A note on modeling retweet cascades on Twitter

Ashish Goel[1], Kamesh Munagala[2], Aneesh Sharma[3], and Hongyang Zhang[4]

[1] Department of Management Science and Engineering, Stanford University, ashishg@stanford.edu
[2] Department of Computer Science, Duke University, kamesh@cs.duke.edu
[3] Twitter, Inc, aneesh@twitter.com
[4] Department of Computer Science, Stanford University, hongyz@stanford.edu **

Abstract. Information cascades on social networks, such as retweet cascades on Twitter, have been often viewed as an epidemiological process, with the associated notion of *virality* to capture popular cascades that spread across the network. The notion of structural virality (or average path length) has been posited as a measure of global spread.
In this paper, we argue that this simple epidemiological view, though analytically compelling, is not the entire story. We first show empirically that the classical SIR diffusion process on the Twitter graph, even with the best possible distribution of infectiousness parameter, cannot explain the nature of observed retweet cascades on Twitter. More specifically, rather than spreading further from the source as the SIR model would predict, many cascades that have several retweets from direct followers, die out quickly beyond that.
We show that our empirical observations can be reconciled if we take *interests* of users and tweets into account. In particular, we consider a model where users have multi-dimensional interests, and connect to other users based on similarity in interests. Tweets are correspondingly labeled with interests, and propagate only in the subgraph of interested users via the SIR process. In this model, interests can be either *narrow* or *broad*, with the narrowest interest corresponding to a star graph on the interested users, with the root being the source of the tweet, and the broadest interest spanning the whole graph. We show that if tweets are generated using such a mix of interests, coupled with a varying infectiousness parameter, then we can qualitatively explain our observation that cascades die out much more quickly than is predicted by the SIR model. In the same breath, this model also explains how cascades can have large size, but low "structural virality" or average path length.

1 Introduction

Information cascades are among the most widely studied phenomena in social networks. There is a vast literature on modeling the spread of these cascades as diffusion processes, studying the kinds of diffusion trees that arise, as well as trying to predict the global spread (or *virality*) of these cascades [16, 9, 4, 11,

** This work was partly done when the author was an intern at Twitter, Inc.

12, 8]. A specific example of such a diffusion process, which is the focus of this paper, are retweet cascades on Twitter.

Extant models of information cascades build on classical epidemiological models for spread of infectious diseases [5]. The simplest of these is the SIR model, where a node in the network can be in one of three states at any time: *Susceptible* (S); *Infected* (I); and *Recovered* (R). Nodes in the network switch their states due to infections transmitted over the network, and the rate of these infections is governed by an infectiousness parameter, p. The SIR model unfolds via the following process: all nodes are initially in state S except the source (or a set of nodes called the "seed set"), which is in state I. Every node which is in state I infects each of its neighbors independently with probability p, before moving itself to state R. If a node in state S gets infected, it moves to state I. This process naturally quiesces with all nodes settling in their final state, and all nodes that were ever in state I are considered to have acquired the infection. There is a natural and trivial mapping of this model to information cascades, where the infectiousness parameter p serves to measure the *interestingness* of the piece of information, in our case, a tweet. In epidemiology, the goal is to differentiate infections that die out quickly from those that spread to the whole network; analogously, information cascades are deemed *viral* if their global reach is large.

The above view of information cascades as the spreading of content through the network is intuitively and analytically appealing. In fact, Goel *et al.* show that when simulated on a scale-free graph, the SIR model statistically mimics important properties of retweet cascades on Twitter. In particular, they use *structural virality*, or average path length in the diffusion tree, as a quantitative measure of "infectiousness" of a cascade, and show that the distribution of cascade sizes (number of users that retweet a tweet plus the author of the tweet) and structural virality are statistically similar to that from the simulations. On the other hand, these empirical studies also show that cascades observed in Twitter are mostly shallow and exceedingly rare: Goel *et al.* [7] show there are no viral cascades in a corpus of a million tweets; and in subsequent work [6], show that viral cascades do indeed exist if the corpus size is increased to a billion tweets. This data contrasts with the observation that social networks like Twitter have a power-law degree distribution [13], and these networks should have low epidemic threshold, so that even with low infectiousness parameter p, most cascades should be viral [2, 1]. Therefore, explaining the low frequency of viral events on Twitter via an SIR model requires that the infectiousness parameter be quite low almost all the time. Finally, this result also begs the question of whether modeling viral events if even of any interest if these events are so rare.

We therefore ask: *Is there something fundamental about real-world information cascades, particularly those on Twitter, that is not captured by the simple SIR model?* Though this question is about a specific social network, and a specific (simplistic) epidemiological model, even understanding this via suitably designed experiments is challenging, and has not been performed before.

1.1 Our Contributions

In the process of answering the above question, we make the following contributions.

Evaluating Epidemic Models Through Twitter Network Our main contribution is to show that the SIR model is a *poor* fit for information flow on Twitter. We show this by empirically testing the hypothesis that retweet cascades on Twitter propagate using the SIR process. Our null hypothesis is that each cascade has an underlying infectiousness p (that could be different for different cascades), and conditioned on receiving the tweet, a user retweets it with probability p. We compare the value of p that we obtain by best-fit for the users directly connected to the source of the tweet (level 1 followers), and those who receive the tweet from a direct follower of the source (level 2 followers). Using a corpus of 8 million cascades, we develop a statistical test to show that these two values of p are different – the second level value is significantly smaller than the first. The technically interesting part of this analysis is the fact that most cascades are shallow. Thus, many tweets generate very few retweets at the first level, and this number dictates the number of tweet impressions and retweets at the second level. The SIR model therefore corresponds to a stochastic process for the retweets that has very low mean but potentially very high variance because of the skewed degree distribution of the graph. We have to therefore devise a statistical test that works around this high variance. Apart from this statistical test, at a coarse level, we find that the median value of first level infection probability is 0.00046, while the median value of second level infection probability is 0 (in other words, half of the tweets do not have second level retweets!). Even among the tweets that have at least 1000 impressions at the first level, more than 80% of them, have that first level p is at least twice the second level p. This suggests that, rather than spreading further from the source, a cascade typically dies out quickly within a few hops.[5] This echoes with the observation that most of the cascades tend to be star-like trees [16]. It also suggests an explanation for truly viral cascades being so rare [6].

Interest-based SIR Model Since the SIR model assumption of fixed propagation probability per cascade is statistically violated on Twitter, we propose an alternative model for retweet cascades. In particular, we present a tweet propagation model that takes *interests* of users and tweets into account. In order to do this, we revisit a Kronecker graph-based model for social networks first considered in [3]. In this attribute based model, users have attribute vectors in some d-dimensions, and interests are specified by a subset of these dimensions along with their attribute values. If fewer dimensions are specified, these interests are *broad* and encompass many users; if many dimensions are specified, these interests are *narrow* with a shallow component around the source. Tweets are also correspondingly labeled with interests, and propagate only in the *subgraph of*

[5] Indeed, the median of first level impressions is 175, while the median of second level impressions is 29!

interested users via a SIR process with infectiousness drawn from a distribution. We show that if tweets are generated using such a mix of narrow and broad interests, then this coupled with a varying infectiousness parameter can qualitatively explain the level-one infectiousness being larger than subsequent levels. As a simple intuition, observe that cascades corresponding to narrow interests only reside in their shallow subgraphs, while those corresponding to broad interests can be "viral" in the usual sense.

As mentioned above, Goel *et al.* [6] define the notion of *structural virality*, or average path length of a cascade as a measure of its virality. They show that this measure is uncorrelated with the size of the cascade, except when structural virality is large. The proposed explanation in their work is an SIR model on a scale-free graph with extremely low infectiousness parameter. Our model leads to a different explanation: cascades corresponding to narrow interests have low structural virality, but can have large size. This explanation does not depend on any specific setting of the infectiousness parameter, and is therefore of independent interest. Finally, we show that cascades arising for broad interests can have large structural virality, but our model would predict a large expected size as well, which again matches previous empirical findings.

1.2 Related Work

Epidemic models on social networks have received a lot of attention in the past decade, and we won't attempt to review the large literature here. Instead, we point the reader to a small set of representative papers and the excellent survey articles and books on the topic [11, 5, 12, 14, 9, 4]. Despite all the attention on studying diffusion, there has been relatively little work evaluating epidemic models on social networks such as Twitter [14, 19, 6]. In particular, we believe that the empirical testing of structural properties of cascades on the Twitter graph (as opposed to a specific generative model) is unique to our work.

Part of the reason, as has been pointed out in [6], is that only recently have large datasets of information contents become available. In the same work, the authors defined the notion of structural virality and observe that it is very rare to observe structurally viral cascades, but they can find these rare cascades by obtaining a large collection of tweets. By carefully choosing the infectiousness parameter of the SIR model on a power law network, they are able to reproduce many empirical statistics of the observed cascades distribution, such as the probability that a piece of content gains at least 100 adopters, and the mean structural virality. However, they also point out that other important statistics does not match with the empirical distribution. For example, the variance is much smaller in the simulated model, compared to the empirical distribution. We present an alternative interest-based model for explaining the same phenomena, while comprehensively refuting the SIR hypothesis.

Similarly, Leskovec et al. [16] were able to fit cascade sizes and degree distributions of a large collection of blogs, with the SIS model defined by an infectiousness parameter. We also want to mention a study of user adoption on Facebook, Ugander *et al.* [19] find that the probability of users joining Facebook

is dependent on the number of connected components in an user's ego network (or neighborhood graph), rather than by the size of the ego network. Note that this work studied user adoption rather than content diffusion, but the observation that sub-structures in the network can dominate network size for adoption is in general agreement with our proposed model.

2 Evaluating the SIR Model on the Twitter Network

In this section, we describe our evaluation of the simple SIR model on eight million retweet cascades observed on Twitter. These retweet cascades are collected from a single week and each cascade is restricted to be started by a user based in the US. In our analysis, we have excluded tweets posted by Twitter accounts that are likely to be spammers using an internal quality detection tool.[6] For each tweet, we collect the information described in Table 1. Note that we use the number of followers of a user as a proxy for the number of impressions of the user's tweets. While we could also count impressions directly on Twitter, this would not correctly represent the significant fraction of users that visit Twitter through third-party clients. All the information described in Table 1 could be collected through the public Twitter APIs. [7] While we used Twitter's internal spam detection mechanism to filter away potential spam users, we believe that exploiting well-known features (for example pagerank values) would also achieve the same results for our task.

2.1 Defining the Null Hypothesis

Let us fix a given set of tweets T. For each tweet $t \in T$, let $p_1(t)$ and $p_2(t)$ denote the underlying retweet rate at the first level and second level of the Twitter graph, respectively. Note that these parameters are fixed but unknown for any given tweet. The dependence of p_1 and p_2 on t models the fact that different tweets can have different infectiousness. Our null hypothesis is that $p_1(t) = p_2(t)$ for all $t \in T$, which corresponds to cascade propagation via the simple SIR model. A different, but equivalent view of the null hypothesis is that it posits $p_1(t)$ is drawn from some distribution, and conditioned on this, we set $p_2(t) = p_1(t)$.

The stochastic process, given a tweet t and corresponding underlying $p_1(t)$ and $p_2(t)$ unfolds as follows (we omit t for notational convenience): let the value v_1 be a (non-random) parameter associated with the tweet source. Then $r_1 \sim B(v_1, p_1)$ is a Binomial random variable with parameters v_1 and the unknown p_1. We will *assume* that v_2 (th) is nonzero whenever r_1 is nonzero. Since v_2 is defined as the total number of followers among those who retweet the source tweet, if this value is zero with r_1 being non-zero, then the source user is very likely to be a spammer. However, since we eliminated spam sources in our filtering step,

[6] A lot of spam tweets have star-like cascade structure that may significantly impact the experiment results while not representing general user behavior.

[7] https://dev.twitter.com/streaming/public

this event is very unlikely in our dataset. Now, r_2 is a random variable that is generated according to $B(v_2, p_2)$. Note that we are modelling r_2 as a Binomial random variable, since it is easier to present than the SIR process. As a matter of fact, there is no difference to our conclusions if r_2 is generated according to the SIR process. The reason for that is Lemma 1 will continue to hold under the SIR process. We observe a realization of the random variables, v_2, r_1, and r_2.

v_1	Number of followers of the source node (the size of $N_1(\tau)$)
r_1	Number of retweets among the set of nodes $N_1(\tau)$ (the size of $R_1(\tau)$)
v_2	Number of nodes that follow any nodes in $R_1(\tau)$ (the size of $N_2(\tau)$)
r_2	Number of retweets among the set of nodes in $N_2(\tau)$

Table 1. A list of observed information for a tweet τ, posted by a node s. Let $N_1(\tau)$ denote the set of nodes that follow the node s. Let $R_1(\tau)$ denote the subset of nodes among $N_1(\tau)$ that retweet the tweet τ. And let $N_2(\tau)$ denote the set of nodes that follow any nodes in $R_1(\tau)$.

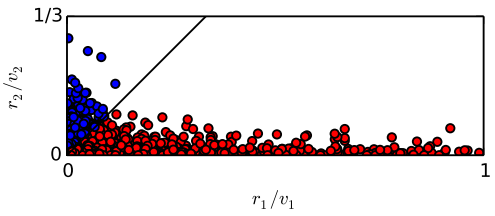

Fig. 1. A scatter plot of ten thousand sampled tweets. The y-axis has been truncated since there are no points beyond 1/3 in the samples.

2.2 Refuting the SIR Model

We will now refute the null hypothesis, *i.e.*, show that $p_1(t) > p_2(t)$ for almost all $t \in T$. Observe that if $r_1(t)$ and $r_2(t)$ are sufficiently large, then by standard concentration bounds, $\frac{r_1(t)}{v_1(t)}$ will be a good approximation to $p_1(t)$, and likewise for $p_2(t)$. A natural approach is therefore to compare the empirical average of $\frac{r_1(t)}{v_1(t)}$ over $t \in T$ to the empirical average of $\frac{r_2(t)}{v_2(t)}$. If these are different, that would refute $p_1(t) = p_2(t)$ for all $t \in T$. In Figure 2.1, we plot these empirical values, and this provides some evidence that the null hypothesis is false. However, this approach is not quite statistically rigorous.

Specifically, the problem with this approach is that when $r_1(t)$ is zero, then $v_2(t)$ is zero and $p_2(t)$ remains undefined. However, if we filter away any tweet

whose $r_1(t) = 0$, then we could potentially bias the estimation of $p_1(t)$ as well. To overcome this issue, we will correct the bias by subtracting a corresponding factor in $\frac{r_1(t)}{v_1(t)}$.

In the lemmas and definitions below, the expectation is over the stochastic process described above, where v_2, r_1, r_2 are random variables. For each tweet $t \in T$ we define the following random variables:

$$X_2(t) = \begin{cases} r_2(t)/v_2(t) \text{ if } v_2(t) > 0 \\ 0 \text{ if } v_2(t) = 0 \end{cases} \tag{1}$$

$$X_1(t) = r_1(t)/v_1(t) - f_0(t) \tag{2}$$

where $f_0(t) = (\frac{v_1(t)}{v_1(t)+1})^{v_1(t)+1}/v_1(t)$.

Lemma 1. *Under the null hypothesis that $p_1(t) = p_2(t)$, we have $\mathbb{E}\, X_2(t) \geq \mathbb{E}\, X_1(t)$, for any $t \in T$.*

Proof. Note that

$$\mathbb{E}\, X_2(t) = p(t) \Pr(v_2(t) \neq 0) = p(t) \Pr(r_1(t) \neq 0),$$

by our assumption that $v_2(t) = 0$ if and only if $r_1(t) = 0$. Further,

$$\mathbb{E}\, X_1(t) = p(t) - f_0(t)$$

The conclusion follows since:

$$p(t) \Pr(r_1(t) = 0) = p(t) \times (1 - p(t))^{v_1(t)} \leq f_0(t).$$

where the last inequality is obtained by observing the maximum value of the function $p(t) \times (1 - p(t))^{v_1(t)}$ of $p(t)$.

For any subset T of tweets, let $\chi_1 = \sum_{t \in T} X_1(t)$ and $\chi_2 = \sum_{t \in T} X_2(t)$. We compute the observed values of χ_1 and χ_2 for several different buckets of tweets T, grouped by ranges over number of first level impressions. These buckets are shown in Table 2. Based on the second and third columns, we conclude that the average observed X_2 is less than the average observed X_1, thereby contradicting the null hypothesis.

Now we examine the significance of the above finding. The idea is that since both χ_1 and χ_2 are sums of independent random variables in the range $[0, 1]$, the observed values should be concentrated around the mean value. While we don't know the mean values, $\mathbb{E}\, \chi_1$ and $\mathbb{E}\, \chi_2$, we can obtain an upper bound of the desired probability, by maximizing over all possible values of $\mathbb{E}\, \chi_1$ and $\mathbb{E}\, \chi_2$, subject to the null hypothesis, Lemma 1. This is summarized in the following Lemma:

Lemma 2. *For a set of tweets T with observed values of $\chi_1 \geq \chi_2$, the probability that such an observation could happen under the null hypothesis, $p_1(t) = p_2(t)$ for all $t \in T$, can be upper bounded by:*

$$2 \exp(-\frac{2\sqrt{2(\chi_1^2 + \chi_2^2)} - 2\chi_1 - 2\chi_2}{3}).$$

v_1	number of tweets	χ_1	χ_2	p-value
$(0, \infty)$	3766k	3017	836	0.0
(100, 1000)	359k	690	109	10^{-100}
(1000, 10000)	2133k	1830	531	10^{-150}
(10000, ∞)	1274k	477	195	10^{-30}

Table 2. Experimental results for several different buckets of tweets. See main text for more details.

Proof. Let $t_1 = \mathbb{E}\,\chi_1$ and $t_2 = \mathbb{E}\,\chi_2$. By Chernoff bound (cf Corollary 4.6 [18]),

$$\Pr(|\chi_1 - t_1| \geq \delta_1 t_1) \leq 2\exp(-t_1\delta_1^2)/3$$
$$\Pr(|\chi_2 - t_2| \geq \delta_2 t_2) \leq 2\exp(-t_2\delta_2^2)/3$$

Hence

$$\begin{aligned}
&\max_{t_2 \geq t_1 > 0} \Pr(|\chi_1 - t_1| \geq \delta_1 t_1, |\chi_2 - t_2| \geq \delta_2 t_2) \\
\leq &\max_{t_2 \geq t_1 > 0} 2\exp(-(t_1\delta_1^2 + t_2\delta_2^2)/3) \\
= &\max_{t_2 \geq t_1 > 0} 2\exp(-(\frac{\chi_1^2}{t_1} + t_1 + \frac{\chi_2^2}{t_2} + t_2 - 2\chi_1 - 2\chi_2)/3) \qquad (3)
\end{aligned}$$

Consider two cases,

1. if $t_2 \leq \chi_1$, then we know that $\frac{\chi_1^2}{t_1} + t_1 \geq \frac{\chi_1^2}{t_2} + t_2$, and (3) can be upper bounded by
$$2\exp(-\frac{2\sqrt{2(\chi_1^2 + \chi_2^2)} - 2\chi_1 - 2\chi_2}{3})$$
when $t_2 = t_1 = \sqrt{\frac{\chi_1^2 + \chi_2^2}{2}}$.
2. if $t_2 > \chi_1$, then we know that $\frac{\chi_1^2}{t_1} + t_1 \geq 2\chi_1$, and $\frac{\chi_1^2}{t_1} + t_1 \geq \frac{\chi_1^2}{\chi_1} + \chi_1$. Then (3) can be upper bounded by
$$2\exp(-(\frac{\chi_1^2}{\chi_1} + \chi_1 - 2\chi_2)/3)$$
when $t_1 = t_2 = \chi_1$. And it's not hard to check that this is smaller than the bound obtained in the first case.

We compute these probabilities and show them as p-values in Table 2. This shows that the observed χ_1, χ_2 are highly unlikely under the null hypothesis.

Finally, we note that the above analysis does not necessarily show that $p_1(t) > p_2(t)$ for almost all t in our corpus. To address this concern, we randomly sample 1% of the tweets, run the same analysis, and repeat for 10000 times. Figure 2.2 plots the histogram of p-values that we obtain. Since we observed consistently low p-values among all the samples, this shows that the null hypothesis of $p_1(t) = p_2(t)$ for all $t \in T$ is very unlikely to hold in our dataset. In fact, our analysis shows that $p_1(t)$ is almost always bigger than $p_2(t)$.

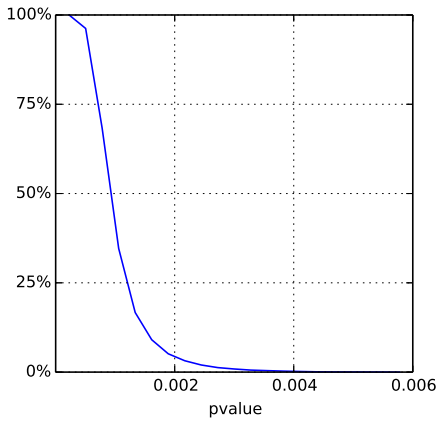

Fig. 2. The histogram of p-values obtained from 10000 random subset of tweets. Each random subset contains 1% of all tweets.

3 An Interest Based Model for Tweet Propagation

We now show that the above empirical observations are consistent with the following model: Users have interests and connect to other users based on similarities in interests. Each tweet corresponds to an interest (either a broad interest or a narrow interest) and is retweeted only by users with the corresponding interest. We formalize this model below, and show how it can qualitatively explain our observations.

We adapt the Kronecker interest model formulated in [3]. This is based on the Kronecker social graph, which has been studied as a reasonable theoretical model for social networks [15, 17]. We note that some of the model assumptions below are not an exact fit for social networks; nevertheless, this model captures most high-level statistical properties observed in reality, in addition to being easy to interpret. In our model, parametrized by a small number K, there are $|V| = n$ users, and $d = \log_K n$ attributes, each with K possible values from the set $S = \{a_1, a_2, \ldots, a_K\}$. Each node $u \in V$ maps to a d-dimensional vector of attribute values $(u_1, u_2, \ldots, u_d)$, where each $u_i \in S$. Therefore, $|V| = K^d = n$. Treat the values in S as the K vertices of an undirected *seed graph* G_0, and denote the adjacency matrix of this graph as A. Assume $A[a_s, a_s] = 1$ for $1 \leq s \leq K$.

For each $u = (u_1, u_2, \ldots, u_d)$ and $v = (v_1, v_2, \ldots, v_d)$, the edge (u, v) exists iff $A[u_j, v_j] = 1$ for all $j = 1, 2, \ldots, d$. We define an interest as a set of pairs of attribute dimensions and their values, where a generic interest $i \in I$ has the following form:

$$i = \{\langle j_1, a_{j_1}\rangle, \langle j_2, a_{j_2}\rangle, \ldots, \langle j_r, a_{j_r}\rangle\} \quad \text{where} \quad j_1, j_2, \ldots, j_r \leq K \text{ and } r \leq d$$

The *consumers* of this interest are defined as:

$$C_i = \{u = (u_1, u_2, \ldots, u_d) \mid A[u_j, a_j] = 1 \ \ \forall \langle j, a_j \rangle \in i\}$$

Similarly, the producers of this interest are defined as:

$$P_i = \{u = (u_1, u_2, \ldots, u_d) \mid u_j = a_j \ \ \forall \langle j, a_j \rangle \in i\}$$

The above interest model has the following interpretation. Since each interest is specified by a subset of attributes along with their values, the graph G_0 and adjacency matrix A specify which interests are related, i.e. which interests specify an *interested in* relationship. We classify interests are *narrow* or *broad*. The narrowest interests have $|i| = d$, and the broadest interest has $|i| = 0$. Further, these interests have a natural hierarchical structure, where the *broader* interests are those specified by fewer attributes. Also note that a producer of an interest needs to align with it's attribute values on all the relevant attribute dimensions, while a consumer of an interest only needs to be *interested in* those attribute values in the relevant attribute dimensions.

We parametrize the tweet propagation process by two distributions: There is an interest distribution F and a SIR parameter distribution G. We choose an interest i at random from distribution F; choose a producer u uniformly at random from P_i, and choose an infectiousness p at random from G. The tweet originates at u, and propagates using the SIR model with parameter p on the subgraph induced by C_i.

We now perform some calculations to understand the behavior of this process for various interest sizes. In order to simplify these calculations, we assume G_0 is regular with degree w, and denote $A = w^d$ as the degree of each user. We assume $A \gg w$. Note that G_0 has K vertices, so $w \leq K$. We denote $d - |i| = s$ as the *size* of the interest. We further assume that the infectiousness parameter p is small so that $wp \ll 1$; on the other hand, we assume it is large enough that $Ap \gg 1$. We note that these assumptions are only to derive simple formulas that can be qualitatively interpreted. We need to use more nuanced parameter settings to model real social networks, but these will not affect the high-level qualitative nature of our conclusions.

Narrowest Interests, $s = 0$ In this case, $|P_i| = 1$, so that there is one user u who is a potential producer. This user is directly connected to all users in C_i. Therefore, for any p, the size of the cascade is Ap, and the structural virality is exactly 2.

Narrow Interest, $s = 1$ In this case, $|P_i| = K$, and these producers are connected as G_0. Assume all these producers have the first $d - 1$ coordinates of their attribute vector fixed to one value, and the final coordinate taking one of K possible values. The consumers C_i are all the neighbors of P_i. For small enough p, let $wp = \delta \in (0, 1)$. Then we approximately have Size $= \frac{A}{w}\delta(1 + \delta)$, and $SV = 2 + \frac{\delta}{2}$. In this case, though structural virality grows very slowly with size, a large structural virality implies a large size but not necessarily the other way around.

Broad Interest, $s = d$ In this case, $|P_i| = n$. Assuming $Ap \gg 1$, the expected size of the cascade is $(Ap)^h$, where $h = \log_A n$ is the depth of the process. The structural virality is $2h$ regardless of p. Therefore, for broad interests with moderate infectiousness p, we expect a high value of structural virality, and a correspondingly high value of size. Therefore, in our model, a high value of structural virality corresponds to a broader interest, and these cascades also have large size.

4 Conclusion

In this paper, we performed an empirical examination of the SIR epidemic model on a large selection of retweet cascades on Twitter. The experimental results refute the null hypothesis, and show that the SIR model does not fit the empirical observations. This is because retweet rates decrease as a cascade propagates further from the source, contradicting the fixed probability per cascade assumption in the SIR model. We also proposed an alternative interest-based diffusion model, where users retweet based on overlapping interests with a tweet. It is an interesting future challenge to empirically test the interest-based diffusion model. Indeed, in preliminary experiments we often found that structurally viral cascades correspond to "broad" topics that also have a very large size. In particular, we tweets containing jokes, appeals for finding a lost person, and "not safe for work" (NSFW) content are common among large structurally viral retweet cascades. On the other hand, tweets that correspond to "narrow" topics (niche sports and other topical content) usually have small structural virality. We leave it to future work to validate these observations on a large scale.

We also emphasize that our work is specific to the flow on information in social networks such as Twitter, and on fitting the simple SIR model (with possibly different levels of infectiousness or interestingness for different tweets) to it. We view this work as one further step towards validating simple models for information spreading. Given the format of retweets on Twitter where multiple retweets to a user can be suppressed, we have not considered threshold models (such as in [10]) that are based on a user receiving multiple copies of the message from different sources. We note that such threshold models have been extensively investigated in other diffusion contexts such as adoption of new technologies, and are likely appropriate for spread of information cascades in other social media. This makes it a good topic for future investigation. We also note that the interest-based model, coupled with SIR on the appropriate interest subgraph, is only one possible explanation for our observations. It is an interesting research direction to see if there are other possible explanations, such as local structure in networks, epidemic thresholds, etc that can be empirically validated. Finally, an interesting direction is to explore alternative notions of virality other than structural virality. In particular, is there a way to capture "viral" events that are specific to a group of friends, or inside a community? We believe that understanding these questions will also provide new insights for content recommendation and targeting on social networks.

Acknowledgment. We are grateful to the anonymous reviewers for very helpful feedbacks. Goel and Zhang are supported by DARPA GRAPHS program via grant FA9550-12-1-0411. Munagala is supported in part by NSF grants CCF-1348696, CCF-1408784, and IIS-1447554, and by grant W911NF-14-1-0366 from the Army Research Office (ARO).

References

1. N. Berger, C. Borgs, J. T. Chayes, and A. Saberi. On the spread of viruses on the internet. In *Proc. ACM-SIAM Symp.on Discrete Algorithms (SODA)*, pages 301–310. Society for Industrial and Applied Mathematics, 2005.
2. M. Boguná, R. Pastor-Satorras, and A. Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters*, 90(2), 2003.
3. R. Bosagh Zadeh, A. Goel, K. Munagala, and A. Sharma. On the precision of social and information networks. In *Proc. ACM Conf. Online Social Networks (COSN)*, pages 63–74, 2013.
4. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. 23rd World Wide Web Conference (WWW)*, pages 925–936, 2014.
5. D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.
6. S. Goel, A. Anderson, J. Hofman, and D. Watts. The structural virality of online diffusion. *Management Science*, 2015.
7. S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proc. ACM EC*, pages 623–638, 2012.
8. B. Golub and M. O Jackson. How homophily affects diffusion and learning in networks. *The Quarterly Journal of Economics*, 2012.
9. M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proc. SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 1019–1028, 2010.
10. D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 137–146, 2003.
11. D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.
12. J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In *Algorithmic game theory*. Cambridge University Press UK, 2007.
13. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
14. J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.
15. J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.
16. J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Symp. Data Mining (SDM)*, volume 7, pages 551–556, 2007.

17. M. Mahdian and Y. Xu. Stochastic kronecker graphs. *Random Struct. Algorithms*, 38(4):453–466, 2011.
18. M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.
19. J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proc. National Academy of Sciences (PNAS)*, 109(16):5962–5966, 2012.