their interests.

Users interact with a Page by first becoming a "fan" of the Page; they can then post messages, photos, and various other types of content depending on the Page's settings. As of March 2009, the most popular Facebook Page was Barack Obama's Page,[2] with over 5.7 million fans.

When users become fans of a particular Page, their action may be broadcast to their friends' News Feeds (see Figure 2). An important exception is that the users may elect to delete the fanning story from their profile feed (perhaps to reduce clutter on their profile, or because they do not want to draw too much attention to their action). In this case, the story will not be publicized on their friends' News Feeds.

Diffusion of Pages occurs when 1) a user fans a Page; 2) this action is broadcast to their friends' News Feeds; and 3) one or more of their friends sees the item and decides to become a fan as well.



Thomas Lento became a fan of Ithaca Beer Company. Yesterday at 4:35pm

ITHACA Other Business - 427 fans

Comment - Like - Become a Fan - See more Pages

Figure 2. Sample News Feed item of Page fanning.

Without News Feed diffusion, we might expect that Pages acquire their fans at a roughly linear pace. However, since Facebook users are so active, actions such as Page fanning are quickly propagated through the network. As a result, we see frequent spikes of Page fanning, presumably driven by News Feed.

To motivate our subsequent analyses, Figure 3 shows the empirical cumulative distribution function of Page fan acquisition over time (the *x*-axis) for a random sample of 20 Pages. Each graph starts at the Page creation date and ends at the end of the sample period (August 19, 2008). From just this small sample of Pages, we see that there is no obvious pattern; clearly, Pages acquire their fans at highly variable rates. This paper will provide some insight into this phenomenon after the following section, which provides a more detailed description of the data used in the analysis.

Data

In this paper, we analyze Page data by creating trees that link *actors* and *followers* for each Page on Facebook. We measure diffusion via *levels* of a chain, as shown in Figure 4. An important note is that due to News Feed aggregation, users may see multiple friends perform a Page fanning action in a single News Feed story. For example, Charlie may see the following News Feed story: "Alice and Bob became a fan of Page XYZ." In this case, Charlie's node on the tree would have two parents. Furthermore, if Alice and Bob were on separate diffusion chains, the two chains have now merged. We extract every such chain from server logs that record fanning events and News Feed impressions for all of the Pages in our sample.

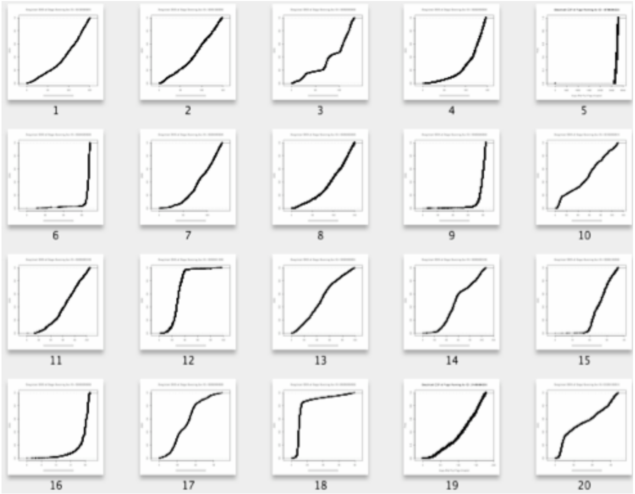


Figure 3. Empirical CDFs of Page fanning over time for a random sample of 20 Pages.

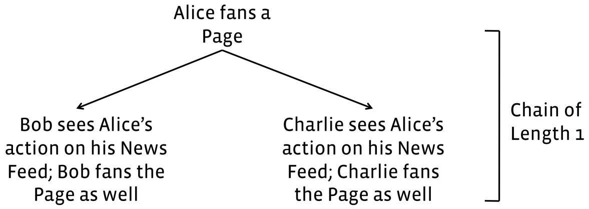

Alice fans a Page
Bob sees Alice's action on his News Feed; Bob fans the Page as well
Charlie sees Alice's action on his News Feed; Charlie fans the Page as well
Chain of Length 1

Figure 4. A possible diffusion chain of length 1.

We infer links of associations based on News Feed impressions of friends' Page fanning activity: if a user U_a saw that friend U_b became a fan of Page P within 24 hours prior of U_a becoming a fan, we record an edge from the follower U_b to the actor U_a. The 24-hour window is a logical breakpoint that allows us to account for various methods of Page fanning (e.g., clicking a link on the News feed, navigating to the Page and clicking "Become a Fan," etc.) without being overly optimistic in assigning links. As we create larger trees, some users (i.e., fans in the middle of a chain) may become both actors and followers; some users may be actors but not followers (the chain-starters); and some users can be followers but not actors (the leaves of the chains).

Our dataset consists of all Facebook Pages created between February 19, 2008 and August 19, 2008, and all of their associated fans (as of August 19, 2008). These data include 262,985 Pages that contained at least one diffusion event. Because any Facebook user can create a Page about any topic, there are a large number of Pages with just a few

[2] http://www.facebook.com/barackobama

fans. We will therefore limit our discussion to the subsets of these Pages that allow us to present more meaningful summary statistics.

Previous cascade models typically have an assumption of a "global cascade," which involves the finite subgraph of susceptible individuals on an infinite graph (Watts 2002). Our data represent the entire network of people who became a fan of a given Page before our cutoff date of August 19, 2008. Since these data do not present a clear definition of susceptibility, we assume that for popular Pages, the susceptible population consists of those that have adopted; in other words, for comparison's sake, we assume these diffusion events are large enough to be considered global cascades.

Chains Dataset

In addition to our general Pages data, we wish to observe the characteristics of chain length variation for different chain-starters. Due to the computational complexity of this particular analysis, we study this phenomenon by creating a *chains dataset* of 10 Pages and all 399,022 of their associated fans as of August 19, 2008 (of the 399,022 fans, 179,010 were chain-starters and 220,012 were followers). Each of these Pages was at least 40 days old as of the end of the analysis period and had at least 7,500 fans at that point. Given these criteria, a set of Pages were randomly selected and filtered to remove overlapping subjects and unrecognizable/foreign content that would be difficult for the authors to interpret. For each of these Pages, we gathered all of their associated fans and calculated the maximum chain length for each fan that started chains. We also collected various user-level features, such as age, gender, friend count, and various measures of Facebook activity. All data were analyzed in aggregate, and no personally identifiable information was used in the analysis. Table 1 shows the summary statistics for these Pages.

The next section discusses an analysis of the "large-clusters" phenomenon using the global Pages dataset. We then present an examination of the maximum chain length for each chain starter using the smaller chains dataset.

Page Name	Date Created	# Nodes	% in Biggest Cluster	% Single-tons	Max Chain Length
The Goonies	6/7	7,936	55.38%	14.37%	57
City of Chicago	4/6	9,277	20.91%	54.97%	23
Fudd-ruckers	3/17	10,269	43.71%	36.30%	27
Tom Cruise	5/28	28,592	56.67%	28.44%	41
Usain Bolt	6/1	33,967	37.03%	31.28%	14
Damian Marley	3/24	40,594	21.05%	45.91%	23
Stanley Kubrick	3/21	41,620	28.82%	40.30%	28
Cadbury	3/18	57,011	76.50%	16.39%	44
Zinedine Zidane	2/24	76,624	59.31%	25.95%	55
NPR	2/20	93,132	24.72%	33.63%	34

Table 1. Summary statistics for the chains dataset.

Large-Clusters Phenomenon

When the process described in Figure 4 is allowed to continue on a large scale, the result is that a flurry of chains, all started by many people acting independently, often merges together into one huge group of friends and acquaintances. This merging occurs when one person fans a Page after seeing two or more friends (who are on separate chains) fan that same Page.

A case in point is a Page devoted to a popular European cartoon, Stripy.[3] The diagram in Figure 5 shows the cartoon's close-knit communities of fans in both Bosnia (squares) and Slovenia (triangles). A few fans serve as the "bridge" that brings the two groups together. A third cluster of Croatian fans (diamonds, shown in the bottom-right cluster) hasn't yet found its connecting bridge. Finally, there are a few fans from other countries (circles) scattered within the two large clouds, perhaps Bosnian and Slovenian expatriates!

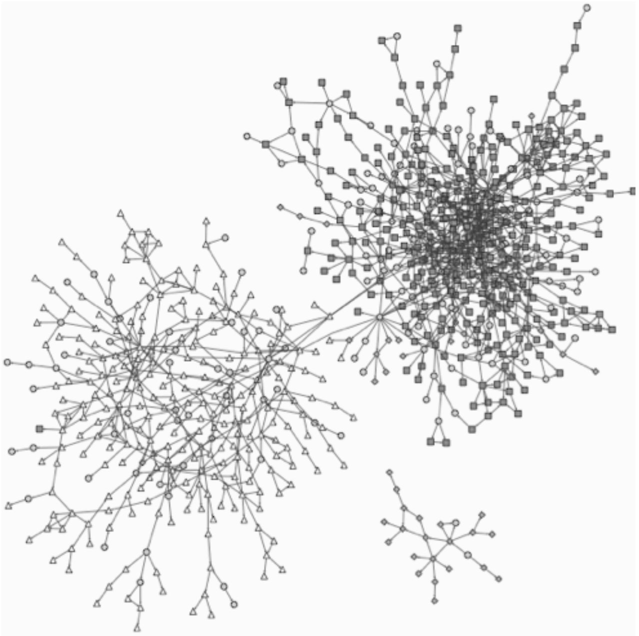


Figure 5. Trees of diffusion for the Stripy Facebook Page.

[3] http://www.facebook.com/pages/Stripy/25208353981

In fact, for some popular Pages (not part of our chains dataset), more than 90% of the fans can be part of a single group of people who are all somehow connected to one another. Typically, each of these close-knit communities contains thousands of separate starting points—individuals who independently decide to fan a particular Page.

As an example, as of August 21, 2008, 71,090 of 96,922 fans (73.3%) of the Nastia Liukin (an American Olympic gymnast) Page were in one connected cluster. Because the Beijing Olympics were going on at the time, there was a high latent interest in the Page. So, users were highly likely to fan the Page if a friend alerted them to its existence via News Feed's passive sharing mechanism.

This large-cluster effect is widespread, especially for recent Pages (it is more likely for older Pages to have distinct waves of fanning): for Pages created after July 1, 2008, for example, the median Page had 69.48% of its fans in one connected cluster as of August 19, 2008. A natural question is to wonder how these large clusters come about: are these clusters started by a very small percentage of the nodes, as is commonly assumed in the literature? Or does a different pattern emerge?

We analyze these data by looking at each Page and calculating the size of each cluster of fans. Each cluster consists of chains that have merged via the aggregated News Feed mechanism described earlier. For many Pages, the size of the largest cluster is orders of magnitude larger than the second-largest cluster: in the Nastia Liukin example, the second-largest connected cluster (after the 71,090-fan connected cluster) has only 30 fans.

After looking at the distribution of "start nodes" and "follower nodes" in these clusters, we find no evidence to support the theory that just a few users are responsible for the popularity of Pages. Instead, across all Pages of meaningful size (>1000 fans), an average of 14.8% (SD 7.9%) of the fans in each Page's biggest cluster were start nodes (for Pages of under 1000 nodes, the effect is also present, though variance increases).

Each of these fans arrived independently (presumably by searching for the Page via Facebook Search or from an advertisement) and started their own chains, which eventually merged together as the rest of the fan base took shape. These patterns hold fairly consistently for Pages with a few thousand fans and for those with more than 50,000: for Pages with 5000 fans or more, the average is 14.9% (SD 6.4%), and for Pages with 50,000 fans or more, the average rises to 17.1% (SD 4.8%).

Chains merge frequently because nodes in the graph typically have more than one parent. For all Pages with at least 100 fans, the average node in the largest cluster for each Page has a degree of 2.676 (SD 0.607). This figure increases to roughly 3 when the number of fans increases beyond 1000.

The rest of this paper investigates the aforementioned start nodes that begin chains of diffusion.

Prediction of Maximum Chain Length

Knowing that a large percentage of a Page's fans start page-fanning chains, we wish to further investigate what qualities separate these individuals from those that adopt via diffusion. Specifically, we wish to investigate the question: *given the demographic and Facebook usage statistics of each start node, can we predict the node's maximum chain length?*

Table 2 presents some summary statistics of our chains dataset to get better acquainted with the users that start chains. In our chains dataset, starters make up 46.32% of the users in the full dataset and 16.91% of the users in the largest cluster of each Page.

	Starters		**Non-Starters**	
	Mean	**SD**	**Mean**	**SD**
Age	24.65	9.82%	24.07	9.00%
Male	53.29%		55.60%	
Facebook-age	411.44	342.94	408.64	276.23
Friend-count	251.38	270.11	198.16	176.11
Activity-count	1.93	8.80	1.54	7.65

Table 2. Summary statistics for the chains dataset. All differences are statistically significant using a two-sample t-test.

Facebook-age denotes how long the user has been a member of Facebook, in days. *Activity-count* is a proxy to user activity, combining the total number of messages, photos, and Facebook wall posts added by the user.

We see that starters and non-starters have fairly similar statistics. However, start nodes tend to have more Facebook friends than their non-starter counterparts and have slightly larger *Facebook-ages*. Furthermore, as evidenced by the higher variation in *activity_count*, more of the start nodes are very active Facebook users. A likely explanation is that starters tend to be more frequent users of Facebook (evidenced by their increased content production), so they are more familiar with the interface and more likely to search for new Pages to fan. Non-starters, on the other hand, are more passive users of Facebook and are thus less likely to start diffusion chains.

Analysis of Start Nodes

For each of the 179,010 start nodes in our data, we calculate all the chains of diffusion and find each user's maximum-length chain. This value, *max_chain*, is the response variable in our data. For the Pages in our dataset, values range from 0 to 56. The data are heavily skewed to the right, indicating the presence of many short chains. Selected percentiles of *max_chain* are given below.

0%-67%	**68%**	**75%**	**90%**	**95%**	**98%**
0	1	1	3	5	10

However, we know for a fact that there are excess zeros in our *max_chain* data: if a user fans a Page and immediately deletes it from their profile feed, the story will no longer be eligible for broadcasting to their friends' News Feeds, regardless of how popular the user is. Thus, *max_chain* will always be exactly zero in this scenario. Data on excess zeros are not available, so for illustrative purposes we present selected percentiles of *max_chain* where all zeros have been deleted:

25%	**50%**	**60%**	**75%**	**95%**	**98%**
1	2	3	3	11	18

Typically, Poisson regression is used to model count response variables. However, Poisson random variables are expected to have a mean equal to its variance, which is clearly not the case here (where the variance far exceeds the mean). Instead, we use negative binomial regression, which is appropriate when variance >> mean. To correct for the excess zeros, we use a *zero-inflation* correction. This procedure allows us, in a single regression, to select variables that contribute to the true content in the response variable ("count model coefficients") and also a (potentially different) set of variables that contribute to the excess zeros in the response ("zero-inflation model coefficients").

The predictor variables for the count model are:

- log *age*
- *gender*
- log *Facebook_age* (number of days the user has been a member of Facebook)
- log *activity_count* (messages sent + photos uploaded + Facebook wall posts sent)
- log *friend_count* (number of Facebook friends)
- log *feed_exposure* (number of friends who saw the News Feed story broadcasting the user's fanning action)
- log *popularity* (number of friends that "care about" the start node high enough that the News Feed algorithm considers broadcasting the start node's Page fanning story)

The variables *age*, *gender*, *Facebook-age*, and *activity-count* are used for the zero-inflation model. We assume that the number of friends and level of News Feed exposure would not impact the probability of deleting the Page fanning story from a user's profile feed.

Table 3 presents the correlation matrix for these variables. There is a fairly high correlation between *friend_count*, *feed_exposure*, and *popularity*, but it may still be useful to include these in the model.

	1	2	3	4	5	6	7	8
1. max_chain	1.00	-0.07	0.00	0.05	0.00	0.17	0.28	0.04
2. age	-0.07	1.00	-0.06	0.11	0.07	-0.15	0.07	0.07
3. gender	0.00	-0.06	1.00	-0.03	-0.08	0.06	-0.01	-0.15
4. facebook_age	0.05	0.11	-0.03	1.00	0.10	0.33	0.37	0.19
5. activity_count	0.00	0.07	-0.08	0.10	1.00	0.20	0.12	0.34
6. friend_count	0.17	-0.15	0.06	0.33	0.20	1.00	0.45	0.51
7. feed_exposure	0.28	0.07	-0.01	0.37	0.12	0.45	1.00	0.32
8. popularity	0.04	0.07	-0.15	0.19	0.34	0.51	0.32	1.00

Table 3. Correlation matrix for model variables.

We run a standard zero-inflation negative binomial ("ZINB") model with starting values estimated by the expectation maximization (EM) algorithm. Standard errors are derived numerically using the Hessian matrix.

The ZINB coefficients for the pooled model (all 10 Pages) are shown in Table 4.

Count model coefficients (negbin with log link):

	Estimate	Std. Error	Pr(>|z|)	
(Intercept)	2.346007	0.083646	< 2e-16	***
age	-0.814130	0.016249	< 2e-16	***
gender==male	-0.084873	0.010606	1.22e-15	***
Facebook_age	-0.379611	0.010994	< 2e-16	***
activity_count	-0.056424	0.007047	1.18e-15	***
friend_count	0.067955	0.008139	< 2e-16	***
feed_exposure	0.929996	0.005766	< 2e-16	***
popularity	-0.206341	0.005110	< 2e-16	***
Log(theta)	-0.960615	0.007173	< 2e-16	***

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	Pr(>|z|)	
(Intercept)	24.42513	1.91443	< 2e-16	***
age	-0.06867	0.20232	0.73430	
gender==male	-0.18038	0.14023	0.19834	
Facebook_age	-5.31638	0.33802	< 2e-16	***
activity_count	-0.51408	0.17925	0.00413	**

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 0.3827
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.045e+05 on 14 Df
```

Table 4. Coefficients for the ZINB model.

To ensure significance of our model, we run a likelihood ratio test:

```
      Df  LogLik  Df  Chisq  Pr(>Chisq)
1     14  -204532
2     3   -224739 -11 40414  < 2.2e-16 ***
```

Here, Model 1 is the ZINB model calculated earlier, and Model 2 is *max_chain* regressed only on a constant term. The p-value is < 0.001, ensuring that our model is significant. However, we also wish to confirm that the ZINB model is an improvement over a standard negative binomial regression model with the same coefficients (that is, without the zero-inflation correction).

This can be accomplished by running a standard negative binomial regression (not shown here) and comparing the two models with the Vuong test (Vuong 1989). The Vuong test gives a small p-value if the zero-inflated negative binomial regression is a statistically significant improvement over a standard negative binomial regression.

```
Vuong Non-Nested Hypothesis Test-Statistic:
7.602584
(test-statistic is asymptotically distributed
N(0,1) under the null that the models are
indistinguishible)
model1 > model2, with p-value 1.454392e-14
```

Model 1 is the ZINB model; model 2 is the regular negative binomial model. The Vuong test reports a p-value < 0.001, so we conclude that the ZINB model is significant.

In Table 5, we present the count model coefficients for ZINB models on a selection of individual Pages from our chains dataset (coefficients significant at the 5% level are in **bold**). For all of these regressions, the likelihood ratio test and Vuong tests were significant at the 5% level.

Variable	Goon.	Fudd	Cruise	Bolt	Zidane
(Intercept)	-0.097	**2.574**	**1.419**	0.596	-0.105
age	0.509	**-0.537**	**-1.006**	**-0.386**	**-0.981**
gender==male	0.186	-0.011	-0.076	**-0.144**	**0.116**
Facebook_age	**-0.654**	**-0.522**	-0.052	**-0.415**	**0.153**
activity_count	0.019	-0.102	**-0.142**	**-0.064**	**-0.100**
friend_count	**-0.480**	**-0.220**	0.087	-0.023	0.009
feed_exposure	**1.379**	**1.279**	**1.008**	**0.860**	**1.053**
popularity	0.084	-0.014	**-0.245**	0.021	**-0.120**

Table 5. Count model coefficients for selected Pages.

Analysis of Chains Regressions

The zero-inflation model coefficients represent the increased probabilities of zero-inflation; that is, the probability that a user immediately removes the Page fanning story from their profile and constrains their *max_chain* to zero. Since this is not a focus of this paper, we concentrate instead on the count model coefficients.

We note that in the pooled model, all count model coefficients are highly statistically significant. However, this is not surprising given the large sample sizes observed here. If we run the same ZINB model using data from each Page separately, we find that the signs on every variable frequently flip from positive to negative, with the exception of *feed_exposure*. Furthermore, our coefficients from the pooled model are generally quite small: for example, the coefficient on log *activity_count* is -0.056424, which implies that a 1% increase in *activity_count* is only expected to decrease the log of *max_chain* by 0.056%. We might expect that highly active users are more likely to have their News Feed actions ignored, but this is a trivial decrease that is not realistically meaningful.

When comparing the results from the pooled model with results from the individual Page models, we see that the only consistently significant effect is for *feed_exposure*, which is a control for the number of friends who saw the News Feed story broadcasting the user's fanning action. This coefficient consistently hovers around 1, which indicates that if the News Feed algorithm decides to publish the user's action to 1% more people, we would expect a 1% longer *max_chain* to result. This result holds even after controlling for distribution (via the *popularity* variable) and for the number of friends.

The most interesting finding seems to be that after controlling for *feed_exposure*, the log *friend_count* variable does not have a realistically meaningful coefficient (0.067955, meaning that a 1% increase in *friend_count* is only expected to increase the log of *max_chain* by 0.068%). That is, after controlling for News Feed exposure variables, neither demographic characteristics nor number of Facebook friends seems to play an important role in the prediction of maximum diffusion chain length.

Conclusions and Future Work

Our examination of Facebook Pages shows that large-scale diffusion networks play a significant role in the spread of Pages through Facebook's social network. For many Pages with large followings, the majority of fans occur in a single connected cluster of diffusion chains that merge together to form a global cascade. The structure of these clusters reveals several important aspects of the empirical nature of global cascades. First, we find that within most clusters, roughly 14-18% of the nodes are chain initiators, which differs from the more restricted start conditions generally assumed in the theoretical literature. While it is easier to discuss theoretical underpinnings of a global cascade by disallowing exogenous diffusion, we find that this may not lead to a completely accurate conceptualization of the media diffusion observed in our studies.

We also find that because of the connectivity of the Facebook network and the ease of Page fanning, the maximum length of diffusion chains from initiator nodes can sometimes be extremely long, especially in comparison to the diffusion chains that have been observed in other empirical studies of real-world phenomena. In fact, we have observed chains of up to 82 levels in our complete dataset. It may be interesting for marketers and practitioners to note that when compared to real-life studies of diffusion, Facebook chains of Page fanning tend to be longer-lasting and involve more people: in a study of word-of-mouth diffusion of piano teachers, Brown and Reingen

(1987) found that only 38% of paths involved at least four individuals. Using the same definition on our data, we find that for a random sample of 82,280 Pages, 86.4% of paths of Page diffusion involve at least four individuals. This result may be useful for potential advertisers considering a Facebook marketing campaign versus more traditional word-of-mouth methods.

The properties of these diffusion clusters on Facebook suggest a new characterization of global cascades: whereas the theoretical literature generally assumes that a global cascade is an event that begins with a small number of initiator nodes that are able to affect vulnerable neighbors, we find that global cascades are in fact events that begin at a large number of nodes who initiate short chains; each of these chains quickly collide into a large single structure.

In addition, we investigate the length of the diffusion chain that each initiator triggers in order to understand whether there are aspects of certain initiator nodes that help determine the eventual impact of those initiators on the overall diffusion cluster. We find that after controlling for distribution access and popularity, a particular initiator's demographic properties and site usage characteristics do not appear to have any meaningful impact on that node's maximum diffusion chain length. The only way to increase maximum diffusion chain length, in fact, is to increase the likelihood that a Page fanning action appears in other users' News Feeds. Thus, there may not be a simple and easy way to identify initiators that "matter most," and it may be that a wide variety of individuals are equally likely to trigger a large global cascade.

These results are undoubtedly tied to the unique setting of our study. Our observations may partially be shaped by social behavior that is specific to Facebook: for example, the distinct characteristics of the Facebook user population, the specific social norms that dictate interaction and influence on the site, and the ways that users perceive and relate to Facebook Pages in contrast to other forms of content. In addition, our findings may hinge on the nature of the News Feed algorithm that determines which activity to surface, how that information is displayed, and for how long it is exposed to the user. Nevertheless, our conclusions remain important; they are the results of the first study of a large number of real contagion events on a social network that accurately captures the genuine social ties that exist between people in the real world.

Further research can expand our empirical understanding of contagion in many ways. First, we can explore the properties of initiator nodes to see whether demographic, social, or structural characteristics shape the ability of a node to trigger a cascade. In addition, we may want to evaluate how accurately various theoretical models of diffusion account for the empirical phenomena we have uncovered. We may also want to test experimental contagion events to better understand how different pieces of content and different start conditions determine the eventual structure of a diffusion cascade. Finally, we may wish to compare the structure and dynamics of the Pages diffusion network to other viral features on Facebook, such as Notes (as illustrated by the recent "25 Things About Myself" meme, for example) and Groups, where diffusion may occur both due to News Feed and due to active propagation (i.e., users may send invitations to join a group).

References

Anderson, R., and May, R. 1991. *Infectious Diseases of Humans: Dynamics and Control*. New York, NY: Oxford University Press.

Brown, J. J. and Reingen, P. H. 1987. Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research,* 14: 350-62.

Centola, D., and Macy, M. W. 2007. Complex Contagions and the Weakness of Long Ties, *American Journal of Sociology,* 113: 702-34.

Cha, M. et al. 2008. Characterizing Social Cascades in Flickr. In *Proceedings of ACM SIGCOMM Workshop on Online Social Networks (WOSN).*

Gladwell, Malcolm. 2000. *The Tipping Point*. Boston, Mass.: Little Brown.

Gould, R. V. 1993. Collective Action and Network Structure. *American Sociological Review*, 58: 182-97.

Granovetter, M. 1978. Threshold Models of Collective Behavior. *American Journal of Sociology* 83: 1420-43.

Gruhl, D. et al. 2004. Information Diffusion Through Blogspace. In *Proceedings of the International World Wide Web Conference (WWW*).

Leskovec, J. et al. 2007. Cascading Behavior in Large Blog Graphs. In *SIAM International Conference on Data Mining.*

Macy, M. 1991. Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review* 56: 730-47.

Moore, C. and Newman, M. E. J. 2000. Epidemics and Percolation in Small-World Networks. *Phys. Rev. E* 61: 5678-82.

Newman, M. E. J. 2002. The Spread of Epidemic Disease on Networks. *Phys. Rev. E* 66: 016128.

Oliver, P., Marwell, G., and Teixeira, R. 1985. A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action. *American Journal of Sociology* 91: 522-56.

Oliver, P., Marwell, G., and Prahl, R. 1988. Social Networks and Collective Action: A Theory of the Critical Mass. III. *American Journal of Sociology* 94: 502-34.

Vuong, Q. H. 1989. Likelihood Ratio Tests For Model Selection and Non-Nested Hypotheses. *Econometrica* 57: 307-33.

Watts, D. J. 2002. A Simple Model of Information Cascades on Random Networks. *Proceedings of the National Academy of Science 99*: 5766-71.

Watts, D. J. and Dodds, P. S. 2007. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research* 34: 441-58.

Watts, Duncan J., and Strogatz, S. H. 1998. Collective Dynamics of "Small-World" Networks. *Nature* 393: 440-2.