

Information Cascades in Feed-based Networks of Users with Limited Attention

Sameet Sreenivasan^{*}, Kevin S. Chan[†], Ananthram Swami[†], Gyorgy Korniss^{*} and Boleslaw Szymanski^{*}

^{*}Rensselaer Polytechnic Institute, Troy, NY 12180

Email: {sameet, korniss, szymab}@rpi.edu

[†]US Army Research Laboratory, Adelphi, MD 20783

Email: kevin.s.chan.civ@mail.mil, a.swami@ieee.org

Abstract—We build a model of information cascades on feed-based networks, taking into account the finite attention span of users, message generation rates and message forwarding rates. Through simulation of this model, we study the effect of the extent of user attention on the probability that the cascade becomes viral. In analogy with a branching process, we estimate the branching factor associated with the cascade process for different attention spans and different forwarding probabilities, and demonstrate that beyond a certain attention span, cascades tend to become viral. The critical forwarding probabilities have an inverse relationship with the attention span. Next, we develop an analytical and numerical approach that allows us determine the branching factor for given values of message generation rates, message forwarding rates and attention spans. The branching factors obtained using this analytical approach show good agreement with those obtained through simulations. Finally, we analyze an event-specific dataset obtained from Twitter, and show that estimated branching factors correlate well with the cascade-size distributions associated with distinct hashtags.

Index Terms—social computing, information cascades, limited attention

I. INTRODUCTION

With the pervasiveness of social networking platforms, users are highly connected and have the ability to generate and forward information across networks. Feed-based networks such as Facebook’s news feed, Twitter, and Instagram have become immensely popular,

Research was in part sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (Network Science CTA) and by the Office of Naval Research Grants No. N00014-09-1-0607 and N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

and allow users to scroll through streams of the latest information from sources that the user chooses to follow. It has become increasingly apparent that the volume of information being generated is far greater than the amount of information that users of these social networks can consume. Despite this preponderance of information, certain messages are able to “go viral”, meaning that the message is forwarded and seen by most (or at least an asymptotically finite fraction) of the users in the social network. Conversely, there are many messages that are seen by few people and forwarded by no one.

Given the preponderance of feed-based social networks, users will only look through a limited number of messages in their feed (before getting bored, fatigued, or interested in another topic or feed). If a particular message is deemed to be interesting, then the message is forwarded to the user’s followers. As a consequence of the variety in the number of feeds one can monitor, individual messages end up competing for limited attention span resulting in a fat-tailed distribution of the popularity of individual messages [1], [2].

Our main contribution is defining a model of networked user behavior within feed-based networks to evaluate the impact of the limited attention of users. With this model, we establish a method to identify if a particular topic will result in large cascades. The proposed method and models are validated using a dataset of tweets for several topics. Our model is based on three parameters that define the behavior of individual users: the probability of a user generating a new message, the probability that the user forwards a given message in his feed, and the length of the user feed. Simulations of our model support the idea that the dynamics of an information cascade follow a branching process [3], and how far the given branching process is above or below criticality depends on the depth of the feed to which a user typically devotes his attention. At low values of forwarding probability, the probability distribution of cascade-size distributions is a power-law with an

exponential cutoff indicative of a sub-critical branching process, while at large values of forwarding probability global cascades become increasingly prevalent. We analyze this model and the resulting branching process in detail, and also investigate a real dataset of tweets for signatures of the behavior suggested by our model.

A brief overview of related work is found in Section II. The model is introduced in Section III and the analysis is described in Section V. In Section IV, we use a set of tweets to demonstrate that signatures of critical, sub-critical and supercritical behavior are also present in empirical data and therefore qualitatively support our model.

II. BACKGROUND AND RELATED WORK

In the past five years, the topic of information cascades has caught the attention of both social network and network analysis researchers. Sociologists have studied attention span for various media and marketing campaigns using social media. Exploring the relationship between the humans and social media platforms has received significant attention especially in terms of its impact on human performance and behavior [4], [5] and economic gains [6], [7]. With proper models of impact of attention on social behavior, designers of social media platforms can control or maximize effectiveness or performance of their users or their platforms. In particular, recent theoretical work by Weng [1] and Gleeson [2] have shown how simple models of message forwarding can give rise to a fat-tailed distribution for the degree of spread (popularity) of a given message. In Section III, we comment on how these models are similar to and different from the one we consider in this work.

Using attributes of the underlying graph as explanatory variables is a common approach to studying information cascades on networks. Typically, these networks have either directional links, through follower and followee relationships or bi-directional links, through friendship statuses. Higham et al. [8] consider a model to predict spikes of activity while considering various graph-theoretic characteristics of these networks. In contrast, our analysis is based on abstracting the cascade as a branching process, and we apply this approach to data generated from simulations on scale-free (SF) networks, and to data collected from Twitter.

Another approach common in studying information cascades in social networks is to consider community structure. Feng et al. use a contagion approach to study attention in social media with a fractional SIR model [9]. Weng et al. [10], [11] investigate the impact of community structure on spreading of memes. The dynamics of these cascades are studied with respect to

parameters of various complex contagion models. Others have used epidemic models to investigate the dynamics of information cascades. Jin et al. [12] study a four-parameter epidemic model and fit the model to empirical data. Analogs to states in this type of model are: the user is *susceptible* when open to viewing messages of a particular topic; *exposed* while a message is in its queue; *infected* after the message is forwarded, *recovered* when it can return to the susceptible state. Our approach does not allow for such a recovered state in that the user retweets a message at most once.

We do not represent or study message content nor do we conduct sentiment analysis on the data. Modeling of these dynamics can provide more insight into the behavior of the users and global dynamics [13]. Further, for tweets containing hashtags of interest, the tweet may represent negative or positive support of the topic, which may also provide additional dimensions of study.

Finally, the model in this work does not consider placement of messages in feeds as conditions for virality. Hodas and Lerman [14], [15] study the impact of the placement of the messages in the queue and its impact on social contagion, where messages at the top of queues are more likely to be viewed. One can also consider other characteristics of the messages such as which other users forwarded the messages as well as their relationship to the current active user. In this work, our model is a FIFO stream.

A work related to biases in data collection, [16] compares the full Twitter feed *Firehose* with the sampled *Gardenhose* Twitter Streaming API to which the majority of researchers have access.

III. MODEL

We abstract the social network under consideration as a graph G . For simplicity we assume the graph to be undirected, but results found using directed edges are qualitatively similar. We assume that the degree distribution is scale-free as indicated by previous analyses of connectivity in social networks [17]. The parameters governing message propagation in our model are as follows:

- p_n : probability of new message generation
- p_r : probability of message forwarding
- L : length of message buffer.

The dynamics of message cascading results from the following rules. In each time slot t , one user (node i) is randomly selected to become active, and then will choose to create a new message with probability p_n . If a new message is generated, then each of the followers of node i will receive this message in their message

feed, and, if needed, their oldest message is removed from their buffers. Then, node i goes through its feed of up to L messages one by one, starting from the most recent message. With probability p_r , each message will be independently forwarded to its neighbors. We assume that a node does not forward the same message twice. When a user receives a message that has been already forwarded, it will ignore this message although the message still occupies a place on its feed. The duration of a time slot is the time required for a node to go through the above set of $L + 1$ actions.

Given this model, we study how the popularity of a message - the number of users forwarding that message - is statistically distributed. Note that in the rest of the paper, we interchangeably use the terms *queue* and *feed*.

Similar models have been studied in [1] and [2], but with the crucial difference that when a node is selected for an update, it either generates a new message or forwards a randomly chosen message from its queue. The former study has additional parameters in the model as well, making it analytically less tractable. In contrast the model in [2], as shown therein, is analytically tractable; however the system approaches criticality only when the message generation probability tends to zero. Furthermore, a regime of supercritical behavior is not attainable. This is another point where our model contrasts with that of [2] as we shall demonstrate in forthcoming sections both via theory, as well as analysis of empirical data.

IV. RESULTS

A. Simulations

In this section, we describe the set of simulations conducted to test the model described in Section III and to gain understanding of the conditions under which one can expect global cascades, *i.e.* a significant fraction of nodes forwarding the message.

We simulate the model on scale free networks consisting of $N = 10,000$ nodes with a power-law exponent of $\gamma = 2.5$ and average degree $\langle k \rangle = 10$, unless otherwise specified. Ten independent simulations were run, each for $T = 40 \times N = 4 \times 10^5$ timesteps. One timestep consists of one user being activated, potentially generating a new message, and going through its queue. To account for transient behavior of the queues, we begin collecting data on the number of users sharing specific messages only for those messages created after $t = 10^5$ timesteps. Increased transient period did not affect the results significantly. Unless otherwise stated, we use $p_n = 0.45$. We refer to the number of users forwarding a given message as the *cascade size* associated with that message.

Figs 1a-1c show results from simulations of various values of forwarding probability p_r and depth of the queue L . Observations are that for high forwarding probabilities, $p_r = 0.98$, global cascades are present for all L shown. For low forwarding probabilities $p_r = 0.02$, the cascade sizes are distributed exponentially. For the intermediate forwarding probability shown $p_r = 0.4$, the cascade-size distribution transitions from power-law to exhibiting super-critical behavior of global cascades.

In Fig. 1a, for the lowest value of forwarding probability, cascade sizes are distributed exponentially, while at the intermediate value of $p_r = 0.4$, the distribution has a robust linear behavior (on the log-log plot) upto a specific cascade size, suggesting a power-law behavior of the cascade-size distribution, modulated by an exponential cutoff. Finally, at the largest value of forwarding probability, the mass of the distribution appears to be increasingly shifting to the tail. The latter behavior becomes more pronounced as the queue length L is increased - see Figs 1b and 1c. We also observe the same general pattern of behavior for other values of p_n , as shown in Figs 2a and 2b. Thus the choice of $p_n = 0.45$ in the simulations is not critical. Simulations of our model on networks using the Erdős-Rényi (ER) random graph generation model indicate behaviors similar to those above; therefore we can assume that the criticality threshold for cascades exists for other random graphs, but the threshold may differ. Simulation results are shown in Fig 3. To obtain further insights, and to explore possible scaling behavior, we simulated our model for various values of L and p_r , keeping their product fixed. As seen from Fig. 4, for small values of this product [$Lp_r = 2$, Fig. 4(a)] there is no sign of scaling, while for larger values of this product [$Lp_r = 40$, Fig. 4(b)] all distributions collapse approximately to the same curve, indicating that the relevant parameter in this regime is Lp_r . In general, the transitions in the form of the distribution are indicative of a branching process like behavior which many systems subject to cascades have been found to follow [18], [19]. Specifically, a branching process can be sub-critical, critical or super-critical, with the respective tree-sizes generated by the branching process being distributed exponentially, as a power-law, and bimodally with more and more of the distribution becoming concentrated at the higher mode.

Next, we estimate the critical forwarding probability p_r at which the cascade distributions can be assumed to attain their power-law behavior as follows. The goal in general is to determine the “branching factor” of the process *i.e.* the average number of copies that a typical node produces for any received message. When the branching factor is 1, the branching process is said to

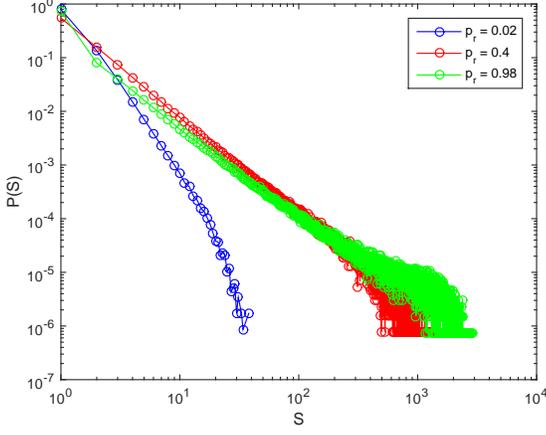
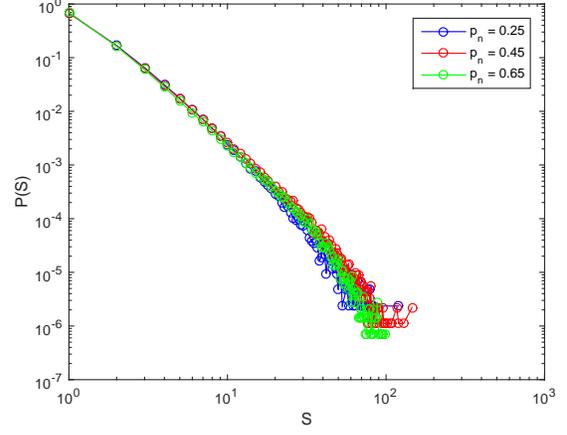
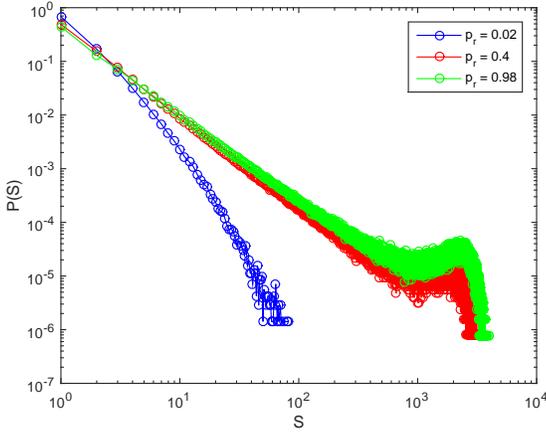
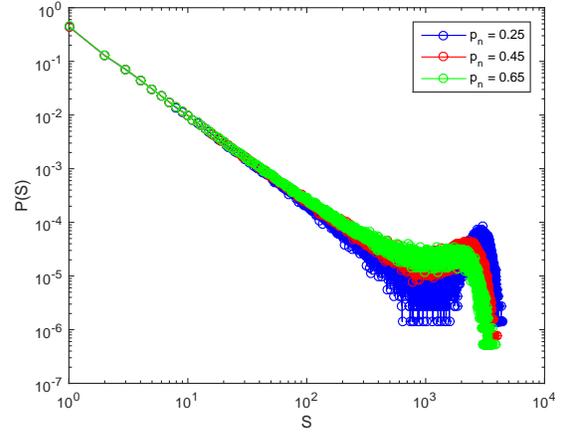
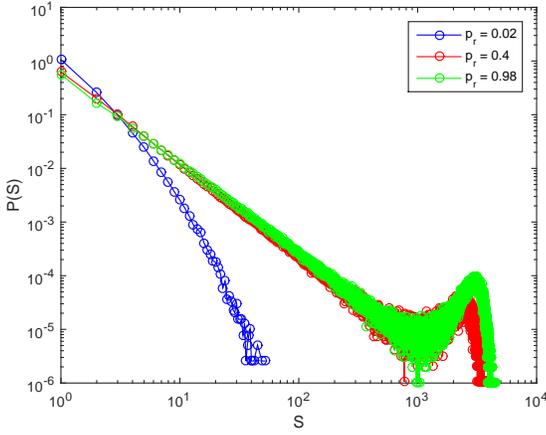
(a) $L = 10$.(a) $L = 40, p_r = 0.02$.(b) $L = 100$.(b) $L = 100, p_r = 0.98$.(c) $L = 200$.

Fig. 1: Distribution of cascade size for SF networks with, $N = 10,000, \langle k \rangle = 10, p_n = 0.45$ for various values of L and p_r .

be critical. The branching factor is estimated as follows::

- For a given message m , put all nodes that generated

Fig. 2: Impact of model parameter p_n on cascade-size distribution for SF networks with $N = 10,000, \langle k \rangle = 10$ for various values of L and p_r .

or received a copy of that message on their queue in the set R_m .

- For every r in set R_m for a specific message m , count the number of copies of the message that the node produces. If the node never forwards the message, the number of copies is $n_m(r) = 0$. For a node that does share the message, the number of copies $n_m(r)$ is equal to the node degree, if the node originates the message, or the node degree minus one if the node received the message, since the neighbor from whom the node received the message will not forward it again.
- The branching factor is estimated as $\mu = \frac{\sum_m \sum_{r \in R_m} n_m(r)}{\sum_m \sum_{r \in R_m} 1}$.

The branching factor estimated in this manner is shown for a range of p_r values for different queue lengths in Fig. 5a. The dashed horizontal line indicates

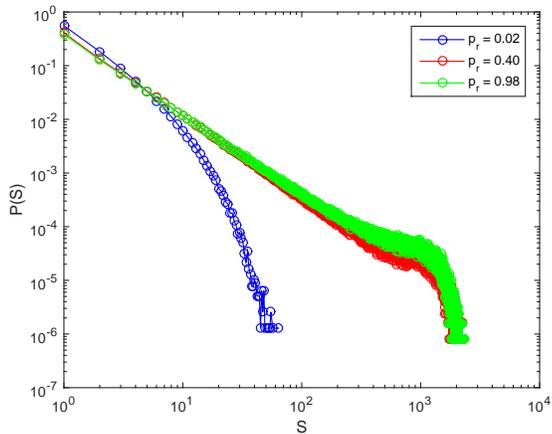
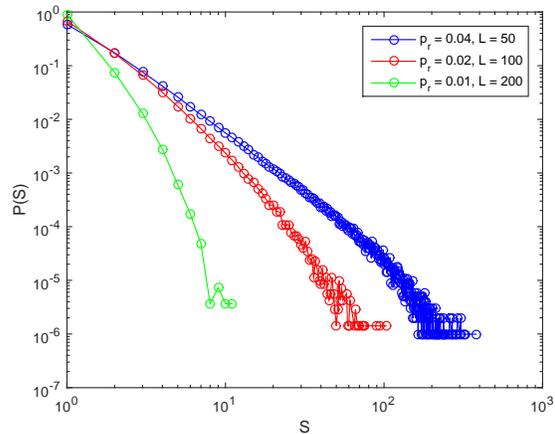
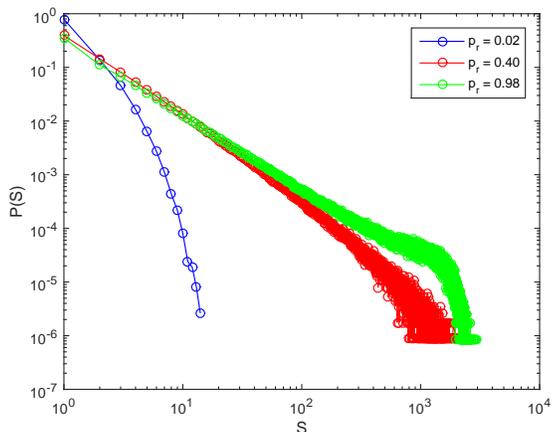
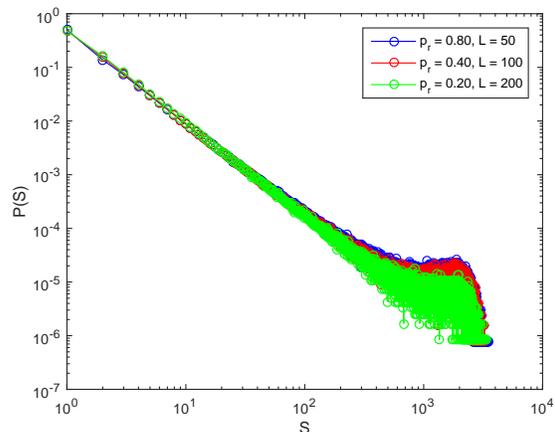
(a) $\langle k \rangle = 5$.(a) $Lp_r = 2$.(b) $\langle k \rangle = 10$.(b) $Lp_r = 40$.

Fig. 3: Distribution of cascade sizes on *ER* graphs with $N = 10,000$, $L = 100$, $p_n = 0.45$ for various values of p_r .

Fig. 4: Impact of control variable Lp_r on cascade-size distributions for *SF* networks with $N = 10,000$, $\langle k \rangle = 10$, $p_n = 0.45$.

the critical branching factor. As a general trend, as L increases, the p_r at which the cascades approach a critical branching process decreases. Also of note is that for $L = 2$, the dynamics of cascades will always operate in the subcritical regime. As L becomes very large, the system will increasingly operate in the supercritical regime. Fig. 5b shows the same analysis with simulation results superimposed to show the relative matching between the theory and simulation for three values of L .

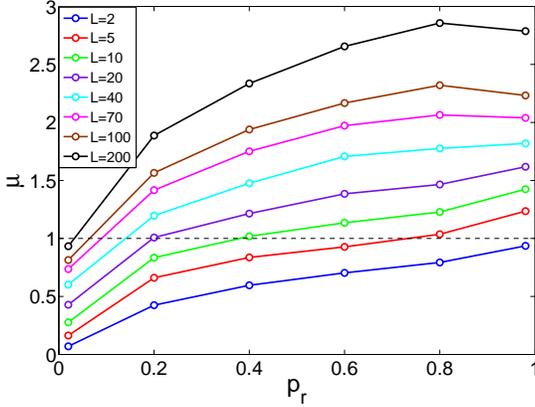
In the next section, we derive an analytical estimate of the critical value of the forwarding probability p_r at which the cascade distributions can be assumed to attain their power-law behavior thus making global cascades rare but possible.

V. ANALYSIS

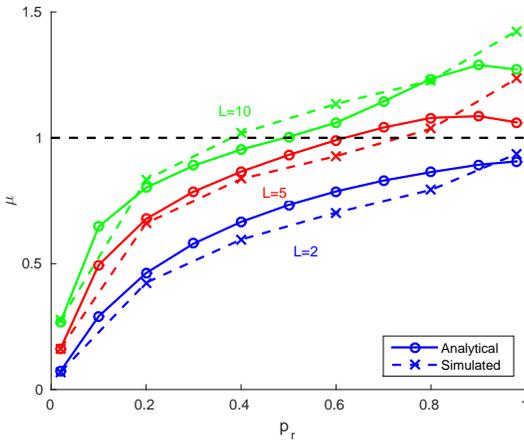
Here we describe our approach to analytically estimating the branching factor for the model given parameters

p_n , p_r and L . The basic idea is to estimate the mean number of copies of a message that a user produces given that it has received the message. For simplicity, and due to the low density of loops in randomly generated scale-free networks [20], we assume for purposes of analysis, that the structure on which the cascades are taking place is a tree with the same degree distribution as the original scale-free network. Under our model assumptions, a node forwards messages only to its children on the tree, which ensures that any message is received by a node only once.

In order to estimate the mean number of copies of a message produced by a node, we first estimate the effective forwarding probability of a received message by a node. A primary consideration here is the fact that the node can have multiple attempts at forwarding the message for as long as the message has not been



(a) Branching factor vs. Forwarding probability for various L . Horizontal line represents criticality point separating the subcritical and supercritical regimes.



(b) Comparison of simulation and analysis of branching factor vs p_r for various L .

Fig. 5: Branching factor analysis for various values of L .

pushed out of the queue by incoming messages. Thus the message gets a number of chances before exiting the user's window of attention. Let ρ_n be the probability that the message survives on the node's queue for exactly n rounds of the node's activity. Since in each round the message can be forwarded with a probability p_r , the effective forwarding probability for a given message can be written as:

$$p_f = \sum_j \rho_j (1 - (1 - p_r)^j) \quad (1)$$

since the term in parentheses on the right hand side represents the probability that the message is forwarded in j updates.

The probability ρ_n can in turn be derived from the survival probability \mathcal{P}_n which captures the probability that the user receives less than L messages in n updates thus

ensuring that the message under consideration remains on the user's queue for at least n updates. The survival probability itself can be written as:

$$\mathcal{P}_n = \sum_j q_{j,n} u(m_j < L) \quad (2)$$

where $q_{j,n}$ denotes the probability that j events are forwarded to the user during n updates of the user itself, and $u(m_j < L)$ is the probability that the number of messages received by the user as a result of these j forwarding events is less than its queue length. To obtain $q_{j,n}$, we treat the update of a node as a Poisson process with rate 1. As a result, the process describing an event among any of the k (incoming) neighbors of a node is a Poisson process with rate k . Utilizing these two independent Poisson process one obtains an expression for $q_{j,n}$,

$$q_{j,n} = \binom{n+j-1}{j} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^n \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^j \quad (3)$$

where λ_1 and λ_2 are the rates of the two interacting Poisson processes. For the results shown in the paper, if a node has k neighbors, $\lambda_1 = 1$ and $\lambda_2 = k$.

We obtain $u(m_j < L)$ by estimating the probability that a given number of messages are forwarded in j neighborhood update events accounting for the fact that the forwarded messages could be newly generated ones (occurring with probability p_n per update) or previously existing messages. In order to account for the fact that messages that have been forwarded once before are not forwarded again, we assume that each message in the queues of the neighborhood is eligible for forwarding with probability α . To complete the calculation of $u(m_j < L)$ we need to estimate α in a self-consistent manner. We do so by first deriving the probability that a message on a queue has age w i.e. has survived for w updates since it was received, and then use it to compute the average probability that a message on a queue has not been forwarded yet. The former probability can be expressed in terms of α , and the latter should be identical to α . To solve this fixed-point equation, we scan for the value of α at which the self-consistency is satisfied.

With $q_{j,n}$ and $u(m_j < L)$ evaluated, we can estimate the survival probability \mathcal{P}_n , and from the latter derive $\rho_n = \mathcal{P}_n - \mathcal{P}_{n+1}$. Finally, using ρ_n , we can estimate the effective forwarding probability of a message. The number of copies of a message produced by a node is equal to the number of descendants that the node has in a tree. Thus, given the degree distribution p_k of the network, we can derive the probability that the node produces k copies of a message that was forwarded to this node.

Following the algorithm described above, we numerically evaluate the branching factor for given values of $p_n = 0.45$, p_r and L .

As a comparison of the simulation results and the analysis, we show the critical branching factor as a function of p_r in Fig. 5b. There is close agreement for low to intermediate values of p_r . At higher values of p_r and L , the true dynamics of cascades becomes incompatible (for small L there is no divergence) with the consequences of the tree-like assumption in the analysis.

VI. TWITTER ANALYSIS

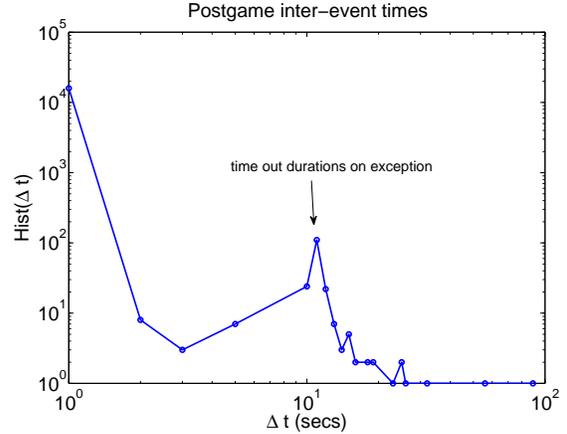
As a complement to the analytical model and simulation results, we performed an empirical study of tweets from Twitter. This study allows us to validate the model and the claim that branching factor is a predictor for virality. We gathered 3,018,944 million tweets, streaming a specific set of hashtags for the 2014 Superbowl (25 Jan 2014 to 4 Feb 2014, with the game held on 2 Feb 2014). The hashtags of interest were related to the Superbowl as well as commercial hashtags, all chosen before data collection. Per Twitter’s convention, a forwarded message is called a retweet. Additionally, two periods of tweets were extracted, pregame and postgame tweets.

A. Data Streaming Limitations

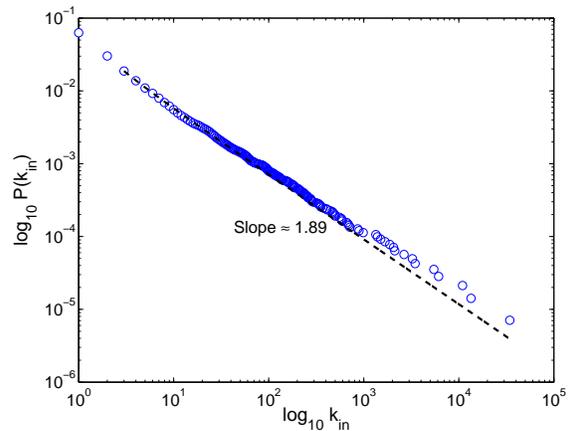
There are some inherent “rate limits” imposed on the collection of data from Twitter which make the data lossy [16]. Also there are interruptions when there is a high rate of requests, and there are “time-out” periods to prohibit getting locked out. Fig. 6a shows the time interval between successive tweets that we captured. “Time out” periods are also clearly visible. The spike in the middle of the data collection corresponds to the middle of the game in which Twitter traffic significantly increased.

B. Followership network vs. Mention network

Another limitation of the data is that we do not have network information (connectivity between users) as the gardenhose API does not provide access to this information. With our collected data, we infer the implicit “mention network” between users, by assuming that if user i mentioned user j in a tweet, then i is a follower of j (i.e. i points to j). The in-degree distribution of the inferred network is shown in Fig. 6b. In our analysis, we focus on only the (undirected) giant component.



(a) Inter-event times for the Postgame dataset.



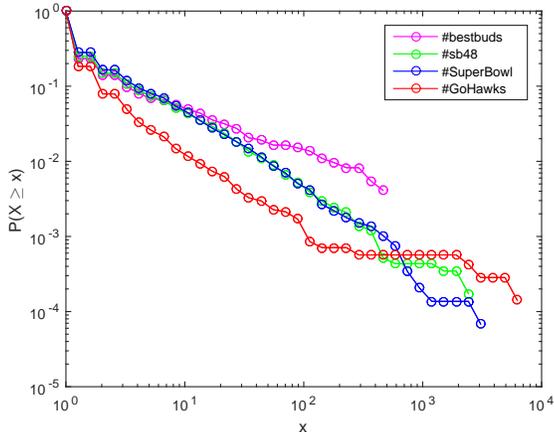
(b) In-degree distribution of mention network constructed from Postgame dataset.

Fig. 6: Postgame dataset inter-event times and in-degree distribution.

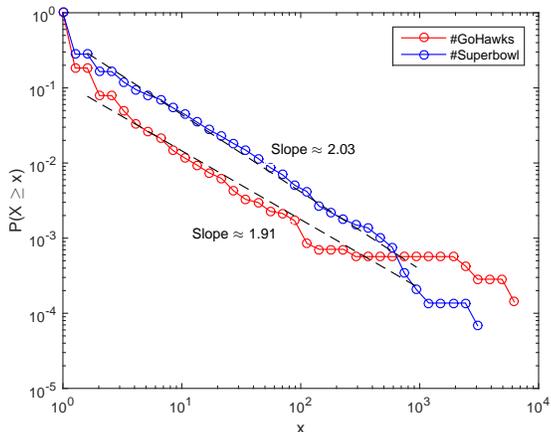
C. Cascade-size distributions

To enable the study, we measured the distribution of cascade sizes for various hashtags. Of the 31 hashtags that we tracked (see Appendix), we selected four hash-tags due to their relative abundance as compared to the rest. For each of these four hashtags, we gathered all tweets in our dataset that contained the hashtag under consideration, and none of the other hashtags that we were tracking. Within these tweets, we analyzed cascade sizes corresponding to each distinct tweet (each retweeted message has an identifier of the original message of which it is a copy). The cascade-size distributions obtained through this analysis are shown in Fig. 7a. A qualitative behavior similar to that seen in our simulations can also be observed here, although care must be taken in the comparison since we have shown the complementary cumulative distribution of cascade sizes here unlike in Figs 1a-1c. However, signatures

of subcritical, critical, and supercritical processes are observable here as well, with the hashtag *bestbuds* having a far smaller tail than those of the rest, while the hashtag *GoHawks* has a much fatter tail resembling a supercritical branching process.



(a) Cumulative cascade-size distributions of selected hashtags.



(b) Cumulative cascade-size distributions with power-law maximum likelihood fits to #GoHawks and #Superbowl.

Fig. 7: Cumulative cascade-size distributions for hashtags #bestbuds, #sb48, #Superbowl and #GoHawks.

We further investigate the cascade-size distributions associated with the two hashtags with the most data, #Superbowl and #GoHawks. A power-law distribution, as would be expected from a critical branching process, yields a good fit to the cascade-size distribution for both hashtags at lower cascade-size values, but shows a poor fit for #GoHawks at larger cascade-sizes; see Fig. 7b. Note that it is the complementary cumulative distribution function for the cascade sizes to which we have fit the power-law. As we see later, the reason for the differences in the quality of the fits is the difference in the branching factors of the two hashtags. As mentioned earlier, for large branching factors, as is the case for #GoHawks,

the process is supercritical, and therefore results in the power-law fit being poor.

D. Branching Factor Analysis

Next, we estimate the branching factor for the cascades to determine how well this metric predicts virality. Due to the limitations of data, we use the following approach:

For a given message x :

- 1) Record all nodes that have retweeted a copy of that message. Count the number of copies they generate i.e. the number of their followers (on the mention network).
- 2) We assume that the set of recorded nodes and their combined neighborhood comprise the set of nodes receiving x . Call this set the “touched” set.
- 3) Branching factor = total number of copies/ Size of touched set.

The order of hashtags according to branching factor in Table I is consistent with ordering based on the fatness of the tails of their cascade-size distributions in Figure 7b.

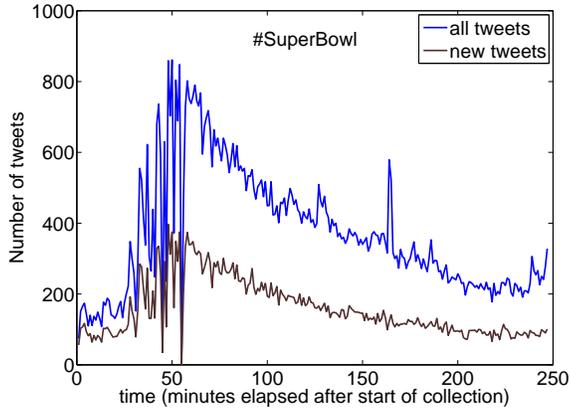
Hashtag	Estimated Branching Factor
bestbuds	1.097
Superbowl	1.756
sb48	5.310
GoHawks	12.135

TABLE I: Branching factors for several hashtags

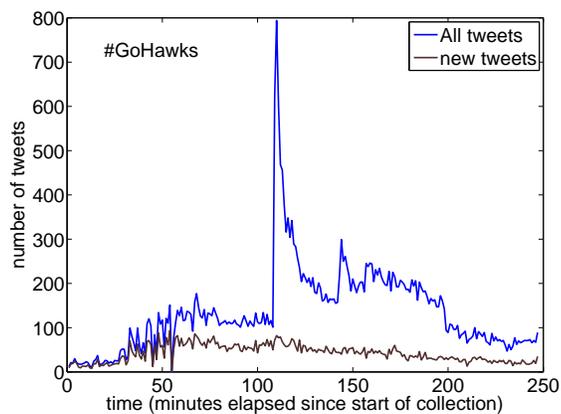
We studied tweets for #Superbowl and #GoHawks, looking at the retweet distributions over time. This is shown in Figures 8a and 8b, where the retweets are shown as a subset of the new tweets over time. Additionally, Tables III and IV (See Appendix) show an excellent match between actual sources of tweet and the first tweeter of tweet. These tables show the 10 most widely spread tweets with #SuperBowl and #GoHawks. Of note are the boldfaced rows, which indicate that the original tweet was captured in the streamed dataset.

VII. DISCUSSION

We have presented and analyzed here a parsimonious model for cascades in feed-based social network environments. Using this model, we show that the cascading behavior can fall into three regimes, corresponding to the analogous regimes of a branching process. The particular regime that the cascading behavior falls into depends on the combination of parameters in the model. What differentiates our results from prior theoretical results [2] is that supercritical behavior is possible in



(a) #SuperBowl



(b) #GoHawks

Fig. 8: Activity of new tweets and retweets for #SuperBowl and #GoHawks over time.

our model. Furthermore, empirical data obtained from Twitter also qualitatively supports our hypothesis that cascading behavior can fall into three regimes.

VIII. APPENDIX

List of hashtags gathered from Twitter

REFERENCES

- [1] L. Weng, A. Flammini, and F. Vespignani, “Competition among memes in a world with limited attention,” *Scientific Reports*, no. 2:335, March 2012.
- [2] J. P. Gleeson, J. A. Ward, K. P. O’Sullivan, and W. T. Lee, “Competition-Induced Criticality in a Model of Meme Popularity,” *Physical Review Letters*, vol. 112, no. 4, p. 048701, Jan. 2014.
- [3] T. Harris, *The Theory of Branching Processes*. Dover Publications, 1963.
- [4] L. D. Rosen, L. M. Carrier, and N. A. Cheever, “Facebook and texting made me do it: Media-induced task-switching while studying,” *Computers in Human Behavior*, vol. 29, no. 3, pp. 948 – 958, 2013.

GoodToBeBad	UpForWhatever
CupTherapy	NoRoomForBoring
FuelYourPleasure	KissForPeace
itsgotime	HowMatters
Halftime	AmericalsBeautiful
sorrycokeandpepsi	StayUncompromised
VW	Wings
sb48	Covered
Uncovered	bestbuds
salute	SuperBowl
12thman	GoHawks
SeaHawks	beastmode
tgibf	youmadbro
teasmallbiz	unitedinorange
broncospride	peyton
nfl	

TABLE II: List of hashtags collected from Twitter stream

Rank	Follower count (original)	Follower count (dataset)
1	1113326	263
2	30303	457
3	4088	631
4	20096	6
5	557775	171
6	874	681
7	481634	46
8	251500	251500
9	24777739	200
10	159544	538

TABLE III: Top retweeted #SuperBowl tweets and comparison of original tweeters and tweeters in dataset. **Bold** entries indicate that the original source was captured in the streamed dataset.

Rank	Follower count (original)	Follower count (dataset)
1	650504	650504
2	467808	20730
3	54370	5
4	898024	898204
5	108	108
6	453854	54
7	642489	232
8	244	244
9	6220	6220
10	440624	165

TABLE IV: Top retweeted #GoHawks tweets and comparison of original tweeters and tweeters in dataset. **Bold** entries indicate that the original source was captured in the streamed dataset.

- [5] R. Junco, “Too much face and not enough books: The relationship between multiple indices of facebook use and academic performance,” *Computers in Human Behavior*, vol. 28, no. 1, pp. 187 – 198, 2012.
- [6] D. Evans, *Social Media Marketing: An Hour a Day*, 2nd ed., I. John Wiley & Sons, Ed., 2012.
- [7] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer, “Quantifying the invisible audience in social networks,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. ACM, 2013, pp. 21–30.
- [8] D. J. Higham, P. Grindrod, A. V. Mantzaris, A. Otley, and P. Laflin, “Anticipating Activity in Social Media Spikes,” *ArXiv e-prints*, Jun. 2014.
- [9] Ling Feng and Yanqing Hu and Baowen Li and H. Eugene Stanley and Shlomo Havlin and Lidia A. Braunstein, “Competing for Attention in Social Media under Information Overload Conditions,” *PLoS ONE*, vol. 10, no. 7, July 2015. [Online]. Available: doi:10.1371/journal.pone.0126090
- [10] L. Weng, F. Menczer, and Y.-Y. Ahn, “Virality Prediction and Community Structure in Social Networks,” *Scientific Reports*, vol. 3, p. 2522, Aug. 2013.
- [11] —, “Predicting Successful Memes using Network and Community Structure,” *Proc. Eighth Intl AAAI Conf. Weblogs and Social media*, Mar. 2014.
- [12] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, “Epidemiological modeling of news and rumors on twitter,” in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ser. SNAKDD '13. New York, NY, USA: ACM, 2013, pp. 8:1–8:9.
- [13] S. Doan, L. Ohno-Machado, and N. Collier, “Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses,” *Proc 2012 IEEE 2nd Int'l Conf. Healthcare, Imaging and Systems Biology (HISB'12)*, Oct. 2012.
- [14] N. Oken Hodas and K. Lerman, “How Visibility and Divided Attention Constrain Social Contagion,” *Proc. 2012 AE/IEEE SOCIALCOM-PASSAT'12*, May 2012.
- [15] K. Lerman, R. Ghosh, and T. Surachawala, “Social Contagion: An Empirical Study of Information Spread on Digg and Twitter Follower Graphs,” *ArXiv e-prints*, Feb. 2012.
- [16] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, “Is the sample good enough? comparing data from twitters streaming api with twitters firehose,” in *Proc. 7th Intl AAAI Conf on Weblogs and Social Media*, 2013.
- [17] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. [Online]. Available: <http://doi.acm.org/10.1145/1772690.1772751>
- [18] K.-I. Goh, D.-S. Lee, B. Kahng, and D. Kim, “Sandpile on Scale-Free Networks,” *Physical Review Letters*, vol. 91, no. 14, p. 148701, Oct. 2003.
- [19] J. M. Beggs and N. Timme, “Being critical of criticality in the brain,” *Frontiers in Physiology*, vol. 3, no. 163, 2012. [Online]. Available: http://www.frontiersin.org/fractal_physiology/10.3389/fphys.2012.00163/abstract
- [20] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Phys Rev E*, vol. 64, no. 2, p. 026118, Aug. 2001.